Conferences in Research and Practice in
Information Technology

Volume 61

Data Mining and Analytics 2006

Data Mining and Analytics 2006

Proceedings of the
Fifth Australasian Data Mining Conference (AusDM2006),
Sydney, Australia, 29-30 November, 2006

Peter Christen, Paul J. Kennedy, Jiuyong Li,
Simeon J. Simoff and Graham J. Williams , Eds.

Volume 61 in the Conferences in Research and Practice in Information Technology Series.
Published by the Australian Computer Society Inc.

Published in association with the ACM Digital Library.

Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia, 29-30 November, 2006

Conferences in Research and Practice in Information Technology, Volume 61.

Editors: **Peter Christen**
Department of Computer Science
Faculty of Engineering and Information Technology
The Australian National University
Canberra ACT 0200 Australia
Email: peter.christen@anu.edu.au

Paul J. Kennedy
Faculty of Information Technology
University of Technology, Sydney
Broadway, NSW, 2007, Australia
Email: paulk@it.uts.edu.au

Jiuyong Li
Department of Mathematics and Computing
The University of Southern Queensland
Toowoomba, QLD, 4350, Australia
Email: jiuyong@usq.edu.au

Simeon J. Simoff
Faculty of Information Technology
University of Technology, Sydney
Broadway, NSW, 2007, Australia
Email: simeon@it.uts.edu.au

Graham J. Williams
Australian Taxation Office
2 Constitution Avenue
Canberra ACT 2601, Australia
Email: Graham.Williams@togaware.com

Series Editors:
Vladimir Estivill-Castro, Griffith University, Queensland
John F. Roddick, Flinders University, South Australia
Simeon Simoff, University of Technology, Sydney, NSW
`crpit@infoeng.flinders.edu.au`

Publisher: Australian Computer Society Inc.
PO Box Q534, QVB Post Office
Sydney 1230
New South Wales
Australia.

Conferences in Research and Practice in Information Technology, Volume 61.
ISSN 1445-1336.
ISBN 1-920-68242-2.

Printed, November 2006 by UTS Printing Services, Sydney, NSW.
Cover Design by Modern Planet Design, (08) 8340 1361.

The *Conferences in Research and Practice in Information Technology* series aims to disseminate the results of peer-reviewed research in all areas of Information Technology. Further details can be found at http://crpit.com/.

Table of Contents

Algorithms

Financial Data Mining

Web Mining

Preface

The Australasian Data Mining Conference series **AusDM**, initiated in 2002, is the annual flagship venue where data mining and analytics professionals - scholars and practitioners, can present the state-of-art in the field. Together with the Institute of Analytics Professionals of Australia, **AusDM** has built a unique profile in nurturing this joint community. The first and second edition of the conference (held in 2002 and 2003 in Canberra, Australia) facilitated the links between different research groups in Australia and some industry practitioners. This year the event has been supported by:

- Togaware, again hosting the website and the conference management system, coordinating the review process and other essential expertise;
- the University of Technology, Sydney, for providing the venue, registration facilities and various other support at the Faculty of Information Technology;
- the Institute of Analytic Professionals of Australia (IAPA) and NetMap Analytics Pty Limited, for facilitating the contacts with the industry;
- the ARC Research Network on Data Mining and Knowledge Discovery, for providing financial support;
- the e-Markets Research Group, for providing essential expertise for the event;
- the Australian Computer Society, for publishing the conference proceedings.

This year the conference has grown in size and for the first time has run parallel sessions. The conference program committee reviewed 58 submissions, out of which 25 submissions have been selected for publication and presentation. **AusDM** follows a rigid blind peer-review process and ranking-based paper selection process. All papers were extensively reviewed by at least three referees drawn from the program committee. We would like to note that the cut-off threshold has been high (4.3 on a 5 point scale) and is higher than for last year. This is testament to the high quality of submissions. We would like to thank all those who submitted their work to the conference. We will continue to extend the conference format to be able to accommodate more presentations.

Data mining and analytics today have advanced rapidly from the early days of pattern finding in commercial databases. They are now a core part of business intelligence and inform decision making in many areas of human endeavour including science, business, health care and security. Mining of unstructured text, semi-structured web information and multimedia data have continued to receive attention, as have professional challenges to using data mining in industry. Accepted submissions have been grouped into seven sessions reflecting these application areas. Four invited industry keynote sessions put the research into context.

Special thanks go to the program committee members. The final quality of selected papers depends on their efforts. The **AusDM** review cycle runs on a very tight schedule and we would like to thank all reviewers for their commitment and professionalism.

Peter Christen, Paul J. Kennedy, Jiuyong Li,
Simeon J. Simoff and Graham J. Williams

AusDM06 Organising Chairs
November 2006

Organising Committee

Conference Chairs

Simeon J Simoff, University of Technology, Sydney
Graham J Williams, Australian Taxation Office and University of Canberra

General Chairs

Peter Christen, Australian National University
Paul Kennedy, University of Technology, Sydney
Jiuyong Li, University of Southern Queensland

Industry Chairs

Eugene Dubossarsky, Ernst-Young
John Galloway, NetMap Analytics
Inna Kolyshkina, Price-Waterhouse Coopers

Programme Committee

Hussein Abbass, University of NSW / Australian Defence Force Academy
Rohan Baxter, Australian Taxation Office, Canberra, Australia
Helmut Berger, University of Technology, Sydney, Australia
Vladimir Estivill-Castro, Griffith University, Queensland, Australia
Mohamed Gaber, Monash University, Melbourne, Australia
Raj Gopalan, Curtin University of Technology, Perth, Australia
Warwick Graco, Australian Taxation Office, Canberra, Australia
Lifang Gu, Australian Taxation Office, Canberra, Australia
Simon Hawkins, University of Canberra, Australia
Hongxing He, CSIRO, Canberra, Australia
Markus Hegland, Australian National University, Canberra, Australia
Robert Hilderman, University of Regina, Canada
Joshua Huang, University of Hong Kong, Hong Kong
Warren Jin, CSIRO, Canberra, Australia
Rao Kotagiri, University of Melbourne, Australia
John Maindonald, Australian National University, Canberra, Australia
Arturas Mazeika, Free University of Bozen, Italy
Damien McAullay, CSIRO, Canberra, Australia
Richi Nayak, Queensland University of Technology, Australia
Christine O'Keefe, CSIRO, Canberra, Australia
Kok-Leong Ong, Deakin University, Victoria, Australia
Mehmet Orgun, Macquarie University, Sydney, Australia
Jon Patrick, University of Sydney, Sydney, Australia
François Poulet, ESIEA, Laval, France
Richard Price, DSTO, South Australia, Australia
Ben Raymond, Australian Antarctic Division, Tasmania, Australia
John Roddick, Flinders University, Adelaide, Australia
Tatiana Semenova, Australian Taxation Office, Australia
Kate Smith-Miles, Deakin University, Victoria, Australia
Andre Skusa, Syskoplan AG, Germany
David Taniar, Monash University, Melbourne, Australia
Jim Warren, University of Auckland, New Zealand
Ying Yang, Monash University, Melbourne, Australia

AusDM Sponsors

We wish to thank the following sponsors for their contribution towards this conference.

http://www.togaware.com

Faculty of Information Technology
http://www.uts.edu.au, http://www.it.uts.edu.au

The e-Markets Research Group
http://www.e-markets.org.au

http://www.iapa.org.au

http://www.netmapanalytics.com

ARC Research Network on Data Mining and Knowledge Discovery
http://www.dmkd.flinders.edu.au

http://www.neuronworks.com

Conference Programme

Wednesday, 29 November, 2006

09:00 - 09:05 **Opening and Welcome**

09:05 - 10:05 **INDUSTRY KEYNOTE: Tom Osborn, Thought Experiments**.

10:05 - 10:30 Coffee break

10:30 - 12:30 **Session 1: Professional Challenges**

- 10:30 - 11:00 SAFELY DELEGATING DATA MINING TASKS,
  Ling Qiu, Kok-Leong Ong, Siu Man Lui
- 11:00 - 11:30 DATA MINING METHODOLOGICAL WEAKNESSES AND SUGGESTED FIXES,
  John Maindonald
- 11:30 - 12:00 ACCURACY ESTIMATION WITH CLUSTERED DATASET,
  Ricco Rakotomalala, Jean-Hughes Chauchat, François Pellegrino
- 12:00 - 12:30 TOWARDS AUTOMATED RECORD LINKAGE,
  Karl Goiser, Peter Christen

12:30 - 13:30 Lunch

13:30-13:40 **INDUSTRY UPDATE: Inna Kolyshkina, IAPA,**
Update about IAPA matters.

13:40-14:30 **INDUSTRY KEYNOTE: Warwick Graco, ATO,**
Some Challenges in Knowledge Discovery.

14:30 - 15:00 Coffee break

15:00 - 17:00 **Session 2: Health Data Mining**

- 15:00 - 15:30 A COMPARATIVE STUDY OF CLASSIFICATION METHODS FOR MICROARRAY DATA ANALYSIS,
  Hong Hu, Jiuyong Li, Ashley Plank, Hua Wang, Grant Daggard
- 15:30 - 16:00 DATA MINING IN CONCEPTUALISING ACTIVE AGEING,
  Richi Nayak, Laurie Buys, Jan Lovie-Kitchins
- 16:00 - 16:30 ANALYSIS OF BREAST FEEDING DATA USING DATA MINING METHODS,
  Hongxing He, Huidong Jin, Jie Chen, Damien McAullay, Jiuyong Li, Tony Fallon
- 16:30 - 17:00 USING A KERNEL-BASED APPROACH TO VISUALIZE INTEGRATED CHRONIC FATIGUE SYNDROME DATASETS,
  Ahmad Al-Oqaily, Paul Kennedy

15:00 - 17:00 **Session 3: Scientific Data Mining**

- 15:00 - 15:30 ANALYZING HARMONIC MONITORING DATA USING DATA MINING,
  Ali Asheibi, David Stirling, Danny Soetanto
- 15:30 - 16:00 DISCOVER KNOWLEDGE FROM DISTRIBUTION MAPS USING BAYESIAN NETWORKS,
  Norazwin Buang, Nianjun Liu, Terry Caelli, Rob Lesslie, Michael J Hill
- 16:00 - 16:30 DATA MINING FOR LIFETIME PREDICTION OF METALLIC COMPONENTS,
  Esther Ge, Richi Nayak, Yue Xu, Yuefeng Li

Thursday, 30 November, 2006

09:00 - 10:00 **INDUSTRY KEYNOTE: Jolie Baasch, Credit Corp Group Limited, Data Mining: Bridging the Gap between Research and Deployment.**

10:05 - 10:30 Coffee break

10:30 - 12:30 **Session 4: Text Mining**

- 10:30 - 11:00 INTEGRATED SCORING FOR SPELLING ERROR CORRECTION, ABBREVIATION EXPANSION AND CASE RESTORATION IN DIRTY TEXT,
  Wilson Wong, Wei Liu, Mohammed Bennamoun
- 11:00 - 11:30 A STUDY OF LOCAL AND GLOBAL THRESHOLDING TECHNIQUES IN TEXT CATEGORIZATION,
  Nayer Wanas, Dina Said, Nevin Darwish, Nadia Hegazy
- 11:30 - 12:00 A CHARACTERIZATION OF WORDNET FEATURES IN BOOLEAN MODELS FOR TEXT CLASSIFICATION,
  Trevor Mansuy, Robert Hilderman
- 12:00 - 12:30 WEIGHTED KERNEL MODEL FOR TEXT CATEGORIZATION,
  Lei Zhang, Debbie Zhang, Simeon Simoff, John Debenham

10:30 - 12:30 **Session 5: Algorithms**

- 10:30 - 11:00 VISUALIZATION OF ATTRACTIVE AND REPULSIVE ZONES BETWEEN VARIABLES,
  Sylvie Guillaume, Leïla Nemmiche Alachaher
- 11:00 - 11:30 ON THE OPTIMAL WORKING SET SIZE IN SERIAL AND PARALLEL SUPPORT VECTOR MACHINE LEARNING WITH THE DECOMPOSITION ALGORITHM,
  Tatjana Eitrich, Bruno Lang
- 11:30 - 12:00 MARKING TIME IN SEQUENCE MINING,
  Carl Mooney, John Roddick

12:30 - 13:30 Lunch

13:30 - 14:30 **INDUSTRY KEYNOTE: Paul Beinat, NeuronWorks Pty. Ltd**.

14:30 - 15:00 Coffee break

15:00 - 17:00 **Session 6: Financial Data Mining**

- 15:00 - 15:30 DISCOVERING DEBTOR PATTERNS OF CENTRELINK CUSTOMERS,
  Yanchang Zhao, Longbing Cao, Yvonne Morrow, Yuming Ou, Jiarui Ni, Chengqi Zhang
- 15:30 - 16:00 WHAT TYPES OF EVENTS PROVIDE THE STRONGEST EVIDENCE THAT THE STOCK MARKET IS AFFECTED BY COMPANY SPECIFIC NEWS?
  Calum Robertson, Shlomo Geva, Rodney Wolff
- 16:00 - 16:30 INVESTIGATING THE SIZE AND VALUE EFFECT IN DETERMINING PERFORMANCE OF AUSTRALIAN LISTED COMPANIES: A NEURAL NETWORK APPROACH,
  Justin Luu, Paul Kennedy

15:00 - 17:00 **Session 7: Web Mining**

- 15:00 - 15:30 EXTRACTION OF FLAT AND NESTED DATA RECORDS FROM WEB PAGES,
  Siddu P. Algur, P. S. Hiremath
- 15:30 - 16:00 TRACKING THE CHANGES OF DYNAMIC WEB PAGES IN THE EXISTENCE OF URL REWRITING,
  Ping-Jer Yeh, Jie-Tsung Li, Shyan-Ming Yuan
- 16:00 - 16:30 A FRAMEWORK OF COMBINING MARKOV MODEL WITH ASSOCIATION RULES FOR PREDICTING WEB PAGE ACCESSES,
  Faten Khalil, Jiuyong Li, Hua Wang
- 16:30 - 17:00 MODELING PROLIFERATION OF IDEAS IN ONLINE SOCIAL NETWORKS,
  Muhammad Ahmad, Ankur Teredesai.

Safely Delegating Data Mining Tasks

Ling Qiu [1, a] **Kok-Leong Ong** [2] **Siu Man Lui** [1, b]

[1] School of Maths, Physics & Information Technology
James Cook University
[a] Townsville, QLD 4811, Australia
[b] Cairns, QLD 4870, Australia
Email: {ling.qiu, carrie.lui}@jcu.edu.au

[2] School of Engineering & Information Technology
Deakin University
Geelong, VIC 3217, Australia
Email: leong@deakin.edu.au

Abstract

Data mining is playing an important role in decision making for business activities and governmental administration. Since many organizations or their divisions do not possess the in-house expertise and infrastructure for data mining, it is beneficial to delegate data mining tasks to external service providers. However, the organizations or divisions may lose of private information during the delegating process. In this paper, we present a Bloom filter based solution to enable organizations or their divisions to delegate the tasks of mining association rules while protecting data privacy. Our approach can achieve high precision in data mining by only trading-off storage requirements, instead of by trading-off the level of privacy preserving.

Keywords: Delegating, privacy preserving, Bloom filter, data mining.

1 Introduction

1.1 Background and Motivation

Data mining, as one of the IT services most needed by organizations, has been realized as an important way for discovering knowledge from the data and converting "data rich" to "knowledge rich" so as to assist strategic decision making. Padmanabhan *et al.* (2003) demonstrated the use of data mining for CRM (customer relationship management) applications in e-commerce. The benefits of using data mining for business and administrative problems have been demonstrated in various industries and governmental sectors, e.g., banking, insurance, direct-mail marketing, telecommunications, retails, and health care (Apte, C., Liu, Pednault & Smyth 2002). Among all the available data mining methods, the discovery of associations between business events or transactions is one of the most commonly used data mining techniques. Association rule mining has been an important application in decision support and marketing strategy (Lin, Q.-Y., Chen, Chen & Chen 2003).

We consider a typical application scenario as follows. In an organization (e.g., a governmental sector), there are several divisions including an IT division which provides IT services for the whole organization. A functional division may have to delegate its data mining tasks to the IT division because of two reasons: lack of IT expertise and lack of powerful computing resources which are usually centrally managed by the IT division. The data used for the data mining usually involves privacy that the functional division may not want to disclose to anyone outside the division. To preserve the data privacy, this division should first convert (or encrypt) the source data to another format of presentation before transferring to the IT division. Therefore, there are two factors which are important for enabling a functional division to delegate data mining tasks to the IT division: (1) the computational time of data conversion is less than that of data mining; otherwise it is not at all worthwhile to do so; and (2) the storage space of converted data should be acceptable (the less the better though, several times more is still acceptable and practical).

This scenario can be extended to a more general circumstance in which all divisions are individually independent organizations or companies. This is because in today's fast-paced business environment, it is impossible for any single organization to understand, develop, and implement every information technology needed. It can also be extended to online scenarios, e.g., a distributed computing environment in which some edge servers undertaking delegated mining tasks may be intruded by hackling activities and may not be fully trusted.

When delegating mining tasks[1], we should protect the following three elements which may expose data privacy: (1) the source data which is the database of all transactions; (2) the mining requests which are itemsets of interests; and (3) the mining results which are frequent itemsets and association rules.

People have proposed various methods to preserve customer privacy in data mining for some scenarios, such as a distributed environment. However, those existing methods cannot protect all three elements simultaneously. This is because when a first party[2] delegates its mining tasks to a third party[3], it has to provide the source database (which might be someway encrypted) together with some additional infor-

[1] Without further specification, we always refer to association rule mining tasks.

[2] This is the party that delegates its data mining tasks. It may be a functional division in the scenario discussed above or a center server in a distributed environment with client-server architecture.

[3] This is the party that is authorized by the first party to undertake the delegated data mining tasks. It may be the IT division of an organization or an edge server in a distributed environment with client-server architecture.

mation (e.g., plain text of mining requests) without which this third party may not be able to carry out the mining tasks. Given this situation, those proposed methods are unable to efficiently prevent the exposure of private information to the third party, or unable to prevent the third party from deciphering further information from the mining results (which would be sent back to the first party) with the additional information.

1.2 Our Solution

In this paper, we present a Bloom filter based approach which provides an algorithm for privacy preserving association rule mining with computation efficiency and predictable (controllable) analysis precision. The Bloom filter (Bloom, B. 1970) is a stream (or a vector) of binary bits. It is a computationally efficient and irreversible coding scheme that can represent a set of objects while preserving privacy of the objects (technical details will be presented in Section 3.1).

With our approach, firstly the source data is converted to Bloom filter representation and handed over to a third party (e.g., the IT division of the organization) together with mining algorithms. Then the first party sends its mining requests to the third party. Mining requests are actually candidates of frequent itemsets which are also represented by Bloom filters. Lastly, the third party runs the mining algorithms with source data and mining requests, and comes out the mining results which are frequent itemsets or association rule represented by Bloom filters. In the above mining process, what the first party exposes to the third party does not violate privacy (Kantarcıoğlu, M., Jin & Clifton 2004); that is, the third party would not be able to distill down private information from Bloom filters. Therefore all the three elements mentioned above are fully protected by Bloom filters.

The goal of privacy preserving can be achieved by Bloom filter because it satisfies simultaneously the following three conditions. First, transactions containing different numbers of items are mapped to Bloom filters with the same length. This prevents an adversary from deciphering the compositions of transactions by analyzing the lengths of transactions. Second, Bloom filters support membership queries. This allows an authorized third party to carry out data mining tasks with only Bloom filters (i.e., Bloom filters of either transactions or candidates of frequent itemsets). Third, without knowing all possible individual items in the transactions, it is difficult to identify what items are included in the Bloom filter of a transaction by counting the numbers of 1's and 0's. This is because the probability of a bit in a Bloom filter being 1 or 0 is 0.5 given that the parameters of the Bloom filter are optimally chosen (see detailed mathematical analysis in (Qiu, L., Li & Wu 2006)).

The experimental results show that (1) the data conversion time is much less than mining time, which supports the worthiness to delegate mining tasks; (2) there is a tradeoff between storage space and mining precision; (3) there is a positive relationship between privacy security level and mining precision; (4) the converted data does not require more storage space compared with its original storage format.

1.3 Organization of the Paper

The remaining sections are organized as follows. Firstly in Section 2 we review the related work on privacy preserving data mining. After that in Section 3 we present our solution which uses a technique of keyed Bloom filters to encode the raw data, the data mining requests, and also the results of data analysis during the data exchanges for privacy preserving. Next, we demonstrate in Section 4 the implementation of the proposed solution over a point-of-sale dataset and a web clickstream dataset. We present experiments which investigate the tradeoffs among the level of privacy control, analysis precisions, computational requirements, and storage requirements of our solution with comparisons over other mining methods. Lastly in Section 5, we conclude the paper with discussions of different application scenarios made possible by the solution, and point out some directions for further study.

2 Literature Review

Association rule mining has been an active research area since its introduction (Agrawal, R., Imilienski & Swami 1993). Various algorithms have been proposed to improve the performance of mining association rules and frequent itemsets. An interesting direction is the development of techniques that incorporate privacy concerns.

One type of these techniques is perturbation based, which perturbs the data to a certain degree before data mining so that the real values of sensitive data are obscured while statistics properties of the data are preserved. An early work of Agrawal and Srikant (2000) proposed a perturbation based approach for decision tree learning. Some recent work (Evfimievski, A., Srikant, Agrawal & Gehrke 2002, Rizvi, S. & Haritsa 2002, Atallah, M., Bertino, Elmagarmid, Ibrahim & Verykios 1999, Oliveira, S. & Zaiane 2003, Saygin, Y., Verykios & Clifton 2001) investigates the tradeoff between the extent of private information leakage and the degree of data mining accuracy. One problem of perturbation based approach is that it may introduce some false association rules. Another drawback of this approach is that it cannot always fully preserve privacy of data while achieving precision of mining results (Kargupta, H., Datta, Wang & Sivakumar 2003), the effect of the amount of perturbation of the data on the accuracy of mining results is unpredictable.

The second type of these techniques is distributed privacy preserving data mining (Pinkas, B. 2002, Vaidya, J. & Clifton 2002, Kantarcıoğlu, M. & Clifton 2002) based on secure multi-party computation. This approach is only applicable when there are multiple parties among which each possesses partial data for the overall mining process and wants to obtain any overall mining results without disclosing their own data source. Moreover, this method needs sophisticated protocols (secure multi-party computation based). These make it infeasible for our scenario.

Both types of techniques are designed to protect privacy by masquerading the original data. They are not designed to protect data privacy from the mining requests or the mining results, which are accessible by data miners.

Recently, Agrawal *et al.* (2004) presented an order-preserving encryption scheme for numeric data that allows comparison operations to be directly applied on encrypted data. However, encryption is time consuming and it may require auxiliary indices. It is only designed for certain type of queries and may not be suitable for complex tasks such as association rule mining.

3 Our Solution: Bloom Filter-Based Approach

From the literature, there is no single method that can enable organizations to delegate data mining tasks

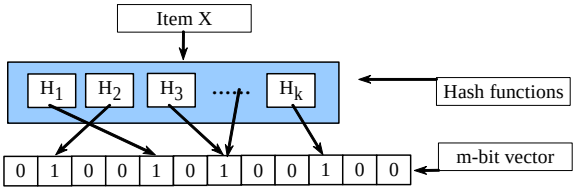


Figure 1: Constructing a Bloom filter of an item

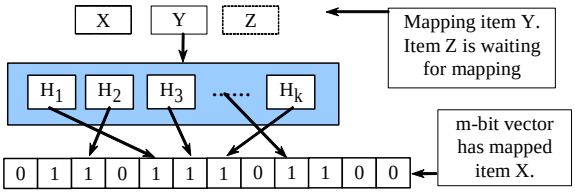


Figure 2: Constructing a Bloom filter of a transaction

while protecting all three elements involving the disclosure of privacy in the process of delegating data mining tasks. Our main objective in this paper is to propose a computationally feasible and efficient solution for the scenario.

A large number of examples from different industries (such as financial, medical, insurance, and retails) can be used for the study of the thread of privacy and business knowledge disclosure. In this paper, we consider the well-known association rule mining (Agrawal, R. et al. 1993), also known as market basket analysis (Chen, Y.-L., Tang, Shena & Hu 2005) in business analytics. Association rule mining can be performed by two steps: (1) mining of frequent itemsets, followed by (2) mining of association rules from frequent itemsets. Currently a well-known algorithm for mining of frequent itemsets is Apriori algorithm proposed by Agrawal and Srikant in (Agrawal, R. & Srikant 1994). Based on Apriori algorithm, in our early study (Qiu, L. et al. 2006) we investigated the feasibility of using a Bloom filter based approach for mining of frequent itemsets with privacy concerns. In this paper, we propose a solution with concerns of privacy protection by extending the Bloom filter-based approach to the whole process of association rule mining and applying to delegating scenario.

In what follows, we first introduce the mechanisms of constructing Bloom filters and membership queries over Bloom filters with discussion on the feature of privacy preserving. We then present algorithms of frequent itemset mining and association rule mining.

3.1 Bloom Filters

The Bloom filter (Bloom, B. 1970) is a computationally efficient hash-based probabilistic scheme that can represent a set of objects with minimal memory requirements. It can be used to answer membership queries with *zero* false negatives (i.e., without missing of useful information) and low false positives (i.e., with incurring of some extra results that are not of interests).

The mechanism of a Bloom filter contains (1) a binary vector (or stream) with length m and (2) k hash functions $h_1, h_2, \ldots, h_k$ of range from 1 to m. Given an item x, the Bloom filter of x, denoted as $B(x)$, is constructed by the following steps: (1) initialize by setting all bits of the vector with 0 and (2) set bit $h_i(x)$ (where $1 \leqslant i \leqslant k$) of the vector with 1. For example, if $h_2(x) = 7$, then bit 7 (i.e., the 7^{th} bit) of the vector is set with 1. It is possible that several hash functions set the same bit of the vector, i.e., $h_i(x) = h_j(x)$. Thus, after conversion, the number of

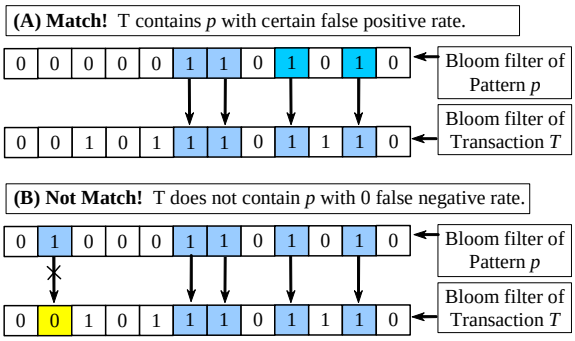


Figure 3: Membership query over Bloom filters

1's in $B(x)$ is not greater than k. Figure 1 illustrates how to construct a Bloom filter of an item.

It is similar to construct a Bloom filter of a transaction $T = \{X, Y, Z\}$ (or an itemset). Figure 2 illustrates the process in which item Y (or Z) is mapped onto the binary vector onto which item X (or items X and Y) has already been mapped. This process can be presented as $B(T) = B(X) \oplus B(Y) \oplus B(Z)$ where operator $\oplus$ stands for bitwise OR[4]. It should be pointed out that converting the data into Bloom filters is an irreversible process. Any unauthorized party accessing to Bloom filters will have no way to know/infer the original value of the data represented by Bloom filters unless they have the access to the original data, the hash functions, and the secrete keys (introduced later in this subsection).

To check whether a pattern p is contained by a transaction T, we examine whether $B(p) \otimes B(T) = B(p)$ holds, where operator $\otimes$ stands for bitwise AND[5]. If $B(p) \otimes B(T) \neq B(p)$, then p is definitely not contained by T; otherwise, p is a member of T with very low probability of false (i.e., false positive rate). Figure 3 shows the process of membership query with Bloom filters.

A Bloom filter does not incur any false negative, meaning that it will not suggest that a pattern is not in T if it is; but it may yield a false positive, meaning that it may suggest that a pattern is in T even though it is not. In our application, the false positive rate is upper-bounded by 0.5^k, where k is the number of hash functions, and the optimal value of k is given by $k = \frac{m}{n} \ln 2$, where m is the length of Bloom filters (i.e., the number of bits in the binary vectors), and n is the average length of transactions (i.e., the average number of items in transactions). Technical details for deriving the optimal value of k can be found in (Qiu, L. et al. 2006). Therefore, the false positive rate decreases exponentially with linear increase of the number of hash functions or the length of Bloom filters. For many applications, this is acceptable as long as the false positive rate is sufficiently small.

The privacy of data can be preserved by Bloom filters due to the irreversible feature. Given the above parameters of a Bloom filter, there are m^k possible mappings (for example, if we set the length of a Bloom filter $m = 80$ and the number of hash functions $k = 25$, then there are totally 80^{25} possible mappings in constructing the Bloom filter). Thus a Bloom filter can against some straightforward attacks (e.g., unknown-text attack and brute-force attack). It is certain that some other encryption algorithms (e.g., DES or RSA) are more secure; however, the computational cost is much more higher than

[4]The bitwise OR operation is defined as: $0 \oplus a = a$ and $1 \oplus a = 1$ where a is a binary variable 0 or 1.

[5]The bitwise AND operation is defined as: $0 \otimes a = 0$ and $1 \otimes a = a$ where a is a binary variable 0 or 1.

that of our method. The length of Bloom filters is a tradeoff between security level and computational cost. To enhance the security level, we insert a secret key K into each itemset or transaction before constructing its Bloom filter. The secret key K should not be chosen from the items. This amendment can be represented as $B_K(T) = B(T) \oplus B(K)$, in which $B_K(T)$ is referred to a keyed Bloom filter (see (Qiu, L. et al. 2006)). Without further mention, we always assume that Bloom filters are constructed with a secret key.

With the membership query mechanism of Bloom filters described above, we are able to conduct association rule mining with access to only Bloom filters. Given the irreversible feature of Bloom filters, a first party can convert all data involving disclosure of privacy to Bloom Filters and safely delegate the mining tasks to a third party without disclosing any value of the data in the database, the mining requests, and the mining results. We do not need to worry about missing of useful information (i.e., frequent itemsets and strong association rules in our application) due to *zero* false negative rate; but we may get some extra information (which may confuse data hacker while not affecting the quality of mining results) with low probability of false positive rate (see detailed mathematical analysis in (Qiu, L. et al. 2006)).

3.2 Mining Processes and Algorithms

The procedure of mining frequent itemsets is the process of membership queries over Bloom filters. Based on Apriori algorithm, a frame work of our method is shown in Algorithm 1. Algorithm 1 can be divided into three phases: *counting phase* (lines 3–5), *pruning phase* (lines 6–8), and *candidates generating phase* (lines 9–10) in each round ℓ, where ℓ indicates the size of each candidate itemset dealt with. In the counting phase, each candidate filter is checked against all transaction filters[6] and the candidate's count is updated. In the pruning phase, any Bloom filter is eliminated from the candidate set if its count (i.e., Support(x)) is less than the given threshold $N \cdot \tau$. Finally, in the candidates generating phase, new candidate Bloom filters are generated from the Bloom filters discovered in the current round. The new candidates will be used for data mining in the next round. With the results of frequent itemsets, the mining of association rules is relatively simple, which is shown in Algorithm 2.

The complete process of association rule mining is given as follows. (1) The first party hands over to the third party the application software that performs frequent itemset mining together with the database of transactions represented by Bloom filters. (2) The first party sends to the third party mining requests which include candidate itemsets and the threshold of minimum support. The generation of candidates is done at the first party side by running Apriori_gen (Agrawal, R. & Srikant 1994) which is the critical step of Apriori algorithm. This step has to be done in the first party side because it involves data privacy (Qiu, L. et al. 2006). (3) The third party carries out mining tasks with the data received and finally returns the mining results which are Bloom filters of frequent itemsets together with their supports. (4) With frequent itemsets and their supports returned from the third party, it is easy to generate strong association rules with thresholds of minimum confidence. This job can be performed by the first party itself, or by the third party. If it is performed by the first party, there is no need to convert frequent itemsets to Bloom filters. If it is performed by the third party, for privacy considerations all data has to be converted to Bloom filters.

[6] All transactions are organized in a tree hierarchy so as to minimize the times of membership queries. See details in (Qiu, L. et al. 2006).

Algorithm 1 Mining of frequent itemsets from Bloom filters

1: $C_1 = \{B(I_1), \ldots, B(I_d)\}$
   // $B(I_i)$ is the Bloom filter of item I_i
2: **for** $(\ell = 1;\ C_\ell \neq \varnothing;\ \ell{+}{+})$ **do**
3: **for** each $B(S) \in C_\ell$ and each transaction filter $B(T_i)$ **do**
4: **if** $B(S) \otimes B(T_i) = B(S)$ **then** Support(S) ++
   // S is a candidate frequent ℓ-itemset
5: **end for**
6: **for** each $B(S) \in C_\ell$ **do**
7: **if** Support$(S) < N \cdot \tau$ **then** delete $B(S)$ from C_ℓ
   // N is transaction number in the database,
   // and τ the threshold of minimum support
8: **end for**
9: $F_\ell = C_\ell$ // F_ℓ is the collection of Bloom filters of all "frequent" ℓ-itemsets
10: $C_{\ell+1} = can\_gen(F_\ell)$ // generate filters of candidate itemsets for the next round
11: **end for**
12: Answer $= \bigcup_\ell F_\ell$ // all filters of frequent itemsets

Algorithm 2 Mining of association rules from Bloom filters of frequent itemsets

1: $AR = \varnothing$; $F = \{B(F_1^1), \ldots, B(F_{d_1}^1), B(F_1^2), \ldots, B(F_{d_2}^2), \ldots, B(F_1^k)), \ldots, B(F_{d_k}^k)\}$
   // $B(F_i^s)$ is the Bloom filter of frequent
   // s-itemset F_i^s where $1 \leqslant s \leqslant k$ and $1 \leqslant i \leqslant d_s$
2: **for** $(s = 1;\ s < k\ ;\ s{+}{+})$ **do**
3: **for** $(t = s + 1;\ t \leqslant k\ ;\ t{+}{+})$ **do**
4: **for** each $B(F_i^s)$ and each $B(F_j^t)$ **do**
5: **if** $B(F_i^s) \otimes B(F_j^t) = B(F_i^s)$ and Support(F_j^t)/Support$(F_i^s) \geqslant \xi$ // ξ is the minimum threshold // of confidence
6: **then** $F_i^s \Rightarrow F_j^t - F_i^s$ is a strong association rule and is added to AR
7: **end for**
8: **end for**
9: **end for**
10: **return** AR // All strong association rules

4 Experiments

4.1 Experimental Settings

We implement the solution and evaluate it with experiments on two real datasets BMS-WebView-2 and BMS-POS which are publicly available for research communities[7]. Dataset BMS-POS contains several years of point-of-sale data from a large electronic retailer; whereas dataset BMS-WebView-2 contains several months of clickstream data from an e-commerce website. Table 1 shows the number of items, the average size of transactions, and the number of transactions included in these datasets. Figure 4 shows the distribution of transaction sizes of the datasets. For dataset BMS-POS, a transaction

[7] Downloadable at http://www.ecn.purdure.edu/KDDCUP.

Table 1: Characteristics of real datasets

Dataset	Distinct items	Max-size	Average size	Number of transactions
BMS-POS	1,657	164	6.53	515,597
BMS-WebView-2	3,340	161	4.62	77,512

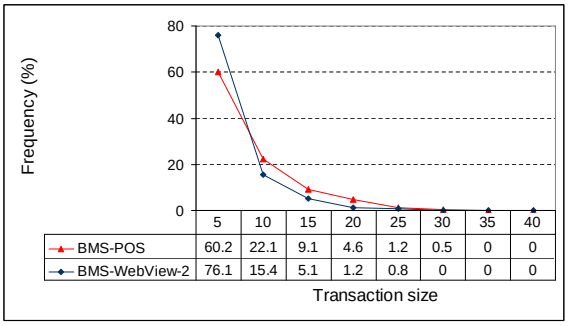


Figure 4: Distribution of transaction sizes

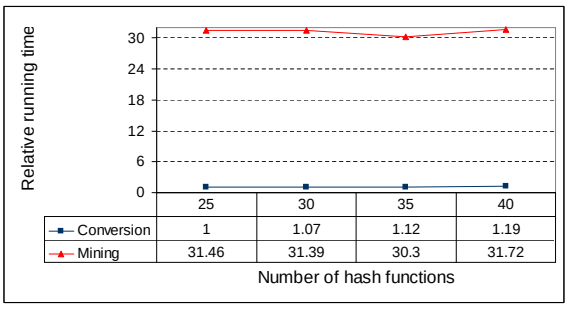


Figure 5: Data conversion time vs. mining time on dataset BMS-POS

is a list of items purchased in a basket; whereas for dataset BMS-WebView-2, a transaction is a browsing session which contains a list of webpages visited by a customer. The experiments are run on a Compaq desktop computer with Pentium-4 CPU clock rate of 3.00 GHz, 3.25 GB of RAM and 150 GB harddisk, with Microsoft Windows XP Professional SP2 as the operating system.

We have qualitatively analyzed the privacy preserving feature of Bloom filters in Section 3.1 (further theoretical analysis and discussions can be found in (Qiu, L. et al. 2006)). Therefore the emphasis of this set of experiments is to investigate the relationship among the level of privacy protection (determined by the number of hash functions), storage requirement, computation time, and analysis precision. In the experiments, we set the threshold of minimum support $\tau = 1\%$ and cluster the transactions in each dataset into 4 groups based on their transaction sizes (refer to (Qiu, L. et al. 2006) for technical details of grouping). We change k the number of hash functions used for Bloom filters from 25 to 40 in the experiments.

4.2 Experimental Results

Figures 5 and 6 show that the time of mining frequent itemsets is much more than the time of converting data to Bloom filter presentations, meaning that the mining process takes the major part of running time. This result verifies the worthiness in terms of running time for data format conversion before delegating mining tasks (satisfying the first factor enabling to delegate mining tasks as mentioned in Section 1).

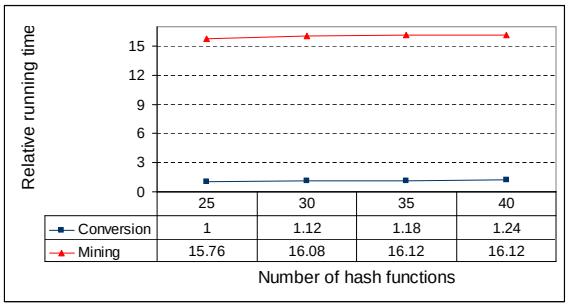


Figure 6: Data conversion time vs. mining time on dataset BMS-WebView-2

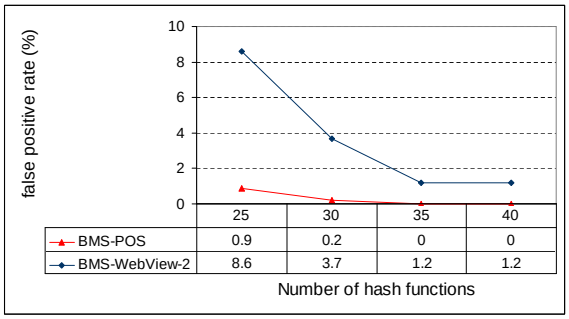


Figure 7: Mining precision vs. number of hash functions

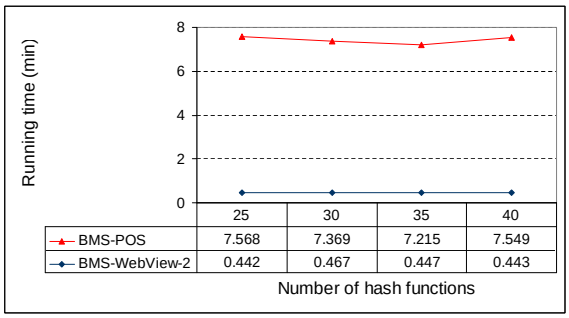


Figure 8: Running time vs. number of hash functions

Figure 7 shows the mining precisions with the change of k. There is a globally decreasing trend of false positive rates for each real dataset. For dataset BMS-POS, the false positive rate is less than 1% for $k \geqslant 25$. For dataset BMS-WebView-2 the false positive rate is below 10% for $k = 25$ and less than 4% for $k \geqslant 30$.

Figure 8 shows that the running time changes slightly with hash function number k. The running time is around 8 minutes for dataset BMS-POS and within 0.5 minute for dataset BMS-WebView-2, because comparatively dataset BMS-POS contains 7 times as many as transactions.

Figure 9 shows that the storage requirement is linearly increasing with k for both datasets. The reason is that the optimal value of k is given by $k = \frac{m}{n} \ln 2$ where m is the length of Bloom filters (Qiu, L. et al. 2006). The results of this experiment show that high mining precision can be achieved by increasing the number of hash functions. Consequently, the storage requirement increases linearly due to the use of longer Bloom filters.

Figure 10 shows a comparison of the average storage space required by a transaction under difference storage formats. The results show that the storage space of Bloom filter format is practical, i.e., it is less

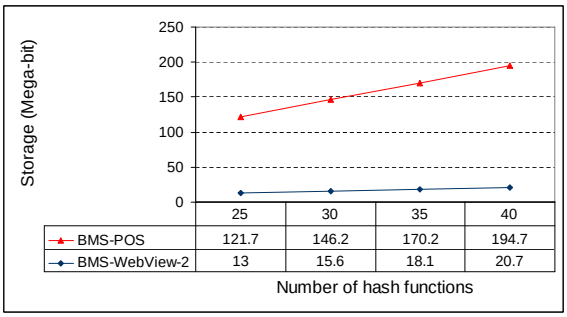


Figure 9: Storage requirement vs. number of hash functions

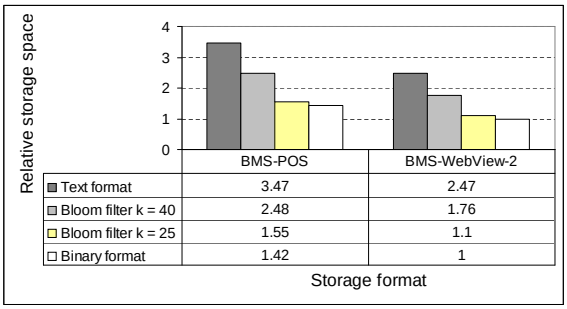


Figure 10: Average storage space of a transaction

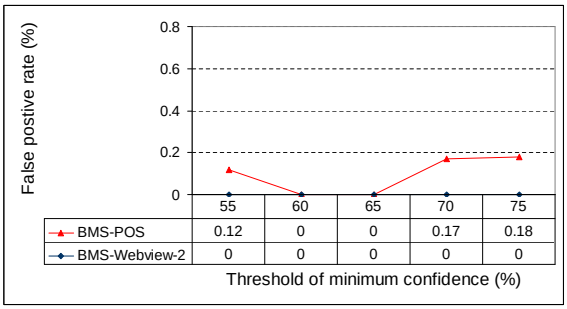


Figure 11: Precision of mining association rules

than text format but a bit more than binary format depending on the precision requirement (adjustable by k). This satisfies the second factor that it is worthwhile in term of storage space to adopt Bloom filter presentations as mentioned in Section 1. We can achieve further saving of storage space without decreasing mining precision with some techniques proposed by Qiu *et al.* (2006) (e.g., δ-folding and grouping).

With the mining results of frequent itemsets returning from the third party, we continue the mining of association rules with given threshold of minimum confidence. In this experiment, we let $k = 30$ under which the false positive rates of frequent itemsets are lower than 5% for both datasets. We vary the threshold of minimum confidence from 55% to 75%. The false negative rates are *zero* for both datasets, meaning that our approach does not miss useful information. As shown in Figure 11, the false positive rate is *zero* for dataset BMS-WebView-2 and lower than 0.2% for dataset BMS-POS. The running time for any dataset is less than 0.1 second.

5 Conclusions

In this paper, we have discussed and identified the risks of exposing data privacy in the scenario of delegating data mining tasks. We have also identified the factors that enable us to delegate mining tasks. We have proposed a privacy persevering data mining method and applied it to association rule mining in this delegation scenario. As compared with other existing methods, the metrics of our method include: (1) our approach is effective in protecting of three elements that can expose data privacy in the process of delegating mining tasks; (2) there is a positive relationship between the privacy security level and the analysis precision; (3) to increase the privacy security level, we only need to sacrifice data storage space; and (4) the solution is scalable, i.e., the storage space increases linearly with the privacy protection level or the analysis precision.

In our current study, we have developed a privacy protection method for association rule mining with a single (centralized) database. We can also apply our method to other mining tasks (e.g., mining of some other rules that are of interest to researchers). Further study to investigate the feasibility and implementation of the proposed solution in a multiple (distributed) databases environment is needed. Another future research direction could be investigating the feasibility of using the keyed Bloom filter approach in other tasks of business analytics.

References

Agrawal, R., Imilienski, T. & Swami, A. (1993), Mining association rules between sets of items in large databases, *in* 'Proceedings of the ACM SIGMOD ICMD', pp. 207–216.

Agrawal, R., Kiernan, J., Srikant, R. & Xu, Y. (2004), Order preserving encryption for numeric data, *in* 'Proceedings of the ACM SIGMOD ICMD', pp. 563–574.

Agrawal, R. & Srikant, R. (1994), Fast algorithms for mining association rules in large databases, *in* 'Proceedings of VLDB'94', pp. 487–499.

Agrawal, R. & Srikant, R. (2000), Privacy-preserving data mining, *in* 'Proceedings of the ACM SIGMOD ICMD', pp. 439–450.

Apte, C., Liu, B., Pednault, E. & Smyth, P. (2002), 'Business applications of data mining', *Communications of the ACM* **45**(8), 49–53.

Atallah, M., Bertino, E., Elmagarmid, A. K., Ibrahim, M. & Verykios, V. S. (1999), Disclosure limitation of sensitive rules, *in* 'Proceedings of the IEEE KDEE', pp. 45–52.

Bloom, B. (1970), 'Space time tradeoffs in hash coding with allowable errors', *Communications of the ACM* **13**(7), 422–426.

Chen, Y.-L., Tang, K., Shena, R.-J. & Hu, Y.-H. (2005), 'Market basket analysis in a multiple store environment', *Decision Support Systems* **40**(2), 339–354.

Evfimievski, A., Srikant, R., Agrawal, R. & Gehrke, J. (2002), Privacy preserving mining of association rules, *in* 'Proceedings of the 8th ACM SIGKDD KDD 2002', pp. 217–228.

Kantarcıoğlu, M. & Clifton, C. (2002), Privacy preserving distributed mining of association rules on horizontally partitioned data, *in* 'Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery', pp. 24–31.

Kantarcıoğlu, M., Jin, J. & Clifton, C. (2004), When do data mining results violate privacy?, *in* 'Proceedings of the 10th ACM SIGKDD KDD 2004', pp. 599–604.

Kargupta, H., Datta, S., Wang, Q. & Sivakumar, K. (2003), On the privacy preserving properties of random data perturbation techniques, *in* 'Proceedings of the 3rd IEEE ICDM', pp. 99–106.

Lin, Q.-Y., Chen, Y.-L., Chen, J.-S. & Chen, Y.-C. (2003), 'Mining inter-organizational retailing knowledge for an alliance formed by competitive firms', *Information & Management* **40**(5), 431–442.

Oliveira, S. & Zaiane, O. (2003), Protecting sensitive knowledge by data sanitization, *in* 'Proceedings of the 3rd IEEE ICDM', pp. 211–218.

Padmanabhan, B. & Tzhilin, A. (2003), 'On the use of optimization for data mining: theoretical interactions and eCRM opportunities', *Management Science* **49**(10), 1327–1343.

Pinkas, B. (2002), 'Cryptographic techniques for privacy preserving data mining', *ACM SIGKDD Explorations* **4**(2), 12–19.

Qiu, L., Li, Y. & Wu, X. (2006), 'Preserving privacy in association rule mining with Bloom filters', *Journal of Intelligent Information Systems* . In press.

Rizvi, S. & Haritsa, J. (2002), Maintaining data privacy in association rule mining, *in* 'Proceedings of VLDB'02', pp. 682–693.

Saygin, Y., Verykios, V. S. & Clifton, C. (2001), 'Using unknowns to prevent discovery of association rules', *Sigmod Record* **30**(4), 45–54.

Vaidya, J. & Clifton, C. (2002), Privacy preserving association rule mining in vertically partitioned data, *in* 'Proceedings of the 8th ACM SIGKDD KDD', pp. 639–644.

Data Mining Methodological Weaknesses and Suggested Fixes

John Maindonald

Centre for Mathematics and Its Applications, Australian National University, Canberra ACT 0200, Australia.
Email: john.maindonald@anu.edu.au

Abstract

Predictive accuracy claims should give explicit descriptions of the steps followed, with access to the code used. This allows referees and readers to check for common traps, and to repeat the same steps on other data. Feature selection and/or model selection and/or tuning must be independent of the test data. For use of cross-validation, such steps must be repeated at each fold. Even then, such accuracy assessments have the limitation that the target population, to which results will be applied, is commonly different from the source population. Commonly, it is shifted forward in time, and it may differ in other respects also.

A consequence of source/target differences is that highly sophisticated modeling may be pointless or even counter-productive. At best, model effects in the target population may be broadly similar. Investigation of the pattern of changes over time is required. Such studies are unusual in the data mining literature, in part because relevant data have not been available.

Several recent investigations are noted that shed interesting light on the comparison between observational and experimental studies, with particular relevance when there is an interest in giving parameter estimates a causal interpretation.

Data mining activity would benefit from wider co-operation in the development and deployment of computing tools, and from better integration of those tools into the publication process.

Keywords: Data mining, statistics, predictive accuracy, target population, observational data, selection bias, reject inference, comparison of algorithms.

1 Introduction

It is now widely though not universally understood that training set accuracy, derived by using the training data for testing also, can be grossly optimistic. Cross-validation or a bootstrap approach is therefore preferred. Where however feature selection and/or model tuning are a component of the model fitting process, care is required to avoid subtler versions of the bias in the training set accuracy measure. For an unbiased assessment, any feature selection and/or model tuning must be repeated at each fold of the cross-validation.

Other important issues relate to the distinction between observational and experimental data, to differences between source and target population, to the stability and interpretability of model parameters, to the comparison of algorithms, to the implications of new technology for the publication process, and to improving cooperation in the development of new tools. The remainder of this section will make preliminary comments on the first two of these issues.

1.1 Observational versus experimental data

Data mining typically uses for prediction or other inferences data that are observational rather than experimental. This introduces hazards that, for data from carefully planned and conducted experiments, are largely absent.

Thus, in a recent study that used a large US car accident database (Meyer and Finney, 2005), the interest is in a model parameter that accounts for the effect of airbag availability on accident mortality. Many factors apart from airbag availability contribute to the outcome. If other factors are ignored, airbags seem to give large benefits. After accounting for the effects of seatbelts and various other factors, benefits appear small or non-existent.

Compare this with a notional experimental study, where cars would be randomly assigned to have airbags fitted, or not, and where other factors (use of a seatbelt, speed of impact, etc) should on average contribute only statistical noise.

Where experimental studies fail, it is typically for one of a small number of reasons, commonly failure of the randomization process. Other possibilities are that experimental subjects (or units) may be untypical of the population to which results will be applied, or that the experiment may answer a question that is different (perhaps subtly different) from the question of interest.

By contrast, it is hardly possible to give a simple and reasonably complete summary of the different ways in which observational studies may fail. See however Rosenbaum (2002). The range and variety of different types of observational study is almost unlimited.

In some business and industrial problems, it may be reasonable to limit attention to a small number of well understood causal factors. The assumption that this is the case should not be made lightly. At the very least, issues such as will be discussed below severely limit the range of problems where relatively automated data mining approaches can be trusted to give useful results. At worst they may make any inferences from available data, however carefully teased out, perilous.

1.2 Accuracy varies with target population

A recurring theme will be that accuracy assessments are specific to a particular target population. A sim-

ple example, using the `Pima.tr` data set that is included with the *MASS* package for R, will illustrate the point. There are seven columns of features that may help explain diabetic status, recorded as `No` or `Yes`. Use of Breiman's random forest algorithm, as implemented in R, gave a classification rule thus:

```
> Pima.rf <- randomForest(type ~ .,
+     data = Pima.tr)
```

The confusion matrix is

```
> Pima.rf$confusion
```

	No	Yes	class.error
No	111	21	0.159
Yes	36	32	0.529

The error rate is estimated as 28%; this is calculated as (132×0.159 + 68×0.529)/ (132 + 68).

If however predictive accuracy is calculated for a population in which the proportions of `No` and `Yes` are equal, the expected error rate changes to 0.5×0.159 + 0.5×0.529 = 34%.

Thus use of a balanced sample in cross-validation accuracy assessment may make a large difference to the assessment. Any report of an overall measure should be accompanied by details of the population composition that is assumed. Better still, accompany the report with details of the confusion matrix.

As an aside, note that balanced samples are a poor use of data, unless the relative proportions, perhaps weighted according to misclassification costs, are equal in the target population. Even then, it is better to use prior weights to train a model that is optimal for the relative frequencies and costs in the target population. See Ripley (1996) for the relevant Bayesian decision theory for classification models.

In the case discussed, the difference was in the relative frequencies in two categories. Differences between source and target population are common, and rarely so straightforwardly handled. This discussion will be pursued in the next section, noting also implications for the comparison of algorithms.

2 Honest use of cross-validation

A cross-validation estimate of accuracy, or an estimate obtained from a random split of the data into two parts, is often the best that is available. In default of anything better, it provides an upper bound on the accuracy that can be expected for predictions for the target population. Such estimates are in any case commonly used when algorithms are compared. Unless done correctly, such comparisons are meaningless, and potentially misleading.

Where there is feature selection and/or significant model selection and/or significant model tuning, the following steps are involved:

1. Select features and/or select model and/or tune the model
2. Fit the model that is in due course selected as "best".

Both these steps must be repeated at each fold of the cross-validation process, using what are the training data for that fold.

Consider the following experiment, leading in due course to Figure 1:

- Set up a matrix X whose n rows are observations, and whose $p >> n$ columns are features.

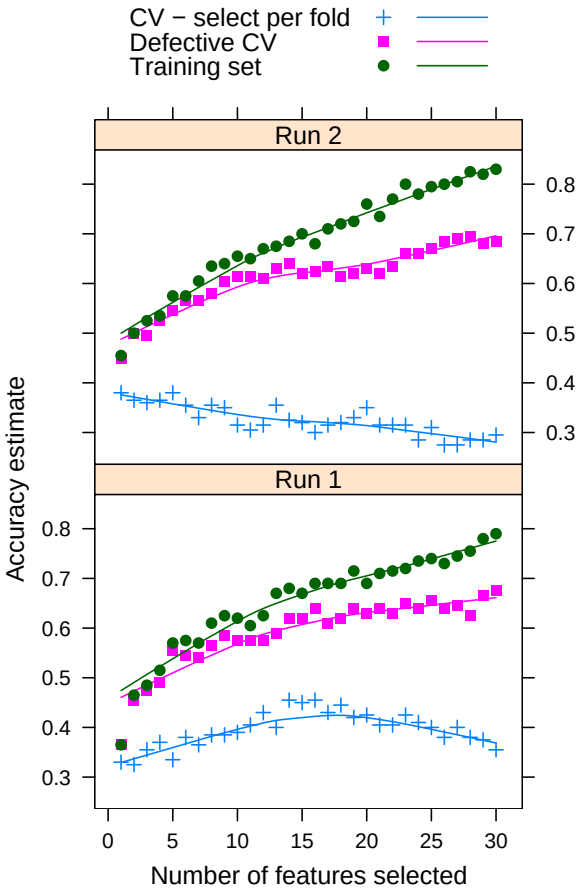


Figure 1: The plots repeat discriminant rule calculations, using different sets of random normal data, to compare different accuracy measures. Each data set had 200 observations, divided approximately equally between three groups, and 1000 features, The training measure (filled circle) is severely biased. Ten-fold cross-validation, with features selected using all the data (filled square), has a less severe bias that is nevertheless unacceptable. An acceptable measure (points shown as +) requires re-selection of features at each fold of the cross-validation.
(The code used to create these plots is available from the web page noted in Section 5.)

- Fill the elements with random normal data, independent (though this is not essential to the demonstration) between elements.
- Generate a categorical variable y, with $k = 3$ categories.

Now do the following:

Defective cross-validation – select once
For each of $m = 2, 3 \ldots, 30$, repeat the following steps

1. Using an analysis of variance F-statistic as the criterion, choose the m features that best separate the rows of X into the k categories.
2. Do a cross-validation of a discriminant analysis with the chosen m features, and determine predictive accuracy.

Cross-validation – select at each fold
At each fold of the cross-validation, there is a local set of training data, on which the model is trained, with remaining data providing the local

test set. Modify the above procedure so that, at each fold, the local training set is used for feature selection (and any model tuning), prior to fitting the model and making predictions for the local test set.

Figure 1 shows, for two different X-matrices where the data are "white noise", and with 200 observations that were divided approximately equally between three groups, the resulting assessments of predictive accuracy, plotted against number of features. Notice the different patterns of change in the correct cross-validated error rate (points shown as +), different between the two sets of random data. Similar results to the points shown as + will be obtained if the overall model is assessed on a completely separate set of randomly chosen test data, i.e. on a new matrix Z of the same dimensions as X and filled in the same way with random normal data.

The trap may seem obvious, but a number of authors, including well-known names, have been caught by it. See Ambroise and McLachlan (2002); Zhu et al. (2006). In statistics as in mathematics, plausible notions and methods that have not been validated with proper care are commonly found to be wrong or misguided. Simulation, with random data as in Figure 1, can often give a useful wake-up call.

Suppose however that there is a genuine signal in X, which will now be treated as a sample from the source population. Suppose also that there are systematic differences from the target population, from which we have a sample Z. Assume now that the cross-validation is done correctly. Close tuning to fit the source population will, at some point, lead to the degrading of performance on the target population. Although he does not make the point in this way, this is implicit in the comments in Hand (2006). I will comment further in the next section.

3 Source and target population

Differences in the relative frequencies of different groups in the data are relatively simple to handle. More generally, accuracy assessments that are based on cross-validation, or that are from a random split of the total data into training and test data, are realistic only if the processes that generated the data are the same processes that will apply when results are put to practical use. The source population, from which the data have been sampled, must be closely identical to the target population to which results will be applied. Or, to use different language, the model that describes the processes that have generated the data must also be an adequate description for the processes that will apply when model predictions are used in practice.

A clear and unequivocal near identity of source and target population is, with observational data, unusual. In a business context data that are derived from the past year's activities, or from the past several years, may be the basis for changes in business practice that will affect future years. This point will be taken up below.

3.1 Reject inference

A common further complication is noted in Hand (2006). In assessing credit risk, the sample is distorted as a sample of the potential population of applicants. The true outcome is known only for those applicants who were given credit, yet the inference is required for all applicants, leading to the term *reject inference*. The methodology discussed in Heckman (1979) can in principle address this problem, but requires assumptions that cannot always be checked.

Even if predictions are stable under temporal or other changes that affect the population of interest, it does not follow that model parameters (e.g., regression coefficients) will be stable. There are subtle and complex issues that affect the interpretation of such coefficients when they are derived from observational data.

3.2 Source and Target – A Taxonomy

The following are typical of situations that may occur in practice. This is a slightly modified version of the classification of the range of possibilities that appears in Maindonald & Braun (2006):

1. The data used to develop the model are, to a close approximation, a random sample from the population to which predictions will be applied. If this can be assumed, a simple use of a resampling method will give an accuracy estimate that is unbiased with respect to the population that is the target for predictions.

2. Test data are available that are from the target population, with a sampling mechanism that reflects the intended use of the model. The test data can then be used to derive a realistic estimate of predictive accuracy.

3. The sampling mechanism for the target data differs from the mechanism that yielded the data in 1, or yielded the test data in 2. However, there is a model that predicts how predictive accuracy will change with the change in sampling mechanism. Thus, in the `Pima.tr` and `Pima.te` datasets that were the basis for the calculations in Subsection 1.2, the predictive accuracy is a function of the relative number that are `Yes`.

4. The connection between the population from which the data have been sampled and the target population may be weak or tenuous. It may be so tenuous that a confident prediction of the score function for the target population is impossible. In other words, a realistic test set and associated sampling mechanism may not be available. An informed guess may be the best that is available.

These four possibilities are not completely distinct; they overlap at the boundaries. The distinction between them, such as it is, is however a good starting point for making a judgment on the closeness of the connection between the source and target populations.

Item 3 covers a wide range of possibilities. One simple possibility was discussed earlier, where the remedy is to give groups within the data weights that reflect the relative frequencies and perhaps costs in the target population, rather than those in the source population. The forest cover dataset from the UCI Machine Learning Database (Newman et al, 1998) is interesting in this connection. The relative numbers of the seven different forest cover types change systematically as a window of perhaps 5000 from the 500,000 observations moves through the data. This presumably reflects systematic changes in geographical location – information not included in the data. As the window moves, there are large changes in local predictive accuracy, i.e., using the data within the window as target. This is the case both for a model fitted to the data as a whole, and when the model is fitted to the data locally. While the confusion matrix from the local model changes somewhat between successive windows, the effect on predictive accuracy is of minor consequence relative to that of changes in the proportions of the different cover types.

For reject inference problems, approaches such as in Heckman (1979) are available, but rely strongly on specific modeling assumptions. Validation is accordingly both more necessary and more difficult.

In another common circumstance, there may for example be very extensive data on house prices in two suburbs of a large city. For predicting house prices in another suburb we have what is effectively a sample of two, and must further assume that this can be treated as a random sample. The assumption that errors are independently and identically distributed across the total sample of prices, as in most software that is explicitly aimed at data mining, will lead to optimistic assessments of predictive accuracy. (For other examples see Maindonald, 2003). Similar issues arise with data that are a time series. Again it is necessary, formally or informally, to account for the "error" part of the model.

3.3 Changes with time

Consider again the use of the current year's data to make assessments that will affect next year's business activity. If data from several previous years are available, then it makes sense to run the analysis separately for each of those years, and check for consistency between the different sets of results. If such data are not available, then there may be no good basis for judging the relevance to the subsequent year's business activity. Even where there does seem to be some modest level of consistency over time, this consistency may be placed in jeopardy by changes in external circumstances. Economic shocks – a dramatic increase in oil prices or an economic recession – may depending on the specific context create discontinuities that invalidate or place in doubt assessments that are based on past data.

Where the source and target populations are separated in time, model refinement is readily taken to a point where improved accuracy for the source population leads to reduced accuracy for the target population. What is signal at one point in time may with the passage of time become bias. Under-fitting, relative to estimates of accuracy that are based on a random split between test and training data or on cross-validation, may lead to improved accuracy for the data that matter.

Hand (2006) has two interesting examples that relate to credit scoring. Hand's Figure 4 shows the error rate over a $3\frac{1}{2}$ period, from a classifier built at the start of the period. The error rate drops to almost zero after 8 months, then after a year is back at the initial level, then rises to be $2\frac{1}{2}$ times the initial level by the end of the period. In a second graph (Hand's Figure 5), the performance of a tree-based classifier is compared with that of a linear discriminant function, over customers 1 to 60,000, using odd-numbered customers in the range 1 to 4999 for training. At the beginning of the series, the misclassification cost was around 0.1 less for the tree-based classifier. This difference had reduced to perhaps 0.05 by the end of the series, with the performance of the linear discriminant staying fairly constant at a cost of around 0.225. Other issues concern inevitable changes in the composition of the target population, arbitrariness and drift in the class definition ("concept drift"), and vagueness in the assignment of relative costs.

3.4 Implications for comparing algorithms

In practice then, there is not the identity between source and target populations that the standard comparisons of algorithms assume. Published comparisons of algorithms are at best broad indications of performance, even once careful attention has been given to advice such as appears on the web site Keogh et al (2006) or in the papers Elkan (2001); Salzberg (1997).

Hand (2006) makes the further point that, in comparisons of different algorithms, users who are more expert with a particular method will have a bias towards obtaining their best results with that method. This, and the inevitability that performance is to an extent data dependent, are yet further reasons for treating published comparisons of algorithms as, at best, broad indications of performance.

A quick check through the UCI Machine Learning repository did not reveal any sizeable data sets that are well suited to studying changes in algorithm performance over time. This is clearly a serious gap in the resources that are currently available for testing and evaluating algorithms. In a number of cases (e.g., the email spam database), it would be highly interesting to have comparable time-stamped data from a period of several years. The newly established UCR Time Series Data Mining Archive (Keogh, 2006) is therefore very welcome. Many practical classification problems have a time-dependent component that should not be ignored.

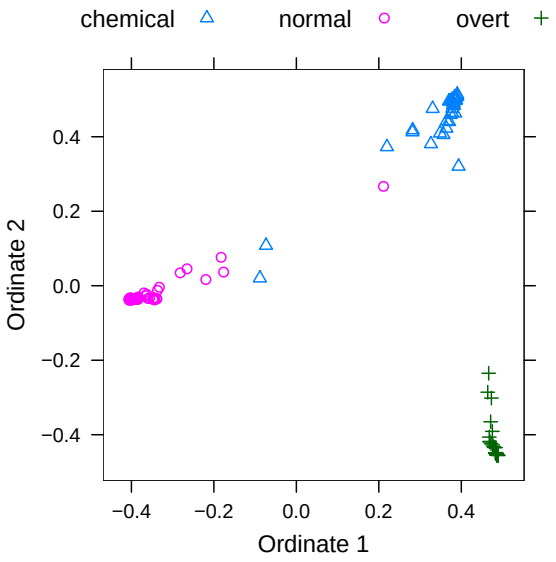


Figure 2: Calculations used the `diabetes` dataset, included in R's *mclust* package. Proximities r_{ij}, calculated for any pair (i, j) of points as the proportion of trees in which they appear at the same terminal node, were derived from use of the `randomForest` function with the `diabetes` dataset. Distances 1-r_{ij} were then used with R's `cmdscale` metric scaling function, yielding a two-dimensional representation.

3.5 Low-dimensional representations

It is helpful to characterize, where possible, conditions under which one or other algorithm is likely to perform well. A low-dimensional representation that shows the separation of groups in supervised classification, or of clusters in a cluster analysis, may give valuable insight. It may indicate gross features of the distribution of data, and give visual clues that highlight differences between one algorithm and another. Where the main effect of tweaking an algorithm is to change which observations are misclassified, the plot will show this. Insight is often more helpful than a 0.1% gain in the cross-validation estimate of predictive accuracy.

Figure 2 was obtained by using the "proximities" from a `randomForest` discriminant rule to derive a low-dimensional representation. The figure legend gives the details. The plot identifies three points where the class labels seem in doubt. Plots of discriminant scores from R's `lda` (*MASS* package) or of the ordination scores from `svm` (*e1071* package) with default parameters, do not show the same clear separation. Why?

4 The Interpretation of Model Parameters

An unequivocal interpretation is usually impossible when there are multiple explanatory features that might be included in the model, perhaps measured with different accuracies. Typically, it is necessary to appeal to other supporting sources of information. Parameter estimates, even if highly significant statistically, cannot necessarily be taken at face value. I will note several instructive case studies. Even if not highly typical of the problems tackled by data miners, they have lessons of which data miners should be aware.

A referee has made the point that whether observational studies are effective in any particular circumstance will depend on the importance, subtlety and nature of the inference. Where the interpretation of parameter(s) is an issue, and there are multiple explanatory features, there is inevitable subtlety.

4.1 Smoking and lung cancer

Notwithstanding the strength of the link between smoking and lung cancer, with papers making the link appearing in the late 1920s, it was not until the 1950s that the connection was placed beyond reasonable doubt. Only when it was clear that multiple independent lines of evidence all pointed in the same direction were the most tenacious critics silenced. See Freedman (1990) for further commentary on the history, on the statistical issues, and for a number of other examples.

Effects that are much smaller than in the connection between smoking and lung cancer may be hard or impossible to tease out, especially if several factors are involved and no one factor strongly predominates.

4.2 Hormone Replacement Therapy

The health effects of hormone replacement therapy (HRT) have been a subject for extensive investigation over a long period of time, with extensive data now available both from observational and from experimental studies. This large collection of studies offers data analysts a unique opportunity to compare results between experimental and observational studies.

Case-control studies, as in Varas-Lorenzo et al (2000), are among the best-regarded of the observational studies. In these "cases", i.e., individuals who have the disease, are first identified. These are then matched with disease-free "controls", chosen to be as similar as possible in all respects except perhaps use of the therapy, in this case HRT. The hope is that over subjects as a whole, disease status will be the same as if the assignment to receive HRT had been done randomly. Almost inevitably the matching is not completely effective, and regression must be used to adjust for remaining differences. If an important explanatory variable is omitted from the adjustment (perhaps, as suggested below, childhood socio-economic status), conclusions may be fatally compromised.

Contrast such studies with experimental studies such as are reported in Rossouw et al (2002). As they enrol, participants are randomly assigned either to HRT or to a placebo, perhaps subject to restrictions that maintain a numeric balance between treatment and control groups. Strict adherence to randomization protocols ensures the identity of the treatment and control populations.

A large meta-analysis of the "best" quality cohort and other observational studies (Stampfer and Colditz, 1991) found a relative reduction in coronary heart disease (CHD) risk of 50% from any use of HRT. Where population based studies gave more or less definitive results, they agreed broadly in their conclusions, to the extent that Stampfer and Colditz could claim

> Overall, the bulk of the evidence strongly supports a protective effect of estrogens that is unlikely to be explained by confounding factors.

Broad agreement across the different studies does not however mean that the estimates are correct. Few would now defend Stampfer and Colditz's conclusion, for reasons that will now be discussed.

The experimental results showed that, far from reducing CHD risk, risk was increased. One large randomized controlled trial (Rossouw et al, 2002) found that HRT use increased CHD hazard by a factor of 1.29 (95% CI 1.02–1.63), after 5 years of follow-up.

This was particularly anomalous because the results of the observational studies have been consistent with the results of randomized trials for other outcomes – breast cancer (increased risk for the combined oestrogen/progesterone HRT; for a 50-year old from 11 in 1000 to maybe 15 in 1000), colon cancer (reduced risk), hip fracture (reduced risk, but diet, exercise and other drugs can achieve the same or better results) and stroke (increased risk; for a 50-year old from 4 in 1000 to 6 in 1000). See again Swan et al (2006) and e.g., Rossouw et al (2002) for further details and references.

Lawlor et al (2004) discuss why there is agreement for most outcomes, but not for CHD. Childhood socio-economic indicators are known to be important as predictors of CHD, independently of adult socio-economic status (SES), behavioural and physiological risk factors. This is not true for the other outcomes considered. Additionally, the use of HRT is "strongly socially patterned"; those with low childhood SES less commonly used HRT. Consider now individuals with low childhood SES, but high adult SES. Their low childhood SES is associated with low use of HRT and consequent lowered risk of CHD. In the analysis, the only adjustment is for their high adult SES. The benefit derived from non-use of HRT is wrongly ascribed, in the analysis and its associated interpretation, to their high adult SES.

If this account is correct, it highlights the importance of accounting properly for socio-economic effects. When studying an outcome of interest from an observational public health study, it is important to ask whether the simpler type of model that can account for breast cancer risk is adequate, or whether the situation that pertains to CHD risk is more likely.

4.3 Other examples and references

Do airbags save lives? The available US data are not encouraging, if analyzed with care. See Meyer and Finney (2005), and articles in a forthcoming issue of *Chance* that will continue the discussion, now with corrected data. The data, although extensive, suffer from a version of the reject inference problem –

they are from accidents that are sufficiently serious that at least one car was towed from the scene. Estimates of the effect of airbags change spectacularly with changes in the other factors that are incorporated into the model.

Leavitt and Dubner (2005) have a number of examples that illustrate the care that must be taken in bringing together multiple sources of evidence, in order to reach conclusions that seem reasonably secure. Their account of the reasons for the reduction in US crime rates in the 1990s, which I find convincing, has attracted huge controversy.

Rosenbaum (2002) teases out practical implications of the use of observational rather than experimental data, using for illustration a number of interesting examples. The insights in this important book have received less attention than they deserve in the statistical community, and scant attention in the data mining community. The brief final chapter, entitled "Some Strategic Issues", makes a number of specific suggestions that merit attention.

5 Re-engineering the publication process

Advances in computer technology allow and demand large changes in the reporting of data, in data analysis, in the total content of publications, and in access to the separate components of the content (Maindonald, 2005). Data mining is among the areas where the potential for change and innovation is greatest. Code and data that are used in papers should be available as a matter of course, preferably as part of a compendium (Gentleman and Lang, 2004) such as will now be discussed, which the reader can readily process through a computer program to create a version of the final paper. The compendium should include or give access to

- the text of the paper
- the data on which it is based, and
- the code used for analysis and for generation of tables and graphs.

The notion of reading a paper is substantially enlarged, to include interaction with the processes involved in moving from data to analysis to published paper.

The `noweb` literate programming syntax (Johnson and Johnson, 1997) is a suitable vehicle for the implementation of these ideas. My experience has been with the implementation in the R system (R Core Development Team, updated regularly). The function `Sweave` (Leisch, 2006) provides a flexible framework for mixing text and R code in an enhanced LaTeX document for automatic report generation. When processed through R's `Sweave` function, markup instructions that surround the R code chunks determine which chunks, and which of the output generated by the code, will be included in the final LaTeX document. Output may include tables and figures.

Gentleman and Lang (2004) argue strongly for the provision of an Sweave type compendium for any paper that presents results of genomic analyses, as a matter of standard practice. Users can then know with certainty the steps that have been followed. Benefits include the opening to scrutiny of any biases in the analysis protocols, and a ready ability to reproduce results and test their sensitivity to analysis choices.

The arguments are surely equally cogent for journals and conferences that publish data mining papers. Provision of `Sweave` type features is a reasonable requirement for any language that is intended for scientific use. A present serious limitation of `Sweave` is that code that appears in the LaTeX document has comments stripped from it.

An `Sweave` version of this present paper is available from the web page `http://www.maths.anu.edu.au/~johnm/dm/ausdm06/ausdm06-jm.Rnw`. The R packages *hddplot*, *mclust* (which includes the `diabetes` dataset), *randomForest* and *xtable* must be installed.

The file **ausdm06-jm.Rnw**, when processed through R's `Sweave` function, yields a LaTeX file and associated graphics file from which this present paper can be generated.

6 Final Comments

The issues that I have raised are all in a sense statistical, though not always receiving the attention that they deserve in statistics courses. Here, I will comment on the different traditions of data mining and of statistics, and on the large area of interest that they have in common.

6.1 Different traditions of data analysis

Statistics started as a discipline that had a strong practical orientation. The small number of statistics departments that predated World War II likewise had a strong practical orientation. The three decades that followed World War II saw the widespread establishment of statistics departments, now with a strong theoretical focus. Many of the teachers saw statistics as primarily a mathematical discipline. Over the intervening years, the teaching of statistics has slowly matured to pay more attention to applications, though this change still has some way to go. Over this same period, theory and computing have moved in synergy to bring been huge advances both in theory and in practical computing tools.

The R system (R Core Development Team, updated regularly) is an outstanding product of the new synergy between theory, computing and practice. It demonstrates what is possible when experts co-operate widely across national boundaries. It promises larger achievements yet, more in tune with modern ideas of computer systems.

Where academic statistics took mathematics and a range of practical demands as its points of departure, data mining has taken computing, algorithmics and data bases as its points of departure. It has thrown out a variety of challenges to statistics – challenges which I think valuable for the future development of statistics. A specific challenge is to make statistical methodology available to those who, while bypassing much of the mathematical theory, wish to have access to the fruits of that theory. Simulation is for this as well as for other purposes highly important, especially as it sometimes offers a way ahead in cases where the theory is intractable.

Data miners face, likewise, challenges from the statistical tradition, beyond those raised earlier in this paper. Among these is the challenge to marshal computing skills and tools effectively. Standalone tools are typically deficient in the data manipulation and graphical abilities needed to use them effectively, require the mastering of their own idiosyncratic user interfaces, and do not penetrate widely into the communities where they might find use. Contrast this with the use of R or another such system, as a framework for developing new software, and as an interface into the end product. Many data miners, sensitive to the benefits of such a common interface, are already using and contributing to R. Those who do not wish to go this route have the challenge of finding or developing a system that can equal or better R: in the

expertise that has contributed to its development, in its range of abilities, in the trustworthiness of its output, in its cohesion, in its linkages into other systems, in automated checks that impose minimal standards of consistency across the system as a whole, in the use of the internet to give access to R and to associated resources, in its relative ease of use, and in the wide extent of its user community.

Hard-won insights from both the practical and theoretical streams of statistical development require the attention of data miners. I attach high importance to issues that I have noted in this paper, centering around source and target population, realistic assessment of predictive accuracy, the interpretation of model parameters, and the insights that may be derived by comparing results from observational studies with results from experimental studies.

6.2 The training of data miners

To what extent is understanding of statistical issues, such as I have canvassed, required for effective data mining? Relatively automated use of data mining tools will give better results for some applications than for others. Without however some sense of what issues are important, how will the data analyst know the difference? Anyone who expects to make data mining a substantial part of their work will do well to take time and effort to get on top of the issues that I have canvassed. They can all be understood without recourse to advanced mathematics. To be effective, these points must be reinforced by exposure to, and understanding of, the practical data analysis contexts in which they arise.

6.3 Course materials

Course materials for a course component that includes a statistically focused commentary on data mining are available from my website (Maindonald, 2006). Data that the laboratory exercises explore include two substantial datasets that were mentioned above – the US forest cover data and the US car accident data.

Acknowledgement

I am grateful to a referee for helpful comments and several useful leads.

References

Ambroise, C and McLachlan, G J, 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, **99**:6262–6266.

Elkan, C 2001. Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000 (postscript) (pdf). In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD'01), pp. 426-431. `http://www-cse.ucsd.edu/users/elkan/kddcoil.pdf`

Freedman, D 1979. From association to causation: some remarks on the history of statistics. *Statistical Science* **14**:243-258.

Gentleman, R and Lang, D T 2004. Statistical Analyses and Reproducible Research. Bioconductor Project Working Papers. Working Paper 2. `http://www.bepress.com/bioconductor/paper2`

Hand, D J 2006. Classifier technology and the illusion of progress. *Statistical Science* **21**:1-14.

Heckman, J J 1979. Sample selection bias as a specification factor. *Econometrics* **47**:153-161.

Johnson, A L and Johnson, B C. 1997. Literate programming using noweb. *Linux Journal*, 64-69, October 1997. `http://members.shaw.ca/andrew-johnson/noweb_lj.pdf`

Keogh, E 2006. The UCR Time Series Data Mining Archive `http://www.cs.ucr.edu/~eamonn/TSDMA/index.html` Riverside CA. University of California – Computer Science & Engineering Department.

Keogh, E, Xi, X, Wei, L & Ratanamahatana, C A 2006. The UCR Time Series Classification/Clustering Homepage `www.cs.ucr.edu/~eamonn/time_series_data/`

Lawlor, D A, Davey Smith, D F and Ebrahim, S 2004. Commentary: The hormone replacement – coronary heart disease conundrum: is this the death of observational epidemiology? *International Journal of Epidemiology* **33**:464–467.

Lawlor, D A, Davey Smith, D F and Ebrahim, S 2004. Socioeconomic position and Hormone Replacement Therapy use: explaining the discrepancy in evidence from observational and randomized controlled trials. *American Journal of Public Health* **94**:2149–2154.

Leavitt, S D and Dubner, S J, 2005. *Freakonomics. A Rogue Economist Explores the Hidden Side of Everything*. William Morrow.

Leisch F 2006. Sweave User Manual. `http://www.ci.tuwien.ac.at/~leisch/Sweave`.

Maindonald, J H 2005. Data, science and new computing technology. *New Zealand Journal of Science* **62**:126-128.

Maindonald, J H 2006. Statistical Commentary on Data Mining: Course Materials. `http://www.maths.anu.edu.au/~johnm/courses/dm/`

Maindonald, J H, 2003. The role of models in predictive validation. Invited Paper. 54th session of the ISI, Berlin, 2003. `http://www.maths.anu.edu.au/~johnm/dm/isi2003-models.pdf`

Maindonald, J H and Braun, W J, 2nd edn, 2006, in press. *Data Analysis and Graphics Using R – An Example-Based Approach.* Cambridge University Press. `http://wwwmaths.anu.edu.au/~johnm/r-book.html`

Maindonald, J H and Burden, C J 2005. Selection bias in plots of microarray or other data that have been sampled from a high-dimensional space. In R May and A J Roberts, eds, *Proceedings of 12th Computational Techniques and Applications Conference CTAC-2004*, volume 46, pp. C59–C74. `http://anziamj.austms.org.au/V46/CTAC2004/Main`.

Meyer, M C and Finney, T 2005. Who wants airbags? *Chance* **18**:3-16. `http://www.stat.uga.edu/~mmeyer/airbags.htm` See also `http://wwwmaths.anu.edu.au/~johnm/datasets/airbags/`

Newman, D J, Hettich, S, Blake, C L and Merz, C J 1998. UCI Repository of machine learning databases `http://www.ics.uci.edu/~mlearn/MLRepository.html` Irvine, CA: University of California, Department of Information and Computer Science.

R Core Development Team. *An Introduction to* R. `http://cran.r-project.org`

Ripley, B D 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rosenbaum, P R 2002. *Observational Studies*, 2nd edn. Springer-Verlag.

Writing Group for the Women's Health Initiative Investigators 2002. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* **288**:321-333.

Salzberg, S L 1997. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining. and Knowledge Discovery* **1**:317–327.

Stampfer M J and Colditz G A 1991. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Preventative Medicine* **20**:47–63.

Swan, N, Fry, R, McPherson, A, Trevena, L and Davis, S 2006. Hormone Replacement Therapy - Part Two - Radio National Summer. `http://www.abc.gov.au/rn/healthreport/stories/2006/1530042.htm#`

Varas-Lorenzo C, Garcia-Rodriguez L A, Perez-Gutthann S, Duque-Oliart A 2000. Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. *Circulation* **101**:2572-2578.

Zhu, X, Ambroise, C, and McLachlan, G J 2006. Selection bias in working with the top genes in supervised classification of tissue samples. *Statistical Methodology* **3**:29–41.

Accuracy Estimation with Clustered Dataset

Ricco Rakotomalala* **Jean-Hughes Chauchat*** **Francois Pellegrino****

*ERIC Laboratory - University of Lyon 2
Lyon - France
Email: {Ricco.Rakotomalala,Jean-Hughes.Chauchat}@univ-lyon2.fr

**Laboratoire Dynamique du Langage - University of Lyon 2
Lyon - France
Email: Francois.Pellegrino@univ-lyon2.fr

Abstract

If the dataset available to machine learning results from cluster sampling (e.g. patients from a sample of hospital wards), the usual cross-validation error rate estimate can lead to biased and misleading results. An adapted cross-validation is described for this case. Using a simulation, the sampling distribution of the generalization error rate estimate, under cluster or simple random sampling hypothesis, are compared to the true value. The results highlight the impact of the sampling design on inference: clearly, clustering has a significant impact; the repartition between learning set and test set should result from a random partition of the clusters, and not from a random partition of the examples. With cluster sampling, standard cross-validation underestimates the generalization error rate, and is deficient for model selection. These results are illustrated with a real application of automatic identification of spoken language.

Keywords: Accuracy estimation, Supervised Learning, Clustered dataset.

1 Introduction

Most of the time, learning is organized on a dataset which is a mere sample taken from the universe to which the results are to be generalized. Concerning a supervised learning task, measuring the quality of the generalization is known as the "assessment" (Stone 1974). Several measures of the quality exist (Lavrac 1999)(generalization error rate, sensitivity, specificity, ROC curve, ...). They are usually obtained through resampling methods which are often applied under the hypothesis that the learning set is a simple random sample of observations (independent and identically distributed - *iid* - observations) from the universe of interest (Efron & Tibshirani 1995).

In practice, this hypothesis is rarely verified; the available dataset is often the result of cluster sampling, or (more generally) two-stage sampling:

- patients from a sample of hospital wards;
- X-ray images, from various angles, of a sample of patients;
- children from a sample of classrooms or schools;
- land samples from an oil drilling rig sample...

From these examples, one understands that access to individuals has been only possible through the cluster. For instance, the patients cannot be selected one by one; a sample of hospitals is selected, in which all patients (cluster sampling) or a sub-sample (two stage sampling) are observed. Clusters are not the result of some computation on the available data. The clustering structure is inherent to the data, part of data collection, "meta-information" needed to understand and to use the data.

In this paper, standard cross-validation results (under simple random sampling assumption) are shown to be overly optimistic when the dataset is actually clustered. A modification to the cross-validation procedure that accounts for the sampling design is suggested, according to the usual applications of resampling methods in sample survey problems (Shao & Tu 1995). The proposed modification is supported, first, by an application to simulation data and secondly, by an application to speech recognition.

In section 2, an adaptation of cross validation to cluster sampling will be examined. In section 3, using a simulation, the sampling distribution of the generalization error rate, under cluster or simple random sampling, will be compared to the respective true values; the consequences of model selection will be examined. In section 4, results of an application to a real life database will be presented: the problem is to automatically recognize the language spoken by a sample of individuals using the physical analysis of the audio signal of their voices. Related work is reviewed in section 5. Lastly, a conclusion and future works are presented in section 6.

2 Adapting cross-validation to cluster samples

The true error rate (Err) is a measure of how accurately the classification, built with the learning sample, would be if they were applied to the whole universe. As the universe cannot be observed completely, a learning set (the sample) is used to infer about the universe, and only an estimate of the error rate ($\hat{Err}$) can be computed. In this case, the sample selection is an integral part of the inference process, and any evaluation should account for it.

In reality, survey design information is seldom made available, notably absent on the benchmark datasets available on Internet servers (for example, those from the UCI repository (Bay 1999)). In fact, statistics and methods proposed to measure the error rate rely on a simple random assumption, hardly ever realized.

The usual cross validation estimate (under *iid*, independently and identically distributed data, hypoth-

esis) of the error rate (Err) and the adaptation to cluster sampling are described in this section.

2.1 Cross-validation on *iid* samples

The cross-validation method (in J folds) usual to machine learning proceeds as follows (Stone 1974), supposing *iid* observations:

1. a sample of n individuals are obtained from the population;
2. the n cases on the learning set are randomly split in J folds of size n_j, usually $n_j = \frac{n}{J}$;
3. the learning algorithm is applied to the whole data set, save part j;
4. the rules learned in (3.) are applied to the cases left out, fold j, and an error rate $\hat{Err}_{(j)}$ is measured;
5. the "generalization error rate" $\hat{Err}$ is estimated as

$$\hat{Err} = \sum \frac{n_j}{n} \hat{Err}_{(j)}$$

The estimator $\hat{Err}$ is biased: $E(\hat{Err}) > Err$ because the samples used for learning in (2.) are approximately size $n\frac{(J-1)}{J} < n$; the bias decreases as J increases (of course with *iid* samples), but the random variation of $\hat{Err}$ and computation time grow with J. But, in spite of these drawbacks, it is proved that in the particular case of leave-one out cross-validation for instance, the worst-case error of this estimate is not much worse than that of the training error estimate (Kearns & Ron 1997).

2.2 Cross-validation on cluster samples

If the learning set was obtained through cluster sampling, then the standard procedure must be modified as follows:

1. A sample of n individuals, grouped in G clusters of respective size n_g observations ($g = 1, \ldots, G$);
2. The G clusters are subdivided in J parts, part j thus comprising G/J clusters, with $n_j = \sum_{g \in j} n_g$ observations;
3. Steps (3.) to (5.) of the standard procedure are then applied.

There is a "cluster effect" when the variability within the clusters is small compared to that of the whole population; then, for a given sample size n, the true generalization error rate increases. This must be apparent in the cross-validation estimation process.

3 Application on simulated data

The main interest of a simulation model is that the true error rate is known. In effect, one can either compute the theoretical distribution-based error rate, or, using a random number generator, create as many individuals as needed to construct the test data set and estimate the error rate with controlled precision. The latter option was chosen for this paper.

The example described here uses a decision tree learning algorithm and two explanatory variables (for an easier interpretation of the charts). We will deepen our analysis later (sec. 3.3) by studying the effect of increasing number of attributes and the choice of the learning algorithm.

3.1 The simulation model

For this problem, the objective of the learning algorithm is to distinguish between two classes: positives (+) and negatives (o). In the universe of reference, individuals are grouped in clusters of size $2 \times m$, of which m are positive ("+") and m are negative ("o"). The positives, independently their cluster, are distributed according to a bivariate normal distribution with zero mean and $s^2 \times I$ covariance matrix, where s is a constant and I is an identity matrix. The negatives of a cluster are also normally distributed with the same covariance matrix but their mean is located on the circle of radius 1. The (negatives) cluster means are uniformly and randomly distributed on the unit circle (for example figure 1.a). The dataset contains g clusters, that is, $n = 2 \times m \times g$ individuals.

Hence there are three parameters: s, the dispersion of each half cluster about its mean, m, the cluster size for one class value, and g, the number of clusters on the learning set.

For each value of these parameters ($s = 0.1$, 0.2 , 0.5 , 1; $m = 5$, 10 , 20 , 40; $g = 10$), 100 learning sets were randomly generated, as well as a testing set of $1,000$ clusters for the precise estimation of the true generalization error rate.

We used the C4.5 decision trees algorithm (Quinlan 1993), that can approximate any linear or non-linear boundary with broken lines made up of segments parallel to either axis.

For each learning sample (Figure 1):

1. the decision trees were constructed using the learning set;
2. the "true" error rate is approximated by the error of the tree model on the large test set (Figure 1.d);
3. the cross-validation error rate estimate was computed by taking the clustering into account (sec. 2.2);
4. the cross-validation error rate estimate was computed as if the data were *iid* (disregarding the clustering).

Let's review the algorithm using the example in figure 1.a which comprises 10 clusters, that is $n = G \times 2 \times m = 10 \times 2 \times 10 = 200$ examples. A tree (figure 1.b) is constructed on that base and the error rate is zero (figure 1.c); estimation of the true error rate, on the large dataset, is shown figure 1.d.

For our experimentation, we choose a $J = 10$ clustered cross-validation, it makes a good bias-variance trade-off of the error evaluation (Dietterich 1998). Because we have 10 clusters in the synthetic dataset, one cluster is removed at each step of our clustered cross-validation. One of those steps is depicted in figure 2: when learning on the 9 leftmost clusters, the resulting tree (figure 2.a) poorly classifies 6 of the 10 negatives set aside (figure 2.b).

3.2 Simulation results

Many interesting elements can be underlined (Figure 3):

- the true error rate Err increases with s, the relative cluster variability, and decreases with m, the cluster size;
- standard cross-validation, disregarding the cluster effect, severely underestimates the error rate;
- estimation bias increases with the cluster effect: bias is maximum when $s = 0$, that is when all individuals are identical ($s = 0.1$ and $s = 0.2$ on figure 3);
- cross-validation accounting for cluster effect slightly overestimates the true error rate; as mentioned earlier, this was expected because the cross-validation uses, at each step, a fraction of the available sample to construct the prediction model.

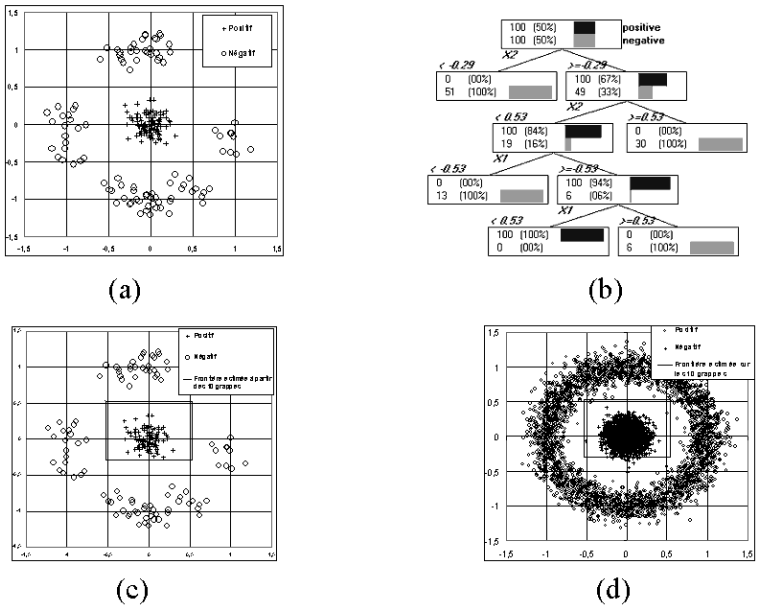


Figure 1: Learning on a sample of 10 clusters (a, b, c), applying classification rules on the generated test set (d)

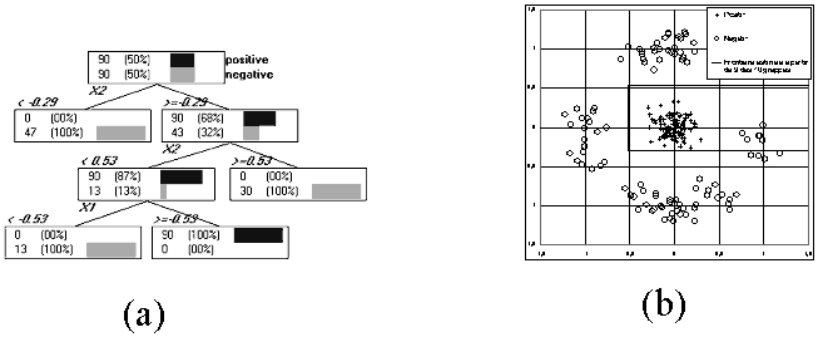


Figure 2: One step of clustered cross-validation

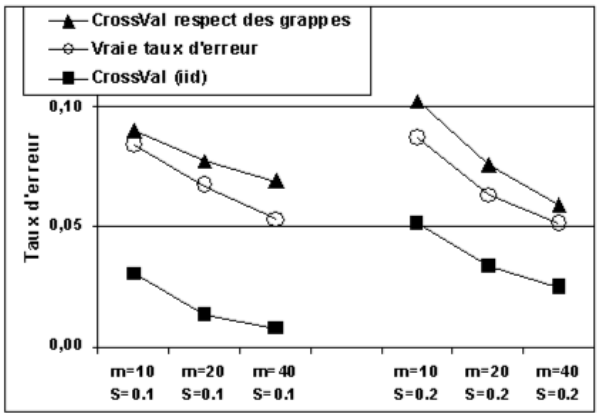


Figure 3: True and estimated error rate for $G = 10$ clusters: $s = 0.1$, $s = 0.2$ et $m = 10$, 20 et 40. Average on 100 simulations for each case.

3.3 Simulations using alternative algorithms and multiple attributes

The preceding simulations illustrated cross-validation with clustered data, and some reasons why clustering must be accounted for. In this section, the error estimation bias is examined under varying experimental set-ups :

- increased number of predictive attributes : dim = 2, 3, 5, 10;
- different learning algorithms : C4.5 Decision Tree, 1-Nearest Neighbour, Naive Bayes (Hastie, Tibshirani & Friedman 2001).

The simulation population is similar to the previous one, but here the negatives (o) are centered on a random point of the hyper-sphere of dimension dim = 2, 3, 5 or 10.

3.3.1 General results for each learning algorithm

Experiments confirm the results of section 3.2; whatever the learning algorithm and whatever the dimension of the space:

- standard cross-validation (assuming iid observations) always under-estimates the true error rate;
- cross-validation accounting for clusters is much closer to the true error rate.

That is :

- With the decision tree (C4.5), standard cross-validation (iid) severely under-estimates the error rate, regardless of the sample size ($n = 10 \times 2 \times m$), the clustering effect (here, noted s), and the dimension of the space;
- The nearest neighbour method (1-NN) is very sensitive to the clustering effect : if, in each class, members of a given cluster are closer to one another that those of other clusters ($s = 0.1$), then standard cross-validation dramatically under-estimates the error rate by giving an estimate of zero; notably, the true error rate with 5 predictive attributes is about 15%; and more than 30% with 10 predictive attributes. The bias of the standard method is important even with moderate clustering effect ($s = 0.3$), especially when many predictive attributes are at play ($dim = 5$ or 10). When clustering effect is smaller ($s = 0.5$) and many predictive attributes are used, the 1-NN method learns almost nothing (the true error rate increases to 50% for two classes, even with samples of size $n = G \times 2 \times m = 10 \times 2 \times 50 = 1000$ cases) yet the standard method remains " optimistic ", and the bias increases with the sample size.
- With Naive Bayes algorithm, the three error estimates are almost identical when clustering effect is important ($s = 0.1$) : it is a peculiar interaction of our synthetic data and the Naive Bayes estimator. In effect, this method discretizes continuous attributes into intervals before learning. Here, because the low variance clusters of negatives orbit around a kernel of positives, the pre-processing of each attribute always give the correct three-class discretization, the positives in the center class flanked by classes of negatives. This artifact decreases when the variance increases, causing positives and negatives to intertwine and classes to be less homogeneous. Under-estimation (under the iid assumption) increases when clustering decreases (larger s), and increases in sample size and number of attributes.

Language	Speaker	Rec./Speaker	Avg.length/Rec.
German	10	20	21,9
English	10	15	17,6
Spanish	10	15	20,9
French	10	10	21,9
Italian	10	15	21,7
TOTAL	50	750	

Table 1: MULTEXT dataset, cluster structure.

3.3.2 Model selection

The results are summarized in Fig. 4 which shows that standard cross-validation can yield to very poor results if the data set is clustered, and if the clustering effect is significant ($s = 0.1$):

- In this case, judging by the standard *iid* cross-validation (figure 4.a), the nearest neighbour (1-NN) is always better than the other strategies because its error rate is always zero, though the true error rate (figure 4.c) may be larger than that of naive Bayes as soon as at least three attributes are involved ; this last result is confirmed by cross-validation under clustering (figure 4.b);
- Still under strong clustering ($s = 0.1$), the decision trees C4.5 seem better than Naive Bayes in every situation, as indicated by standard cross-validation (figure 4.a). This is deceiving, for the Naive Bayes method outperforms the three other methods (figure 4.b and 4.c) when at least three predictive attributes are present (see 3.3.1).

The simulation is, of course, particular. Still, depending on the data, standard cross-validation can lead to a very poor choice of the learning method.

4 Application to real data: speech recognition

4.1 Statement of the problem

Language identification from sound bites is an emerging domain of automatic speech processing. In this era of international and global media, the stakes are numerous, be it man-machine interface or computer-assisted human dialogue.

Most of the approaches developed so far use statistical modelling of phonetic (nature of sounds) and phonotactic (how the sounds are assembled) characteristics of the various languages (Zisman & Berkling 2001). Such approaches require vast amounts of sound recordings along with their phonetic transcriptions (entirely supervised learning).

Data mining techniques, with innovative parameterization, can give convincing results with partially supervised learning and smaller learning data sets.

4.2 The task and the data

The experiments were conducted on the multilingual set MULTEXT (Campione & Veronis 1998). This database contains audio recordings in five European languages (English, French, German, Italian and Spanish) spoken by 50 individuals (5 mens and 5 women per language). Each recording represents a 5-sentence text and each speaker read between 10 and 20 of those short texts. Table 1 summarizes the clustered structure of the data, one cluster corresponding to one individual's recordings.

The task is to identify the language spoken on a recording different from those used for learning.

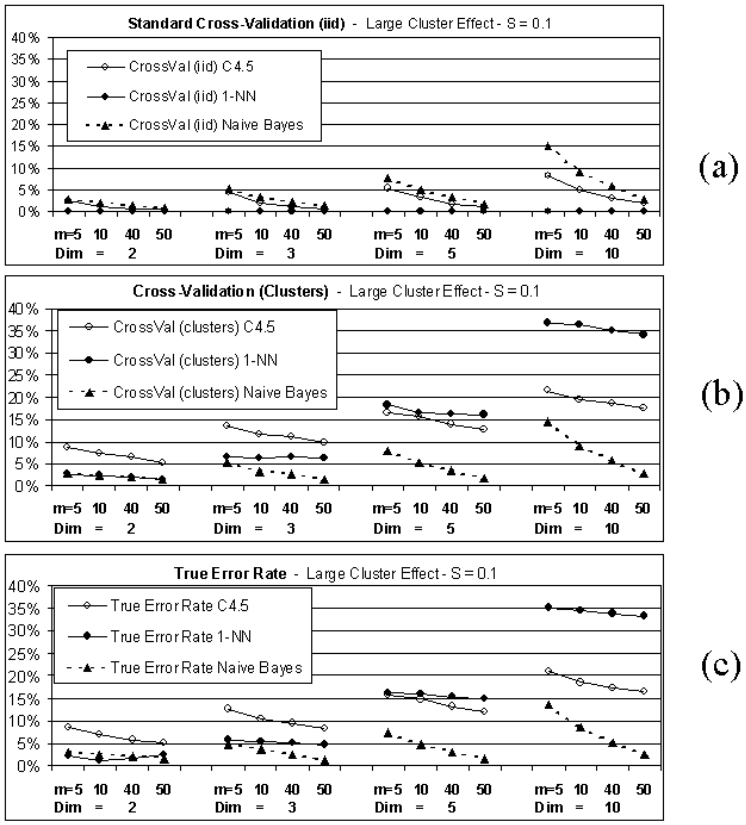


Figure 4: Standard *iid*, cluster cross-validation error rate estimates and true error rates with large cluster effects

4.3 The descriptors

Classical approaches are based on the spectral analysis of the audio signal for which the cluster effect is well-recognized (the spectrum informs on the identity and the language of the speaker). The approach followed here used a parameterized rhythmic space for which the cluster is theoretically less important.

An automatic segmentation of the audio signal in pseudo-syllables was first realized (Farinas & Pellegrino 2001). These units are composed of one or more consonant segments followed by one vocalic segment. They are correlated to the rhythmic structure of the language and can thus be used to identify the language. Each pseudo-syllable is set in a 5-dimensional space:

- Dc (total duration of the pseudo-syllable consonants, in ms);
- Dv (duration of the vocalic segment, in ms);
- Nc (number of consonant segments in the pseudo-syllable, unit free);
- Fo (fundamental frequency of the pseudo-syllable vowel, in Hertz);
- E (relative energy of the vowel, in dB).

For each recording, the parameter means, variances and covariances were computed using all the pseudo-syllables of the sound excerpt. Hence, 20 parameters are available for each statistical individual.

4.4 Comparing various approaches

Many learning algorithms were tested, they have different representation and learning characteristics (Hastie et al. 2001): C4.5 Decision Tree (DT); 1-Nearest Neighbour (1-NN); Naive Bayes (NB); linear discriminant analysis (LDA); multi-layer perceptron (MLP).

Algorithme	**Standard (iid)**	**Clustered**
Decision Tree	25%	35%
1-NN	26%	37%
Naive Bayes	36%	48%
Linear Disc. Analysis	15%	20%
Multi-layer Perceptron	16%	21%
GMM	-	20%

Table 2: Cross-validation error rate estimate for each induction method

In all cases, cross-validation was performed, first not accounting for clustering (different recordings of the same speaker can be used for both learning and testing), and secondly accounting for the clusters (speakers for learning and testing are different).

Finally, a comparison with the EM (Expectation-Maximization) estimation algorithm on Gaussian mixture model (GMM), usually used in speech processing, was performed. Table 2 summarizes the results.

In spite of the small number of characteristics accounted for (average duration of the consonant segments, etc.), rhythmic modeling gives interesting results, in the order of 20% of erroneous identification, whether pattern recognition (GMM) or two usual data mining techniques are used. The complex parameter space seems to handicap DT, 1-NN and NB, more than LDA or MLP.

Moreover, accounting for the clustering modifies the estimated error rate whatever the learning algorithm used. These experiments illustrate that, when working with real data, clustering must be factored in to avoid serious underestimation of the generalization error rate.

5 Related work

The key point of this work is the necessity of accounting for the sample design of the learning dataset when defining the resampling procedure (cross-validation, bootstrap, etc.). Thus, subdividing (learning set, testing set) must be done randomly on the clusters rather than on the individuals; the same is true for leave-one-out.

There is little similar work. Most of the discussion has focussed on the optimal number of parts for cross-validation (Kohavi 1995), on the introduction of sophisticated resampling schemes (Dietterich 1998), or on bias correction with respect to the classifier (Efron & Tibshirani 1997). Some authors have introduced stratified cross-validation, the objective being to maintain the distribution of classes among the subdivisions. There is no sound justification to this, the underlying idea being the reduction of the variability of the models produced at each step. They think that, and the work presented here agrees with their hypothesis, the strategy is only efficient when the initial sample is itself stratified, that is the frequency of each class is explicitly reproduced in the sample (Kohavi 1995).

In this paper, we note that the *iid* cross-validation systematically underestimates the true error rate when we have clustered dataset. It is biased. But, the behavior of the variance of the clustered estimator is not clear. It appears that the standard estimators of the variance of the cross-validation error rate is often underestimated (Bengio & Grandvalet 2004). If we use the same estimator in our context, one can think that we obtain the same result. But it is obvious that it will be necessary to study in detail this assertion which rests only on one intuition.

Another problem is model selection. Some works show that a bad estimation of the generalization error rate can be sufficient for model selection if we obtain a clear picture of the relative performance of the learning algorithms (Petrak 2000). The author claims that using a moderate subsample of the dataset allows to ranking the algorithms. When we want to adapt this approach to clustered dataset, the unit of the sample must be the clusters and not the individuals.

Following the same idea, a bad generalization rate estimation may be sufficient for model selection. One can wonder whether the bias of the standard cross-validation method (*iid* assumption), which underestimates the generalization error rate when instances were sampled from clusters of data, is constant across different learning methods. If it is true, it appears that the *iid* cross-validation can nevertheless used for model selection, whatever the sampling scheme. The answer appears complex, depending on the algorithm characteristics, on the nature of the discrepancy, bias (a systematic difference with respect to the true value) or variance (discrepancies due solely to sampling). Our first results (see section 3.3) show that, using the synthetic data, there is no apparent correlation between the rank of the methods as ordered by standard cross-validation, and their rank as ordered by the cross-validation under clustering. In-depth analysis are still awaited; in the meantime, caution dictates that sampling design be accounted for cross-validation, even if the ultimate goal is model selection.

6 Conclusion

In this paper, it is shown that correct assessment of the predictive model by resampling methods must account for the sampling scheme used for the construction of the learning set. With cluster sampling, standard cross-validation significantly underestimates the generalization error rate and seems not efficient for model selection.

The approach proposed here can be extended to other sampling strategies (stratification, unequal probability sampling). Then, under those different schemes, the size and direction of the standard cross-validation estimation bias must be determined.

References

Bay, S. (1999), 'The UCI KDD archive [http://kdd.ics.uci.edu]', Irvine, CA: University of California, Department of Computer Science.

Bengio, Y. & Grandvalet, Y. (2004), 'No unbiased estimator of the variance of k-fold cross-validation', *Journal of Machine Learning Research* **5**, 1089–1105.

Campione, E. & Veronis, J. (1998), A multilingual prosodic database, *in* 'Proc. of ICSLP'98', Sydney.

Dietterich, T. (1998), 'Approximate statistical tests for comparing supervised classification learning algorithms', *Neural Computation* **10**(7), 1895–1924.

Efron, B. & Tibshirani, R. (1995), Cross-validation and the bootstrap : Estimating the error rate of a prediction rule, Technical Report 176, Department of Statistics, University of Toronto.

Efron, B. & Tibshirani, R. (1997), 'Improvements on cross-validation: The 0.632+ bootstrap method', *JASA* **92**(438), 548–560.

Farinas, J. & Pellegrino, F. (2001), Automatic rhythm modeling for language identification, *in* 'Proc. of Eurospeech '01', Aalborg, Scandinavia, pp. 2539–2542.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Kearns, M. J. & Ron, D. (1997), Algorithmic stability and sanity-check bounds for leave-one-out cross-validation, *in* 'Computational Learing Theory', pp. 152–162.

Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* 'Proceedings of the International Joint Conference on Artificial Intelligence - IJCAI'95'.

Lavrac, N. (1999), 'Selected techniques for data mining in medicine', *Artificial intelligence in medicine* **16**, 3–23.

Petrak, J. (2000), Fast subsampling performance estimates for classification algorithm selection, *in* 'ECML-00 Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination', pp. 3–14.

Quinlan, J. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Shao, J. & Tu, D. (1995), *The Jackknife and Boostrap*, Springer.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions', *Journal of the Royal Statistical Society* **B 36**, 111–147.

Zisman, M. & Berkling, K. (2001), 'Automatic language identification', *Speech Communication* **35**(1).

Towards Automated Record Linkage

Karl Goiser **Peter Christen**

Department of Computer Science,
The Australian National University,
Canberra ACT 0200, Australia
Email: {Karl.Goiser, Peter.Christen}@anu.edu.au

Abstract

The field of Record Linkage is concerned with identifying records from one or more datasets which refer to the same underlying entities. Where entity-unique identifiers are not available and errors occur, the process is non-trivial. Many techniques developed in this field require human intervention to set parameters, manually classify possibly matched records, or provide examples of matched and non-matched records. Whilst of great use and providing high quality results, the requirement of human input, besides being costly, means that if the parameters or examples are not produced or maintained properly, linkage quality will be compromised. The contributions of this paper are a critical discussion on the record linkage process, arguing for a more restrictive use of blocking in research, and evaluating and modifying the farthest-first clustering technique to produce results close to a supervised technique.

1 Introduction

Record Linkage is concerned with the process of identifying records from one or more datasets which refer to the same entities (e.g. people, organisations or objects) (Winkler 2006). Where applied to a single dataset, the process is known as *de-duplication*. The utility of Record Linkage lies in its ability to provide information that would otherwise be impossible or too costly to obtain. For example, a linkage of hospital records with motor vehicle accident data could provide information about the required procedures and outcomes for different types and severities of accidents (Christen & Goiser 2005, Winkler 2006). Record linkage is often used in the initial, preprocessing phase of data mining projects in order to enrich data or remove duplicate records. This paper is divided into three parts: this introduction, a critical discussion of the nature and some of the major issues of Record Linkage, and an experimental part which leverages the discussion and examines the possibility of conducting record linkage without human intervention.

1.1 The Record Linkage Process

Historically, the process predates computers (Gill 2001), but it wasn't until their advent and common use, together with increasing storage of information, that significant advances were made. Newcombe developed the basic idea of probabilistic linkage in the 1950's, and the mathematical foundation was set down by Fellegi and Sunter in 1969 (Fellegi & Sunter 1969).

Consider a population of *entities* (people, businesses, products, etc.) from which are drawn one or more *datasets*, some of whose *records* may refer to the same entities. The drawing down process may be the entering of information relating to a patient's hospital visit, the result of a credit card transaction, adding a new customer into a database, or recording a birth. Each entity may appear more than once in a single dataset (e.g. multiple credit card transactions, mothers giving birth, etc.) and in more than one dataset. The process of entering information about the entities into a dataset may be subject to errors such as typing mistakes, miss-spellings, optical character recognition errors, etc. There may also be differences due to the use of abbreviation or varying amounts of detail recorded. Thus finding the records which relate to the same entities can be seen to be non-trivial in the sense that no simple exact search, database join or sort could find them (Christen & Goiser 2005).

This paper assumes the linkage of two datasets, **A** and **B**, with neither dataset containing duplicate records (Christen & Goiser 2005). There may be some records in **A** which refer to the same entities as records in **B**, and it is the task of Record Linkage to find them. The linkage process involves two basic steps, *comparison* and *classification* (Winkler 2006). The comparison step takes pairs of records from the cross-product of the datasets, $\mathbf{A} \times \mathbf{B}$, and, for each pair, produces a vector of one or more values indicating the level of similarity or difference between attributes of the records which were compared. The values can be categorical, ordinal or numeric, but are generally real values in the range [0,1] with increasing values representing increasing similarity (Winkler 2006). The vectors place the comparison of a record pair into a space whose dimension is equal to the number of attributes compared. From these vectors, the classification step determines the class of each pair as either a *match*, or a *non-match* (Christen & Goiser 2005) (the *possible-match* classification used in the Fellegi and Sunter approach will be discussed below). Classification techniques are generally either *supervised* or *unsupervised* (Mitchell 1997). Supervised techniques use pre-classified data to generate a *classifier* which is then used to determine the class. Variations on these steps are possible, however, there must be some form of comparison between the records as well as a determination of the class.

Considering each attribute comparison, there are two distributions, one each for the matches and non-matches. Figure 1 shows this for a dataset used in the experiments in Section 3. The normal process is to select a threshold value which places as many matches as possible above, and as many non-matches below it.

Table 1: Confusion matrix of record pair classification

Actual	Classifications	
	Match	Non-match
Match	True match True positive (TP)	False non-match False negative (FN)
Non-match	False match False positive (FP)	True non-match True negative (TN)

Errors occur where comparisons of matches fall below the threshold, or non-matches above it. The confusion matrix in Table 1 describes the types of errors than can be produced (Christen & Goiser 2005).

It can be noted that, if the cost of making a false match is different from the cost of making a false non-match, it could be possible to find threshold values which, while increasing the number of false classifications, could decrease the overall costs. For example, in a cancer screening test, lowering the threshold value could decrease the number of undetected cancers (false negatives) at the expense of a higher number of erroneous detections (false positives), thus increasing the cancer detection rate.

In the traditional Fellegi and Sunter approach a third class is allowed: *possible-matches* (Fellegi & Sunter 1969). Compared record pairs are assigned to this class where the criteria are not strong enough to make a definitive match or non-match decision. The resulting pairs are referred to manual *clerical review* – where human judgement, with possible reference to more information, can allow a determination to be made. However, this can be problematic, as comparing even medium-sized datasets with only a few thousand records results in millions of comparisons, and a possible-match rate of only 1% will require thousands of manual classifications. These will take time – delaying the project – and adding to costs. Also, human decision-making has an inherent bias and error rate. As the motivation of this paper is to work towards an automated record linkage process, this class will not be considered further.

1.2 Blocking

Comparing the cross-product of $\mathbf{A} \times \mathbf{B}$ results in quadratic complexity, and is thus difficult or impossible for large datasets. Different *blocking* techniques such as standard blocking, sorted-neighbourhood, bigram indexing and canopy clustering have been developed to ameliorate this problem (Baxter, Christen & Churches 2003, Elfeky, Verykios & Elmagarmid 2002). The idea is to cheaply filter out obvious non-matches before executing the more detailed and time-consuming comparisons.

As an example of the process, standard blocking criteria uses record attributes such as postal/zip code to create blocks of similar records, with the comparisons then only being between records in the same block. A problem with this approach is that inaccuracy in the criteria can lead to records being placed in the wrong blocks, thus removing them from being compared with any potential matches. This can be solved by conducting several passes using different criteria each time (Winkler 2006). In (Winkler 2005), it is suggested that a 10-pass blocking strategy will reduce the number of record pair comparisons required for the linkage of two datasets with one million records each from 10^{12} to around $10^7 - 10^8$.

1.3 Modern Approaches

In recent years, researchers have started to incorporate techniques from Machine Learning, Data Mining, Information Retrieval, and Artificial Intelligence research to improve the linkage process. A popular approach (Bilenko & Mooney 2003, Chaudhuri, Ganjam, Ganti & Motwani 2003, Cohen, Ravikumar & Fienberg 2003, Yancey 2004, Zhu & Ungar 2000) has been to learn distance measures (like edit-distance) that are used for approximate string comparisons (Christen 2006). As shown in (Cohen et al. 2003), combining different learned string comparison methods can result in improved linkage classification. An Information Retrieval–based approach (Cohen 1998) is to represent records as document vectors and to compute the cosine distance between such vectors, while (Nahm, Bilenko & Mooney 2002) explore the use of support vector machines to classify record pairs.

Active learning is used in (Sarawagi & Bhamidipaty 2002) and (Tejada, Knoblock & Minton 2002) to address the problem of lack of training data. The basic idea is to iteratively select for human determination, a comparison which is the hardest for the technique to classify, then learn from that and build a better classifier. This has the effect of significantly reducing the number of manual determinations required. A hybrid system is described in (Elfeky et al. 2002) which utilises both unsupervised (clustering) and supervised (instance-based learning and decision trees) machine learning techniques. Active learning is also used in (Michalowski, Thakkar & Knoblock 2004), however, secondary data sources are used in place of human input, thus making their approach unsupervised. This work is part of an increasing interest in the application of record linkage to web-based data and technologies.

High-dimensional overlapping clustering is applied in (McCallum, Nigam & Ungar 2000) as an alternative to traditional blocking (in order to reduce the number of record pair comparisons to be made), while in (Gu & Baxter 2004*a*), the use of simple K-means clustering together with a user-tunable fuzzy region for the class of possible-matches is investigated. Methods based on nearest neighbours are explored in (Chaudhuri, Ganti & Motwani 2005), with the idea being to capture local structural properties instead of a single global distance approach. Graphical models (Ravikumar & Cohen 2004) are another unsupervised technique that aims at using the structural information available in the data to build hierarchical probabilistic models for record pair classification.

Many of these new approaches are based on supervised learning techniques and require training data which is often not available in real world situations, or only obtainable via manual preparation (a costly process similar to manual clerical review). Additionally, many of the recent publications in this area present experimental studies that are based on small datasets of up to a couple of thousand records (Christen 2005).

2 Critical Discussion

This section discusses some problematic issues affecting the record linkage process: task complexity, parameter freedom, the availability of training data, and blocking - all affecting the ability to automate the linkage process.

2.1 Task Complexity but Match Rarity

With at most one true match between each record in **A** and **B** (assuming no duplicates), the largest possible number of matches is the smaller of the size of **A** and **B**, $n = min(|\mathbf{A}|, |\mathbf{B}|)$, with $|\cdot|$ denoting the number of records in a dataset. Where $n = |\mathbf{A}| = |\mathbf{B}|$, there can be no more than n matches: every record in **A** is linked to a different record in **B** (Christen & Goiser 2005). However, as every record in **A** potentially needs to be compared against every record in **B**, the number of comparisons required is $|\mathbf{A}| \times |\mathbf{B}|$ which is n^2 for $n = |\mathbf{A}| = |\mathbf{B}|$. Thus, while the number of matches increases linearly with the size of the data, the number of comparisons required increases quadratically. When de-duplicating a dataset, all records potentially need to be compared with all others, thus requiring $n(n-1)/2$ comparisons.

This result has significant ramifications for Record Linkage. For example, in many areas such as social security, health, tax, corporate customer information, a dataset of one million records is considered to be small, yet conducting a million times a million, 10^{12}, complicated comparisons would not be considered viable due to the time it would take. For example, if a comparison requires 0.1 milliseconds per attribute and there are 10 attributes to compare, the linkage of two datasets with one-million records each (assuming no blocking) would require 10^9 seconds which is nearly 32 years!

The other aspect of the linear increase in matches versus the quadratic increase in the problem size is that the matches become rare. In the above example, while 10^{12} comparisons are potentially required, the maximum number of matches is 10^6. The rate of matches to comparisons – the 'hit' rate – for these datasets is one-in-a-million.

Figure 1 shows density plots of a real-world sampled dataset (n = 996,166, comprising 353 matches and 995,813 non-matches, a match rate of 1/2,822) (Centre for Epidemiology and Research, NSW Department of Health 2001). It can be seen that, while matches form a distinct grouping with their own modal peak, they hardly register in comparison to the large number of non-matches. This must be considered when selecting and using classification methods. Even when blocking is applied, the number of matches to non-matches is often very different. Thus, while the complexity of the task becomes difficult for medium- to large-sized datasets, the matches themselves become increasingly rare.

2.2 Parameter Freedom

Many comparison and classification techniques allow tuning in order to increase their accuracy, or to allow trading off one benefit in order to increase another. For example, in the Fellegi and Sunter model, the threshold values between non-matches, possible-matches and matches are user-adjustable – changing the values will alter the linkage quality as well as the number of record pairs set aside for clerical review (Fellegi & Sunter 1969).

Accurate setting of parameters requires a high degree of knowledge of the techniques used as well as of the characteristics of the data in question, and can take a significant amount of time and effort to determine properly. Once set, the parameters can only be re-used in further linkages with confidence if the characteristics of the new data were such that the same values would be generated. This paper addresses this problem by choosing to research methods which do not require parameters to be set.

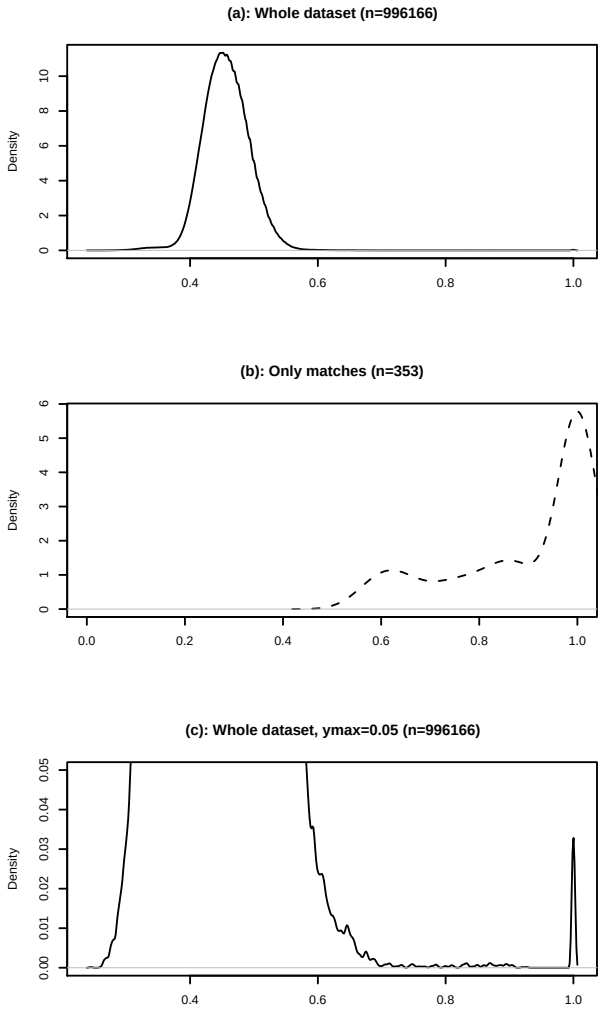


Figure 1: Density plots of one of the sampled datasets for: (a) the whole dataset, (b) just the matches, (c) a 20× y-axis magnification of (a) showing the peak for the matches.

2.3 Availability of Example Data

In order to set parameters, some knowledge of the data must be obtained, whether through data analysis, or supervised Machine Learning techniques (Mitchell 1997). This requires the availability of *example* or *training* data - that is, data for which the match/non-match state is known in advance.

One way of generating such example data is to randomly sample pairs of records and manually classify them. However, considering the size of datasets and the rarity of matches within them, many thousands of record pairs may need to be examined in order to obtain a few examples of matched records, and it may not be known if they form a representative sample of all the matches. A solution to this is to use a technique like active learning to bias the sampling of examples in order to increase the representation of matches (Sarawagi & Bhamidipaty 2002).

Sampling bias and representation must be considered when making use of the example data.

Example data are thus subject to the same sorts of problems as associated with parameters (as discussed in Section 2.2). Further, given that example cases are provided as input to generate parameters, they can be seen as parameters in themselves - different examples will generate different parameter values.

2.4 Blocking

To the extent that blocking removes comparisons in a consistent fashion, it becomes a source of bias - a confounding factor - which, in fact, is its intent: to consistently remove obvious non-matches. However, it must be recognised that biased data will have an effect on the results of classification methods which adapt to or learn from that data. Where blocking removes true matches, it can be seen as a failing, and the extent to which this occurs with consistency is, again, a source of bias. As an example, (Gu & Baxter 2004*a*) block four small datasets and show that between 8% and 30% of the true matches are removed by blocking.

Other issues with blocking include the question of the amount of time saved: from the above, does the reduction from 10^{12} to 10^{8} (or similar amounts) merely change the problem from impossible to unfeasible? Also, errors in the data or the blocking criteria may mean that no matter how many passes are conducted, some true matches won't be passed through (e.g., where they aren't assigned to the same block in standard blocking). Thus, setting up the blocking criteria requires knowledge of the data and the blocking technique used. As an example, standard blocking works optimally when the data is evenly distributed into a moderate number of blocks. However, if only a single block is too large (e.g. a block with surname value of "Smith"), the quadratic issue returns. See (Gu & Baxter 2004*b*) for further discussion, and a potential solution. Note that blocking criteria, again, are parameters - see the discussion on parameter freedom above.

To the extent that different blocking methods and blocking criteria (including not blocking) result in different data, they produce different biases. Thus comparisons between classification methods become invalid or difficult if different blocking methods or criteria are used. The important question to ask about the results of research which uses blocking then becomes: if different blocking methods or criteria were used, can it be shown that the results would be the same? If not, blocking can be regarded as integral to the process, and cannot be divorced from it.

Given these problems, it is strongly recommended that researchers only block their data if it is too large to feasibly conduct a linkage on, or if the research is into blocking techniques. With the ready availability of relatively fast personal computers with large main memory, it cannot be seen how it could be justified to use blocking in research which would require fewer than one million comparisons.

It is noted, however, that without blocking, the linkage or de-duplication of large datasets could not be accomplished. It is thus necessary but problematic, and careful attention must be paid to its use. In using blocking, it must be understood that potential matches will be removed, and that the data will be biased which may affect the results of further procedures.

3 Experiments

Given the above discussion, it was decided to examine the feasibility of parameter-free techniques for record linkage - that is, investigate if techniques which don't require parameters can produce results comparable to those which do. All the presented experiments were conducted without using blocking.

3.1 Experimental Setup

Three comparison and three classification methods were chosen. The *Febrl* (Freely Extensible Biomedical Record Linkage) (Christen, Churches & Hegland 2004) open source record linkage system was used for the comparison step, while the *Weka* (Witten & Frank 2005) open source data mining package was used for the classification step.

3.1.1 String Comparison Methods

Of the three methods, the first two were chosen because they are commonly used in record linkage, and the third because it is novel in this field and could prove to be potentially useful. None of these methods require parameters.

In the **Jaro-Winkler** (**JW**) comparison method (Winkler 1990, Winkler 2006, Yancey 2006) a similarity score based on the number of common characters, character transpositions, and string lengths, as well as giving a higher score for having a common prefix of length up to 4 characters, is calculated.

In the **edit-distance** (**ED**) method, a similarity score is calculated using the normalisation of the minimal number of single-letter insertions, deletions and substitutions required to transform one string into the other (Winkler 2006, Yancey 2006).

Compression comparison (**CC**) (Cilibrasi & Vitanyi 2005, Keogh, Lonardi & Ratanamahatana 2004) uses the fact that the compression of the concatenation of two similar strings is shorter than that of dissimilar ones. The normalised compression distance was used:

$$NCD(x,y) = \frac{C(xy) - min(C(x), C(y))}{max(C(x), C(y))}$$

where $C()$ is a compression algorithm such as `zlib` or `GZip`, and xy is the concatenation of the strings, x and y.

3.1.2 Classification Methods

One supervised and two unsupervised classification methods were chosen. The supervised method requires training data, and, being partitioning clustering techniques, the unsupervised methods require the specification of the number of clusters. As the aim is to have a cluster of matches and a cluster of non-matches, this value is fixed at two. Being fixed, the value doesn't change meaning it is not supplied as a parameter.

Decision trees (**DT**) are one of the major Machine Learning techniques (Mitchell 1997). The normal procedure is to use training data to build a *classifier*, which is then used to classify further data. For this work, they are used as a base line – to compare the results of the unsupervised methods against. As such, to give the best possible results, *all* the data in the dataset under investigation is used for both training *and* testing.

***K*-means** (**KM**) is a commonly used simple unsupervised clustering technique (Han & Kamber 2001). Previous papers which have looked at K-means in the context of record linkage include (Elfeky et al. 2002, Gu & Baxter 2004*a*).

The **farthest-first** (**FF**) clustering technique was first presented in (Gonzalez 1985). It is a very simple algorithm and very fast: assign the centroid for the first cluster to a random point. For the second centroid, choose the point which is farthest from it. For all following centroids, choose the point which is farthest from all the centroids chosen so far.

It is interesting in that, unlike **KM** which can halt at local minima, it is guaranteed to provide a solution within two times the optimal solution value of the objective function used to choose the clusters - the *2-approximation problem* (Gonzalez 1985).

3.2 Data Sources

Three data sources were used in the experiments, two synthetically generated, and one real-world dataset.

3.2.1 Synthetic Data

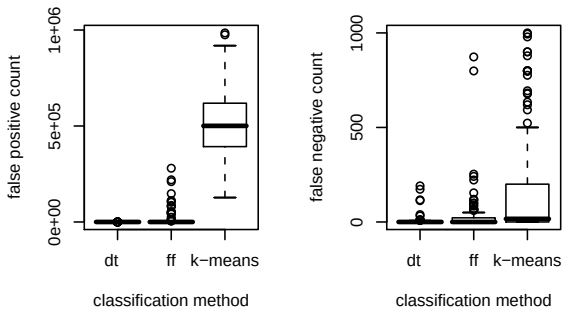
Synthetic data was generated using the dataset generator from *Febrl* (Christen et al. 2004), which allows probabilistic creation of records, as well as simulation of common types of errors at specifiable rates (Christen 2005). Two groups of datasets were generated, one with a maximum of one error per record, and a second with up to 3 errors in any attribute. The generated attributes were: given-name, surname, street-number, street-type, street-value, suburb, postcode, state, date-of-birth, age, phone-number, and social security identifier. For each group, seven pairs of datasets were generated with 0%, 10%, 20%, 50%, 80%, 90% and 100% overlap – the amount of duplication between the datasets, with 0% being no common record, and 100% being duplicate datasets. Each of the pairs of datasets contained one thousand records, thus requiring one million comparisons. From these datasets, variations were generated using different concatenations of the original attributes: **one**: the attributes were concatenated into a single attribute; **three**: the attributes were concatenated into 'name', 'address', and 'other' attributes; **all**: all attributes were kept. That is, generating the variation involved taking the attributes to use and joining them together, delimited by a space, into a new attribute.

The total number of linkages on synthetic datasets were thus: 2 different numbers of errors per record, 7 types of overlap, 3 combinations of attributes, 3 comparison methods, 3 classification methods, giving a total of **378** discreet record linkages of a million comparisons each.

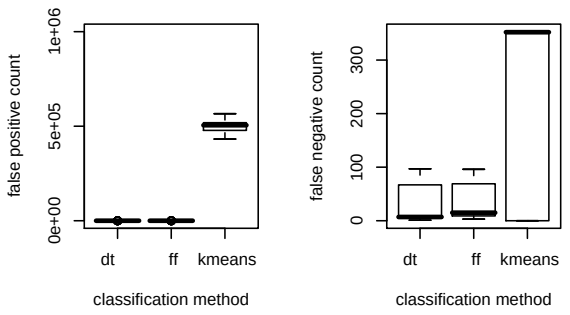
3.2.2 Real-World Data

Access was available to a confidential dataset in which extensive effort had previously been made to correctly classify record duplicates. This dataset, the *New South Wales Midwives Data Collection* (MDC) (Centre for Epidemiology and Research, NSW Department of Health 2001), was provided by the New South Wales Department of Health. It contains 175,211 records relating to births in the years 1999 and 2000 and had been de-duplicated using the commercial probabilistic software, *AutoMatch* (MatchWare Technologies 1998), including post-linkage manual clerical review. Of the 175,211 records, it had been determined that 158,081 mothers appeared once, 8,295 appeared twice, 176 appeared three times, and 3 appeared four times in the dataset.

As an exhaustive de-duplication would require 15,349,359,655 comparisons, it was decided to sample the data, and have about the same number of comparisons as the synthetic data. A sample size



(a) combined synthetic data (all three comparison methods)



(b) MDC (**CC** only)

Figure 2: Errors by classification method for (a) the combined synthetic data, and (b) the MDC data.

of $n = 1,412$, resulting in $n(n-1)/2 = 996,166$ comparisons was decided on. Each sample comprised 353 randomly selected matched pairs and 706 randomly selected non-matched records. Fifty samples were drawn, the three attribute combinations were generated, and the **CC** comparison method was used with all three classification methods. For the MDC data, there were thus a total of **450** discrete record linkages of 996,166 comparisons each.

3.3 Results

Note that unless described as density plots (where the area under the curve is 1), the graphs all use boxplots. A boxplot is a concise graphical device showing the upper and lower quartile (the box), the median (the line crossing the box), and the range (the ends of the lines extending from the box, often called whiskers). Where a data point is suspected to be an outlier, it is plotted as a point, and the whisker is then set at 1.5 times the inter-quartile range. Where the median is offset within the box, it is an indication of skewness in the data.

How does the choice of classification method affect the results?

The boxplots in Figure 2 describe the errors produced by the classification techniques used for the synthetic datasets (combined), and the MDC data. It can be seen that **KM** performs worse than the other two. In fact, for **KM**, the mean number of false positives for all the linkages in the synthetic data is 509,064. Since there were one million comparisons, this means that the method does slightly worse than chance. (For brevity, **KM** will not be further discussed when comparing classifiers.) Otherwise, it is of interest that **FF** has comparable results with **DT**.

Why does *K*-means do so badly?

Han and Kamber point out that, "the K-means method is not suitable for discovering clusters with non-convex shapes or clusters of very different size" (Han & Kamber 2001) (p. 350). For a linkage without blocking as in these experiments, there is an overwhelming number of non-matches (for the synthetic data: $10^6 - (10^3 \times overlap\%/100)$, and for the sampled MDC data: $996,166 - 353$). These different cluster sizes can be seen to be the cause of why **KM** does so badly.

The reason that **KM** has been used successfully in record linkage, e.g. (Gu & Baxter 2004*a*), is that it has been preceded by blocking which has had the effect of evening-up the class sizes. Thus, for **KM** to be successful in record linkage, blocking must have been previously used – and had the effect of evening up the class sizes.

How does the choice of comparison method affect the results?

Figure 3 shows boxplots of the error counts for the combined synthetic experiments. For the false negatives, it can be seen that **FF** performs comparably with **DT** except for two cases of **JW** comparisons (which may be due to **FF** randomly selecting centroids which are not be near the mode of the true negatives). With the false positives, the errors appear larger, increasingly for **ED** and **JW**. As most of these errors are also associated with the **all** concatenation group, it can be seen that **JW** does not provide good discrimination for **FF** when provided with a larger number of comparisons. Note that the median number of false positive errors is 1.5 and 1.0 for **CC** and **ED** comparisons respectively, while the median false negative values are 0, so the results are actually very good for most of the datasets.

Density plots of the results of the different comparison methods are shown in Figure 4. The modes of the distributions of the two classes are closest together for **JW**. As **FF** uses a threshold midway between the centroids to determine the class, it can be seen that the more the overlap between the modes, the more errors would be generated. Other aspects that can be seen to affect accuracy are the skewness of the distributions, and the combination of spread and distance between centroids. For example, the larger the spread and the closer the centroids the more the overlap, so the greater the number of errors. Thus, it can be seen that **FF** performs better in conjunction with **CC** or **ED**.

Can a reasonable result be obtained with parameter-less techniques?

Linking synthetic datasets with some overlap, using **ED** for comparisons of the attributes concatenated together and **FF** for classification, resulted in 0 false positives and a mean of 6.833 (median 0.5) false negative errors. For the MDC dataset, use of **CC** on the three attribute datasets with **FF** classification resulted in a mean of 4.68 (median 3.00) false positives and a mean of 9.76 (median 9.00) false negatives. Regularly having less than ten errors in around one million comparisons is a very small error rate. This is a strong indication that a linkage of reasonable quality can be obtained without requiring parameters.

Why does farthest-first do so well?

Examining the **FF** algorithm and considering Figure 1, it can be seen that a random choice of item will almost always make a selection near the mode

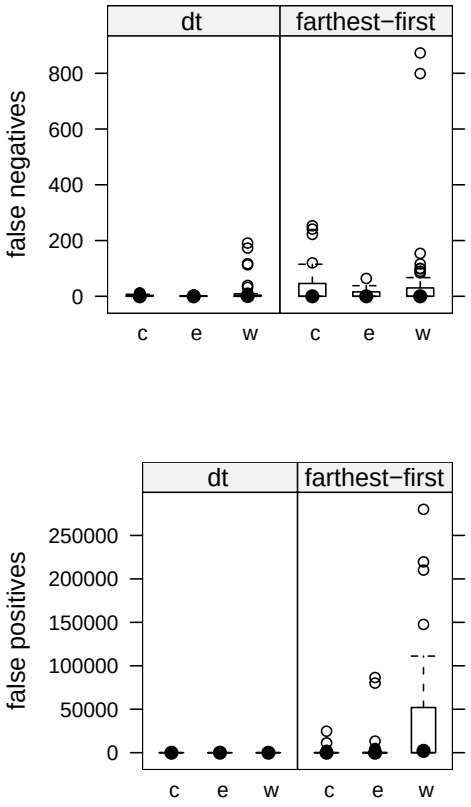


Figure 3: Errors by comparison methods (c=**CC**, e=**ED**, w=**JW**) for **DT** and **FF** classification methods using the combined synthetic datasets.

of the true-negatives for the first centroid, and for the second, it follows that the farthest from it will be the item with the maximum value - the pair with the highest similarity results. As the threshold value is midway between these two centroids, and since both **ED** and **CC** fulfil triangle inequality (Cilibrasi & Vitanyi 2005, Marzal & Vidal 1993), it can be seen that all results on one side of the threshold will be closer to its centroid than any on the other side. Thus, by not conducting blocking, an important characteristic of the data has been retained – one which allows **FF** to return very positive results.

3.4 Attempt to Improve Farthest-First

An improvement would be to remove the random choice for the first centroid, thus removing the chance of choosing an item away from the mode of the true negatives. To examine this, the original Weka **FF** algorithm was translated into the Python programming language[1], and four alternatives for picking the centroids were implemented:

- **Default**: the normal selection process for **FF**.
- **Mode**: for each attribute, bin the values, then choose the largest bin as the first centroid. This is equivalent to producing a histogram plot and selecting the longest bar – a crude density-based selection process. One thousand bins were used across the range of the data values.
- **0-1**: choose the values, 0 and 1 as the centroids. The idea here is that the comparison routines

[1] `www.python.org`

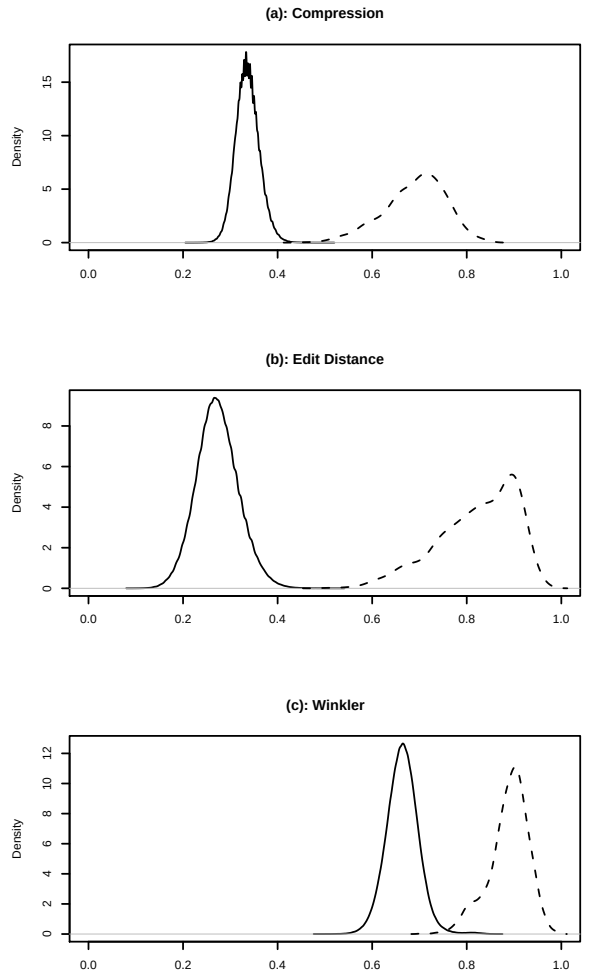


Figure 4: Density plots of the results for the different comparison methods when applied to the dataset with 100% overlap, and up to 3 errors in any attribute. The left peak is that of the non-matches, and the right is that of the matches. Note that each graph contains two separate density plots superimposed (as denoted by the solid and dashed lines).

return values in the range, 0..1. Thus, no processing is required to discover these centroids.

- **Range**: use the minimum and maximum values of the data as the centroids. This can be seen to be the same as normalising the comparison results and choosing the **0-1** modification.

Figure 5 shows the results of applying these modifications to the combined synthetic datasets. Note that **Default** differs slightly from the Weka version in that the latter normalises the attributes, but the Python implementation doesn't since the input data is already in the range 0..1. Also, as both Weka and **Default** use random selection, there is inherent variation in the results.

It can be seen that false positive rates for modifications **0-1** and **Range** appear higher than the others, while their false negative rates are very low. This indicates that the threshold values are set too low. However, the **ED** results for modification **0-1** show little or no errors. This modification uses 0 and 1 for the centroids, and earlier discussion has shown that **ED** separates the modal peaks more than the other methods. **Mode** does appear to show advantages over **Default** in that the three outliers in the false negative graph have been removed. However, differences are very minor as would be expected - except that possible selection of centroids away from the mode had been eliminated. These results were similar for the MDC dataset.

4 Conclusion and Future Work

Record Linkage was introduced and some characteristics and challenges to the field were presented. Blocking was discussed and it was noted that it can bias results unless controlled-for in experiments. This can also compromise the comparison of record linkage techniques. It was therefore recommended that blocking not be used in research unless necessary.

Experiments using unsupervised techniques in the linkage process were conducted and it was found that the K-means clustering technique was not suitable for the linkage of data which had not previously been blocked. The use of the farthest-first classification technique on non-blocked data was found to produce very promising results.

In terms of further research, the complexity of the edit-distance comparison method is the product of the lengths of the strings being compared, which is costly where the strings are long. An improvement would be to use other versions of the technique, such as those used in genetics (Christen 2006). Given the positive results for edit-distance with the modification to farthest-first classification on the synthetic datasets, further examination using real-world datasets will be conducted. The Expectation Maximisation (EM) algorithm (Winkler 1988) provides a method whereby parameters such as the thresholds in the Fellegi and Sunter model (Fellegi & Sunter 1969) can be estimated. Further work will include comparing EM against the farthest-first classification method.

Copies of this paper, the Febrl record linkage system, and other publications can be obtained from: http://datamining.anu.edu.au/linkage.html.

Acknowledgments

This work is supported by an Australian Research Council (ARC) Linkage Grant LP0453463 and partially funded by the NSW Department of Health.

References

Baxter, R. A., Christen, P. & Churches, T. (2003), A comparison of fast blocking methods for record linkage, *in* 'ACM SIGKDD'03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation', Washington, DC, USA, pp. 25–27.

Bilenko, M. & Mooney, R. J. (2003), Adaptive duplicate detection using learnable string similarity measures, *in* 'Proceedings of ACM SIGKDD', ACM Press, Washington DC, pp. 39–48.

Centre for Epidemiology and Research, NSW Department of Health (2001), 'New South Wales mothers and babies 2001', *NSW Public Health Bull* **13:S-4**.

Chaudhuri, S., Ganjam, K., Ganti, V. & Motwani, R. (2003), Robust and efficient fuzzy match for online data cleaning, *in* 'Proceedings of ACM SIGMOD', San Diego, pp. 313–324.

Chaudhuri, S., Ganti, V. & Motwani, R. (2005), Robust identification of fuzzy duplicates, *in* 'Proceedings of the 21st international conference on

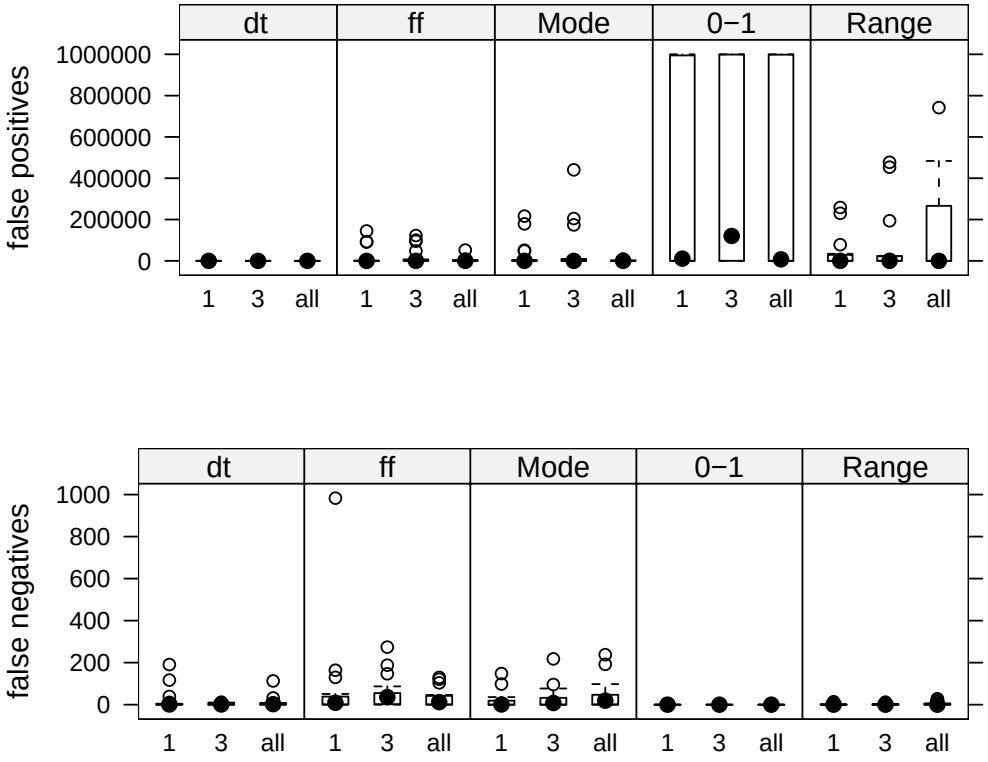


Figure 5: Errors by concatenation method for the classification methods (dt = **DT**, ff = unmodified Python **FF**) for data from the synthetic experiments combined.

data engineering (ICDE'05)', Tokyo, pp. 865–876.

Christen, P. (2005), Probabilistic data generation for deduplication and data linkage, *in* 'IDEAL'05', Springer LNCS 3578, Brisbane, pp. 109–116.

Christen, P. (2006), A comparison of personal name matching: Techniques and practical issues, *in* 'The Second International Workshop on Mining Complex Data (MCD'06)'.

Christen, P., Churches, T. & Hegland, M. (2004), Febrl – A parallel open source data linkage system, *in* 'Proceedings of the 8th PAKDD', pp. 638–647.

Christen, P. & Goiser, K. (2005), Assessing deduplication and data linkage quality: What to measure?, *in* 'Proceedings of the fourth Australasian Data Mining Conference (AusDM 2005)', Sydney.

Cilibrasi, R. & Vitanyi, P. (2005), Clustering by compression, *in* 'IEEE Trans. Information Theory', Vol. 51, pp. 1523– 1545.

Cohen, W. W. (1998), Integration of heterogeneous databases without common domains using queries based on textual similarity, *in* 'Proceedings of ACM SIGMOD', Seattle, pp. 201–212.

Cohen, W. W., Ravikumar, P. & Fienberg, S. (2003), A comparison of string distance metrics for name-matching tasks, *in* 'Proceedings of IJCAI-03 workshop on information integration on the Web (IIWeb-03)', Acapulco, pp. 73–78.

Elfeky, M. G., Verykios, V. S. & Elmagarmid, A. K. (2002), TAILOR: A record linkage toolbox, *in* 'Proceedings of ICDE', San Jose, pp. 17–28.

Fellegi, I. P. & Sunter, A. B. (1969), A theory for record linkage, *in* 'Journal of the American Statistical Association', Vol. 64, pp. 1183–1210.

Gill, L. (2001), Methods for automatic record matching and linking and their use in national statistics, *in* 'National Statistics Methodology Series', number 25.

Gonzalez, T. F. (1985), Clustering to minimize the maximum intercluster distance, *in* 'Theoretical Computer Science', Vol. 38, pp. 293–306.

Gu, L. & Baxter, R. (2004*a*), Decision models for record linkage, *in* 'AusDM 2004, Springer LNAI 3755', Cairns, Australia, pp. 146–160.

Gu, L. & Baxter, R. A. (2004*b*), Adaptive filtering for efficient record linkage, *in* 'Proceedings of the Fourth SIAM International Conference on Data Mining (SDM-04)', Orlando, Florida.

Han, J. & Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Fransisco, CA.

Keogh, E., Lonardi, S. & Ratanamahatana, C. (2004), Towards parameter-free data mining, *in* '2004 ACM SIGKDD international conference on knowledge discovery and data mining', pp. 206–215.

Marzal, A. & Vidal, E. (1993), 'Computation of normalized edit distance and applications.', *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(9), 926–932.

MatchWare Technologies (1998), *AutoStan and AutoMatch, User's Manuals*, Kennebunk, Maine.

McCallum, A., Nigam, K. & Ungar, L. (2000), Efficient clustering of high-dimensional data sets with application to reference matching, *in* 'Proceedings of ACM SIGKDD', Boston, pp. 169–178.

Michalowski, M., Thakkar, S. & Knoblock, C. A. (2004), Exploiting secondary sources for automatic object consolidation, *in* 'Proceedings of the 2004 VLDB Workshop on Information Integration on the Web'.

Mitchell, T. M. (1997), *Machine Learning*, McGraw-Hill, Boston.

Nahm, U., Bilenko, M. & Mooney, R. (2002), Two approaches to handling noisy variation in text mining, *in* 'Proceedings of the ICML-2002 workshop on text learning (TextML'2002)', Sydney, pp. 18–27.

Ravikumar, P. & Cohen, W. W. (2004), A hierarchical graphical model for record linkage, *in* 'roc. of the 20th Conference on Uncertainty in Artificial Intelligence', Banff, Canada, pp. 454–461.

Sarawagi, S. & Bhamidipaty, A. (2002), Interactive deduplication using active learning, *in* 'Proceedings of ACM SIGKDD', ACM Press, Edmonton, pp. 269–278.

Tejada, S., Knoblock, C. & Minton, S. (2002), Learning domain-independent string transformation weights for high accuracy object identification, *in* 'Proceedings of ACM SIGKDD', Edmonton, pp. 350–359.

Winkler, W. E. (1988), Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage, *in* 'Proceedings of the Survey Research Methods Section, American Statistical Association'.

Winkler, W. E. (1990), String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage, *in* 'Section on Survey Research Methods', American Statistical Association, pp. 354–359.

Winkler, W. E. (2005), Approximate string comparator search strategies for very large administrative lists, Technical Report RRS2005/02, US Bureau of the Census.

Winkler, W. E. (2006), Overview of record linkage and current research directions, Technical Report RRS2006/02, US Bureau of the Census.

Witten, I. H. & Frank, E. (2005), *Data Mining: Practical machine learning tools and techniques*, 2nd edn, Morgan Kaufmann, San Francisco.

Yancey, W. E. (2004), An adaptive string comparator for record linkage, Technical Report RR2004/02, US Bureau of the Census.

Yancey, W. E. (2006), Evaluating string comparator performance for record linkage, Technical Report RRS2005/05, US Bureau of the Census.

Zhu, J. & Ungar, L. (2000), String edit analysis for merging databases, *in* 'KDD workshop on text mining, held at ACM SIGKDD', Boston.

A Comparative Study of Classification Methods for Microarray Data Analysis

Hong Hu[1] **Jiuyong Li**[1] **Ashley Plank**[1] **Hua Wang**[1] **Grant Daggard**[2]

Department of Mathematics and Computing[1]
Department of Biological and Physical Sciences[2]
University of Southern Queensland,
Toowoomba, QLD 4350, Australia
Email: huhong@usq.edu.au

Abstract

In response to the rapid development of DNA Microarray technology, many classification methods have been used for Microarray classification. SVMs, decision trees, Bagging, Boosting and Random Forest are commonly used methods. In this paper, we conduct experimental comparison of LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5, and Random Forest on seven Microarray cancer data sets. The experimental results show that all ensemble methods outperform C4.5. The experimental results also show that all five methods benefit from data preprocessing, including gene selection and discretization, in classification accuracy. In addition to comparing the average accuracies of ten-fold cross validation tests on seven data sets, we use two statistical tests to validate findings. We observe that Wilcoxon signed rank test is better than sign test for such purpose.

Keywords: Microarray data, classification.

1 Introduction

In recent years, the rapid development of DNA Microarray technology has made it possible for scientists to monitor the expression level of thousands of genes with a single experiment (Schena, Shalon, Davis & Brown 1995, Lockhart, Dong, Byrne & et al. 1996). With DNA expression Microarray technology, researchers will be able to classify different diseases according to different expression levels in normal and tumor cells, to discover the relationship between genes, to identify the critical genes in the development of disease. There are many active research applications of Microarray technology, such as cancer classification (Golub, Slonim, Tamayo & et al. 1999, Veer, Dai, de Vijver & et al. 2002, PetricoinIII, Ardekani, Hitt, Levine & et al. 2002), gene function identification (Lu, Patterson, Wang, Marquez & Atkinson 2004, Santin, Zhan, Bellone & Palmieri 2004), clinical diagnosis (Yeang, Ramaswamy, Tamayo & et al. 2001), and drug discovery studies (Maron & Lozano-Pérez 1998).

This project was partially supported by Australian Research Council Discovery Grant DP0559090.

A main task of Microarray classification is to build a classifier from historical Microarray gene expression data, and then it uses the classifier to classify future coming data. Many methods have been used in Microarray classification, and typical methods are Support Vector Machines (SVMs) (Brown, Grundy, Lin, Cristianini, Sugnet, Furey, Jr & Haussler 2000, Guyon, Weston, Barnhill & Vapnik 2002), k-nearest neighbor classifier (Yeang et al. 2001), C4.5 decision tree (Li & Liu 2003, Li, Liu, Ng & Wong 2003), rule-base classification method (Yeang et al. 2001) and ensemble methods, such as Bagging and boosting (Tan & Gibert 2003, Dietterich 2000).

SVMs, decision trees and ensemble methods are most frequently used methods in Microarray classification. Reading through the literature of Microarray data classification, it is difficult to find consensus conclusions on their relative performance. We are very interested in classifying Microarray data using C4.5 since it provides more interpretable results than other methods do. Therefore, we design an experiment to find out the classification performance of C4.5, AdaBoostingC4.5, BaggingC4.5, Random Forests, Libsvms on seven Microarray cancer data sets.

In the experimental analysis, we use sign test and Wilcoxon signed rank test to compare classification performance of different methods. We find that Wilcoxon signed rank test is better than sign test for such comparison. We also find inconsistent results in accuracy test and Wilcoxon signed rank test, and we interpret the results in a reasonable way.

The rest of this paper is organized as follows. In Section 2, we describe the relevant methods in this comparison study. In Section 3, we introduce our experimental design. In Section 4, we show our experimental results and present discussions. In Section 5, we conclude the paper.

2 Algorithm selected for comparison

Numerous Microarray data classification algorithms have been proposed in recent years. Most of them have been adapted from current data mining and machine learning algorithms.

C4.5 (Quinlan 1993, Quinlan 1996) was proposed by Quinlan in 1993 and it is a typical decision tree algorithm. C4.5 partitions a training data into some disjoint subsets simultaneously, based on the values of an attribute. At each step in the construction of the decision tree, C4.5 selects an attribute which separates data with the highest information gain ratio (Quinlan 1993). The same process is repeated on all subsets until each subset contains only one class. To simplify the decision tree, the induced decision tree is pruned using pessimistic error estimation (Quinlan 1993).

SVMs was proposed by Cortes and Vapnik (Cortes

& Vapnik 1995) in 1995 and It has been a most influential classification algorithm in recent years. SVMs are classifiers which transform the input samples into a high dimensional space by a kernel function and use a linear hyperplane to separate two classes mapped to that high dimensional space by support vectors which are selected vectors from training samples. SVMs has been applied to many domains, for example, text categorization (Joachims 1998), cancer classification (Furey, Christianini, Duffy, Bednarski, Schummer & Hauessler 2000, Brown et al. 2000, Brown, Grundy, Lin, Cristianini, Sugnet, Ares & Haussler 1999).

In the past decade, many researchers have devoted their efforts to the study of ensemble decision tree methods for Microarray classification. Ensemble decision tree methods combine decision trees generated from multiple training data sets by re-sampling the training data set. Bagging, Boosting and Random forests are some of the well-known ensemble methods in the machine learning field.

Bagging was proposed by Leo Breiman (Breiman 1996) in 1996. Bagging uses a bootstrap technique to re-sample the training data sets. Some samples may appear more than once in a data set whereas some samples do not appear. A set of alternative classifiers are generated from a set of re-sampled data sets. Each classifier will in turn assign a predicted class to an incoming test sample. The final predicted class for the sample is determined by the majority vote. All classifiers have equal weights in voting.

The boosting method was first developed by Freund and Schapire (Freund & Schapire 1996) in 1996. Boosting uses a re-sampling technique different from Bagging. A new training data set is generated according to its sample distribution. The first classifier is constructed from the original data set where every sample has an equal distribution ratio of 1. In the following training data sets, the distribution ratios are made differently among samples. A sample distribution ratio is reduced if the sample has been correctly classified; otherwise the ratio is kept unchanged. Samples which are misclassified often get duplicates in a re-sampled training data set. In contrast, samples which are correctly classified often may not appear in a re-sampled training data set. A weighted voting method is used in the committee decision. A higher accuracy classifier has larger weight than a lower accuracy classifier. The final verdict goes along with the largest weighted votes.

Based on Bagging, Leo Breiman introduced another ensemble decision tree method called Random Forests (Breiman 1999) in 1999. This method combines Bagging and random feature selection methods to generate multiple classifiers.

3 Experimental design methodology

3.1 Ten-fold cross-validation

Tenfold cross-validation is used in this experiment. In tenfold cross-validation, a data set is equally divided into 10 folds (partitions) with the same distribution. In each test 9 folds of data are used for training and one fold is for testing (unseen data set). The test procedure is repeated 10 times. The final accuracy of an algorithm will be the average of the 10 trials.

3.2 Test data sets

Seven data sets from Kent Ridge Biological Data Set Repository (**?**) are selected. These data sets were collected from very well researched journal papers, namely Breast Cancer (Veer et al. 2002), Lung Cancer (Gordon, Jensen, Hsiao, Gullans & et al. 2002), Lymphoma (Alizadeh, Eishen, Davis, Ma & et al. 2000), ALL-AML Leukemia (Golub et al. 1999), Colon (Alon & et al. 1999), Ovarian (PetricoinIII et al. 2002) and Prostate (Singh & et al. 2002). Table 1 shows the summary of the characteristics of the seven data sets. We conduct our experiments by using tenfold cross-validation on the merged original training and test data sets.

Data set	Genes	Class	Record
Breast Cancer	24481	2	97
Lung Cancer	12533	2	181
Lymphoma	4026	2	47
Leukemia	7129	2	72
Colon	2000	2	62
Ovarian	15154	2	253
Prostate	12600	2	21

Table 1: Experimental data set details

3.3 Softwares used for comparison

We have done our experiments with C4.5, C4.5AdaBoosting, C4.5Bagging, Random forests, LibSVMs with the Weka-3-5-2 package which is available online (`http://www.cs.waikato.ac.nz/ml/weka/`). Default settings are used for all compared ensemble methods. We were aware that the accuracy of some methods on some data sets can be improved when parameters were changed. However, it was difficult to find another uniform setting good for all data sets. Therefore, we did not change default settings since the default produced high accuracy on average.

3.4 Microarray data preprocessing

We used information gain ratio for gene selection and used Fayyad and Irani's MDL discretization method provided by Weka to discretize numerical attributes. Our previous results (Hu, Li, Wang & Daggard 2006) show that with preprocessing, the number of genes selected affects the classification accuracy. The overall performance is better when data sets contain 50 to 100 genes. For our experiment, we set the number of genes as 50. After the data preprocessing, each data set contains 50 genes with discretized values.

3.5 Sign test

Sign test (Conover 1980) is used to test whether one random variable in a pair tends to be larger than the other random variable in the pair. Given n pairs of observations. Within each pair, either a plus, tie or minus is assigned. The plus corresponds to that one value is greater than the other, the minus corresponds to that one value is less than the other, and the tie means that both equal to each other. The null hypothesis is that the number of pluses and minuses are equal. If the null hypothesis test is rejected, then one random variable tends to be greater than the other.

3.6 Wilcoxon signed rank test

Sign test only makes use of information of whether a value is greater, less than or equal to the other in a pair. Wilcoxon signed rank test (Conover 1980, Daniel 1978) calculates differences of pairs. The absolute differences are ranked after discarding pairs with the difference of zero. The ranks are sorted in ascending order. When several pairs have absolute differences that are equal to each other, each of these

Data set	C4.5	Random Forests	AdaBoostC4.5	BaggingC4.5	LibSVMs
Breast Cancer	84.5	88.7	90.7	85.6	72.2
Lung Cancer	98.3	99.5	98.3	97.8	100.0
Lymphoma	74.5	93.6	89.4	89.4	55.3
Leukemia	88.9	98.6	95.8	95.8	100.0
Colon	88.7	83.9	90.3	90.3	90.3
Ovarian	96.8	99.2	98.8	98.0	100.0
Prostate	95.2	100	95.2	95.2	100.0
Average	89.6	94.8	94.1	93.2	88.3

Table 2: Average accuracy of seven preprocessed data sets with five classification algorithms based on tenfold cross-validation

	C4.5	Random Forests	AdaBoostC4.5	BaggingC4.5	LibSVMs
C4.5	–				
Random Forests	0.063	–			
AdaboostC4.5	**0.031**	0.63	–		
BaggingC4.5	0.11	**0.0088**	**0**	–	
LibSVMs	0.23	0.34	0.34	0.34	–

Table 3: Summary of sign test between any two of the compared classification methods. P-values of the test are given, and significant p-values at 95% confidence level are highlighted.

several pairs is assigned as the average of ranks that would have otherwise been assigned. The hypothesis is that the differences have the mean of 0.

4 Experimental results and discussions

Table 2 shows the individual and average accuracy results of all the compared methods based on seven preprocessed data sets with the tenfold cross-validation method. Table 5 shows the individual and average accuracy results of the compared methods based on seven original data sets with tenfold cross-validation method.

Based on Table 2, we have the following conclusions: with preprocessed data sets, all ensemble methods on average perform better than C4.5 and LibSVMs. Both C4.5 and LibSVM perform similar to each other.

Those results demonstrate that the ensemble decision tree methods can improve the accuracy over single decision tree method on Microarray data sets. These results are consistent with most machine learning study.

To determine whether the ensemble methods consistently outperform single classification methods, we also conducted a sign test. The results are shown in Table 3. Based on the sign test, we have the following conclusions.

1. AdaBoostC4.5 is the only one among the all compared classification algorithms that outperforms C4.5.
2. Comparing between ensemble methods, Random Forests and AdaBoostC4.5 outperform BaggingC4.5 significantly.
3. No sufficient evidence supports that any ensemble method and C4.5 outperform LibSVMs.

We have the following observations from the sign test. The average difference of 6.5% (between Random Forest and LibSVM) may not be statistically significant, but the average difference of 0.9% (between AdaBoostC4.5 and BaggingC4.5) are statistically significant. This may sounds strange, but is understandable. The average accuracy indicates the average performance of a method on the data sets. However, the sign test indicates if a method is consistently better than another on each test data set. The accuracy difference can be very small. For example, each accuracy value of AdaboostC4.5 is slightly higher than Bagging C4.5, and hence Sign test shows that AdaBoostC4.5 is significantly better than BaggingC4.5. However, the accuracy improvement is marginal.

This also indicates a limitation of the sign test: the difference of 0.01 and 10.0 are considered the same in the sign test since only plus or minus is used. We conducted a Wilcoxon signed rank test based on Table 2. The results of Wilcoxon signed rank test is shown in Table 4

Table 4 shows that all ensemble methods, Random Forest, AdaBoostC4.5 and BaggingC4.5, are significantly more accurate than C4.5. This conclusion is consistent with most research literature. Though AdaBoostC4.5 performs marginally better than BaggingC4.5 on each data set. The Wilcoxon signed rank test does not support that the differences are significant. We tend to believe that the Wilcoxon signed rank test is better than sign test for our purpose.

Based on Table 2 and table 4, we can conclude that all ensemble methods significantly outperform C4.5. We do not have sufficient evidence to show whether LibSVM and another method is better. Though Table 2 give a large average accuracy difference between an ensemble method and LibSVM, we do not know wether LibSVM and an ensemble method will perform better on a data set. This is because that SVM and decision trees are two different types of classification methods. They are suitable for different data sets.

To show that all methods benefit from data preprocessing, we conducted experiments on original data sets, and show their accuracy results in Table 5.

Table 2 and Table 5, clearly indicate that all classification methods on data preprocessed by discretization and gene selection methods achieve higher average accuracy over themselves on data without data preprocessing. After data preprocessing, accuracy performance has been improved significantly for all compared classification algorithms with up to 17.4% improvement.

To show that this improvement is significant, we conducted Sign test and Wilcoxon signed rank test on differences between accuracies on preprocessed data and original data. The test results are shown in Table 6 and Table 7.

Based on a sign test of 95% confidence level, All methods except C4.5 improve the predictive accuracy on the preprocessed Microarray data sets than the original data sets. Not enough evidence supports that C4.5 performs significantly better on the preprocessed

	C4.5	Random Forests	AdaBoostC4.5	BaggingC4.5	LibSVMs
C4.5	–				
Random Forests	$\leq$ **0.05**	–			
AdaboostC4.5	**0.005**	0.2-0.3	–		
BaggingC4.5	**0.025**	0.1-0.2	0.091	–	
LibSVMs	0.5	0.4-0.5	0.4-0.5	0.4-0.5	–

Table 4: Summary of Wilcoxon signed rank test between any two of the compared classification methods. P values are shown and significant p-values at 95% confidence level are highlighted.

Data set	C4.5	Random Forests	AdaBoostC4.5	BaggingC4.5	LibSVMs
Breast Cancer	62.9	61.9	61.9	66.0	52.6
Lung Cancer	95.0	98.3	96.1	97.2	82.9
Lymphoma	78.7	80.9	85.1	85.1	55.3
Leukemia	79.2	86.1	87.5	86.1	65.3
Colon	82.3	75.8	77.4	82.3	64.5
Ovarian	95.7	94.1	95.7	97.6	87.0
Prostate	33.3	52.4	33.3	42.9	61.9
Average	75.3	78.5	76.7	79.6	67.1
Difference	14.3	16.3	17.4	13.6	21.2

Table 5: Average accuracy on seven original data sets of five classification methods based on tenfold cross-validation. The last row shows the differences in average accuracy between the average accuracy based on preprocessed data and original data for every compared classification method

Microarray data sets than the original data set.

These results show that the data precessing method improves the predictive accuracy of classification. As we mentioned before, Microarray data contains irrelevant and noisy genes. Those genes do not help classification but reduce the predictive accuracy . Microarray data preprocessing is able to reduce the number of irrelevant genes in Microarray data classification and therefore can generally help to improve the classification accuracy.

Apart from predictive accuracy, the representation of predictive results is another important fact for determining the quality of a classification algorithm. Among the compared algorithms, the classifier of C4.5 is a tree, and the classifier of an ensemble method is formed by a group of trees. Trees are more easier to be evaluated and interpreted by users. By contrast, the outputs of SVMs are numerical values and are less interpretable.

5 Conclusion

In this paper, we conducted a comparative study of classification methods for Microarray data analysis. We compared five classification methods, namely LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5, and Random Forest, on seven Microarray data sets, with or without gene selection and discretization. The experimental results show that all ensemble methods are significantly more accurate than C4.5. Data preprocessing significantly improves accuracies of all five methods. We conducted both sign test and Wilcoxon signed rank test to evaluate the performance differences of comparative methods. We observed that the Wilcoxon signed rank test is better than the sign test. We also found that there is no sufficient evidence to support the performance difference between the SVM and an ensemble method although the average accuracy of SVM is much lower than that of an ensemble method. A possible explanation is that they are two different classification schemes, and hence one may be able to suits for a data set whereas the other does not.

References

Alizadeh, A., Eishen, M., Davis, E., Ma, C. & et al. (2000), 'Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling', *Nature* **403**, 503–511.

Alon, U. & et al. (1999), 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays', *PNAS* **96**, 6745–6750.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Breiman, L. (1999), Random forests–random features, Technical Report 567, University of California, Berkley.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Jr, M. & Haussler, D. (2000), Knowledge-based analysis of microarray gene expression data by using suport vector machines, *in* 'Proc. Natl. Acad. Sci.', Vol. 97, pp. 262–267.

Brown, M., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, M. & Haussler, D. (1999), Support vector machine classification of microarray gene expression data, Technical Report UCSC-CRL-99-09, University of California, Santa Cruz, Santa Cruz, CA 95065.

Conover, W. J. (1980), *Practical nonparametric statistics*, Wiley, New York.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks.', *Machine Learning* **20**(3), 273–297.

Daniel, W. W. (1978), *Applied nonparametric statistics*, Houghton Mifflin, Boston.

Dietterich, T. G. (2000), 'An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization', *Machine learning* **40**, 139–157.

Freund, Y. & Schapire, R. E. (1996), Experiments with a new boosting algorithm, *in* 'International Conference on Machine Learning', pp. 148–156.

Furey, T. S., Christianini, N., Duffy, N., Bednarski, D. W., Schummer, M. & Hauessler, D. (2000), 'Support vector machine classification and validation of cancer tissue samples using microarray expression data.', *Bioinformatics* **16**(10), 906–914.

		with original data			
with preprocessed data	C4.5	Random Forests	AdaBoostC4.5	BaggingC4.5	LibSVMs
C4.5	0.0625				
Random Forests		**0.0078**			
AdaboostC4.5			**0.0078**		
BaggingC4.5				**0.0078**	
LibSVMs					**0.0156**

Table 6: Summary of sign test between accuracy of the compared classification methods on original and preprocessed data sets. P values at 95% confidence level are highlighted.

		with original data			
with preprocessed data	C4.5	Random Forests	AdaBoostC4.5	BaggingC4.5	LibSVMs
C4.5	**0.025**				
Random Forests		**0.005**			
AdaboostC4.5			**0.005**		
BaggingC4.5				**0.005**	
LibSVMs					**0.01**

Table 7: Summary of Wilcoxon signed rank test between accuracy of the compared classification methods based on original and preprocessed data sets. P values at 95% confidence level are highlighted

Golub, T., Slonim, D., Tamayo, P. & et al. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Gordon, G., Jensen, R., Hsiao, L.-L., Gullans, S. & et al. (2002), 'Translation of microarray data into clinically relevant cancer diagnostic tests using gege expression ratios in lung cancer and mesothelioma', *Cancer Research* **62**, 4963–4967.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002), 'Gene selection for cancer classification using support vector machines', *Machine Learning* **46**(1-3), 389–422.

Hu, H., Li, J., Wang, H. & Daggard, G. (2006), Combined gene selection methods for microarray data analysis, *in* '10th International Conference on KnowledgeBased & Intelligent Information & Engineering Systems. To appear'.

Joachims, T. (1998), Text categorization with support vector machines: learning with many relevant features, *in* 'Proceedings of 10th European Conference on Machine Learning', number 1398, pp. 137–142.

Li, J. & Liu, H. (2003), Ensembles of cascading trees, *in* 'ICDM', pp. 585–588.

Li, J., Liu, H., Ng, S.-K. & Wong, L. (2003), Discovery of significant rules for classifying cancer diagnosis data, *in* 'ECCB', pp. 93–102.

Lockhart, D., Dong, H., Byrne, M. & et al. (1996), 'Expression monitoring by hybridization to high-density oligonucleotide arrays', *Nature Biotechnology* **14**, 1675–1680.

Lu, K., Patterson, A. P., Wang, L., Marquez, R. & Atkinson, E. (2004), 'Selection of potential markers for epithelial ovarian cancer with gene expression arrays and recursive descent partition analysis', *Clin Cancer Res* **10**, 291–300.

Maron, O. & Lozano-Pérez, T. (1998), A framework for multiple-instance learning, *in* M. I. Jordan, M. J. Kearns & S. A. Solla, eds, 'Advances in Neural Information Processing Systems', Vol. 10, The MIT Press, pp. 570–576.

PetricoinIII, E., Ardekani, A., Hitt, B., Levine, P. & et al. (2002), 'Use of proteomic patterns in serum to identify ovarian cancer', *The lancet* **359**, 572–577.

Quinlan, J. (1996), 'Improved use of continuous attributes in C4.5', *Artificial Intelligence Research* **4**, 77–90.

Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California.

Santin, A., Zhan, F., Bellone, S. & Palmieri, M. (2004), 'Gene expression profiles in primary ovarian serous papillary tumors and normal ovarian epithelium: idnetification of candidate molecular markers for ovarian cancer diagnosis and therapy', *International Journal of Cancer* **112**, 14–25.

Schena, M., Shalon, D., Davis, R. & Brown, P. (1995), 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science* **270**, 467–470.

Singh, D. & et al. (2002), 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell* **1**, 203–209.

Tan, A. C. & Gibert, D. (2003), 'Ensemble machine learning on gene expression data for cancer classification', *Applied Bioinformatics* **2**(3), s75–s83.

Veer, L. V., Dai, H., de Vijver, M. V. & et al. (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature* **415**, 530–536.

Yeang, C., Ramaswamy, S., Tamayo, P. & et al. (2001), 'Molecular classification of multiple tumor types', *Bioinformatics* **17**(Suppl 1), 316–322.

Data Mining In Conceptualising Active Ageing

Richi Nayak[1], Laurie Buys[2] and Jan Lovie-Kitchin[3]

[1]School of Information Systems, [2] Centre for Social Change Research, [3] Faculty of Health
Queensland University of Technology, Brisbane, Australia

{r.nayak, l.buys, j.kitchin@qut.edu.au}

Abstract

The concept of older adults contributing to society in a meaningful way has been termed 'active ageing'. We present applications of data mining techniques on the active ageing data collected via a survey of older australian on a wide range of social and behavioural variables. The goal is to understand the underlying relationships and attributes which characterise active ageing. The data mining results indicate that an individual's health, attitude to learning, social network support and (positive) emotional feelings are significant contributors to achieving active ageing.

Keywords: Classification, clustering, association mining, Active ageing

1 Introduction

The concept of older adults contributing to society in a meaningful way has been termed 'active ageing' (Kinsella and Phillips 2005). The policy framework for active ageing developed by the WHO emphasizes that health, participation and security are important for quality of life for older adults (WHO 2005). This framework is guiding the conceptual development of "active ageing" in diverse national contexts.

To conceptualise active ageing in terms of complex issues that intertwine and converge with the ageing experience, rather than in a singular health/social dimension, the Australian Active Ageing (Triple A) study at the Queensland University of Technology (QUT) has conducted a national-wide postal survey to collect the responses of older people on a wide range of questions related to 'work', 'learning', 'social', 'spiritual', 'emotional', 'health and home', 'life events' and 'demographics'. Previously, studies such as the Australian Longitudinal Study of Ageing (ALSA) (Andrews, Clark, and Luszcz 2002), and the Dubbo Study of Ageing (Simons et al. 1990) included social aspects as well as the more usual psychological and behavioral (ALSA), and bio-medical issues (Dubbo Study). This study at QUT is the first of its kind reflecting a wide variety of aspects of older adults life.

Data Mining (DM) techniques have been successfully applied to a number of application domains including finance, marketing, health, Internet and others (Han 2001). This paper extends the use of DM to understand how the older population in Australia is ageing. The large number of interrelated variables and the complex relationships between various life aspects of older people's experience greatly enhance the potential of applying data mining (DM) techniques.

We use different data mining tools available within the 'SAS 9.1 Enterprise Miner' to conduct experiments. In particular, we used clustering, predictive modelling and association mining to carry out an exploratory analysis of the data. The results successfully highlight interesting trends in the data and describe that an individual's health, attitude to learning, social network support and (positive) emotional feelings are significant contributors to achieving active ageing. The trends and patterns of this data set will assist to develop the concept of active ageing and contribute significantly to future policy directions.

2 Data Mining Application to Active Ageing

The goal is to pre-process the national survey data; apply an appropriate DM technique or a combination of techniques; and evaluate and interpret the meaningful rules to enhance the understanding of active ageing.

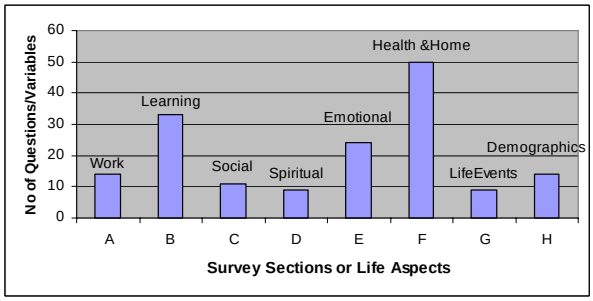


Figure 1: Distribution of Variables

2.1. Data Pre-processing

The first task was to understand the nature of the data set to decide the strategies and methods of data mining. The Triple A study data includes responses of members of a large Australia wide seniors organisation on a wide range of questions related to their life. A total of 2655 surveys

were returned at 46% response rate. Of these 32 incomplete surveys were excluded due to missing age, postcode and/or non-response to survey questions; others were excluded due to respondents being younger than 50 years old.

Survey questions were divided into groups/aspects such as 'work', learning', 'social', 'spiritual', 'emotional', 'health', 'home', 'life events' and 'demographics'. Each question is treated as a variable/attribute and its responses become the values of the variable in the data mining analysis. Figure 1 lists the number of variables in each aspect. There were many transformations done in the data set in order to analyse with data mining algorithms.

2.1.1 Empty Fields

A major problem in preparing data for mining was that some questions requested the respondents to skip questions depending on their response to certain questions. These variables were subsequently modified to limit the effect of the large percentage of respondents who didn't answer them. For example, variable A1 requested respondents to skip to A4 if they answered no, which led to a large percentage of respondents leaving variables A2 and A3 blank. Therefore A2 and A3 were modified to include a response of N/A for those respondents who were not required to do so.

2.1.2 Value Transformation

There were some variables that were transformed to fulfil the goal of the mining process. For example, the variable H2 asked respondents to supply their age. Since the Triple A project is only interested in analysing the various age groups, therefore the data was modified to reflect the group that the age variable belongs to (55-64; 65-74; 75 years and over). Similarly, the variable H3 asked respondents to supply their postcode. We found that the postcode had little effect on the data, and therefore used it to calculate the state that the respondent lived in, whether they lived in a metropolitan or regional area, and to which socio-economic status they belong.

Finally, the processed data consists of 2,623 cases and 165 variables. Majority of variables are categorical. The distribution analysis showed that the data set is suitable for any data mining technique to be applied. The result of the data mining would highly depend upon how we formulate the problem.

2.2 Conducting Data Mining

The main objective in this study was to find the inter-relation between various aspects of life to achieve active ageing. We decided to first apply the clustering analysis to understand the nature of the data set overall. We utilized the most commonly used centroid based approach *k-means* (Jain, Murty and Flynn 1999). Next, we applied the association and predictive modelling to understand the deeper relationships between various sections and various variables.

We had many choices for the predictive modelling (Lim and Loh 2000). Since, our main goal is to understand how the variables from different life aspects are related to each other, good comprehensibility and high accuracy were the main requirements in a method. *Decision tree* is chosen due to their ability to obtain reasonable accuracy, good comprehensibility, efficiency, robustness and scalability.

The *Apriori algorithm* is used for the association analysis. As support is lowered the processing time to find association rules dramatically increases and it is often infeasible to find all association rules. Consequently only maximal association rules are chosen as it reduces the number of rules that need to be analyzed. A maximal association rule is a supported association rule that isn't contained in another supported association rule (Agrawal, Imielinski and Swami 1994). Typically interesting maximal association rules are only supported by a low percentage of entries.

3. Analysis of generated clusters

To understand if there is a clear segmentation in the data set, clustering analysis is performed. The k-means clustering algorithm divided the data set into 7 distinct but overlapped clusters. There is no segmentation of the value ranges of the variables, such that all values within a continuous interval of each variable belong to the same cluster. The clusters are quite close to each other. This indicates that variables don't define distinct boundaries, though the given variables are slightly more interesting than others. There exist some highly correlated relationships within the data which should lead to interesting rules for active ageing. The figure 2 shows the number of respondents grouped in 7 clusters.

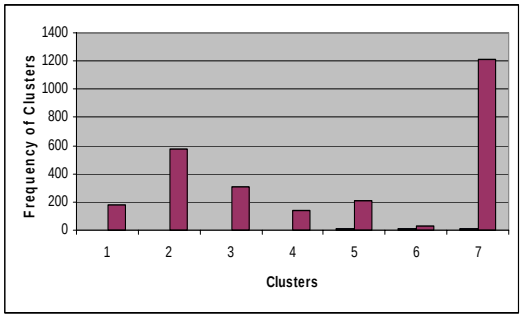


Figure 2: Cluster Distribution

The cluster 7 that groups the majority of respondents showed that the people feel contented in the current life when they enjoy the social interactions, and they have more intention to learning new things. Also, health did not limit them in doing the outside activities.

The cluster 2 grouped the people who do not feel contented in their current life. Respondents in this group have lower social interaction, and are less capable of handling their social relationships .The cluster 3 grouped the people according to the difficulty level at performing daily activity. Majority of people were from the age cohort 75 years and over, and some from 65-74 years.

The figure 3 shows (1) the distribution of the 44 variables that appear in clusters across each aspect of life, and (2) the distribution of the 44 cluster-variables in each aspect in comparison to the original distribution of the total 165 variables in each aspect of the data set. For example, the

aspect B contributes 27% (12 out of 44) of variables in the total number of variables appearing in the clusters, whereas the entire data set includes only 33 (i.e. 20%) aspect B variables (figure 1).

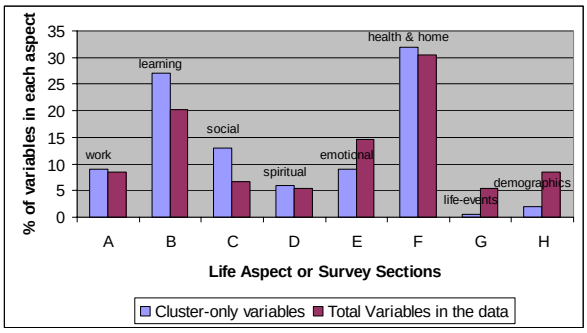


Figure 3: Distribution of Variables in Clusters

The figure 3 shows that the 'Health & Home' and 'Learning' are the most prominent in terms of number of variables in clusters. This indicates that an individual's health and attitude to learning are the most significant contributors to active ageing. However, when comparing the cluster variables and original variable distribution, the figure 3 shows that the variables on 'Social', 'Learning' and 'Health & Home' dimensions have more impact, since their proportion in the clusters is much higher than the original.

The Variable D4, which asked whether the responders feel contented in their present life, is found to be the most influential to form clusters based on the relative performance measure of the variables in determining the clusters.

The results of the cluster analysis indicate that there are correlated patterns within the data. However it is necessary to perform the predictive modelling or association analysis on a selection of variables to determine whether those correlations have any meaning.

	# Variables		# Rules		Misclassification Rate	
Selected target from all targets	All	Cluster	All	Cluster	All	Cluster
A1: Full time or part time paid work	29	33	57	90	3.33%	4.06%
A2: #Hours paid work per week	79	43	145	192	11.79%	19.53%
A12: Voluntary activities or Not	64	38	116	143	9.26%	18.14%
A13: Involved in tertiary degree courses or diplomas	84	42	156	206	13.49%	16.43%
B4: Interested in new or current event program	98	41	196	206	21.30%	27.32%
B10: Extent to open to new technology	79	41	187	222	15.13%	19.55%
B11: Need to learn to enjoy learning new things	78	39	191	190	14.66%	16.70%
B12: Need to learn to organise holiday/travel arrangements	82	39	183	204	23.74%	38.38%
C1: # (Emotionally) close People within one hour travel away	109	43	207	274	55.02%	55.59%
C3: #family and friends who had phone conversation in the past week	113	42	231	245	43.35%	46.91%
C5: Family and friends understands the respondent or Not	103	42	209	217	14.78%	18.81%
C6: Feel useful to family or friends or Not	112	42	225	225	17.28%	20.51%
D1: Believe in a higher being or Not	104	43	207	262	30.50%	40.37%
D3: Feel in control of your life or Not	95	42	208	230	20.15%	23.31%
D4: Feel contented in your life	110	43	226	223	20.31%	24.68%
D9: Feel as searching for personal meaning	108	42	221	246	36.32%	37.81%
E1: Confidence in the opinion or Not	105	42	206	228	21.90%	25.55%
E12: Happy with the personality or Not	103	43	214	244	18.70%	23.20%
E23: Satisfied with accomplished in life	113	42	256	230	26.49%	30.63%
F3f: Limitation of health in bending	101	42	215	241	17.65%	22.06%
F3g: Limitation of health in walking	78	42	138	171	10.09%	14.54%
F4b: Limitation of health in accomplishing the wishes	105	43	210	207	14.82%	19.05%
F4c: Limitation of health in work	78	39	160	166	11.87%	14.23%
G1:Effect of a major illness in life	97	41	205	237	16.78%	20.46%
G2: Effect of change of work in life	90	41	208	209	19.20%	23.19%
G3: Effect of a new study course in life	100	41	182	195	16.88%	19.13%
G6: Effect of child-care activities in life	94	40	171	168	12.67%	15.16%
H1: Gender	92	42	186	225	13.99%	18.77%
H2: Age group	122	42	222	274	17.19%	20.18%
H3a: Residential State in Australia	117	41	228	211	39.74%	45.54%
H4: Country of Birth	93	42	169	203	12.97%	15.33%

Table 1: Predictive Models Details

4. Analysis of Predictive Models

Predictive modelling or classification is performed to establish relationships that exist between various sections (life aspects) and various variables present in them. We wanted to examine how a variable in one aspect is related to other variables. We have chosen four of the target variables in each life aspect based on (1) the equal distribution among their representative class values, and (2) their appearance in clusters.

We build the decision tree models for each target with two kinds of input attributes: (1) all the rest of attributes as dependent labelled as 'All', and (2) only the rest of the attributes that appear in the clusters labelled as 'Cluster'. The results in Table 1 show the details of the models for each target with the average performance on 10-fold CV experiments. We report the number of variables contained in the rules, the number of classification rules, and the misclassification rate for the classifiers.

The relatively small misclassification rates (26 out of 32 have < 30% misclassification) in Table 1 show that the models can be found that can accurately describe some of the data. This indicates that there are strong patterns within the data. The large range of required variables for the 'All' group in Table 1, and the poor misclassification rates in the models using only 'Cluster' variables (A12, B12, D1 and E16 in particular), indicates that the subset of variables from the cluster are insufficient to accurately describe all of the given targets. This also helps to reveal the nature of overlapped clusters. These results confirm that since respondents have very similar backgrounds (such as similar life style as supported by the statistical analysis of responses), the k-means clustering is not able to categorize them in distinct and disjoint classes. Accordingly, clustering variables alone can not be used for building a classifier model.

The Figure 4 illustrates how each life aspect influences in building predictive models by showing the percentage of variables from each aspect that form the model for the given target. The average of the targets for each aspect is calculated by dividing the number of variables for the given aspect by the number of variables required to classify the target. We compare the percentage of variables that exist in each section (or life aspect) in the survey with the mean percentage of dependence for each aspect based on experiments. The Figure 4 shows that the Learning, Emotional, and Health and Home aspects (B, E, F) are the most significant, as combined they contribute 64.49% of all classification variables. This indicates that physical and emotional health combined with the desire to learn are the most significant factors when considering active ageing. However, these aspects also contribute 65.24% of the total variables in the data set so this could be expected. It is therefore necessary to conduct further analysis to measure the significance of each life aspect.

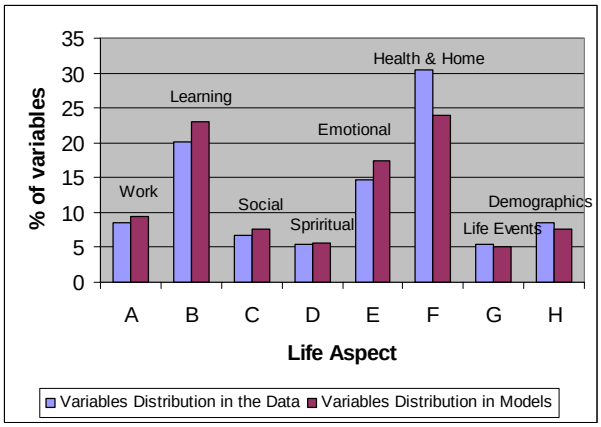


Figure 4: Average Aspect Dependencies

The Table 2 shows the analysis to indicate the aspects on which some targets (life aspects) rely more so that the variables from those aspects can be deemed important in determining those targets. The results in Table 2 that are shaded dark grey highlight results that have a 20% or more difference from the "Variables" row. The "Variables" row shows the percentage of variables that exist in each life aspect in the data. The "Mean" row shows the mean percentage of dependence for each aspect based on experiments. The highlighted results indicate that the target from the given aspect rely on the prescribed aspect more than is statistically expected. Therefore the variables from those aspects are quite important in determining those targets.

Target	**Section Breakdown**							
Section Average	**A (%) Work**	**B Learning**	**C Social**	**D Spiritual**	**E Emotional**	**F Health & Home**	**G Life Events**	**H Demographics**
Variables	8.54	20.12%	6.71%	5.49%	14.63%	30.49%	5.49%	8.54%
A	12.48%	20.76%	9.30%	5.19%	14.07%	18.46%	7.03%	12.71%
B	10.99%	27.36%	6.20%	5.58%	19.78%	19.15%	5.38%	5.56%
C	8.91%	23.83%	7.99%	5.08%	18.33%	25.86%	4.08%	5.93%
D	8.84%	22.93%	8.02%	6.28%	17.58%	24.17%	4.74%	7.44%
E	6.93%	23.11%	7.42%	6.54%	19.48%	24.23%	4.66%	7.64%
F	8.53%	20.87%	7.32%	5.94%	15.06%	29.88%	4.75%	7.65%
G	9.23%	23.67%	8.47%	4.71%	15.11%	26.29%	6.01%	6.50%
H	9.13%	22.25%	6.65%	5.96%	19.55%	24.10%	4.53%	7.84%
Mean	9.38%	23.10%	7.67%	5.66%	17.37%	24.02%	5.15%	7.66%

Table 2: Detailed Analysis of Dependencies of Each Aspect

Targets from aspect A are more reliant on variables in aspects A, C, G, and H and less reliant on aspect F. Targets from aspect B are more reliant on variables in aspect A, B, and E and less reliant on aspects F and H. Targets from aspect C are more reliant on variables in aspect E, and less reliant on aspects G and H. Targets from aspects D, E, and H are more reliant on variables in aspect E, and less reliant on aspect F. Targets from aspect G are more reliant of variables in aspect H.

Therefore it can be concluded that variables in E (Emotional) aspect are more significant, and variables in F (Health & Home) are less significant than any other aspects (according to the difference in values of "Variables" and "Mean" row). Despite this it is still evident that aspects B, E, and F are the most important in this survey.

To further investigate the dependencies of various life aspects and variables, we generated models to predict the selected target considering only the variables from each aspect. We compare the dependencies of each aspect in building the predictive models independently with the model that includes all variables from all aspects. The misclassification rates of the model based on all variables versus models reliant on only variables from the specified aspect are shown in Figure 5.

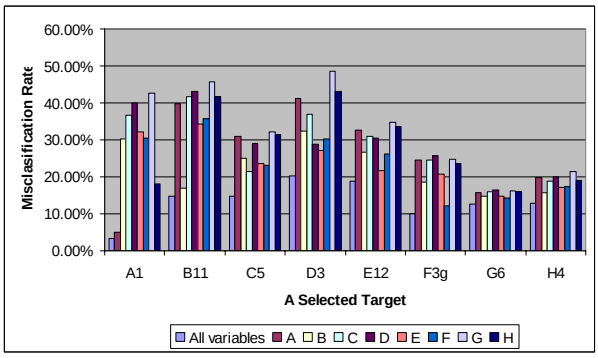


Figure 5: Models versus Life Aspects

The figure 5 shows that every target except G6 and H4 produces significantly more accurate models when only relying on variables from their respective aspect. The G6 and H4 targets can be accurately predicted (<= 21.4% misclassification) using variables from any aspect. A total of 81.81% of respondents answered "No" to G6: "Were they involved in any childcare activities in the past year?" and 76.07% of respondents answered "Australia" to H4: "Country of Birth". Therefore when such a high percentage of respondents answered the same way it is highly likely that strong relationships exist between many variables that can explain the data.

The G6 target is within 30% of accuracy of the base model for every aspect except aspect D, which reiterates the previous conclusion that G6 can be explained by several variables. However, excluding G6, only B11, E12, and F3g have models that rely on their own aspect so strongly. In simpler words, only these models are within 30% of the accuracy of the base classifier. Therefore it is reasonable to assume that strong correlations exist within an aspect. Again 'Learning', 'Emotional', and 'Health and Home' (B, E, and F) appear to be relied upon the most heavily.

5 Analysis of Association Rules

Association rule mining was performed to establish dependencies between various variables in certain selected sections without any preconditions given.

No of Association Rules for the given Support			
	All	Male	Female
100%	0	0	0
90%	44	0	0
80%	636	0	0
70%	5,388	0	0
60%	43,507	0	0
50%	N/A	0	64
40%	N/A	19	3,811
30%	N/A	9,288	N/A

Table 3: Association Rules versus Support

The results in Table 3 show the number of maximal association rules found for the different data sets at the given support requirements. The "All data set" includes all of the respondents whilst the "Male and Female data sets" contain only male and female respondents respectively. The higher number of rules inferred for female than male shows that female respondents are in more agreement. Results for the Female data set with 50% minimum support are supported by 28.3% of respondents, as 56.7% of respondents are female in the survey.

Note that as support is lowered there is a dramatic increase in the number of maximal association rules found. It becomes infeasible after some point due to excessive computation efforts (e.g., at and below 50% support for the All data set). Therefore to find interesting association rules it is necessary to create meaningful subsets such as the Male and Female groups.

Variables from various Aspects (in %)				
	At 90% support	80%	70%	60%
A	0	0	0.2	1.7
B	0	0	4.1	12.3
C	0	0	0	0.3
D	0	0	0	0
E	0	0	0	0.02
F	100	99.8	99.9	99.9
G	25	58.3	72.8	77.5
H	0	6.9	20.9	30

Table 4: Percent of Association Rules Containing Variables from Various Aspects

The results in Table 4 show the percentage of rules that contain at least one variable from the specified section for the All data set. Clearly section F is important as it is contained in over 99% of all maximal association rules for the given tests. Section G is the next most important whist sections A, B, C and H are not as important. However section D has no variable, and section E has only one variable where 60% of respondents gave the same answer. Therefore to find interesting variables containing sections D and E a much lower minimum support value is required. As stated previously though it is infeasible to find all maximal association rules with low support, and therefore specific subsets must be chosen to find interesting association rules within these sections.

6. Discussion and Conclusion

The Triple A project at the Queensland University of Technology, Brisbane, Australia incorporates a wide scope of issues significant to the quality of life for older people. With the use of the data mining techniques, this paper examined the inter-relationships of a wide range of 'work', 'learning', 'social', 'spiritual', 'emotional', 'health and home', 'life events' and 'demographics' variables in order to identify those that contributed most strongly to positive responses and could therefore be indicators of "Active Ageing".

Firstly, clustering was applied to understand the nature of the data set. The respondents were put together into seven groups according to their feeling of 'contented' in their life, physical health limitations, etc. Moreover, the overlapped clusters showed that variables related to the 'Health and Home', 'Learning', 'Social' and 'Emotional' sections are prominent in deciding clusters of common characteristics. This indicates that an individual's health, attitude to learning, social network support and emotional feelings are significant contributors to positive overall wellbeing which is an indicator of active ageing. The results of the cluster analysis indicate the need to perform predictive modelling on a selection of variables to determine whether those correlations have any meaning.

A number of decision tree models were built to further explore the relationships between various variables targeted to certain variables. The results show that the subsets of variables as they appeared in clusters alone are insufficient to accurately describe all of the given targets. This emphasizes that a complex relationship exists between variables to describe a concept, and all the variables do contribute significantly.

We also modelled several decision trees considering all variables to predict a number of targets chosen from each aspect. The analysis of number of variables appearing in various models shows that the 'Learning', 'Emotional', and 'Health and Home' aspects are more significant, as combined they contribute 64.49% of all model variables. Further analysis shows that 'Emotional' feelings are more significant, and 'Health and Home' is less significant than any other aspect. Factors related to 'Learning' are also significant. This was evident when we closely examined the trees. Majority of rules have dominance of 'emotional feelings' and 'learning needs and interests' in explaining active ageing concepts.

Association analysis was also performed to find the rules of conceptualising active ageing without any preconditions or bias given as in predictive modelling. Association rules are inferred based on the rate of agreement in overall responses. The variables related to 'Health and Home' appeared in over 99% of all rules for the given tests. The 'Life Events' information was the next important. This indicates that majority of respondents in this survey do agree on questions about two aspects of life, which is related to the fact that respondents were selected from the same database. The majority shared similar characteristics such as being comfortable financially, living with a companion, in their own home in a metropolitan area and being well educated.

A number of lessons are learned by applying data mining to a sample survey data for the purpose of knowing more about the complex ageing process.

- The survey data includes a homogenous population of respondents in terms of education and finance status. For more insight, we need to mix various populations such as those with diverse living status, lower education, lower financial status etc. Regardless, the data mining techniques were able to capture the essence of the data set to reflect general positive engagement in life. It reinforces the fact that the inferred patterns are as good as the data are.
- Clustering is a good technique to understand the basic nature of the data set. However to understand the data in depth, the application of predictive modelling and association analysis techniques were required.
- The main goal of this data mining application was to understand the survey data and learn if there are any patterns and rules exit that reflect the population. This emphasizes the need for good comprehensibility of output results, which demonstrates the reason that decision tree was chosen for predictive modelling.
- Spending significant time in preprocessing, formulating the various subsets of problem and analyzing the comprehensible results were the main reasons for the success of the project. The data mining output alone is not sufficient; there were a large number of decision tree rules. However, a detailed analysis of these models resulted into a detailed comparison of various life aspects. So the analysis of the output result should get equal importance.
- Data mining analysis successfully highlighted complex issues that intertwine and converge with the ageing experience, rather than in singular health or economic dimensions. Our results present a portrait of Australian older adults that is distinctly different from the stereotype generated by models which negatively portray ageing as a process of decline.

The identified relationships assist us to better understand the impact of ageing upon older Australians' engagement in society and their general sense of well being. The identified patterns will help policy makers and service providers to provide services that meet the needs of older people in the future.

The future research includes (1) the analysis of data with hierarchical clustering as many variables are strongly correlated and (2) the association mining analysis to shift from general criteria to more specific ones by regrouping the variables into interesting types such as communication technologies, TV, car driving etc.

7. Acknowledgement

This work was partly supported by the QUT ECR and the QUT Strategic Collaborative Grant Scheme. We would like to thank all the team members of the Triple A project. We would also like to thank Calum Robertson, Lei Zhang, Xin Li and Ying Luo for their assistance in conducting experiments.

8. References

Agrawal, R., Imielinski, T. and Swami, A. (1993): Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD International Conference on Management of Data*, Washington DC, USA, **22**:207-216, ACM Press.

Andrews G, Clark M, Luszcz M. (2002) Successful aging in the Australian Longitudinal Study of Aging: Applying the MacArthur model cross-nationally. Journal of Social Issues; 58(4): 749-765.

Han J. (2001) *Data Mining: Concepts and Techniques* Morgan Kaufmann Publishers, San Francisco

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys (CSUR), 31*(3), 264-323.

Kinsella K. & Phillips, DR (2005). Global aging: The challenge of success. Population Bulletin 60(1): 3-40.

Lim, T. S. & Loh, W. Y. (2000). A comparison of prediction accuracy, complexity and training time of thirty three old and new classification algorithms. *Machine Learning, 40*(3), sep. 203-228.

Simons LA, McCallum J, Simons J, Powell I, Ruys J, Heller RLC (1990). The Dubbo study: An Australian prospective community study of the health of elderly. Australian & New Zealand Journal of Medicine; 20: 783-789.

World Health Organisation [WHO]. Active Ageing: A Policy Framework. [Online] [Accessed 2005 April]. Available from URL http://www.who.int/hpr/ageing/ActiveAgeingPolicyFrame.pdf

Analysis of Breast Feeding Data Using Data Mining Methods

Hongxing He[1] **Huidong Jin**[1,2] **Jie Chen**[1] **Damien McAullay**[1]
Jiuyong Li[3] **Tony Fallon**[3]

[1]CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra ACT 2601, Australia.
[2]National ICT Australia (NICTA), Canberra Lab, Canberra, Australia.
[3]University of Southern Queensland, Toowoomba QLD, 4350, Australia.
Email: hongxing.he@csiro.au, huidong.jin@nicta.com.au,jiechen@ieee.org,
Damien.McAullay@csiro.au,jiuyong@usq.edu.au, fallon@usq.edu.au

Abstract

The purpose of this study is to demonstrate the benefit of using common data mining techniques on survey data where statistical analysis is routinely applied. The statistical survey is commonly used to collect quantitative information about an item in a population. Statistical analysis is usually carried out on survey data to test hypothesis. We report in this paper an application of data mining methodologies to breast feeding survey data which have been conducted and analysed by statisticians. The purpose of the research is to study the factors leading to deciding whether or not to breast feed a new born baby. Various data mining methods are applied to the data. Feature or variable selection is conducted to select the most discriminative and least redundant features using an information theory based method and a statistical approach. Decision tree and regression approaches are tested on classification tasks using features selected. Risk pattern mining method is also applied to identify groups with high risk of not breast feeding. The success of data mining in this study suggests that using data mining approaches will be applicable to other similar survey data. The data mining methods, which enable a search for hypotheses, may be used as a complementary survey data analysis tool to traditional statistical analysis.
Keywords: Data Mining, Survey Data, Features Selection, Classification, Association Rule

1 Introduction

Breast feeding is acknowledged by the World Health Organisation (WHO 2001) to be the optimal method of infant feeding. It has been shown in past literature to provide physical and psychological benefits to both mother and child (Kramer & Kakuma 2003). Additionally, there is evidence to suggest that increasing initiation of breastfeeding and breastfeeding duration have environmental benefits, as well as economic benefits, both to health care systems and individual families (Riodan 1997, Smith 2001, Smith et al. 2002).

It is therefore important to study the factors leading to decisions on baby feeding method. A research team at the University of Southern Queensland has conducted research on this issue (Hegney et al. 2003). In the project, all mothers giving birth in the two Toowoomba hospitals between July 10 and November 30 in 2001 were approached to participate in the study prior to being discharged from the hospital. Mothers who decided to participate agreed to fill out a pre-discharge questionnaire prior to discharge or by telephone shortly after discharge. They were then contacted via telephone at three-months and six-months postpartum to complete follow-up surveys. Of the 940 mothers eligible to participate at discharge, 625 (67%) chose to participate. 554 (89%) mothers were able to be contacted at the three-month follow-up. Of the 372 mothers who were breastfeeding at three-month follow-up, 329 (88%) mothers could be contacted at six-month follow-up.

An extensive study has been carried out on the data collected in the project to study the factors which influence the decision on whether or not breast feeding is given by mothers. In the study, detailed uni-variate analyses were carried out to evaluate the role of individual factors on the output variable. Logistic regression has also been applied to the problem (Hegney et al. 2003). We take an alternative data mining approach in this paper. We apply a feature selection module to select the most discriminating and least redundant features from the original feature set. In selecting the feature subset we do not assume any prior domain knowledge; we let the data speak for themselves. The features selected are then used to train a decision tree model or logistical regression to classify the individuals with respect to an output variable. The output variable is used as the class variable of the individual subjects. In this approach, we do not consider features individually, rather we consider feature sets or subsets as a whole that will influence the decision of whether breast feeding or other feeding method will be used. We also apply a risk pattern mining approach to identify groups of mothers who are not likely to breast feed their babies. These rules should be able to help to conduct a targeted education program to promote breast feeding more effectively.

The remainder of this paper is organised as follows. Section 2 discusses the feature selection methods used as a data pre-processing step of data mining. Section 3 explains two types of classification tools used in the study. Section 4 describes the risk pattern mining method. Section 5 presents main results and discussions of the data mining application to the breast feeding data. Section 6 concludes the paper.

2 Feature Selection

Questionnaires often incorporate a large number of questions to capture as much information from respondents as possible. It is important to do so as there may be limited opportunities to gather this information from the respondents, so the study should be over-inclusive rather than exclusive. However, not all of this information is going to be useful when trying to answer a particular question. Some irrelevant or redundant features are likely to be included in the

survey design. Therefore, feature selection becomes an important step before the data can be properly analysed . In the current study we consider two different approaches used in our feature selection procedure. One of them is a selection algorithm based on information theory and the other is a statistical approach (Chi-square).

2.1 Information Theory Based Feature Selection

Feature selection methodology based on information theory is a type of *filter* approach (Fleuret 2004, Wang et al. 2004). A filter approach is classifier independent, where relevance of features to the class variable and correlation between features are studied in order to select the most important features. The other type of general approach is *wrapper* (Kohavi & John 1997), which is classifier dependent and various subsets of the original features are compared to identify the best option in terms of the size of feature subset and classification accuracy using a classifier. For our studies, we use the filter approach. The feature selection method does not rely on any classifier. The feature subset selected can then be used to do classification with various classifiers. They can also be used for other data mining tasks, say, clustering Jin et al. (2005) and visualisation Jin et al. (2004).

The information theory based feature selection method uses concepts from *entropy* and *mutual information* (Shannon. 1948, Cover & Thomas. 1991) as a basis for selecting discriminating and non-redundant features. The entropy, measuring uncertainty of a variable, is defined in Equation 1.

$$H(x) = -\sum_{i=1}^{n} P_{x_i} \log P_{x_i} \tag{1}$$

Where P_{x_i} is the probability of x taking the value x_i. Variable x takes n distinct mutually exclusive values. The Mutual Information (MI) or information gain is defined in Equation 2.

$$\begin{aligned} I(y;x) &\triangleq IG(x \mid y) = H(x) - H(x \mid y) \\ &= H(y) - H(y \mid x) = H(x) + H(y) - H(x,y) \end{aligned} \tag{2}$$

Where $H(x,y)$ is defined by Equation 3.

$$H(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{x_i,y_j} \log P_{x_i,y_j} \tag{3}$$

We apply a feature selection algorithm FIEBIT (Feature Inclusion and Exclusion Based on Information Theory) (He et al. 2005), developed recently to select the most discriminating and least redundant features. FIEBIT uses Conditional Mutual Information (CMI) while excluding irrelevant and redundant features according to the comparison among Individual Symmetrical Uncertainty (ISU) and Combined Symmetrical Uncertainty (CSU). The Conditional Mutual Information of y, x_n given x_m can be defined as:

$$\begin{aligned} I(y;x_n|x_m) &= H(y \mid x_n) - H(y \mid x_n, x_m) \\ &= H(y, x_m) - H(x_m) \\ &= -H(y, x_n, x_m) + H(x_n, x_m) \end{aligned} \tag{4}$$

Furthermore, if we normalise mutual information, we may introduce some symmetric measures. For example, following Yu & Liu (2004), we may use Individual Symmetric Uncertainty (ISU) to describe the correlation between a feature x and class variable y. Basically, it is the mutual information (or information gain) between two variables normalised by the sum of their individual entropy.

$$ISU(x;y) = 2\frac{I(y;x)}{H(x) + H(y)}. \tag{5}$$

The ISU compensates for mutual information bias toward features with more values and restricts its values to the range [0,1]. In addition, it still treats a pair of features symmetrically.

Similar to the ISU, we can treat feature $x_j \times x_i$ as the domain $x_{j,i}$ to define Combined Symmetric Uncertainty (CSU) with respect to class variable y.

$$CSU(x_j, x_i; y) = 2\frac{I(y; x_j, x_i)}{H(x_j, x_i) + H(y)}. \tag{6}$$

The feature selection method uses Conditional Mutual Information Maximisation (CMIM) introduced by (Fleuret 2004, Wang et al. 2004) to select the $(k+1)^{th}$ feature based on Equation 7 when k features have been selected.

$$f(k+1) = \arg\max_{n}(\min_{1 \le l \le k} I(y; x_n \mid x_{f(l)})) \tag{7}$$

The process continues until the desired number of features are selected.

Feature Inclusion and Exclusion Based on Information Theory (FIEBIT) chooses the features with the highest minimum conditional mutual information of the features not selected-so-far. If k features have already been selected, Equation 7 selects the $(k+1)^{th}$ feature.

After each new feature is selected, FIEBIT excludes the redundant features using ISU criteria Yu & Liu (2004). The candidate set then becomes smaller for each step. FIEBIT can therefore efficiently select a near-optimal feature subset without pre-defining the number of features to be selected. The detail of the algorithm implementing FIEBIT can be found in (He et al. 2005).

2.2 Statistical Approach in Feature Selection

We compare our data mining approach to a statistical approach which uses the Chi-square test for selecting the features (Yang & Pedersen 1997). The χ^2 defined by Equation 8 quantitatively measures the relevance of a condition to the outcome.

$$\chi^2 = \sum_{i=1}^{n} \frac{(E_i - O_i)^2}{E_i} \tag{8}$$

where O_i is the ith actual value while E_i is its expected value. It takes value 0 if the feature has no effect whatsoever on the outcome, which is commonly called the null hypothesis in statistics. In other words, the output variable is independent of the input variable. A large χ^2 value implies a great importance in deciding the value of the output variable by the variable. Therefore, we select the variables which have high χ^2 values with the output variable. In order to overcome the bias of the population selection, we calculate the p-values which indicates the statistical significance of the χ^2 value.

Chi-square is a test of statistical significance for bi-variate tabular analysis. We use the following two criteria to select n_f features to form a selected feature subset. n_f is a user predefined number.

1. The P value is lower than 0.05.

2. Top n_f features in the list sorted by χ^2 in descending order.

The second criterion selects the features unlikely to be independent to the output variable. The first condition guarantees the result to be statistically significant at the level 0.05.

3 Classification

Classification is one of the most popular data mining tasks. It aims to classify subjects automatically by labelling each subject as a class index. All subjects are then divided into distinct classes. For example, in the breast feeding survey data we can use a feature indicating that the mother chooses to breast feed her baby or not as the class variable. The objective of classification is to predict the class variable using descriptive variables automatically. In order to classify automatically, a reliable model needs to be created. There is a learning process to train the model to perform the classification task.

There are generally two types of learning systems for training the model. Supervised learning uses training samples to optimise the parameters in the training model. Unsupervised learning automatically divides the subjects into various classes in such a way that the subjects belonging to the same class are similar to each other and subjects belonging to different classes are dissimilar. Supervised learning is the most popular approach in classification when study samples are available. It is therefore applied in the current study. Classification models may suffer from the over-fitting problem. The model may achieve very high accuracy on the training samples, however, it may perform poorly on generalisation. Therefore, we need some kind of validation method to test its generalisation accuracy. We use *leave-one-out* as a validation method. In the *leave-one-out* approach we use one data record in turn as the test data. All the other data records are used to train the classification model. The trained model is then applied to the single subject, which is not used in the training process. In general, there will be some correctly and wrongly classified subjects after N runs. The average error rates are then calculated on N runs (N is the total number of data records).

3.1 Decision Tree

Decision tree is a popular supervised learning method used in data mining. Decision tree describes a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications. It is easy to visualise a decision tree. It has advantages over other so called black box classification tool, such as neural network, for having more explanatory power. It not only gives the decision on the classification but also presents the reasoning behind the decision. In our breast feeding data application, output variable y is a categorical variable. The decision is made by a classification tree rather than a regression tree where y takes continuous values. As mentioned above, we use the output variable as a class variable. The selected subset of features are used as input variables. We use the binary variable "M3FEEDAT" (Any breast feeding at 3 months postpartum") as a class variable. It takes two possible values; "Breast Feeding" and "Not Breast Feeding".

We use the commonly used C4.5 (Quinlan 1993) software to train and validate our model. C4.5 creates pruned and un-pruned decision trees based on the training data set. The decision trees are then used to predict the class of test data. The attractiveness of the decision model is judged by its prediction accuracy on the test data.

3.2 Generalised Linear Model

We compare decision tree model with the commonly used statistical approach logistic regression. In logistic regression, the dependent variable is a logit, which is the natural log of the odds.

$$logit(P) = \ln(\frac{P}{1-P}) = a + b\mathbf{X} \qquad (9)$$

The log odds (logit) is assumed to be linearly related to $\mathbf{X}$, where $\mathbf{X}$ is short notation for all input variables used in the model. We use the freely available statistical package R to perform the generalised linear modeling. What makes our study different from traditional statistical approaches is that we do not use all data records in our training. Instead, we use data mining methodology to test the accuracy of the generalised linear model. As mentioned previously, we divide the data set into training and test data sets. The training data set is used to train all the parameters in Equation 9. The test data set tests the generalisability of the model.

4 Risk Pattern Mining (RPM)

Risk pattern mining (Li et al. (2005), Gu et al. (2003)) deals with data consisting of two unbalanced classes. The minor class (usually the high risk class) is the primary study group. Unlike the classification method, risk pattern mining is not used to build classifiers, but to generate an optimal risk pattern set. A pattern is excluded from the optimal risk pattern set when its relative risk is lower than a simpler pattern with fewer variables in it. Therefore, the optimal pattern set does not give highly accurate prediction, but indicates all interesting cohorts that are more likely to belong to the high risk class. In our risk pattern mining approach, the targeted class of the study is the mothers who do not breast feed their babies. RPM enables us to identify mothers with certain characteristics, leading to a high risk of not breast feeding their babies. Bi-variate analysis certainly helps to identify the single characteristic, which may lead to the baby feeding method decision. This has been done extensively by an earlier study (Hegney et al. 2003). The risk pattern mining method can find the factors associated with not only a single variable alone but also a combination of factors. The combination of these factors leads to the decision of not breast feeding their babies. Therefore, the risk pattern mining approach may complement the statistical approach. The risk pattern mining method also has the advantage of identifying the group automatically from available data alone. It does not assume any prior knowledge on what may pose a high risk. It allows the data to speak for themselves. The main interest of this study is in finding groups of higher occurrences of mothers not breast feeding their babies than the average. These mothers are classified as class C (target class). The other class is called class $\overline{C}$. We define the support of A, $supp(A)$, as the number of subjects satisfying the condition A. We can represent the population by contingency table 1.

The Risk Ratio (RR) values for Class C is defined as ratio of cross products of terms in Table 1 or expressed as follows.

$$RR(A \to C) = \frac{supp(A \to C)}{supp(A)} / \frac{supp(\overline{A} \to C)}{supp(\overline{A})} \qquad (10)$$

Table 1: Contingency Table

	C	$\overline{C}$	**Total**
A	$supp(A \rightarrow C)$	$supp(A \rightarrow \overline{C})$	$supp(A)$
$\overline{A}$	$supp(\overline{A} \rightarrow C)$	$supp(\overline{A} \rightarrow \overline{C})$	$supp(\overline{A})$
Total	$supp(C)$	$supp(\overline{C})$	

RR specifies how many times more likely the subjects satisfying pattern A and belonging to the target class are than others. Its 95% Confidence Interval (CI) can be calculated by Equation 11 (Fleiss (1981)).

$$I(A \rightarrow C) = RR(A \rightarrow C) \pm \frac{supp(A \rightarrow C)supp(\overline{A} \rightarrow \overline{C}) - supp(A \rightarrow \overline{C})supp(\overline{A} \rightarrow C))}{\sqrt{supp(\overline{C})supp(C)supp(A)supp(\overline{A})}}. \quad (11)$$

5 Results and Discussion

We use the feature "M3FEEDAT" (Type of feeding at 3 months postpartum) as the output variable. There are 53 descriptive variables and 498 subjects. The purpose of the data mining is to decide the factors influencing the decision on feeding method.

5.1 Feature Selection and Classification

FIEBIT and Chi-square methods are applied in the feature selection step. C4.5 and logistic regression are used as classification models for data with features selected by a feature selection module. In logistic regression, the functions provided in R package is used to establish and apply the model in deciding the classification. We use *leave-one-out* as the testing method to decide the accuracy of the models. The classification accuracies on training and test data are listed in Table 2.

The results of logistic regression using all features are not available due to the restrictions of the software. From the decision tree analysis, the following branch covers 123 subjects, of which 106 are breast feeders. Only 7 are not breast feeders.

```
Q3.9.11 (Breast Feeding is convenient) =Yes
Q3.9.1  (My mother breastfed)          =Yes
Q3.9.14 (I do not want to have to mix
        formula/sterilize bottles)     =Yes
```

It is likely that these factors are important in deciding the feeding method.

The following observations can be made out of the results of various feature selection and classification methods.

- Classification accuracy is improved by using a kind of feature selection as a data preprocessing step. The use of the whole feature set leads to high classification accuracy on training data, but low accuracy on test data. This implies the over-fitting problem associated with irrelevant or redundant features included.
- Feature subsets selected by the information theory based method lead to a bit higher classification accuracy on test data than that selected by the Chi-square method. More experiments, say using an n-fold cross-validation, may help draw a sound conclusion, which are left as future work.
- The highest classification accuracy on test data is achieved by using FIEBIT as the feature selection method followed by C4.5 as the classifier. The generalisation accuracy is 77.91%. The accuracy on training data is only slightly higher (80.94%) than that of the test data. This indicates that the over training problem is largely overcome by selecting a good feature subset.
- A decision tree can be used as a feature selection method since it can be applied to the original data set. However, features selected by decision trees are not as good as FIEBIT and ChiSQ on this data set since the accuracy on the test data set is lower. More experiments may help draw a sound conclusion. For example, we may impose some sort of regularisation on the classification models (ridge regression or something similar in the logistic regression, or more aggressive pruning of the decision tree) which would be likely to lead to nice results too. This is left as a future work direction.

5.2 Risk Pattern Mining

The rules identified by the algorithm can find the cohort with high risk ratio of not breast feeding. We use 8 features selected by FIEBIT in the following example.

Results of simple rules with one or two variables from Risk Pattern Mining (RPM) can be also found by statistical analysis, and both findings agree. However, in a statistical approach it is difficult to explore interactions of three or more variables systematically whereas RPM method can. The following rules are some cohorts that are overlooked by the previous statistical analysis.

```
Rule 1
 Q3.9.13  (I enjoy breast feeding)  = No
 Q3.9.14  (I do not want to have to
           mix formula/sterilize
           bottles)                 = No
 FEEDDECI (At what time did the
           mother make the decision
           about feeding method)    = Late
                    in pregnancy/after baby
                        born/still deciding
     Cohort size = 11
 Contingency table
              not breast  breast
              feeding     feeding
 pattern         10         1
 non-pattern    119         368
```

Length = 3, $RR = 3.72 \pm 0.22$

There are a total of 11 subjects in the cohort, 10 of them are not breast feeding. The subjects in the cohort are 3.72 times more likely not to be breast feeding than the subjects not satisfying the rule.

```
 Rule 2
 Q3.9.11  (Breastfeeding is more
          convenient)             = Yes
 Q3.9.1   (My mother breastfed)   = No
 MOAGEREC (Mother's age)          = Under 25
 Q3.9.13  (I enjoy breast
           feeding)               = No
 Q3.9.14  (I do not want to have
          to mix formula/sterilize
          bottles)                = No
 Q3.8.1  (general anaesthetic)    = No
     Cohort size = 8
 Contingency table
            not breast   breast
            feeding      feeding
 pattern         7          1
 non-pattern   122         368
```

Length = 6, $RR = 3.51 \pm 0.18$

There are total 8 subjects in the cohort, 7 of them are not breast feeding. The subjects in the cohort are

Table 2: Prediction accuracies of various feature selection and classification methods

Feature Selection	Number of Features	Classification Method	Accuracy(%) Training	Accuracy(%) Testing
FIEBIT	8	C4.5	80.94	77.91
FIEBIT	8	LogReg	77.70	76.49
None	53	C4.5	87.70	71.66
None	53	LogReg	NA	NA
ChiSQ	11	C4.5	81.97	75.05
ChiSQ	11	LogReg	77.58	76.20

3.51 times more likely not to be breast feeding than the the subjects not satisfying the rule.

```
Rule 3
Q3.9.11  (Breastfeeding is more
          convenient)            = no
MOAGEREC (Mother's age)          = 25-30
Q3.9.14  (I do not want to have
          to mix formula/sterilize
           bottles)              = no
FEEDDECI (At what time did the
          mother make the
          decision about feeding
          method)                = Before
                                  pregnancy
   Cohort size = 18
Contingency table
                 not breast  breast
                 feeding     feeding
pattern            14          4
non-pattern       115         365
```

Length $= 4$, $RR = 3.25 \pm 0.23$

There are total 18 subjects in the cohort, 14 of them are not breast feeding. The subjects in the cohort are 3.25 times more likely not to be breast feeding than the subjects not satisfying the rule.

These results show that the risk pattern mining method enables us to identify a cohort of mothers who are more likely not to breast feed their babies. This will enable us to conduct a focused education program on a targeted group of mothers to increase the rate of breast feeding. For example, based on rule 2, the cohort identified are 3.51 times more likely not to breast feed their babies. We should therefore target the young mothers (under 25) whose mother did not breast feed their babies. Specific information can be provided to address the concerns of each cohort.

6 Conclusions

Data mining aims at extracting novel, valuable and actionable knowledge from a database. In this study, we attempt to use data mining techniques on a survey data set, rather than traditional statistical analyses. We claim that the application of data mining methods may extract knowledge that statistical analysis may find difficult to identify, say, logistic regression for all the variables. As a complementary approach to statistical analysis, data mining methods can be used as a viable tool in survey data analysis.

1. Feature selection is an important step in data preprocessing as it enables irrelevant and redundant features to be eliminated. It substantially reduced the data dimensions for modelling. It not only makes the modelling procedure more efficient but also improves the accuracy of the model developed by data mining or statistical methods.

2. The feature selection process followed by a classification module is able to build a classifier which classifies the data on whether or not the breast feeding method is selected automatically with reasonable accuracy. The classification accuracy using properly selected feature subsets reached over 75% on test data. This implies that when we apply the model to a new subject (a mother), we can predict her likelihood of breast feeding or not with reasonable confidence. We can therefore take proper measures to provide education surrounding this prediction.

3. Risk pattern mining is able to discover a number of rules which identify groups of patients with a high risk of not beast feeding. The knowledge discovered by risk pattern mining may be used by doctors or nurses to assess the risk associated with a new mother based on her characteristics. A real world application is expectable by using the information discovered by risk pattern mining. For example, we may use a graphic interface McAullay et al. (2005), Chen et al. (2005), to enable medical practitioners to gain knowledge effectively based on the risk patterns mined, which is left as future work. A proper targeted education or other measures may then be taken.

4. The application of various data mining methods to breast feeding survey data helps to discover knowledge and the understanding of the decisions made by mothers. It complements and enhances the statistical analysis. Statistical analysis is powerful in hypothesis testing. Data mining methods, on the other hand, help in hypothesis generating Jin et al. (2006) It generates decision trees, rule sets etc. automatically without assuming any prior knowledge. The knowledge discovered becomes more comprehensive when both statistical analysis and data mining methods are applied to the same data set.

References

Chen, J., He, H., Li, J., Jin, H., McAullay, D., Williams, G., Sparks, R. & Kelman, C. (2005), Representing association classification rules mined from health data, in *Proceedings of 9th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES2005)*, Melbourne, Australia, pp. 1225–1231.

Cover, T. M. & Thomas., J. A. (1991), *Elements of Information Theory*, Wiley-Interscience.

Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions*, Wiley.

Fleuret, F. (2004), 'Fast binary feature selection with conditional mutual information', *Journal of Machine Learning Research* **5**, 1531–1555.

Gu, L., Li, J., He, H., Williams, G., Hawkins, S. & Kelman, C. (2003), Association rule discovery with unbalanced class, in *Proceedings of the 16th Australian Joint Conference on Artificial Intelligence (AI03), Lecture Notes in Artificial Intelligence*, Perth, Western Australia, pp. 221–232.

He, H., Jin, H. & Chen, J. (2005), Automatic feature selection for classification of health data, in *Proceedings of The 18th Australian Joint Conference on Artificial Intelligence (AI2005)*, Sydney, Australia, pp. 910–913.

Hegney, D., Fallon, T., O'Brien, M., Plank, A., Doolan, J., Brodribb, W., Hennessy, J., Laurent, K. & Baker, S. (2003), *The Toowoomba Infant Feeding Support Service Project: Report on Phase 1 A Longitudinal Needs Analysis of Breastfeeding Behaviours and Supports in the Toowoomba Region.*

Jin, H., Chen, J., Kelman, C., He, H., McAullay, D. & O'Keefe, C. M. (2006), Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases, in *PAKDD'06*, pp. 867–876.

Jin, H.-D., Shum, W., Leung, K.-S. & Wong, M.-L. (2004), 'Expanding self-organizing map for data visualization and cluster analysis', *Information Sciences* **163**, 157–173.

Jin, H., Wong, M.-L. & Leung, K.-S. (2005), 'Scalable model-based clustering for large databases based on data summarization', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(11), 1710–1719.

Kohavi, R. & John, G. (1997), 'Wrappers for feature selection', *Artificial Intelligence* pp. 273–324.

Kramer, M. S. & Kakuma, R. (2003), *Optimal duration of exclusive breastfeeding*, The Cochrane Library.

Li, J., Fu, A. W.-C., He, H., Chen, J., Jin, H., McAullay, D., Williams, G., Sparks, R. & Kelman, C. (2005), Mining risk patterns in medical data, in *Proceedings of KDD'05*, pp. 770–775.

McAullay, D., Williams, G., Chen, J., Jin, H., He, H., Sparks, R. & Kelman, C. (2005), A delivery framework for health data mining and analytics, *in* V. Estivill-Castro, ed., *Twenty-Eighth Australasian Computer Science Conference (ACSC2005)*, Vol. 38 of *CRPIT*, ACS, Newcastle, Australia, pp. 381–390.

Quinlan, J. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Riodan, J. M. (1997), 'Commentary. the cost of not breastfeeding: a commentary.', *Journal of Human Lactation* **13**(2), 93–97.

Shannon., C. E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**, 379–423,623–656.

Smith, J. (2001), *Mothers milk, money and markets*, Ann Congress Perinatal Society Australia and New Zealand.

Smith, J. P., Thompson, J. F. & Ellwood, D. A. (2002), 'Hospital system costs of artificial infant feeding: Estimates for the australian capital territory', *Australian and New Zealand Journal of Public Health* **26**(6), 543–551.

Wang, G., Lochovsky, F. H. & Yang, Q. (2004), Feature selection with conditional mutual information maxmin in text categorization, in *Proceedings of CIKM'04*, Washington, US, pp. 8–13.

WHO (2001), *The optimal duration of exclusive breastfeeding*, World Health Organization.

Yang, Y. & Pedersen, J. O. (1997), A comparative study on feature selection in text categorization, in *Proceedings of International Conference on Machine Learning*, Nashville, TN, USA.

Yu, L. & Liu, H. (2004), Redundancy based feature selection for microarray data, in *Proceedings of KDD'04*, ACM Press, New York, NY, USA, pp. 737–742.

Using a kernel–based approach to visualize integrated Chronic Fatigue Syndrome datasets

Ahmad Al–Oqaily **Paul J. Kennedy**

Faculty of IT,
University of Technology, Sydney,
PO Box 123, Broadway, NSW 2007,
AUSTRALIA,
Email: aaoqaily@it.uts.edu.au

Abstract

We describe the use of a kernel–based approach using the Laplacian matrix to visualize an integrated Chronic Fatigue Syndrome dataset comprising symptom and fatigue questionnaire and patient classification data, complete blood evaluation data and patient gene expression profiles. We present visualizations of the individual and integrated datasets with the linear and Gaussian kernel functions. An efficient approach inspired by computational linguistics for constructing a linear kernel matrix for the gene expression data is described. Visualizations of the questionnaire data show a cluster of non–fatigued individuals distinct from those suffering from Chronic Fatigue Syndrome that supports the fact that diagnosis is generally made using this kind of data. Clusters unrelated to patient classes were found in the gene expression data. Structure from the gene expression dataset dominated visualizations of integrated datasets that included gene expression data.
Keywords: kernel–based visualization, Laplacian matrix, data integration, biomedical datasets.

1 Introduction

Chronic Fatigue Syndrome (CFS) (Afari & Buchwald 2003) is an illness with a primary symptom of debilitating fatigue over a six month period. Currently diagnosis of CFS is generally made by clinical assessment of symptoms using a number of questionnaires or surveys measuring functional impairment, quantifiable measurements of fatigue and occurrence, duration and severity of the symptoms (Reeves et al. 2005). One goal of current research is to derive a definition of the syndrome, which goes beyond a clinical assessment of symptoms to an empirical diagnosis founded on measurements such a gene expression profiles. The motivation for this kind of research is to gain a clearer understanding of the illness and to find empirical guidelines for its diagnosis.

The question we examine in this paper is whether data visualization methods, specifically a method based on the eigenvectors of the Laplacian matrix (Shawe-Taylor & Cristianini 2004), can be used to discover patterns in biomedical datasets associated with CFS patients. Also, because there are several datasets from different sources, we are interested in creating integrated datasets and visualizing the combined data.

In the biomedical domain it is commonplace for data to be generated by high–throughput technology. One example is microarray technology (Baldi & Hatfield 2002) which generates gene expression profiles that simultaneously measure the level of expression of thousands of genes in biological samples. In general, biomedical datasets derived from high–throughput technology are described by a small number of samples (patients) and a large number of features or attributes (i.e. genes) per sample. This results in what is often referred to the 'curse of dimensionality' which makes building classifiers troublesome. For analysis of this type of data, algorithms are often applied to select features from the data to reduce its dimensionality. As a first step towards working to building classifiers for this kind of data, we initially just visualize it and look for patterns in the dataset.

In our case the data is represented by different datasets and kinds of measurements including questionnaires, complete blood evaluations and gene expression data so we also create integrated datasets by combining the individual datasets.

We apply a kernel based method to visualize the individual and combined datasets. The method (Shawe-Taylor & Cristianini 2004) we use is similar to kernel PCA (kPCA). Kernel PCA is kernel–based extension of the well known Principal Component Analysis (PCA) algorithm (Haykin 1999) and is used to reduce the dimensionality of datasets in a principled way. PCA forms a new dataset where each attribute is a linear combination of attributes from the original dataset. When the dataset is reduced to two or three dimensions it can be graphed and this allows PCA and kPCA to be used to visualize the data. Kernel PCA differs from PCA in that the data is transformed using a kernel function before the new attributes are derived. The benefit of doing this is that the attributes are not limited to being linear combinations of the original attributes and can therefore "see" nonlinear relationships in the original data. Also, using a kernel function allows visualization of non–vectorial data. We use a method based on kPCA but it differs in that whilst kPCA uses eigenvectors of the kernel matrix, the method we employ uses eigenvectors of a slightly different matrix: the Laplacian matrix. A similar approach to clustering of text with Laplacian matrices in done in (Li, Ng & Lim 2004). That approach, however, did not apply kernel functions to the data.

In section 2 we describe the datasets used for this study and the preprocessing steps applied to the data. Next, in section 3 we describe in detail the data mining and visualization approach used to identify potential patterns in the CFS datasets. In section 4 we present and results of applying the kernel based visualization method to the datasets. In section 5 we discuss these results and describe future directions for

the research. Finally, in section 6 we summarize the paper.

2 Data

In this section we describe in detail the datasets used and the preprocessing steps applied.

We used publicly available data generated as the 2006 Critical Assessment of Microarray Data Analysis (CAMDA 2006) competition datasets (CDC Chronic Fatigue Syndrome Research Group 2005). The data consists of separate datasets for patients linked by a patient identifier ("ABTID"). There were datasets with (i) survey results from fatigue and symptom questionnaires; (ii) complete blood evaluations; (iii) gene expression profiles; (iv) single nucleotide polymorphism (SNP) data; and (v) proteomics data.

The sources of data used in this study are significant because they cover the full biological spectrum from genotype through to phenotype. That is, ranging from data concerning genes through to data about their expression in the body in the form of proteins. The researchers who generated the data for the CAMDA competition hypothesize that the gene expression profile data will allow identification of "prognostic indicators" or biomarkers for diagnosis of CFS (National Center for Infectious Diseases 2005). As mentioned above, CFS is currently diagnosed using symptom questionnaires, so identification of biomarkers is potentially very significant. The analysis in this paper explores this hypothesis.

The SNP and proteomics datasets were not analyzed in this study. Analysis of the SNP and proteomics data is straightforward with our methods but will be analyzed in future studies. The SNP and proteomics data will not be mentioned further in this paper.

Data was integrated between the three datasets used in the study simply by linking of records using the patient identifier i.e. ABTID. Not all patients have data in each of these datasets. In cases where data was not available across the integrated dataset (e.g. when linking the gene expression and blood work datasets) we omitted the patients affected. This was acceptable in our situation because, as individuals in the gene expression data are a subset of patients in the clinical datasets, there was considerable overlap between the datasets and not many patients were lost.

Patients were classified into a number of categories which we grouped into three different classes: (i) those classified by physicians as suffering from Chronic Fatigue Syndrome (CFS); (ii) those classified as suffering from symptoms associated with CFS but with insufficient severity to be classified as CFS (IFS); and (iii) non fatigued individuals (NF).

In the following subsections we describe each dataset in more detail.

2.0.1 Clinical Datasets: Illness Classification and Complete Blood Work

The clinical data comprised two datasets: (i) an Illness Classification and Symptoms dataset consisting of information about patient symptoms and fatigue and (ii) evaluation of blood samples taken from patients. Each individual indicated with a patient identifier ("ABTID") has a record in both of these datasets. There were 139 CFS/ISF patients and 73 NF individuals.

The "Illness Classification SF36 MFI and Symptoms" (illness) dataset is generated based on survey results for the above mentioned patients (CDC Chronic Fatigue Syndrome Research Group 2005). The dataset includes (i) attributes that describe the general information of the patient like sex, date of birth, race, and ethnicity; (ii) the Medical Outcomes Survey Short Form–36 (SF–36) (Ware & Sherbourne 1992), as a measurement criteria for functional impairment, such as physical function, role emotional, and mental health; (iii) Multidimensional Fatigue Inventory (MFI) (Smets, Garssen, Bonke & DeHaes 1995), to obtain reproducible quantifiable measures of fatigue including "General Fatigue", "Physical Fatigue", "Active Reduction" and "Mental Fatigue"; and (iv) the CDC Symptom Inventory (Wagner, Nisenbaum, Heim, Jones, Unger & Reeves 2005) to document the occurrence, duration and severity of the symptom complex including attributes such as "Sore Throat", "Tender Nodes", "Muscle Pain, and Depression". The "Complete Blood Evaluation" dataset (blood) contained measurements of components of individual's blood as well as flags for when these measures were out of normal range.

2.0.2 Gene Expression Datasets

Microarray technology allows the high throughput analysis of global gene expression within a biological specimen. Gene expression measurements are made simultaneously for many thousands of genes. The gene expression profile of diseased cells may reflect the unique genetic alterations present and has been shown to be predictive of clinical and biological characteristics of illness for many diseases (Baldi & Hatfield 2002). A major issue in these data is the unreliable variance estimation, complicated by the intensity–dependent technology–specific variance (Weng, Dai, Zhan, He, Stepaniants & Bassett 2006). Below we describe our approach to normalizing this data. The gene expression profiles used in this study measured the level of expression of genes in blood samples from patients.

Data collected was for a subset of individuals: 118 CFS/ISF patients and 53 NF individuals. Generally there is one gene expression profile for each of these patients. A few individuals had more than one sample. The gene expression profile for a sample contains data for around ten thousand genes and data for each gene comprised around 15 attributes.

2.0.3 Preprocessing of the Clinical datasets

Most of the attributes in the questionnaire and blood evaluation datasets were used without much preprocessing.

Some attributes of the "illness" dataset, the clinical dataset containing the patient's answers to the illness questionnaires, are omitted because they are (i) skewed with almost all individuals having the same attribute value, (ii) not deemed useful for the data mining effort, or (iii) are calculated by the original researchers and would bias our efforts. The attributes concerned are "DOB", "intake classific", "cluster", "onset", "yrs ill", "race" and "ethnic". The dependent variable "Empiric" is used as the patient class and patient subtypes are combined to make three classes CFS, ISF and NF.

In the blood evaluation dataset, we add a copy of the "Empiric" attribute so that the dataset has the patient class.

All attributes of the clinical datasets apart from the patient class "Empiric" were converted to numeric values as the kernel visualisation method employed requires strictly numeric data. Binary attributes were converted to -1 and +1 for "false" and "true" respectively. Similarly, in the questionnaire dataset, categorical data values such as "mild", "moderate" and "severe" were coded to 1, 2 and 3 respectively.

Missing values were universally converted to 0. This is consistent with the coding scheme used for binary and categorical attributes. Coding of missing values to 0 is appropriate for kernel based schemes because the value 0 does not adversely effect the dot products used to build kernels. Missing values were very infrequent in the dataset and we believe that this simple approach to dealing with them is effective.

Data items in both clinical datasets were centred by subtracting the mean and attributes were normalized.

2.0.4 Preprocessing of the Gene Expression data

Data for each gene comprised a spot label (the name of the gene) and several measurements describing the level of expression of the gene as well as quality control indications of the expression measurement. We extracted the "Spot labels", "ARM Dens — Levels", "MAD — Levels" and "SD — Levels" fields. We discarded the other fields. The three statistical measures of gene expression are normalized over all arrays (samples) and patients by multiplying values with the average value of every gene over all arrays divided by the average value of every gene over the individual array.

Data for each sample was in a separate text file with filename indicative of the identifier of the patient sampled. The data for all samples was concatenated into a single gene expression file and with the patient identifier as the initial field. Additionally, the patient class "Empiric" was associated with the gene expression data through linking with the patient identifier although it was not added as an attribute to the large concatenated file.

3 Approach

The approach we have used is to visualize patients in the datasets with a kernel–based visualization method and to look for interesting features in the visualization. As shown in Fig. 1, we visualize each of the datasets (i.e. illness, blood and gene expression) in isolation, in integrated pairs (i.e. illness and blood, illness and gene, and blood and gene) and finally the integrated triplet.

As we mentioned above, the integration between datasets is done using the patient identifier. The patient class (CFS, ISF or NF) is excluded from the attributes used in the visualization because this is what we want to see in the visualization. However, each patient class is plotted with a different symbol and color in the visualization. If patients of the same class are grouped together in a visualization, this lends support to the claim that there is a relationship in the dataset. This potential relationship can be investigated in future work.

3.1 Kernel–based Visualization Approach

The approach we use is related to the kernel–based extension to Principal Component Analysis (PCA) (Haykin 1999). Principal Component Analysis is a well established method that transforms a dataset into a different coordinate system. The transformation is essentially a rotation of the dataset. The coordinates of the transformed dataset (called principal components) are orthogonal linear combinations of the original coordinates. The principal components are ordered in descending order by the amount of variance they explain in the data. Often much of the variance in the dataset can be explained by many fewer coordinates than in the original dataset (e.g. less than

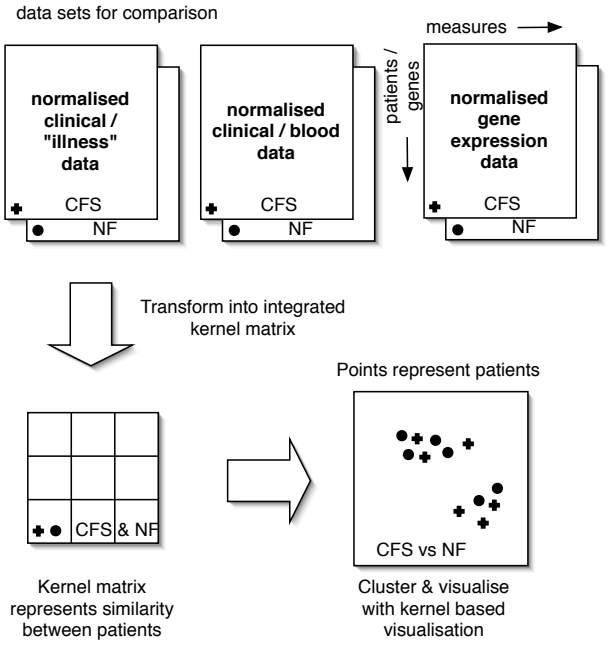


Figure 1: Approach used to visualize the individual and integrated CFS datasets.

ten). This fact means that PCA is often used for compression of data or feature selection. It also facilitates visualization of datasets by plotting the first two or three principal components of the dataset. However, as principal components are linear combinations of the original dataset, PCA has the limitation that it can only model linear relationships in the data.

There have been several approaches to extending PCA to handle nonlinear relationships. One approach is kernel PCA (kPCA) ((Müller, Mika, Rätsch, Tsuda & Schölkopf 2001), (Haykin 1999) or (Shawe-Taylor & Cristianini 2004)) which transforms the dataset $\mathbf{X}$ into a feature space using a kernel function κ before the PCA is done. Kernel PCA returns the principal components of data items in the feature space. It takes as input a Gram kernel matrix $\mathbf{K}$ which is a representation of the original dataset transformed with the kernel function. Each element $\mathbf{K}_{ij}$ of the kernel matrix is defined as

$$\mathbf{K}_{ij} = \kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{1}$$

where x_i and x_j are the data items, $\phi(x_i)$ is the transformation of x_i into the "feature" space and $\langle \cdot, \cdot \rangle$ is the dot product operator. Generally it is not necessary to compute $\phi(x_i)$ explicitly. Instead, $\mathbf{K}$ is computed directly from the dataset. This is called the "kernel trick" and it means that the feature space can be very large without making generation of $\mathbf{K}$ inefficient. It also means that non–vectorial data types can be handled using special kernels such as string kernels (e.g. (Leslie, Kuang & Eskin 2004)).

Two commonly used kernel functions are the *linear kernel* and the *Gaussian kernel*. The linear kernel is defined as

$$\kappa(x_i, x_j) = \langle x_i, x_j \rangle \tag{2}$$

and is simply the dot product of the two data items. The Gaussian kernel explicitly considers the distance between data items and is defined as

$$\kappa(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{3}$$

where σ is a control parameter.

Finding the principal components in kPCA amounts to deriving the eigenvalues and eigenvectors of the kernel matrix $\mathbf{K}$. Shawe–Taylor and Cristianini describe another technique in (2004) that they say more explicitly controls the correlation between the points in the original and feature spaces. The technique is essentially the same as kPCA except that it uses (non–zero) eigenvalues and matching eigenvectors of the Laplacian matrix $\mathbf{L}(\mathbf{K})$ of the kernel matrix instead of the kernel matrix. The Laplacian matrix is defined as

$$\mathbf{L}(\mathbf{K}) = \mathbf{D} - \mathbf{K} \tag{4}$$

where $\mathbf{D}$ is the diagonal matrix with entries

$$\mathbf{D}_{ii} = \sum_{j=1}^{l} \mathbf{K}_{ij} \tag{5}$$

where l is the size of the kernel matrix.

In this study, we employ this last method using the Laplacian matrix to identify the first three principal components of datasets for visualization. We examine the data using both the linear kernel in (2) and the Gaussian kernel in (3).

3.2 Applying Kernel–based Visualization to the CFS data

Application of the kernel–based visualization scheme to our datasets was, in general, fairly straightforward. As described above, the method requires a kernel matrix representing the dataset to be visualized.

We used the linear kernel and the Gaussian kernel to make kernel matrices for the clinical datasets (illness and blood) both individually and in the integrated pair. Due to its large size, application of a kernel function to the gene expression dataset required a special approach similar to that used in computational linguistics.

Each row of the gene expression dataset represents an individual gene measurement for a particular microarray (for each patient). The straightforward approach of calculating the linear kernel matrix is to concatenate the rows of the gene expression dataset into a matrix consisting of one row for each array with a set of attribute values for each spot label ("ARM Dens - Levels", "MAD - Levels" and "SD - Levels") then to calculate the linear kernel by multiplying the matrix with its transpose.

Clearly this approach is impractical in our situation because of the large number of genes on each of the arrays. A more efficient approach, motivated by computational linguistics (see, for example, the description of generating the vector–space kernel in (Shawe-Taylor & Cristianini 2004)), for direct computation of the linear kernel matrix from the gene expression data is more appropriate.

The kernel value for two samples (i.e. microarrays) is calculated from sorted lists of genes (spot labels) associated with each array. The kernel value is calculated as the sum of the product of the attribute values for genes matching in both lists.

Computation of the linear kernel for the integrated datasets (of gene expression combined with illness and/or blood) is trivial. The linear kernel for the integrated dataset is simply the sum of the linear kernel matrices for the individual datasets.

Unfortunately this simple method of computing the linear kernel for the integrated datasets does not apply for the Gaussian kernel. In this case it is necessary to have the entire data vector. Since we never compute the data vector for the gene expression data (it is too large), we are unable to easily use the Gaussian kernel for the gene expression dataset or any integrated dataset containing the gene expression dataset.

4 Results

4.1 Visualization of individual datasets

We visualise each dataset individually. Figure 2 shows visualizations of the illness dataset. That is, the dataset containing patient's answers to the fatigue and symptom questionnaires. Figure 2a shows a visualization of the dataset with the linear kernel and Fig. 2b shows a visualization with the Gaussian kernel. The σ parameter of the Gaussian kernel was set to 100 in this case. We examined other settings for this parameter, but the value 100 gave the clearest images. As can be seen clearly in both figures, the NF individuals cluster together on the left hand side, with the ISF in the middle and the CFS patients on the right. These pleasing results are expected, as the data in this dataset is used to make the patient classifications. For the illness dataset, there is not a great deal of difference between the visualization with the linear kernel and with the Gaussian kernel.

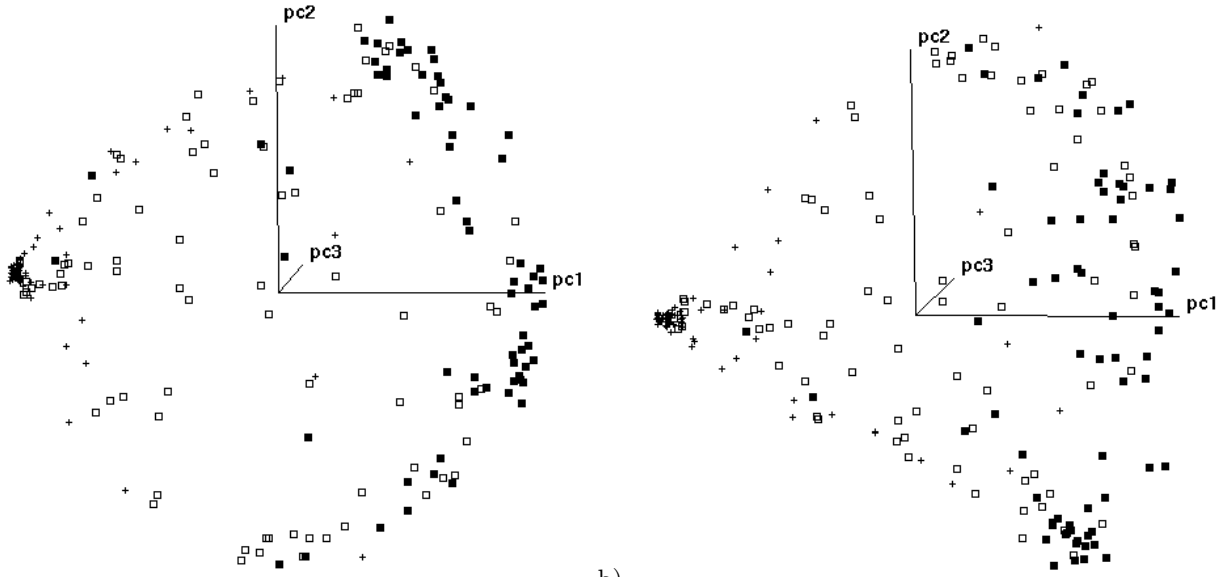
We investigated some of the patients marked as ISF that clustered with the NF individuals on the left. Patients in the dataset are actually classified with two different schemes. When we examine some of the patients that appeared to cluster incorrectly, they are classified correctly using the other scheme.

Figure 3 shows a visualization with the linear kernel of the complete blood evaluation dataset. In Fig. 4 we present visualizations of the same dataset using the Gaussian kernel. As with the illness dataset above, we tried different values for the σ parameter of the Gaussian kernel but found that the value of 100 produced the best images. Fig. 4a plots the first two principal components of the projection of the data in feature space and Fig. 4b shows the three–dimensional view. In both the linear and Gaussian visualizations of the complete blood evaluation dataset there are no clear groupings of patient classes into distinct or separable regions. This suggests to us that there are no simple "biomarkers" in the blood evaluation dataset associated with CFS. In Fig. 4a there are some small groupings of CF patients particularly five clustered in the center of the diagram that may warrant further investigation.

Finally, in Fig. 5 we present visualizations using the linear kernel for the gene expression dataset. Figure 5a shows the two–dimensional plot with the first two principal components and Fig. 5b gives the three–dimensional view. Three fairly distinct clusters of patients can be seen. However, these are not naturally associated with the classes of patient, although the top right grouping in Fig. 5a contains the least number of CF patients compared the NF individuals. The same clusters are visible in the three–dimensional plot. Although not clearly shown in the diagram (Fig. 5b), these clusters are mostly embedded in a plane with only a few data points extending further along the pc3 axis. Although the clusters do not seem to be associated with the classes of patient, it would be interesting to further investigate relationships within the clusters.

4.2 Visualization of Integrated datasets

In this section we investigate patterns in the integrated datasets. First we look at the pairs of datasets. Then we examine the combination of all three datasets. We visualise the integrated datasets



a) b)

Figure 2: Kernel visualization of illness classification (questionnaire) dataset. a) linear kernel. b) Gaussian kernel with $\sigma = 100$. Axes in both graphs are the first three principal components. + = NF individuals, □ = ISF patients and ■ = CF patients.

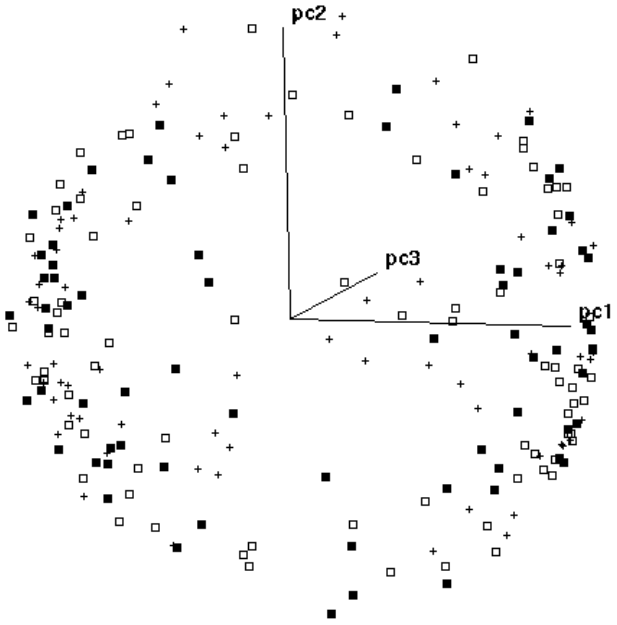


Figure 3: Kernel visualization of the complete blood evaluation dataset using the linear kernel. Axes are the first three principal components. + = NF individuals, □ = ISF patients and ■ = CF patients.

with only the linear kernel rather than both the linear and Gaussian kernel functions because (i) the use of the Gaussian kernel function did not add much to the visualizations in the previous section and (ii) because the Gaussian kernel was not applied to the gene expression dataset for the reasons described above.

Figure 6 shows visualizations of the integrated illness and blood evaluation datasets. As before, the visualization on the left (Fig. 6a) shows the two–dimensional view and on the right (Fig. 6b) the three dimensional view. It is interesting to compare these images with the visualizations for the individual datasets to see the effect of integrating the data (i.e. with Fig. 2a and Fig. 3). Recall that the NF individuals were tightly clustered in the visualization of the illness dataset but not in the blood evaluation dataset. In the integrated dataset the visualization shows the NF patients again clustered. However, instead of the compact clustering of the illness dataset, the NF individuals are now clustered in a line. In the three–dimensional visualization most of the data points are on the surface of a sphere and the NF individuals appear as line of "longitude".

Next we integrate the blood and gene expression datasets and visualise with the linear kernel function in Fig. 7. Comparing with the graphs of the individual datasets in Figures 3 and 5 it can be seen that the structure of the gene expression dataset dominates. The diagram of Fig. 7a seems to be the reflection across the horizontal of Fig. 5a. Similarly, the visualization of the integrated illness and gene expression datasets in Fig. 8 are essentially the same with the structure completely controlled by the gene expression dataset. We speculate that the reason that the gene expression dataset dominates the structure is that it contains many more attributes than either the illness or blood evaluation datasets.

Finally, in Fig. 9 we present the linear kernel visualization of the integrated triplet of datasets. Again, the situation is the same as above and the gene expression dataset dominates the visualization. There are again three fairly distinct clusters that are not naturally associated with the patient classes.

5 Discussion and Future Work

The kernel–based visualization approach allows us to integrate datasets in a straightforward way, particularly if we are content to limit ourselves to the linear kernel. This limitation, at least in the context of our study, does not seem to be onerous because the differences between visualizations of the individual blood and illness datasets with the linear and Gaussian kernel functions seemed to be relatively unimportant.

Being able to visualise integrated datasets in this way supports a "constructionist" approach to data analysis where we look for "global" patterns in the integrated dataset. The opposite method, a "reductionist" approach, looks for "local" patterns over subsets of attributes in individual datasets. An example of the latter approach in the context of this paper is

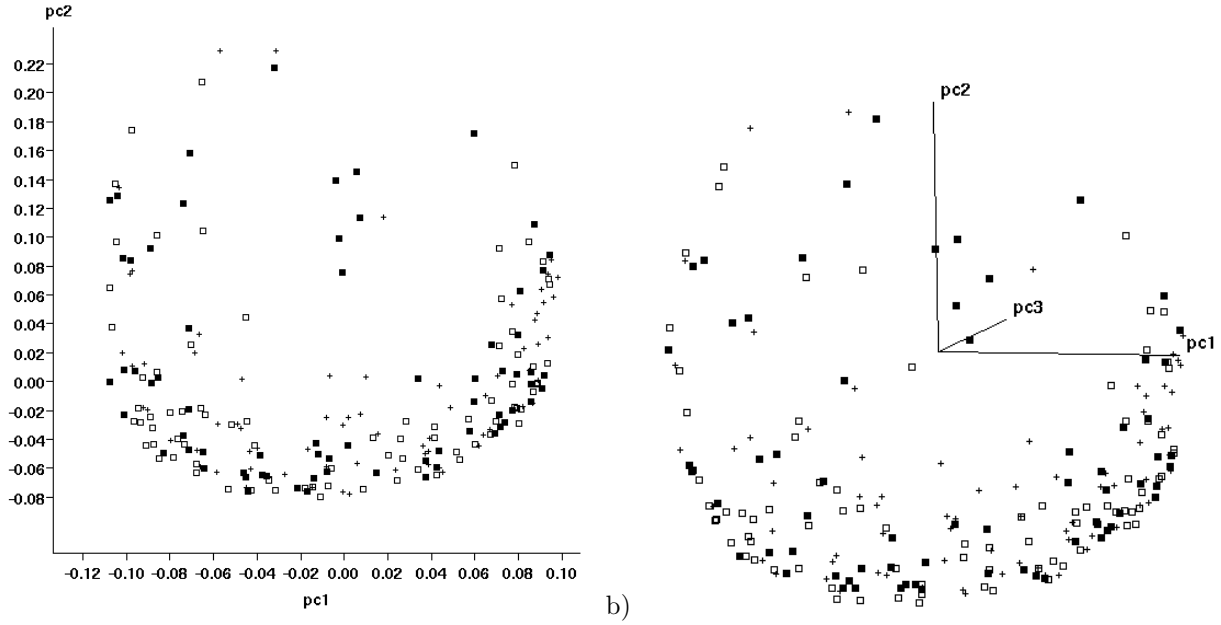


Figure 4: Kernel visualizations of complete blood evaluation dataset using the Gaussian kernel with $\sigma = 100$. a) Axes are the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.

the search for small sets of genes indicative of CFS. We do not necessarily advocate a "constructionist" approach over "reductionist" ones. Rather, it is better to apply both approaches which look at different ends of the problem. A combined approach would identify global patterns, which can be used to focus the search effort for local patterns.

It was heartening that the kernel–based visualization was able to distinguish patterns in the illness data as this data was used to make the CFS diagnosis. Visualizations of the other datasets did not show such clear patterns. However, the three clusters in the datasets containing gene expression data warrant further scrutiny.

The kernel–based visualization technique is a general purpose approach and can be applied to other biomedical datasets. Indeed we intend to examine the domain of acute lymphoblastic leukaemia next. Previous work exists linking the cancer to genes so we expect to see clear clusters in this domain.

Use of specialized kernels with the technique allows visualization of non–vectorial data. We essentially used this kind of approach to efficiently build a linear kernel for the gene expression data in Section 3.2.

We are also interested in the issue of the structure from the gene expression dataset dominating over the other datasets. It is important to understand why this is the case: is it the result of the structure of the integrated dataset or is it due to the relative numbers of attributes in the individual datasets? This question must be addressed before there can be more widespread use of the technique for visualization of integrated datasets. We think that the issue here is indeed the large discrepancy between the numbers of gene expression attributes compared to the number of clinical attributes. Any method, such as the one we use, that treats attributes as equally important, will be dominated by the dataset with the larger number of attributes. One approach we plan to use to overcome this problem is to apply feature selection on the datasets *before* the data integration and visualisation. This feature selection will even up the relative numbers of attributes in the different datasets.

The linear kernel used in this study is able only to identify linear relationships in the data. Use of other kernels (such as the polynomial or Gaussian kernels) allows visualization of nonlinear relationships. In this work we were restricted to use of the linear kernel because of the size of the gene expression data. Efficient calculation of other kernels for the gene expression data is another area of future investigation.

6 Conclusion

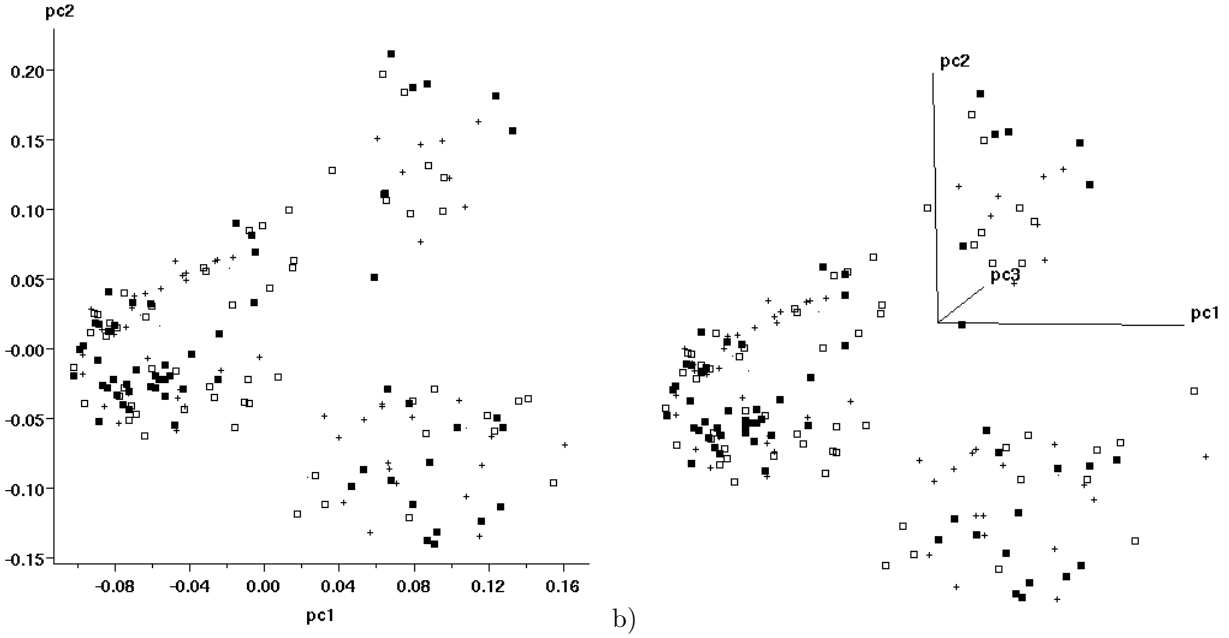
This study describes the use of a kernel–based approach using the Laplacian matrix to visualise an integrated Chronic Fatigue Syndrome dataset with symptom and fatigue questionnaire and patient classification data, complete blood evaluation data and patient gene expression profiles. Visualizations were produced for individual and integrated datasets with linear and Gaussian kernel functions. We described an efficient approach to constructing a linear kernel matrix for the gene expression data. The visualizations of the questionnaire data showed a cluster of non–fatigued individuals distinct from those suffering from Chronic Fatigue Syndrome. This observation supports the fact that diagnosis is generally made using this kind of data. The method was unable to find clusters in the other datasets that related to patient classes, although three distinct clusters were found in the gene expression data. Structure from the gene expression dataset dominated visualizations of integrated datasets that included gene expression data.

References

Afari, N. & Buchwald, D. (2003), 'Chronic fatigue syndrome: A review', *American Journal of Psychiatry* (160), 221—236.

Baldi, P. & Hatfield, G. W. (2002), *DNA Microarrays and Gene Expression: from experiments to data analysis and modeling*, Cambridge University Press.

CDC Chronic Fatigue Syndrome Research Group (2005), 'CAMDA 2006 conference contest datasets', cited; Avail-



a) b)

Figure 5: Kernel visualization of gene expression dataset using the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.

able from: http://www.camda.duke.edu/camda06/datasets/, viewed at 10 March 2006.

Haykin, S. (1999), *Neural networks: a comprehensive foundation*, 2nd edition edn, Prentice–Hall.

Leslie, C., Kuang, R. & Eskin, E. (2004), Inexact matching string kernels for protein classification, *in* B. Schölkopf, K. Tsuda & J.-P. Vert, eds, 'Kernel methods in computational biology', MIT Press, pp. 95–112.

Li, W., Ng, W.-K. & Lim, E.-P. (2004), Spectral analysis of text collection for similarity–based clustering, *in* H. Dai, R. Srikant & C. Zhang, eds, 'Proceedings of Pacific–Asia Knowledge Discovery and Data Mining Conference (PAKDD) 2004', LNAI 3056, Springer–Verlag, Berlin Heidelberg, pp. 389–393.

Müller, K., Mika, S., Rätsch, G., Tsuda, K. & Schölkopf, B. (2001), 'An introduction to kernel–based learning algorithms', *IEEE Transactions on Neural Networks* **12**, 181–201.

National Center for Infectious Diseases (2005), 'Proposal: clinical assessment of subjects with Chronic Fatigue Syndrome and other fatiguing illnesses in Wichita', cited; Available from: ftp://ftp.camda.duke.edu/CAMDA06_DATASETS/wichita_clinical_irb_protocol.doc.

Reeves, W. C. et al. (2005), 'Chronic fatigue syndrome — a clinically empirical approach to its definition and study', *BMC Medicine* **3**(19).

Shawe-Taylor, J. & Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge.

Smets, E. M., Garssen, B. J., Bonke, B. & DeHaes, J. C. (1995), 'The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue', *J. Psychosom. Res.* **39**, 315–325.

Wagner, D., Nisenbaum, R., Heim, C., Jones, J. F., Unger, E. R. & Reeves, W. C. (2005), 'Psychometric properties of a symptom–based questionnaire for the assessment of chronic fatigue syndrome', *BMC Health Quality Life Outcomes* **3**(8).

Ware, J. E. & Sherbourne, C. D. (1992), 'The MOS 36–item short form health survey (sf–36): conceptual framework and item selection', *Med. Care* **30**, 473–483.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B. & Bassett, D. E. (2006), 'Rosetta error model for gene expression analysis', *Bioinformatics* **22**(9), 1111–1121.

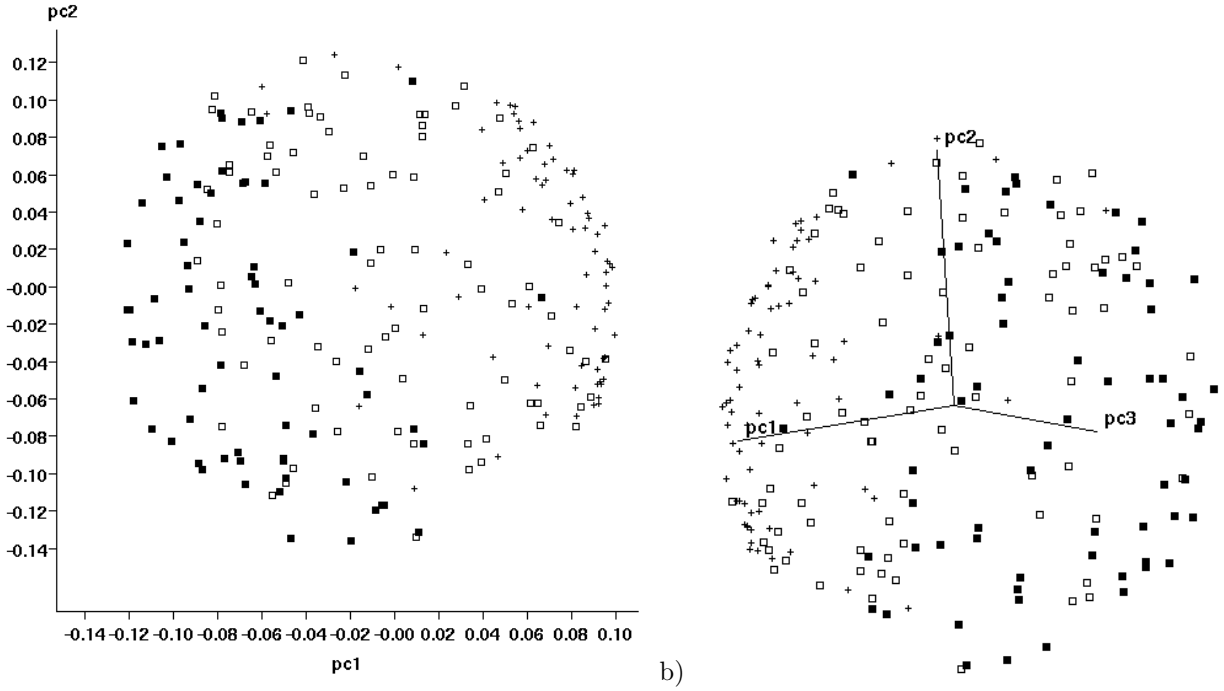


Figure 6: Kernel visualization of integrated illness and blood datasets using the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.

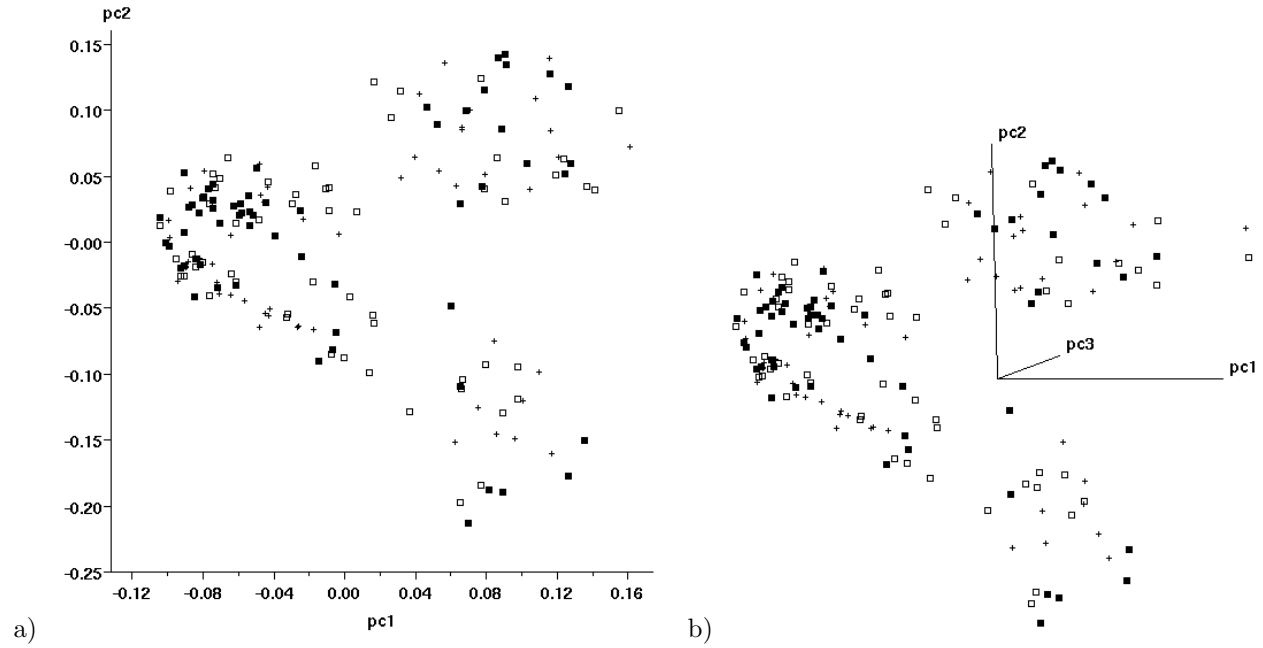
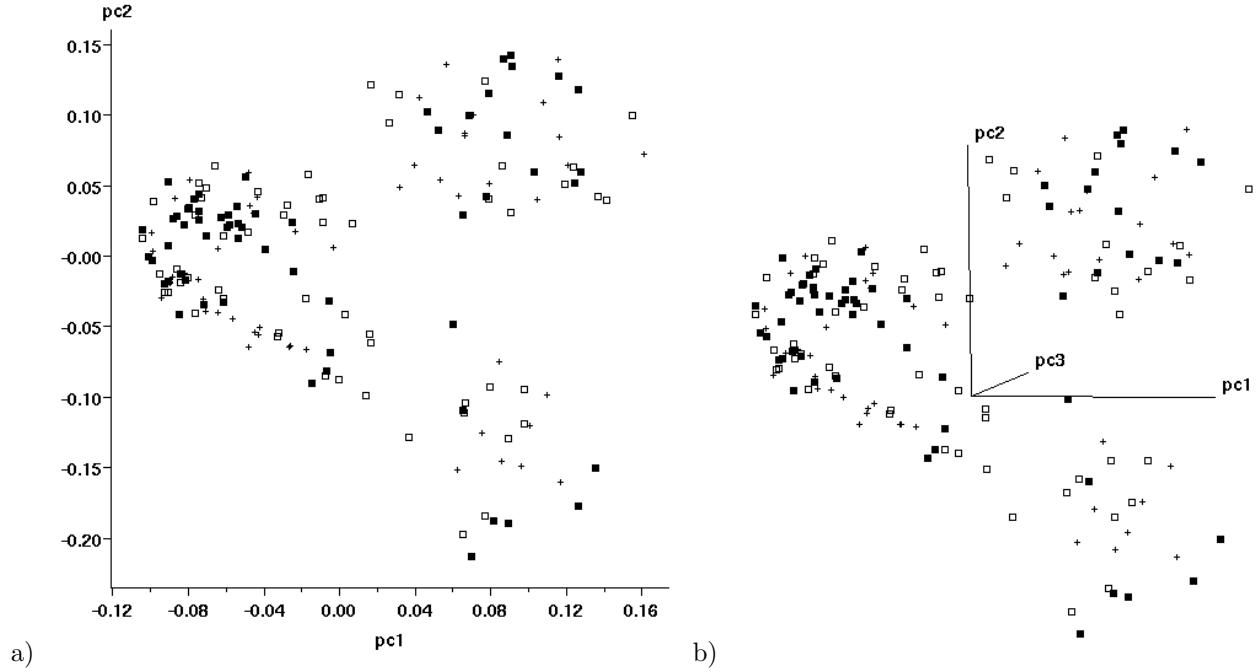
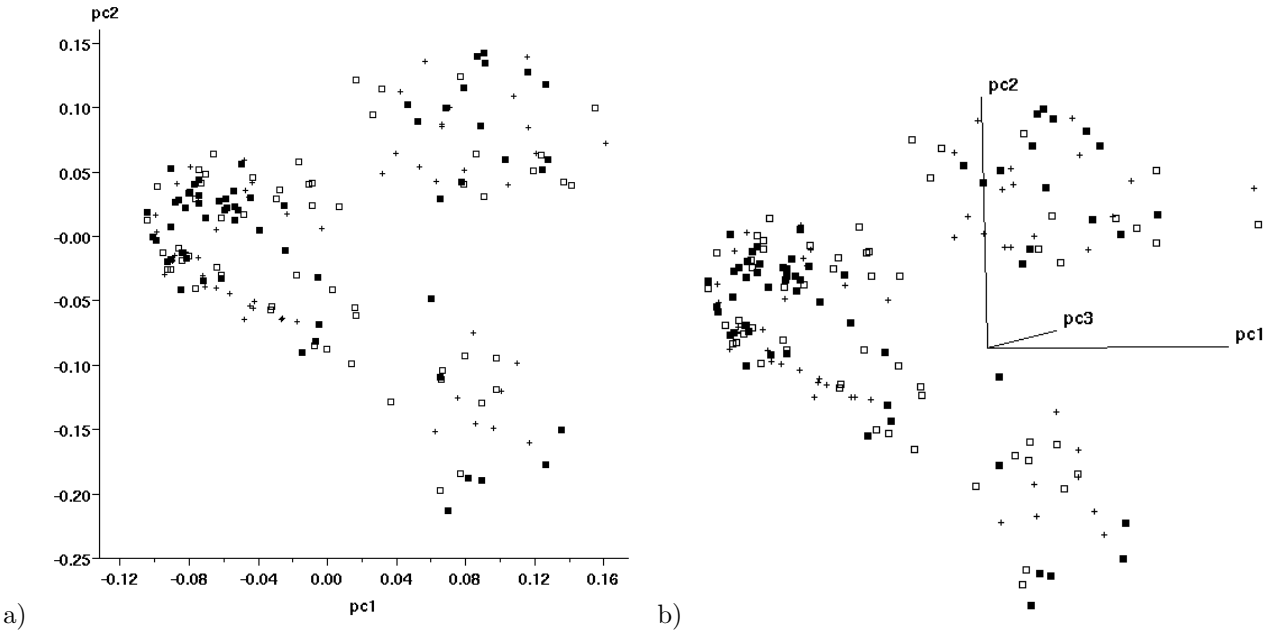


Figure 7: Kernel visualization of integrated blood and gene expression datasets using the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.



a) b)

Figure 8: Kernel visualization of integrated illness and gene expression datasets with the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.



a) b)

Figure 9: Kernel visualization of integrated blood, illness and gene expression datasets with the linear kernel. a) Two dimensional visualization with axes being the first two principal components. b) Three–dimensional visualization. + = NF individuals, □ = ISF patients and ■ = CF patients.

Analyzing Harmonic Monitoring Data using Data Mining

Ali Asheibi, David Stirling, Danny Soetanto

Integral Energy Power Quality and Reliability Centre
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong
Northfields Ave, Wollongong 2522, NSW

Email: atma64@uow.edu.au

Abstract

Harmonic monitoring has become an important tool for harmonic management in distribution systems. A comprehensive harmonic monitoring program has been designed and implemented on a typical electrical MV distribution system in Australia. The monitoring program involved measurements of the three-phase harmonic currents and voltages from the residential, commercial and industrial load sectors. Data over a three year period has been downloaded and available for analysis. The large amount of acquired data makes it difficult to identify operational events that impact significantly on the harmonics generated on the system. More sophisticated analysis methods are required to automatically determine which part of the measurement data are of importance. Based on this information, a closer inspection of smaller data sets can then be carried out to determine the reasons for its detection. In this paper we classify the measurement data using data mining based on clustering techniques which can provide the engineers with a rapid, visually oriented method of evaluating the underlying operational information contained within the clusters. The paper shows how clustering can be used to identify interesting patterns of harmonic measurement data and how these relate to their associated operational issues.

Keywords: Harmonics, Power Quality, Monitoring system, Data Mining, Classification, Clustering, Segmentation.

1 Introduction

With the increased use of power electronics in residential, commercial and industrial distribution systems, combined with the proliferations of highly sensitive micro-processor controlled equipment, more and more distribution customers are sensitive to excessive harmonics in the supply system (Heydt 1998), some even leading to failure of equipment. An increasing number of electric distribution network service providers are installing harmonic monitoring equipment to measure the three Phase harmonic voltage and current waveforms in their power system to detect and mitigate the harmonic distortion problems (Shuter, Vollkommer et al. 1989; Duggan and Morrison 1993; Lachaume, Deflandre et al. 1993; Morrison and Duggan 1993; Khan 2001; McGranagahn 2001).

Recently, a harmonic monitoring program was designed and implemented in a medium voltage (MV) distribution system in Australia (Gosbell, Mannix et al. 2001; Robinson 2003). The monitoring involved simultaneous measurements of the three-phase harmonic current and voltage from the residential, commercial, and industrial load sectors. The simultaneous measurements of three-phase harmonic currents and voltages from the different load sectors allow for the effect on the net distribution system harmonic voltage and current to be determined. The coordinated approach in obtaining the results has overcome some of the problems with synchronizing and reporting data (Emanuel and et al 1993; Sabin, Brooks et al. 1999).

An enormous amount of data over a three year period has been downloaded and available for analysis. However, it is difficult to analyze the data using visual inspection of the acquired voltage and current waveforms. It is also difficult to identify operational issues that generate the harmonics produced at varying operation time. A more sophisticated analysis method is required to automatically segment the data into manageable data set for analysis to understand the causes and effects of the harmonics obtained. These can then be used for resolving possible existing problems and to predict future problems. This implies that some decisions regarding data selection and acquisition must be made.

In this paper, a data mining tool is used for the automatic segmentation of the harmonic database. Segmentation (or clustering) is the discovery of similar groups of multidimensional records in databases and is a powerful mining tool. The data mining tool is based on the successful AutoClass (Cheeseman and Stutz 1995) and Snob research programs (Oliver, Baxter et al. 1996) and uses mixture models (McLachlan 1992) to represent clusters. The tool allows for the automated selection of the number of clusters and for the calculation of means, variances and relative abundance of the clusters. The paper describes how the data mining tool is used to search and analyze the multidimensional patterns on a cluster basis for the three-phase harmonic voltage and current data acquired from the harmonic monitoring program. By

observing the data in each cluster, several very interesting operational data can be deduced.

The paper will first define harmonics, describe the design and implementation of the harmonic monitoring program and the data obtained. Results from the harmonic monitoring program are then analyzed using the data mining tool. The paper discusses the significance of the clusters obtained and how the associated operational conditions can be deduced from the resultant clusters. Several observations on the use of such data mining tools for analyzing large amounts of harmonic data are also discussed.

2 Understanding Harmonics

Ideally, the waveforms of the voltage supplied by the utility and the current utilized by consumers are perfect 50HZ sine wave. However, in practice these waveforms are deformed, producing multiple frequencies other than the 50Hz sine wave This phenomenon of wave deformation due to multiple frequencies is often referred to as '*harmonic distortion*'. The sources of the harmonic are electronic based equipment, such as computers, fax machines, TVs etc and arcing devices, such as arc furnaces, arc welders and dischargeable lighting. Other sources of harmonic are saturable devices like transformers and motors (Shwehdi, Mantawy et al. 2002). Harmonic distortion can cause serious long term and short term problems to power systems, as well as to consumers. Harmonics can cause capacitor failure, overheating of neutral conductors and false tripping of electrical equipment in the utility. Harmonic losses in industrial systems can increase the operational cost and decrease the useful life of the system equipment (Carpinelli, Caramia et al. 1996). There are several practical mitigation methods, such as introducing filters, modifying loads and adjusting the frequency response of the system. For harmonic management of a network, it is mandatory to carry out careful harmonic monitoring analysis in order to determine the harmonic levels at any point in the power systems or at customer sites. The next section explains the steps of the harmonic monitoring program that has been carried out in this study.

3 Harmonic Monitoring Program

A harmonic monitoring program (Gosbell, Mannix et al. 2001; Robinson 2003) utilized a typical MV distribution system in Australia in a typical 33/11kV zone substation that supplies ten 11kV radial feeders. The zone substation is supplied at 33kV from the bulk supply point of a transmission network. Figure 1 gives the layout of the zone substation and feeder system for the harmonic monitoring program.

Seven monitors were installed, a monitor at each of the residential, commercial and industrial sites (site ID 5-7), a monitor at the sending end of the three individual feeders (site ID 2-4) and a monitor at the zone substation incoming supply (site ID 1). Sites 1-4 in Figure 1 are all within the substation at the sending end of the feeders identified as being of a predominant load type. Site 5 was along the feeder route approximately 2km from the zone substation, feeds residential area. Site 6 supplies a shopping centre with a number of large supermarkets and many small shops. Site 7 supplies a factory manufacturing paper product such as paper towels, toilet paper and tissues. Based on the distribution customer details, it was found that site 2 comprises 85% residential and 15% commercial, site 3 comprises 90% commercial and 10% residential and site 4 comprises 75% industrial, 20% commercial and 5% residential. The monitoring equipment used is the EDMI Mk3 Energy Meter from Electronic Design and Manufacturing Pty. Ltd. as shown in Figure 2 (EDMI 2000). All three line-to-neutral voltages and line currents were recorded at the LV locations. The memory capabilities of the above meter at the time of purchase limited recordings to the fundamental current and voltage in each phase, the current and voltage THD in each phase, and three other individual harmonics in each phase.

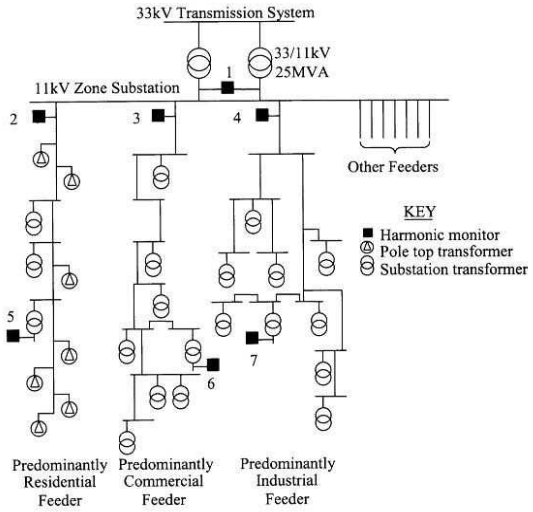


Figure 1: Single line diagram illustrating the zone distribution system

For the harmonic monitoring program, the harmonics chosen to be recorded were the 3rd, 5th and 7th harmonic currents and voltages at each monitoring site, since these are the most significant harmonics. The memory restrictions of the monitoring equipment dictated that each parameter was recorded only every 10 minutes. The data retrieved from the harmonic monitoring program spans from August 1999 to December 2002.

Figure 2: EDM1 2000-04XX Energy Meter

Figures 3-4, show a typical output data of the fundamental, 3rd, 5th and 7th harmonic currents in Phase 'a' at sites 1 and 2, taken from 6 -19 January 2001. It is obvious that for the engineers to realistically interpret such large amounts of data, it will be necessary to cluster the data into meaningful segments.

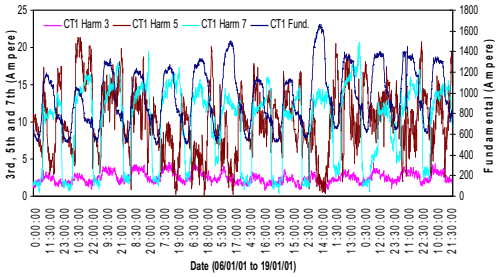


Figure 3: Phase 'a' Fundamental, 3rd, 5th and 7th Harmonic current waveforms in site 1 (zone substation)

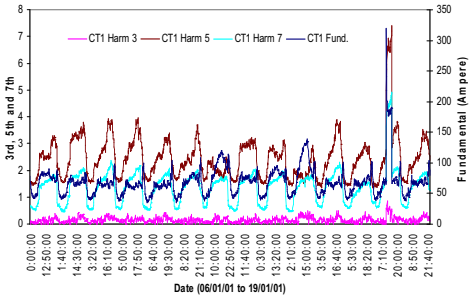


Figure 4: Phase 'a' Fundamental, 3rd, 5th and 7th Harmonic current waveforms in site 2 (residential)

4 Using Unsupervised Clustering with MML

Unsupervised clustering is based on the premise that there are several underlying classes that are hidden or embedded within a data set. The objective of such processes is to identify an optimal model representation of these intrinsic classes, by separating the data into multiple clusters or subgroups.

The partitioning of data into candidate subgroups is usually subject to some objective function like a probabilistic model distribution, e.g. Gaussian. From any arbitrary set of data, several possible models or segmentations might exist with a plausible range of clusters. Accordingly an appropriate evaluation scheme, such as Minimum Message Length (MML), or, Minimum Description Length (MDL) encoding, is used to evaluate each successive set of segmentations and monitor their progression towards a globally best model. This methodology is also known as finite mixture modelling or intrinsic classification (Wallace 1968; Wallace and Dowe 1994; R. Xu and D. Wunsch II 2005) . Algorithms such as "SNOB" or "Auto Class", which employ an MML approach have been found to form clusters that are statistically distinct (unique) within the multivariate parameter space of the data. The specific software used in this work was ACPro, one of several data mining algorithms that have yielded valuable insights for a range of scientific, and industrial real-time data (Oliver, Roush et al. 1998; Stirling and Zulli 2004). Finite mixture models provide a more formal (probabilistic-based) mechanism with which to fit arbitrary complex probability density functions (pdf's) of the data. In addition, from a practical perspective, such models also provide relief from the inherent constraints (priors and initial conditions) that accompany heuristic (distance) methods such as k-means or hierarchical agglomerative approaches (M. A.T. Figueiredo and A. K. Jain 2002).

5 Results and outcomes

A specialized data mining software package for the automatic segmentation of databases, ACPro, was used in this work. ACPro is essentially an unsupervised clustering that utilize minimum message length (MML) (Oliver, Baxter et al. 1996) encoding metric. It has the ability to discover similar groups of records in the database in the form of clusters. ACPro was applied to the measured harmonic data from the monitoring program for the test system in Figure 1. The data from different sites (sites 1, 2, 3 and 4) were used as input data to the software and 5 different clusters (s0, s1, s2, s3 and s4) each with specific abundance, mean and standard deviation were obtained. The clusters were then sorted in ascending order based on the mean value of the fundamental current, such that cluster (s0) is associated with the off peak load period and cluster (s4) is the cluster related to the on-peak load period. Figure 5(a) shows these clusters superimposed on the fundamental current waveform at site 3 (commercial). Figure 5(b) shows the abundance, mean and standard deviation for the clusters of the three harmonic currents per one phase.

By observing how the measured data are classified into various clusters, the Utility Engineer can more readily deduce the power quality event that may have trigger a change from one cluster to another. To confirm that, other available data in the utility can be used, such as temperature and reactive power measurements.

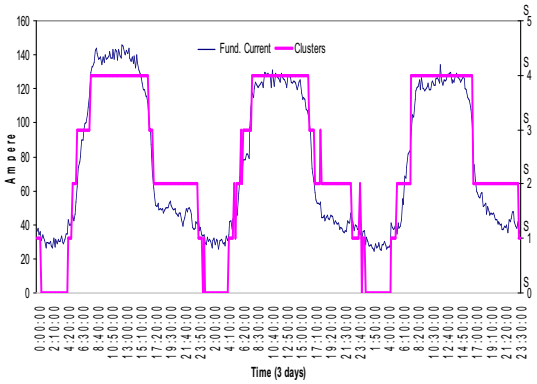


Figure 5(a): The clusters obtained superimposed on the phase 'a' fundamental waveform at site 3 (commercial)

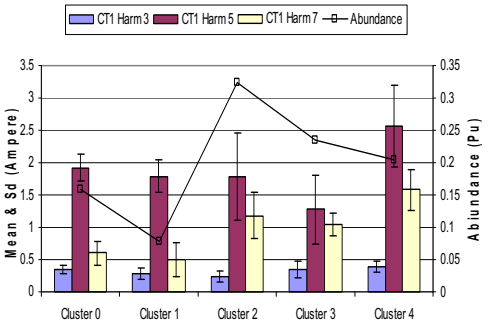


Figure 5(b): Abundance, mean and standard deviation for each cluster of harmonic currents (3rd, 5th and 7th) on phase 'a' at commercial site

Figure 6(a) shows the clusters obtained from substation site (site 1) superimposed on the fundamental current measurement data for two days. Figure 6(b) shows the 7th harmonic current and 7th harmonic voltage at the substation. Figure 6(c) shows the MVAr measurement at the 33kV side of the power system. From Figures 6(b) and (c), it can be observed that the second cluster (s2) is related to the capacitor switching event. Early in the morning, when the system MVAr is high as shown in Figure 6(c), the capacitor is switched on in the 33kV side to reduce bus voltage and late at night when the system MVAr is low, the capacitor is switched off to avoid excessive voltage rise.

The switching-on and switching-off of the capacitor is clearly reflected in the 7th harmonic current as shown in Figure 6(b). The capacitor switching operation in the 33KV side can also be detected at the other sites (residential, commercial and Industrial). Figure 7 shows the case for the commercial site (site 3), where the effect of the capacitor switching operation can be easily observed.

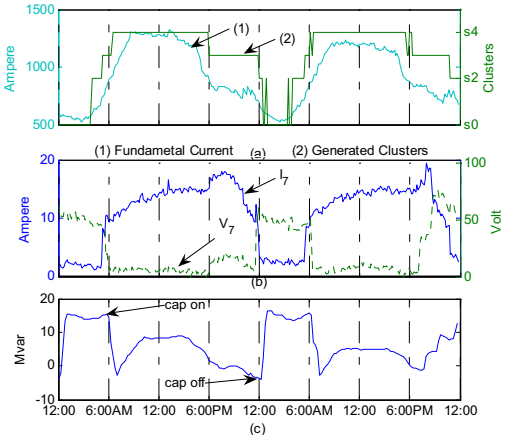


Figure 6: Clusters at substation site in two working days (a) Clusters superimposed on the fundamental current waveform,. (b) 7th harmonic current and voltage data. (c) MVAr load at the 33kv busbar.

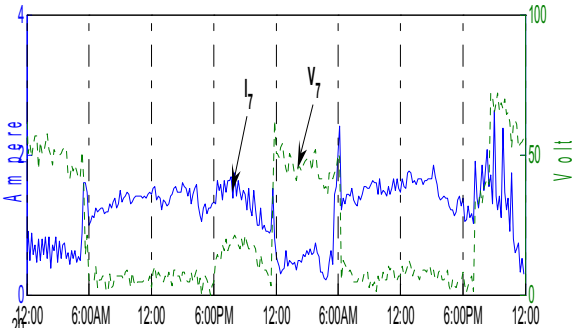


Figure 7: 7th harmonic current and voltage at Commercial site in two working days

The third cluster (s3) appears to occur in the evening period, when Figure 6(b) shows that the 7th harmonic currents rise around 6pm. and drops off around 11 pm. Upon discussion with the Utility Engineer, we were informed that this may be due to the reduction of industrial load and hence the percentage of the harmonics against the fundamental current becomes higher. It was also suggested that this could be due to the switching-on of TVs and appliances when people come home from work.

The same cluster is now applied to a week data from the residential site (site 2). The fundamental current during that week and the temperature data are shown in Figures 8(a) and (b). In this particular week, the Monday temperature (day 3) is relatively high, showing high fundamental current in day 3, resulting in significant use of air conditioners. Also in day 6, the fundamental current can be observed to increase significantly. To analyse these data, initially we concentrate looking at day 3 and 4 and observe how the data mining program cluster the data in these two days. This is then followed by the observation of the day 6 data to see if the data mining program can discriminate the high current data accurately.

Figure 9(a) shows the clusters superimposed on the fundamental current data of days 3 and 4. Cluster s0 is off peak and cluster s3 is on peak. Cluster 1 is representing the switching on and off of the capacitor at the 33KV side. Cluster s2 is representing turning on of TV's and any switch mode power supply (SMPS), in which period the 5th and 7th harmonic currents tend to be high (Fig 9-b). The reason this cluster start early in the morning rather than afternoon is because of the school holiday.

From Figure 9(a) it can be observed that there is a period of peak load (cluster s3) around midnight, and following discussion with the utility engineer, we were told that this is related to the turning-on of off-peak water heaters. Another on peak load (cluster s3) is from 11:00am to 8:00 pm, and from the temperature measurements (Figure 8-b) near the substation area, it can be concluded that air conditioners are the suggested load at that time.

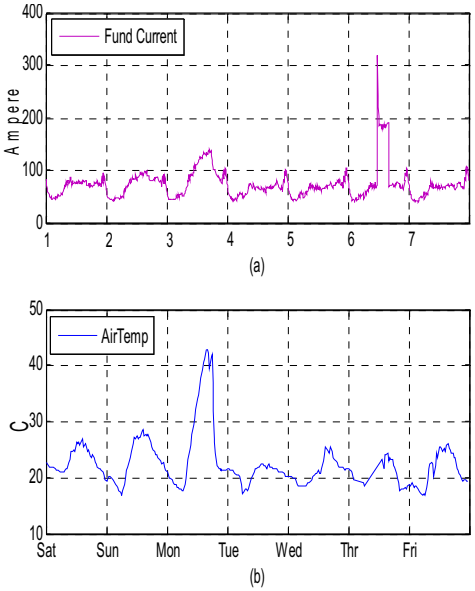


Figure 8: Fundamental current and air temperature in the area of the substation under study for one week from 13/01/2001 to 19/01/2001

Figure 10 shows the clusters obtained for day 6 when the fundamental load current rises in an unusual manner. It can be observed that the data mining program accurately identifies the abnormal event as a separate cluster (cluster 4) between 11:30 am until 4:30 pm in Figure 10(a). Further discussions with the utility engineer, revealed that this is a load transfer event from a faulty feeder to the residential feeder at that time. 5th and 7th harmonic currents are found to be high at that time.

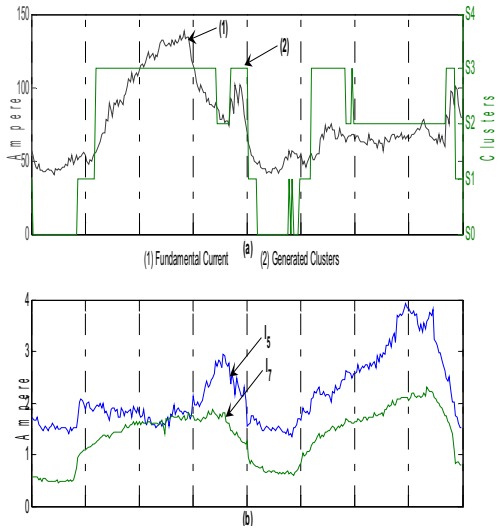


Figure 9: Clusters at residential site for two days (a) Clusters superimposed on the fundamental current waveform, (b) 7th harmonic current and voltage data

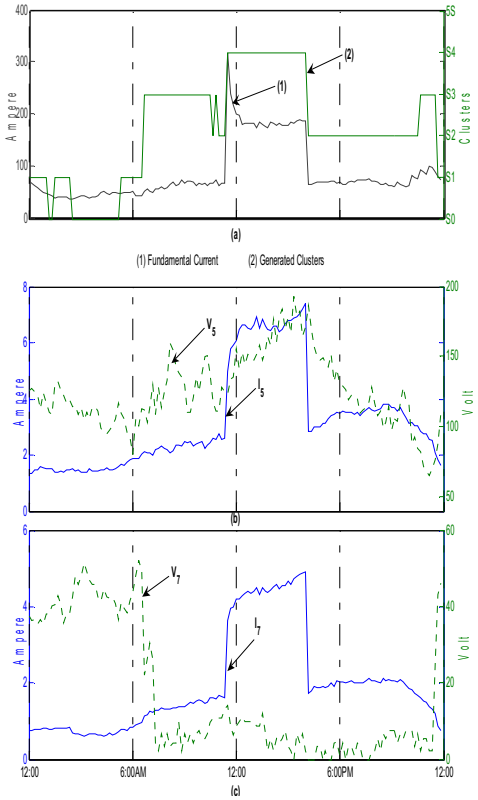


Fig.10: Overload on residential feeder due to transferring load from faulty feeder (a) Clusters superimposed on the fundamental current waveform, (b) 5th harmonic current and voltage data, (c) 7th harmonic current and voltage data

6 Conclusion

Power quality (PQ) data from a harmonic monitoring program in an Australian MV distribution system containing residential, commercial, and industrial customers has been analyzed using data mining techniques. Data mining, and in particular cluster analysis has been shown to be able to identify useful patterns within the data set. The utility engineers as experts in the domain can make ready use of the clustered data to quickly interpret these patterns, and in particular to detect unusual power quality events. The availability of this rapid, visually oriented method of evaluating the underlying information contained in large data sets can be an invaluable tool for power utility engineers.

7 References

Carpinelli, G., P. Caramia, et al. (1996). Probabilistic evaluation of the economical damage due to harmonic losses in industrial energy system. *Power Delivery, IEEE Transactions on* **11**(2): 1021-1031.

Cheeseman, P. and J. Stutz (1995). Bayesian Classification (AUTOCLASS): Theory and Results. In *Advances in Knowledge Discovery and Data Mining*. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusanny. Menlo Park, AAAI press**:** 154-180.

Duggan, E. and R. E. Morrison (1993). Prediction of harmonic voltage distortion when a nonlinear load is connected to an already distorted supply. *IEEE* **40**(3): 161-165.

EDMI (2000). Users Manual - EDMI 2000-04XX Energy Meter, Electronic Design and Manufacturing International.

Emanuel, A. E. and et al (1993). A Survey of Harmonic Voltages and Currents in Customer's bus. *IEEE Trans. Power Delivery* **8**(1): 411-421.

Gosbell, V., D. Mannix, et al. (2001). Harmonic Survey of an MV distribution system. *AUPEC, 23-26 September 2001*, Perth.338-342

Heydt, G. T. (1998). Electric Power Quality: A Tutorial Introduction. *IEEE Computer Applications in Power* **11**: 15-19.

Khan, A. K. (2001). Monitoring Power for the future. *IEEE Power Engineering Review Journal* **15**(2): 81-85.

Lachaume, J., T. Deflandre, et al. (1993). Harmonics in MV and LV distribution systems - Present and Future Levels. *IEE Disturbances and protection in Supply Systems*.

M. A.T. Figueiredo and A. K. Jain (2002). Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Learning* **24**(3): 381-396.

McGranagahn, M. (2001). Trends in Power Quality Monitoring. *IEEE Power Engineering Review Journal* **21**(10): 3-9, 21.

McLachlan, G. (1992). Discriminant Analysis and Statistical Pattern Recognition. New York, Wiley.

Morrison, R. E. and E. Duggan (1993). Long Term Monitoring of power system harmonics. *IEEE Colloquium (Digest)*(120): 2/1-2/3.

Oliver, J., R. Baxter, et al. (1996). Unsupervised Learning using MML, in "Machine Learning. *13th Int. Conf ICML '96*.364-372

Oliver, J., T. Roush, et al. (1998). Analysis Rock Samples for the Mars Lander. *American Association for Artificial Intelligence*: 299-303.

R. Xu and D. Wunsch II (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks* **16**(3): 645-678.

Robinson, D. (2003). Harmonic Management of MV Distribution Systems. PhD Thesis, School of Elec., Comp. and Telecom. Engineering, University of Wollongong.

Sabin, D. D., D. L. Brooks, et al. (1999). Indices for assessing harmonic Distortion from Power Quality Measurements: Definitions and benchmark Data. *IEEE Trans. Power Delivery* **14**(2): 489-496.

Shuter, T. C., H. T. Vollkommer, et al. (1989). Survey of harmonic levels on the American Electric Power Distribution System. *IEEE Trans. Power Delivery* **4**(4): . 2204-2213.

Shwehdi, M. H., A. H. Mantawy, et al. (2002). Harmonic flow study and evaluation of a petrochemical plant in Saudi Arabia. *Power Engineering 2002 Large Engineering Systems Conference on, LESCOPE 02*.165-172

Stirling, D. and P. Zulli (2004). Ontology Trend Analysis of Dynamic Signals. *International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP2004*, Melbourne, Australia.445-449

Wallace, C. (1968). Intrinsic Classification of Spatially Correlated Data. *The Computer Journal* **41**(8).

Wallace, C. and D. Dowe (1994). Intrinsic classification by MML - the snob program. *Proc. 7th Australian Joint conf. on Artificial intellegence*, Armidale, Australia, World Scientific Publishing Co

Discover Knowledge from Distribution Maps Using Bayesian Networks

Norazwin Buang
Research School of Information Sciences and Engineering (RSISE), Australian National University
Canberra, Australia
norazwin.buang@rsise.anu.edu.au

Nianjun Liu, **Terry Caelli**
National ICT Australia (NICTA)
Canberra Lab, ACT, Australia
nianjun.liu@nicta.com.au
terry.caelli@nicta.com.au

Rob Lesslie and Michael J. Hill
Bureau of Rural Sciences (BRS)
Canberra, Australia
rob.lesslie@brs.gov.au
hillmjdr@hotmail.com

Abstract

This paper applies a Bayesian network to model multi criteria distribution maps and to discover knowledge contained in spatial data. The procedure consists of three steps: pre processing map data, training the Bayesian Network model using distribution maps of Australia and testing the generalization and diagnosis of the model using individual states' maps. The Bayesian network that we used in this study is known as naïve Bayesian network. Results show that this environmental Bayesian network model can generalize the classification rules from training data for good prediction and diagnosis of a distribution map.

Keywords: Bayesian network, multi-criteria analysis, combining evidence, distribution maps.

1 Introduction

Analysis of spatial information in natural resource management is crucial to support a decision making process. However, with the advent of various technologies to acquire the data, analysis of multiple spatial data becomes a very challenging area. These technologies will produce the data with different accuracy and different resolution in the data. In spite of multi representations of spatial data, evidences of an area can be from different time intervals and different observer's views which make combination of those evidences complicated. Spatial data can have spatial attributes and non-spatial attributes. The former one is incorporated in spatial topological systems or relationships, and the latter is also called thematic data.

Environmental distribution map is crucial for decision support systems as it helps to monitor resource condition and also to identify the potential areas for investment. The information discovered from distribution maps is also contributing to applications such as land value determination, local and regional planning, pest and disease control, emergency response planning, agricultural productivity assessment and agricultural diversification (BRS, 2006). The complexity of a distribution map depends on the number of classes used to represent the data. It can be classified into as few as two classes or an infinite number of classes to represent the data. The more classes is used to quantise the data, the more precise the produced distribution map is, but also the more complexity of computation arises. In this study, we quantise the pixel values in a raster distribution map into five classes, where one colour is used to represent the associated class of a pixel.

In order to analyse the information contained in distribution maps, we need to discover as much knowledge as we can. Knowledge discovery on distribution maps includes combining multi criteria maps in order to extract general knowledge and interesting patterns from non-spatial attributes. The dependency between data that are usually uncertain make the analysis more complicated. Besides, different experts might have different opinions about the dependencies and factors involved in deriving any knowledge from spatial data. As a beginning, in this study, we focus on raster distribution maps as spatial data and discover knowledge from non-spatial attributes where we extract non-spatial attributes for each pixel in a distribution map. We apply Bayesian network model to do the generalization and diagnosis in predicting the class distribution of the data in a target map.

The aim of this study is to apply a Bayesian network in modelling multi criteria distribution maps. This paper provides a very basic introduction about Bayesian network. The technical details can be found in (Heckerman, 1995). The structure of this paper is as follows. We start in section 2, presenting several applications of Bayesian network in spatial data analysis. Section 3 describes the sources of the data used in this experiment. The process of the experiment is explained in section 4, including modelling distribution maps by naïve Bayesian network, data pre processing and model training and testing. The experimental results are discussed in section 5 and discussions are discussed in section 6. Finally, conclusions and future works are presented in section 7.

2 Literature Review

Spatial data can be derived from different type of sources. It has been used for discovering interesting knowledge and for making a decision. In order to do prediction and to produce a decision, the data needs to be represented in

a particular approach and one of them is Bayesian network. We will also describe several applications of Bayesian networks for prediction in this section.

A Bayesian network, also called belief net, is a directed acyclic graph (DAG) which consists of nodes to represent variables and arcs to represent dependencies between variables (Pearl 1986, Charniak 1991). Arcs or links also represent causal influences among the variables. The strength of an influence between variables is represented by the conditional probabilities which are summarized in a conditional probability table (CPT). Bayesian network is one of the graphical modelling techniques and it has been used widely in various applications including computer vision, medicine and spatial data analysis. There are two fundamentals idea in a Bayesian network according to De Vel et. al (2006). First, the notion of modularity where a complex system is decomposed into simpler parts and the second fundamental idea is their connections. The model also can deal with two main problems: uncertainty and complexity and therefore have an explanatory power for the modelling data.

In Bayesian networks, one can predicts the target values or the missing values given the model and other evidences. The class value for each pixel in a target node is obtained by finding the class associated with the highest posterior probability for that node. Bayesian networks work under the assumption of conditional independence where it estimates the posterior probabilities of the classification occurred in the training data. Bayesian networks have several advantages for data analysis. First, it can handle situations where some of the input data are missing. This is a great advantage, as having incomplete data is unavoidable in real world applications. Second, there are a few algorithms that have been developed for both structure and parameter training to learn the Bayesian networks from data. Third, Bayesian networks can be extended to model the structured data. This method is called probabilistic relational model (PRM) (Friedman 1999)

Representing spatial data using Bayesian networks for prediction has been applied successfully in many applications. Margaret et. al (2005) modelled satellite images using a Bayesian network for estimating leaf area index (LAI). The network was evaluated on a per pixel basis and the predicted results showed better classification than other classifiers such as neural networks and spectral vegetation indices, one of the pixel classifier approaches. On the other hand, Stassopoulou et al. (1998) used a Bayesian network to infer the risk of desertification of some burned forest in the Mediterranean region by combining several related evidences. The evidences used were from various sources with different resolution and accuracy. The network was also evaluated on a per pixel basis for modelling the data that came from different resolution.

Swayne (2004) used a Bayesian network and extended it into an influence diagram for multi objective modeling and decision support for a nonpoint source pollution model in Southern Ontario, Canada. They adjusted the conditional probability values of a Bayesian network to get a better decision based on the specific criteria preference. Besides prediction, the ability of Bayesian networks in detection applications has been successfully developed and the details can be found in Stassopoulou et al (2000) for building detection and Sebe et al. (2004) for skin detection.

3 Spatial Data Source

The data used in this study is acquired from the Multi Criteria Analysis Shell for Spatial decision support system (MCAS-S) provided by the Bureau Rural Sciences (BRS) of Australia (Michael et. al, 2005). Figure 1 and 2 show the distribution maps used to train the Bayesian network model. Figure 1(a) – (d) describe the distributions maps of population total, elevation, taxable income and accessible/remoteness index of Australia (ARIA) for 2001. The data has been quantized into five different classes and the scope for each class is also shown in the figure. One colour is used to represent a class while white pixel belongs to the background. These distribution maps are in a raster format where each pixel is associated with either a class or background.

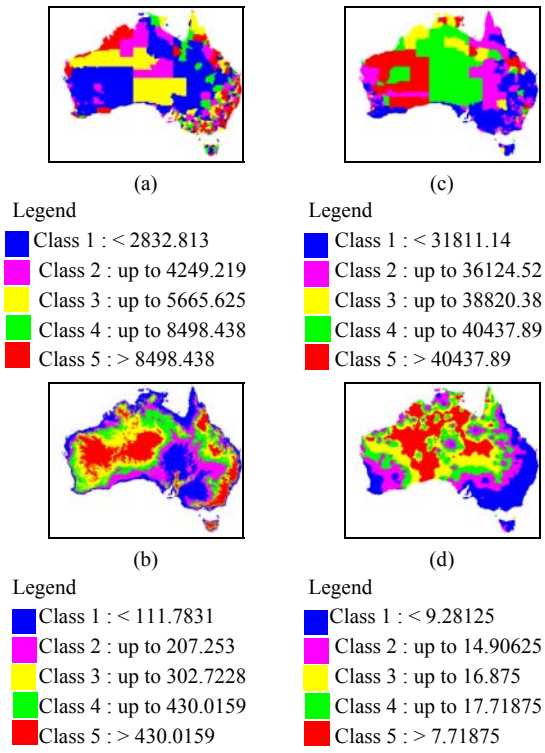


Fig. 1. Distribution maps of Australia (a) population total, (b) elevation, (c) taxable income, and (d) accessible remoteness index of Australia (ARIA).

In addition to the data shown in Figure 1, we are also given a distribution map of development potential as shown in Figure 2. Data of development potential distribution is also quantized into five different classes and the same colours are used to label the classes in this map. Figure 2(a) shows the distribution map of development potential of Australia while Figure 2(b) is a set of classes corresponding to the specified window in the map. All these five distribution maps are used in

parameter training. The detail of the training process is discussed in section 4.

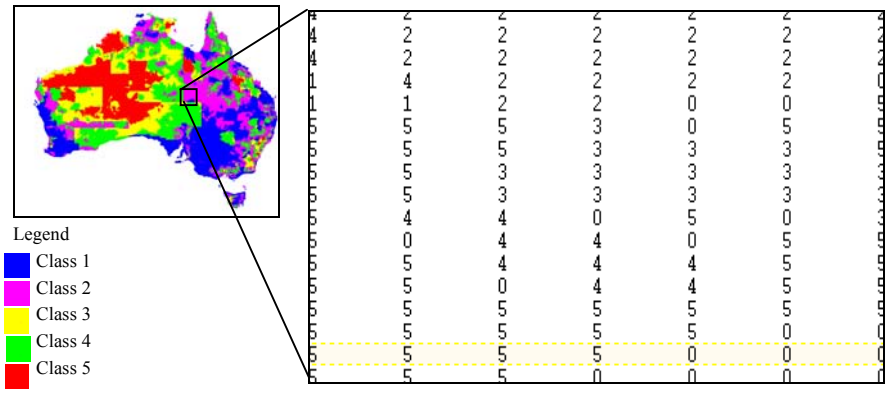


Fig. 2. (a) Distribution map of development potential of Australia (b) pixels's classes in the specified region.

Besides the distribution map of the whole Australia, we are also provided with the distribution maps of the same criteria (population total, elevation, taxable income and ARIA) for the individual states in Australia (Queensland, Victoria, New South Wales, Australian Capital Territory, Western Australia and Adelaide). In this study, Victoria distribution maps are used for testing the Bayesian network model and the details are discussed in section 5.

To get the class value of each pixel, we pre processed each distribution map using Matlab and Java. The size of the distribution map image is 270 pixels height and 340 pixels width. Therefore, for each image, there are 91800 pixel values. However, only some parts of the image contain class values while others are just the background. Section 4.1 describes the process in details.

4 Bayesian Network (BN) for Modelling Spatial Data

4.1 General Framework

Figure 3 shows the flow chart of the experiment. The first step is data pre processing. The RGB values of each pixel are extracted from a raster distribution map and for each of them a class number is assigned based on its RGB values. A class number '0' is set to a background pixel, while class '1' to '5' are assigned to five different states or classes for each node. In the second step, the pixel values in the distribution map of Australia are used for training the Bayesian network model to obtain the model parameters- the Conditional Probability Table (CPT). In this study, we assume the structure of the Bayesian network is known and we focus only on parameter training. The next two steps are testing. We apply an individual state-Victoria distribution maps to test the performance of the proposed Bayesian Network Model. In the first experiment, we test the trained model to infer the distribution map of development potential of the whole state of Victoria, while in the second experiment we apply the model to infer the Victoria elevation distribution map. The proposed Bayesian network model will be discussed in details in the following subsections.

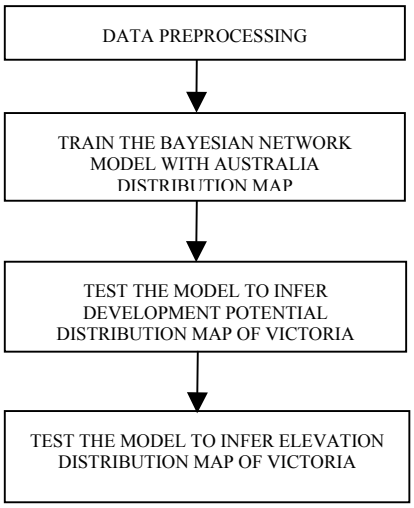


Fig. 3. General framework of the experiment

4.1.1 Modelling structure.

In the Bayesian network model used in this experiment, each distribution map is represented by a node. Figure 4 shows the structure of the Bayesian network. The *Development Potential* node is a child node for *population*, *elevation*, *tax-income* and *ARIA*. *Population* node, *elevation* node, *tax-income* node and *ARIA* node are parent nodes for *Development Potential* and also known as root nodes. There is no link between root nodes. Each node has five states that are 'vlow', 'low', 'med', 'high' and 'vhigh' associated with five classes in the distribution map, and the state's value corresponds to the probability of that class in the map. Class '1' corresponds to state '*vlow*'(very low), while class '2' corresponds to state '*low*'. Class '3' is the state '*med*' (medium) and class '4' equals to state '*high*'. Finally, class '5' corresponds to state '*vhigh*' (very high).

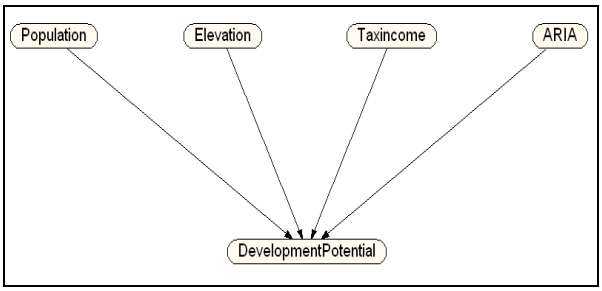


Fig. 4. The proposed Bayesian network model

4.1.2 Training data and Trained Model

The non-spatial attributes from the distribution maps of Australia are used to train the model. We only select the value of each pixel that contains the available class for making up the training dataset as the format shown in Figure 5. Norsys's Netica Java software toolkit is used to train the Bayesian network. The data is incorporated into the network with *'.cas'* file format. There are six columns in the file including the IDnum. The .cas file is created by using Java codes.

Idnum	Population	Elevation	Taxincome	ARIA	DevelopPotential
1	vhigh	vlow	low	vhigh	med
2	med	low	med	low	med
3	vhigh	vlow	vhigh	med	low
:					
:					
:					
35881	low	high	vhigh	vlow	low
35882	vhigh	vlow	vlow	vhigh	vlow

Fig. 5. Format of the input data into Bayesian network (.cas file)

After training the model, a conditional probability table (CPT) is assigned to children nodes while prior probability is assigned to root nodes. According to the structure, *Development Potential* has a CPT while all other nodes have prior probabilities. As we are using the Netica learning algorithm, the prior distributions are Dirichlet functions. The *Development Potential* node consists of five states and four links directed into this node has $5^4 = 625$ rows in its CPT. Figure 6 shows parts of the CPT for *Development Potential* node derived from Netica (2006).

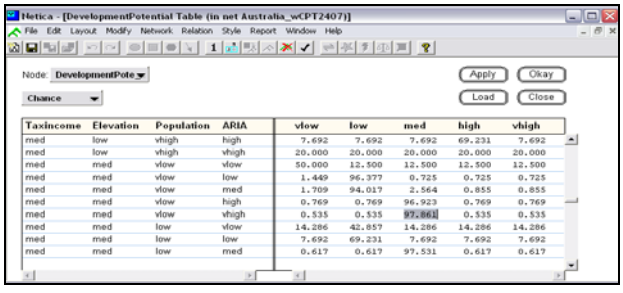


Fig. 6. Conditional Probability Table (CPT) for *Development Potential* node.

4.1.3 An example of proposed Bayesian Network Model

Here, an example of the mode in the pixel level is explained. Figure 7 shows the posterior probabilities for each class in *Development Potential* node when the state of population node is '*very low*', state of *Elevation* node is '*medium'*, state of *Taxable Income* node is '*medium'* and state of *AccessRemoteIndexAustralia (ARIA)* node is '*very high'*. From this sample, we can see that the posterior probabilities from class '1' to class '5' are 0.53, 0.53, 97.9, 0.53 and 0.53 respectively. As a state '*med'* holds the maximum probability (97.9%), the class of the pixel is inferred as class 3, where green colour is set to represent that pixel in the target distribution map.

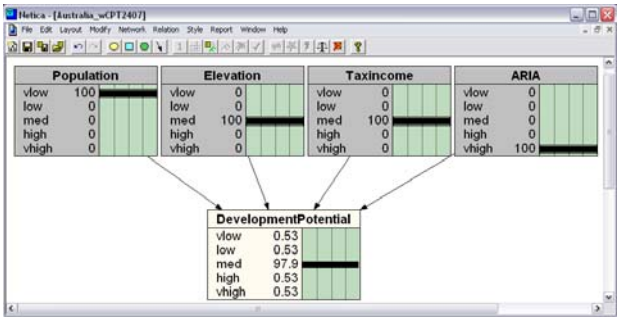


Fig. 7. Posterior probability inferred for each class of Development Potential node in Netica.

5 Experimental Results

This section presents the results of two testing experiments: generalization and diagnosis. The distribution maps of Victoria are used to test the model in both experiments. We compared on the values for each pixel. After the states' probabilities of each pixel are inferred, the pixel is set to the class/state with maximum probability and the pixel's colour is set accordingly. We only present the visual comparison in this paper for both testing experiments. We use the fastest known inference algorithm associated with Netica, a junction tree of clique algorithm for exact general probabilistic inference.

The first experiment is to test the model to infer class value of *Development Potential* node given other nodes (Population, Elevation, TaxableIncome and ARIA). Figure 8 shows the data and the result of the experiment. Figure 8(a)–(d) presents the data used as inputs to the model. Figure 8(e) is the development potential distribution map inferred by the Bayesian network model while Figure 8(f) is an empirical development potential distribution map, also as an expert knowledge. From the result, we can see that there are no big differences between the distribution map in Figure 8(e) and 8(f).

The second experiment is to infer the uncertainty—the class of the pixel with missing value of *Elevation* node given other nodes (Population, TaxableIncome, ARIA, and Development Potential). Figure 9 shows the data and the results of the experiment. Figure 9(a), (c), (d), (e) shows the data used as inputs to the model. Parts of the distribution map of elevation is removed and we used

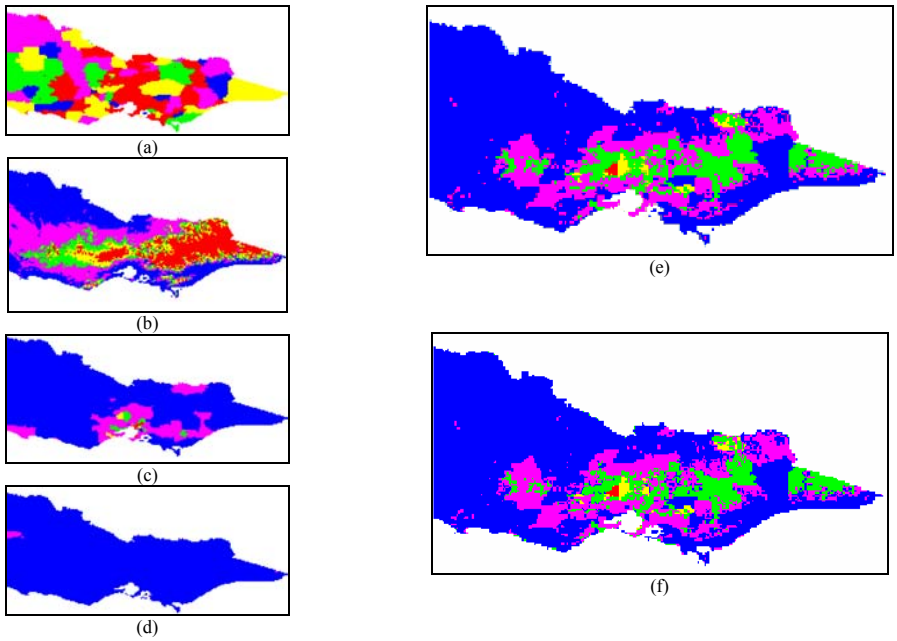


Fig. 8. Data and result of testing the generalization of the training data (a) Data Population (b) Data Elevation (c) Data Taxable Income (d) Data ARIA (e) Development potential distribution map inferred by the Bayesian network model (f) Empirical Development Potential

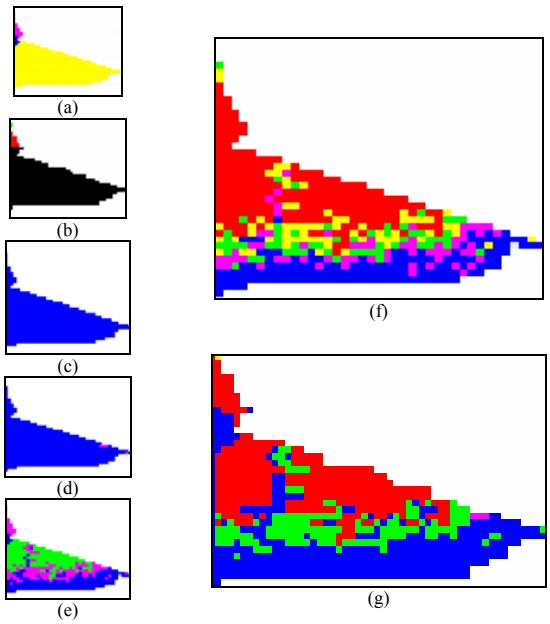


Fig. 9. Data and result of testing the diagnosis of the training data (a) Data Population (b) Data Elevation (missing data) (c) Data Taxable Income (d) Data ARIA (e) Data Development Potential (f) Data Elevation inferred by the model (g) Complete Data Elevation

black pixels to represent the missing values as shown in Figure 9(b). Figure 9(f) is the Elevation distribution map inferred by the Bayesian network model while Figure 9(g) is the elevation distribution map acquired from the MCAS-S application. More than 80% of pixel's colour matches quite well between the distribution maps in Figure 9(f) and 9(g).

6 Discussions

The use of Bayesian network in this context assumes that all data in distribution maps can be quantized into the same number of classes where each class has more or less the same number of pixels. In real world, this assumption might not be true. Moreover, we do not use a truncated function to limit the range in each class. For example, there might be a limit value for elevation that does not influence other nodes. This limitation can be solved if the expert opinions are included into the modelling process.

This problem is also related to the use of different modes between equal area distribution maps and equal interval distribution maps. In the future, it might be worth to understand how both modes affect the parameter estimation process in Bayesian networks.

The process to extract the class or state value for each pixel is not very accurate but still reliable. This is because of the difficulties to extract the values for each pixel especially at the boundaries between different states/classes. In this study, the class value for each pixel is only based on RGB values. It is quite hard to identify the exactly class value for each pixel that have similar colours or quite similar RGB values. As a result, some of the pixels are considered as missing if the program cannot distinguish the class values for that pixel.

This study only considered parameter training and the structure training is not included. There is a lot of research that focus on developing algorithms for structure training. In the future, we are planning to incorporate structure learning when constructing a Bayesian network. Since the dependencies between spatial environmental distributions are in diversity, the future work will include more complex model structure and model parameters' training. For the future, we are also planning to include learning algorithms when hidden variables are present. Despite all three problems discussed above, we are also facing difficulties to get more data for testing the model that we build.

7 Conclusions and Future Work

This paper describes how a Bayesian network can be used to model raster environmental distribution maps and their dependencies. It demonstrates that a Bayesian network is quite robust to discover knowledge from distribution maps, even though the present Bayesian network used in this study is only naïve Bayesian network. As their full potential has not yet been explored, we recommend it as one of the future works.

Acknowledgments The authors acknowledge Bureau Rural Sciences (BRS), Department of Agriculture, Fisheries, and Forestry for providing the data in the experiment. Financial support for this project from the Australian Research Council is gratefully acknowledged. National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology, and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Research Centre of Excellence programs. The authors would also like to thank Mr. Nathan Brewer for his supports in reviewing and helping correcting this article.

8 References

BRS (Bureau of Rural Sciences) (2006): Guidelines for land use mapping in Australia: principles, procedures and definition.

Heckerman, D. (1995): A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research (MSR).

Pearl, J (1986): Fusion, propagation and structuring in belief networks. Artificial Intelligence, 29, 241-288.

Charniak, E (1991): Bayesian networks without tears. AI Magazine. Vol. 12(4). 50-63.

N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. (1999): Learning probabilistic relational models. *In Proceedings of the Sixteenth International Joint Conferences on Artificial Intelligence* (IJCAI99).

De Vel, O, Liu, N., Caelli, T. and Caetano, T. (2006): An Embedded Bayesian Network Hidden Markov Model for Digital Forensics: *in IEEE Intelligence and Security Informatics Conference (ISI 2006)*, pp459-465, May 23-24, San Diego, USA

Kalacska, M., Sanchez-Azofeifa, G. A., Caelli, T., Rivard, B., Boerlage, B. (2005): Estimating leaf area index from satellite imagery using Bayesian Networks, *IEEE Transactions on Geoscience and Remote Sensing.*

Stassopoulou, A., Petrou, M., Kittler, J. (1998): Application of a Bayesian network in a GIS based decision making system. *International Journal of Geographical Information Science* 12 (1), 23±45.

Swayne, D.**,** Jie Shi. (2004): Possible Courses: Multi-Objective Modelling and Decision Support Using a Bayesian Network Approximation to a Nonpoint Source Pollution Model. Complexity and Integrated Resources Management, *Transactions of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society*, iEMSs: Manno, Switzerland, :568-573.

Stassopoulou, A. and Caelli, T. (2000): Building Detection using Bayesian networks. *International J.Pattern Recognition Artificial Intelligence*, vol. 14, 715-733.

N. Sebe, I. Cohen, T. Gevers, T.S. Huang. (2004): Skin Detection: A Bayesian Network Approach. *International Conference on Pattern Recognition*, Cambridge, UK.

Michael J. Hill, Rob Lesslie, Andrew Barry, Simon Barry. (2005): A Simple, Portable, Spatial Multi-Criteria Analysis Shell – MCAS-S. In Zerger, A. and Argent, R.M. (eds) MODSIM 2005 *International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand* .

Netica Java API Library, http://www.norsys.com/. Accessed 1 August 2006.

Data Mining for Lifetime Prediction of Metallic Components

Esther Ge[1] Richi Nayak[1] Yue Xu[2] Yuefeng Li[2]

[1] School of Information Systems, [2] School of Software Engineering and Data Communications
Queensland University of Technology
GPO Box 2434, Brisbane Qld 4001, Australia

{e.ge,r.nayak,yue.xu,y2.li}@qut.edu.au

Abstract

The ability to accurately predict the lifetime of building components is crucial to optimizing building design, material selection and scheduling of required maintenance. This paper discusses a number of possible data mining methods that can be applied to do the lifetime prediction of metallic components and how different sources of service life information could be integrated to form the basis of the lifetime prediction model.

Keywords: Data Mining, prediction, corrosion, civil engineering

1 Introduction

Our globe is increasingly challenged by growing populations and aging infrastructure. An escalating demand to maintain the infrastructure is always at place. Service life of building components is a key issue in predictive and optimizing design and management of buildings and civil infrastructures. It is influenced by many factors like materials, environment and maintenance etc. The corrosion of metallic components is the main factor that influences the service life of building.

Recent Australian research found there are over 300 metallic components with 2-3 materials and 2-3 coatings in a standard Australian house (I. Cole et al., 2006). Those components in general have been exposed to environments. Corrosion decay is very serious in metallic components due to sunlight, rain and salt deposition. Therefore, it is necessary to develop efficient means of estimating corrosion rate and then the service life. The need to develop accurate methods to predict the lifetime of metallic components has become an international recognition. For example, the European Performance Based Building network and the CIB working group W80 on design life of buildings is working on the further development of the Factorial Approach to predict the service life of building components (I. Cole et al., 2006).

The material should be selected to match the severity of the environment. For example, in severe marine locations, very durable materials need to be selected while in benign environments lower quality products can be used. As with materials selection, the timing of maintenance and building design would be tailored to the severity of the environment. Through these ways, substantial cost savings can be made. For example, it has been estimated that nearly $5 million was spent by Queensland Department of Public Works (QDPW) in 03/04 in replacing corroded metallic components in Queensland schools (I. Cole et al., March 2005).

Data mining is a powerful technology to solve prediction problem (Fayyad, Piatetsky-Shapiro, & Smyth, 1995b). It has been effectively applied to civil engineering for corrosion prediction. For example, Kessler et al. (1994) improved prediction of the corrosion behavior of car body steel using a Kohonen self organizing map. Furuta et al. (1995) developed a practical decision support system for the structural damage assessment due to corrosion using Neural Network. More recently, Morcous et al. (2002) proposed a case-based reasoning system for modelling infrastructure deterioration. Melhem and Cheng (2003) first used KNN and Decision Tree for estimating the remaining service life of bridge decks. And later Melhem et al. (2003) investigated the use of wrapper methods to improve the prediction accuracy of the decision tree algorithm for the application of bridge decks. However, limited research was conducted on comparing the prediction accuracy of various methods and how we can get the best prediction accuracy to the corrosion rates.

This paper discusses a number of possible data mining methods that can be applied to do the lifetime prediction of metallic components and how different sources of service life information could be integrated to form the basis of the lifetime prediction model. Experiments are conducted with Weka, a public data mining tool.

2 Data Acquisition

The data sets include three different sources of service life information: Delphi Survey, Maintenance database, and Holistic Corrosion Model. The Delphi Survey includes the estimation of service life for a range of metallic components by experts in the field such as builders, architects, academics and scientists. Maintenance database is derived from the maintenance records that provide a repository of past experiences on component lifetime predictions under specific conditions. Holistic Model is based on a theoretical understanding of the basic corrosion processes (I. S. Cole, Furman, & Ganther, 2001). It provides the required knowledge for computing the lifetime of metallic components. An independent model for Colorbond is included in the Holistic model since Colorbond has different features from other materials. Details of these data sets are presented in Table 1.

Data Set	Number of Cases	Number of attributes	Target attribute
Delphi Survey	683	10	Mean
Maintenance Database	1297	18	Zincalume Life
			Galvaniz-ed Life
Holistic Model	9640	11	MLannual
Colorbond	4780	20	Life of gutter at 600um

Table 1: Details of Data Sets

The Delphi survey data set contains the predicted life information for over 30 components, 29 materials, for marine, industrial and benign environments of both service (with and without maintenance) and aesthetic life. They are knowledge of domain experts. The output of this data set is an estimated components life. The estimated life was stored in two forms: the mode and the mean as well as a standard deviation for the mean. The mean is the average years of service life, aesthetic life or time to first maintenance. As the Delphi dataset is the result of surveys, the final dataset was examined in three ways to determine its accuracy and reliability. They were analysis for internal consistency of the data, analysis for consistency with expected trends based on knowledge of materials performance and correlation with existing databases on component performance. In all of these comparisons, the Delphi dataset showed good agreement (I. Cole et al., March 2005).

The maintenance data set contains life information of roof component for schools in Queensland. They are the results of analysing over 10000 records with regard to significant maintenance. The outputs are service life of Zincalume and Galvanized Steel materials for roofs.

The holistic data set contains theoretical information of corrosion for gutters in Queensland schools. The overall model is a reflection of influence of climatic conditions and material/environment interactions on corrosion. The output of this data set is the annual mass loss of Zincalume or Galvanized steel. Once the mass loss of material is determined, its service life is measured with appropriate formulas (I. Cole et al., March 2005).

Because Holistic model has no facility for handing the particular material Colorbond, the rules for the degradation of Colorbond is devised separately. The Colorbond data set includes this information. The output of Colorbond is service life of Colorbond for gutters.

In general, Delphi Survey is expert opinions; Maintenance database is operational while Holistic Model is theoretical. They form three important source of information for predicting lifetime of metallic components. They are independent but complement each other. Delphi Survey can be used for analysing correlation with other two data sets on component performance and consistency with expected trends based on knowledge of materials performance while Maintenance database and Holistic Model provide de facto and theoretical proof respectively for prediction. Maintenance, Holistic and Colorbond relate to different component types with different material while Delphi contains all component types with all material, which can be used to check for consistency. More specifically, Maintenance is for roofs with Galvanized Steel and Zincalume, Holistic is for gutters with Galvanized Steel and Zincalume, Colorbond dataset is for gutters with Colorbond and Delphi is for a range of components including roofs and gutters with different materials including Galvanized Steel, Zincalume and Colorbond. There is no overlap of predicted outcome from Maintenance, Holistic and Colorbond while the predicted outcome from them can be compared with the outcome from Delphi.

3 Data Mining for Lifetime Prediction

In this section, we explore various predictive data mining techniques to apply for lifetime prediction problem and to find a best one. Before a learning algorithm is applied, the data must be pre-processed (Olafsson, 2006).

3.1 Data Pre-processing

Data quality is a key aspect in performing data mining on a real-world data. Raw data generally include many noisy, inconsistent and missing values and redundant information. In this section, we explain how data is pre-processed to make data ready for mining.

3.1.1 Feature Selection

Feature selection is for removing those attributes irrelevant to mining results. In our data sets, some attributes like Centre Code, Centre Name and LocID only provide identification information. They have no mining value. Similarly, some attributes such as Building Type and Material in Colorbond contain only one value. They were also ignored during mining.

For Delphi Survey, the estimated life was stored in two forms: the mode and the mean as well as a standard deviation (SD) for the mean. As we want a real value to be the final predicted result, the attribute ‘mean’ is chosen as the target attribute. All other attributes are kept as inputs to know their influence to the target value. They are as follows:

Building type | Component | Measure | Environment | Material | Maintenance | Criteria | Mean

For maintenance database, there are two target attributes: Zincalume Life and Galvanized Life. After examining all attributes carefully, we found that some attributes are only related to ‘Zincalume Life’ while others are only related to ‘Galvanized Life’. Therefore, we divided maintenance database into two parts: one is for ‘Zincalume Life’ and the other is for ‘Galvanized Life’. The attribute ‘Centre Code’ and ‘Centre Name’ are removed since they are

identification information. The final attributes for 'Zincalume Life' and 'Galvanized Life' are as follows:

Longitude | Latitude | Salt Deposition | Zincalume Mass Loss | Marine | N | Zincalume Life

Longitude | Latitude | Salt Deposition | Zinc Mass Loss | Steel Mass Loss | Marine | Nzinc | Nsteel | L | M | Zinc Life | Steel Life | Galvanized Life

For Holistic Model, as we describe in Data Acquisition section, the service life is calculated based upon 'MLannual'. We create a target variable named 'Service Life' which is calculated from formulas (I. Cole et al., March 2005). Similarly, 'LocID' and 'Location' are removed because they are identification information. 'State' and 'Building Type' are also ignored since they only have one value. Therefore, the final attributes are as follows:

XLong | YLat | SALannual | Material | Gutter Position | Gutter Maintenance | MLannual | Service Life

Similar process has been done for Colorbond. 'LocID', 'Building Type', 'Position', 'Material', 'Building Face' and 'BuildingFacePos' are ignored because they are either identification information or only have one value. The final attributes are as follows:

SALannual | Exposure | PositionVsExposure | Gutter Type | rain_annual_mm | cum_MZa_2ndYear | cum_dSTEEL_2ndYear | remCr | normCr | accelerated_corrosion_rate | Time to White Rust of Zincalume | Time to penetration of Zincalume | Time to onset of Red Rust | Life of gutter at 600um

'Life of gutter at 600um' is the target attribute.

3.1.2 Data Cleaning

In our data sets, the percent of missing values are very low. For example, for Delphi Survey, only the attribute 'mode' has 8% missing values while all other attributes have no missing values. For Colorbond, all attributes have no missing values. However, inconsistent values do exist in every data set. An example is the use of lowercases and capitals such as 'Steel' and 'steel'. More examples are different spellings but same meaning like 'Galvanised' and 'Galvanized' or different words but same meaning like 'Steel in Hardwood' and 'Steel-Hardwood'. More spaces are included in values could be another reason to cause inconsistency like 'Residential ' and 'Residential '. Data mining tool will treat those kinds of values as different values and hence influence the predicted results. All such errors are recovered during data cleaning. For example, the 'Material' attribute in Delphi Survey originally has 36 values. After cleaning, there are only 29 values.

3.1.3 Data Discretization

Data discretization is considered because some learning algorithms are better able to handle discrete data. We discretized all numeric attributes including target attributes to nominal type by dividing them into ranges before applying Naïve Bayes and Decision Tree mining algorithm. For example, 'Mean' contains values from 3 to 58. It is divided into 10 ranges: [3-13], (13-17], (17-21], (21-25], (25-29], (29-33], (33-37], (37-41], (41-45], (45-58]. While for other classification data mining methods like Neural Network and SVM, we keep all continuous values.

3.2 Data Modelling and Mining

Our main objective in this research is to make an accurate prediction for the lifetime of metallic components. Therefore, our problem is a prediction data mining problem. The overflow of prediction model is given in Figure 1.

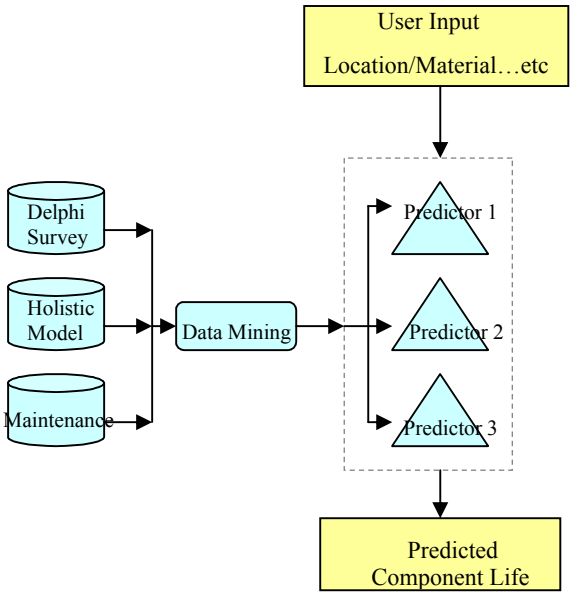


Figure 1: Overflow of Prediction Model

Data mining methods are applied to all three data sets to build three predictors first. After that, these three predictors can make predictions for user's inputs. The final predicted life is either a multiple choice provided by three predictors or a value combined from the outputs of three predictors.

In order to get accurate predicted results, we have applied various data mining methods including Naïve Bayes, K Nearest Neighbors (KNN), Decision Tree (DT), Neural Network (NN), Support Vector Machine (SVM) and M5 Model Trees on these data sets. Naïve Bayes is a statistical-based algorithm. It is useful in predicting the probability that a sample belongs to a particular class or grouping (Fayyad, Piatetsky-Shapiro, & Smyth, 1995a). KNN is based on the use of distance measures. Both DT (Quinlan, 1986) and NN are very popular methods in data mining. DT is easy to understand and better in classification problems while NN can not produce comprehensible models in general and is more efficient for predicting numerical target.

Support vector machine (Vapnik, 1995) is relatively new method. It can solve the problem of efficient learning from a limited training set. M5 Model trees (Quinlan, 1992) is an effective learning method for predicting real values. Model trees, like regression trees, are efficient for large datasets. However, model trees are generally much smaller

than regression trees and prove more accurate (Quinlan, 1992).

All those are traditional data mining methods. We also used bagging (Breiman, 1996) to improve the performance of these methods. Bagging generates multiple predictors and uses these to get an aggregated predictor, which has better performance.

4 Experimental Results and Discussion

Although there can be many performance measures for a predictor such as the training time and comprehensibility, the most important measure of performance is the prediction accuracy in real-world predictive modelling problems (Zhang, Eddy Patuwo, & Y. Hu, 1998). For classification problem, prediction accuracy is defined as the number of correctly classified instances divided by total number of instances. For regression problem, correlation coefficient is often used to evaluate the performance. Correlation coefficient measures the statistical correlation between the predicted and actual values.

Data set	Prediction Accuracy	
	Naive Bayes	Decision Tree
Delphi Survey	30.0587%	36.217%
Holistic Model	89.744%	90.125%
Colorbond	94.728%	96.548%
Maintenance for Galvanized	93.138%	94.603%
Maintenance for Zincalume	91.904%	93.215%

Table 2: Prediction accuracy of Naïve Bayes & DT

In our data sets, all targets are continuous values. However, Naïve Bayes and Decision Tree implemented in Weka can only work for classification problem. Therefore, before using these two methods, all numeric attributes are discretized to nominal type. The average accuracy over 10-CV of these algorithms on these data sets is reported in Table 2.

KNN, NN, SVM and M5 implemented in Weka can work for regression problem. The average correlation coefficients over 10-CV of these algorithms on these data sets are reported in Table 3.

The results in Table 2 show that for Naive Bayes and Decision Tree, prediction accuracy is around 90% except Delphi Survey. Both Naive Bayes and Decision Tree are not good for Delphi Survey (only 30.0587% and 36.217% prediction accuracy that means more than half cases are not classified correctly). The highest accuracy is for Colorbond (94.728% from Naïve Bayes and 96.548% from DT). Decision Tree is a very good classification method but seems less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute. Transforming our prediction problem to classification problem by discretizing continuous values to categorical values proved not suitable on our datasets, especially for Delphi Survey.

Data set	Correlation coefficient (cc)			
	KNN	NN	SVM	M5
Delphi Survey	0.797	0.9299	0.928	0.9333
Holistic Model	0.9960	0.979	0.8412	0.9892
Colorbond	0.9962	1	0.9999	1
Maintenance for Galvanized	0.9915	0.9994	0.9737	0.9883
Maintenance for Zincalume	0.9886	0.999	0.9889	0.9971

Table 3: Correlation coefficient of KNN, NN, SVM & M5

Table 4 shows the number of classes after discretizing the target attribute. We can find that the numbers of classes for all data sets are almost same while the number of cases varies from 683 to 9640. There are 10 classes while only 683 cases in Delphi Survey. Therefore, it may be the truth that decision tree is prone to errors in classification problems with many classes and relatively small training set.

Data Set	No. of cases	No. of target classes	No. of input attribu-tes	Num-erical attrib-utes (%)	Catego-rical attribu-tes (%)
Delphi Survey	683	10	7	0%	100%
Holistic Model	9640	10	6	50%	50%
Colorbond	4780	10	13	76.92 %	23.08%
Maintenance for Galvanized	1297	10	12	91.67 %	8.33%
Maintenance for Zincalume	1297	9	6	83.33 %	16.67%

Table 4: Details of Data Sets

The results in Table 3 show that for KNN, NN, SVM and M5, very good results are achieved. Most of Correlation

coefficients (cc) are above 0.95. The lowest cc is 0.797 (KNN for Delphi Survey) and the highest is 1 (NN and M5 for Colorbond). NN works very well for all data sets, getting very high cc for all data sets. This result proves that NN is very efficient for handling numerical values and well-suited for predicting numerical target because most of attributes in our data sets are numerical values (The last two columns of Table 4 show the percentage of numerical and categorical attributes. We can find that almost all data sets have more than 50% numerical attributes). M5 is learned efficiently as NN. Especially, it is better for Delphi Survey and Holistic Model than NN.

The results from SVM are similar to NN, but reduced more for Holistic Model. The results from KNN are also similar to NN, even better for Holistic Model. But KNN got the worst result for Delphi Survey. This may prove that KNN is quite effective if the training set is large. Because there are 9640 cases in Holistic Model, 4780 cases in Colorbond, 1297 cases in maintenance while only 683 cases in Delphi Survey.

From the view of each data set, Colorbond gets the best result. The cc from all methods for Colorbond is very high (The highest reaches 1 while the lowest is also 0.9962).

The results for Delphi Survey are the worst (The highest is only 0.9333 while the lowest is 0.797).

All results indicate those methods which can deal with continuous values directly like KNN, NN, SVM and M5 are better than those that have to discretize continuous values like Naïve Bayes and DT.

However, the interesting fact is that no one method is always best for all three data sets. M5 is the best method for Delphi Survey (cc is 0.9333), KNN is the best method for Holistic Model (cc is 0.9960), NN and M5 are the best methods for Colorbond (cc is 1) and NN is the best method for Maintenance database (cc is 0.999).

In next step, we experiment bagging (Breiman, 1996) to improve the prediction performance. Further experiments were performed on the best method for each data set. Results are shown in Table 5.

Data set	Correlation coefficient	
	M5 / KNN / NN	**Bagged M5 / KNN / NN**
Delphi Survey	0.9333	0.9454
Holistic Model	0.9960	0.9967
Colorbond	1	1
Maintenance for Galvanized	0.9994	0.9997
Maintenance for Zincalume	0.999	0.9995

Table 5: Results from Bagging

From the results in Table 5, we find that bigger correlation coefficient can be obtained using bagging for M5, KNN and NN. It indicates that bagging is more accurate than the individual predictors.

So far, we have got five best models for our data sets. In order to see if the predicted service lives from different data sets are consistent, we choose some test cases as input data to produce predicted service lives from those models. Some examples of test cases are as follows:

1 | Windsor State school | Roof | Zincalume | Maintenance: Yes | Not Marine

2 | Bald Hills State School | Roof | Zincalume | Maintenance: Yes | Marine

3 | Beenleigh State School | Roof | Galvanized Steel | Maintenance: No | Not Marine

4 | Allora State School | Gutters | Galvanized Steel | Maintenance: Yes | Not Marine

5 | Calliope State School | Gutters | Colorbond | Maintenance: No | Not Marine

These test cases are in different environments (Marine or Not Marine), using different materials (Zincalume, Galvanized Steel or Colorbond) for different components (Roof or Gutters) with or without maintenance. They are selected in order to verify the predicted service lives under different conditions. The predicted service lives from different data sets for these test cases are shown in Table 6.

Because Holistic Model is only for Gutters of Galvanized Steel and Zincalume, Colorbond data set is only for Gutters of Colorbond, Maintenance data set is only for Roof of Galvanized Steel and Zincalume and Delphi is for a range of components and different materials, we compare the results of case 1, 2, 3 from Delphi and Maintenance and the results of case 4, 5 from Delphi, Holistic and Colorbond. From the above results, we found that for the first test case we got 51.877 from Delphi while only 29.928 from Maintenance. Similar contradiction happened to the fifth test case (36.64 from Delphi and 68.786 from Colorbond). For case 2,3,4, almost consistent results are achieved.

ID	**Delphi**	**Mainten-a nce**	**Holistic**	**Color-b ond**
1	51.877	29.928	N/A	N/A
2	27.185	30.449	N/A	N/A
3	35.929	26.338	N/A	N/A
4	33.151	N/A	30.951	N/A
5	36.64	N/A	N/A	68.786

Table 6: Predicted Service Life (years) for Test Cases

4.1 Existing Problems

Although no one method is always best for all three data sets, we can build independent model using the most suitable method for each data set. Sometimes a conflict exists among predicted values from three predictors for a given situation. One example is the first and the fifth test cases as shown in Table 6. There are twofold reasons for these contradictions (1) an inconsistency exists among three data sets, and (2) there exists an error during the mining process. If the first one is the case, the

inconsistencies need to be fixed with an expert opinion. The problem also arises how to choose the most appropriate answer for a given situation in case of inconsistencies.

The expert (or knowledgeable) user will have some prior knowledge to indicate the right choice. However, a naive user will not be able to make a decision depending on the result of the system.

The ideal way is to do some post-processing for the predicted result before presenting it to users. The post-processing should eliminate the conflict and select a best answer for users from multiple choices provided by multiple predictors.

Moreover, as mentioned in the data cleaning section, the data sets contains few missing values. The predictors are built based on data sets with few missing values. The user, however, may not be able to provide all inputs to get a precise answer from the use of data mining system. A method to deal with incomplete and vague queries is also required.

4.2 Possible Solutions

A knowledge base will be built in this solution. This knowledge base is a set of rules which are extracted from three predictors built already. They should identify the service life of a component using a material in a location. The framework of this solution is shown in Figure 2.

When the user queries the framework, the knowledge base is first consulted to search for matching between existing rules and user inputs. If user inputs are matched to a rule, we produce the result directly from the rule. If we can not find a matching rule, new data should be input into predictors to produce a result. Before doing this, user inputs should be pre-processed first for missing values.

Although some data mining algorithms can handle missing values automatically, for example, they replace missing values using most frequent value or average value; the ways they are using usually are not suitable for our case.

Case-based reasoning (Maher, Balachandran, & Zhang, 1995) is chosen to deal with the missing values of user inputs in this solution because the values are very close for similar cases. For example, if user only provides location and material, we can get mass loss of this material from other case using the same material and get salt deposition from other case in close location. After that, user inputs are fed into the predictors to produce the results.

To deal with the conflictions in the results of the three independent predictors and with the rules in knowledge base, post-processing of results is conducted. First we check for the consistency of the results. If they are consistent, we compare them with rules in knowledge base to see if the results are reasonable. For example, a roof in a severe marine location will not last longer than one in benign environment, and stainless steel should last longer than galvanized steel etc. We check the results to see if they match such rules. If they are not logical, adjust the results according to knowledge base. Otherwise, output the results.

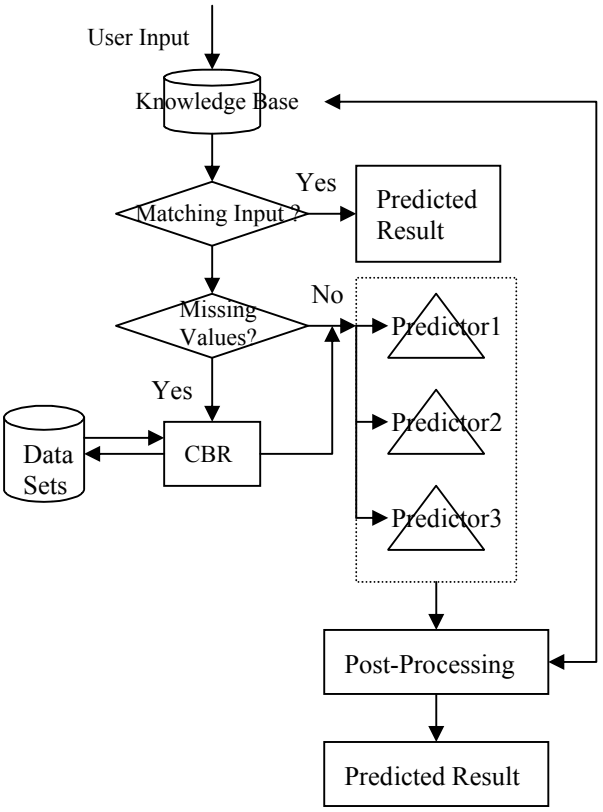


Figure 2: Framework of Solution

If the three predictors' results are inconsistent, the most reasonable (closest) result according to knowledge base is selected. In other words, the predictors' results compared with rules in knowledge base and the illogical results are deleted.

Finally, the result of each new case will be saved as a new rule in the knowledge base for later use.

The key of this solution is the construction of knowledge base. Experience and knowledge of domain experts can guide to construct the knowledge base. The cases covered by rules in knowledge base should be as many as possible. As a result, this solution can be human cooperated mining (Cao & Zhang, 2006).

5 Conclusions and Future work

Lifetime prediction of metallic components is significant in civil engineering. This paper has demonstrated that it is possible to apply data mining methods to solve this problem. We compare a number of data mining methods on the data sets provided by our industry partners and analyse what kind of methods are suitable for what kind of data.

Firstly, traditional data mining methods like Naïve Bayes, Decision Tree, Neural Network, SVM and M5 etc were applied to build a number of independent predictors for each data set. The results indicate that the best method for predicting the service life depends on the data set used to train the model. Moreover, those methods which can deal with continuous values directly like KNN, NN and M5 are better than those that have to discretize continuous values like Naïve Bayes and Decision Tree.

Further experiments for improving the performance were done on those best methods for each data set. We found the improvement to KNN, NN and M5 using bagging was obvious. We also analysed the predicted service life from each data set using certain test cases. The testing shows that in some situations, inconsistent predicted results may be presented by the data mining systems due to using three different data sets (information) for a same test case.

To solve this problem, we propose a possible solution. The key of this solution is the construction of knowledge base which contains a set of rules. When we evaluated the data mining methods, we focused on the prediction accuracy. However, if we need to extract rules from those models for building a knowledge base, we should consider the comprehensibility of those models. For example, we have found Neural Network is almost the best method for our data sets. However, it is very difficult to extract rules from Neural Network model. Moreover, what kind of rules shall we really need? It should be a case-based rule (eg. Contain many attributes like location, material etc and a service life) or if-else rules? Those questions should be answered by future research.

In summary, this study is a preliminary work. The aim of this paper was to explore various data mining methods to predict lifetime of metallic components and find a best one. Future research needs to prove the feasibility and effectiveness of the possible solution described above and develop a real lifetime prediction tool.

6 Acknowledgement

This work is partly supported by the CRC-CI project 2005-003-B funding. We would like to thank all the team members from CSIRO, Queensland Department of Public Works (QDPW) and Queensland Department of Main Roads (QDMR).

7 References

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123-140.

Cao, L., & Zhang, C. (2006). Domain-Driven Actionable Knowledge Discovery in the Real World. In *Lecture Notes in Computer Science* (3918 ed., pp. 821-830).

Cole, I., Ball, M., Carse, A., Chan, W. Y., Corrigan, P., Ganther, W., et al. (2006). *Methods for the Service Life Estimation of Metal Building Products as Applied to Selected Facilities and Bridges*.

Cole, I., Ball, M., Carse, A., Chan, W. Y., Corrigan, P., Ganther, W., et al. (March 2005). *Case-Based Reasoning in Construction and Infrastructure Projects - Final Report* (No. 2002-059-B).

Cole, I. S., Furman, S. A., & Ganther, W. D. (2001). A holistic model of atmospheric corrosion. *Elec Soc S, 2001*(22), 722-732.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1995a). Bayesian Networks for Knowledge Discovery. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 273 - 305). Menlo Park: AAAI Press.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1995b). From Data Mining to Knowledge Discovery: An Overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1 - 34). Menlo Park: AAAI Press.

Furuta, H., Deguchi, T., & Kushida, M. (1995). *Neural network analysis of structural damage due to corrosion.* Paper presented at the Proceedings of ISUMA - NAFIPS '95 The Third International Symposium on Uncertainty Modeling and Analysis and Annual Conference of the North American Fuzzy Information Processing Society.

Kessler, W., Kessler, R. W., Kraus, M., Kubler, R., & Weinberger, K. (1994). *Improved prediction of the corrosion behaviour of car body steel using a Kohonen self organising map.* Paper presented at the Advances in Neural Networks for Control and Systems, IEE Colloquium on.

Maher, M. L., Balachandran, M. B., & Zhang, D. M. (1995). *Case-Based Reasoning in Design*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.

Melhem, H. G., & Cheng, Y. (2003). Prediction of remaining service life of bridge decks using machine learning. *Journal of Computing in Civil Engineering, 17*(1), 1-9.

Melhem, H. G., Cheng, Y., Kossler, D., & Scherschligt, D. (2003). Wrapper Methods for Inductive Learning: Example Application to Bridge Decks. *Journal of Computing in Civil Engineering, 17*(1), 46-57.

Morcous, G., Rivard, H., & Hanna, A. M. (2002). Case-Based Reasoning System for Modeling Infrastructure Deterioration. *Journal of Computing in Civil Engineering, 16*(2), 104-114.

Olafsson, S. (2006). Introduction to operations research and data mining. *Computers & Operations Research*

Part Special Issue: Operations Research and Data Mining, 33(11), 3067-3069.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106.

Quinlan, J. R. (1992). *Learning with Continuous Classes.* Paper presented at the 5th Australian Joint Conference on Artificial Intelligence.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Zhang, G., Eddy Patuwo, B., & Y. Hu, M. (1998). Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting, 14*(1), 35-62.

Integrated Scoring for Spelling Error Correction, Abbreviation Expansion and Case Restoration in Dirty Text

Wilson Wong, Wei Liu and Mohammed Bennamoun

School of Computer Science and Software Engineering
University of Western Australia
Crawley WA 6009
Email: {wilson,wei,bennamou}@csse.uwa.edu.au

Abstract

An increasing number of language and speech applications are gearing towards the use of texts from online sources as input. Despite such rise, not much work can be found in the aspect of integrated approaches for cleaning dirty texts from online sources. This paper presents a mechanism of **I**ntegrated **S**coring for **S**pelling error correction, **A**bbreviation expansion and **C**ase restoration (ISSAC). The idea of ISSAC was first conceived as part of the text preprocessing phase in an ontology engineering project. Evaluations of ISSAC using 400 chat records reveal an improved accuracy of 96.5% over the existing 74.4% based on the use of Aspell only.

Keywords: Spelling error correction, abbreviation expansion, case restoration, dirty text, text preprocessing, text cleaning

1 Introduction

The tasks of correcting spelling errors, expanding abbreviations and restoring cases are essential to many language and speech applications such as text and data mining (Castellanos 2003, Tang, Li, Cao & Tang 2005), and automatic or semi-automatic ontology engineering (Maedche & Volz 2001, Xu, Kurz, Piskorski & Schmeier 2002, Degeratu & Hatzivassiloglou 2002, Novacek & Smrz 2005*b*). These tasks are typically performed as part of the text preprocessing phase in these applications, and are usually known by other names such as *text cleaning* and *text normalization*. Despite the importance of these cleaning tasks, the existing applications have been relying on ad-hoc techniques (e.g. pattern matching, handcrafted rules) designed to serve individual needs when problems arise.

Text preprocessing, especially spelling error correction, abbreviation expansion and case restoration, is beginning to attract the attention of language and speech applications as they gear towards using online sources for textual input. Examples include corporate intranet (Kietz, Volz & Maedche 2000) and documents retrieved by search engines (Cimiano & Staab 2005, Sanchez & Moreno 2005, Sombatsrisomboon, Matsuo & Ishizuka 2003, Turney 2001) for automatic or semi-automatic ontology engineering, and chat records (Castellanos 2003) and emails (Tang et al. 2005) for text and data mining. The quality of texts from such sources, in particular blogs, emails and chat records can be extremely poor. The constructions of sentences in blogs, emails and chat records are filled with spelling errors, ad-hoc abbreviations and improper casing. As the different quality of texts will pose different requirements during the preprocessing phase, dirty texts can be very demanding. With the prevalence of online sources, this *"...annoying phase of text cleaning..."*(Mikheev 2002) becomes much more important and relevant than ever before. Accordingly, an increasing number of researchers, particularly in the field of ontology engineering (Maedche & Volz 2001, Novacek & Smrz 2005*b*, Novacek & Smrz 2005*a*), has began to acknowledge the impact of the cleanliness of input on the quality of ontology. In a text mining research, Tang et al. (Tang et al. 2005) reported an improved terms extraction accuracy of 38-45% (based on F1-measure) after the input (i.e. emails) has been cleaned.

In this paper, we propose a mechanism for **I**ntegrated **S**coring for **S**pelling error correction, **A**bbreviation expansion and **C**ase restoration (ISSAC). ISSAC is built on top of the spell checker Aspell (Atkinson 2006) for automatically correcting spelling errors, expanding ad-hoc abbreviations and restoring case in dirty texts (e.g. chat records). ISSAC makes use of six weights based on different information sources, namely, original rank by Aspell, reuse factor, abbreviation factor, normalized edit distance, domain significance and general significance. Evaluations performed on four different set of chat records yield an average of 96.5% accuracy in the automatic correction of spelling errors, expansion of ad-hoc abbreviations and restoration of casing.

2 Background and Related Work

Spelling error detection and correction is the task of recognizing misspellings in texts and providing suggestions for correcting the errors. For example, detecting *"cta"* as an error and suggesting that the error be replaced with *"cat"*, *"act"* or *"tac"*. Hence, it is obvious that suggestions can only be made after detecting the errors, and more information is usually required to perform a correct replacement. There are four basic types of single-error misspelling (Chan, He & Ounis 2005, Damerau 1964), namely, insertion (e.g *"receivee"* with an extra *'e'*), deletion (e.g. *"receiv"* with the missing *'e'*), substitution (e.g. *"receiva"* where *'a'* is supposed to be *'e'*) and transposition (e.g. *"recieve"* where *'i'* and *'e'* switched). The task of abbreviation expansion deals with recognizing shorter forms of words (e.g. *"abbr."* or *"abbrev."*), acronyms (e.g. *"NATO"*) and initialisms (e.g. *"HTML"*, *"FBI"*), and expanding them to their corresponding words[1]. The many-to-many relation-

[1]Some researchers refer to this relationship as *abbreviation and definition* or *short-form and long-form*

ship between abbreviations and their corresponding expansions makes this task equally difficult. For case restoration, improper casing in words are detected and restored. For example, detecting the letter *'j'* in *"jones"* as improper and correcting the word to produce *"Jones"*.

Most of the work in detection and correction of spelling errors, and expansion of abbreviations are carried out separately. The single-error misspellings, together with other more complex errors (Kukich 1992), have given rise to and shaped a wide range of techniques since 1960s. Two of the most studied classes of techniques for detecting and correcting spelling errors are *minimum edit distance* and *similarity key*. The idea of minimum edit distance techniques began with Damerau (Damerau 1964) and Levenshtein (Levenshtein 1966). Damerau-Levenshtein distance is the minimal number of insertions, deletions, substitutions and transpositions needed to transform one string into the other. For example, to change *"wear"* to *"beard"* will require a minimum of two operations, namely, a substitution of *'w'* with *'b'*, and an insertion of *'d'*. Many variants were developed subsequently such as the algorithm by Wagner & Fischer (Wagner & Fischer 1974). The second class of techniques is similarity key. The main idea behind similarity key techniques is to map every string into a key such that similarly spelled strings will have identical keys (Kukich 1992). Hence, the key, computed for each spelling error, will act as a pointer to all similarly spelled words (i.e. suggestions) in the dictionary. One of the earliest implementation is the SOUNDEX system by Odell & Russell (Odell & Russell 1918, Odell & Russell 1922). SOUNDEX maps a word into a key consisting of its first letter followed by a sequence of numbers. For every of the remaining letter l, a number is assigned according to the rules:

0 if $l \in \{A, E, I, O, U, H, W, Y\}$

1 if $l \in \{B, F, P, V\}$

2 if $l \in \{C, G, J, K, Q, S, X, Z\}$

3 if $l \in \{D, T\}$

4 if $l \in \{L\}$

5 if $l \in \{M, N\}$

6 if $l \in \{R\}$

Zeros are eliminated and repeated numbers are collapsed. For example, $wear \rightarrow w006 \rightarrow w6$ and $ware \rightarrow w060 \rightarrow w6$. Since then, many improved variants were developed such as the Metaphone and the Double-metaphone algorithm by Philips (Philips 1990), Daitch-Mokotoff Soundex (Lait & Randell 1993) and others (Holmes & McCabe 2002).

Many work on detecting and expanding abbreviations are conducted in the realm of named-entity recognition and word-sense disambiguation. Due to the intensive use of abbreviations in medical texts, most researches were initiated and tested in the medical domain. Schwartz & Hearst (Schwartz & Hearst 2003) presented a simple two-stage process for extracting abbreviations and their definitions. The first stage involves the extraction of all abbreviations and definition candidates based on the adjacency to parentheses. Stage two identifies the correct definition out of the many candidates for each abbreviation. The identification is based on two criteria: the correct definition must appear in the same sentence, and the correct definition should have no more than $min(|A|+5, |A|*2)$ words where $|A|$ is the number of characters in an abbreviation A. Park & Byrd (Park & Byrd 2001) presented an algorithm based on rules and heuristics for extracting definitions for abbreviations from texts. Some of the rules and heuristics employed for this purpose include syntactic cues, priority of rules, distance between abbreviation and definition, word casing, and the number of words and stopwords in the definition. Pakhomov (Pakhomov 2001) proposed a semi-supervised approach that employs a hand-crafted table of abbreviations and their definitions for training a maximum entropy classifier. Last but not least, Sproat et al. (Sproat, Black, Chen, Kumar, Ostendorf & Richards 2001) have undertaken a project that attempts to address deficiencies in existing aproaches for various aspects in abbreviation expansion. In particular, the project focuses on the difficulty of normalizing text from online sources such as newsgroups.

In the context of ontology engineering and other related areas such as text mining, spelling errors correction, abbreviations expansion and case restoration (Mikheev 2002, Lita, Ittycheriah, Roukos & Kambhatla 2003, Mikheev 1999) are mainly carried out as part of the text preprocessing (i.e. text cleaning, text filtering, text normalization) phase. A review by Wong et al. (Wong, Liu & Bennamoun 2006) shows that many existing systems require the input texts to be clean and hence, the techniques for extracting plain text from various format, correcting spelling errors and expanding abbreviations becomes unnecessary. Ontology engineering approaches such as Xu et al. (Xu et al. 2002), Text-to-Onto (Maedche & Volz 2001), OntoStruct (Degeratu & Hatzivassiloglou 2002) and OLE/BOLE (Novacek & Smrz 2005*b*, Novacek & Smrz 2005*a*) are the few exceptions. In a text mining approach for extracting topics from chat records, Castellanos (Castellanos 2003) presented a very comprehensive list of techniques for text preprocessing. In addition to the common text preprocessing tasks, the approach employs a thesaurus, constructed using the Smith-Waterman algorithm (Smith & Waterman 1981), for correcting spelling errors and identifying abbreviations. Tang et al. (Tang et al. 2005) presented a cascaded approach for cleaning emails prior to any text mining processing. The approach is composed of four passes including non-text filtering, paragraph normalization, sentence normalization, and word normalization.

3 Scoring Mechanism

The idea of ISSAC was initially conceived as part of an ontology engineering project that uses multiple online sources (i.e. media articles and chat records) with varying cleanliness. In particular, the use of chat records as input has required us to place more effort during the text preprocessing phase. Figure 1 highlights the various spelling errors, ad-hoc abbreviations and improper casing that occur very frequently in chat records which are not present in clean text. ISSAC provides an integrated approach for simultaneously correcting spelling errors, expanding abbreviations and restoring cases without expert participation. Along the same line of thought, Clark (Clark 2003) defended that *"...a unified tool is appropriate because of certain specific sorts of errors"*. To illustrate this idea, consider the error word *"cta"*. Do we immediately take it as a spelling error and correct it as *"cat"*, or is it a problem with the letter casing, which makes it a probable acronym? Hence, it is obvious that the problems of spelling error, abbreviation and letter casing are inter-related to a certain extent. ISSAC provides an integrated approach for solving the three problems (i.e. spelling error correction, abbreviation expansion and case restoration) in the following ways:

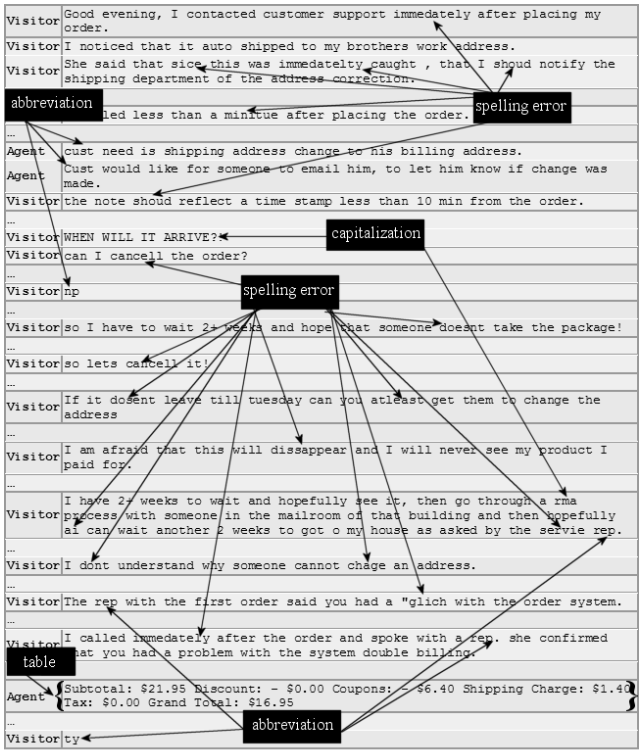


Figure 1: Example of spelling errors, ad-hoc abbreviations and improper casing in a chat record

- Spelling error correction: The foundation for correcting spelling errors is the spell checker Aspell (Atkinson 2006). Whenever a word is considered as an error, Aspell will provide a list of suggestions for correction. This list is the key component of ISSAC. The ability of ISSAC to find the correct replacement will be dependent on the quality of the suggestions.
- Abbreviation expansion: The foundation for expanding abbreviations is the online abbreviation dictionary `www.stands4.com`. When the scoring process takes place and the corresponding expansions for potential abbreviations are required, `www.stands4.com` is consulted. A copy of the expansion is stored in a local *abbreviation dictionary* for future reference. The expansions for potential abbreviations will be added to the suggestion list produced by Aspell. Later, the various weights in ISSAC will help in determining if the error (i.e. potential abbreviation) is an actual abbreviation and that replacement should be done using the corresponding expansion.
- Case restoration: There are two ways of restoring cases using ISSAC. The first employs the inherent capability of Aspell to recognize proper nouns without appropriate casing as errors. This will allow such words to be thrown into ISSAC for restoration. The second way is based on the heuristic that valid words rarely appear as acronyms, and those who do will not fit into the surrounding context. For example, consider the phrase with improper casing, *"shipping TIME frame"*. Appearing as an independent word, *"TIME"* has an equally likely chance of being a word (with improper casing) or an acronym for *"Timed Interactive Multimedia Extensions"*[2]. When the neighbouring words *"shipping"* and *"frame"* are taken into considerations, then the probability of *"TIME"* being an acronym becomes significantly less. Based on this idea, words with all uppercase letters are first turned into lowercase and then automatically disambiguated using ISSAC.

[2]Source from `http://www.stands4.com/bs.asp?st=time&SE=1`

Prior to the scoring, each sentence in the input text (e.g. chat record) is tokenized to obtain a list of words $T = \{t_1, ...t_w\}$ which will be fed into Aspell. For each word e that Aspell considers as erroneous, a list of ranked suggestions S is produced. The list S is generated by Aspell based on the Metaphone algorithm (Philips 1990) and near-miss strategy by its predecessor Ispell (Kuenning 2006). Initially, $S = \{s_{1,1}, ..., s_{n,n}\}$ is an ordered list of n suggestions where $s_{j,i}$ is the j^{th} suggestion with rank i (smaller i indicates higher rank). If e appears in the abbreviation dictionary, the list S is augmented by appending all the corresponding m expansions at the front of S as additional suggestions, all with rank 1. In addition, the error word e is appended at the end of S with rank $n+1$. These augmentations result in an extended list $S = \{s_{1,1}, ..., s_{m,1}, s_{m+1,1}, ..., s_{m+n,n}, s_{m+n+1,n+1}\}$, which is a combination of m suggestions from the abbreviation dictionary (if e is a potential abbreviation), n suggestions by Aspell, and the error word e itself. Placing the error word e back into the list of possible replacements serves one purpose: to ensure that if no better replacement is available, we keep the error word e as it is. Once the extended list S is obtained, each suggestion $s_{j,i}$ is re-ranked using ISSAC based on a combination of weights, including the existing rank i. The other weights include the reuse factor RF, abbreviation factor AF, normalized edit distance NED, domain significance DS and general significance GS. The attempt to measure the significance of suggestions also takes into account the neighbouring words l (i.e. word to the left) and r (i.e. word to the right) of error e. This is based on the assumption that *"...errors are isolated and surrounded by clean context that can be used to correct the errors"* (Clark 2003). The new score for each suggestion $s_{j,i}$ is defined as

$$\begin{aligned} NS(s_{j,i}) &= i^{-1} + NED(e, s_{j,i}) \\ &\quad + RF(e, s_{j,i}) + AF(s_{j,i}) \\ &\quad + DS(l, s_{j,i}, r) + GS(l, s_{j,i}, r) \end{aligned}$$

The different weights are defined in the following subsections.

3.1 Normalized Edit Distance

$NED(e, s_{j,i}) \in (0, 1]$ is the *normalized edit distance* obtained by normalizing the value of the minimum edit distance between error e and suggestion $s_{j,i}$. The normalized edit distance is defined as

$$NED(e, s_{j,i}) = \frac{1}{ED(e, s_{j,i}) + 1}$$

The minimum edit distance ED between two words is obtained using the Wagner-Fischer algorithm (Wagner & Fischer 1974). Wagner-Fischer distance is preferred over other metrics because it covers multi-error misspellings (Kukich 1992) and also accounts for transpositions in addition to insertions, deletions and substitutions. *NED* provides higher importance to suggestions that are lexically similar to the error. During the evaluation, this weight has shown to be useful especially for maintaining proper nouns that

Aspell considers as errors. For example, Aspell mistakenly identified *"Carrollton"*[3] as an error and suggested the replacement *"Carillon"*. Similarly, Aspell suggested that *"iPod"* be replaced with *"pod"*.

3.2 Reuse Factor and Abbreviation Factor

$RF(e, s_{j,i}) \in \{0, 1\}$ is the boolean *reuse factor* for providing more weight to suggestion $s_{j,i}$ that has been previously used for correcting error e. This weight is to ensure consistency for correcting similar errors. Certain spelling errors in chat records are repetitive in nature. For example, the word *"received"* and *"receive"* are often misspelled as *"recieved"* and *"recieve"* respectively. Out of the 2016 spellings errors in an evaluation of 400 chat records, there are about 40 occurrences of *"recieve"* and its variants. The reuse factor is obtained through a lookup into a *history list* that consists of previous corrections. $RF(e, s_{j,i})$ will provide factor 1 if the error e has been previously corrected with $s_{j,i}$ and 0 otherwise.

$AF(s_{j,i}) \in \{0, 1\}$ is the *abbreviation factor* for denoting that $s_{j,i}$ is a potential abbreviation. A lookup into the *abbreviation dictionary*, $AF(s_{j,i})$ will yield factor 1 if suggestion $s_{j,i}$ is found to exists in the dictionary and 0 otherwise. The abbreviation dictionary is automatically updated on demand using `www.stands4.com`.

3.3 Domain Significance

$DS(l, s_{j,i}, r) \in [0, 1]$ measures the *domain significance* of suggestion $s_{j,i}$ based on its appearance in the domain corpora. The domain significance weight is inspired by the *TF-IDF* (Robertson 2004) measure which is commonly used in Information Retrieval. In addition to individual occurrence, the frequency of appearance of $s_{j,i}$ together with its neighbouring words l and r is also taken into consideration. The weight is defined as

$$DS(l, s_{j,i}, r) = \frac{f_{s_{j,i}} + f_{ls_{j,i}r}}{\sum_{k=1}^{m+n+1}(f_{s_{k,i}} + f_{ls_{k,i}r})} e^{-\frac{ND_{s_{j,i}}}{ND}}$$

where $f_{s_{j,i}}$ is the frequency of occurrence of suggestion $s_{j,i}$ in the domain corpora and $f_{ls_{j,i}r}$ is the frequency of occurrence of suggestion $s_{j,i}$ together with neighbours l and r in the domain corpora. $\sum_{k=1}^{m+n+1}(f_{s_{k,i}} + f_{ls_{k,i}r})$ is the sum of the frequencies of occurrences of all individual suggestions, and of the frequencies of occurrences of all suggestions together with neighbours l and r in the domain corpora. ND is the total number of documents in the domain corpora and $ND_{s_{j,i}}$ is the number of documents in the domain corpora that contain suggestion $s_{j,i}$. The significance of $s_{j,i}$ will increase proportionally to the number of times the suggestion appears in the domain corpora. Raising e to the power of $-\frac{ND_{s_{j,i}}}{ND}$ has similar effect as the traditional *IDF* (Robertson 2004), namely, as an offset to suggestions that occur too frequently.

3.4 General Significance

$GS(l, s_{j,i}, r) \in [0, 1]$ measures the *general significance* of suggestion $s_{j,i}$ based on its appearance in the general collection (e.g. Goggle). The purpose of general significance is similar to that of the domain significance. The weight is defined as

$$GS(l, s_{j,i}, r) = \frac{NG_{ls_{j,i}r}}{\sum_{k=1}^{m+n+1} NG_{ls_{k,i}r}} e^{-\frac{NG_{s_{j,i}}}{\sum_{k=1}^{m+n+1} NG_{s_{k,i}}}}$$

where $NG_{ls_{j,i}r}$ is the number of documents in the general collection that contains suggestion $s_{j,i}$ within the neighbours l and r, and $NG_{s_{j,i}}$ is the number of documents in the general collection that contains suggestion $s_{j,i}$ alone. This weight is especially useful for two reasons. Firstly, this weight will provide due consideration for the use of proper names such as *"iPod"* and *"XBox"* that are not part of Aspell's dictionary. Secondly, the contemporary use of English in areas like business and computing has given rise to various entirely new words that results from combinations of existing ones. General collection of documents such as `www.google.com` is the best candidate when considering the spelling for new words.

4 Evaluation and Results

Evaluations are conducted using chat records provided by `247Customer.com`[4]. As a provider of customer lifecycle management services, the chat records by `247Customer.com` offer a rich source of domain information in a natural setting (i.e. conversations between customers and agents). Consequently, these chat records are filled with spelling errors, ad-hoc abbreviations, improper casing and many other problems that are considered as intolerable by many of the existing language and speech applications. Consequently, these chat records become the ideal source for evaluating the scoring mechanism presented in this paper. Four sets of test data, each comes in an XML file of 100 chat records, are employed for evaluations. Each XML file has an average of 10,000 words. The chat records and the Google search engine constitutes the domain corpora and general collection respectively while GNU Aspell version 0.60.4 (Atkinson 2006) is employed for detecting errors and generating suggestions. Four evaluations are performed, one for each set based on the steps described in Algorithm 1.

Determining whether $cISSAC_{u,r}$ and $cAspell_{u,r}$ is a correct replacement for $error_{u,r}$ is a delicate process that must be performed manually. To illustrate, consider the error *"itme"*. It is difficult to automatically determine whether *"itme"* should be replaced with *"time"* or *"item"*. Without neighbouring words, both replacements are of equally likely nature. Appearing as an error (*"shipping itme frame"*) in the first evaluation, Aspell's first choice for $error_{1,r} = itme$ is $cAspell_{1,r} = item$, while ISSAC ranked $cISSAC_{1,r} = time$ as the replacement. It is only proper for us to rate the replacement by Aspell as wrong given the error's appearance in the context of *"shipping"* and *"frame"*. In another error (*"Chad amateau <"* where $<$ is the end-of-sentence character) in the third evaluation, Aspell suggested *"amateur"* as the most ideal replacement while ISSAC suggested *"Amateau"*. As there is no such word as *"Amateau"* in the dictionary, we would be tempted to immediately rate the suggestion by Aspell as correct if we did not take into consideration the fact that *"Chad Amateau"* is a proper name.

The evaluation of the errors and replacements are conducted in an integrated manner. The errors are not classified into spelling errors, ad-hoc abbreviations and improper casing. For example, should the error *"az"* (*"AZ"* is the abbreviation for the state of

[3] A city in the state of Texas. Source from `http://www.ci.carrollton.tx.us/`

[4] `http://www.247customer.com/`

Algorithm 1 Evaluation of ISSAC

1: **input** four sets of chat records CR_1, CR_2, CR_3, CR_4
2: **for** each set of chat records CR_u **do**
3: **initialize** EVA_u, an array of triplets ($error_{u,r}$,$cISSAC_{u,r}$,$cAspell_{u,r}$) where $error_{u,r}$ is the r^{th} error in the u^{th} evaluation, $cISSAC_{u,r}$ is the correction proposed by ISSAC for the r^{th} error in the u^{th} evaluation, and $cAspell_{u,r}$ is the first suggestion proposed by Aspell for the r^{th} error in the u^{th} evaluation
4: **for** each sentence in CR_u **do**
5: Tokenize the sentence to produce a set of words $T = \{t_1, ..., t_w\}$
6: **for** each word $t \in T$ **do**
7: **if** t consists of all uppercase **then**
8: Turn all letters in t to lowercase
9: **else if** t consists of all digits **then**
10: continue with next term
11: **end if**
12: Feed t to Aspell
13: **if** t is identified as error by Aspell **then**
14: **initialize** NSC, an array of new scores for all suggestions for error word t
15: A set of n suggestions for word t, $S = \{s_{1,1}, ..., s_{n,n}\}$ is generated by Aspell
16: Append error t at the end of S
17: Perform lookup in the *abbreviation dictionary* and retrieve all corresponding m expansions for t
18: Append the m expansions at the front of S
19: The additional suggestions will produce an extended list $S = \{s_{1,1}, ..., s_{m,1}, s_{m+1,1}, ..., s_{m+n,n}, s_{m+n+1,n+1}\}$
20: **for** each suggestion $s_{j,i} \in S$ **do**
21: Execute ISSAC to obtain the new score $NS(s_{j,i})$
22: Push $NS(s_{j,i})$ into NSC
23: **end for**
24: Sort NSC in descending order
25: Form the triplets (t,NSC_1,$s_{m+1,1}$) where NSC_1 is the first element in the sorted NSC (i.e. the replacement proposed by ISSAC) and $s_{m+1,1}$ is the first suggestion by Aspell
26: Push the triplets (t,NSC_1,$s_{m+1,1}$) into the array EVA_u
27: **else**
28: continue with next term
29: **end if**
30: **end for**
31: **end for**
32: **for** each triplets ($error_{u,r}$, $cISSAC_{u,r}$, $cAspell_{u,r}$) in EVA_u **do**
33: **if** $cISSAC_{u,r}$ is the correct replacement for $error_{u,r}$ **then**
34: Rate $cISSAC_{u,r}$ as 1
35: **else**
36: Rate $cISSAC_{u,r}$ as 0
37: **end if**
38: **if** $cAspell_{u,r}$ is the correct replacement for $error_{u,r}$ **then**
39: Rate $cAspell_{u,r}$ as 1
40: **else**
41: Rate $cAspell_{u,r}$ as 0
42: **end if**
43: **end for**
44: **end for**
45: Count the number of $cISSAC_{u,r}$ and $cAspell_{u,r}$ rated as 1 for all the four evaluations $EVA_{u=1}$, $EVA_{u=2}$, $EVA_{u=3}$ and $EVA_{u=4}$

Table 1. Accuracy of Aspell and ISSAC across four evaluations

	Evaluation 1, $EVA_{u=1}$	Evaluation 2, $EVA_{u=2}$	Evaluation 3, $EVA_{u=3}$	Evaluation 4, $EVA_{u=4}$	Average
number of correct replacements using ISSAC, $cISSAC_{u,r}=1$	97.06%	97.07%	95.92%	96.20%	96.56%
number of correct replacements using Aspell, $cAspell_{u,r}=1$	74.61%	75.94%	71.81%	75.19%	74.39%

"Arizona") in the context of *"Glendale az <"* be considered as an abbreviation or improper casing? The boundaries between the different types of dirtiness that occur in real-world texts, especially those from online sources, are not clear. This is the main reason behind the increasing number of efforts that attempt to provide techniques to handle various dirtiness in an integrated manner (Tang et al. 2005, Mikheev 2002, Sproat et al. 2001, Clark 2003). After a careful evaluation of all replacements suggested by Aspell and by ISSAC for all 2016 errors, we discovered a promising improvement in accuracy using the latter. The accuracy is obtained by dividing the number of errors with correct replacement by the total number of errors identified by Aspell. As shown in Table 1, the use of the first suggestion by Aspell as replacement yields an average of 74.4%. With the addition of the various weights that form ISSAC, an average increase of 22% was achieved, resulting to an improved accuracy of 96.5%.

5 Discussion

The list of suggestions and the initial ranks provided by Aspell are integral parts of ISSAC. The achievement of an average of 74.4% accuracy by Aspell itself, given the extremely poor nature of the texts shows the existing strength of the Metaphone algorithm and near-miss strategy. The further average increase of 22% in accuracy demonstrates the potential of the combined weights with regard to spelling error correction and other related areas. In the course of analyzing the remaining 3.5% of errors which have been wrongfully replaced, we have discovered several interesting points as explained below.

Firstly, half of the errors with wrong corrections are actually manifestations of certain inadequacies that ISSAC has inherited from Aspell. In other words, the accuracy of correction by ISSAC is bounded by the coverage of the list of suggestions S produced by Aspell. About 2% of wrong replacements is due to the absence of the correct replacement from the list of suggestions produced by Aspell. For example, the error *"dotn"* in the context of *"i dotn have"* was wrongfully replaced by both Aspell and ISSAC as *"do-tn"* and *"do tn"* respectively. After a look into the evaluation log, we realized that the correct replacement *"don't"* was not in S. In another case, error *"everyhitng"* (as in *"over everyhitng again"*) was wrongfully replaced with *"overhung"* and *"everyhitng"* by Aspell and ISSAC respectively. In such cases, there is no way for ISSAC to propose the correct replacement except that error e is an abbreviation. If e is an abbreviation, then the correct replacement (i.e. expansion) would have made its way into the extended list S as one of the first m suggestions $\{s_1, ..., s_m\}$ (i.e. all possible expansions for potential abbreviation e).

Secondly, the use of the two immediate neighbouring words l and r to inject more contextual consideration into domain and general significance has con-

tributed to the large portion of the increase in accuracy. This claim is based on the result of a separate evaluation similar to those presented in the previous section with one exception: we omit the domain DS and general GS significance from ISSAC. A stunning drop of accuracy was observed, with an average of only 77%. Despite the contribution of l and r to the overall performance of ISSAC, it is by no means foolproof. About 1% out of the total errors with wrong replacement are due to two flaws related to l and r.

- In the first one, the neighbouring words themselves are not correctly spelled. For example, the error *"iberal"* (in the context of *"morel iberal return"*) is incorrectly replaced by both Aspell and ISSAC. This is due to the low values of DS and GS which fail to capture the actual significance of the correct replacement (i.e. *"liberal"*) with respect to the erroneous left word *"morel"*. Nonetheless, as shown by the low percentage of such flaw, this problem is not drastic. An error *"gto"* (in the context of *"lookin gto buy"*) was correctly replaced with *"to"* by ISSAC even though the left word *"lookin"* is erroneous.

- In the second case, the left and right words are inadequate. This is especially true when the errors to be corrected are located at the start and end of sentence. For example, there are no left and right words for the error *"winsted"* in the context of *"> winsted <"*. Such phenomena are the same as not using contextual information when attempting to correct an error. In such cases, the most popular suggestion $s_{j,i} \in S$ in both domain and general collection will triumph. Even with one or both neighbouring words present, incorrect replacement is also possible due to the indiscriminative nature of the neighbours. In the example *"both ocats <"*, the left word *"both"* does not provide much clue as to adequately discriminate between suggestions such as *"coats"*, *"cats"* and *"acts"*. Such neighbouring words are in sharp contrast to better ones such as *"wood"* as in *"the mindi wood"*.

Lastly, the remaining 0.5% can be seen as anomalies where ISSAC does not apply. There are two cases for these anomalies:

- Similar to throwing a dice, the first group of anomalies is characterized by the equally likely nature of all the possible replacements. For example, in the context *"Janice cheung <"*, the left word is correctly spelled and has adequately confined the suggestions to proper names even though the right word is absent. In addition, the correct replacement *"Cheung"* is present as a suggestion $s_{j,i} \in S$. Despite all these, both Aspell and ISSAC decided to replace *"cheung"* with *"Cheng"*. A look into the evaluation log reveals that the surname *"Cheung"* is as common as *"Cheng"*. In such cases, the probability of replacing e with the correct replacement is c^{-1} where c is the number of suggestions with approximately same $NS(s_{j,i})$.

- The second group of anomalies are due to contrasting value of certain weights, especially NED and i^{-1}, that causes wrong replacements to be made. For example, in the case *"cannot chage an"*, ISSAC replaced the error *"chage"* with *"charge"* instead of *"change"*. All the other weights for *"change"* are comparatively higher (i.e. DS and GS) or the same (i.e. RF, NED and AF) as *"charge"*. Such inclination indicates that *"change"* is the most proper replacement given the various cues. Nonetheless, the original rank by Aspell for *"charge"* is *i=1* while *"change"* is *i=6*. As smaller i indicates higher rank, the inverse of the original rank by Aspell i^{-1} results in the plummeting of the combined weight for *"change"*.

6 Conclusion and Future Work

As more and more language and speech applications open up to the use of online sources, the need to handle dirty texts becomes inevitable. Regardless of whether we acknowledge this fact, the quality of output and the proper functioning of such applications are, to a certain extent, dependent on the cleanliness of the input texts. Most of the existing techniques for correcting spelling errors, expanding abbreviations and restoring cases are studied separately. We, along with an increasing number of researchers, have acknowledged the fact that many errors in texts are composite in nature. As we have demonstrated during our evaluation and discussion in this paper, many errors are difficult to be classified as either spelling errors, ad-hoc abbreviations or improper casing. In this paper, we presented ISSAC, an integrated scoring mechanism that builds upon the famous spell checker Aspell for simultaneously providing solution to spelling errors, abbreviations and improper casing. ISSAC combines weights based on various information sources, namely, original rank by Aspell, reuse factor, abbreviation factor, normalized edit distance, domain significance and general significance. Four evaluations conducted over 40,000 words have demonstrated a promising accuracy at an average of 96.5%.

Even though the idea for ISSAC was first motivated and conceived within the paradigm of ontology engineering, we see great potential in further improvements and fine-tuning for a wide range of uses, especially in language and speech applications. We hope that an integrated approach such as ISSAC will pave the way for more research in providing a complete solution for text preprocessing (i.e. text cleaning) in general. At this moment, the various factors and weights are considered as equally important in this version of ISSAC. Depending on the applications or dirtiness of the texts, changes to the existing weights or even adding new weights may be necessary. We are planning to vary the importance of the different weights and fine-tune ISSAC to possibly improve the remaining 3.5% flaws pointed out in the discussion section. In addition, we have plans to extend the evaluation of ISSAC using other types of data such as news articles from online media sites. Last but not least, the assessment of the scalability of ISSAC in terms of runtime and accuracy using much larger datasets will also be explored.

Acknowledgement

This research was supported by the Australian Endeavour International Postgraduate Research Scholarship, and the Research Grant 2006 by the University of Western Australia. The authors would like to thank `247Customer.com` for providing the evaluation data. Gratitude to the developer of GNU Aspell, Kevin Atkinson.

References

Atkinson, K. (2006), 'Gnu aspell 0.60.4', http://aspell.sourceforge.net/.

Castellanos, M. (2003), Hotminer: Discovering hot topics on the web, *in* M. Berry, ed., 'Survey of Text Mining', Springer-Verlag.

Chan, S., He, B. & Ounis, I. (2005), An in-depth survey on the automatic detection and correction of spelling mistakes, *in* 'Proceedings of the 5th Dutch-Belgian Information Retrieval Workshop (DIR)'.

Cimiano, P. & Staab, S. (2005), Learning concept hierarchies from text with a guided agglomerative clustering algorithm, *in* 'Proceedings of the Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods', Bonn, Germany.

Clark, A. (2003), Pre-processing very noisy text, *in* 'Proceedings of the Workshop on Shallow Processing of Large Corpora at Corpus Linguistics'.

Damerau, F. (1964), 'A technique for computer detection and correction of spelling errors', *Communications of the ACM* **7**(3), 171–176.

Degeratu, M. & Hatzivassiloglou, V. (2002), Building automatically a business registration ontology, *in* 'Proceedings of the ACM Annual National Conference on Digital Government Research'.

Holmes, D. & McCabe, C. (2002), Improving precision and recall for soundex retrieval, *in* 'Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC)'.

Kietz, J., Volz, R. & Maedche, A. (2000), Extracting a domain-specific ontology from a corporate intranet, *in* 'Proceedings of the 4th Conference on Computational Natural Language Learning', Lisbon, Portugal.

Kuenning, G. (2006), 'International ispell 3.3.02', http://ficus-www.cs.ucla.edu/geoff/ispell.html.

Kukich, K. (1992), 'Technique for automatically correcting words in text', *ACM Computing Surveys* **24**(4), 377– 439.

Lait, A. & Randell, B. (1993), An assessment of name matching algorithms, Technical report, University of Newcastle upon Tyne.

Levenshtein, V. (1966), 'Binary codes capable of correcting deletions, insertions, and reversals', *Soviet Physics Doklady* **10**(8), 707–710.

Lita, L., Ittycheriah, A., Roukos, S. & Kambhatla, N. (2003), truecasing, *in* 'Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics', Japan.

Maedche, A. & Volz, R. (2001), The ontology extraction & maintenance framework: Text-to-onto, *in* 'Proceedings of the IEEE International Conference on Data Mining', California, USA.

Mikheev, A. (1999), A knowledge-free method for capitalized word disambiguation, *in* 'Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics'.

Mikheev, A. (2002), 'Periods, capitalized words, etc.', *Computational Linguistics* **28**(3), 289–318.

Novacek, V. & Smrz, P. (2005*a*), Bole - a new bio-ontology learning platform, *in* 'Proceedings of the Workshop on Biomedical Ontologies and Text Processing'.

Novacek, V. & Smrz, P. (2005*b*), Ole - a new ontology learning platform, *in* 'Proceedings of the International Workshop on Text Mining'.

Odell, M. & Russell, R. (1918), U.s. patent numbers 1,261,167. U.S. Patent Office, Washington, D.C.

Odell, M. & Russell, R. (1922), U.s. patent numbers 1,435,663. U.S. Patent Office, Washington, D.C.

Pakhomov, S. (2001), Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts, *in* 'Proceedings of the 40th Annual Meeting on Association for Computational Linguistics'.

Park, Y. & Byrd, R. (2001), Hybrid text mining for finding abbreviations and their definitions, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)'.

Philips, L. (1990), 'Hanging on the metaphone', *Computer Language Magazine* **7**(12), 38–44.

Robertson, S. (2004), 'Understanding inverse document frequency: On theoretical arguments for idf', *Journal of Documentation* **60**(5), 503–520.

Sanchez, D. & Moreno, A. (2005), Automatic discovery of synonyms and lexicalizations from the web, *in* 'Proceedings of the 8th Catalan Conference on Artificial Intelligence'.

Schwartz, A. & Hearst, M. (2003), A simple algorithm for identifying abbreviation definitions in biomedical text, *in* 'Proceedings of the Pacific Symposium on Biocomputing (PSB)'.

Smith, T. & Waterman, M. (1981), 'Identification of common molecular subsequences', *Journal of Molecular Biology* **147**(1), 195–197.

Sombatsrisomboon, R., Matsuo, Y. & Ishizuka, M. (2003), Acquisition of hypernyms and hyponyms from the www, *in* 'Proceedings of the 2d International Workshop on Active Mining', Japan.

Sproat, R., Black, A., Chen, S., Kumar, S., Ostendorf, M. & Richards, C. (2001), 'Normalization of non-standard words', *Computer Speech & Language* **15**(3), 287–333.

Tang, J., Li, H., Cao, Y. & Tang, Z. (2005), Email data cleaning, *in* 'Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining'.

Turney, P. (2001), Mining the web for synonyms: Pmi-ir versus lsa on toefl, *in* 'Proceedings of the 12th European Conference on Machine Learning (ECML)', Freiburg, Germany.

Wagner, R. & Fischer, M. (1974), 'The string-to-string correction problem', *Journal of the ACM* **21**(1), 168–173.

Wong, W., Liu, W. & Bennamoun, M. (2006), Progress and open problems in ontology engineering from text. Submitted to the Journal of Web Semantics.

Xu, F., Kurz, D., Piskorski, J. & Schmeier, S. (2002), An domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping, *in* 'Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)', Canary island, Spain.

A Study of Local and Global Thresholding Techniques in Text Categorization

Nayer M. Wanas; Dina A. Said; Nadia H. Hegazy; and Nevin M. Darwish
Pattern Recognition and Information Systems Group, Informatics Department, Electronics Research Institute, Cairo, Egypt
Department of Computer Engineering, Faculty of Engineering, Cairo University, Cairo, Egypt
nwanas@ieee.org, dsaid@ieee.org, nhegazy@mcit.gov.eg, ndarwish@ieee.org

Abstract

Feature Filtering is an approach that is widely used for dimensionality reduction in text categorization. In this approach feature scoring methods are used to evaluate features leading to selection. Thresholding is then applied to select the highest scoring features either locally or globally. In this paper, we investigate several local and global feature selection methods. The usage of Standard Deviation (STD) and Maximum Deviation (MD) as globalization schemes is suggested. This work provides a comparative study among fourteen thresholding techniques using different scoring methods and benchmark datasets of diverse nature. This includes investigation of normalizing feature scores before combining them in the global pool. The results suggest that normalized MD outperforms other methods in thresholding Document Frequency (DF) scores using even and moderate diverse data-sets. Furthermore, the results indicated that normalizing feature scores improves the performance of rare categories and balances the bias of some techniques to frequent categories.

Keywords:{Text Categorization, Thresholding Techniques, Dimensionality Reduction, Feature Filtering, Support Vector Machine}

1 Introduction

Text Categorization (TC) is the process of assigning a given text to one or more categories. This process is considered as a supervised classification technique, since a set of labeled (pre-classified) documents is provided as a training set. The goal of TC is to assign a label to a new, unseen, document(Yang 1995). TC can play an important role in a wide variety of areas such as information retrieval(Freitas-Junior, Ribeiro-Neto, Vale, Laender & Lima 2006), news recommendation(Antonellis, Bouras & Poulopoulos 2006), word sense disambiguation(Montoyo, Suarez, Rigau & Palomar 2005), topic detection and tracking(Tsay, Shih & Wu 2005), web pages classification(Yang, Slattery & Ghani 2002, Shen, Sun, Yang & Chen 2006), as well as any application requiring document organization.

Text categorization can be divided into four major stages:

1. **Document pre-processing** where words are extracted from the documents and presented in a form of bag of words (BOW)(Salton, Wong & Yang 1975). Song et al. showed that performing stop word removal and stemming during the pre-processing step may harm the performance a little bit but it has a great effect on reducing the feature space significantly(Song, Liu & Yang 2005).
2. **Dimensionality reduction (DR):** is performed using either feature extraction or feature selection approaches(Sebastiani 2002). In feature extraction, new features are generated based on complex methods such as latent Semantic Indexing (LSI)(Wiener, Pedersen & Weigend 1995), Independent Component Analysis(Kolenda, Hansen & Sigurdsson 2000), and word clustering(Cohen & Singer 1999). On the other hand, the feature selection approach is based on selecting high-relevance features by either filtering irrelevance features(Yang & Pedersen 1997) or wrapping the features around the classifier used(Kohavi & John 1997).
3. **Feature Weighting** where the selected features are weighted by commonly using term frequency inverse document frequency ($tfidf$) technique(Sebastiani 2002) where tf measures the importance of the term within the document and idf measures the general importance of the term(Salton & Buckley 1988). In order to handle the diversity in document lengths, normalization of the $tfidf$ measure could be performed which has proved to lead to a significant improvement in the performance(Song et al. 2005).
4. **Classification:** The weighted feature set is used to learn how to distinguish among the different document categories. Several supervised classifiers have been used in TC. Among them are Decision trees(Johnson, Oles, Zhang & Goetz 2002), K-nearest neighbors(Bang, Yang & Yang 2006) , Naive Bayes(Peng, Schuurmans & Wang 2004), Neural Network(Chen, Lee & Hwang 2005), Maximum Entropy(Kazama & Tsujii 2005), and Support Vector Machine(SVM)(Díaz, Ranilla, Montañes, Fernández & Combarro 2004). SVM has been shown to be among the best performing classifiers in TC applications(Joachims 1998, Dumais, Platt, Heckerman & Sahami 1998, Yang 1999, Yang, Zhang & Kisiel 2003, Li & Yang 2003, Debole & Sebastiani 2005).

DR is one of the most important challenges in TC. This reduction is necessary to avoid overfitting, as well as decreasing the computational resources, storage , and the memory required to manage these features(Sebastiani 2002, Guyon & Elisseeff 2003). This study concentrates on the filter approach of DR due to its simplicity compared to the other

approach(Blum & Langley 1997, Combarro, Montanes, Diaz, Ranilla & Mones 2005). Filtering features is performed first by applying feature scoring methods locally to each category in the training set in order to evaluate features. This is followed by thresholding, which is performed either locally, or globally to select the features that have the highest feature scores(Debole & Sebastiani 2005).

This work proposes new techniques for thresholding feature scores in order to address some shortcomings in the existing techniques. In addition, few studies have been conducted to compare the performance of different thresholding techniques. In order to evaluate the proposed thresholding techniques, a comparative study is conducted among these techniques and the existing state-of-art thresholding techniques. Since the performance of the thresholding techniques is highly affected by the feature scoring method used, the comparative study is performed using four high-performing feature scoring methods and different thresholding values. These feature scoring methods are the Document Frequency(DF)(Yang & Pedersen 1997), Information Gain(IG)(Sebastiani 2002), Mutual Information(MI)(Yang & Pedersen 1997), and Correlation Coefficient(CC)(Ng, Goh & Low 1997). These methods have been widely used, and have shown to be among the top performing methods(Yang & Pedersen 1997, Sebastiani 2002). It might be worth noting that CC is a modification of χ^2 and has been shown to outperform it in the literature(Ng et al. 1997, Ruiz & Srinivasan 2002). The following sections elaborate together to show the rational beyond the new thresholding methods as well as their performance evaluation using different benchmark data-sets.

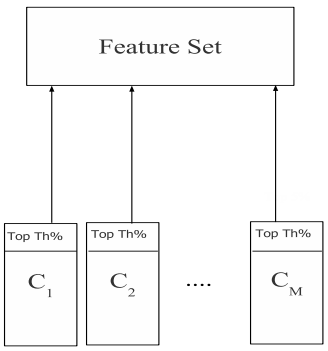
2 Thresholding Policies for Feature Filtering

Thresholding is an important issue in performing feature selection using the filter approach. After applying feature scoring methods to each category, thresholding is performed to select the final set of features that represents the training set. Selection is done according to either (a) a local policy, or (b) a global policy. In the local policy, thresholding is applied locally on each category to compose the final representative feature set(Díaz et al. 2004). While a globalization scheme is performed to extract a single global score for features in the global policy. Then, features with the highest global scores are selected(Yang & Pedersen 1997). Figure 1 illustrates the difference between local and global thresholding policies.

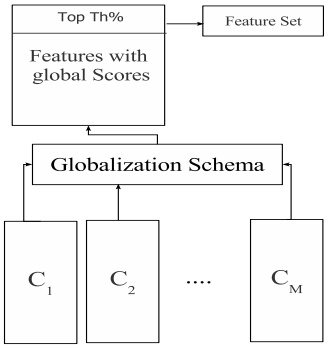
2.1 The local policy

Several researchers have suggested usage of a local policy, where a different set of features is chosen from each category independent on other categories. The local policy tends to optimize the classification process for each category by selecting the most relevant features in that category.

Local selection policy was used by(Apté, Damerau & Weiss 1994) where a local dictionary of the most important words in each topic was used. They then selected from each category the words that matched the category dictionary. However, using local dictionaries suffers from being a domain-dependent and language-dependent approach. Alternatively, Lewis and Ringuette selected the top IG features from each category(Lewis & Ringuette 1994). On the other hand, an implementation to local feature extraction was proposed by(Wiener et al. 1995) where they applied LSI on each category separately. An extension to the usage of the local strategy was presented by(Ng et al. 1997) where they applied this policy to three feature scoring methods, namely DF, CC, and χ^2.



(a) Fixed Local thresholding



(b) Global thresholding

Figure 1: Local and Global Thresholding

The previous approaches select the same number of features from each category. In this work we refer to this policy as the fixed local approach (FLocal). On the other hand, Soucy and Mineau(Soucy & Mineau 2003), proposed selecting features in proportion to the category distribution, or what can be considered a weighted local approach (WLocal). The rational behind this approach is that in highly-skewed datasets the ratio among words of frequent and infrequent categories maybe very large. Hence, selecting the same number of features from both distributions may degrade the performance.

In (Forman 2004), Forman proposed a similar idea to fixed local thresholding, which he called round-robin. In this approach features are selected from each class in a round-robin manner. Additionally, Forman proposed another idea for local thresholding called rand-robin. In this method, the next class to select features from is determined randomly by a process that is controlled by the probability of the class distribution. This seems very similar to the idea of WLocal. However, one drawback of rand-robin is that in highly-skewed datasets it might not select any document from rare-categories. This is due to its random nature in selection, which depends on the category probability.

2.2 The global policy

As an alternative to the local policy, the global policy aims at providing a global view of the training set by extracting a global score from local feature scores. Thresholding is then applied to these global scores,

where features with the highest global score are retained. Yang and Pedersen(Yang & Pedersen 1997) used Maximization (Max) and Weighted Averaging (WAvg) for extracting global scores from χ^2 and MI. Additionally, they used Averaging (Avg) for DF and IG. Calvo and Ceccatto proposed the usage of Weighted Maximum (WMax), where features are weighted by the category probability(Calvo & Ceccatto 2000). Equations 1, 2, 3, and 4 provide the mathematical definitions of Avg, Max, WAvg, and WMax respectively.

$$F_{Avg}(w_k) = \frac{\sum_{i=1}^{M} f(w_k, c_i)}{M} \quad (1)$$

$$F_{Max}(w_k) = \max_{i=1}^{M} \{f(w_k, c_i)\} \quad (2)$$

where $f(w_k, c_i)$ is the score of the word w_k w.r.t. the category c_i, and M is number of categories in the training set.

$$F_{WAvg}(w_k) = \frac{\sum_{i=1}^{M} p(c_i) f(w_k, c_i)}{M} \quad (3)$$

$$F_{WMax}(w_k) = \max_{i=1}^{M} \{p(c_i) f(w_k, c_i)\}. \quad (4)$$

2.3 Proposed Thresholding techniques

Feature scoring methods such as the DF, have an inherent biased to frequent categories. Similarly global techniques such as the Max or Avg tends to select features that are biased to frequent categories. Moreover, weighting the feature scores based on the category probability increases this bias and is excepted to degrade the performance mainly in the classification of rare categories. These rare categories may have high importance and their number maybe large especially in highly skewed datasets. Alternatively, we propose normalizing feature scores before applying the globalization scheme in order to enhance the classification process of rare categories and balance the bias of feature scoring methods such as DF. Accordingly, Normalized Average (NAvg), and Normalized Maximization (NMax) are defined as shown in equations 5, and 6 respectively:

$$F_{NAvg}(w_k) = \frac{\sum_{i=1}^{M} \frac{f(w_k, c_i)}{p(c_i)}}{M} \quad (5)$$

$$F_{NMax}(w_k) = \max_{i=1}^{M} \frac{f(w_k, c_i)}{p(c_i)}. \quad (6)$$

However, an interesting question that arises is how to define a good feature. It is usually assumed to be the feature that has the maximum or the maximum average score in the training set. As a matter of fact, when using a feature scoring method such as DF, this definition is rendered inappropriate. Based on this definition, the features that exist in all categories with the same DF will be considered as good features. However, a good feature is the one that has a score in one category that is substantially different from its score in all other categories. In order to identify such features, we propose the usage of the Standard Deviation (STD) as a globalization scheme. STD gives an estimated measure of how diverse the data is from the mean. Although intuitively the STD should capture good features, as opposed to the Max and Avg, it has its shortcomings. To illustrate this, suppose that a certain feature w_1 occurs in only one category with DF= x while another feature w_2 occurs in two categories with the same DF=x. According to the definition of a good feature w_1 should be considered better than w_2. When using DF as a scoring method, the mean of w_2 will be higher than the mean of w_1. Accordingly, the STD of w_2 will be greater than the STD of w_1 since the scores of w_2 will be more diverse from its mean. In order to overcome this pitfall, we propose the Maximum Deviation (MD) as a globalization scheme. Contrary to the STD, MD gives an estimation of how diverse is the data from the Max which makes it closer to realize the definition of a good feature compared to the STD. Equations 7, and 8 show the mathematical definitions of the STD, and MD respectively.

$$F_{STD}(w_k) = \sqrt{\frac{\sum_{i=1}^{M} \left[f(w_k, c_i) - f_{Avg}(w_k)\right]^2}{M}}. \quad (7)$$

$$F_{MD}(w_k) = \sqrt{\frac{\sum_{i=1}^{M} \left[f(w_k, c_i) - f_{Max}(w_k)\right]^2}{M}}. \quad (8)$$

Accordingly, Weighted STD (WSTD), Normalized STD (NSTD), Weighted MD (WMD), and Normalized MD (NMD) could be systematically defined.

2.4 Comparative studies of Thresholding Techniques

Table 1 provides a summary of different studies conducted to evaluate the performance of current thresholding techniques. The comparisons among the Max and WAvg performed by(Yang & Pedersen 1997, Galavotti et al. 2000) indicated that using the Max is better than WAvg. However, the two measures used in these experiments, $MicroF_1$ and 11-averaging point, are mainly affected by the performance of frequent categories especially in a highly skewed dataset such as Reuters-21578. The study performed by(Calvo & Ceccatto 2000) took $MacroF_1$ into consideration which is a measure of the performance of rare categories. However, the conclusion of this study contradicted the results of(Yang & Pedersen 1997, Galavotti et al. 2000) in $MicroF_1$ despite using the same dataset.

While the previous three studies used only Reuters dataset in the evaluation, Soucy and Mineau(Soucy & Mineau 2003) used four different datasets and four feature scoring methods. However, they reported only the best performance achieved for each dataset regardless of the feature scoring method and threshold value used. Additionally, the evaluation measure used in this study, $MicroF_1$, does not generally affected by the performance of rare categories. The study of(Forman 2004) was conducted using different small threshold values and nineteen different datasets. However, the results reported were only the average performance of the experiments conduct on these datasets. Both studies however fall short of presenting the complete picture required to determine the most suitable thresholding techniques for different datasets. Díaz et. al.(Díaz et al. 2004) presented a comparison between FLocal and Avg. Although this study didn't evaluate the WLocal or even the traditional Max, however, it supported the results of(Soucy & Mineau 2003, Forman 2004) that local thresholding maybe better than global thresholding.

In order to avoid the pitfalls of these studies, in this work we present a comparative study among the proposed thresholding techniques and the existing six techniques. This study is conducted using different threshold values, different feature scoring methods, and datasets of diverse natures. The results of this study are reported using $MicroF_1$ and $MacroF_1$ in order to evaluate the performance of frequent and rare categories.

Table 1: A Summary of Comparative Studies among Thresholding Techniques

	Scoring Methods	Datasets	Conclusions
Yang and Pedersen(Yang & Pedersen 1997)	IG, χ^2	Reuters-21578	11-average point: Max > WAvg
Galavotti et. al.(Galavotti, Sebastiani & Simi 2000)	Simplified χ^2	Reuters-21578	$MicroF_1$: Max > WAvg
Calvo and Ceccatto(Calvo & Ceccatto 2000)	χ^2	Reuters-21578	$MicroF_1$: WAvg, WMax > Max $MacroF_1$: Max,WAvg > WMax
Soucy and Mineau(Soucy & Mineau 2003)	IG, DF χ^2, Cross Entropy	Reuters-21578 lingSpam, DigiTrad, Ohsumed	$MicroF_1$: FLocal > WLocal > Max
Forman(Forman 2004)	IG, χ^2	19 datasets	IG ($MacroF_1$): FLocal > WLocal > Max > Avg χ^2 ($MacroF_1$): FLocal > Max > Avg > WLocal
Díaz et al.(Díaz et al. 2004)	IG, tf , $tfidf$	Reuters-21578, Ohsumed	Recall and Precision: FLocal > Avg

3 Experimental Setup

Three different data-sets were used in this study, they are the 20 Newsgroups Collection (20NG)[1](Joachims 1997), Ohsumed data-set[2] (Joachims 1998, Combarro et al. 2005), and Reuters-21578 ModeApté split[3] (Reuters(90))(Joachims 1998). For the Ohsumed dataset, the first 20,000 documents with abstracts published in 1991 were considered which resulted in 23 categories. The first 10,000 have been used as training set and the rest as test set.

While 20NG is an evenly distributed data-set, Ohsumed data-set can be considered as a moderate diverse data-set. Ohsumed has 23 categories where the largest category has about 1800 document and the smallest one has 65 documents. On the other hand, Reuters (90) can be considered as a highly diverse dataset as it has about 30 categories whose number of documents is below 10.

The F_1 measure is used as an effective measure. F_1 was first proposed as a measure of effectiveness in TC by Lewis(Lewis & Ringuette 1994). $MicroF_1$ and $MacroF_1$ tests are two measures based on F_1. The $MacroF_1$ test equally weights all categories, and thus it is influenced by the performance of rare categories. On the other hand, $MicroF_1$ test equally weights all the documents, and therefore it is affected by the performance of frequent categories(Yang 1999).

4 Results

The study compares the performance of six feature thresholding techniques, namely Averaging (Avg), Maximization (Max), Fixed Local (FLocal), Weighted Local (WLocal) in addition to our proposed techniques Standard Deviation (STD), and Maximum Deviation (MD). The globalization techniques are evaluated using the original, weighted, and normalized scores. This amounts to fourteen thresholding methods evaluated using four feature scoring methods. In this study we apply classification using the SVM^{light}(Joachims 1999)[4]. The experimental results are evaluated using 20 news group (20NG), Ohsumed, and Reuters-21578 ModeAptè (Reuters(90)) data-sets. In the following, we will focus on the evaluation of the thresholding techniques for each data-set separately.

[1]The 20 Newsgroups and the bydate split can be found at http://people.csail.mit.edu/people/jrennie/20Newsgroups

[2]Ohsumed is available http://trec.nist.gov/data/t9_filtering/.

[3]Reuters-21578 is available at http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[4]SVM^{light} is publicly published at http://svmlight.joachims.org.

4.1 20NG Data-set

Since the 20NG dataset is an evenly distributed data-set, the $MicroF_1$ and $MacroF_1$ values match. Therefore, only the results of $MicroF_1$ are reported in Figure 2. Additionally, due to the equal distribution of categories in the data-set, weighting or normalizing feature scores will not be of value since all the categories have the same probability. Therefore, the evaluated thresholding techniques on this data-set are limited to only the Avg, Max, STD, MD and FLocal. The following observations can be seen from the results:

- The results show that the MD has a limited improvement when using DF as a scoring method, especially noticeable at low threshold values. On the hand, FLocal has the best performance compared to the remaining methods.
- Generally, the Max, and MD exhibit an almost identical performance in feature scoring methods except DF. As a matter of fact, the correlation between the feature sets they produce show a high similarity, generally above 90%. This is due to the fact that these scoring methods take into account the relevance of the feature in the category under investigation and other categories in the training set. Therefore, using either the Max, or MD tends to attain nearly the same set of features. This is an expected result since these scoring methods follow from the rational of a good feature.
- Generally, FLocal shows a performance superior to the Max and MD in thresholding feature scoring methods except DF. This is due to the diversity in the number of good features from one category to the other. Therefore, globalization schemes such as the Max, or MD tend to be biased to certain categories in accordance to the number of good features they include. On the other hand, FLocal is an unbiased technique that forces the selection of the same number of features from all the categories.
- The MD is usually slightly better than the STD which shows the potential of MD to capture discriminative features.
- The Avg thresholding technique exhibits the worst performance since it sums the scores of the features across categories and hence it tends to select non-discriminative features.

4.2 Ohsumed Data-set

Tables 2, and 3 show the performance of $MicroF_1$, and $MacroF_1$ for the Ohsumed data-set respectively.

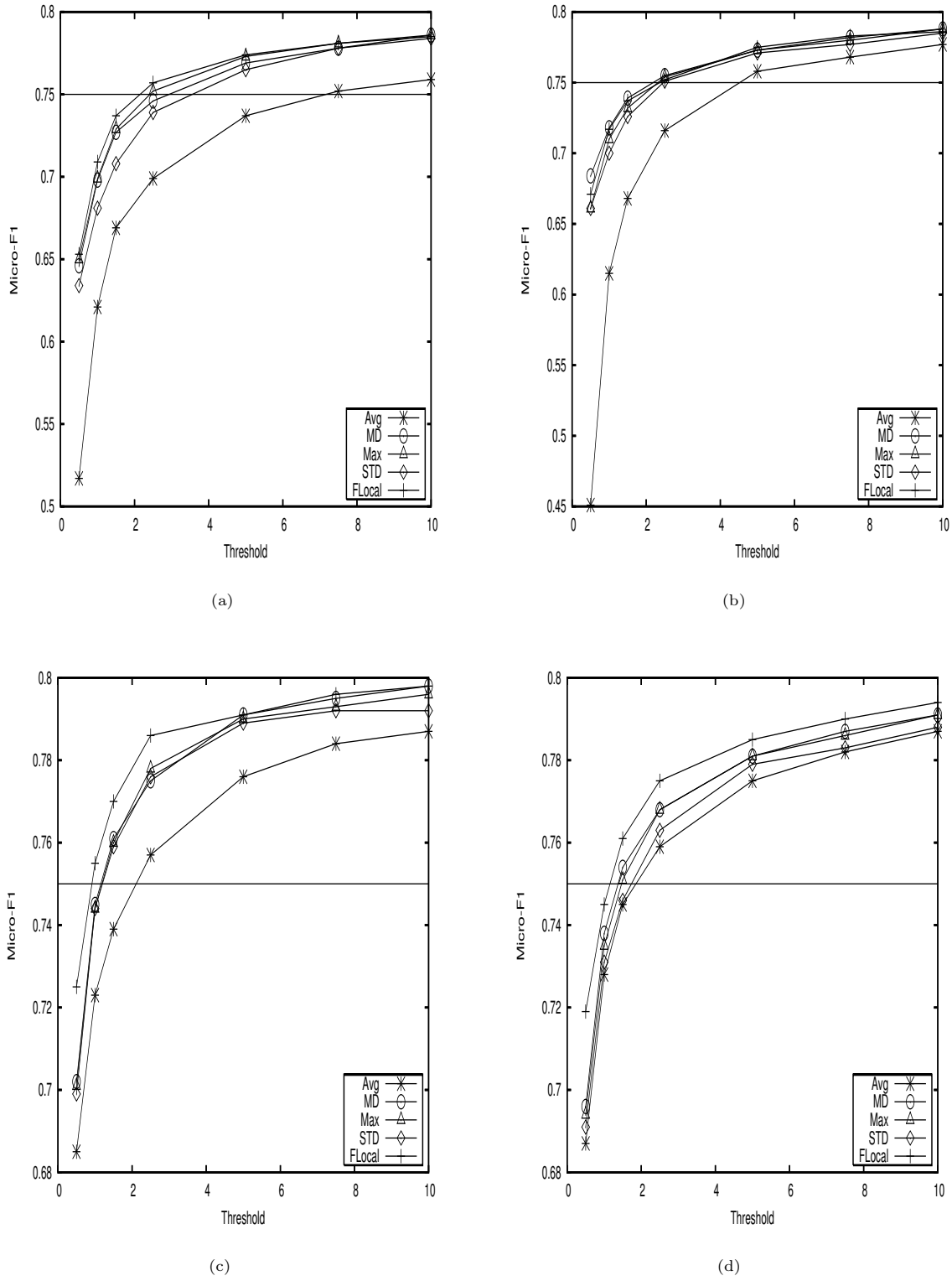


Figure 2: $MicroF_1$ of the 20NG using Thresholding Techniques (a) CC, (b) DF, (c) IG, and (d) MI

Analyzing these results we can make the following observations:

- NMD achieves the best $MicroF_1$ and $MacroF_1$ in thresholding DF scores.
- For CC, MI, and IG, FLocal has the best performance for $MacroF_1$. On the other hand, WLocal is better than FLocal considering the $MicroF_1$ and small threshold values. However, starting at a certain threshold, the performance of FLocal almost matches that of WLocal. This threshold differs from one scoring method to the other.
- Consistent with the results from the 20NG, the MD and Max show a high degree of similarity in performance when thresholding the CC, IG, and MI.
- The poor performance of the Avg in thresholding DF scores is not surprising. Selecting the highest average DF features at small threshold values tends to retain the features that exist in all categories. These features are not useful to discriminate among categories.
- Generally, weighting the scores performs poorly compared to both the original and normalized scores. On the other hand, normalizing scores is better than using the original score at the macro level, since it adds bias to rare categories.

4.3 Reuters(90) Data-set

Figures 3, and 4 illustrate the performance of $MicroF_1$, and $MacroF_1$ for the Reuters (90) dataset respectively. Due to space limitations, we reduced the figure to include only the best six thresholding techniques(MD, Max, NMD, NMax, FLocal and WLocal). The analysis of the performance of these techniques yields the following observations:

- The performance of NMD and NMax is very poor specially for small threshold values. Since Reuters(90) is a highly skewed dataset, normalizing scores will be highly biased to rare categories at small threshold values.
- At the micro level, MD, Max, FLocal, and WLocal have similar performance with a limited superiority of WLocal at small threshold values due to the added bias to frequent categories.
- At the macro level, one can conclude that the WLocal is the best method in thresholding IG and MI at threshold values greater than 1%. It is surprising to observe that the WLocal is generally better than the FLocal on the macro level. This is despite the fact that the FLocal allows the selection of more features from rare categories. Examining the F_1 of individual categories shows that FLocal in fact slightly enhances the performance of rare categories. However, WLocal enhances the performance of frequent categories significantly as it selects more features due to the highly skewed nature of Reuters(90). Since $MacroF_1$, as mentioned in section **??**, gives the same weight to all categories , the WLocal is generally better than the FLocal in thresholding IG and MI.
- On the other hand, FLocal is the best method considering the CC and DF for threshold values less than 7.5%. Beyond this threshold, WLocal outperforms FLocal. Using WLocal in thresholding CC and DF scores in small threshold values leads to a selection of a very small number of features from rare categories. If the scoring method used is not in itself a high-performing scheme, then adding these limited number of selected features will not be enough to discriminate rare categories. This argument is asserted by examining the F_1 of individual categories.

5 Conclusions

In this study we propose performing global feature selection using the Standard Deviation (STD) and Maximum Deviation (MD). In order to balance the bias of some scoring features to frequent categories, we suggest normalizing feature scores before applying the globalization scheme. We conducted a comparative study using our new techniques and state-of-art thresholding techniques. The study was performed using four feature scoring methods and various datasets of different nature.

Generally, localization techniques are better than globalization methods, which supports the results obtained by(Soucy & Mineau 2003, Díaz et al. 2004, Forman 2004). Additionally, localization techniques are much faster since thresholding is done on each category individually as opposed to the whole training set. Furthermore, there might be a possibility to perform thresholding on local categories independently in parallel, which may help speedup the process of dimensionality reduction. With the exception of the DF scoring method, FLocal is the best thresholding method in even distributed data-set and generally moderate diverse data-set. On the other hand, WLocal is the best in the $MicroF_1$ and some cases of the $MacroF_1$ of highly diverse data-set. Furthermore, WLocal achieves a performance that is either better than or equivalent to that of FLocal in the $MicroF_1$ of moderate diverse data-set.

MD shows an enhancement in the performance of thresholding using the DF as a scoring method in an even data-set. Normalized MD shows also potential in improvement the performance of moderate diverse data-set and large threshold values of highly diverse data-set. However, MD has a performance similar to the Max in all other feature scoring methods.

Generally, the MD outperforms STD in most cases. This supports that MD is more efficient in identifying good features. On the other hand, the Avg shows a poor performance comparable to other thresholding techniques. This is in consistent with the results of (Yang & Pedersen 1997, Galavotti et al. 2000).

Normalizing the features scores before applying the globalization method shows also an enhancement in the $MacroF_1$ of moderate diverse data-set. However, when using IG and MI in highly-skewed data-sets, the performance does not match that of the original unmodified score.

As a general conclusion, the results suggest the usage of MD in an even distributed data-set and NMD for a moderate diverse data-set for thresholding DF scores. For methods except DF, the results recommend FLocal for even data-set. Furthermore, WLocal should be used in moderate diverse data-set in case that frequent categories are more important and the number of desired features is small. Otherwise, FLocal is recommended. However for a highly skewed data-set, WLocal is recommended for IG and MI as it generally enhances the classification of both frequent and rare categories. Additionally, WLocal is suggested for CC and DF if frequent categories are more important while FLocal is recommended for rare ones.

Table 2: $MicroF_1$ of the Ohsumed data-set

		Avg			MD			Max			STD			Local	
%	Metric	Avg	WAvg	NAvg	MD	WMD	NMD	Max	WMax	NMax	STD	WSTD	NSTD	Flocal	WLocal
0.5		0.193	0.248	0.150	0.375	0.299	0.366	0.377	0.299	0.365	0.356	0.311	0.347	0.380	**0.413**
1		0.277	0.289	0.242	0.444	0.356	0.416	0.455	0.354	0.424	0.420	0.407	0.411	0.447	**0.457**
1.5		0.328	0.326	0.298	0.488	0.420	0.460	0.487	0.406	0.453	0.456	0.445	0.438	0.490	**0.505**
2.5	CC	0.441	0.392	0.358	0.530	0.472	0.528	0.537	0.472	0.534	0.509	0.488	0.504	0.542	**0.546**
5		0.511	0.448	0.456	0.571	0.546	0.574	0.573	0.539	0.579	0.559	0.543	0.558	**0.584**	**0.585**
7.5		0.552	0.484	0.503	0.587	0.568	0.589	0.589	0.569	0.596	0.581	0.569	0.581	**0.601**	0.596
10		0.567	0.514	0.527	0.596	0.580	0.600	0.599	0.589	**0.603**	0.592	0.583	0.595	**0.604**	**0.602**
0.5		0.000	0.062	0.000	0.349	0.251	**0.423**	0.277	0.243	0.381	0.298	0.255	0.391	0.385	0.395
1		0.141	0.177	0.089	0.389	0.293	**0.486**	0.378	0.293	0.444	0.379	0.315	0.458	0.455	0.430
1.5		0.370	0.356	0.384	0.431	0.324	**0.507**	0.402	0.334	0.491	0.406	0.329	0.498	0.485	0.477
2.5	DF	0.370	0.356	0.384	0.494	0.401	**0.548**	0.486	0.401	0.526	0.481	0.418	0.529	0.527	0.516
5		0.485	0.471	0.475	0.544	0.513	**0.591**	0.544	0.515	0.574	0.541	0.510	0.576	0.573	0.557
7.5		0.534	0.520	0.534	0.575	0.541	**0.604**	0.576	0.541	0.596	0.574	0.538	0.591	0.597	0.576
10		0.553	0.547	0.556	0.592	0.561	**0.605**	0.589	0.559	0.600	0.585	0.565	0.601	0.601	0.586
0.5		0.480	0.445	0.432	0.493	0.445	0.441	0.493	0.445	0.441	0.491	0.444	0.438	0.486	**0.505**
1		0.520	0.495	0.493	0.537	0.504	0.501	0.535	0.504	0.501	0.530	0.504	0.500	0.532	**0.553**
1.5		0.545	0.533	0.529	0.562	0.531	0.524	0.562	0.531	0.525	0.558	0.533	0.529	0.567	**0.578**
2.5	IG	0.580	0.553	0.563	0.598	0.565	0.560	0.596	0.566	0.561	**0.603**	0.566	0.566	**0.605**	**0.603**
5		0.606	0.590	0.595	**0.628**	0.593	0.592	**0.627**	0.592	0.608	0.625	0.598	0.601	**0.629**	**0.628**
7.5		0.610	0.603	0.600	**0.633**	0.612	0.615	**0.633**	0.612	0.617	0.629	0.613	0.619	**0.635**	**0.630**
10		0.611	0.606	0.607	**0.633**	0.620	0.622	**0.632**	0.620	0.624	**0.633**	0.622	0.625	**0.636**	**0.632**
0.5		0.470	0.400	0.274	0.483	0.400	0.435	0.484	0.399	0.436	0.470	0.404	0.431	0.475	**0.495**
1		0.520	0.459	0.410	0.525	0.476	0.490	0.523	0.478	0.491	0.514	0.474	0.488	0.518	**0.535**
1.5		0.551	0.489	0.469	0.543	0.504	0.518	0.543	0.503	0.518	0.536	0.498	0.511	**0.562**	**0.563**
2.5	MI	0.587	0.536	0.524	0.582	0.537	0.551	0.585	0.539	0.555	0.568	0.531	0.565	**0.595**	**0.592**
5		0.616	0.583	0.591	0.610	0.573	0.604	0.610	0.577	0.608	0.602	0.571	0.600	**0.620**	**0.617**
7.5		0.624	0.603	0.610	0.617	0.592	0.611	0.616	0.595	0.614	0.611	0.592	0.609	**0.621**	**0.621**
10		0.625	0.608	0.610	**0.620**	0.604	0.614	**0.622**	0.607	0.616	0.612	0.598	0.610	**0.623**	0.618

Table 3: $MacroF_1$ of the Ohsumed data-set

		Avg			MD			Max			STD			Local	
%	Metric	Avg	WAvg	NAvg	MD	WMD	NMD	Max	WMax	NMax	STD	WSTD	NSTD	Flocal	WLocal
0.5		0.108	0.083	0.145	0.195	0.093	0.269	0.198	0.093	0.274	0.176	0.119	0.203	**0.273**	**0.273**
1		0.154	0.099	0.225	0.289	0.162	0.326	0.295	0.154	0.339	0.239	0.216	0.265	**0.362**	0.316
1.5		0.202	0.127	0.285	0.329	0.226	0.372	0.333	0.202	0.369	0.275	0.246	0.318	**0.395**	0.365
2.5	CC	0.304	0.185	0.331	0.381	0.291	0.427	0.391	0.272	0.434	0.333	0.289	0.396	**0.444**	0.407
5		0.391	0.246	0.422	0.429	0.375	0.482	0.429	0.353	0.488	0.417	0.366	0.470	**0.499**	0.465
7.5		0.439	0.302	0.463	0.464	0.409	0.498	0.467	0.415	0.506	0.451	0.414	0.491	**0.517**	0.482
10		0.455	0.341	0.480	0.473	0.440	0.517	0.481	0.449	0.519	0.479	0.438	0.503	**0.521**	0.497
0.5		0.000	0.018	0.000	0.160	0.067	**0.332**	0.084	0.065	0.242	0.112	0.067	0.254	0.272	0.230
1		0.041	0.049	0.023	0.205	0.098	**0.393**	0.197	0.097	0.344	0.199	0.118	0.353	0.359	0.274
1.5		0.109	0.095	0.098	0.238	0.127	**0.422**	0.214	0.135	0.387	0.220	0.130	0.396	0.392	0.332
2.5	DF	0.185	0.159	0.208	0.304	0.196	**0.462**	0.300	0.196	0.430	0.295	0.218	0.427	0.440	0.389
5		0.310	0.276	0.320	0.379	0.336	**0.511**	0.379	0.336	0.488	0.375	0.331	0.485	0.490	0.435
7.5		0.374	0.349	0.401	0.440	0.386	**0.524**	0.441	0.386	0.510	0.436	0.368	0.503	0.514	0.460
10		0.405	0.390	0.442	0.470	0.416	**0.521**	0.466	0.412	0.514	0.462	0.412	0.514	0.518	0.477
0.5		0.310	0.244	0.342	0.360	0.239	0.371	0.360	0.239	0.371	0.340	0.244	0.366	**0.405**	0.390
1		0.391	0.283	0.409	0.421	0.286	0.448	0.418	0.287	0.448	0.414	0.289	0.446	**0.457**	0.445
1.5		0.413	0.355	0.448	0.456	0.336	0.478	0.454	0.336	0.478	0.453	0.352	0.483	**0.508**	0.476
2.5	IG	0.463	0.372	0.492	0.500	0.388	0.513	0.497	0.389	0.515	0.504	0.391	0.518	**0.546**	0.500
5		0.510	0.444	0.523	0.554	0.429	0.537	0.548	0.429	0.546	0.546	0.432	0.540	**0.561**	0.549
7.5		0.517	0.478	0.520	0.555	0.469	0.548	0.555	0.469	0.549	0.553	0.478	0.550	**0.566**	0.552
10		0.517	0.488	0.523	0.558	0.486	0.549	0.555	0.488	0.552	0.553	0.503	0.549	**0.562**	0.553
0.5		0.295	0.192	0.281	0.317	0.192	0.359	0.317	0.192	0.364	0.303	0.196	0.356	**0.397**	0.386
1		0.387	0.242	0.372	0.396	0.266	0.422	0.395	0.266	0.424	0.359	0.265	0.412	**0.446**	0.427
1.5		0.425	0.261	0.413	0.418	0.287	0.462	0.419	0.286	0.462	0.407	0.285	0.445	**0.498**	0.448
2.5	MI	0.471	0.334	0.473	0.468	0.349	0.495	0.473	0.350	0.499	0.445	0.340	0.501	**0.528**	0.488
5		0.521	0.410	0.520	0.520	0.406	0.534	0.520	0.410	0.538	0.505	0.404	0.528	**0.548**	0.532
7.5		0.533	0.464	0.529	0.530	0.436	0.538	0.530	0.440	0.538	0.519	0.440	0.531	**0.548**	0.540
10		0.537	0.477	0.530	0.534	0.466	0.536	0.534	0.468	0.536	0.521	0.458	0.527	**0.545**	0.535

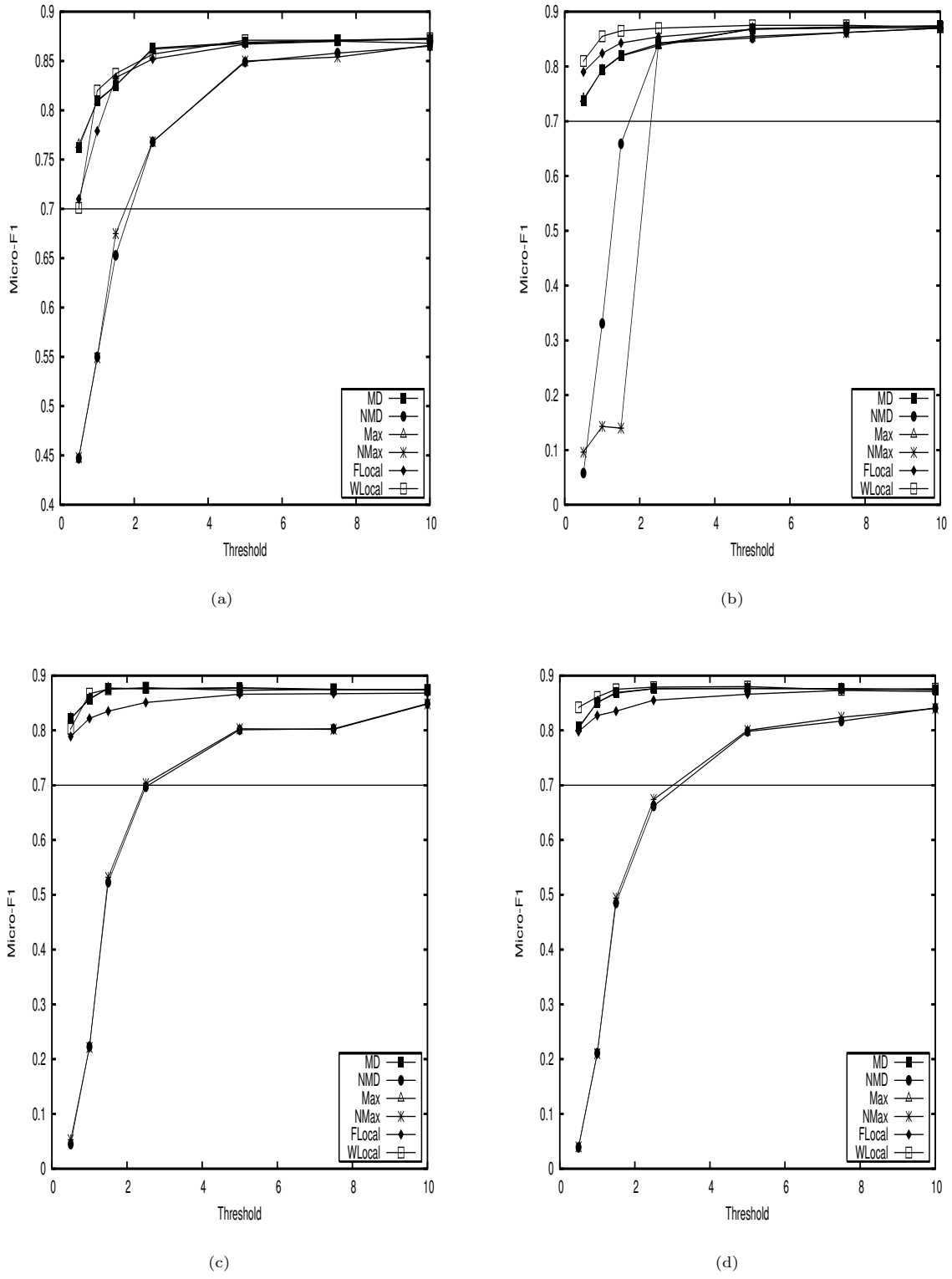


Figure 3: $MicroF_1$ of the Reuters(90) data-set using the Thresholding Techniques (a) CC, (b) DF, (c) IG, and (d) MI

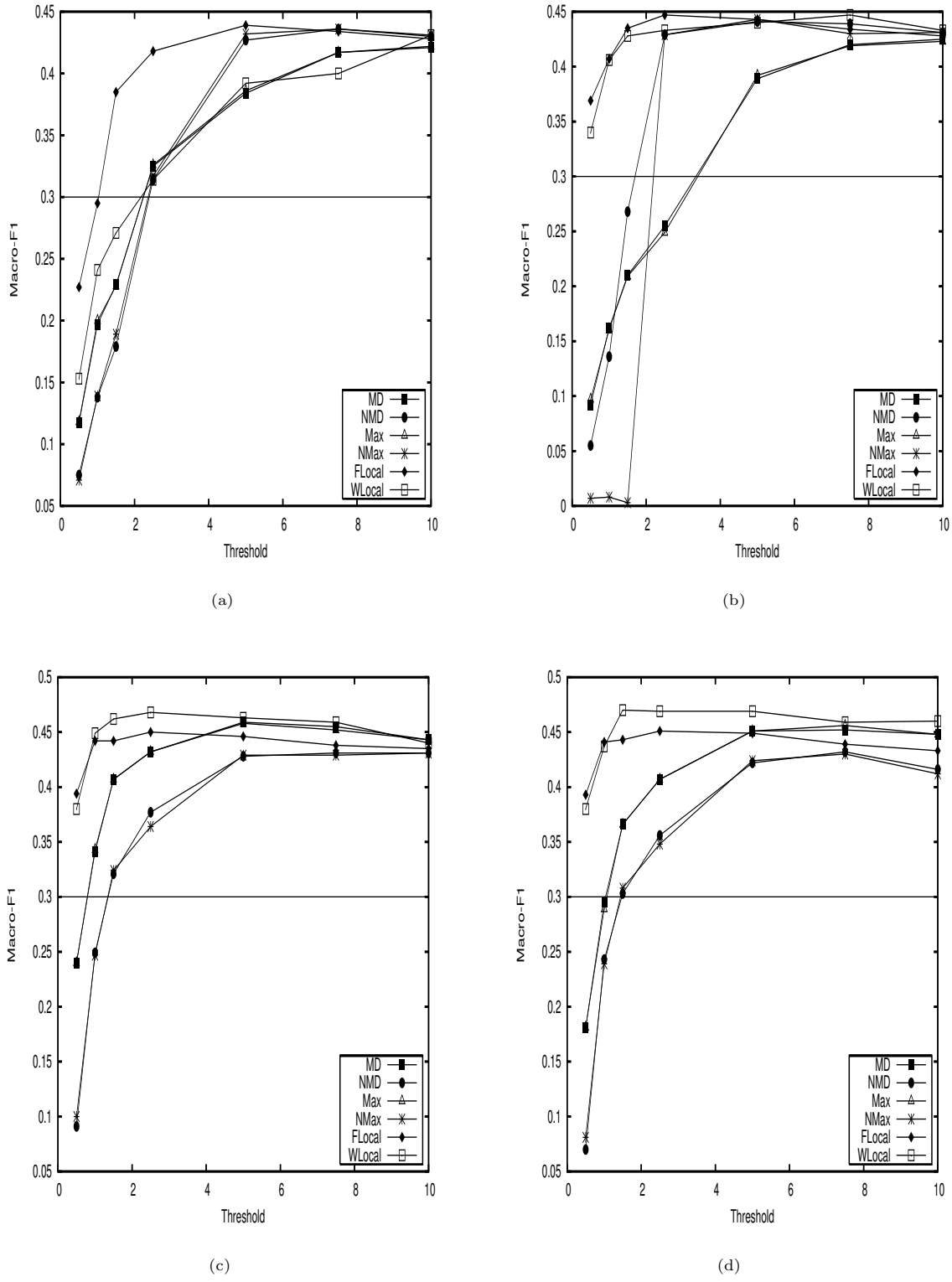


Figure 4: $MacroF_1$ of the Reuters(90) data-set using Thresholding Techniques (a) CC, (b) DF, (c) IG, and (d) MI

References

Antonellis, I., Bouras, C. & Poulopoulos, V. (2006), Personalized news categorization through scalable text classification., *in* 'Proc. of the 8th Asia-Pacific Web Conference- Frontiers of WWW Research and Development (APWeb 2006)', Vol. 3841 of *Lecture Notes in Computer Science*, Springer-Verlag New York, Inc., Harbin, China, pp. 391–401.

Apté, C., Damerau, F. & Weiss, S. (1994), 'Automated learning of decision rules for text categorization', *ACM Transactions on Information Systems (TOIS)* **12**(3), 233–251.

Bang, S., Yang, J. & Yang, H. (2006), 'Hierarchical document categorization with k-NN and concept-based thesauri', *Information Processing and Management* **42**(2), 387–406.

Blum, A. & Langley, P. (1997), 'Selection of relevant features and examples in machine learning.', *Artificial Intelligence* **97**(1-2), 245–271.

Calvo, R. & Ceccatto, H. (2000), 'Intelligent document classification', *Intelligent Data Analysis* **4**(5), 411–420.

Chen, C.-M., Lee, H.-M. & Hwang, C.-W. (2005), 'A hierarchical neural network document classifier with linguistic feature selection', *Applied Intelligence* **23**(3), 277–294.

Cohen, W. & Singer, Y. (1999), 'Context-sensitive learning methods for text categorization', *ACM Transactions on Information Systems* **17**(2), 141–173.

Combarro, E., Montanes, E., Diaz, I., Ranilla, J. & Mones, R. (2005), 'Introducing a family of linear measures for feature selection in text categorization', *IEEE Transaction on Knowledge and Date Engineering* **17**(9), 1223–1232.

Debole, F. & Sebastiani, F. (2005), 'An analysis of the relative hardness of reuters-21578 subsets', *Journal of the American Society for Information Science and Technology (JASIST)* **56**(6), 584–596.

Díaz, I., Ranilla, J., Montañes, E., Fernández, J. & Combarro, E. (2004), 'Improving performance of text categorization by combining filtering and support vector machines', *Journal of the American Society for Information Science and Technology (JASIST)* **55**(7), 579–592.

Dumais, S., Platt, J., Heckerman, D. & Sahami, M. (1998), Inductive learning algorithms and representations for text categorization, *in* 'Proc. of the 7th ACM International Conference on Information and Knowledge Management (CIKM'98)', ACM Press, New York, United States, Bethesda, United States, pp. 148–155.

Forman, G. (2004), A pitfall and solution in multi-class feature selection for text classification, *in* 'Proc. of the 21st International Conference on Machine Learning (ICML'04)', Vol. 69 of *ACM International Conference Proceeding Series*, ACM Press, New York, United States, Banff, Alberta, Canada, p. 38.

Freitas-Junior, H., Ribeiro-Neto, B., Vale, R., Laender, A. & Lima, L. (2006), 'Categorization-driven cross-language retrieval of medical information', *Journal of the American Society for Information Science and Technology (JASIST)* **57**(4), 501 – 510.

Galavotti, L., Sebastiani, F. & Simi, M. (2000), Experiments on the use of feature selection and negative evidence in automated text categorization, *in* 'Proc. of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'00)', Vol. 1923 of *Lecture Notes in Computer Science*, Springer-Verlag New York, Inc., Lisbon, Portugal, pp. 59–68.

Guyon, I. & Elisseeff, A. (2003), 'An introduction to variable and feature selection', *Journal of Machine Learning Research (JMLR), Special issue on special feature* **3**, 1157–1182.

Joachims, T. (1997), A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, *in* 'Proc. of the 14th International Conference on Machine Learning (ICML'97)', Morgan Kaufmann Publishers, San Francisco, United States, Nashville, United States, pp. 143–151.

Joachims, T. (1998), Text categorization with support vector machines: learning with many relevant features, *in* 'Proc. of the 10th European Conference on Machine Learning (ECML'98)', Vol. 1398 of *Lecture Notes in Computer Science*, Springer-Verlag New York, Inc., Chemnitz, Germany, pp. 137–142.

Joachims, T. (1999), Making large-scale support vector machine learning practical, *in* 'Advances in Kernel Methods – Support Vector Learning', MIT Press, Cambridge, MA, United States, pp. 169–184.

Johnson, D., Oles, F. J., Zhang, T. & Goetz, T. (2002), 'A decision-tree-based symbolic rule induction system for text categorization', *IBM systems Journal* **41**(3), 428–437.

Kazama, J. & Tsujii, J. (2005), 'Maximum entropy models with inequality constraints: A case study on text categorization', *Machine Learning* **60**(1-3), 159–194.

Kohavi, R. & John, G. H. (1997), 'Wrappers for feature subset selection', *Artificial Intelligence* **97**(1-2), 273–324.

Kolenda, T., Hansen, L. & Sigurdsson, S. (2000), Independent Components in text, *in* M. Girolami, ed., 'Advances in Independent Component Analysis', Springer-Verlag New York, Inc., pp. 229–250.

Lewis, D. & Ringuette, M. (1994), A comparison of two learning algorithms for text categorization, *in* 'Proc. of the 3rd Symposium on Document Analysis and Information Retrieval (SDAIR'94)', ISRI; University of Nevada, Las Vegas, United States, pp. 81–93.

Li, F. & Yang, Y. (2003), A loss function analysis for classification methods in text categorization, *in* 'Proc. of the 20th International Conference on Machine Learning (ICML'03)', AAAI Press, Menlo Park, United States, Washington, DC, USA, pp. 472–479.

Montoyo, A., Suarez, A., Rigau, G. & Palomar, M. (2005), 'Combining knowledge- and corpus-based word-sense-disambiguation methods', *Journal of Artificial Intelligence Research* **23**, 299–330.

Ng, H., Goh, W. & Low, K. (1997), Feature selection, perceptron learning, and a usability case study for text categorization, *in* 'Proc. of the 20th ACM International Conference on Research and Development in Information Retrieval (SIGIR'97)', ACM Press, New York, United States, Philadelphia, United States, pp. 67–73.

Peng, F., Schuurmans, D. & Wang, S. (2004), 'Augmenting Naive Bayes classifiers with statistical language models', *Information Retrieval* **7**(3-4), 317–345.

Ruiz, M. & Srinivasan, P. (2002), 'Hierarchical text classification using neural networks', *Journal of Information Retrieval* **5**(1), 87–118.

Salton, G. & Buckley, C. (1988), 'Term-weighting approaches in automatic text retrieval', *Information Processing and Management* **24**(5), 513–523.

Salton, G., Wong, A. & Yang, C. S. (1975), 'A vector space model for automatic indexing', *Communications of the ACM* **18**(11), 613–620.

Sebastiani, F. (2002), 'Machine learning in automated text categorization', *ACM Computing Surveys (CSUR)* **34**(1), 1–47.

Shen, D., Sun, J.-T., Yang, Q. & Chen, Z. (2006), A comparison of implicit and explicit links for web page classification, *in* 'Proc. of the 15th international conference on World Wide Web (WWW '06)', ACM Press, New York, United States, Edinburgh, Scotland, pp. 643–650.

Song, F., Liu, S. & Yang, J. (2005), 'A comparative study on text representation schemes in text categorization', *Pattern Analysis & Applications* **8**(1-2), 199–209.

Soucy, P. & Mineau, G. W. (2003), Feature selection strategies for text categorization., *in* 'Proc. of the 16th Conference of Canadian Conference on AI, the Canadian Society for Computational Studies of Intelligence', Vol. 2671 of *Lecture Notes in Computer Science*, Springer-Verlag New York, Inc., Halifax, Canada, pp. 505–509.

Tsay, J.-J., Shih, C.-Y. & Wu, B.-L. (2005), Autocrawler: An integrated system for automatic topical crawler, *in* 'Proc. of the 4th Annual ACIS International Conference on Computer and Information Science (ICIS'05)', IEEE Computer Society, Washington, DC, USA, Jeju Island, South Korea, pp. 462–467.

Wiener, E., Pedersen, J. & Weigend, A. (1995), A neural network approach to topic spotting, *in* 'Proc. of the 4th Symposium on Document Analysis and Information Retrieval (SDAIR'95)', Las Vegas, United States, pp. 317–332.

Yang, Y. (1995), Noise reduction in a statistical approach to text categorization, *in* 'Proc. of the18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95)', ACM Press, New York, United States, Seattle, Washington, United States, pp. 256–263.

Yang, Y. (1999), 'An evaluation of statistical approaches to text categorization', *Journal of Information Retrieval* **1**(1/2), 69–90.

Yang, Y. & Pedersen, J. (1997), A comparative study on feature selection in text categorization, *in* 'Proc. of the 14th International Conference on Machine Learning (ICML'97)', Morgan Kaufmann Publishers, San Francisco, United States, Nashville, Tennessee, United States, pp. 412–420.

Yang, Y., Slattery, S. & Ghani, R. (2002), 'A study of approaches to hypertext categorization', *Journal of Intelligent Information Systems* **18**(2/3), 219–241. Special Issue on Automated Text Categorization.

Yang, Y., Zhang, J. & Kisiel, B. (2003), A scalability analysis of classifiers in text categorization, *in* 'Proc. of the 26th ACM International Conference on Research and Development in Information Retrieval (SIGIR'03)', ACM Press, New York, United States, Toronto, Canada, pp. 96–103.

A Characterization of WordNet Features in Boolean Models for Text Classification

Trevor Mansuy **Robert J. Hilderman**

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{mansuy1t, robert.hilderman}@uregina.ca

Abstract

Supervised text classification is the task of automatically assigning a category label to a previously unlabeled text document. We start with a collection of pre-labeled examples whose assigned categories are used to build a predictive model for each category. In previous research, incorporating semantic features from the WordNet lexical database is one of many approaches that have been tried to improve the predictive accuracy of text classification models. The intuition is that words in the training set alone may not be extensive enough to enable the generation of a universal model for a category, but through WordNet expansion (i.e., incorporating words defined by various relationships in WordNet), a more accurate model may be possible. In this paper, we report preliminary results obtained from a comprehensive study where WordNet features, part of speech tags, and term weighting schemes are incorporated into two-category text classification models generated by both a Naive Bayes text classifier and an SVM text classifier. We characterize the behaviour of these classifiers on fifteen document collections extracted from the Reuters-21578, USENET, DigiTrad, and 20-Newsgroups text corpora. Experimental results show that incorporating WordNet features, utilizing part of speech tags during WordNet expansion, and term weighting schemes have no positive effect on the accuracy of the Naive Bayes and SVM classifiers.

Keywords: text classification, WordNet lexical database

1 Introduction

Text classification, the task of automatically assigning a category label to a previously unlabeled document, has been the focus of much recent research (de Buenaga Rodriguez, Gomez-Hidalgo & Diaz-Agudo 1997), (Hotho & Bloehdorn 2004), (Jensen & Martinez 2000), (Joachims 1998), (Kehagias, Petridis, Kaburlasos & Fragkou 2003), (Lewis 1998), (Mansuy & Hilderman 2006), (Peng & Choi 2005), (Rosso, Ferretti, Jiminez & Vidal 2004), (Scott & Matwin 1998), (Tan, Wang & Lee 2002), (Wiebe & O'Hara 2003). When performing supervised text classification, we use a collection of pre-labeled examples to build a predictive model for each distinct category contained in the examples. The classification accuracy we observe on the previously unlabeled documents largely depends on the quality of the training sets we have used to build the category models. That is, if the training information for a category model is sparse, then we can expect the category model to be a poor representation of the category, and the classification accuracy to be poor. Similarly, a training set may not necessarily be sparse, but it can contain important word relationships that a simple vector of words is not capable of modeling.

In an attempt to address the issue of related concepts in text classification models, a number of researchers have previously incorporated features derived from word relationships in the WordNet lexical database (de Buenaga Rodriguez et al. 1997), (Hotho & Bloehdorn 2004), (Jensen & Martinez 2000), (Kehagias et al. 2003), (Mansuy & Hilderman 2006), (Peng & Choi 2005), (Rosso et al. 2004), (Scott & Matwin 1998). WordNet is a database of words containing a semantic lexicon for the English language that organizes words into groups called synsets (i.e., synonym sets) (Miller 1995). A *synset* is a collection of synonymous words linked to other synsets according to a number of different possible relationships between the synsets (e.g., *is-a*, *has-a*, *is-part-of*, and others). When building a category model for a document, words related to a feature already in the model (and satisfying some desired WordNet relationship) are extracted from the WordNet database and incorporated into the model (called WordNet *expansion*).

In addition to incorporating related concepts to increase accuracy, other approaches have been considered. One of these approaches considers part of speech tags associated with words contained in a document (Jensen & Martinez 2000), (Scott & Matwin 1998). Since some words in the English language can have multiple meanings depending upon how and where they are used in a sentence, the *part of speech* may be relevant to a classification task. WordNet provides facilities for extracting only those words associated with a particular part of speech through tags assigned to words in the source document.

Another approach that has been considered effective for increasing accuracy is feature or term weighting (Kehagias et al. 2003). Term weighting is used to determine the relative importance or impact that a word contributes to the overall category model. That is, it determines the degree to which a word helps in distinguishing the category label.

In this paper, we attempt to address some of the uncertainty around the utilization of WordNet in text classification tasks by characterizing the behaviour that can be expected in a variety of situations. We evaluate the effect of incorporating different combinations of WordNet features, part of speech tags, and term weighting schemes in two-category text classification models generated by both a Naive Bayes text classifier and an SVM text classifier. We character-

```
Synset(pup)                Synset(dog)                Synset(cat)
 syn={dog}                  syn={pup}                  syn={kitten}
 hypr={canine}              hypr={canine}              hypr={feline}
 hypo={Terrier,black}       hypo={Terrier,black}       hypo={Persian,black}
 mero={claws,teeth}         mero={claws,teeth}         mero={claws,teeth}
 holo={pack,litter}         holo={pack,litter}         holo={litter}
```

Figure 1: Sample synsets

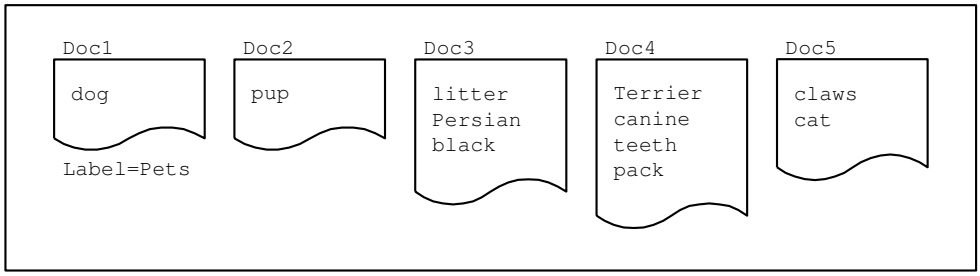


Figure 2: Sample documents

ize the behaviour of these classifiers on fifteen document collections extracted from the Reuters-21578, USENET, DigiTrad, and 20-Newsgroups text corpora.

The remainder of this paper is organized as follows. In the Section 2, we briefly describe how WordNet relationships can be used to augment category models. In Section 3, we review related work. In Section 4, we present experimental results from a comprehensive series of classification tasks. In Section 5, we discuss areas of future work that need to be considered within the context of using WordNet in text classification tasks. We conclude in Section 6, with a summary of our findings.

2 Background

In WordNet, synsets are connected according to a number of lateral, hierarchical, and compositional relationships. In this work, we are concerned with the relationships defined between noun synsets, which we now review, as follows. A synonym is a lateral relationship where a concept X is similar to a concept Y (i.e., an X is a Y and a Y is an X). A hypernym is a hierarchical relationship where a concept Y is a superclass of a concept X (i.e., every X is a kind of Y). A hyponym is a hierarchical relationship where a concept X is a subclass of a concept Y (i.e., a Y includes every X). A meronym is a compositional relationship where a concept X is a component of a concept Y (i.e., an X is part of Y). A holonym is a compositional relationship where a concept Y has a concept X as a component (i.e., a Y has an X).

To illustrate the general approach to text classification incorporating WordNet features, we present a simple, representative classification task. Sample synsets for the concepts `pup`, `dog`, and `cat` are shown in Figure 1 and sample documents are shown in Figure 2. In Figure 1, the sample synsets are connected to other synsets by the various WordNet relationships. In Figure 2, the documents are shown to contain keywords that will be used to facilitate classification. `Doc1` is the only labeled document and represents the training set. Our task is to attempt to automatically assign a category label to `Doc2` through `Doc5`.

We begin by attempting to classify the documents without incorporating WordNet features. Without WordNet features, `CategoryModel (Pets)` = ⟨`dog`⟩. Since `dog` is not found in ⟨`Words (Doc2), Words (Doc3), Words (Doc4), Words (Doc5)`⟩, we are unable to classify any of the documents.

We now consider situations where WordNet features are incorporated. For example, when incorporating synonyms, we have `CategoryModel (Pets)` = ⟨`dog, pup`⟩. That is, from the `dog` synset in Figure 1, we extract the synonym `pup` and add it to the category model for `Pets`. In this case, `pup` is found in ⟨`Words (Doc2)`⟩, so `Label (Doc2) = Pets`. The synonym relationship enabled the classification of a document that did not refer to any concepts found in the training set, but did contain a similar concept.

Similarly, to assign the same category label to documents that refer to more specific concepts, hyponyms can be incorporated. For example, when incorporating synonyms and hyponyms, we have `CategoryModel (Pets)` = ⟨`dog, pup, Terrier, black`⟩. Since `pup` is found in ⟨`Words (Doc2)`⟩, `black` is found in ⟨`Words (Doc3)`⟩, and `Terrier` is found in ⟨`Words (Doc4)`⟩, then `Label (Doc2) = Label (Doc3) = Label (Doc4) = Pets`.

3 Related Work

WordNet has been applied to a variety of problems in machine learning, natural language processing, information retrieval, and artificial intelligence (WordNet 2005). In this section, we discuss a number of relevant contributions that describe approaches to incorporating WordNet semantic features into text classifiers.

One of the first efforts toward the integration of WordNet features into a text classifier is described in (de Buenaga Rodriguez et al. 1997). Here it is proposed that accuracy may be increased if the category model for a document is expanded by incorporating WordNet synonyms of the category label. In this work, since the number of features actually incorporated by WordNet expansion was small, manual word sense disambiguation was used to determine the correct word sense. To evaluate their approach, Rocchio and Widrow-Hoff classification algorithms were used. It was found that accuracy, in general, was

Table 1: The 15 Document Collections Used

Corpora	*Category 1*	*No. of Documents*	*Category 2*	*No. of Documents*
USENET	Micro	163	Neuro	117
USENET	Taxes	170	History	79
Reuters-21578	Corn	168	Wheat	221
Reuters-21578	Livestock	113	Gold	134
DigiTrad	Marriage	200	Murder	224
Digitrad	Politics	194	Religion	238
20-Newsgroups	Atheism	200	Graphics	200
20-Newsgroups	PC	200	Macintosh	200
20-Newsgroups	Cryptography	200	Medicine	200
20-Newsgroups	MS Windows	200	X Windows	200
20-Newsgroups	Automobiles	200	Baseball	200
20-Newsgroups	Motorcycles	200	Guns	200
20-Newsgroups	Hockey	200	Religion	200
20-Newsgroups	Mideast	200	Electronics	200
20-Newsgroups	Space	200	Forsale	200

increased by incorporating synonyms, and, in particular, was increased when the number of categories in the training documents was sparse.

In (Scott & Matwin 1998), an approach is described where all words found in a document are considered for WordNet expansion rather than just the category label. Here, however, both synonyms and hypernyms of a category label are incorporated into the category model. A different representation of a category model is proposed where the features actually correspond to WordNet synsets rather than words. No word sense disambiguation is done in this approach, rather all senses of a word are incorporated into the category model. Using the RIPPER classification algorithm for evaluation of their approach, results were mixed, showing both statistically significant increases and decreases on various document collections.

A similar approach incorporating both synonyms and hypernyms is proposed in (Jensen & Martinez 2000). Noting that words in a synset are organized in occurrence frequency order, in their approach to word sense disambiguation, they only select the most likely sense for incorporation into the category model. Coordinate Matching, TF*IDF, and Naive Bayes classification algorithms were used to evaluate their approach, where different combinations of synonyms, hypernyms, and bigrams were incorporated into the category models. They found that incorporating hypernyms into category models is almost always appropriate.

The work described in (Kehagias et al. 2003) evaluates the merits of modeling senses as features rather than words. The Brown Semantic Corpus, a document collection whose words have been tagged with the correct word sense, is used such that only synsets corresponding to the features found in the document are incorporated into the category model. Consequently, hypernyms are not incorporated in this approach. Of course, word sense disambiguation is not necessary since the document collection has previously been tagged with the correct sense. They used MAP, Naive Bayes, and $k-$NN classifiers to evaluate their approach. An increase in accuracy was obtained on most document collections considered. However, the increases were small, leading the authors to conclude that the benefits from using their approach are marginal.

The application of WordNet as an ontology in text classification problems is explored in (Hotho & Bloehdorn 2004). Their approach incorporates both synonyms and hypernyms in the category model. Three different word sense disambiguation strategies are studied in their approach. These include strategies incorporating all senses and the most likely sense. A third strategy, context, measures the degree of overlap of different WordNet features in relation to how close these features occur to one another in the document being classified. Using an AdaBoost classifier, an increase in accuracy is reported on most document collections considered.

In (Rosso et al. 2004), an approach is proposed whereby vector of words in a category model are replaced with a vector of WordNet synsets. Manual word sense disambiguation is done before classification. A $k-$NN classifier showed an increase in accuracy on most document collections considered.

An approach is proposed in (Peng & Choi 2005) where synonyms, hypernyms, and hyponyms are incorporated into a category model as well as features found in the document. Category models attempt to capture the relationship between synsets rather than simply measuring the density of the various WordNet features. A variation of the most likely sense disambiguation strategy is used where the frequency of each sense in the context of the document collection is considered before WordNet expansion. Using a TF*IDF classifier where only hypernyms are incorporated into a category model, increases in accuracy on particular document collections are obtained that greatly exceed those reported by any other authors. However, the methodology seems simplistic and is not well-documented (e.g., was 10-fold cross-validation used, for instance), so the reported results are suspect.

4 Experimental Results

We implemented a text classifier shell that can incorporate the various WordNet features into a category model. It was constructed so that the classification algorithm and other features can be configured at run time. The shell was implemented in Visual C++ Version 6.0 and run under Windows XP on an IBM compatible PC with a 3.0 GHz Pentium 4 processor and 1 GB of memory. The Naive Bayes and SVM algorithms that we used are part of the Weka 3.4 open source software issued under the GNU General Public License.

4.1 Document Collections

The classification tasks were run on 15 different document collections drawn from four different text corpora: Reuters-21578, USENET, DigiTrad, and 20-Newsgroups. These particular document collections have been used extensively in previous text classification studies (e.g., see Related Work). The 15 document collections are shown below in Table 1. In Table 1, the *Corpora* column describes the origin of the corresponding document collections, the *Category 1* and *Category 2* columns describe a single semantic label attached to documents in the dataset, and the

Table 2: Relative Change in Accuracy from Incorporating Synonyms, Hypernyms, Hyponyms, Meronyms, and Holonyms using the Naive Bayes Classifier

Dataset	*Base*	*Syn*	*Syn+Hypr*	***Syn+Hypo***	***Syn+Mero***	*Syn+Holo*
Livestock/Gold	97.02	-0.38	+1.23	**-2.06**	**0.00**	+0.53
Micro/Neuro	61.07	+0.57	+5.26	**-1.25**	**-1.53**	-0.73
Murder/Marriage	75.72	0.00	-1.27	**-2.11**	**-2.84**	-1.00
Political/Religion	80.13	+1.23	-1.16	**-5.03**	**-1.83**	+0.05
Taxes/History	87.81	-0.71	-1.60	**-16.34**	**-19.58**	-3.11
Wheat/Corn	75.53	+0.60	-3.57	**-4.09**	**+2.14**	+0.83
Atheism/Graphics	89.00	+1.25	+3.00	**-8.25**	**-1.00**	+0.75
PC/Mac	71.50	-0.50	-3.50	**-4.25**	**0.00**	+0.75
Crypt/Medicine	87.75	-0.75	+0.25	**-10.25**	**-1.75**	+0.75
MS Win/X Win	67.25	-1.25	-6.25	**-4.75**	**-2.00**	-0.50
Autos/Baseball	91.50	-0.25	+1.50	**-3.00**	**-3.25**	+1.00
Motorcycles/Guns	87.00	+3.00	+4.00	**-6.00**	**0.00**	+2.25
Hockey/Religion	96.00	-1.25	-3.75	**-8.25**	**-6.75**	-1.75
Mideast/Electro	90.75	+1.00	+1.50	**-9.25**	**-2.00**	+1.00
Space/Forsale	79.75	-1.75	-2.00	**-7.50**	**-3.25**	-1.00

Table 3: Relative Change in Accuracy from Incorporating Synonyms, Hypernyms, Hyponyms, Meronyms, and Holonyms using the SVM Classifier

Dataset	*Base*	*Syn*	*Syn+Hypr*	***Syn+Hypo***	*Syn+Mero*	*Syn+Holo*
Livestock/Gold	97.07	+0.41	+0.41	**-3.58**	-0.89	+0.02
Micro/Neuro	64.14	-1.67	+0.14	**+0.12**	+0.91	-1.22
Murder/Marriage	83.40	-1.29	-0.36	**-1.68**	-2.04	-0.36
Political/Religion	82.44	-0.26	+1.85	**-4.23**	-0.79	+0.84
Taxes/History	71.97	+2.14	+5.84	**+8.11**	+6.26	+1.85
Wheat/Corn	84.77	+1.16	+0.94	**-1.05**	+1.30	+2.07
Atheism/Graphics	92.75	+0.25	-0.25	**-4.25**	-1.25	0.00
PC/Mac	76.75	-0.75	-2.00	**-8.75**	-3.50	-2.00
Crypt/Medicine	83.50	-0.75	0.00	**-1.50**	-0.85	+0.50
MS Win/X Win	81.75	+0.25	+1.50	**-2.75**	-0.75	+1.00
Autos/Baseball	93.50	+0.25	-1.00	**-3.00**	-0.25	+0.50
Motorcycles/Guns	89.25	+0.50	+1.00	**-1.50**	+1.25	+0.75
Hockey/Religion	99.75	0.00	0.00	**0.00**	0.00	+0.00
Mideast/Electro	96.25	-1.25	-4.00	**-3.75**	-3.75	-1.75
Space/Forsale	88.50	+0.75	+0.50	**-2.25**	-1.00	+1.00

No. of Documents columns describes the number of documents assigned to the corresponding category.

The USENET and 20-Newsgroups text corpora are collections of postings to newsgroups. A document is assigned a category based upon the newsgroup to which it is posted. The Reuters-21578 text corpora is a collection of Reuters news stories where a story is assigned to a category based upon the specific topic of the story. The DigiTrad text corpora is a collection of folk song lyrics. A song is assigned to a category based upon the theme of the song. The documents contained in each category consist of a randomly selected subset of documents selected from the original corpora.

4.2 Methodology

Text classification in this, and other work, is typically a two-step process. In the first step, a category model for each category in a corpus of labeled training documents is built. In the second step, a classification algorithm compares unlabeled documents to the learned category model to assign a category label to the unlabeled documents.

In text classification, the characteristics of the environment under which an algorithm is run, can affect the results. For example, something as simple as using a different stoplist can affect the accuracy results generated by two otherwise identical text classification tasks. In this work, all text classification tasks were run according to the following general criteria:

- The same stoplist was used all runs and the corresponding words were removed from the training and evaluation sets for each run.
- 10-fold cross validation was used was used for each run.
- Accuracy was calculated as the number of correct classifications divided by the total number of classifications.
- If it was determined that a document was equally likely to belong to either category (i.e., a tie), this was considered an incorrect classification.
- A word is considered any sequence of alphabetic characters surrounded by non-alphabetic characters such as spaces, numbers, and punctuation.
- A word is always converted to its morphological base using the WordNet morphword function.

For each series of runs, the accuracy obtained for various configurations of the classifier is compared to the accuracy obtained by a "base" classifier (i.e., one whose configuration remains fixed). In the "base" classifier, when part of speech tags were available, only nouns were used for WordNet expansion, all WordNet queries returned the most likely sense of a word, and term frequency weighting was used.

4.3 The Effect of WordNet Features

In this section, we present the results obtained from a series of runs where the effect of incorporating the various WordNet features into the category models for the Naive Bayes and SVM classifiers was evaluated. Incorporating synonyms, hypernyms, and hyponyms into category models has previously been found to increase accuracy in some cases (de Buenaga Rodriguez et al. 1997), (Hotho & Bloehdorn 2004), (Jensen & Martinez 2000), (Kehagias et al. 2003), (Peng & Choi 2005), (Rosso et al. 2004), (Scott & Matwin 1998). We were the first to extend the use of WordNet in text classification by considering meronyms and holonyms (Mansuy & Hilderman 2006). In (Mansuy & Hilderman 2006), we found that incorporating these features into category models increases accuracy in some cases.

Table 4: Relative Change in Accuracy With and Without Part of Speech Tags using the Naive Bayes Classifier

		Syn		*Syn+Hypr*		***Syn+Hypo***		*Syn+Mero*		*Syn+Holo*	
Dataset	*Base*	*POS*	*No POS*	*POS*	*No POS*	***POS***	***No POS***	*POS*	*No Pos*	*POS*	*No POS*
Livestock/Gold	97.02	-0.38	-0.38	+1.23	+0.77	**-2.06**	**-2.13**	0.00	+0.39	+0.53	+0.77
Micro/Neuro	61.07	+0.57	-0.45	+5.26	+4.49	**-1.25**	**-2.32**	-1.53	-1.05	-0.73	-1.96
Murder/Marriage	75.72	0.00	-1.38	-1.27	+0.16	**-2.11**	**-2.22**	-2.84	-3.34	-1.00	-2.91
Political/Religion	80.13	+1.23	+0.05	-1.16	+1.60	**-5.03**	**-5.99**	-1.83	-2.05	+0.05	-0.39
Taxes/History	87.81	-0.71	+0.04	-1.60	-6.47	**-16.34**	**-18.07**	-19.58	-19.58	-3.11	-2.81
Wheat/Corn	75.53	+0.60	+1.31	-3.57	-1.78	**-4.09**	**4.14**	+2.14	+3.30	+0.83	+1.00

Table 5: Relative Change in Accuracy With and Without Part of Speech Tags using the SVM Classifier

		Syn		*Syn+Hypr*		***Syn+Hypo***		*Syn+Mero*		*Syn+Holo*	
Dataset	*Base*	*POS*	*No POS*	*POS*	*No POS*	*POS*	***No POS***	*POS*	*No Pos*	*POS*	*No POS*
Livestock/Gold	97.07	+0.41	+0.41	+0.41	-0.05	-3.58	**-1.59**	-0.89	-0.89	+0.02	+0.41
Micro/Neuro	64.14	-1.67	-2.61	+0.14	+0.43	+0.12	**-2.36**	+0.91	+0.12	-1.22	-2.30
Murder/Marriage	83.40	-1.29	-0.79	-0.36	+0.87	-1.68	**-0.86**	-2.04	-1.31	-0.36	+0.69
Political/Religion	82.44	-0.26	-0.48	+1.85	+1.72	-4.23	**-2.31**	-0.79	-0.21	+0.84	-0.87
Taxes/History	71.97	+2.14	+3.28	+5.84	+4.54	+8.11	**+5.38**	+6.26	+7.27	+1.85	+2.98
Wheat/Corn	84.77	+1.16	+1.70	+0.94	+1.79	-1.05	**+2.27**	+1.30	+2.87	+2.07	+3.32

The accuracy obtained using the Naive Bayes and SVM classifiers with each combination of WordNet feature is shown below in Tables 2 and 3. In Tables 2 and 3, the *Base* column describes the accuracy obtained when no WordNet features are incorporated into the classifier. The *Syn*, *Syn+Hypr*, *Syn+Hypo*, *Syn+Mero*, and *Syn+Holo* columns describe the relative difference in accuracy obtained when the various WordNet features are incorporated. For example, in Table 2, the accuracy obtained for Micro/Neuro in the *Syn+Hypr* column is 66.33, which we show as relative increase of +5.26 over the 61.07 shown in the *Base* column. In all of the results that follow, columns shown in bold represent a significant difference in accuracy from those obtained by the base classifier. The Wilcoxon Signed Rank Test was used to determine statistical significance using a 90% level of significance (i.e., $\alpha = 0.10$) and a null hypothesis that there is no difference between medians (i.e., a two-tailed test).

In Table 2, the *Syn+Hypo* and *Syn+Mero* columns show a statistically significant decrease in accuracy from the *Base* classifier. The accuracy in the *Syn+Hypo* column decreased for all 15 datasets. In the *Syn+Mero* column, accuracy decreased for 11 of the 15 datasets, remained the same for three, and increased for one. In Table 3, the *Syn+Hypo* column shows a statistically significant decrease in accuracy from the *Base* classifier. The accuracy decreased in 12 of the 15 datasets, remained the same in one and increased in two. There was no statistically significant difference between the accuracy of the *Base* classifier and the accuracy shown in any of the other columns.

In previous work by other authors, where synonyms and hypernyms are incorporated in RIPPER (Scott & Matwin 1998), AdaBoost (Hotho & Bloehdorn 2004), Naive Bayes (Jensen & Martinez 2000), (Kehagias et al. 2003), and $k-$NN (Kehagias et al. 2003), (Rosso et al. 2004) classifiers, they suggest that these WordNet features will increase the accuracy of the classifiers, but their results are not validated statistically. In (Kehagias et al. 2003), where synonyms and hypernyms are incorporated into both a Naive Bayes and a $k-$NN classifier, it is suggested that while these WordNet features usually result in an increase in accuracy, the increases are not large enough to justify the additional complexity.

The results reported in the previous paragraph are contrary to our results showing that, generally across many datasets, incorporating synonyms and hypernyms (and synonyms and holonyms) can be expected to make no difference in the accuracy obtained from a Naive Bayes classifier. Further, incorporating synonyms and hyponyms, and synonyms and meronyms can be expected generally, to decrease accuracy across many datasets.

As far as the authors of this paper know, there is no previous research incorporating WordNet features of any kind into an SVM classifier, other than some preliminary work reported in (Mansuy & Hilderman 2006). Consequently, there are no comparative results available. However, based on our results alone, incorporating combinations of synonyms and hypernyms, meronyms, and holonyms can be expected to make no difference in the accuracy obtained from an SVM classifier, and synonyms and hyponyms can be expected to decrease accuracy.

4.4 The Effect of Part of Speech Tags

In this section, we present the results obtained from a series of runs where the effect of part of speech tags were evaluated. WordNet provides facilities for returning only those terms relevant to the part of speech in which a word occurs in the text document. Part of speech tags were available in only six of the 15 datasets: Livestock/Gold, Micro/Neuro, Murder/Marriage, Politics/Religion, Taxes/History, and Wheat/Corn.

The accuracy obtained using the Naive Bayes and SVM classifiers, with and without utilizing the part of speech tags from the six datasets, is shown below in Tables 4 and 5. In Tables 4 and 5, the columns have the same meaning as previously described. The *POS* and *No POS* columns describe the relative difference in accuracy from incorporating WordNet features with and without utilizing part of speech tags, respectively. The values in the *POS* columns of Tables 4 and 5 contain the same values as shown in the *Syn*, *Syn+Hypr*, *Syn+Hypo*, *Syn+Mero*, and *Syn+Holo* columns in Tables 2 and 3, respectively, because part of speech tags (when available) were part of the base configuration. These values are repeated here for reader convenience.

In Tables 4 and 5, there is no significant difference between the accuracy of the classifiers, whether part of speech tags are utilized or not, and the base classifiers, except in the *Syn+Hypo No POS* column, where a statistically significant decrease in accuracy was observed.

While part of speech was considered in some previous work (Jensen & Martinez 2000), (Scott & Matwin 1998), it is not clear or explicitly stated in other work whether part of speech tags were utilized (de Buenaga Rodriguez et al. 1997), (Hotho & Bloehdorn 2004), (Kehagias et al. 2003), (Peng & Choi 2005), (Rosso et al. 2004), (Wiebe & O'Hara 2003). As far as the authors of this paper know, there is no previous research evaluating the accuracy

Table 6: Relative Change in Accuracy for Boolean and Term Frequency Weighting Schemes using the Naive Bayes Classifier

	Base		*Syn*		*Syn+Hypr*		***Syn+Hypo***		***Syn+Mero***		*Syn+Holo*	
Dataset	*BW*	*TW*	*BW*	*TW*	*BW*	*TW*	***BW***	***TW***	***BW***	***TW***	*BW*	*TW*
Livestock/Gold	99.23	97.02	-0.84	-0.38	+0.77	+1.23	**-8.43**	**-2.06**	**-0.39**	**0.00**	0.00	+0.53
Micro/Neuro	64.85	61.07	+4.26	+0.57	+7.39	+5.26	**+0.49**	**-1.25**	**+4.92**	**-1.53**	+4.72	-0.73
Murder/Marriage	85.34	75.72	-0.28	0.00	-0.59	-1.27	**-11.34**	**-2.11**	**-5.80**	**-2.84**	-0.39	-1.00
Political/Religion	86.60	80.13	+0.26	+1.23	-0.09	-1.16	**-8.86**	**-5.03**	**-4.75**	**-1.83**	+0.13	+0.05
Taxes/History	93.23	87.81	+2.98	-0.71	+3.15	-1.60	**-19.54**	**-16.34**	**-11.85**	**-19.58**	-1.30	-3.11
Wheat/Corn	86.60	75.53	+1.09	+0.60	-0.81	-3.57	**-10.70**	**-4.09**	**+2.26**	**+2.14**	+3.48	+0.83
Atheism/Graphics	95.00	89.00	0.00	+1.25	-0.25	+3.00	**-14.25**	**-8.25**	**+0.25**	**-1.00**	+0.25	+0.75
PC/Mac	81.00	71.50	-3.50	-0.50	-3.00	-3.50	**-14.25**	**-4.25**	**-4.75**	**0.00**	-2.25	+0.75
Crypt/Medicine	90.00	87.75	+0.25	-0.75	+0.75	+0.25	**-9.00**	**-10.25**	**-2.50**	**-1.75**	0.00	+0.75
MS Win/X Win	81.50	67.25	-1.75	-1.25	-5.25	-6.25	**-15.75**	**-4.75**	**-5.25**	**-2.00**	-2.75	-0.50
Autos/Baseball	96.25	91.50	+0.25	-0.25	+0.50	+1.50	**-5.00**	**-3.00**	**-3.00**	**-3.25**	-0.50	+1.00
Motorcycles/Guns	92.75	87.00	+0.25	+3.00	-1.25	+4.00	**-11.50**	**-6.00**	**-2.25**	**0.00**	-0.25	+2.25
Hockey/Religion	100.00	96.00	0.00	-1.25	-0.50	-3.75	**-12.25**	**-8.25**	**-0.50**	**-6.75**	0.00	-1.75
Mideast/Electro	89.50	90.75	+0.25	+1.00	0.00	+1.50	**-6.50**	**-9.25**	**+1.75**	**-2.00**	+1.00	+1.00
Space/Forsale	88.50	79.75	-1.00	-1.75	-5.00	-2.00	**-8.25**	**-7.50**	**-2.75**	**-3.25**	-2.00	-1.00

Table 7: Relative Change in Accuracy for Boolean and Term Frequency Weighting Schemes using the SVM Classifier

	Base		*Syn*		*Syn+Hypr*		***Syn+Hypo***		***Syn+Mero***		*Syn+Holo*	
Dataset	*BW*	*TW*	*BW*	*TW*	*BW*	*TW*	***BW***	***TW***	***BW***	*TW*	*BW*	*TW*
Livestock/Gold	100.00	97.07	0.00	+0.41	0.00	+0.41	**-0.46**	**-3.58**	**0.00**	-0.89	0.00	+0.02
Micro/Neuro	66.10	64.14	+0.29	-1.67	-0.57	+0.14	**-1.62**	**+0.12**	**+1.51**	+0.91	+3.18	-1.22
Murder/Marriage	87.84	83.40	-0.64	-1.29	-0.66	-0.36	**-3.00**	**-1.68**	**-2.48**	-2.04	-1.28	-0.36
Political/Religion	82.18	82.44	+0.09	-0.26	+0.58	+1.85	**-4.24**	**-4.23**	**-0.87**	-0.79	-0.82	+0.84
Taxes/History	93.40	71.97	+1.13	+2.14	-6.09	+5.84	**-4.54**	**+8.11**	**-0.76**	+6.26	+0.54	+1.85
Wheat/Corn	96.98	84.77	+0.63	+1.16	-0.71	+0.94	**-1.39**	**-1.05**	**-0.68**	+1.30	+0.17	+2.07
Atheism/Graphics	94.00	92.75	-0.50	+0.25	+0.75	-0.25	**-2.25**	**-4.25**	**-1.50**	-1.25	-0.50	0.00
PC/Mac	82.75	76.75	-3.75	-0.75	-6.00	-2.00	**-9.25**	**-8.75**	**-4.50**	-3.50	-3.25	-2.00
Crypt/Medicine	91.25	83.50	-0.50	-0.75	+2.25	0.00	**-2.75**	**-1.50**	**-0.25**	-0.85	-0.25	+0.50
MS Win/X Win	88.25	81.75	-1.00	+0.25	-3.50	+1.50	**-7.00**	**-2.75**	**-2.00**	-0.75	-0.75	+1.00
Autos/Baseball	94.00	93.50	0.00	+0.25	0.00	-1.00	**-4.00**	**-3.00**	**-2.50**	-0.25	+0.50	+0.50
Motorcycles/Guns	91.75	89.25	+0.75	+0.50	+3.75	+1.00	**+0.50**	**-1.50**	**0.00**	+1.25	+1.00	+0.75
Hockey/Religion	99.75	99.75	0.00	0.00	+0.25	0.00	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00
Mideast/Electro	93.75	96.25	-5.75	-1.25	-0.25	-4.00	**-7.00**	**-3.75**	**-4.75**	-3.75	-6.75	-1.75
Space/Forsale	91.75	88.50	-0.50	+0.75	-1.75	+0.50	**-1.25**	**-2.25**	**-1.00**	-1.00	-0.50	+1.00

of Naive Bayes and SVM classifiers utilizing part of speech tags. But again, based on our results alone, utilizing part of speech tags can be expected to make no difference to the Naive Bayes or SVM classifiers.

4.5 The Effect of Term Weighting Schemes

In this section, we present the results obtained from a series of runs where the effect of term weighting schemes was evaluated. During WordNet expansion, the importance of each term added to a category model is represented by a value called its weight. Here we evaluate Boolean weighting and term frequency weighting schemes.

Boolean weighting considers only the occurrence or non-occurrence of a word in a document, with no regard to the number of times the word actually occurs. Thus, with Boolean weighting, all words are considered to have equal importance. In contrast, term frequency weighting not only considers the occurrence of a word in a document, it also considers how often the word occurs.

The accuracy obtained using the Naive Bayes and SVM classifiers, and the two term weighting schemes is shown in Tables 6 and 7. In Tables 6 and 7, the columns have the same meaning as previously described. The *BW* and *TW* columns describe the relative difference in accuracy for the Boolean weighting and term frequency weighting schemes, respectively. The values in the *TW* columns of Tables 6 and 7. contain the same values as shown in the *Syn*, *Syn+Hypr*, *Syn+Hypo*, *Syn+Mero*, and *Syn+Holo* columns in Tables 2 and 3, respectively, because term frequency weighting is part of the base configuration. Again, these are repeated here for reader convenience.

In Tables 6 and 7, there is a statistically significant decrease in accuracy in the *Syn+Hypo BW* and *Syn+Mero BW* columns for both the Naive Bayes and SVM classifiers. The statistically significant difference in the *Syn+Hypo TW* and *Syn+Mero TW* columns for both classifiers is simply repeating the results shown in Tables 2 and 3. Although not shown in bold to avoid confusion, in Tables 6 and 7, there is a statistically significant increase in accuracy in the *Base BW* in relation to the *Base TW* column. Some of the increases are quite large (e.g., 21.43% for Taxes/History in Table 7).

5 Discussion and Future Work

The results reported in the previous section contradict the results previously reported by others. In an effort to understand why, further analysis is currently ongoing and we hope to soon have a precise explanation supported by solid empirical evidence. In the meantime, we speculate on possible factors, as follows.

- **Word Sense Disambiguation:** The lack of an intelligent word sense diambiguation scheme is a potential source for increasing classification error. For example, the default scheme used in this work was the most likely sense. However, in each case where the most likely sense is not the correct sense (i.e., the sense being used in the text is some other sense), then this will result in poor information being added to the model. The worst case scenario would occur when two documents belonging to different classes use different senses of the same word. In this case, information being added to the models has the potential to actually reduce the discriminating power of the classifier.
- **Context:** The context within which a WordNet feature is added to a model could be a potential source of problems. That is, since WordNet can

be considered a general ontology for English, it does not contain any domain-specific knowledge. For example, in an agricultural context, the concept *field* should have a strong relationship to other agricultural concepts related to a *field*. But in baseball, the concept *field* has a strong relationship to a totally different set of concepts. However, WordNet is not able to capture this kind of domain-specific knowledge. Thus, even if WordNet does model a concept accurately for one class, this can be offset by a poor model for the other class.

- **Feature Weighting:** WordNet does not provide a feature weighting scheme. However, an attempt at feature weigthing was addressed in (Jensen & Martinez 2000), where a hypernym feature was weighted according to the depth in which it occurred in the WordNet hierarchy. When WordNet is used to augment a classification task, many features are generally added to a model. However, these features are being added only for those words in a document that can be found in the WordNet hierarchy. For example, features like *diamond*, *ball*, and *bat* could be added to a model dealing with baseball, while not adding features for related proper nouns (that WordNet does not contain) like *Babe Ruth* and *Yankee Stadium*. Proper nouns like these can be quite descriptive, but eventually, their relative discriminative power may be overwhelmed in the model by all the WordNet features. Perhaps a weighting scheme that gives special weight to features that don't exist in WordNet should be investigated.

- **Feature Selection:** A feature selection scheme that prevents the loss in accuracy from adding too many poor WordNet features may be appropriate for reducing classification error. For example, adding many features to a model increases the number of dimensions in a classification problem. In many cases, such as when adding a very general hypernym, the feature space represented by each model is likely to overlap on those dimensions related to the general hypernym, making the model less discriminative.

6 Conclusion

We evaluated the effect on accuracy of incorporating features from WordNet relationships, part of speech tags, and term weighting schemes. We found that incorporating the various WordNet features and part of speech tags had no statistically significant positive effect on the accuracy of the Naive Bayes and SVM classifiers. Similarly, in combination with the various WordNet features, there was no statistically significant positive effect on accuracy for either the Boolean or term frequency weighting schemes. However, there was a statistically significant increase for Boolean weighting in relation to term frequency weighting.

References

de Buenaga Rodriguez, M., Gomez-Hidalgo, J. & Diaz-Agudo, B. (1997), Using WordNet to complement training information in text categorization, *in* 'Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP'97)', Tzigov Chark, Bulgaria, pp. 150–157.

Hotho, A. & Bloehdorn, S. (2004), Boosting for text classification with semantic features, *in* 'Proceedings of the Workshop on Mining for and from the Semantic Web at the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)', Seattle, U.S.A., pp. 70–87.

Jensen, L. & Martinez, T. (2000), Improving text classification by using conceptual and contextual features, *in* 'Proceedings of the Workshop on Text Mining at the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00)', Boston, U.S.A., pp. 101–102.

Joachims, T. (1998), Text categorization with support vector machines: Learning with many relevant features, *in* 'Proceedings of the European Conference on Machine Learning', pp. 137–142.

Kehagias, A., Petridis, V., Kaburlasos, V. & Fragkou, P. (2003), 'A comparison of word- and sense-based text classification using several classification algorithms', *Journal of Intelligent Information Systems* **21**(3), 227–247.

Lewis, D. D. (1998), Naive (bayes) at forty: The independence assumption in information retrieval, *in* 'Proceedings of the Tenth European Conference on Machine Learning', London, UK, pp. 4–15.

Mansuy, T. & Hilderman, R. (2006), Evaluating WordNet features in text classification models, *in* 'Proceedings of the 19th International Florida Artificial Intelligence Research Symposium (FLAIRS'06)', Melbourne Beach, U.S.A. To appear.

Miller, G. (1995), 'WordNet: A lexical database for English', *Communications of the ACM* **38**(11), 39–41.

Peng, X. & Choi, B. (2005), Document classifications based on word semantic hierarchies, *in* 'Proceedings of the International Conference on Artificial Intelligence and Applications (AIA'05)', Innsbruck, Austria, pp. 362–367.

Rosso, P., Ferretti, E., Jiminez, D. & Vidal, V. (2004), Text categorization and information retrieval using WordNet senses, *in* 'Proceedings of the 2nd Global Wordnet Conference (GWC'04)', Brno, Czech Republic, pp. 299–304.

Scott, S. & Matwin, S. (1998), Text classification using WordNet hypernyms, *in* 'Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems (Coling-ACL'98)', Montreal, Canada, pp. 45–52.

Tan, C., Wang, Y. & Lee, C. (2002), 'The use of bigrams to enhance text categorization', *Information Processing and Management: An International Journal* **38**(4), 529–546.

Wiebe, J. & O'Hara, T. (2003), Classifying functional relations in factotum via WordNet hypernym associations, *in* 'Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'03)', Mexico City, Mexico, pp. 347–359.

WordNet (2005), WordNet Bibliography. [http://mira.csci.unt.edu/wordnet].

Weighted Kernel Model for Text Categorization

Lei Zhang **Debbie Zhang** **Simeon J. Simoff** **John Debenham**

Faculty of Information Technology
University of Technology, Sydney
PO Box 123 Broadway NSW 2007 Australia
Email{leizhang, debbiez, simeon, debenham}@it.uts.edu.au

Abstract

Traditional bag-of-words model and recent word-sequence kernel are two well-known techniques in the field of text categorization. Bag-of-words representation neglects the word order, which could result in less computation accuracy for some types of documents. Word-sequence kernel takes into account word order, but does not include all information of the word frequency. A weighted kernel model that combines these two models was proposed by the authors [1]. This paper is focused on the optimization of the weighting parameters, which are functions of word frequency. Experiments have been conducted with Reuter's database and show that the new weighted kernel achieves better classification accuracy.

Keywords: Bag-of-words Kernel, Word-sequence Kernel, Weighted Kernel Model, Text Categorization

1 Introduction

Text categorization is the task of assigning documents into predefined categories (classes), specified by their topics. For example, the documents might be news items and the classes might be national news, sports news and business news. Documents are classified based on their content[2]. As documents are characterized by the words that appear in each document, they are firstly transformed into a representation that is suitable for the classification task. Then learning algorithms are applied to perform the classification task. Automated text categorization has been successfully demonstrated in many applications [3] [4].

One of the widely used representation of documents is known as the bag-of-words model [5], which is a set of words contained in the documents. Bag-of-words is based on both frequency of a word in a document and the corpus. However, bag-of-words model has many shortcomings. Particularly, it ignores both syntax and semantics of the documents. Without considering the word position, the information of the sequence of words is lost.

Lodhi proposed the use of string kernels [6], which was the first significant departure from the bag-of-words model. In string kernels, the features are not word frequencies or related expansions, but the extent to which all possible ordered subsequences of characters are presented in the document. Cancedda [7] proposed the use of string kernel with sequences of words rather than characters, known as the word-sequence kernel. The word-sequence kernel has several advantages, in particular it is more computationally efficient and it ties in closely with standard linguistic pre-processing techniques. Although word-sequence kernel takes into account word positions, it does not include the information about word frequency. This issue will be discussed further in Section 3.

To make both the word frequency and position information available for the learning algorithms, a combined weighted kernel model was proposed [1]. However, not every combination of the bag-of-words approach and word-sequence kernel approach will result in improved computational accuracy. This is because these two kernels, which represent the similarity between documents respectively, have different contribution to construct the new kernel. Moreover simply combining these two kernels do not satisfy valid kernel conditions [8].

This paper is based on our previous work of combining kernels. It emphasizes on the optimization of the weighting parameters, which is critical to the categorization accuracy. The rest of this paper is organized as follows: In Section 2, the basic idea of bag-of-words kernel is reviewed. The word-sequence kernel and issues of this kernel are presented in Section 3. The detail implementation of proposed new kernel model and algorithm for determining the weighting parameter are described in Section 4. Section 5 presents the experimental results using the Reuters data set, followed by the conclusions in Section 6.

2 Bag-of-words Kernel

Bag-of-words model is the traditional approach for representing a document as a term vector. A bag is a set of a dictionary. As repeated elements are allowed, this representation takes into account not only the presence of a word but also its frequency [9]. A document is represented by a row vector

$$\phi(d) = [tf(t_1, d), tf(t_2, d), ..., tf(t_N, d)] \in \mathbb{R}^N \quad (1)$$

where $tf(t_i, d)$ is the frequency of the term t_i in the document d. Hence, a document is mapped into a space of dimensionality N being the size of the dictionary. Each entry records how many times a particular term is used in the document. The vector space kernel or bag-of-words kernel is given by the following definition

$$k(d_1, d_2) = \langle \phi(d_1), \phi(d_2) \rangle = \sum_{j=1}^{N} tf(t_j, d_1) tf(t_j, d_2) \quad (2)$$

The value of N in equation 1 is related to the length of the documents. Excessive number of irrelevant words and terms will not only increase the computational cost but also decrease the accuracy of the classification. Therefore the most frequent words and

terms are usually selected in order to construct the bag-of-words kernel. As mentioned in Section 1, this technique only takes into account the word frequencies, ignoring the information on word positions. In many language modeling applications, such as speech recognition and short message classification, word order is extremely important. Furthermore, it is likely that word order can assist in topic inference. For example, consider the following two sentences

"The interest rate goes up, US dollar goes down."
"The interest rate goes down, US dollar goes up."

These two sentences [10] have exactly the same words and word frequencies. It is the different word order that results in opposite meanings, which cannot be distinguished by the bag-of-words method. Therefore the string kernel and word-sequence kernel were introduced to tackle with the word order issue.

3 Word-sequence Kernels

In string kernels, the features are not word frequencies. The document is represented by all possible ordered subsequences of characters. However, this method is computationally expensive for long documents. More recently, Cancedda et al. extended the string kernel to word-sequence kernel, where all possible sequences of words are used instead of sequences of characters. This novel way to compute the document similarity based on matching subsequence has outperformed the string kernel in many applications [11] [12].

Following the definition of Lodhi [6], let Σ be a finite alphabet set. A string is a finite sequence of characters from Σ, including the empty sequence. For strings s and t, l_s denotes the length of the string s, where $s = s_1...s_{l_s}$. The string $s[i:j]$ is the substring $s_i...s_j$ of s. u is a subsequence of s, if there exist indices $\mathbf{i} = (i_1, ..., i_{l_u})$, with $1 \leq i_1 < ... < i_{l_u} \leq l$, such that $u = s[\mathbf{i}]$. The length $l(\mathbf{i})$ of the subsequence in s is $i_{l_u} - i_1 + 1$. Σ^n denotes the set of all finite strings of length n, and Σ^* is the set of all strings $\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$. The feature mapping ϕ for a string s is given by defining the u coordinate $\phi_u(s)$ for each $u \in \Sigma^n$.

$$\phi_u(s) = \sum_{\mathbf{i}:u=s[\mathbf{i}]} \lambda^{l(\mathbf{i})} \tag{3}$$

where λ is the decay factor. These features measure the number of occurrences of subsequences in the string s weighting them according to their lengths. Hence, the inner product of the feature vectors for two strings s and t gives a sum over all common subsequences weighted according to their frequency of occurrence and lengths

$$K_n(s,t) = \sum_{u \in \Sigma^n} \langle \phi_u(s) \cdot \phi_u(t) \rangle = \sum_{u \in \Sigma^n} \sum_{\mathbf{i}:u=s[\mathbf{i}]} \sum_{\mathbf{j}:u=t[\mathbf{j}]} \lambda^{l(\mathbf{i})+l(\mathbf{j})} \tag{4}$$

Following the example of Lodhi, we examine different values of λ when n = 2, 3. We can compute the similarity between the following two sentences:

K("science is organized knowledge", "wisdom is organized life")

The similarity with $\lambda = 0.5$ was calculated by Lodhi, which were 0.580 when n equals to 2 and 0.478 when n equals to 3. We examine this algorithm by choosing different values of the λ as shown in the following table.

We could conclude that the string kernel algorithm is not much influenced by the value of λ. Therefore, the median value 0.5 was chosen for λ for the rest of experiments in this paper.

	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
kernel (n = 2)	0.557	0.580	0.620
kernel (n = 3)	0.483	0.478	0.483

Table 1: Result by using string kernel

	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$
BOW	1	1	1
n=2	0.874	0.837	0.825
n=3	0.492	0.487	0.558
BOW + n=2	0.999	0.992	0.994
BOW + n=3	0.999	0.993	0.878
BOW + n=2 + n=3	0.999	0.985	0.865

Table 2: Result by simply combination of two kernels

However string kernel and word-sequence kernel do not include the information of word frequency. Considering the term "kernel methods" and "kernel model", if we choose word-sequence kernel with $n = 2$, the similarity is zero. However, in many situation, these two terms are considered the same.

4 Weighted and Combined Kernel Model

Bag-of-words representation could be considered as a special case of word-sequence kernel. Consider the length $l = 0$, $\lambda^l = \lambda^0 = 1$. When $l = 0$, word-sequence kernel only contains the information of word itself, which is essentially the representation of bag-of-word.

However, the simply combination of $l = 0$ with word-sequence kernels is not a feasible approach. Let us consider the interest rate example in Section 2. We examine the bag-of-words and word-sequence kernels with n=2, 3, and combined n = 2, 3 with bag-of-word. The similarity of these two sentences is shown in the following table.

Simply combination of bag-of-word and word-sequence kernel will result in less computational accuracy as the entries of the element of these two kernel are in different scale. For example, when λ equals 0.5 and n equals 3, the similarity is 0.487, which makes sense while 0.99 is not properly. That is because the two sentences in the "interest rate" example are two related sentences, but the meaning is different. Given a high value near one is not properly, neither given a zero value. Therefore a value in the middle is very ideal. This is also where the idea come of choosing the parameter based on the word frequency information.

Moreover simply combine these two kernel wouldn't result in a new valid kernel, because the similarity of the same document may be greater than one.

Here we propose an approach to combine the word-sequence kernel and bag-of-word kernel.

$$K_{combined} = (1-\lambda) * K_{BOW} + \lambda * K_{Word-sequence} \tag{5}$$

The λ in the above formula is no longer the decay factor in Lodhi's word-sequence kernel. The parameter λ is now for balancing the contribution of the word frequency and word order information. It may be fixed, or determined from the data. Fixed parameters are more likely rely on the domain knowledge, which is different according to different projects. In this paper, we propose a technique to determine the parameter from the data by using the word frequency information.

Since the common understanding of a binary classification problem, the more words occurs in both

	BOW + n=2	BOW + n=3	BOW + n=2,3
$\lambda = 0.5$	0.862	0.695	0.792

Table 3: Result by using new combined kernel

Frequent Word	BOW	Word-sequence	Weighted Kernel
50	80 %	80%	85 %
100	80 %	80%	85 %

Table 4: Results on C15 and C22 data based on 100 Documents

Frequent Word	BOW	Word-sequence	Weighted Kernel
50	70%	60%	80 %
100	75%	60%	85 %

Table 5: Results on C21 and C22 data based on 100 Documents

classes, the less important the bag-of-words is. The ideal situation for bag-of-words would be no words occurs in two classes. In such a situation there is no need for word order information, and bag-of-words itself would accurately classify these two classes. And only the situation of many words occurs in both classes, we need give more attention to the word order. Therefore the parameter is determined by how many words occurs in both classes defined as follows

$$\lambda = n/N \tag{6}$$

where n is the number of words occurs in both classes for two classes classification or all classes for multi-classification. N_i is the sum of words in class i, and N is the average number of words of all classes defined by $N = (\sum_{i=1}^{c} N_i)/c$, where c is the number of classes. This new approach contain both word information and word sequence information, and does not require switching between bag-of-word kernel and word-sequence kernel. Compute the interest rate example by using the new kernel, we have the result as shown in Table 3.

There may be other techniques to determine the value of the parameter. The parameter could also be influenced by the domain knowledge by human beings. In this paper, the proposed algorithm for λ has been demonstrated successful in the example and the experiment presented in the next section.

5 Experiment

Experimental studies have been carried out to compare the performance of bag-of-words kernel, word-sequence kernel and proposed weighted kernel approach. The Reuters News Data Sets, which are frequently used as benchmarks for classification algorithms, was used in this paper for the experiments. The Reuters 21578 collection is a set of 21,578 short (average 200 words in length) news items, largely financially related, that have been pre-classified manually into 118 categories.

The experiments were conducted using 100 documents from three news group: C15 (performance group), C22 (new products/services group) and C21 (products/services group). The first set of experiments used C15 and C22 data, while the second set of experiments used C21 and C22. The second set of data is more difficult to classify than the first set since data sets C21 and C22 are closely related. This is confirmed by the experimental results by using the bag-of-word kernel and word-sequence kernel separately. However, as shown in Tables 4 and 5, the combined and weighted kernel achieves similar results.

50 and 100 frequent keywords were chosen for the bag-of-word kernel. For the word-sequence kernel, frequent words sequences were used in instead of a full list of words. The first column shows the number of selected frequent words. The second column shows the bag-of-word classification result. The third column shows the result of simple combination of the bag-of-word kernel and word-sequence kernel. The last column shows the result of the proposed combined and weighted kernel approach presented in section 4.

The above results show that classification based on bag-of-word model is better than word-sequence kernel in C21 and C22 group. This implies that the word-sequence kernel does not include the bag-of-word information. Although there are many identical keywords in C21 and C22, there is little information on the keyword sequence. Because keywords are not always occur in the same order in a sentence. Therefore word-sequence kernel alone does not reveal any feature representing the documents.

The bag-of-word kernel works better for the C15 and C22 data sets. This is because these two groups are not very close and have not many keywords in common. While the simple combination of the above two approaches results in poorer accuracy in all experiments, the proposed new kernel produces far better results.

6 Conclusion

Bag-of-words model and word-sequence kernel are two important techniques applied in the field of text categorization. Combining word frequency and word order is taking the advantage of both bag-of-words kernel and word-sequence kernel. The parameter based on the word frequency information makes the weighted kernel valid and high computational accuracy. Experiments was conducted with Reuter's database and show the new weighted kernel achieves better classification accuracy.

References

[1] Zhang, L., Zhang, D., Simoff, J.S.: Combined kernel approach for text categrization (submitted). In: the 19th ACS Australian Joint Conference on Artificial Intelligence, Sydney, Australia (2006)

[2] Sebastiani, F.: Machine learning in automated text categorisation. ACM Computing Surveys **34** (2002) 1–47

[3] Amasyali, M., Yildirim, T.: Automatic text categorization of news articles. In: Signal Processing and Communications Applications Conference, 2004. Proceedings of the IEEE 12th, Turkish (2004) 224 – 226

[4] Basu, A., Walters, C., Shepherd, M.: Support vector machines for text categorization. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Hawaii (2003)

[5] Shawe-Taylor, J.: Kernel Methods for Pattern Analysis. University of Cambridge, Cambridge (2004)

[6] Lodhi, H.: Text classification using string kernels. Journal of Machine Learning Research **2** (2002) 419–444

[7] Cancedda, N.: Word-sequence kernels. Journal of Machine Learning Research **3** (2003) 1059–1082

[8] Schlkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge, Massachusetts (2002)

[9] Jalam, R., Teytaud, O.: Kernel-based text categorisation. In: International Joint Conference on Neural Networks. Volume 3., Washington, DC (2001) 1891 – 1896

[10] Zhang, D., Simoff, J.S., Debenham, J.: Exchange Rate Modelling for e-Negotiators. Springer, Computational Intelligence (2006)

[11] Sato, K.: Extracting word sequence correspondences with support vector machines. In: the 19th international conference on Computational linguistics. (2002)

[12] Li, Y.: Text document clustering based on frequent word sequences. In: the 14th ACM international conference on Information and knowledge management. (2005)

Visualization of attractive and repulsive zones between variables

Sylvie Guillaume, Leïla Nemmiche Alachaher

Laboratoire LIMOS, UMR 6158 CNRS
Université Blaise Pascal, Complexe scientifique des Cézeaux
63177 AUBIERE Cedex - France

{sylvie.guillaume, nemmiche}@isima.fr

Abstract

This paper presents a preprocessing step in mining association rules which uses tables to summarize synthetically the way variables interact by highlighting any zones which are attractive. Attractive zones are those which guarantee that potentially interesting rules will be extracted, and any irrelevant rules removed. These attractive zones will also make it possible to carry out a contextual discretization. In addition they constitute the starting point for mining association rules thereby decreasing the space where rules have to be searched for. Finally, this tabular representation of the behaviour of associations is particularly interesting in the case of quantitative variables where knowledge is no longer parsed.

Keywords: Data mining, association rules, quantitative variables, discretization, interestingness measures.

1 Introduction

Algorithms for the discovery of association rules [Agrawal, T. Imielinski and A. Swami 1993], [J. Han and Y. Fu 1995], [H. Mannila, H. Toivonen and A.I. Verkamo 1994], [J.S. Park, M.S. Chen and P.S. Yu 1995], [A. Savasere, E. Omiecinski and S. Navathe 1995] generate a prohibitive number of implications because they are unsupervised and the user neither expresses his or her goals nor selects endogenous variables. This profusion of a large number of rules obviously gives rise to a number of problems. First of all, any expert inundated with all this knowledge could have trouble adapting them and any attempt at analysis and synthesis could fail. Moreover, not all the generated rules are of equal interest and most of them are redundant, weakly significant and irrelevant. [L. Dumitriu, C. Tudorie, E. Pecheanu and A. Istrate 2000] [. R. Lehn, F. Guillet and H. Briand 1998] decrease the number of extracted rules by eliminating those that are redundant whereas [R.J. Bayardo and R. Agrawal 1999], [S. Brin, R. Motwani and C. Silverstein 1997], [M. Kamber and R. Shinghal 1995], [M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and A.I. Verkamo 1994], [S. Lallich and O. Teytaud 2004] identify measures allowing the interesting rules to be filtered.

With quantitative variables, the problem is even worse firstly because one discretization as a preliminary step must be carried out [R. Srikant and R. Agrawal 1996], and this discretization is done without taking the context into account i.e. their association with the other variables, and secondly because the generated knowledge for this kind of variable is parsed.
Since the human brain can easily process visual information and quickly extract a large quantity of information and knowledge from it, the proposed approach tries to reproduce the process of the expert: going from the general to the particular by presenting the extracted knowledge in a visual and synthetic form.

The paper presents a preprocessing step in mining association rules which uses tables to summarize synthetically the way variables interact by highlighting any zones which are attractive. Attractive zones are those which guarantee that potentially interesting rules will be extracted, and any irrelevant rules removed. These attractive zones will also make it possible to carry out a contextual discretization. In addition they constitute the starting point for mining association rules thereby decreasing the space where rules have to be searched for. Finally, this tabular representation of the behaviour of associations is particularly interesting in the case of quantitative variables where knowledge is no longer parsed.
The remainder of the paper is organized as follows. In *section 2* we present the attractive zones between variables. In *section 3* we explain the selected technique to obtain synthetic views of the behaviour of associations between quantitative variables and in *section 4* we explain how that point is reached adapting an existing objective measure, *intensity of inclination*. In *section 5* this technique is evaluated using databases and we conclude with a summary in *section 6*.

2 Attractive Zones

In this section, we present attractive zones between variables which will enable us to discard some irrelevant rules. The first part of this section is limited to binary variables in order to better understand the meaning of these zones and in the second part, we extend our study to quantitative variables.

2.1 Binary Variables

Confidence, the measure used to extract association rules starting from frequent itemsets[1], has the advantage of decreasing the space where rules have to be searched for thanks to its property of anti-monotonicity but on the downside it does retain some irrelevant rules. The upper part of *figure 1* shows an example of an irrelevant rule extracted using confidence. The table in *figure 1* gives the number of customers (*or transactions*) that have bought pancakes and cider. If the user-specified threshold of confidence is equal to *0,80* (*minConf = 0,80*), the rule *"pancake → cider"* is relevant since confidence is equal to *0,88* (*22/25*). However the probability of buying cider is equal to *0,90* and consequently when we know that a customer has bought pancakes the chances of him (*actually*) buying cider are lower than when we have no information concerning a customer's purchases!

This phenomenon occurs when the probability *Pr(Y)* of the appearance of frequent itemset *Y* is higher than the confidence of rule $X \rightarrow Y$ i.e. when $\Pr(Y) > \Pr(Y/X) \geq minConf$.

The lower hand side of *figure 1* represents the evolution of the confidence of rule $X \rightarrow Y$ going through three characteristic points : incompatibility ($(X)_{e_i \in \Omega} \cap (Y)_{e_i \in \Omega} = \varnothing$ *thus Pr(Y/X) = 0*), independence (*Pr(Y/X) = Pr(Y) i.e. the previous appearance of X does not change the probability of appearance of Y*) and logical implication ($(X)_{e_i \in \Omega} \subseteq (Y)_{e_i \in \Omega}$ *thus Pr(Y/X) = 1*).

The zone ranging between independence and logical implication is the zone where we are sure we will obtain interesting rules since the appearance of *X* increases the chances of itemset *Y* occurring. This zone is called the attractive zone of *Y* per *X*.

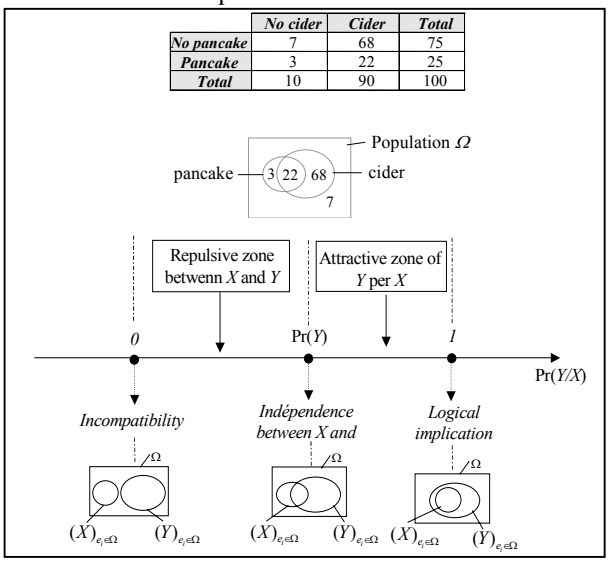


Figure 1: Example of irrelevant rule extracted with confidence (*upper part of figure*) and evolution of confidence of rule $X \rightarrow Y$ (*lower part of figure*).

To discard these irrelevant rules, solutions have been proposed. Lift [IBM 1996] and conviction [S. Brin, R. Motwani and C. Silverstein 1997] are measures that can evaluate the deviation from independence and thus to know if we are in the attractive zone.

In order to guarantee that we will obtain potentially relevant rules, we carry out a preprocessing pruning of irrelevant rules but not a postprocessing pruning with other measures: we highlight these attractive zones between variables and using these extracted zones as our starting point, we begin extracting association rules.

2.2 Quantitative Variables

The framework of our study being quantitative variables, we will now transpose these attractive zones to this kind of variable and then we will explain the technique used to detect them.

Let *X* and *Y* be quantitative variables taking values respectively in the intervals $[x_{min}, x_{max}]$ and $[y_{min}, y_{max}]$ and having respectively *r* and *s* distinct values $x_1 = x_{min}, x_2, \ldots, x_r = x_{max}$ and $y_1 = y_{min}, y_2, \ldots, y_s = y_{max}$. Let *N* be the number of transactions in the database or in the population Ω and let n_{ij} ($i \in \{1,..,r\}$ *and* $j \in \{1,..,s\}$) be the number of transactions which verify simultaneously $X = x_i$ and $Y = y_j$. Let *T* be the contingency table *(represented in table 1)* of the differences between observed n_{ij} and expected $n_{i.}n_{.j}/N$ frequencies under an assumption of independence between variables *X* and *Y*.

	$X = x_1$	…	$X = x_i$	…	$X = x_r$	
$Y = y_1$	n_{11}-$n_{.1}n_{1.}$/N	…	n_{i1}-$n_{.1}n_{i.}$/N	…	n_{1r}-$n_{.1}n_{r.}$/N	*0*
…	…	…	…	…	…	…
$Y = y_j$	n_{1j}-$n_{.j}n_{1.}$/N	…	n_{ij}-$n_{.j}n_{i.}$/N	…	N_{rj}-$n_{.j}n_{r.}$/N	*0*
…	…	…	…	…	…	…
$Y = y_s$	n_{1s}-$n_{.s}n_{1.}$/N	…	n_{is}-$n_{.s}n_{i.}$/N	…	n_{rs}-$n_{.s}n_{r.}$/N	*0*
	0	…	*0*	…	*0*	*0*

Table 1: Contingency table *T* of differences between observed and expected frequencies.

If we consider again the case of binary variables, the attractive zone represented in the lower side of *figure 1* is that whose $\Pr(Y/X) > \Pr(Y)$. Let n_X be the number of transactions verifying *X*, n_Y be the number of transactions verifying *Y* and n_{XY} be the number of transactions simultaneously verifying *X* and *Y*, thus the attractive zone is that where $n_{XY} / n_X > n_Y / N$ or $n_{XY} - n_X n_Y / N > 0$. In the case of quantitative variables, we have to search for cases (i, j) of the contingency table *T* of differences verifying $n_{ij} - n_{i.}n_{.j} / N > 0$ i.e. cases where $\Pr(Y = y_j / X = x_i) > \Pr(Y = y_j)$ or $\Pr(X = x_i / Y = y_j) > \Pr(X = x_i)$.

In order to illustrate these remarks, let us take a real-life example from some banking data where the database consists of *47,112* transactions described by three categories of variables, namely information about customers (*age, number of years with bank,...*), information about various accounts opened with the bank (*bonds, mortgages, savings accounts, ...*) and statistics about various accounts (*rate of indebtedness, total income, …*). Variables *X* and *Y* represent respectively the number of accounts *"stocks"* and the number of *"house purchase saving plans"* opened by a household.

Table 2 represents the contingency table *T* of differences between observed and expected frequencies. Values of variables *X* and *Y* are respectively in intervals [*0..5*] and

[1] or conjunctions of variables which cover a range which is superior to a user-specified threshold.

[*0..2*]. For $X = 1$ and $Y = 0$ there are *694* fewer transactions compared to what could be expected under the assumption that X and Y are independent. On the contrary, for $X = 1$ and $Y = 1$, there are *684* more transactions compared to what could be expected.

	X = 0	***X = 1***	***X = 2***	***X = 3***	***X = 4***	***X = 5***	***Total***
Y = 0	988	- 694	- 270	- 21	- 4	0	*-1*
Y = 1	- 976	684	268	21	4	0	*1*
Y = 2	- 12	10	2	0	0	0	*0*
Total	*0*	*0*	*0*	*0*	*0*	*0*	*0*

Table 2: Contingency table *T* of differences for the banking example.

The attractive zones between X and Y are the following : $(X = 0, Y = 0)$, $(X = [1, 4], Y = 1)$ et $(X = [1, 2], Y = 2)$.

The discovery of association rules will be limited to these extracted zones.

3 Parsed Views of Associations

The transformation stage (*i.e. the discretization stage and the stage of complete disjunctive coding*) of quantitative variables not only results in irrelevant intervals being obtained since the discretization is independent of the association with variables, but also generates specific rules (*i.e. verified by a subset of the population*) which only contain a part of the behaviour of variables. For example, we have the following rules:

$\boldsymbol{R_1} : X = [x_1, x_2] \rightarrow Y = [y_1, y_2]$, $\boldsymbol{R_2} : X = x_1 \rightarrow Y = y_6$,

$\boldsymbol{R_3} : X =]x_3, x_4] \rightarrow Y =]y_2, y_3]$, $\boldsymbol{R_4} : X = x_5 \rightarrow Y =]y_4, y_6]$,

$\boldsymbol{R_5} : X =]x_6, x_7] \rightarrow Y = [y_1, y_3]$, $\boldsymbol{R_6} : X = x_8 \rightarrow Y =]y_2, y_6]$.

It is very difficult to have a global view of the implication between X and Y in this form. The human mind finds it difficult to assimilate: the only form which can be understood easily is the summarized and therefore less accurate form.

In order to have a global view of associations between variables (*bearing in mind that our starting point is frequent associations between variables to find association rules*) and more particularly quantitative variables where knowledge is parsed, we search for attractive zones, in a cursory way, in order to make this summarized form easily accessible to experts. It will then be possible for experts to visualize these zones in a more accurate way.

The four selected attractive zones are represented in *figure 2*.

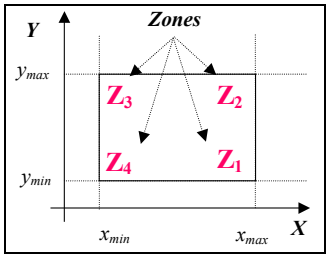


Figure 2: The four selected attractive zones.

Zones where transactions verify simultaneously:

zone 1 : a high value for X and a low value for Y.

zone 2 : a high value for X and Y.

zone 3 : a low value for X and a high value for Y.

zone 4 : a low value for X and Y.

For each zone, we would like to know if we are in an attractive zone ($n_{ij} - \frac{n_{i.}n_{.j}}{N} > 0$), a repulsive zone ($n_{ij} - \frac{n_{i.}n_{.j}}{N} < 0$) or an independence zone ($n_{ij} = \frac{n_{i.}n_{.j}}{N}$).

We symbolize the attractive zone by the character "+", the repulsive zone by "-" and the independence zone by any character.

In order to obtain these maps of associations between variables, we adapt an existing measure, intensity of inclination, which verifies that in the zone Z_1, we have fewer transactions compared to what could be expected.

4 Adaptation of Intensity of Inclination

In this section we first remind the reader of the definition of intensity of inclination, a measure allowing implications between conjunctions of quantitative variables to be mined [S. Guillaume 2002] and then we show how this measure has been adapted to reveal the behaviour of variables in the four zones mentioned in *figure 2*.

4.1 Intensity of Inclination

Now we will give a more general definition to variables X and Y : they are conjunctions of quantitative variables. Let X and Y be respectively two conjunctions of p and q quantitative variables. We suppose that $X = X_1, .., X_p$ and $Y = Y_1, .., Y_q$, where $X_1, .., X_p, Y_1, .., Y_q$ are quantitative variables taking values $x_{1_i}, .., x_{p_i}, y_{1_i}, .., y_{q_i}$ ($i \in \{1..N\}$) respectively in intervals $[x_{1_{\min}} .. x_{1_{\max}}], .., [x_{p_{\min}} .. x_{p_{\max}}], [y_{1_{\min}} .. y_{1_{\max}}], .., [y_{q_{\min}} .. y_{q_{\max}}]$.

Intensity of inclination evaluates whether the number of transactions not strongly verifying the rule $X \rightarrow Y$ (*i.e. the number of transactions verifying simultaneously a high value for each attribute $X_1, .., X_p$ and a low value for each attribute $Y_1, .., Y_q$*) is significantly small compared to the expected number of transactions under the assumption that X and Y are independent. These transactions that do not strongly verify the rule are called negative transactions.

Let x_i and y_i be respectively values taken by variables X and Y in the database Ω for transaction e_i ($e_i \in \Omega$) and let x_{min} and y_{max} be respectively the minimum and maximum values taken by variables X and Y.

The number t_0 of negative transactions, or raw measure of non-inclination, is defined by:

$$t_o = \sum_{i=1}^{N} (x_i - x_{\min})(y_{\max} - y_i) \text{ with}$$

$$x_i = \sum_{j=1}^{p} x'_{j_i}, \quad x_{\min} = \sum_{j=1}^{p} x'_{j_{\min}}, \quad y_i = \sum_{k=1}^{q} y'_{k_i}, \quad y_{\max} = \sum_{k=1}^{q} y'_{k_{\max}},$$

$$x'_{j_i} = \frac{x_{j_i} - \mu_{X_j}}{\sigma_{X_j}} \ (j \in \{1..p\}), \quad y'_{k_i} = \frac{y_{k_i} - \mu_{Y_k}}{\sigma_{Y_k}} \ (k \in \{1..q\})$$

Let μ_{X_j} and μ_{Y_k} be respectively the means of variables X_j ($j \in \{1, .., ,p\}$) and Y_k ($k \in \{1, .., ,q\}$) and let σ_{X_j} and σ_{Y_k} be respectively standard deviations of X_j and Y_k.

The random variable T, whose t_0 is an observed value, can be approximated asymptotically by a normal distribution $\mathcal{N}(\mu, \sigma)$ with $\mu = N\,(\mu_X - x_{min})(y_{max} - \mu_Y)$ and $\sigma^2 = N\,[v_X v_Y + v_Y\,(\mu_X - x_{min})^2 + v_X (y_{max} - \mu_Y)^2]$.

The means and variances of attributes X and Y are given by the following expressions:

$$\mu_X = \sum_{j=1}^{p} \mu_{X_j}, \quad \mu_Y = \sum_{k=1}^{q} \mu_{Y_k}, \quad v_X = \sum_{j=1}^{p} v_{X_j} + 2\sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} \mathrm{cov}(X_j, X_{j'})$$

and $$v_Y = \sum_{k=1}^{q} v_{Y_k} + 2\sum_{k=1}^{q-1} \sum_{k'=k+1}^{q} \mathrm{cov}(Y_k, Y_{k'}) \quad \text{with}$$

$\mathrm{cov}(X_i, X_{i'}) = \mu_{X_i X_{i'}} - \mu_{X_i} \mu_{X_{i'}}$.

If the probability $Pr(T \leq t_o)$ of having a number inferior or equal to t_o is high, we can say that t_o is not significantly small because this occurrence can happen fairly frequently and then this implication $X \rightarrow Y$ is not relevant.

To evaluate this implication in increasing order, the measure $\varphi(X \rightarrow Y) = 1 - F(t_o) = Pr(T > t_o)$ has been retained where F is the cumulative distribution of T. Then, the implication $X \rightarrow Y$ can be admitted with a level of confidence ($1-\alpha$) if and only if $Pr(T \leq t_o) \leq \alpha$ or $Pr(T > t_o) \geq 1-\alpha$.

The intensity of inclination is given by:

$$\varphi(X \rightarrow Y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{t_0}^{+\infty} e^{\frac{(t-\mu)^2}{2\sigma^2}}\, dt$$

Thus the intensity of inclination evaluates the *"smallness"* of the number of negative transactions i.e. if we obtain a small number of transactions in the zone Z_1 as compared with independence. For that, the measure calculates the sum of weights for all transactions in the database, weight being given by the raw measure of non-inclination. This weight is maximum for transactions verifying $X = x_{max}$ and $Y = y_{min}$ (*corresponding perfectly to the concept of negative transactions*) and decreases until it is equal to zero for transactions verifying $X = x_{min}$ ou $Y = y_{max}$.

4.2 Adaptation

Since the intensity of inclination evaluates if there are fewer transactions in the zone Z_1 compared to what could be expected, we start by adapting this measure for the zones Z_2, Z_3 et Z_4. After this, we then use the previous results to know if there are more transactions compared to what could be expected for the four zones.

Discovery of repulsive zones

First, we adapt the raw measure of non-inclination t_0 for the three zones Z_2, Z_3 and Z_4. For the zone Z_1, we know that this measure calculates the number of transactions verifying simultaneously a high value for X and a low value for Y, then the raw measure is equal to $t_o = \sum_{i=1}^{N} (x_i - x_{\min})(y_{\max} - y_i)$. Consequently, for the other three zones, we have the following raw measures :

zone *2* : $t_o = \sum_{i=1}^{N} (x_i - x_{\min})(y_i - y_{\min})$, number of transactions verifying simultaneously a high value for X and Y.

zone *3* : $t_o = \sum_{i=1}^{N} (x_{\max} - x_i)(y_i - y_{\min})$, number of transactions verifying simultaneously a low value for X and a high value for Y.

zone *4* : $t_o = \sum_{i=1}^{N} (x_{\max} - x_i)(y_{\max} - y_i)$, number of transactions verifying simultaneously a low value for X and Y.

Consequently, the following expressions of the intensity of inclination allow us to detect the non-attraction between variables X and Y in the various zones:

zone *2* : $\varphi(X \rightarrow (y_{max} + y_{min} - Y))$

zone *3* : $\varphi(Y \rightarrow X)$

zone 4 : $\varphi((y_{\max} + y_{\min} - Y) \rightarrow X)$

Discovery of attractive zones

In this section, we have adapted the intensity of inclination in order to know if there are more transactions compared to what could be expected for the four zones.

The implication $X \rightarrow Y$ can be admitted with a level of confidence ($1-\alpha$) if and only if $Pr(T \leq t_o) \leq \alpha$. On the contrary, there will be many negative transactions if $Pr(T \leq t_o) \geq 1-\alpha$ and we can deduce from this that we are in an attractive zone.

Thus, we can detect if the four zones are attractive thanks to the following expressions:

zone *1* : $1 - \varphi(X \rightarrow Y)$

zone *2* : $1 - \varphi(X \rightarrow (y_{max} + y_{min} - Y))$

zone *3* : $1 - \varphi(Y \rightarrow X)$

zone *4* : $1 - \varphi((y_{max} + y_{min} - Y) \rightarrow X)$

5 Experimental Results

This detection and visualization step of the various zones is a preprocessing step for the discovery of association rules (*see figure 3*).

The process of mining association rules has been presented and developed in [S. Guillaume 2003] with the following difference: we started from ordinal association rules i.e. associations XY where the zone Z_1 is a repulsive area.

During these experiments, we only present the first step in this process and show results obtained using three UCI databases [P.M. Murphy and D.W. Aha 1995]: Wages, Abalone and Fisher's Irises [R. Fisher 1936].

The Wages database consists of *534* transactions described by *11* variables including *4* quantitative variables: Education (*number of years of education*), Experience (*number of years of work experience*), Wage (*dollars per hour*) and Age (*years*). Categorical variables are: Region (*person who lives in the South or elsewhere*), Sex, Union membership, Race (*Hispanic, White and Other*), Occupation (*Management, Sales, Clerical, Service, Professional and Other*), Sector (*Manufacturing, Construction and Other*) and Married (*marital status*).

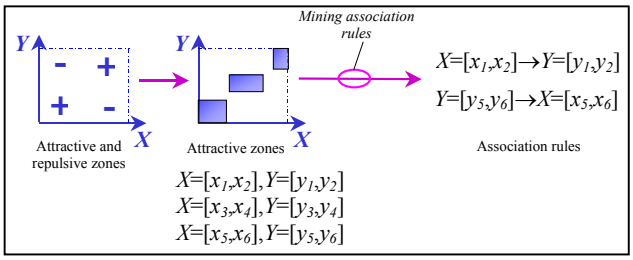


Figure 3: Process for discovering association rules.

Figure 4 presents some attractive and repulsive zones for the Wages database.

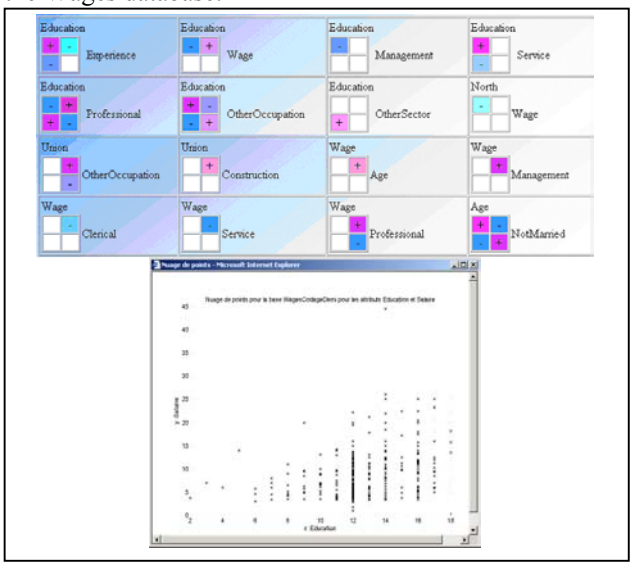


Figure 4: Some attractive and repulsive zones for the *Wages* database (*upper part of figure*) and a two-dimensional scatter diagram for variables "*Education - Wage*" (*lower part of figure*).

If we study the association between the variables "*Education*" and " *Wage*", the upper hand side of *figure 4* shows us that Z_2 is an attractive zone and Z_3 is a repulsive zone. The lower hand side of *figure 4* shows a two-dimensional scatter diagram for these two variables. We check that transactions are uniformly distributed in the zones Z_1 and Z_4 and that transactions verifying a high value for Wage (Y) are concentrated more in the zone Z_2. We notice an exceptional person with *14* years of work experience and earning a very high wage (*approximately 45 dollars per hour*).
This result reinforces our belief that the greater the number of years of work experience, the higher the wage.

The upper hand side of figure 4 also reveals to us :
- In the management and professional sectors, the number of persons with not many years of undergraduate studies is low compared to the expected number under the assumption of independence (*see the associations " Education-Management" and "Education-Professional"*), whereas for services and other occupations (*"Education-Service" and " Education-Other occupation"*) the opposite phenomenon exists, with more transactions compared to the expected number.
- Wages are higher for the oldest people (*" Wage-Age"*), in the management and professional categories (*" Wage-Management" and "Wage-Professional"*). On the contrary, they are lower in the clerical and service sectors (*" Wage-Clerical" and " Wage-Service"*).

The Abalone dataset consists of *4,177* abalones described by *9* variables including *8* quantitative variables (*Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight and the number of Rings*) and one categorical variable (*Sex taking the values "Male", "Female" and "Infant"*).
Figure 5 gives some associations for the Abalone dataset.

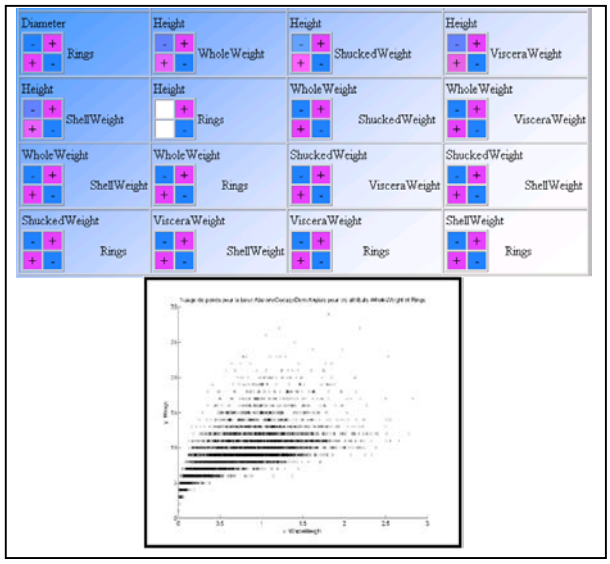


Figure 5: Some associations for the *Abalone* dataset (*upper part of figure*) and a two-dimensional scatter diagram for variables "*Whole weight - Rings*" (*lower part of figure*).

We note that these quantitative variables are strongly positively correlated among themselves. Thus, for example, the higher the number of rings, the higher the diameter, the height, the whole weight, the shucked weight, the viscera weight and the shell weight. The lower hand side of *figure 5* shows a two-dimensional scatter diagram for variables "*Whole weight*" and "*Rings*" in order to confirm the positive correlation detected between these variables.

The last dataset consists of *150* irises described by *5* variables including *4* quantitative variables (*sepal length, sepal width, petal length and petal width*) and one categorical variable (*class of iris plant : Iris Setosa, Iris Versicolour and Iris Virginica*).
Figure 6 gives associations for this database.

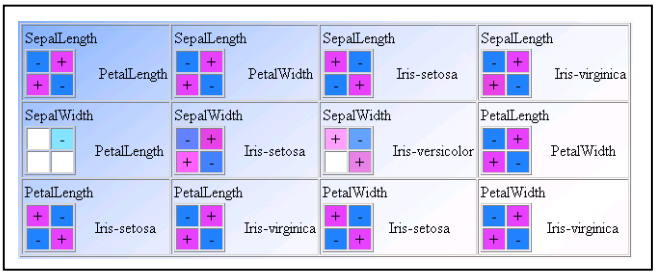


Figure 6: Associations for the *Iris* dataset.

We learn from this extraction that:

- the sepal and petal length and the petal width for the Setosa class generally have low values contrary to the Virginica family which has high values. *Figure 7* shows the contrary distribution of the variable ″sepal length″ for these two classes.
- the sepal width for the Setosa class generally has high values contrary to the other quantitative variables.

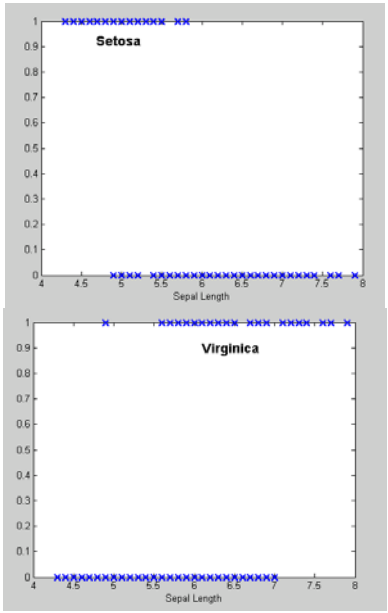


Figure 7: Scatter diagram for variables ″Sepal Length″ - ″Setosa″ (*upper part of figure*) and variables ″Sepal Length″ - ″Virginica″ (*lower part of figure*).

6 Conclusion and Further Work

This visualization step of attractive and repulsive zones allows us to obtain a synthetic view of associations between variables. It is a preliminary step for the process of mining association rules and is used as a basis for discretization and decreases the space where rules have to be searched for. It can also be used in a process of selection of relevant variables. This method is particularly suitable for quantitative variables which take a moderate number of values. When this number is too high, we no longer obtain a global measure but a local measure. A solution would be to increase the number of zones in order to better appreciate the behaviour of these variables.

Acknowledgments. We thank STÉPHANE COUTURAUD for making the web site and CAROL FYNN for her invaluable assistance with the writing of this paper in English.

7 References

R. Agrawal, T. Imielinski and A. Swami,″Mining association rules between sets of items in large databases″, In *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data (SIGMOD'93)*, Washington, D.C., ACM Press, pp. 207-216, May 1993.

R.J. Bayardo and R. Agrawal, ″Mining the most interesting rules″, *In Proceedings KDD 99* pp. 145-154, 1999.

S. Brin, R. Motwani and C. Silverstein, ″Beyond market baskets : generalizing associations rules to correlations″, In *Proceedings of ACM SIGMOD'97*, pp. 265-276, 1997.

L. Dumitriu, C. Tudorie, E. Pecheanu and A. Istrate, ″A new algorithm for finding association rules″, In *Proceedings of Data Mining 2000,* WIT Press, vol. 2, pp. 195-202, 2000.

R. Fisher, The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*, 1936.

S. Guillaume, ″Discovery of ordinal association Rules″, Dans *Proceedings of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD02), pp.322-327, Taipei, Taiwan, 6-8 May 2002, ISBN 3-540-43704-5..

S. Guillaume, ″Ordinal association rules towards association rules″, Dans *Proceedings of the 5th International Conference on Data Warehousing and Knowledge Discovery* (DaWaK 2003), 3-5 September 2003, pp. 161-171, Prague, Czech Republic, ISBN 3-540-40807-X.

J. Han and Y. Fu, ″Discovery of multiple-level association rules from large databases″, *In Proceedings of the VLDB Conference,* Zurich, Switzerland, September 1995.

IBM 96 ″*IBM Intelligent Miner User's Guide*″, Version 1 Release 1, SH12-6213-00 edition.

M. Kamber and R. Shinghal, ″Evaluating the interestingness of characteristic rules″, *In Proceedings KDD 95* pp. 263-266, 1995.

M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and A.I. Verkamo, "Finding interesting rules from large sets of discovered association rules" – In *Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM'94)*, ACM Press, pp. 401-407, November 1994.

S. Lallich et O. Teytaud, 2004. Evaluation et validation de l'intérêt des règles d'association, 2004.

R. Lehn, F. Guillet and H. Briand, ″Eliminating redundancy in a rule system″, In *Proceedings of the 4th European Meeting on Cybernetics and System Research*, Austrian Soc. Of Cybernetic Studies, vol. 2, pp. 793-798, 1998.

H. Mannila, H. Toivonen and A.I. Verkamo, ″Efficient algorithms for discovering association rules″. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in Databases*, pp. 181-192, Seattle, Washington, 1994.

P.M. Murphy and D.W. Aha, *UCI Repository of Machine Learning Databases*. Machine-readable collection, Dept of Information and Computer Science, University of California, Irvine, 1995. [Available by anonymous ftp from ics.uci.edu in directory pub/machine-learning-databases]

J.S. Park, M.S. Chen and P.S. Yu, "An effective hash-based algorithm for mining association rules" - In Proc. 1995 *ACM-SIGMOD Int. Conf. Management of Data*, San Jose, CA, May 1995.

A. Savasere, E. Omiecinski and S. Navathe, ″An efficient algorithm for mining association rules in large databases″, In *Proceedings of the VLDB Conference*, Zurich, Switzerland, September 1995.

R. Srikant, R. Agrawal, ″Mining quantitative association rules in large relational tables″, *Proceedings 1996 ACM-SIGMOD International Conference Management of Data*, Montréal, Canada, Juin 1996.

On the Optimal Working Set Size in Serial and Parallel Support Vector Machine Learning with the Decomposition Algorithm

Tatjana Eitrich⋆ **Bruno Lang**‡

⋆ Central Institute for Applied Mathematics, Research Centre Juelich, Germany
Email: t.eitrich@fz-juelich.de
‡ Applied Computer Science and Scientific Computing Group, Department of Mathematics, University of Wuppertal, Germany

Abstract

The support vector machine (SVM) is a well-established and accurate supervised learning method for the classification of data in various application fields. The statistical learning task – the so-called training – can be formulated as a quadratic optimization problem. During the last years the decomposition algorithm for solving this optimization problem became the most frequently used method for support vector machine learning and is the basis of many SVM implementations today. It is characterized by an internal parameter called working set size. Usually small working sets have been assigned. The increasing amount of data used for classification led to new parallel implementations of the decomposition method with efficient inner solvers. With these solvers larger working sets can be assigned. It was shown, that for parallel training with the decomposition algorithm large working sets achieve good speedup values. However, the choice of the optimal working set size for parallel training is not clear. In this paper, we show how the working set size influences the number of decomposition steps, the number of kernel function evaluations and the overall training time in serial and parallel computation.

1 Introduction

The support vector machine for classification and regression is a powerful machine learning method. Its popularity is mainly due to the applicability in various fields of data mining, such as text mining (Joachims 1998), biomedical research (Yu, Yang, Wang & Han 2003), and many more. SVM test accuracy is excellent and in many cases it outperforms other machine learning methods such as neural networks. SVM has its roots in the field of statistical learning which provides the reliable generalization theory (Vapnik 1998). Several properties that make this learning method successful are well-known, e.g. the kernel trick (Schölkopf 2001) for nonlinear classification and the sparse structure of the final classification function. In addition, SVM has an intuitive geometrical interpretation, and a global minimum can be located during the training phase.

Most current SVM implementations are based on the well known decomposition algorithm for solving the optimization problem of SVM training (Hsu & Lin 2002*b*). It repeatedly selects a subset of the free variables and optimizes over these variables. Thus, decomposition provides a framework for handling large SVM training tasks, where the kernel matrix does not fit into the available memory. Its main advantage is the flexibility concerning the size of the subproblems – the working set size. All values larger than one and smaller or equal to the training set size are possible. The limitation for large working sets is due to memory requirements of the machine and the characteristics of the inner solver. A widely used decomposition method called SMO (Platt 1999) uses the extreme case of only two free variables in each iteration. Other approaches use larger working sets. However, the optimal choice of the working set size is not clear, especially for large data sets. In this paper, we will show, how the training time is influenced by the size of the subproblems for the inner solver.

Real world data sets are becoming increasingly large. The main drawback of current SVM models is their high computational complexity for large data sets (Chen, Lu, Yang & Li 2004). Parallel processing is essential to provide the performance required by large-scale data mining tasks. Therefore the development of highly scalable parallel SVM algorithms is a new important topic of current SVM research. Recently, new parallel decomposition algorithms have been proposed. For parallel computation large working set sizes are possible and lead to good speedup values. However, it is not clear, which influence the working set size has onto the overall training time for large data sets in serial and parallel mode. One goal of this paper is to show how the working set size controls the performance of SVM training in general. In addition, the results of serial computation are broaden to the parallel mode.

The paper is organized as follows. In Sect. 2 we review basics of binary SVM classification. In Sect. 3 we describe the scheme of the serial decomposition algorithm and explain the importance of the working set size. We present a survey of parallel data mining methods in Sect. 4. Our parallelization of the support vector machine learning method is explained in Sect. 5. In Sect. 6 we show results of various tests using small and large data sets in serial and parallel mode. In view of overall learning time we discuss the issue of the optimal working set size for parallel SVM training.

2 Basic Concepts of the Support Vector Machine

Support vector machine learning means to determine functions that can be used to classify data points. Here, we discuss binary classification, but the SVM learning framework also works for multi-class and regression problems (Hsu & Lin 2002*a*). The supervised SVM learning method is based on so-called reference

data of given input–output pairs (training data)

$$(\boldsymbol{x}^i, y_i) \in \mathbb{R}^n \times \{-1, 1\}, \quad i = 1, \ldots, l,$$

that are taken to find an optimal separating hyperplane (Cristianini & Shawe-Taylor 2000)

$$f_{\mathrm{l}}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b = 0.$$

Using assumptions of statistical learning theory (Vapnik 1998), the desired classifier is then defined as

$$h(\boldsymbol{x}) = \begin{cases} +1, & \text{if } f_{\mathrm{l}}(\boldsymbol{x}) \geq 0, \\ -1, & \text{if } f_{\mathrm{l}}(\boldsymbol{x}) < 0, \end{cases}$$

with the linear decision function f_{l}, see Fig. 1.

If the two classes are not linearly separable, then f_{l} is replaced with a nonlinear decision function (Schölkopf & Smola 2002)

$$f_{\mathrm{nl}}(\boldsymbol{x}) = \sum_{i=1}^{l} y_i \alpha_i K(\boldsymbol{x}^i, \boldsymbol{x}) + b,$$

where $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a (nonlinear) kernel function (Schölkopf 2001). The classification parameters α_i $(i = 1, ldots, l)$ can be obtained as the unique global solution of a suitable (dual) quadratic optimization problem (Schölkopf & Smola 2002)

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^l} \quad g(\boldsymbol{\alpha}) := \frac{1}{2} \boldsymbol{\alpha}^T H \boldsymbol{\alpha} - \sum_{i=1}^{l} \alpha_i \qquad (1)$$

with $H \in \mathbb{R}^{l \times l}$, $H_{ij} = y_i K(\boldsymbol{x}^i, \boldsymbol{x}^j) y_j$ $(1 \leq i, j \leq l)$, constrained to

$$\boldsymbol{\alpha}^T \boldsymbol{y} = 0, \quad 0 \leq \alpha_i \leq C.$$

In the final solution, only a part of the entries in $\boldsymbol{\alpha}$ are positive, whereas all others are zero. This is due to the Karush-Kuhn-Tucker conditions for convex optimization problems (Fletcher 1981). In SVM theory, the corresponding training points are called support vectors, see Fig. 1. The parameter C controls the trade-off between the width of the classifier's margin and the number of weak and wrong classifications in the training set. This parameter has to be chosen by the user. Due to space limitations, we omit a detailed introduction to the SVM theory. For readers who are not familiar with this topic we refer to Burges (1998).

Usually, the Hessian H is dense, and therefore the complexity of evaluating the objective function g in (1) scales quadratically with the number l of training pairs, leading to very time-consuming computations.

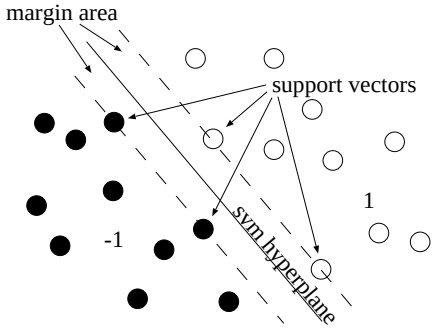


Figure 1: Support vector machine classification function based on the support vectors in the training data.

3 Decomposition and the Working Set Size

The high cost for solving (1) is due to the size and the density of the quadratic matrix H. Classical solvers for QP problems with simple constraints are the so called active set methods (Fletcher 1981, Leyffer 2005), which in each iteration minimize the objective function over the active set (a subset of the constraints that are locally active), until a solution is reached. A variation of the active set approach for support vector machine training is the well known decomposition method described in Osuna *et al.* (1997). It repeatedly splits the original optimization problem (1) into active and inactive parts. In each step, the active data points (or working set points) are used to define a new QP subproblem, which is then handled by an inner solver. This method is very flexible, since the user can choose the desired number of active points $\tilde{l} \leq l$ to control the size of the QP subproblems. Due to space limitations, the decomposition algorithm cannot be discussed here in detail, we refer to Hsu & Lin (2002*b*), Chang *et al.* (2000) and Laskov (2002) for a detailed description. To summarize, in each iteration of the decomposition algorithm the following five steps need to be processed:

1. Select a working set of $\tilde{l}$ "active" variables from the l free variables α_i.
2. Solve the quadratic subproblem of size $\tilde{l}$ that results from restricting the optimization in (1) to the active variables and fixing the remaining variables.
3. Update the global solution vector $\boldsymbol{\alpha}$.
4. Update the gradient of the overall problem.
5. Check a stopping criterion.

Implementation of a sophisticated working set selection scheme, an efficient subproblem solver and a fast but accurate gradient update are crucial for good overall performance of decomposition methods. Several approaches have been proposed and improved during the last decade. Promising suggestions are given in

- Serafini & Zanni (2005) – for the working set selection,
- Ruggiero & Zanni (2001, 1999) – for fast inner solvers,
- Zanghirati & Zanni (2003*a*) and Serafini *et al.* (2004) – for sparse gradient updates.

As already mentioned, one important model parameter of the decomposition method is the working set size $\tilde{l}$. In is known, that for larger working sets the number of decomposition steps will decrease, but a single step may be more expensive. In contrast, for small working sets the quadratic subproblem may be solved very fast, but the number of steps will increase heavily. Choosing $\tilde{l} = 2$ results in the well-known Sequential Minimal Optimization (SMO) scheme (Platt 1999). SMO decomposition steps are fast, but this method suffers from a very large number of optimization steps even when using moderate problem sizes. Today, it is not clear which working set sizes are preferable. In this work, we will analyze the overall behavior of this two competing effects. The training time not only depends on the working set selection scheme and the inner solver, but also on the data set and the characteristics of the machine used

for running the software. However, as we will show in Sect. 6, the working set size has enormous influence onto the training time of SVM training and needs to be optimized primarily. We will show, that indeed a global minimum of the training time does exist and needs to be located, especially when using expensive parallel computing resources.

4 A Survey of Parallel Data Mining and SVM Methods

Most sequential data mining algorithms have large runtimes, but the volume of data available for analysis is growing rapidly, i.e. the number of attributes as well as the number of instances both increase. In addition to improvements of the serial algorithms, the development of parallel techniques may help to avoid computational bottlenecks. This section gives an overview of activities concerning large scale data mining, particularly the problem of classification using machine learning techniques like SVMs.

4.1 Parallel Data Mining

The first parallel data mining algorithms have emerged a decade ago. In Skillicorn (1998) the general differences between parallel data mining and other numerical parallel algorithms are explained. The design of scalable data mining algorithms requires meeting several challenges, e.g., the enormous memory requirements have to be supported by the computing system.

Various algorithms, especially for supervised learning methods, have been parallelized.

- A parallel algorithm for data mining of association rules was presented in Parthasarathy (2001). It has been designed for shared memory multiprocessors.
- The *ScalParC* software (Joshi, Karypis & Kumar 1998), designed in 1998, was one of the first methods for parallel decision tree classification. Parallel decision tree applications are still of interest, mainly in the important field of Grid computing (Hofer 2004).
- Clustering is useful in various fields, i.e., pattern recognition and learning theory. The runtime complexity of a serial k-means clustering algorithm is high for problems of large size. Therefore parallel clustering methods have been developed. We refer to Kantabutra (2000) for a master-slave approach.
- K-nearest neighbor methods have received a great deal of attention since they are applied frequently in bioinformatics, but performance is a serious problem for many implementations. In Callahan (1993) a parallel algorithm was introduced to overcome the problem of runtime.
- Artificial neural networks (ANNs) are well-known data mining methods with high learning cost when the models are large. An approach for speeding up their implementation by using parallel environments is given in Misra (1997).
- Bayesian networks for unsupervised classification tasks include time consuming steps which can be parallelized. A description is given in Jin *et al.* (2005).
- Boosting is a method for improving the accuracy of any given learning algorithm (Schapire 1999) and is often used within the context of supervised learning. A framework for distributed boosting is presented in Lazarevic & Obradovic (2001). The method requires less memory and computational time than serial boosting packages.

4.2 Parallel Support Vector Machine Approaches

Efficient and parallel support vector machine learning is a young and emerging field of research, but the number of truly parallel implementations is small. Most approaches just try to increase the efficiency of the serial algorithms and to overcome the problem of large scale applications by dividing the data into subsets. Different approaches for splitting a large data set into small subsets have been implemented (Graf, Cosatto, Bottou, Dourdanovic & Vapnik 2005). Usually, results of the individual training stages are merged to finally obtain a single SVM model. The individual optimization steps can be run in parallel. A fast SVM algorithm, which uses caching, digest and shrinking policies is given in Dong & Suen (2003). The clustering-based SVM (Yu, Yang & Han 2003) is a learning method that scans the data set before training the SVM. It selects the data which are supposed to maximize the benefit of learning and is useful for very large problems when a limited amount of computing resources is available. So far it is only applicable for linear problems. In addition, various projects exist where a simple parallelization scheme is used to speed up the learning process. In Dong *et al.* (2003) a parallel optimization step is proposed. It approximates the kernel matrix by block diagonal matrices and splits the original problem into sub-problems which can be solved independently from each other with standard algorithms. This step is used to remove non-support vectors before SVM training. Parallel training of several binary SVMs for solving multiclass problems is described in Poulet (2003). Parallel cross validation methods do exist for the *WEKA* machine learning package (Celis & Musicant 2002). Parallel parameter optimization techniques such as grid search or pattern search have been studied for SVM parameter fitting (Eitrich & Lang 2005). These approaches can be interpreted as coarse grained parallelization techniques for SVM methods at a high level which is independent from the inner solver. However, the computational bottleneck of a single SVM training on a large data set can be avoided only by implementing a fine grained parallel support vector machine training. The following methods have been proposed. Parallel computation of the kernel matrix for high dimensional data spaces is implemented in Qiu & Lane (2005). The speedup is limited because of high communication costs. Therefore an approximation method, that reduces the kernel matrix, was implemented. The method is applicable only for commonly used kernels which are inner product-based and requires changes in the algorithm for each kernel. A distributed SVM algorithm for row-wise and column-wise data distribution is described in Poulet (2003), which so far can be used for linear SVMs only. A promising parallel MPI-based decomposition solver for training support vector machines has been implemented recently (Serafini, Zanghirati & Zanni 2004). A parallel support vector machine for multi-processor shared memory (SMP) clusters has been introduced in Eitrich & Lang (2006*a*).

5 Parallel Support Vector Machine Decomposition

The SVM training, i.e. the solution of the quadratic program (1), suffers from large data sets (Graf et al.

	australian	*fourclass*	*adultpart*	*adult*
# features	14	2	123	123
# training points	690	862	15000	32561
# positive points	307	307	3562	7841
# negative points	383	555	11438	24720

Table 1: Characteristics of the data sets.

$\hat{l}$	4	6	10	50	80	100	120	150	200	350	500	600	650	690
$\#D$	7455	6193	4349	543	90	31	14	9	7	5	3	4	3	1
$\#E$	19.9	24.2	28.8	18.9	5.1	2.2	1.2	0.9	0.9	1.1	1.4	1.4	1.6	0.7
$\#sv$	246	247	246	247	247	247	246	247	247	247	247	246	247	247
t	5.60	6.27	7.58	7.08	2.78	1.24	**0.94**	1.01	2.89	5.88	12.48	18.17	23.54	7.55

Table 2: Results for the *australian* data set.

2005). In this work, we target an already fruitful approach for serial and parallel SVM training with the decomposition method. In Zanghirati & Zanni (2003*a*) and Serafini *et al.* (2004) a parallelized MPI-based decomposition algorithm has been proposed. It is based on a variable projection method as inner solver (Zanghirati & Zanni 2003*b*, Ruggiero & Zanni 2000). In Eitrich & Lang (2006*a*) we have presented an implementation of parallel support vector machine decomposition training for shared memory systems based on this solver. The parallel SVM algorithm uses library and loop level parallelism. Calls to the ESSLSMP library (Engineering Scientific Subroutine Library for Shared Memory Parallel Machines) (IBM n.d.) as well as OpenMP loop level parallelism lead to a scalable training method. In both approaches the main parallel parts of the decomposition method belong to the time consuming computations in each step (Eitrich & Lang 2006*a*), i.e.

1. the computation of the kernel matrix for the new subproblem,
2. expensive matrix-vector multiplications in the decomposition routine and the inner solver,
3. the gradient update for the overall optimization problem.

For details to the parallelization schemes we refer to Zanghirati & Zanni (2003*a*) and Eitrich & Lang (2006*a*). Tests for both packages were run using large working sets, which led to promising speedup values, e.g. for 3600 in Zanghirati & Zanni (2003*a*) and $\hat{l} = 5000$ in Eitrich & Lang (2006*a*). However, the improved serial algorithms have not been tested for smaller data sets or large data sets with small working set sizes. This aspect needs to be analyzed and will bring improvements for serial as well as parallel support vector machine training.

6 Experimental Results

We performed our tests on the Juelich Multi Processor (JUMP) (Detert 2004) using our parallel SVM software which also provides a parallel validation loop (Eitrich, Frings & Lang 2006). JUMP is a distributed shared memory parallel computer consisting of 41 frames (nodes). Each node contains 32 IBM Power4+ processors running at 1.7 GHz, and 128 GB shared main memory. The 1312 processors have an aggregate peak performance of 8.9 TFlop/s. Our software is written using Fortran90 and the ESSLSMP library. For each thread we chose the following characteristics:

- 3.5 GB consumable memory,
- 3.0 GB data limit,
- 0.5 GB stack limit,
- 1h wall clock limit.

6.1 Description of the Data Sets

The so-called *australian* data set is available from Hettich *et al.* (1998). It contains credit card applications with 14 attributes – 6 numerical and 8 categorical. The number of instances is 690. The class distribution is 44.5% vs. 55.5%.

The *fourclass* data set was introduced in Ho & Kleinberg (1996). It is artificial and is aimed at testing the performance of classifiers. 862 data points in a two-dimensional space have been assigned to four classes in the original data set. The classes are distributed irregularly and show isolated regions to complicate classification. The data set was transformed into a binary classification problem and is available from Chang & Lin (2001).

The *adult* data set is available under Hettich *et al.* (1998). The task is to predict whether a household has an income greater than $50000 (Platt 1999). Originally, 14 attributes of a census form of a household were given. We use the data set in the discretized form with 123 binary features, which is available from Chang & Lin (2001) under the name *a9a*. 32561 training points are available. In this paper, we use two versions of the data – the whole data set as well as a subset of 15000 points which we call *adultpart*.

A summary of the data sets characteristics is shown in Table 1.

6.2 Influence of the Working Set Size for Serial Computation

In Tables 2 and 3 we show the number of decomposition steps D, the number of kernel function evaluations E, the number of support vectors sv, as well as the training time t (in seconds) assigning various working set sizes for the *australian* as well as the *fourclass* data set.

Starting with 4 active points, we increased the working set size until the maximal value (the whole training set) was reached. For a better interpretation we show the plot of the training times (with some more values) in Fig. 2. The results show similar behavior. For small values of $\hat{l}$ the number of decomposition steps is large and decreases with increasing $\hat{l}$. The number of kernel evaluations is not monotone

$\hat{l}$	4	10	20	30	50	100	200	300	400	500	600	700	800	862
$\#D$	2260	640	41	25	15	13	5	4	4	5	3	4	3	1
$\#E$	7.76	5.33	0.52	0.39	0.34	0.42	0.39	0.54	0.81	1.43	1.22	2.11	2.04	0.83
$\#sv$	98	98	98	98	98	98	98	98	98	98	98	98	98	98
t	1.90	1.42	0.28	**0.26**	0.72	1.10	4.28	5.89	5.01	16.18	12.87	21.78	16.93	9.21

Table 3: Results for the *fourclass* data set.

$\hat{l}$	10	50	100	300	500	650	800	1000	1400	2000	2400	3000	3500	4000
$\#D$	6094	1465	563	109	47	32	28	23	18	13	12	10	9	7
$\#E$	810	1028	777	402	260	212	213	213	222	228	243	260	275	261
$\#sv$	5526	5549	5546	5546	5541	5547	5539	5544	5539	5543	5542	5558	5550	5555
t	310	375	286	156	109	**97**	108	122	187	316	443	639	925	1116

Table 4: Results for the *adultpart* data set.

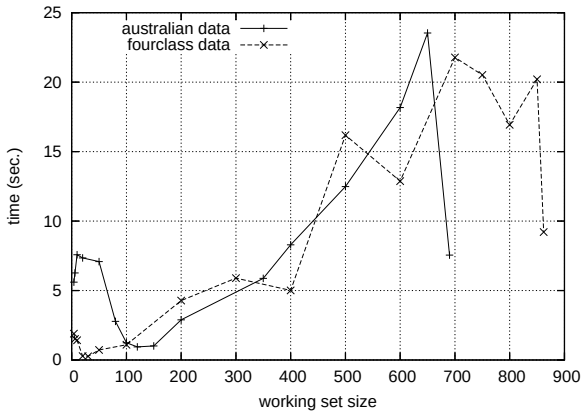


Figure 2: Training times for different working set sizes (small data sets).

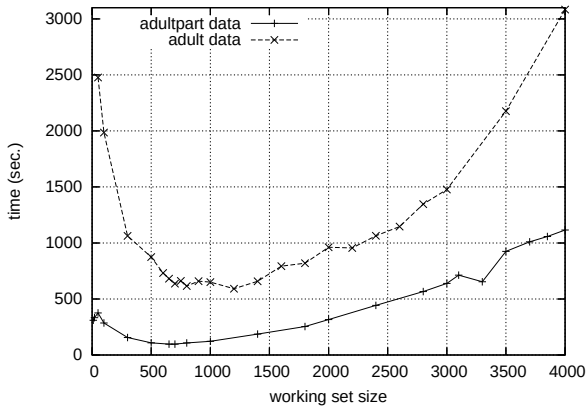


Figure 3: Training times for different working set sizes (huge data sets).

in such a way. For both data sets there is an interval of working set sizes that leads to small numbers of kernel evaluations. For the *australian* data set this interval is approx. $[120, 350]$ and for the *fourclass* data set $[30, 200]$. Please note that for both data sets the largest possible value of $\hat{l}$ led to another minimum for the sum of kernel evaluations. This is due to the fact, that in this extreme case only a single decomposition step is required which saves a lot of overhead. This is an interesting result. However, the choice of $\hat{l} = l$ seems not to be useful. First, it does not always lead to a global minimum of kernel computations as for the *fourclass* data set and second, the training times are not optimal. Although training times show local minima for $\hat{l} = l$, the optimal training times for both data sets are smaller. For the *fourclass* data set we can save a factor of 40 in training time. This behavior comes from the fact, that we implemented a sparse gradient update, as it was introduced in Zanghirati & Zanni (2003*a*). Thus, each iteration of the decomposition method is extremely cheap for small data sets. It seems, that together with a fast inner solver and sophisticated working set selection the small working set sizes beat the bigger ones.

In Tables 4 and 5 we show the number of decomposition steps D, the number of kernel function evaluations E, the number of support vectors sv as well as the training time t (in seconds) for the *adultpart* as well as the *adult* data set using different working set sizes. The corresponding plot with some more values is shown in Fig. 3. Results for both data sets again show similar behavior. The training time curve is convex and has a minimum at $\hat{l} = 650$ for the *adultpart* and $\hat{l} = 1200$ for the *adult* data set. The numbers of kernel evaluations are high for small data sets and decrease for increasing working set sizes. They reach a first minimum for the minimal training times, but, then they remain somehow stable while the training time increases dramatically. This is caused by the inner solver and gradient update costs. Working set sizes between 500 and 1400 led to acceptable training times, and for the largest training set larger working set sizes are preferable.

6.3 Influence of the Working Set Size for Parallel Computation

The parallel decomposition scheme is based on parallel computations of the kernel matrix and parallel gradient update in each decomposition step as well as parallel matrix-vector multiplications in the inner solver. Usually applied to very large working sets (Zanghirati & Zanni 2003*a*, Eitrich & Lang 2006*b*) we showed, that smaller working set sizes lead to better results in the serial case. In this section, we will broaden our analysis to parallel training and run the same tests for the largest data set (*adult*) to show the parallel behavior of the decomposition method for smaller working set sizes.

In Table 6 some results for parallel runs using 1, 2 and 4 threads are given. With $t(\cdot)$ we denote the training time against the number of threads and with $s(\cdot)$ the corresponding speedup value, where the speedup is defined in the usual way as $s(n) = t(1)/t(n)$, where n is the number of threads. In Fig. 4

$\hat{l}$	50	100	300	500	650	800	1200	1600	2000	2400	3000	3500
$\#D$	4472	1774	332	145	100	70	40	35	27	22	18	16
$\#E$	6803	5428	2832	1903	1591	1310	1013	1062	1042	992	1001	1031
$\#sv$	11779	11758	11768	11759	11755	11766	11766	11752	11771	11765	11751	11782
t	2477	1986	1063	874	681	616	**592**	793	962	1064	1477	2176

Table 5: Results for the *adult* data set.

$\hat{l}$	50	100	300	500	650	800	1000	1200	1600	2000	2400	2800	3000	3500
$t(1)$	2477	1986	1063	874	681	616	651	**592**	793	962	1064	1346	1477	2176
$t(2)$	1141	1006	571	460	388	370	379	**347**	458	558	617	795	879	1224
$s(2)$	2.2	2.0	1.9	1.9	1.9	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.9
$t(4)$	659	610	318	276	234	201	200	**191**	250	318	356	457	486	713
$s(4)$	3.8	3.3	3.3	3.2	2.9	3.1	3.3	3.1	3.2	3.0	3.0	2.9	3.0	3.1

Table 6: Parallel training results for the *adult* data set with 1, 2 and 4 threads.

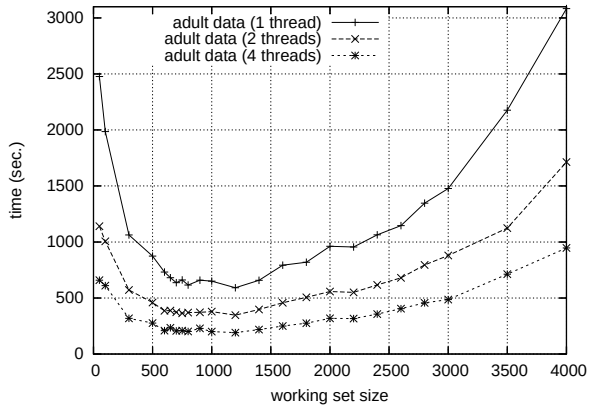


Figure 4: Training times for different working set sizes with 1, 2 and 4 threads (*adult* data set).

we show the corresponding plots for all our tests.

The minimal training time for the parallel runs is again achieved with $\hat{l} = 1200$. The speedup values in this area are comparable to those with larger working sets. Thus, working set sizes around 1000 are preferable in this case, since they lead to the best (smallest) training times in serial as well an parallel mode.

In parallel data mining, the interest is in efficiently using the available resources. In our tests we observed acceptable speedup values for all working set sizes we had chosen. This is due to the fact, that the parallel kernel matrix evaluation is perfectly scalable and the problem size for the parallel gradient update is not dependent on the working set size, so that the parallel scheme works fine. From our tests we conclude, that the optimal working set size is indeed dependent on the data set size and increases for large data sets. Very large working sets lead to high training times in general. In our tests we observed global minima for the training time, that are smaller than we expected so far.

7 Summary and Future Work

We have analyzed the decomposition algorithm for SVM training with a fast inner solver. For several small and large data sets we have tested, how the working set size influences the training time. For small data sets we observed enormous differences, whereas the variability was smaller for the large data sets. However, differences of one order of magnitude may occur easily if choosing a non-optimal working set size. For small optimal working sets, like in our results, the attainable speedups for the parallel SVM training will be limited due to the fact, that the efficiency of parallel matrix-vector multiplications decreases with increasing numbers of CPUs or threads. However, this is not a critical point in SVM learning. As expensive parameter tuning experiments need to be done, remaining CPUs can be assigned to either parallel cross validation schemes (Eitrich et al. 2006) or to parallel parameter optimization methods (Eitrich & Lang 2005).

In the future, we will continue with the first experiments presented in this paper using even larger data sets. In addition, we will work on methods that automatically compute the optimal working set size for parallel SVM learning depending on the data set, the characteristics of the machine used, as well as the validation and/or parameter optimization scheme.

Acknowledgments

The computations were performed on the IBM p690 cluster JUMP at Research Centre Jülich. The authors would like to thank the ZAM team at Jülich for support.

References

Burges, C. J. C. (1998), 'A tutorial on support vector machines for pattern recognition', *Data Mining and Knowledge Discovery* **2**(2), 121–167.

Callahan, P. B. (1993), Optimal parallel all-nearest-neighbors using the well-separated pair decomposition, *in* 'Proceedings of the 34th Symp. Foundations of Computer Science, IEEE', pp. 332–340.

Celis, S. & Musicant, D. R. (2002), Weka-parallel: machine learning in parallel, Computer Science Technical Report 2002b, Carleton College.

Chang, C.-C., Hsu, C.-W. & Lin, C.-J. (2000), 'The analysis of decomposition methods for support vector machines', *IEEE Transactions on Neural Networks* **11**(4), 1003–1008.

Chang, C.-C. & Lin, C.-J. (2001), *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, N., Lu, W., Yang, J. & Li, G. (2004), *Support vector machine in chemistry*, World Scientific Pub Co Inc.

Cristianini, N. & Shawe-Taylor, J. (2000), *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, UK.

Detert, U. (2004), *Introduction to the JUMP architecture.* *http://jumpdoc.fz-juelich.de

Dong, J.-X., Krzyzak, A. & Suen, C. Y. (2003), A fast parallel optimization for training support vector machines, *in* P. Perner & A. Rosenfeld, eds, 'Proceedings of 3rd International Conference on Machine Learning and Data Mining', pp. 96–105.

Dong, J.-X. & Suen, C. Y. (2003), 'A fast SVM training algorithm', *International Journal of Pattern Recognition* **17**(3), 367–384.

Eitrich, T., Frings, W. & Lang, B. (2006), HyParSVM – a new hybrid parallel software for support vector machine learning on SMP clusters, *in* W. E. Nagel, W. V. Walter & W. Lehner, eds, 'Euro-Par 2006, Parallel Processing, 12th International Euro-Par Conference', Vol. 4128 of *LNCS*, Springer, pp. 350–359.

Eitrich, T. & Lang, B. (2005), Parallel tuning of support vector machine learning parameters for large and unbalanced data sets, *in* M. R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher & I. Fischer, eds, 'Computational Life Sciences, First International Symposium (CompLife 2005), Konstanz, Germany', Vol. 3695 of *Lecture Notes in Computer Science*, Springer, pp. 253–264.

Eitrich, T. & Lang, B. (2006*a*), Data mining with parallel support vector machines for classification, *in* T. Yakhno & E. Neuhold, eds, 'ADVIS 2006', Vol. 4243 of *LNCS*, Springer, pp. 197–206.

Eitrich, T. & Lang, B. (2006*b*), Efficient implementation of serial and parallel support vector machine training with a multi-parameter kernel for large-scale data mining, *in* 'Proceedings of the 11. International Conference on Computer Science (ICCS), February 24-26, 2006, Prague, Czech Republic', pp. 6–11.

Fletcher, R. (1981), *Practical methods of optimization, Vol II: constrained optimization*, John Wiley & Sons, Chichester and New York.

Graf, H. P., Cosatto, E., Bottou, L., Dourdanovic, I. & Vapnik, V. (2005), Parallel support vector machines: the cascade SVM, *in* L. K. Saul, Y. Weiss & L. Bottou, eds, 'Advances in Neural Information Processing Systems 17', MIT Press, Cambridge, MA, pp. 521–528.

Hettich, S., Blake, C. L. & Merz, C. J. (1998), *UCI repository of machine learning databases.* http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Ho, T. K. & Kleinberg, E. M. (1996), Building projectable classifiers of arbitrary complexity, *in* 'Proc. of the 13th Int. Conf. on Pattern Recognition, Vienna, Austria', pp. 880–885.

Hofer, J. (2004), Distributed induction of decision tree classifier within the grid data mining framework: Gridminer-core, AURORA Technical Report 2004-04, Institute for Software Science, University of Vienna, Vienna.

Hsu, C. & Lin, C. (2002*a*), 'A comparison of methods for multi-class support vector machines', *IEEE Transactions on Neural Networks* **13**, 415–425.

Hsu, C.-W. & Lin, C.-J. (2002*b*), 'A simple decomposition method for support vector machines', *Machine Learning* **46**(1-3), 291–314.

IBM (n.d.), 'ESSL - Engineering and Scientific Subroutine Library for AIX version 4.1'.

Jin, R., Yang, G. & Agrawal, G. (2005), 'Shared memory parallelization of data mining algorithms: techniques, programming interface, and performance', *IEEE Transactions on Knowledge and Data Engineering* **17**(1), 71–89.

Joachims, T. (1998), Text categorization with support vector machines: learning with many relevant features, *in* 'Proceedings of the 10th European Conference on Machine Learning (ECML 1998), Chemnitz', Vol. 1398 of *Lecture Notes in Computer Science*, Springer, pp. 137–142.

Joshi, M. V., Karypis, G. & Kumar, V. (1998), ScalParC: a new scalable and efficient parallel classification algorithm for mining large datasets, *in* 'IPPS: 11th International Parallel Processing Symposium', IEEE Computer Society Press.

Kantabutra, S. & Couch, A. L. (2000), 'Parallel k-means clustering algorithm on NOWs', *NECTEC Technical Journal* **1**, 243–248.

Laskov, P. (2002), 'Feasible direction decomposition algorithms for training support vector machines', *Machine Learning* **46**(1-3), 315–349.

Lazarevic, A. & Obradovic, Z. (2001), The distributed boosting algorithm, *in* 'KDD 2001: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining', ACM Press, New York, NY, USA, pp. 311–316.

Leyffer, S. (2005), 'The return of the active set method'. to appear in Oberwolfach Report.

Misra, M. (1997), 'Parallel environments for implementing neural networks', *Neural Computing Surveys* **1**, 48–60.

Osuna, E., Freund, R. & Girosi, F. (1997), An improved training algorithm for support vector machines, *in* 'IEEE Workshop on Neural Networks and Signal Processing'.

Parthasarathy, S., Zaki, M., Ogihara, M. & Li, W. (2001), 'Parallel data mining for association rules on shared-memory systems', *Knowledge and Information Systems* **3**(1), 1–29.

Platt, J. (1999), Fast training of support vector machines using sequential minimal optimization, *in* B. Schölkopf, C. J. C. Burges & A. J. Smola, eds, 'Advances in Kernel Methods — Support Vector Learning', MIT Press, Cambridge, MA, pp. 185–208.

Poulet, F. (2003), Multi-way distributed SVM algorithms, *in* 'Proc. of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003), Cavtat-Dubrovnik, Croatia, 2003', Vol. 2838 of *Lecture Notes in Computer Science*, Springer.

Qiu, S. & Lane, T. (2005), Parallel computation of RBF kernels for support vector classifiers, *in* 'SDM'.

Ruggiero, V. & Zanni, L. (1999), 'On the efficiency of splitting and projection methods for large strictly convex quadratic programs', *Applied Optimization* pp. 401–413.

Ruggiero, V. & Zanni, L. (2000), 'Variable projection methods for large convex quadratic programs', *Recent Trends in Numerical Analysis* pp. 299–313.

Ruggiero, V. & Zanni, L. (2001), 'An overview on projection-type methods for convex large-scale quadratic programs', *Nonconvex Optimization and Its Applications* **58**, 269–300.

Schapire, R. E. (1999), A brief introduction to boosting, *in* 'Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence'.

Schölkopf, B. (2001), The kernel trick for distances, *in* 'Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA', MIT Press, pp. 301–307.

Schölkopf, B. & Smola, A. J. (2002), *Learning with kernels*, MIT Press, Cambridge, MA.

Serafini, T., Zanghirati, G. & Zanni, L. (2004), Parallel decomposition approaches for training support vector machines, *in* 'Proceedings of the International Conference on Parallel Computing (ParCo 2003), Dresden, Germany', Elsevier, pp. 259–266.

Serafini, T. & Zanni, L. (2005), 'On the working set selection in gradient projection-based decomposition techniques for support vector machines', *Optimization Methods and Software* **20**(4-5). Special issue on Numerical Methods for Local and Global Optimization: Sequential and Parallel Algorithms.

Skillicorn, D. B. (1998), Strategies for parallelizing data mining, *in* 'Proceedings of the Workshop on High-Performance Data Mining, in association with IPPS/SPDP'.

Vapnik, V. N. (1998), *Statistical learning theory*, John Wiley & Sons, New York.

Yu, H., Yang, J. & Han, J. (2003), Classifying large data sets using SVMs with hierarchical clusters, *in* L. Getoor, T. E. Senator, P. Domingos & C. Faloutsos, eds, 'Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003', ACM, pp. 306–315.

Yu, H., Yang, J., Wang, W. & Han, J. (2003), Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines., *in* '2nd IEEE Computer Society Bioinformatics Conference (CSB 2003), 11-14 August 2003, Stanford, CA, USA', IEEE Computer Society, pp. 220–228.

Zanghirati, G. & Zanni, L. (2003*a*), 'A parallel solver for large quadratic programs in training support vector machines', *Parallel Computing* **29**(4), 535–551.

Zanghirati, G. & Zanni, L. (2003*b*), Variable projection methods for large quadratic programs in training support vector machines, Technical report, Department of Mathematics, University of Modena and Reggio Emilia, Italy.

Marking Time in Sequence Mining

Carl H. Mooney and John F. Roddick

School of Informatics and Engineering
Flinders University of South Australia,
PO Box 2100, Adelaide, South Australia 5001,
Email: {carl.mooney, roddick}@infoeng.flinders.edu.au

Abstract

Sequence mining is often conducted over static and temporal datasets as well as over collections of events (episodes). More recently, there has also been a focus on the mining of streaming data. However, while many sequences are associated with absolute time values, most sequence mining routines treat time in a relative sense, only returning patterns that can be described in terms of Allen-style relationships (or simpler).

In this work we investigate the accommodation of timing marks within the sequence mining process. The paper discusses the opportunities presented and the problems that may be encountered and presents a novel algorithm, $INTEM_{TM}$, that provides support for timing marks. This enables sequences to be examined not only in respect of the order and occurrence of tokens but also in terms of pace. Algorithmic considerations are discussed and an example provided for the case of polled sensor data.

1 Introduction

Frequent-pattern (sequence) mining from static databases has been conducted for a number of years and algorithms for this form of mining are relatively mature (Pei et al. 2001, Srikant & Agrawal 1996, Wang & Han 2004, Yan et al. 2003). Transaction datasets commonly include a *time-stamp* for each transaction and it is this that can be used, in conjunction with a *transaction_id*, to constrain the mining activity with respect to time.

However, sequence mining is not limited to data stored in transaction-structured datasets and there are other domains where an implicit time-stamp may or may not be included such as web logs, alarm data in telecommunications networks, sensor data, and so on. In such domains, the data can be viewed as a series of events occurring at specific times and therefore the problem becomes a search for collections of events (episodes) that occur frequently together. Solving this problem requires a different approach, and several types of algorithm have been proposed for different domains (Mannila & Toivonen 1996, Mannila et al. 1997, Mooney & Roddick 2004, Spiliopoulou 1999).

Such datasets can also be very similar in nature to, or are themselves, streaming datasets, an area of research that is gaining significant interest at present (Gaber et al. 2005, Giannella et al. 2003, Lin et al. 2003). However, the datasets used in these domains do not always include a *time-stamp* and this reduces the problem to those that occur close to each other in the sequence. This changes the semantics of *frequent* and makes mining more problematic if time constraints are required, or if information relative to the pace of the activity is of interest. However, in some datasets, the passage of time, while not being available as a full time-stamp, may be marked by a token representing a *timing tick*.

In this paper we address this problem by introducing the notion of a ***timing mark*** (or ***timing tick***) to accommodate the passage of time within the sequence mining process. This allows the process not only to provide information relative to order and occurrence of sequences but also the pace at which they occurred.

The remainder of the paper is organised as follow. Section 2 briefly discusses background material on sequence mining and related work. Section 3 introduces the concept of the *timing mark* and discusses the opportunities presented and potential problems that may be encountered. Section 4 deals with algorithmic considerations and presents a novel algorithm $INTEM_{TM}$ that provides support for the concept of *timing marks*. Section 5 provides experimental results resulting from the implementation while Section 6 offers some conclusions and suggestions for future work.

2 Background and Related Work

The sequential pattern mining problem can be viewed from both a static dataset and episodic point of view; the latter being the area most closely related to the mining discussed in this work. We outline below some definitions of the related areas and previous research in sequence, episodic and time series mining is briefly discussed.

2.1 Sequential Pattern Mining

Given a dataset of sequences, where each sequence consists of a list of transactions ordered by transaction time and each transaction is a set of items, the aim of sequential pattern mining is to discover all sequential patterns with a user-specified minimum support, where the support of a pattern is the number of data-sequences that contain that pattern (Agrawal & Srikant 1995).

The problem of mining sequential patterns can be stated as follows: Let $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, termed *items*, which comprise the alphabet. An *event* is a non-empty unordered collection of items. It is assumed without loss of generality that items of an event are sorted in lexicographic order. A *sequence* is an ordered list of events. An event is

denoted as $(i_1, i_2, \ldots, i_k)$, where i_j is an item. A sequence α is denoted as $\langle \alpha_1 \rightarrow \alpha_2 \rightarrow \cdots \rightarrow \alpha_q \rangle$, where α_i is an event. A sequence with k-items ($k = \sum_j |\alpha_j|$) is called a *k-sequence*. For example, $\langle B \rightarrow AC \rangle$ is a 3-sequence. A sequence $\langle \alpha_1 \rightarrow \alpha_2 \ldots \rightarrow \alpha_n \rangle$ is a *subsequence* of another sequence $\langle \beta_1 \rightarrow \beta_2 \ldots \rightarrow \beta_m \rangle$ if there exist integers $i_1 < i_2 < \ldots < i_n$ such that $\alpha_1 \subseteq \beta_{i_1}, \alpha_2 \subseteq \beta_{i_2}, \ldots, \alpha_n \subseteq \beta_{i_n}$. For example the sequence $\langle B \rightarrow AC \rangle$ is a subsequence of $\langle AB \rightarrow E \rightarrow ACD \rangle$, since $B \subseteq AB$ and $AC \subseteq ACD$, and the order of events is preserved. However, the sequence $AB \rightarrow E$ is not a subsequence of ABE and vice versa.

The process is provided with a dataset $\mathcal{D}$ of input-sequences where each input-sequence in the dataset has the following fields: sequence-id, event-time and the items present in the event. It is assumed that no sequence has more than one event with the same time-stamp, so that the time-stamp may be used as the event identifier. In general, the *support* or *frequency* of a sequence, denoted $\sigma(\alpha, \mathcal{D})$, is defined as the total number of input-sequences in the dataset $\mathcal{D}$ that contain α[1]. Given a user-specified *minimum support* threshold (denoted *min_supp*), a sequence is said to be *frequent* if it occurs at least *min_supp* times and the set of frequent k-sequences is denoted as $\mathcal{F}_k$. A frequent sequence is deemed to be *maximal* if it is not a subsequence of any other frequent sequence. The task then becomes the discovery of all frequent sequences from a dataset $\mathcal{D}$ of input-sequences and a user supplied *min_supp*.

2.2 Episodic Mining

The first algorithmic framework developed to mine datasets that were episodic in nature was introduced by Mannila et al. (1995). The task was to find all episodes that occur frequently in an event sequence, given a class of episodes and an input sequence of events. In their framework an episode is defined to be:

> *"... a collection of events that occur relatively close to each other in a given partial order*, and ... *frequent episodes as a recurrent combination of events"* (Mannila et al. 1997)

The notation used is as follows.

E is a set of event types and an event is a pair (A, t), where $A \in \mathbf{E}$ is an event type and t is the time (occurrence) of the event. There are no restrictions on the number of attributes that an event type may contain, but the paper only considers single values with no loss of generality. An event sequence **s** on **E** is a triple (s, T_s, T_e), where $s = \langle (A_1, t_1), (A_2, t_2), \ldots, (A_n, t_n) \rangle$ is an ordered sequence of events such that $A_i \in \mathbf{E}$ for all $i = 1, \ldots, n$, and $t_i \leq t_{i+1}$ for all $i = 1, \ldots, n-1$. Further $T_s < T_e$ are integers, T_s is called the starting time and T_e the ending time, and $T_s \leq t_i < T_e$ for all $i = 1, \ldots, n$.

2.3 Time Series Mining

Data mining of time series datasets includes not only sequence mining but also clustering, classification, and association mining (Lin et al. 2003, Das et al. 1998, Guralnik & Srivastava 1999, Höppner 2001, Keogh et al. 1993). As would be expected, the constraints available each time are those appropriate for the form of mining and the rules that emerge from this type of analysis are similarly aligned with the mining method chosen. For the case of sequences, typical rules are based on (*a priori* supplied) calendric, or cyclic patterns and have some similarity to those addressed in this paper.

[1] This general definition has been modified as algorithmic development has progressed and different methods for calculating support have been introduced, see the work of Joshi et al. (1999) for a complete summary of counting techniques.

3 Timing Marks

The concept of *timing marks* introduced here refers to embedded tokens that indicate the passage of time. They are not *time-stamps* in that they do not record absolute time values but rather *ticks* which can be referenced to determine the pace of events [2]. For example, the notion of polled data infers a (fixed) time interval during which the polling occurs. During this interval, it may be possible that not all sensors are read and/or some do not return data. Moreover, many sequences of events have a time-stamp, either inherently in how they are reported, or overlaid by a system that needs to interrogate the data. How this fixed time element is encoded in the data is of interest. In traditional sequence mining, time-series mining and web-log mining each element to be mined has a time-stamp associated with it and therefore encoding an additional *timing mark* is not necessary. With sensor data and other data streams there is usually no time-stamp and therefore it is necessary to include a time-stamp or *timing mark* into the data.

In our recent work on mining interacting episodes we have implicitly assumed (in common with other researchers) that each token (sensor reading) occurred for a fixed period of time and that the time between tokens was zero (or alternatively, that events are instantaneous and the time between tokens was constant). That is, we could view a sequence of n tokens as occurring over n time periods of equal length, no matter what the period/granularity was (Mooney & Roddick 2004). This work relaxes this assumption. That is, although the time between events may remain unchanged – equal length intervals – the number of tokens (events) that occur within that time can vary. To accommodate this assumption we have introduced *timing marks* into the data. These *timing marks* may have different properties depending on the data they are associated with and more generally *timing marks* can be viewed as having the following possibilities[3]:

Timing marks as tokens. One of the polled sensors is used as the timing mark which would mean that all time-referenced sequences would be reported with reference to this sensor. One problem with this option is that the sensor used as the timing mark may not fire regularly and as such any rules that are reported may or may not have value. If however the sensor is firing in every cycle then its usefulness from a reporting standpoint is valuable in the same way as if it were a delimiter.

Timing marks added as delimiters. In this option, timing marks are added as additional tokens to the sequence. This is necessary where all other tokens are sporadic as is the case with many types of sequence. For our purposes, this is our assumed format.

[2] The consensus glossary (Jensen et al. 1998) delineates between two forms of time - absolute and relative. Timing marks are, in many respects, both and neither of these possessing an absolute time period relative to each other but little else. Certainly, little of the current temporal data mining research (Roddick & Spiliopoulou 2002) can handle timing marks.

[3] In the following examples the period '.' is used as the notation of the *timing mark*.

Timing marks as absolute time. In some cases, each token carries with it an absolute time stamp. In this case there is more information that is required for our purposes here and it would be trivial to convert such a sequence to one that contained timing marks as delimiters.

The value of timing marks becomes apparent when queries can be issued and results reported with respect to timing marks. A given sequence, for example, could be deemed as "fast/bursty" or as "slow". For instance, the difference between the two sequences **ABC** and **A...B..C** may be significant even if they occur within the normal lookahead.

More significantly, the semantics of rules using temporal relationships such as $A - during \rightarrow B$ or $A - meets \rightarrow B$ may change depending on the number of timing marks that have been encountered. For instance, to allow for recording latency, two intervals may be deemed to *meet* if they occur within n timing marks.

4 Algorithmic Considerations

Timing marks can be either present or absent in a data stream and as such users should have the opportunity to include or exclude the timing marks in their search for frequent episodes. Consequently, the timing marks feature has necessarily been implemented as a constraint, thus allowing the user to select the token that is the timing mark and, in addition, choose whether to report those episodes that contain exactly the prescribed number of timing marks or all episodes up to and including the prescribed number of timing marks.

Since we provide the user with the choice, it makes sense for the implementation of this constraint to be post episode discovery. To further reinforce this decision, the token used for the timing mark may be one of the data tokens, not one that is orthogonal to the data, in which case the user may not wish to remove the token from those episodes that are reported. In order to facilitate the fact that the timing mark may be one of the data tokens, the input file is scanned upon selection to generate a list of those tokens available. This incurs no overhead since the file has to be read before further processing can be undertaken. Since this is an added constraint, the impact on the existing algorithm is minimal.

The two parameters used are:

- The **lookahead** (or window) parameter used in previous work (Mooney & Roddick 2004) (similar to Mannila *et al.*'s window concept (Mannila et al. 1997)), defines the maximum length episode to be mined) is included, together with
- a *timing mark count* (**tmc**), which defines either the maximum number of timing marks that can be included *or* the exact number of timing marks that should occur in the sequence.

Since both of these measures can be used, for the purpose of "frequent", a sequence up to *lookahead* must also occur within the prescribed number of marks – ie. the cut-off is either the *lookahead* or *timing mark count* whichever is the smaller.

4.1 Timing Mark Pruning

If the user has chosen to include the timing marks in their search then the following will occur after the frequent episodes have been discovered. First, pruning will be conducted on the frequent sequences so that only those that contain the prescribed number of timing marks, see Algorithm 4.1, will remain. If the timing mark is not determined to be one of the tokens in the data, then removal of the timing marks from those remaining sequences will ensue, and finally reassignment of them to the correct output containers, see Algorithm 4.2.

For timing marks to remain unambiguous to the user and therefore be consistent throughout the application then the following convention is adopted:

1) Within one mark means that there are no *timing marks* allowed in the sequence. Algorithmically this can be described by – assuming the timing mark is "." –

   if (tmc = 1 $\wedge$ cand.indexOf(".") $\neq -1$) **then**
   set output to null
   end if
   return

   This also leads to an added pruning technique – i.e. in the case of one timing mark, if we are looking for an x length sequence and the last item in the sequence is a *timing mark*, then the next x sequences are not viable candidates so can be eliminated from the search.

2) Within one or more *timing mark*s. During one timing mark is as described above while n marks indicates that there are n distinct sections in the sequence which would have embedded $n-1$ timing marks.

Algorithm 4.1 Algorithm for imposing timing marks on sequence discovery

Input: A set of frequent sequences that are to be pruned for timing marks.
Output: the collection of frequent sequences according to the timing mark constraints.

```
1: procedure PRUNEFORTIMINGMARKS(ArrayList aList)
2:     for (i := 0; i < aList.size(); i++) do
3:         TreeMap tm := aList.get(i);
4:         TreeMap clTm := tm.clone();
5:         for all (String cand : clTm.keySet()) do
6:             int numMarks = countTimingMarks(cand);
7:             if (exactly_selected) then
8:                 if (numMarks ≠ maxMarks) then
9:                     tm.remove(cand);
10:                end if
11:            else
12:                if (numMarks > maxMarks) then
13:                    tm.remove(cand);
14:                end if
15:            end if
16:        end for
17:    end for
18: end procedure
```

Algorithm 4.2 Removes the timing marks from the frequent sequences and reassigns them to the correct output containers.

Input: a list of frequent episodes that have been pruned for the required number of timing marks.
Output: the required frequent episodes without timing marks.

```
1: procedure REMOVEALLTIMINGMARKS(ArrayList aList)
2:     ArrayList modList := new ArrayList();
3:     for all (TreeMap tmap : aList) do
4:         TreeMap modTree := new TreeMap();
5:         for all (String cand : tmap.keySet()) do
6:             String newCand := removeTimingMarks(cand);
7:             if (!newCand.equals("")) then
8:                 modTree.put(newCand, tmap.get(cand));
9:             end if
10:        end for
11:        modList.add(modTree);
12:    end for
13:    frequentList := reassignEpisodes(modList);
14: end procedure
```

4.2 Rule semantics

Typically rules from sequence discovery are of the type that can described in terms of Allen-style (Allen 1983) relationships (or simpler). This is the case not only for market-basket mining (Agrawal & Srikant 1995, Ayres et al. 2002, Garofalakis et al. 1999, Han et al. 2000), but also episodic mining (Mannila et al. 1997, Mooney & Roddick 2004). In the case of episodic mining both parallel and serial episodes yield these types of rules. When using *timing marks* as delimiters the following possibilities, similar to those of episodic mining, must be considered:

1) if the sequences occur within the interval delimited by a pair of *timing marks*, for example, **.ABCDEF.**, then this is analogous to parallel episodes[4], or

2) if the sequences must occur within a certain number of *timing marks*. For example, **.AB.CDE.F.**, then it is analogous to serial episodes.

In the first case above, the discovered sequences could be treated as transactions (if order is irrelevant) and therefore further processing may be conducted using other data mining methods, such as association rule mining (Ceglar & Roddick 2006). In the second case details about the 'speed' of discovered sequences can be obtained with respect to the number of timing marks that are contained, allowing for a better understanding of the data.

If the *timing mark* is viewed as a fixed length period with no absolute time-stamp associated with it then we can search for sequences that occur under both of the above conditions. For example, given the sequence **.ABCDEFG.HIJ..** with a maximum look ahead of 5:

1) For sequences that occur within one cycle – **ABCDE**, **BCDEF**, and **CDEFG** are all valid while **FGHIJ** is not. This may be useful to determine if certain sensors did not fire during a particular cycle.

2) For sequences that occur over a period of x time cycles – sequence **FGHIJ** occurs over two time cycles.

Given this information, the knowledge of the position of the *timing mark* allows for added semantics to be attached to the sequence – not only can we say that **FGHIJ** occurs over two time cycles but also that first cycle is ended with **FG** and the second is begun with **HIJ**. This may have added interest, depending on the application, to any resulting output that may be derived.

5 Experimental Results

The algorithm was implemented in the $Java^{TM}$ programming language and all experiments were conducted on a 2.6GHz AMD machine running $Windows^{®}$ XP with 1Gb of RAM (see Figure 1). The $INTEM_{TM}$ implementation represents an extension to the $INTEM$ (**INT**eracting **E**pisode **M**iner), which also includes graphical output, as well as text, of the discovered sequences. Furthermore, interactions may also be discovered and reported using Allen-style relationships (Allen 1983), Freksa's conceptual neighbourhoods (Freksa 1992) or more fine grained Midpoint relationships (Roddick & Mooney 2005). Since the user has control over the number and method of timing mark inclusion the "speed" of the discovered sequences is known and can be reported easily.

[4]This would be data dependent and would rely on whether the order within the marks is relevant.

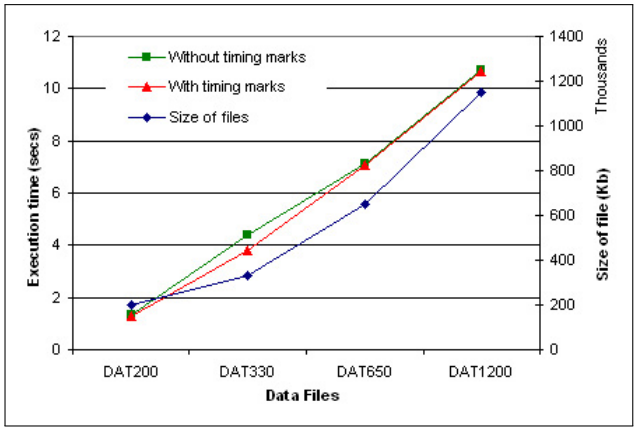


Figure 2: Execution times with and without timing marks for the test files.

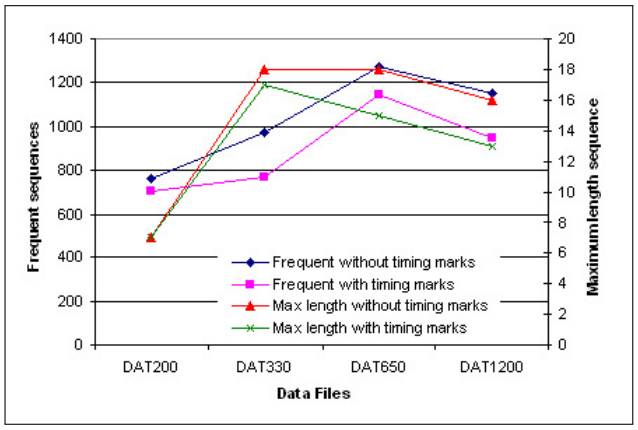


Figure 3: Number of frequent sequences and maximum length sequences with and without timing marks.

All tests were conducted using a low support level (0.005), a *lookahead* value of 20 and, when including timing marks, a *timing mark count* (tmc) of 2 with the reporting option set to exclude the marks from the output. The results show that there is no overhead incurred when using the timing mark option, see Figure 2. Indeed, since the constraint is implemented deeper in the process there is a slight speed increase when looking for sequences containing timing marks. The reason for the speed up can also be seen, (see Figure 3), by e fact there are less sequences discovered with the timing mark option selected and in the majority of cases the maximum length of the discovered sequences is smaller.

6 Conclusions and Future Work

In this paper we have discussed the inclusion of timing marks for dealing with data that have no absolute time attached to the events to be mined. We have shown that the implementation of the algorithm incurs negligible added overhead and that the benefits associated with the rules that may be reported are important in terms of being able to determine the pace of a sequence.

Future research is necessary in this area to accommodate this feature into algorithms that can deal with streaming data – an already complex domain (see (Gaber et al. 2005)) Further research is also needed in the area of rule generation, together with some consideration of the resultant semantics of the

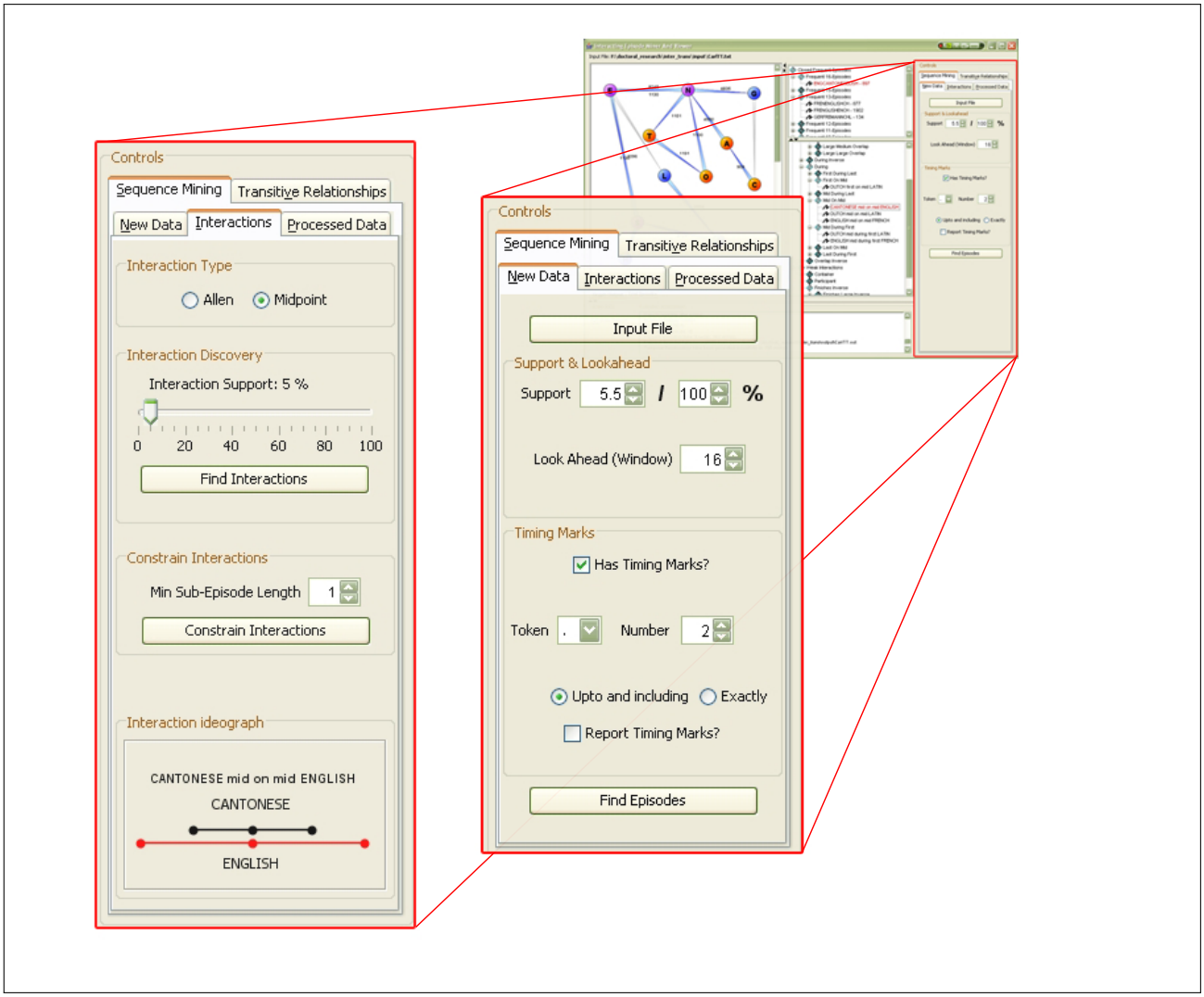


Figure 1: Screenshot of the experimental system.

rules. This latter issue is highly dependent on the data being mined and therefore careful consideration is needed.

References

Agrawal, R. & Srikant, R. (1995), Mining sequential patterns, *in* P. S. Yu & A. S. P. Chen, eds, '11th International Conference on Data Engineering (ICDE'95)', IEEE Computer Society Press, Taipei, Taiwan, pp. 3–14.

Allen, J. (1983), 'Maintaining knowledge about temporal intervals', *Communications of the ACM* **26**(11), 832–843.

Ayres, J., Flannick, J., Gehrke, J. & Yiu, T. (2002), Sequential pattern mining using a bitmap representation, *in* '8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, Edmonton, Alberta, Canada, pp. 429–435.

Ceglar, A. & Roddick, J. F. (2006), 'Association mining', *ACM Computing Surveys* **38**(2).

Das, G., Lin, K.-I., Mannila, H., Renganathan, G. & Smyth, P. (1998), Rule discovery from time series, *in* '4th International Conference on Knowledge Discovery and Data Mining (KDD-98)', AAAI Press.

Freksa, C. (1992), 'Temporal reasoning based on semi-intervals', *Artificial Intelligence* **54**(1-2), 199–227.

Gaber, M. M., Zaslavsky, A. & Krishnaswamy, S. (2005), 'Mining data streams: a review', *SIGMOD Record* **34**(2), 18–26.

Garofalakis, M. N., Rastogi, R. & Shim, K. (1999), SPIRIT: Sequential pattern mining with regular expression constraints, *in* M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik & M. L. Brodie, eds, '25th International Conference on Very Large Data Bases, VLDB'99', Morgan Kaufmann, Edinburgh, Scotland, UK, pp. 223–234.

Giannella, C., Han, J., Pei, J., Yan, X. & Yu., P. (2003), Mining frequent patterns in data streams at multiple time granularities, *in* H. Kargupta, A. Joshi, K. Sivakumar & Y. Yesha, eds, 'Next Generation Data Mining'.

Guralnik, V. & Srivastava, J. (1999), Event detection from time series data, *in* S. Chaudhuri & D. Madigan, eds, '5th International Conference on Knowledge Discovery and Data Mining', ACM Press, San Diego, CA, USA, pp. 33–42.

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U. & Hsu, M.-C. (2000), Freespan: frequent pattern-projected sequential pattern mining, *in* '6th ACM SIGKDD International Conference

on Knowledge Discovery and Data Mining', ACM Press, Boston, MA, USA, pp. 355–359.

Höppner, F. (2001), 'Discovery of temporal patterns - learning rules about the qualitative behaviour of time series'.

Jensen, C. S., Clifford, J., Elmasri, R., Gadia, S. K., Hayes, P., Jajodia, S., Dyreson, C., Grandi, F., Kafer, W., Kline, N., Lorentzos, N., Mitsopoulos, Y., Montanari, A., Nonen, D., Peressi, E., Pernici, B., Roddick, J. F., Sarda, N. L., Scalas, M. R., Segev, A., Snodgrass, R. T., Soo, M. D., Tansel, A., Tiberio, P. & Wiederhold, G. (1998), A consensus glossary of temporal database concepts - february 1998 version, *in* O. Etzion, S. Jajodia & S. Sripada, eds, 'Temporal Databases - Research and Practice', Vol. 1399, Springer, pp. 367–405.

Joshi, M. V., Karypis, G. & Kumar, V. (1999), Universal formulation of sequential patterns, Technical Report Under Preparation #99-21, Department of Computer Science, University of Minnesota.

Keogh, E., Chu, S., Hart, D. & Pazzani, M. (1993), Segmenting time series: A survey and novel approach, *in* 'Data Mining in Time Series Databases', World Scientific Publishing Company.

Lin, J., Keogh, E., Lonardi, S. & Chiu, B. (2003), A symbolic representation of time series, with implications for streaming algorithms, *in* '8th ACM SIGMOD Workshop on Research issues in Data Mining and Knowledge Discovery, DMKD'03', ACM Press, pp. 2–11.

Mannila, H. & Toivonen, H. (1996), Discovering generalised episodes using minimal occurrences, *in* '2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)', AAAI Press, Menlo Park, Portland, Oregon, pp. 146–151.

Mannila, H., Toivonen, H. & Verkamo, A. I. (1995), Discovering frequent episodes in sequences, *in* U. M. Fayyad & R. Uthurusamy, eds, '1st International Conference on Knowledge Discovery and Data Mining (KDD-95)', AAAI Press, Menlo Park, CA, USA, Montreal, Quebec, Canada, pp. 210–215.

Mannila, H., Toivonen, H. & Verkamo, A. I. (1997), 'Discovery of frequent episodes in event sequences', *Data Mining and Knowledge Discovery* **1**(3), 259–289.

Mooney, C. H. & Roddick, J. F. (2004), Mining relationships between interacting episodes, *in* M. W. Berry, U. Dayal, C. Kamath & D. Skillicorn, eds, '4th SIAM International Conference on Data Mining (SDM'04)', SIAM, Lake Buena Vista, Florida.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. & Hsu, M.-C. (2001), PrefixSpan: Mining sequential patterns efficiently by prefix projected pattern growth, *in* '2001 International Conference of Data Engineering (ICDE'01)', Heidelberg, Germany, pp. 215–226.

Roddick, J. F. & Mooney, C. H. (2005), 'Linear temporal sequences and their interpretation using midpoint relationships', *IEEE Transactions on Knowledge and Data Engineering* **17**(1), 133–135.

Roddick, J. F. & Spiliopoulou, M. (2002), 'A survey of temporal knowledge discovery paradigms and methods', *IEEE Transactions on Knowledge and Data Engineering* **14**(4), 750–767.

Spiliopoulou, M. (1999), Managing interesting rules in sequence mining, *in* 'Principles of Data Mining and Knowledge Discovery', pp. 554–560.

Srikant, R. & Agrawal, R. (1996), Mining sequential patterns: generalisations and performance improvements, *in* P. M. G. Apers, M. Bouzeghoub & G. Gardarin, eds, 'International Conference on Extending Database Technology, EDBT'96', Vol. 1057 of *LNCS*, Springer, Avignon, France, pp. 3–17.

Wang, J. & Han, J. (2004), BIDE: Efficient mining of frequent closed sequences, *in* '20th International Conference on Data Engineering', IEEE Press, pp. 79–90.

Yan, X., Han, J. & Afshar, R. (2003), Clospan: Mining closed sequential patterns in large datasets, *in* '3rd SIAM International Conference on Data Mining (SDM'03)', San Francisco, CA.

Discovering Debtor Patterns of Centrelink Customers

Yanchang Zhao[1], Longbing Cao[1], Yvonne Morrow[2], Yuming Ou[1], Jiarui Ni[1], and Chengqi Zhang[1]

[1] Faculty of Information Technology, University of Technology, Sydney
PO Box 123, Broadway, NSW 2007, Australia

{yczhao, lbcao, yuming, jiarui, chengqi}@it.uts.edu.au

[2] Business Integrity Strategy Branch, Centrelink, Australia
PO Box 312, Sunshine, VIC 3020, Australia

yvonne.y.morrow@centrelink.gov.au

Abstract

Data mining is currently becoming an increasingly hot research field, but a large gap still remains between the research of data mining and its application in real-world business. As one of the largest data users in Australia, Centrelink has huge volume of data in data warehouse and tapes. Based on the available data, Centrelink is seeking to find underlying patterns to be able to intervene earlier to prevent or minimize debt. To discover the debtor patterns of Centrelink customers and bridge the gap between data mining research and application, we have done a project on improving income reporting to discover the patterns of those customers who were or are in debt to Centrelink. Two data models were built respectively for demographic data and activity data, and decision tree and sequence mining were used respectively to discover demographic patterns and activity sequence patterns of debtors. The project produced some potentially interesting results, and paved the way for more data mining applications in Centrelink in near future.

Keywords: Data mining, decision tree, association rule, sequence mining

1 Introduction

Data mining is currently becoming an increasingly hot research field, but a large gap still remains between the research of data mining and its application in real-world business. As one of the largest data users in Australia, Centrelink has huge volume of data in data warehouse and tapes. Centrelink raised over $900 million worth of customer debts, excluding Child Care Benefits and Family Tax Benefits, in the year 2004-05 (Centrelink 2005). Moreover, Customer contact generates massive quantities of activity transactions. For example, Centrelink processed 5.2 billion transactions in the year 2004-2005 (Centrelink 2005). These transactions may contain important information related to both debt prevention and the achievement of Government Social Security objectives.

To discover the debtor patterns of Centrelink customers and bridge the gap between data mining research and application, we have done a project on improving income reporting to discover the patterns of those customers who were or are in debt to Centrelink. Two data models were built respectively for demographic data and activity data, and decision tree and sequence mining were used respectively to discover demographic patterns and activity sequence patterns of debtors.

We used the following analysis methods to discover the demographic characteristics and activity sequence patterns of debtors, which may provide information to help know who the debtors are, why the debts occur, and under what conditions debts has a high probability of occurring.

- decision tree for classification of debtor/non-debtor
- association mining for frequent customer circumstances
- sequence mining for activity sequence patterns

Two softwares, Teradata Warehouse Miner (TWM) (Teradata 2005) and Teradata SQL Assistant, were used for the above analysis. Most of our analytical models for mining debtor patterns were either directly based on the modules of TWM or on the improved SQL codes generated by TWM.

With the help of the above tools, we studied Centrelink data related to debtors, built data models and analytical models, and then produced some potentially interesting results, such as the demographic characteristics and the activity patterns of debtors. Our models and methodologies and the experience acquired in this project paved the way for more data mining applications in Centrelink in near future.

The rest of the paper is organised as follows. The business problem and the data available in this project are discussed in Section 2. Then we present the demographic data model, data mining model and results in Section 3. Section 4 describes the activity sequence data model, data mining model and results. The last section concludes the paper and discusses some future work.

2 Business Problem and Available Data

In this section, the business problem will be introduced and the available data will be described.

2.1 Business Problem

Centrelink is a government agency delivering a range of Commonwealth services to the Australian community. It distributes approximately $63 billion in social security payments on behalf of policy departments. The following are some statistics (Centrelink 2006). Centrelink:

- has 6.4 million customers, or approximately one-third of the Australian population,
- administers more than 140 different products and services for 25 government agencies, and
- pays 9.98 million individual entitlements each year and records 5.2 billion electronic customer transactions each year.

From the above figures, we can see that it has not only a very large population but also very large volume of customer data. Moreover, it has huge volume of transaction data which records its activities, which is shown by the following statistics. It

- has sent 87.2 million letters customers each year,
- receives more than 32.68 million telephone calls each year,
- receives 39.5 million website page views each year,
- grants 2.77 million new claims each year,
- conducts more than 98,700 Field Officer reviews each year, and
- has more than 650,000 booked office appointments each month.

Among the large population, there are some people purposely trying to maximize their payments from Centrelink by reporting less income or inadvertently claiming more benefits than they are entitled, which leads to a large amount of debts. The fraud statistics from Centrelink (Centrelink 2006) show that: between 1 July 2004 and 30 June 2005,

- Centrelink conducted 3.8 million entitlement reviews, which resulted in 525,247 payments being cancelled or reduced.
- Almost $43.2 million a week was saved and debts totaling $390.6 million were raised as a result of this review activity.
- Included in these figures were 55,331 reviews of customers from tip-offs received from the public, resulting in 10,022 payments being cancelled or reduced and debts and savings of $103.1 million.
- There were 3,446 convictions for welfare fraud involving $41.2 million in debts.

All the above figures show that debt prevention is a very important task for Centrelink. From the above statistics, we can see that about 14% out of all entitlement reviews resulted in debts, with a lot of money saved. However, 86% reviews resulted in nil debt. Therefore, much effort can be saved if we can target those customers who are debtors or are of high probability to have debts and then conduct reviews only on those customers. From the perspective of activity, if we can detect some activity sequence patterns are associated with debts, then something can be done in advance to prevent or reduce debts. Based on the above idea, we conducted a project to discover demographic characteristics and activity sequence patterns of debtors, expecting that the results may help to target specific customer groups or activity sequence patterns associated with high probability of debts. On the basis of the discovered patterns, more data mining work can be done on debt detection and then a debt prevention and/or reduction system can be built in near future.

2.2 Data Available

The Centrelink corporate database contains information about Centrelink customers and activities used to support them as well as activities to manage the government business. Centrelink delivers help and payments to customers on behalf of the Government and at the same time maintains tight controls over those payments to keep the integrity of the system accountable and transparent.

Newstart Allowance (NSA) is an allowance for those unemployed residents who are aged 21 or over and under Age Pension age. The population of this project is those customers who benefit from NSA. NSA is composed of up to four parts: Basic rate, RA (Rent assistance), PHA (Pharmaceutical allowance) and RAA (Remote area allowance). NSA customers can be classified into three groups: those with income debt, those with other debts and those without debt. This project concentrates on those NSA customers with income debt and analyses their characteristics and activities.

There are three kinds of data which relates to the above problem: customer data from Newsnap database, debt data from Debt database and activity data from AMHS database.

- Newsnap database: It keeps summary information of Newstart Sytems (NSS) cluster population, including personal information about the customer, such as gender, age, marital status, and indigenous indicator as well as address movements, income and migrant information.
- Debt database: It is used by debt management and contains debt details, co-debtors, recoverable debts and overpayments.
- Transaction database: It contains data about transactions and activity management before they are applied to the database. For the activities, what's available in mainframe is those activities after the beginning of 2006. Those data before that are stored in tapes. Therefore, only the activities from 1/1/2006 to 31/3/2006 are used in this project because of the time limit.

The debts under consideration are from 1/7/2004 to 28/2/2006 for debt file. The data from 1/1/2006 to 31/3/2006 in transaction database are considered. The customer data is a snapshot of Newsnap file on 30/6/2005. To compute the history summary, eg., address change times, the summary of the previous financial year is used.

All the three databases are internal databases stored in IBM mainframe, and can be accessed via SAS. The data related to NSA will be extracted from the above three databases and then loaded into Teradata warehouse.

Personal ID will be used to link the three databases together. From Newsnap, those whose benefit type is NSA

Table 1. Demographic data model

Fields	Notes
Customer current circumstances	These fields are from the current customer circumstances in customer data, which are personal id, indigenous code, medical condition, sex, age, birth country, migration status, education level, postcode, language, rent type, method of payment, etc.
Aggregation of debts	These fields are derived from debt data by aggregating the data in the past financial year (from 1/7/2004 to 30/06/2005), which are debt indicator, the number of debts, the sum of debt amount, the sum of debt duration, the percentage of a certain kind of debt reason, etc.
Aggregation of history circumstances	These fields are derived from customer data by aggregating the data in the past financial year (from 1/7/2004 to 30/06/2005), which are the number of address changes, the number of marital status changes, the sum of income, etc.

Table 2. Summary of demographic data

Time frame	# of NSA customers	# of NSA debtors*
Snapshot on 30/06/2005	498,597	189,663

* Number of NSA customers on 30/06/2005 who had NSA debts during the previous financial year (from 1/7/2004 to 30/06/2005)

will be extracted. With those personal IDs from NSA data, those related transactions and debt records will be extracted from transaction database and debt database, respectively.

3 Discovering Demographic Patterns of Debtors

This section will present the construction of a customer demographic data model based on customer data and debt data, the data mining model and the results.

3.1 Building Demographic Data Model

This data model is to organise customer circumstances data and debt information into one table (see **Table 1**), based on which the characteristics of debtors and non-debtors will be discovered. In this data model, each customer has one record, which shows its latest or aggregated information of customer circumstances and debt. There are three kinds of attributes in this data model: customer current circumstances, the aggregation of debts, and the aggregation of customer history circumstances (say, the number of address changes), which are shown in **Table 1**. Debt indicator is defined as a binary attribute which indicates whether a customer had any debts in the previous financial year. The summary of the built demographic data model is given in **Table 2**.

3.2 Feature Selection

There are over 80 features in the constructed demographic data model, which is too much for available data mining software due to the huge search space. The following methods were used to select features.

- Correlation: the correlation between variables and debt indicator (see **Table 3**),
- Chi-square: the contingency difference of variables to debt indicators (see **Table 4**),
- Data exploration based on statistics: the impact difference of a variable on debtors and non-debtors.

Chi-square analysis is used to find the relationship between debt indicator and customer circumstances. It is implemented in module “Test based on Contingency Tables” with TWM. In those modules, debt indicator is set as first column, while customer circumstances variables are set as second columns. The statistical test style is set to “Chi Square”.

Based on correlation, chi-square test and data exploration, 15 features, such as ADDRESS_CHANGE_TIMES, RENT_AMOUNT, RENT_TYPE, CUSTOMER_SERVICE_CENTRE_CHANGE_TIMES and AGE, are selected as input for the following decision tree and association rule analysis.

3.3 Decision Tree Mining on Demographic Data

Based on the above feature selection, decision tree is used to build a classification model for debtors/non-debtors. It is implemented in TWM module “Decision Tree”. In those modules, debt indicator is set to dependent column, while customer circumstances variables are set as independent columns. The best result we got is a tree of 676 nodes, and its accuracy is shown in **Table 5**, where “0” and “1” stand for “no debt” and “debt”, respectively. The accuracy is poor (63.71%), and the error of false negative is high (30.53%). It is difficult to further improve the accuracy of decision tree on the whole population, however, some leaves of higher accuracy can be discovered by focusing on smaller groups. We found that the current version of

Table 3. Correlation between debt indicator and customer circumstances

Attributes	Correlation
CUSTOMER_SERVICE_CENTRE_CHANGE_IND	0.143
ADDRESS_CHANGE_TIMES	0.139
CUSTOMER_SERVICE_CENTRE _CHANGE_TIMES	0.135
ADDRESS_CHANGE_IND	0.129
RENT_AMOUNT	0.090
MARITAL_CHANGE_TIMES	0.085
MARITAL_CHANGE_IND	0.083
RA_ENTITLEMENT_AMOUNT	0.058
AGE	-0.092
LODGEMENT_FREQUENCY	-0.098

Table 4. Result of chi-square analysis for customer circumstances and debt indicator

Attributes	Chi-square
CUSTOMER_SERVICE_CENTRE _CHANGE_TIMES	13026
ADDRESS_CHANGE_TIMES	10889
LODGEMENT_FREQUENCY	6057
RENT_TYPE	4276
AGE	3940
RENT_AMOUNT	3903
MARITAL_CHANGE_TIMES	3745
RA_RATE_EXPLANATION	3424
RA_PRECLUSION_A13SON	3043
ACCOMMODATION	2924

Table 5. Confusion matrix of decision tree results

	Actual 0	Actual 1
Predicted 0	280,200 (56.20%)	152,229 (30.53%)
Predicted 1	28,734 (5.76%)	37,434 (7.51%)

Teradata warehouse miner cannot output leaves above a given accuracy threshold automatically and that it is difficult to navigate through the huge tree manually, we then turned to association rule mining to discover interesting patterns with higher accuracy on small groups of population.

3.4 Association Rule Mining on Demographic Data

Association mining (Agrawal, Imielinski, and Swami 1993) is used to find frequent customer circumstances patterns that are highly associated with debt or non-debt. It is implemented with “Association” module of TWM. In the module, personal ID is set as group column, while item_code is set as item column, where item_code is derived from customer circumstances and their values. In order to apply association rule analysis to our customer data, we regard each value of each feature as an item. Taking feature DEBT_IND as example, it has 2 values, which are DEBT_IND_0 and DEBT_IND_1. So DEBT_IND_0 is regarded as an item and DEBT_IND_1 is regarded as another item.

Table 6. Results of association rule mining

Association Rule	Support	Confidence	Lift
RA_RATE_EXPLANATION=P and age 21 to 28 =>debt	0.003	0.65	1.69
MARITAL_CHANGE_TIMES =2 and age 21 to 28 =>debt	0.004	0.60	1.57
age 21 to 28 and PARTNER_CASUAL_INCOME_SUM>0 and rent amount ranging from $200 to $400 => debt	0.003	0.65	1.70
MARITAL_CHANGE_TIMES =1 and PARTNER_CASUAL_INCOME_SUM>0 and HOME_OWNERSHIP=NHO => debt	0.004	0.65	1.69
age 21 to 28 and BAS_RATE_EXPLAN=PO and MARITAL_CHANGE_TIMES=1 and rent amount in $200 to $400 =>debt	0.003	0.65	1.71
CURRENT_OCCUPATION_STATUS=CDP => no debt	0.017	0.827	1.34
CURRENT_OCCUPATION_STATUS=CDP and SEX=male => no debt	0.013	0.851	1.38
HOME_OWNERSHIP=HOM and CUSTOMER_SERVICE_CENTRE_CHANGE_TIMES =0 and REGU_PAY_AMOUNT in $400 to $800 => no debt	0.011	0.810	1.31

Due to the limitation of spool space, we cannot run association rule analysis on the whole customer data. Therefore, we conduct association rule analysis on a 10% sample of the original data, and the discovered rules are then tested on the whole customer data. We select the top 15 features to run association rule analysis with minimum support as 0.003, and selected results are shown in **Table 6**. For example, the first rule in **Table 6** shows that 65% out of those customers with RA_RATE_EXPLANATION as "P" (Partnered) and aged from 21 to 28 have had debts in the previous financial year, and the lift of the rule is 1.69.

4 Discovering Activity Sequence Patterns of Debtors

The previous section describes our work on "static" demographic data, which tries to find the demographic patterns of debtors. In addition to the demographic characteristics, a debtor may have interesting activity patterns which show the interactions between him/her and Centrelink. This section will present the work on mining "dynamic" activity data, which is composed of customer activities sequences from 1/1/2006 to 31/3/2006.

An activity represents a unit of work. It records: 1) which customer is the activity related to (e.g. personal id), 2) the type of actions (e.g. activity code), 3) the date and the time the actions were done, 4) the reason for the actions (e.g. receipt of source documents and manual notes), etc. An activity can be caused by new claims or circumstance changes. It can also be caused by transaction of somebody else's. For example, when a customer is granted Newstart, an activity will be created to check the partner's record to ensure that the Family Tax Benefit is paid correctly. An activity can also be generated automatically by the system or manually registered. For example, in cases where maintenance action is in progress, automatic activities are generated to ensure the progress is reviewed.

An activity sequence is a series of actions for a customer, and mining such sequences may help to discover which sequences are highly associated with debts and which are not. In the following an activity sequence data model, data mining models and the results will be presented.

4.1 Building Activity Sequence Data Model

This data model organises activity data into baskets or sequences for association and sequence analysis and two

different kinds of baskets/sequences are built respectively for debt and non-debt. According to domain expert's opinion and the frequency of activity, the following strategy is used to build the baskets/sequences. For debt baskets/sequences, we look back one month before each debt to build up the basket/sequence for analyzing the activities. That is, if there is a debt of a customer, those activities of the customer happened in 30 days immediately before the debt are put into one basket. For debt basket, those debts beginning in January 2006 and their associated activities are excluded because the available activities before them are less than one month. For non-debt basket, those activities in January and February 2006 can be taken as a basket for a customer having no debts in the first 3 months of 2006. Those activities from 16/1/2006 to 15/2/2006 are used to build baskets for non-debt, because there are too few activities in the beginning days of the year.

An activity sequence is built for each debt of a person, and personal ID and debt ID are concatenated as the identifier of the sequence. This is to make the data ready for both association mining and sequence mining with TWM. The following are two examples of debt activity sequence and non-debt activity sequence, where A_i stands for an activity, and D and N denote "debt" and "no debt", respectively.

Debt activity sequence: A_1, A_2, A_3, A_5, A_6, A_7, D.

Non-debt activity sequence: A_4, A_6, A_3, A_2, A_5, A_9, A_1, N.

Table 7 summarizes the results of constructing debt and non-debt activity baskets/sequences. Activities are also reorganised in terms of income and non-income, where debt baskets are only built for all income-related debts, and all other types of debts are ignored while non-debt baskets are constructed for all persons not having a debt. This is based on that income-related debt is one of the concerns of this project, and also because of the imbalance of debt and non-debt classes of activities.

4.2 Sequence Mining Models and Results

The following data mining approaches are used to discover activity sequence patterns of debtors. Some examples of the discovered results of sequence patterns will follow. Note that the activity codes are replaced with A_i for the consideration of business confidentiality.

4.2.1 Sequence Mining on Activity Sequences on Mixed Data

Because the two classes, debt and non-debt, are highly unbalanced, it is difficult to find meaningful patterns from such data, and the measures of confidence and lift are skewed. Therefore, we draw a sample from non-debt data, which is of the same size as that of debt data, and conducted sequence mining on the sampled data. Then sequence mining was employed to discover those activity sequences co-occurring frequently with debt or non-debt. The support threshold is set to 0.01 and some mined results are given in **Table 8**.

4.2.2 Sequence Mining on Activity Sequences on Separated Data

To avoid the undesirable effect of unbalanced class sizes, we mined separately on debt activity sequences and non-debt activity sequences. The difference from mining on the mixed data is that this task discovers frequent activity patterns without worrying about debt or non-debt and that only support is computed for these patterns. Note that the support here is local support (Chen et al 2005), that is, the support of the pattern on debt data or non-debt data, not on the whole dataset. The support threshold is also set to 0.01.

4.2.3 Discovering Dual-Target and Contrast-Target Activity Patterns

For those activity patterns discovered on separated data, we look for those patterns which are frequent in both datasets, and those patterns which are frequent in one dataset but infrequent in the other. The former is named as *dual-target pattern*, while the latter is as *contrast-target pattern*. The way to find dual-target patterns is simply selecting those common frequent patterns in both debtor dataset and non-debtor dataset. The method to discover contrast-target patterns is computing the ratio of the supports of a pattern on the two datasets and then selecting those patterns with ratios much different from one.

Dual-target activity patterns given in **Table 9** are those frequent activity patterns identified in both debt group and non-debt group. For these activities and patterns, the local support indicates how frequently the activity/pattern occurs in each group. Activity A13 occurs in 86.3% of the debt records and in 69.4% of the non-debt records. For activity A17, a debt is raised within 30 days after A17 for 25.5% activity sequences, and no debt is raised within 30 days of activity A17 in 17.8% of activity sequences with A17. In this case, probably there are some other activities that make the sequences with A17 lead to different results. Further investigation is needed to tell what makes the sequences result in debt or non-debt. Obviously more research needs to be conducted into these activities. The inference is that an activity or sequence occurring in both debt and non-debt groups and strongly associated with debt is associated with missing information in the non-debt group. The potential in this area is for activities/sequences strongly associated with debt to act as a trigger for some kind of information where they occur in the non-debt group.

Contrast-target activity patterns are those activity associations or sequences much more frequent in debt group (or non-debt group) than in non-debt group (or debt group), which are shown in **Table 10**. For example, the first pattern "A13, A19" is 1.6 times more associated with debt than with no debt. For the activities of a customer, if A13 happens first and then A19 occurs, the customer should be checked carefully in case he may have a debt in the near future, or the activity sequences with "A13, A19" should be checked to make sure what can be done to prevent the occurrence of debt. These sequences, if confirmed by further research, can act as markers of debt or potential debt. Where they occur they could be a clear

trigger for targeted intervention strategies to prevent or minimize debt.

4.2.4 Discovering Reverse-Target Activity Patterns

For *reverse-target activity patterns*, we look for those frequent pattern pairs like "P=>debt" and "PQ=>no debt", or "P=>no debt" and "PQ=>debt", where P and Q are activities or activity sequences. In both cases, the occurrence of Q has a significant impact on the result, by changing the result to its opposite. This kind of patterns help to find which activity or sequence Q has significant impact on the result. For frequent pattern pairs like "P=>debt" and "PQ=>no debt", when activity sequence P happened, then activities in Q are suggested to conduct to reduce or prevent debt. For frequent pattern pairs like "P=>debt" and "PQ=>no debt", activities in Q are suggested not to conduct to reduce or prevent debt.

To measure the interestingness of the above patterns, we designed a measure of "impact" for pattern pair "P=>result1" and "PQ=>result2"as follows (see (Cao, Zhao, and Zhang 2006) for details).

$$\text{Impact} = \frac{Sup_2 / Sup_4}{Sup_3 / Sup_1},$$

where Sup_1 is the local support of an underlying pattern, e.g., "P=>result1", Sup_2 is the local support of a derivative pattern, e.g., "PQ=>result2", Sup_3 is the local support of the rule ("P=>result2") contrary to the underlying pattern, and Sup_4 is the local support of the rule ("PQ=>result1") contrary to the derivative pattern. Impact should be greater than 1.0 for useful patterns. The larger impact is, the more interesting is the pattern pair.

The top-ranking reverse-target activity patterns are given in **Table 11**. It is found that A1, A22, A23, A20 are not related to debts in our data, while A13, A14 or A15 is more likely to be associated with debt than non-debt. However, when A1, A22, A23 or A20 follow A13, A14 or A15, the composite patterns have higher likelihood of resulting in non-debt rather than debt.

4.2.5 Pruning Redundant Patterns

There are many redundant patterns in the discovered rules. Assume that "A=>C" and "AB=>C" are two rules of the same confidence, then "AB=>C" is redundant because it provides no more information given "A=>C" is known. The method is to remove those patterns if there are any shorter sub-patterns having the same or roughly the same confidence. For contrast-target patterns, those patterns whose support ratios are the same or less than the support ratios of their sub-patterns are removed. For reverse-target patterns, those pattern pairs whose impacts are the same or less than the impacts of their sub-patterns are removed. For more details on pruning redundant patterns, please refer to (Zaki 2004).

5 Concluding Remarks

Data mining techniques have been used to discover the debtor patterns of Centrelink customers. Two data models, customer demographic data model and activity sequence data model, were first built and then techniques of decision tree and sequence mining were employed to mine the demographic data and activity data. The discovered patterns maybe used to find those customer groups with high probability of debts, so that reviews on those customers can be conducted or letters can be mailed to them to help reduce debts. Moreover, by discovering activity sequence patterns associated with debt/non-debt, appropriate actions can be suggested for next step under a given situation to reduce the probability of leading to debt. This is one of our efforts to solve real-world business problems with advanced data mining techniques, and it shows promising applications of data mining to solve real-life problems in near future.

However, there are still many open problems to be solved. First, there are still hundreds or even thousands of discovered rules after redundant patterns have been pruned. How can interesting patterns be efficiently selected from them? Second, most rules obtained with existing statistical measures of interestingness are not interesting at all from business perspective, and many business interesting rules may be pruned during data mining procedure, so post-mining rules pruning helps little. The use of domain knowledge when mining can not only help to find "business interesting" patterns, but also help to reduce the search space and running time of data mining algorithms. How can domain knowledge be effectively incorporated in data mining procedure? Third, the business data is complicated and the customers and customer debts/activities are linked to many other customers, such as spouse, dependents and tenants. How can existing data mining approaches be improved to discover more useful patterns by utilizing those additional information? For example, an activity A1 of a customer C1 may lead to an activity A2 of his/her spouse C2, and A2 may activate activity A3 of a third customer C3. How can existing approaches for sequence mining be improved to take into consideration the linkage and interaction between activity sequences of different customers? Last and the most important, how to use these discovered rules to help predict and prevent debt? How to build an efficient debt prevention system to effectively detect debt in advance and give appropriate suggestions to help reduce or prevent debt? How to evaluate the risk of debt when an action is taken? All the above problems remain open and will be part of our future work.

6 Acknowledgments

The authors would like to thank Dr. Jie Chen, Mr. Peter Newbigin, Mr. Fernando Figueiredo, Mr. Rick Schurmann and Mr. Mark Norrie for their work and support. This work was supported in part by the Australian Research Council (ARC) Discovery Projects (DP0449535 and DP0667060), National Science Foundation of China (NSFC) (60496327) and Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01).

Table 7. Summary of activity sequences

	# of activities	# of sequences	# of debt-related sequences
All debts	6,063,703	454,934	16,540 (3.6%)
Income debts	5,770,523	439,953	1,559 (0.35%)

Table 8. Results of activity sequence mining

Sequence Rule	Support	Confidence	Lift
A1 => no debt	0.0529	0.753	1.51
A2 => debt	0.0173	0.831	1.66
A3 => debt	0.0712	0.716	1.43
A4, A3 => debt	0.0157	0.845	1.69
A5, A3 => debt	0.0144	0.833	1.67
A6, A3 => debt	0.0141	0.815	1.63
A7, A4 => debt	0.0141	0.759	1.52
A8, A7 => debt	0.0388	0.738	1.48
A8, A8 => debt	0.0260	0.704	1.41
A8, A6 => debt	0.0173	0.692	1.38
A9, A10 => debt	0.0154	0.686	1.37
A11, A9 => debt	0.0138	0.682	1.36
A12, A9 => debt	0.0157	0.681	1.36
A8, A4 => debt	0.0209	0.677	1.35
A8, A8, A7 => debt	0.0138	0.860	1.72

Table 9. Dual-target activity sequence patterns

Activity	Local support of “activity=>debt”	Local support of “activity=>no debt”
A13	0.863	0.694
A14	0.845	0.601
A15	0.711	0.569
A16	0.322	0.257
A4	0.286	0.155
A17	0.255	0.178
A6	0.248	0.203
A8	0.226	0.155
A5	0.164	0.120
A12	0.162	0.138
A18	0.133	0.128

Table 10. Contrast-target activity patterns

Sequence patterns	DSUP: support of "pattern=>debt"	NSUP: support of "pattern =>no debt"	DSUP/NSUP
A3	0.142	0.053	2.7
A15, A3	0.094	0.029	3.2
A13, A19	0.033	0.013	2.6
A14, A19	0.028	0.011	2.6
A8, A4	0.042	0.017	2.6
A19, A14	0.026	0.010	2.4
A14, A4, A4	0.042	0.015	2.9
A4, A4, A14	0.038	0.014	2.7
A8, A7, A13	0.051	0.018	2.7
A8, A7, A14	0.055	0.020	2.7
A8, A8, A7	0.028	0.010	2.7
A14, A4, A14	0.126	0.047	2.7
A13, A4, A5	0.028	0.011	2.6
A20	0.026	0.035	0.8
A1	0.035	0.093	0.4

Table 11. Reverse-target activity patterns

Sequence patterns	Impact	SUP1, e.g., support of A1=> no debt	SUP2, e.g., support of A1,A14 =>debt	SUP3, e.g., support of A1=>debt	SUP4, e.g. Support of A1,A14=>no debt
A1 => no debt A1, A14=> debt	3.73	0.093	0.017	0.035	0.013
A1 => no debt A1, A15 => debt	2.93	0.093	0.019	0.035	0.017
A13 => debt A13, A1 => no debt	2.44	0.863	0.024	0.694	0.012
A15, A14 => debt A15, A14, A22 => no debt	2.00	0.527	0.015	0.340	0.012
A14, A15 => debt A14, A15, A22 => no debt	1.87	0.499	0.017	0.344	0.014
A13, A17 => debt A13, A17, A18 => no debt	1.64	0.166	0.012	0.108	0.012
A13, A17 => debt A13, A17, A21 => no debt	1.58	0.166	0.011	0.108	0.010

7 References

Agrawal, R., Imielinski, T. and Swami, A. (1993): Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington DC, USA, 22:207-216, ACM Press.

Cao, L., Zhao, Y. and Zhang, C. (2006): Mining Impact-Targeted Activity Patterns in Unbalanced Data, submitted to IEEE TKDE special issue on Intelligence and Security Informatics, 30 April 2006.

Centrelink (2005): Centrelink Annual Report 2004-2005.

Centrelink (2006): Centrelink Fraud Statistics and Centrelink Facts and Figures, http://www.centrelink.gov.au/internet/internet.nsf/about_us/fraud_stats.htm, http://www.centrelink.gov.au/internet/internet.nsf/about_us/facts.htm.

Chen, J., He, H., Li, J., Jin, H., McAullay, D., Williams, G., Sparks, R. and Kelman C. (2005): Representing Association Classification Rules Mined from Health Data. *Proc. of 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2005)*, Melbourne, Australia, September 14-16, 2005, pp. 1225-1231.

Teradata (2005): Teradata Warehouse Miner User's Guide - Release 04.01.00, 2005.

Zaki, M. (2004): Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223-248, 2004.

What Types of Events Provide the Strongest Evidence that the Stock Market is Affected by Company Specific News?

Calum Robertson
School of Software Engineering and Data Communications
Faculty of Information Technology
Queensland University of Technology
Level 7, S Block, 2 George Street
Brisbane, Queensland Australia 4000
cs.robertson@qut.edu.au

Shlomo Geva
School of Software Engineering and Data Communications
Faculty of Information Technology
Queensland University of Technology
Level 7, S Block, 2 George Street
Brisbane, Queensland Australia 4000
s.geva@qut.edu.au

Rodney Wolff
School of Economics and Finance
Faculty of Business
Queensland University of Technology
Level 8, Z Block, 2 George Street
Brisbane, Queensland Australia 4000
r.wolff@qut.edu.au

Abstract

The efficient market hypothesis states that an efficient market immediately incorporates all available information into the price of the traded entity. It is well established that the stock market is not an efficient market as it consists of numerous traders with differing strategies and interpretations of information. However there is substantial evidence to suggest that the stock market does incorporate new information into prices. Unfortunately little research has focussed on the high frequency effect of real time news, across a broad base of assets. This paper investigates how the US, UK, and Australian markets incorporate all real time news, not just Press Announcements, Annual Reports, etc. We find that there is strong evidence to suggest that the markets do incorporate news quickly.

Keywords. Stock Market, News, Return, Volatility, Market Reaction.

1. Introduction

A plethora of research is available which shows that the occurrence of news does effect the market, with the majority of research focusing on macroeconomic news, which provides an indication of the state of the economy (Almeida, Goodhart and Payne 1998, Bomfim 2003, Brannas and De Gooijer 2004, Ederington and Lee 1993, 1995, 2001, Ewing 2002, Graham, Nikkinen and Sahlstrom 2003, Han and Ozocak 2002, Hess 2004, Kim 1998, 2003, Kim, McKenzie and Faff 2004, Nikkinen and Sahlstrom 2004a, 2004b, Nofsinger and Prucyk 2003, Simpson and Ramchander 2004, Sun and Sutcliffe 2003, Tse 1999). However macroeconomic news is relatively infrequent compared to asset specific information, e.g. Simpson and Ramchander (2004) state that the United States of America releases 23 macroeconomic reports regularly, usually monthly, whilst Fung, Yu and Wai (2003) found an average over 373 news articles per asset per month. Not only is asset specific information more frequent but it has been shown to have a noticeable effect on the given asset (Chan 2003, Donders and Vorst 1996, Dungey, Fry and Martin 2004, Fung, Yu and Wai 2003, Goodhart 1989, Goodhart and Figliuoli 1992, Goodhart, Hall, Henry and Pesaran 1993, Hong, Lim and Stein 2000, Melvin and Yin 2000, Michaely and Womack 1999, Mitchell and Mulherin 1994, Mittermayer 2004, Roll 1984, Womack 1996, Wuthrich, Permunetilleke, Leung, Cho, Zhang and Lam 1998).

Previous research has shown that the market reacts quickly to macroeconomic news (Ederington and Lee 1993, 1995, 2001, Han and Ozocak 2002, Nofsinger and Prucyk 2003). However as macroeconomic news is scheduled (i.e. the market is aware exactly when the news is released) the market anticipates the content of news and can react quickly based on whether the actual news matches analyst forecasts. It is not known how rapidly the stock market reacts to non-macroeconomic news and therefore it would be interesting to determine if the market responds in a timely manner, if at all, to non-macroeconomic news.

Most research to date which investigates the intraday effect (reaction of the market on the day which the news was released) of news has focussed on the Foreign Exchange markets (Almeida, Goodhart and Payne 1998, Bollerslev and Domowitz 1993, Ederington and Lee 1993, 1995, 2001, Goodhart 1989, Goodhart and Figliuoli 1991, 1992, Goodhart, Hall, Henry and Pesaran 1993, Han and Ozocak 2002, Melvin and Yin 2000, Peiers 1997), or Futures markets (Hess 2004, Tse 1999), whilst little has focussed on the Stock market (Mittermayer 2004, Nofsinger and Prucyk 2003). Nofsinger and Prucyk (2003) investigated the intraday effect of macroeconomic news on the S&P 100 Index Option and found that bad news with high information surprise is responsible for most abnormal volume associated with macroeconomic news. Mittermayer (2004) investigated the effect of Press Announcements on stocks on the New York Stock Exchange and NASDAQ and found evidence to suggest that the market does react to the news, and furthermore the content of the news which triggered the reaction. However Press Announcements are relatively rare compared to other types of news available from real time news providers, so it bears further investigation.

The purpose of this paper is to identify events which occur with a high correlation to the occurrence of real time news, such that they can be used to identify "interesting" news articles. Furthermore this paper aims to investigate the percentage of news articles which the

market appears to find significant, as the sheer volume of news available would prohibit an individual from reading all available news.

2. Data

All data for this research was obtained using the Bloomberg Professional® service. The dataset consists of stocks which were in the S&P 100, FTSE 100, and ASX 100 indices as at the 1st of July 2005 and continued to trade through to the 1st of September 2006, which is a total of 286 stocks. For each stock the Trading Data, and News were collected for the period beginning 1st of May 2005 through to and including the 31stof August 2006.

The set defined in Eq. (1) consists of each distinct minute where trading occurred for the stock (s), within all minutes for the period of data collection (T_A). For each minute the average price, the volume, and the number of ticks (number of trades) for trades during that minute are also stored. However we are only interested in the business time scale (minutes which occurred during business hours for the market on which the stock trades). Furthermore we want a heterogeneous time series (i.e. an entry for every business trading minute for the stock, regardless of whether any trading occurred). Therefore we produce the date, price, volume, and tick time series for all minutes in the business time scale (T_B) with the definitions in Eqs. (2)-(5). We define the price at time t as the price of the last actual trade for the stock prior to or at the given time. We set the volume and ticks equal to 0 if there wasn't a trade at time t. Note that if the stock was suspended from trading for a whole day then the day is excluded from T_B.

$$I_{(s)} = \{I_1, I_2, \ldots, I_m\} \mid I_{(s,z)} = (d_{(s,z)}, p_{(s,z)}, v_{(s,z)}, k_{(s,z)}) \wedge z \in \mathrm{T}_A \quad (1)$$

$$D_{(s)} = \{D_1, D_2, \ldots, D_n\} \mid D_{(s,t)} > D_{(s,t-1)} \wedge D_{(s,t)} \in \mathrm{T}_B \ \wedge \ \mathrm{T}_B \subseteq \mathrm{T}_A \quad (2)$$

$$P_{(s)} = \{P_1, P_2, \ldots, P_n\} \mid P_{(s,t)} = (p_{(s,z)} \mid z = \max(z \mid d_{(s,z)} \le D_{(s,t)})) \quad (3)$$

$$V_{(s)} = \{V_1, V_2, \ldots, V_n\} \mid V_{(s,t)} = (\exists! d_{(s,z)} = D_{(s,t)} \ ? \ v_{(s,z)} \ : \ 0) \quad (4)$$

$$T_{(s)} = \{T_1, T_2, \ldots, T_n\} \mid T_{(s,t)} = (\exists! d_{(s,z)} = D_{(s,t)} \ ? \ k_{(s,z)} \ : \ 0) \quad (5)$$

The news search facility within the Bloomberg Professional® service was used to download all relevant articles for each stock within the dataset. These articles include Press Announcements, Annual Reports, Analyst Recommendations and general news which Bloomberg has sourced from over 200 different news providers. The set defined in Eq. (6) consists of each distinct news article for the stock and contains the time and content of the article. However we are only interested in the business time scale and are only concerned whether news occurred at the given time. Therefore we produce the news time series defined in Eq. (7) such that each business trading minute for the stock contains the count of the articles which occurred during it. If an article occurs after hours then it is stored in the first trading minute of the next trading day, as defined in Eq. (7).

$$A_{(s)} = \{A_1, A_2, \ldots, A_p\} \mid A_{(o)} = (d_o, c_o) \ \wedge \ s \in \mathrm{T}_A \quad (6)$$

$$N_{(s)} = \{N_1, N_2, \ldots, N_q\} \mid N_{(s,t)} = \prod \{\forall A_{(s)} \mid D_{(t-1)} < d_o \le D_{(t)}\} \quad (7)$$

3. Methodology

The return time series for a stock (s), defined in Eq. (8), is formed by taking the difference in the log prices from the trading data defined in Eq. (3) over the period Δt. The return time series identifies periods of high return which may indicate that the market is reacting to news.

The change in volume time series for a stock (s), defined in Eq. (9), is formed by taking the difference in the log of the average volume, over n minutes, between the time t and t-Δt. The volume defined in Eq. (4) is averaged over n minutes to limit the effect of trading minutes where no trade occurred. The conditional log of the average volume is used such that in the case where no trade occurs during the given n minutes, the function still produces an answer. The change in volume time series detects periods where there is a sudden increase in the volume of the stock traded, which might suggest that the market is reacting to news.

$$R_{(s,\Delta t)} = \{R_1, \ldots, R_m\} \mid R_{(s,t,n,\Delta t)} = \log(P_{(s,\ t)}) - \log(P_{(s,\ t-\Delta t)}) \quad (8)$$

$$CV_{(s,n,\Delta t)} = \{CV_1, \ldots, CV_m\} \mid CV_{(s,t,n,\Delta t)} = L(V_{(s,t,n)}) - L(V_{(s,t-\Delta t,n)}) \quad (9)$$
$$\wedge \ V_{(s,t,n)} = \frac{1}{n}\sum_{i=0}^{n-1} V_{(s,t-i)} \ \wedge \ L(x) = ((x > 0) \ ? \ \log(x) \ : \ \log(0.5))$$

$$CT_{(s,n,\Delta t)} = \{CT_1, \ldots, CT_m\} \mid CT_{(s,t,n,\Delta t)} = L(T_{(s,t,n)}) - L(T_{(s,t-\Delta t,n)}) \quad (10)$$
$$\wedge \ T_{(s,t,n)} = \frac{1}{n}\sum_{i=0}^{n-1} T_{(s,t-i)} \ \wedge \ L(x) = ((x > 0) \ ? \ \log(x) \ : \ \log(0.5))$$

$$\upsilon_{(s,n,\Delta t)} = \{\upsilon_1, \ldots, \upsilon_m\} \mid \upsilon_{(s,t,n,\Delta t)} = \sqrt{\frac{y}{\Delta t}} \times \left[\frac{1}{n}\sum_{i=0}^{n-1} \left(R_{(s,t-i,n,\Delta t)} - M_{(s,t,n,\Delta t)}\right)^p\right]^{\frac{1}{p}} \quad (11)$$
$$\wedge \ M_{(s,t,n,\Delta t)} = \frac{1}{n}\sum_{i=0}^{n-1} R_{(s,t-i,\Delta t)} \ \wedge \ y = 250 \times \begin{pmatrix} 390 \mid s \in US \\ \vee \ 510 \mid s \in UK \\ \vee \ 360 \mid s \in AU \end{pmatrix}$$

The change in ticks time series for a stock (s), defined in Eq. (10), is formed by taking the difference in the log of the average ticks, over n minutes, between the time t and t-Δt. The ticks defined in Eq. (5) are averaged, and the conditional log is used for the same reasons as for the volume. The change in ticks time series pinpoints periods where there is a sudden increase in the number of trades, which implies that the market is reacting to news.

The volatility time series for a stock (s), defined in Eq. (11), calculates the annualised volatility of the stock. When p is set to 2 it is simply the annualised variance of the return of the stock. The y value is calculated by multiplying the number of trading days per year (generally set to 250), by the number of trading minutes for the day (390 minutes for the US, 510 for the UK, and 360 for Australia). The volatility is annualised such that the results of each country can be directly compared, as suggested by Dacorogna, Gencay, Muller, Olsen and Pictet (2001). The volatility time series discovers points where the stock price changes rapidly, which could insinuate that the market is reacting to news.

The stocks are grouped together as per Eq. (12) so that we can examine the effect of news on an individual country. We divide the trading day into equally sized time windows ΔT, as defined in Eq. (13), in order to examine the intraday effect of the news. Note that the first period is ignored because we don't want the after hours news and market behaviour to skew results. The first period is the larger of ΔT and Δt.

We define a generalised time series F, where F is one of the return, volume, tick, and volatility time series. We use the generalised time series to define the Event Point Process (EPP) in Eq. (14). In this point process a point value of 1 indicates that the generalised time series for the given stock exceeded the specified threshold (x), which we will refer to as an event. It should be noted that the return, volume and tick time series are log values about 0, so a threshold of 10% means that the value should be ≥ to log(11/10) or ≤ to log(10/11). The volatility time series is always positive when p is even so the ≤ condition is ignored.

$$S = \{S_1, S_2, \ldots, S_h\} \mid h \geq 1 \tag{12}$$

$$W = \{W_1, W_2, \ldots, W_g\} \mid W_{(w,\Delta t,\Delta T)} = \{\forall T_B \mid t_0 \leq time(T_{(B,\,i)}) < t_0 + \Delta T \wedge t_0 = t_f + (w-1)\times\Delta T\} \wedge t_f = \min(time(T_B)) + \max(\Delta T, \Delta t) \tag{13}$$

$$E_{(w,\Delta T,S,n,\Delta t,x)} = \{E_1, \ldots, E_m\} \mid E_{(w,\Delta T,s,t,n,\Delta t,x)} = \left(F_{(s,t,n,\Delta t)}\begin{pmatrix}\geq x\\ \vee \leq -x\end{pmatrix}\right) ? 1 : 0 \wedge s \in S \wedge d_{(s,t)} \in W_{(w,\Delta t,\Delta T)} \tag{14}$$

$$EN_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)} = \{EN_1, \ldots, EN_m\} \mid EN_{(w,\Delta T,s,t,n,\Delta t,x,\Delta\tau)} = \left(\left(E_{(w,\Delta T,s,t,n,\Delta t,x)} = 1\right) \wedge \left(\prod_{i=1}^{\Delta\tau} A_{(s,\,t-i)} > 0\right)\right) ? 1 : 0 \tag{15}$$

$$E\overline{N}_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)} = \{E\overline{N}_1, \ldots, E\overline{N}_m\} \mid E\overline{N}_{(w,\Delta T,s,t,n,\Delta t,x,\Delta\tau)} = \left(\left(E_{(w,\Delta T,s,t,n,\Delta t,x)} = 1\right) \wedge \left(\prod_{i=1}^{\Delta\tau} A_{(s,\,t-i)} = 0\right)\right) ? 1 : 0 \tag{16}$$

The Event given News Point Process (ENPP) is defined in Eq. (15), where a point value of 1 symbolises that an event occurred at time t for the stock s in the EPP and at least one company specific news article arrived between t-Δτ and t-1. The Event Without news Point Process (EWPP) defined in Eq. (16) has a point value of 1 when an event occurred at time t for the stock s in the EPP and no company specific news arrived between t-Δτ and t-1.

The Ratio of Events Related to News to Events (RERNE) defined in Eq. (17) indicates the percentage of events which are preceded by news. A high RERNE value suggests that most events for the given parameters are preceded by news, which would imply that news is responsible for these events. A low RERNE value can denote that the market takes longer than the specified Δτ time to react to news, or that the events are caused by other factors, or that the events themselves are merely noise.

The Benchmark defined in Eq. (18) provides a measure of the likelihood of news arriving within the specified Δτ time. This is achieved by calculating a return of 0% in Eq. (14), which produces a point process where every point has a value of 1, and therefore the point process in Eq. (15) simply indicates the points when news occurs within the specified Δτ time.

This Benchmark is then used to calculate the Likelihood that Events are Related to News (LERN) in Eq. (19). A high LERN value implies that it is more likely for news to occur prior to an event that it is normally. A LERN value equal to 100% indicates that it is just as likely for news to occur before an event as it is to occur at any other time. A low LERN value signifies that it is less likely for news to occur prior to an event than normal, which would imply that news isn't responsible for the event.

$$RERNE_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)} = \frac{\sum EN_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)}}{\sum E_{(w,\Delta T,S,n,\Delta t,x)}} \tag{17}$$

$$B_{(w,\Delta T,S,n,\Delta t,\Delta\tau)} = \frac{\sum}{\prod} EN_{(w,\Delta T,S,n,\Delta t,0,\Delta\tau)} \tag{18}$$

$$LERN_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)} = \frac{RERNE_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)}}{B_{(w,\Delta T,S,n,\Delta\tau)}} \tag{19}$$

The Event T-Test (ETT) defined in Eq. (20) performs a Student t-Test on the period distribution of the chance of an event occurring with news versus the chance of an event occurring without news during each period. The purpose of the ETT is to test the null hypothesis that the occurrence of events is not influenced by the occurrence of news.

The News T-Test (NTT) defined in Eq. (21) performs a Student t-Test on the period distribution of the chance of news occurring prior to an event versus the chance of news occurring during each period. The intent of the NTT is to test the null hypothesis that the occurrence of news before events is the same as the occurrence of news normally.

$$ETT_{(W,\Delta T,S,n,\Delta t,x,\Delta\tau)} = tTest\left(\left\{\forall \frac{\sum}{\prod} EN_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)} \mid w \in W\right\}, \left\{\forall \frac{\sum}{\prod} E\overline{N}_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)} \mid w \in W\right\}\right) \tag{20}$$

$$NTT_{(W,\Delta T,S,n,\Delta t,x,\Delta\tau)} = tTest\left(\left\{\forall \frac{\sum}{\prod} EN_{(w,\Delta T,S,n,\Delta t,x,\Delta\tau)} \mid w \in W\right\}, \left\{\forall B_{(w,\Delta T,S,n,\Delta t,\Delta\tau)} \mid w \in W\right\}\right) \tag{21}$$

4. Results

Table 1 shows some of the characteristics of the dataset. The number of trading minutes during standard business days, week, month, and year can be used by the reader to appreciate how frequently events occur in the Table 2.

The Average Minutes without a Trade gives an indication of how many minutes within normal trading hours have no trades for each country. Clearly trading on the US market is more frequent than on the others. Bearing in mind that the Australian market is far smaller than the other two, it shouldn't be a surprise that there is less activity than on the others.

The News Articles in Dataset gives the reader an idea of the frequency of news in the different markets.

Variable	US	UK	AU
Trading Minutes Per Business Day	390	510	360
Trading Minutes Per Typical Business Week	1,950	2,550	1,800
Trading Minutes Per Typical Business Month	8,125	10,625	7,500
Trading Minutes Per Typical Business Year	97,500	127,500	90,000
Average Minutes without a Trade (%)	2.36%	37.65%	50.68%
News Articles in Dataset	293,416	136,627	130,988
Average After Hours News Articles (%)	57.76%	43.22%	76.61%

Table 1. This table shows some characteristics of the dataset for each country to help the reader appreciate later results.

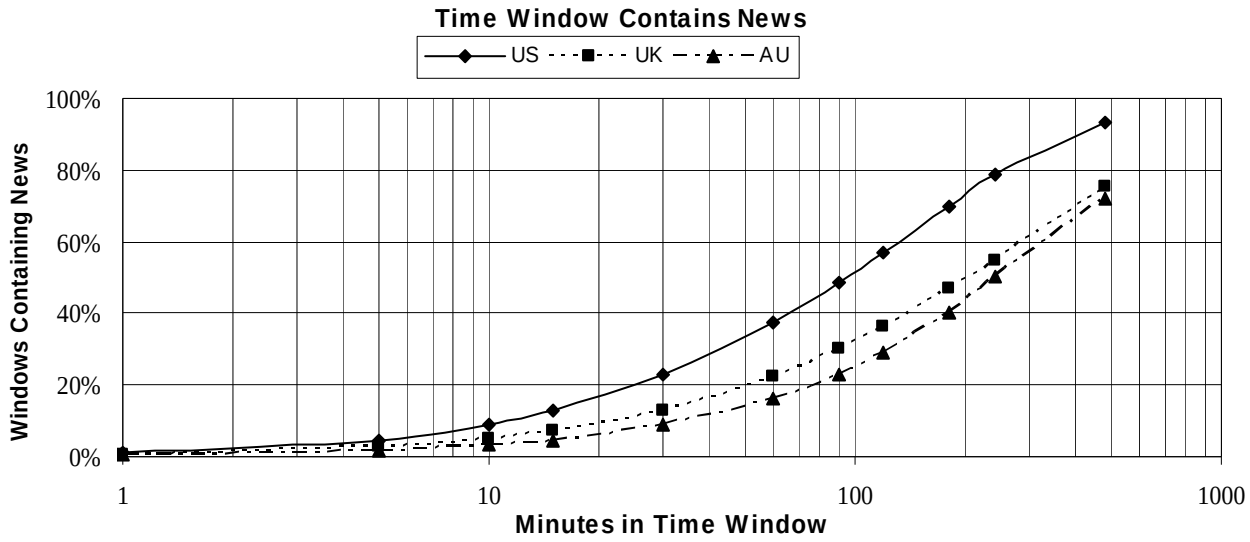


Fig. 1. As the size of the time window is increased the percentage of windows which contain news also does. Clearly there is more news distributed in the US market than the others.

Obviously the US receives far more information than the other markets, which means it should be easier to identify events linked to news articles within that market.

The Average After Hours News articles give the percentage of the articles which occur outside business hours in the different markets. The UK market has the lowest ratio but this could be a factor of longer trading hours than the other markets. The Australian market has the least business hours, and the highest after hours news ratio, though it is a far smaller market than the other two, so it shouldn't be a surprise that less information is made available during the trading day.

The choice of a time window $\Delta\tau$ in which news can be found is significant. If the percentage of minutes which contain a news article within the time window is too high it is difficult to establish whether the news was responsible for the event, or whether it was a coincidence. The Benchmark results using Eq. (18), shown in Fig. 1, identify the chance of finding news within the given time window. The figure shows that a reasonable amount of minutes are preceded by news within 60 minutes ($\Delta\tau$=60), without every minute being preceded by news.

It is logical to assume that if news causes the event then the reaction should be within the same time window as that for the news, and therefore the values $\mathbf{n=\Delta t=\Delta\tau=60}$ are used for all tests in this paper. Furthermore we use **p=2** for the volatility tests, which means that we are calculating the annualised variance. Finally we set $\mathbf{\Delta T=30}$ in order to analyse the intraday effect of news, which means that we ignore the first 60 minutes of the trading day. Further tests using different variations may yield more intriguing results but the aim of this paper is to identify the types of events which appear to be linked to news, rather than finding the ideal parameters.

The results in Table 2 show the average number of minutes between events during the first period (60 minutes), and the rest of the day. Return and volatility events are rarer during the rest of the day than during the first period, with return events a lot rarer. However volume and tick events tend to be rarer in the first period than during the rest of the day in the UK and Australia. This is probably due to the number of minutes without trades in these countries, which limits the average volume and number of ticks. Alternatively it could indicate that these markets exhibit a fairly steady rate of trade in the opening hour of business.

The results of the RERNE and LERN tests are shown in Table 3 where bolded LERN values highlight results where the value exceeds 100%. As the return threshold is increased to 5% there is an increase in both RERNE and LERN values for all countries. Only the US fails to show a further increase by the 10% threshold. Whilst the 10% return results have higher values for the UK and Australian markets than the 5% threshold it should be

Event Type	Threshold	First Period			Rest of Day		
		US	UK	AU	US	UK	AU
Return	0.1%	1.14	1.16	1.14	1.36	1.41	1.40
	0.2%	1.32	1.35	1.25	1.88	2.02	1.81
	0.5%	2.16	2.30	1.85	5.20	6.02	4.90
	1.0%	5.27	5.93	3.63	23.34	30.96	20.47
	2.0%	23.20	28.64	13.18	187.38	296.73	176.53
	5.0%	172.61	452.16	190.78	2,986.46	6,669.99	4,177.65
	10.0%	644.54	3,102.39	1,674.15	30,892.04	57,065.47	37,419.38
Volume	10%	1.1	1.1	1.1	1.2	1.1	1.1
	20%	1.2	1.2	1.2	1.4	1.2	1.2
	50%	1.7	1.7	1.7	2.6	1.6	1.4
	100%	3.6	2.7	2.7	7.0	2.5	2.0
	200%	13.9	5.3	6.0	35.7	4.8	3.7
	500%	80.8	17.4	24.0	269.9	15.0	10.9
	1000%	142.8	44.9	74.7	561.4	35.7	27.5
Ticks	10%	1.2	1.2	1.1	1.3	1.2	1.1
	20%	1.6	1.4	1.3	1.8	1.4	1.3
	50%	3.6	2.4	2.0	5.2	2.5	1.9
	100%	15.2	4.8	4.8	30.8	5.8	4.1
	200%	63.0	13.8	22.1	191.4	21.0	15.4
	500%	147.8	97.6	280.5	433.9	144.0	117.7
	1000%	167.5	533.9	1,417.3	513.2	572.1	360.7
Volatility	1%	1.0	1.0	1.0	1.0	1.0	1.0
	2%	1.0	1.0	1.0	1.0	1.0	1.0
	5%	1.1	1.0	1.0	1.3	1.2	1.2
	10%	1.8	1.4	1.4	3.0	2.5	2.4
	20%	5.8	4.2	3.7	12.9	11.1	9.9
	50%	52.4	53.4	40.5	149.9	191.4	148.8
	100%	264.7	501.3	401.6	936.3	1,832.4	1,516.9

Table 2. This table shows the average number of minutes between events of each type during the first 60 minute period, and the rest of the day for the given thresholds.

noted that these are over 8.5 times rarer than 5% return events when excluding the first 60 minutes, and therefore the 5% return values are the most interesting. These results signify a strong correlation between the advent of news and subsequent return events.

There appears to be a steady decrease in correlation between news and events as the threshold for volume events is increased. Only the 10-100% threshold range for LERN tests for the US show that there is an increased likelihood of news prior to a volume event. None of these has a value sufficiently high to suggest that there might be some correlation between the arrival of news followed by a volume event. Therefore there appears to be little correlation between news and volume events.

The UK market reveals that there is a stable decline in correlation between news and events as the threshold for tick events is increased. The US market shows a slight rise to the 100% threshold and then a stable decline afterwards. The 10%-200% threshold range for LERN tests for the US, and the 10%-20% threshold range, and 500%-1000% threshold range for LERN tests for the Australian markets imply that there is some correlation between news and tick events. However only the 500%-1000% threshold tests for the Australian market have values high enough to denote that could be a link between news and tick events. Therefore, whilst tick events appear to be a better indicator than volume events they don't appear to be too reliable.

All bar the 1-10% threshold range for LERN results for the Australian market and the 2%-5% threshold range for the UK suggest that volatility events are linked to news. There is a steady increase in correlation for all countries as the threshold is increased to 100%. Therefore it appears that there is a strong correlation between the arrival of news and later volatility events, with the 50% threshold providing the strongest evidence.

The results in Table 4 show the p-values for the ETT and NTT tests, where ΔT is set to 30 minutes, and therefore

Event Type	Threshold	RERNE			LERN		
		US	UK	AU	US	UK	AU
Return	0.1%	29.19%	17.52%	7.01%	**101.06%**	**100.25%**	**101.84%**
	0.2%	29.44%	17.80%	6.84%	**101.93%**	**101.87%**	99.35%
	0.5%	30.64%	18.83%	7.49%	**106.08%**	**107.77%**	**108.74%**
	1.0%	34.42%	22.12%	9.80%	**119.15%**	**126.57%**	**142.24%**
	2.0%	45.80%	30.90%	19.37%	**158.54%**	**176.86%**	**281.25%**
	5.0%	66.35%	55.29%	67.82%	**229.66%**	**316.41%**	**984.76%**
	10.0%	60.64%	79.76%	73.44%	**209.91%**	**456.48%**	**1066.41%**
Volume	10%	29.04%	17.37%	6.86%	**100.52%**	99.41%	99.64%
	20%	29.15%	17.26%	6.83%	**100.92%**	98.79%	99.16%
	50%	29.35%	16.88%	6.74%	**101.61%**	96.60%	97.93%
	100%	29.03%	16.23%	6.50%	**100.50%**	92.91%	94.45%
	200%	27.85%	15.08%	5.93%	96.39%	86.32%	86.06%
	500%	25.76%	13.64%	5.04%	89.16%	78.07%	73.18%
	1000%	22.63%	13.83%	4.83%	78.33%	79.16%	70.19%
Ticks	10%	29.19%	17.30%	6.91%	**101.04%**	98.99%	**100.36%**
	20%	29.51%	17.07%	6.94%	**102.15%**	97.68%	**100.81%**
	50%	30.51%	16.28%	6.84%	**105.62%**	93.15%	99.35%
	100%	33.03%	14.78%	6.29%	**114.33%**	84.56%	91.34%
	200%	32.75%	12.05%	5.70%	**113.35%**	68.94%	82.73%
	500%	21.22%	10.82%	8.24%	73.47%	61.93%	**119.61%**
	1000%	18.65%	11.80%	13.58%	64.55%	67.55%	**197.24%**
Volatility	1%	28.90%	17.47%	6.88%	**100.05%**	**100.00%**	99.89%
	2%	28.96%	17.40%	6.84%	**100.24%**	99.59%	99.37%
	5%	29.48%	17.39%	6.51%	**102.06%**	99.54%	94.53%
	10%	31.01%	17.81%	6.87%	**107.35%**	**101.93%**	99.79%
	20%	36.36%	21.03%	10.39%	**125.87%**	**120.35%**	**150.87%**
	50%	57.31%	37.62%	26.55%	**198.37%**	**215.32%**	**385.51%**
	100%	71.13%	61.24%	47.25%	**246.23%**	**350.47%**	**686.17%**

Table 3. This table shows the results of the RERNE, and LERN tests for each type for the given thresholds. The LERN results in bold have values over 100% which indicate that news is more likely prior to the given event type and threshold, than it is normally.

the first 60 minutes of the day are ignored. The tests in which we have at least 95% confidence to reject the null hypothesis are in bold. The 5% threshold range for every country and the 2% and 10% thresholds for the UK and the Australian markets are the only ETT results where we can reject the null hypothesis that return events are not influenced by news.

The 2%-5% threshold range for every country, and the 1%-2% threshold range and the 10% threshold for the UK and Australian markets are the only NTT results where we can reject the null hypothesis that the arrival of news before return events is the same as the arrival of news normally. Combining the two indicates that the 5% threshold for every country and the 2% and 10% thresholds for the UK and Australian markets have a strong correlation between news and return events. This concurs with the results of the LERN and RERNE tests, which imply that return events are linked to the arrival of news.

No ETT test results can be used to reject the null hypothesis that volume events are not linked to the arrival of news. Furthermore only the 500-1000% threshold range for the Australian market have NTT results which can reject the null hypothesis that the occurrence of news prior to a volume event is the same as news normally. Therefore these results appear to confirm the RERNE, and LERN tests that volume events do not imply that the market has reacted to news.

Only the 1000% threshold for the Australian market provides ETT test results which can reject the null hypothesis that tick events are not related to the advent of news. Furthermore only the 200% thresholds for the US, and 1000% threshold for the Australian market for the NTT test provide enough evidence to reject the null hypothesis that tick events do not imply that the market has reacted to news. This suggests that there is weak evidence that tick events are related to the arrival of news. However low RERNE and LERN values reveal

Event Type	Threshold	ETT			NTT		
		US	UK	AU	US	UK	AU
Return	0.1%	87.29%	85.15%	88.81%	97.22%	94.47%	94.46%
	0.2%	93.88%	62.19%	66.29%	97.31%	71.68%	71.26%
	0.5%	86.30%	42.14%	86.73%	80.35%	14.90%	62.05%
	1.0%	55.00%	16.31%	38.16%	8.07%	**0.01%**	**0.00%**
	2.0%	7.34%	**2.02%**	**4.22%**	**0.00%**	**0.00%**	**0.01%**
	5.0%	**2.15%**	**0.07%**	**0.18%**	**0.00%**	**0.00%**	**0.00%**
	10.0%	25.29%	**0.17%**	**1.17%**	5.09%	**0.03%**	**0.38%**
Volume	10%	95.32%	32.23%	88.29%	99.24%	93.65%	98.02%
	20%	98.17%	32.21%	83.90%	99.47%	87.16%	94.69%
	50%	89.48%	27.67%	84.73%	91.03%	65.75%	86.28%
	100%	76.17%	24.78%	73.12%	71.78%	39.02%	55.93%
	200%	58.37%	21.54%	51.10%	50.65%	14.75%	8.25%
	500%	75.37%	20.21%	25.12%	50.69%	6.16%	**0.23%**
	1000%	77.80%	20.64%	23.18%	86.15%	10.79%	**0.28%**
Ticks	10%	61.02%	49.98%	86.53%	92.69%	87.54%	96.66%
	20%	60.42%	47.79%	84.08%	85.10%	72.26%	92.54%
	50%	38.56%	47.83%	94.86%	55.47%	34.38%	92.37%
	100%	14.21%	47.48%	57.02%	5.52%	12.18%	39.67%
	200%	40.53%	45.23%	42.76%	**1.84%**	11.68%	20.74%
	500%	58.37%	50.59%	39.09%	8.91%	97.50%	5.57%
	1000%	49.22%	57.04%	**2.54%**	34.37%	92.87%	**0.84%**
Volatility	1%	36.25%	65.22%	7.26%	99.56%	99.96%	98.78%
	2%	47.01%	**0.12%**	11.14%	98.55%	95.00%	92.48%
	5%	93.38%	99.25%	18.48%	95.70%	98.90%	38.24%
	10%	99.90%	69.93%	71.52%	99.18%	39.44%	32.14%
	20%	63.53%	34.59%	52.36%	**2.05%**	**0.02%**	**0.00%**
	50%	10.56%	10.01%	9.26%	**0.00%**	**0.00%**	**0.00%**
	100%	13.92%	**4.74%**	**4.57%**	**0.01%**	**0.00%**	**0.00%**

Table 4. This table shows the p-values of the ETT, and NTT tests for each type for the given thresholds. The results in bold indicate that there is 95% confidence that the null hypothesis can be rejected for the given test, country, and parameters.

Event Type	Threshold	US	UK	AU
Return	2%	3.58%	3.94%	6.68%
	5%	0.37%	0.35%	1.29%
	10%	0.03%	0.06%	0.24%
Volatility	100%	0.62%	0.50%	1.07%

Table 5. This table shows the percentage of articles which occur outside the first period of the day which correlate to the given event type and threshold.

that, whilst tick events are more reliable than volume events, tick events aren't strongly correlated to news.

The 2% and 100% thresholds for the UK, and the 100% threshold for the Australian market for the ETT test provide evidence to reject the null hypothesis that volatility events are not affected by news. The 20-100% threshold range for all countries for the NTT test can be used to reject the null hypothesis that the advent of news prior to a volatility event are the same as the advent of news normally. It may bear further investigation into the values n, Δt, and $\Delta\tau$ for the US, as the RERNE, LERN, and NTT tests all strongly imply that news is correlated to volatility events. It could be that a shorter period is required to obtain better ETT results. Apart from the US ETT test, it appears that there is a strong correlation between the arrival of news and subsequent volatility events. This is strongly supported by the evidence of the RERNE and LERN tests, and therefore we conclude that volatility events are linked to news.

Finally Table 5 shows that only a fraction of articles which occur during the trading day, excluding the first 60 minutes, correlate with the given return and volatility events.

5. Conclusions

We have found strong evidence to suggest that the stock market does react to real time news. Return and volatility appear to give the most compelling evidence. However only the 5% threshold for all countries, and the 2% and 10% thresholds for the UK and Australian markets have return events which are supported by all of the RERNE, LERN, ETT, and NTT tests. This implies that further research is required to determine if volatility events occur differently when the market reaction period is changed. However there appears to be some weak evidence that news affects volume and tick events. Furthermore the most significant tests, shown in Table 5 imply that only a fraction of news is responsible for the most significant market reactions. This reveals that the market interprets news differently, and considers some news more significant than others.

Further research into when events and news occur is necessary to establish if the market behaves in a uniform manner. Furthermore the content of news which the market reacts to should be investigated, as Fung, Yu and Wai (2003) and Mittermayer (2004) have studied, in order to highlight potentially significant news articles for investors.

6. References

Almeida, A., Goodhart, C. A. E. and Payne, R. (1998): The Effects of Macroeconomic News on High Frequency Exchange Rate Behavior. *Journal of Financial & Quantitative Analysis*, **33**(3):383-408.

Bollerslev, T. and Domowitz, I. (1993): Trading Patterns and Prices in the Interbank Foreign Exchange Market. *Journal of Finance*, **48**(4):1421-1443.

Bomfim, A. N. (2003): Pre-announcement Effects, News Effects, and Volatility: Monetary Policy and the Stock Market. *Journal of Banking and Finance*, **27**(1):133-151.

Brannas, K. and De Gooijer, J. G. (2004): Asymmetries in conditional mean and variance: modelling stock returns by asMA-asQGARCH. *Journal of Forecasting*, **23**(3):155-171.

Chan, W. S. (2003): Stock Price Reaction to News and No-news: Drift and Reversal After Headlines. *Journal of Financial Economics*, **70**:223-260.

Dacorogna, M. M., Gencay, R., Muller, U., Olsen, R. B. and Pictet, O. V. (2001) *An Introduction to High-Frequency Finance*, Academic Press, London.

Donders, M. W. M. and Vorst, T. C. F. (1996): The Impact of Firm Specific News on Implied Volatilities. *Journal of Banking and Finance*, **20**(9):1447-1461.

Dungey, M., Fry, R. and Martin, V. L. (2004): Currency Market Contagion in the Asia-Pacific Region. *Australian Economic Papers*, **43**(4):379-95.

Ederington, L. H. and Lee, J. H. (1993): How markets process information: News releases and volatility. *Journal of Finance*, **48**(4):1161-1191.

Ederington, L. H. and Lee, J. H. (1995): The short-run dynamics of the price adjustment to new information. *Journal of Financial & Quantitative Analysis*, **30**(1):117-134.

Ederington, L. H. and Lee, J. H. (2001): Intraday Volatility in Interest-Rate and Foreign-Exchange Markets: ARCH, Announcement, and Seasonality Effects. *Journal of Futures Markets*, **21**(6):517-552.

Ewing, B. T. (2002): Macroeconomic news and the returns of financial companies. *Managerial and Decision Economics*, **23**(8):439-446.

Fung, G. P. C., Yu, J. X. and Wai, L. (2003): Stock Prediction: Integrating Text Mining Approach using Real-Time News. *Proc. IEEE International Conference on Computational Intelligence for Financial Engineering*, Hong Kong, 395-402.

Goodhart, C. A. E. (1989): News and the foreign exchange market. *Proc. Manchester Statistical Society*, 1-79.

Goodhart, C. A. E. and Figliuoli, L. (1991): Every Minute Counts in Financial Markets. *Journal of International Money and Finance*, **10**(1):23-52.

Goodhart, C. A. E. and Figliuoli, L. (1992): The Geographical Location of the Foreign Exchange Market: A Test of an 'Island' Hypothesis. *Journal of International and Comparative Economics*, **1**:13-27.

Goodhart, C. A. E., Hall, S. G., Henry, S. G. B. and Pesaran, B. (1993): News Effects in a High-Frequency Model of the Sterling-Dollar Exchange Rate. *Journal of Applied Econometrics*, **8**:1-13.

Graham, M., Nikkinen, J. and Sahlstrom, P. (2003): Relative Importance of Scheduled Macroeconomic News for Stock Market Investors. *Journal of Economics and Finance*, **27**(2):153-165.

Han, L.-M. and Ozocak, O. (2002): Risk-return relationships in foreign-currency futures following macroeconomic announcements. *Journal of Futures Markets*, **22**(8):729-764.

Hess, D. (2004): Determinants of the Relative Price Impact of Unanticipated Information in U.S. Macroeconomic Releases. *Journal of Futures Markets*, **24**(7):609-629.

Hong, H., Lim, T. and Stein, J. C. (2000): Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *Journal of Finance*, **55**(1):265-95.

Kim, S.-J. (1998): Do Australian and the US Macroeconomic News Announcements Affect the USD/AUD Exchange Rate? Some Evidence from E-GARCH Estimations. *Journal of Multinational Financial Management*, **8**(2-3):233-248.

Kim, S.-J. (2003): The Spillover Effects of US and Japanese Public Information News in Advanced Asia-Pacific Stock Markets. *Pacific-Basin Finance Journal*, **11**(5):611-630.

Kim, S.-J., McKenzie, M. D. and Faff, R. W. (2004): Macroeconomic News Announcements and the Role of Expectations: Evidence for US Bond, Stock and Foreign Exchange Markets. *Journal of Multinational Financial Management*, **14**(3):217-232.

Melvin, M. and Yin, X. (2000): Public Information Arrival, Exchange Rate Volatility, and Quote Frequency. *Economic Journal*, **110**(465):644-661.

Michaely, R. and Womack, K. L. (1999): Conflict of Interest and the Credibility of Underwriter Analyst Recommendations. *Review of Financial Studies*, **12**(4):653-86.

Mitchell, M. L. and Mulherin, J. H. (1994): The Impact of Public Information on the Stock Market. *Journal of Finance*, **49**(3):923-50.

Mittermayer, M.-A. (2004): Forecasting Intraday Stock Price Trends with Text Mining Techniques. *Proc. 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*, Big Island, Hawaii, 30064b.

Nikkinen, J. and Sahlstrom, P. (2004a): Impact of the Federal Open Market Committee's Meetings and Scheduled Macroeconomic News on Stock Market Uncertainty. *International Review of Financial Analysis*, **13**(1):1-12.

Nikkinen, J. and Sahlstrom, P. (2004b): Scheduled Domestic and US Macroeconomic News and Stock Valuation in Europe. *Journal of Multinational Financial Management*, **14**(3):201-215.

Nofsinger, J. R. and Prucyk, B. (2003): Option volume and volatility response to scheduled economic news releases. *Journal of Futures Markets*, **23**(4):315-345.

Peiers, B. (1997): Informed Traders, Intervention, and Price Leadership: A Deeper View of the Microstructure of the Foreign Exchange Market. *Journal of Finance*, **52**(4):1589-1614.

Roll, R. (1984): Orange Juice and Weather. *American Economic Review*, **74**(5):861-80.

Simpson, M. W. and Ramchander, S. (2004): An examination of the impact of macroeconomic news on the spot and futures treasuries markets. *Journal of Futures Markets*, **24**(5):453-478.

Sun, P. and Sutcliffe, C. (2003): Scheduled announcements and volatility patterns: The effects of monetary policy committee announcements on LIBOR and short sterling futures and options. *Journal of Futures Markets*, **23**(8):773-797.

Tse, Y. (1999): Market microstructure of FT-SE 100 index futures: An intraday empirical analysis. *Journal of Futures Markets*, **19**(1):31-58.

Womack, K. L. (1996): Do Brokerage Analysts' Recommendations Have Investment Value? *Journal of Finance*, **51**(3):137-67.

Wuthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J. and Lam, W. (1998): Daily Stock Market Forecast from Textual Web Data. *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, 2720-2725.

Investigating the size and value effect in determining performance of Australian listed companies: a neural network approach

Justin Luu **Paul J. Kennedy**

Faculty of IT,
University of Technology, Sydney,
PO Box 123, Broadway, NSW 2007,
AUSTRALIA,
Email: paulk@it.uts.edu.au

Abstract

This paper explores the size and value effect in influencing performance of individual companies using backpropagation neural networks. According to existing theory, companies with small market capitalization and high book to market ratios have a tendency to perform better in the future. Data from over 300 Australian Stock Exchange listed companies between 2000–2004 is examined and a neural network is trained to predict company performance based on market capitalization, book to market ratio, beta and standard deviation. Evidence for the value effect was found over longer time periods but there was less for the size effect. Poor company performance was also observed to be correlated with high risk.
Keywords: multilayer perceptron, size and value effect, company performance prediction.

1 Introduction

The stock exchange is an exceedingly fluid, dynamic and engaging entity. It facilitates thousands of transactions which occur simultaneously from traders striving to outbid and outsell each other. From the moment it opens there is unceasing activity until the second it closes. Decisions to buy, sell or hedge are based on analysis of sophisticated theoretical models or the instinct of a speculator. New information about company developments and stock recommendations are continuously made available while papers are released on new and different ways in which the market can be exploited. But can the market really be exploited?

Eugene F. Fama (1965) described how an active market filled with well informed and "intelligent participants" leads to a situation where the stock price reflects its actual value. This is due to the situation in which investors compete for new available information about the stock for profit. The stock will then promptly reflect the new price that the information retains. This is known today as the Efficient Market Hypothesis (EMH).

The EMH is a controversial idea, even today, as many investors and active fund managers truly believe that there is value in exploiting the timing of market. However, the great irony of the EMH is the market's ability to promptly correct itself when presented with news regarding a new inefficiency or mis–pricing: news which many investors attempt to exploit. This implies that it is not possible to make above-average returns. Once new information becomes available it "triggers a rapid process of adjustments, and re-prices the stock to its "correct" level" (Kingdon 1997).

However, there also exist anomalies in the market which contradict the EMH such as the size and value effect as described in the work by Fama and French in (1993). The size effect states that stocks with smaller portfolios of companies will perform better in the future while the value effect suggests that firms with a high book ratio in relation to its market price will also outperform.

The aim of this work is to investigate the EMH by testing existence of the size and value effects using a backpropagation multilayer perceptron (Reed & Marks II 1999) (Bishop 1995). In the process we examine the attributes from the three factor model developed by Fama and French that describe the size and value effects.

There has also been other work in the prediction of stock performance including studies by Gaunt (2004) and Albanis and Batchelor (2000).

Evidence of the three factor model as an effective pricing model in an Australian context can be seen with Gaunt (2004) which updates the study Halliwell, Heaney and Sawicki (1999) by examining Australian companies from the period of 1991–2000. The analysis shows that the three factor model has more explanatory power in predicting the future return on assets than a simpler one factor model, CAPM (see Section 2.2). Unlike our study, Gaunt (2004) divided the dataset into 25 portfolios ranking them into varying amounts of book to market value and market capitalization. Gaunt's results were consistent with Halliwell (1999) in that the three factor model provided better explanatory power than the traditional CAPM model for performance of portfolios. He found evidence of both the size and value effects and observed that less risky stocks offered better raw return. Our work is similar to Gaunt in that we investigate Australian companies. However, our study extends to more recent data and we also apply a neural network to predict individual stock performance.

Albanis and Batchelor (2000) describe different models of analysis for identifying high performing shares. They used analysis techniques including probabilistic neural networks, vector quantization, recursive partitioning and rule induction to investigate stock performance on the London stock exchange from 1991 to 1997. They observed that nonlinear approaches gave better classification performance than linear methods. Our work differs from this study in terms of recency of the data and the Australian context.

In section 2 we describe the Efficient Market Hypothesis and the Three Factor Model of return. Next, in section 3 we describe the data attributes used as proxies to components of the Three Factor Model and

the source and preprocessing applied to the data. In section 4 we describe the multilayer perceptron used to learn the relationship between data attributes and the performance of the company. Next in section 5 we detail the experiments run and their results. That section includes an account of initial trials predicting the performance class of companies and investigations into relationships between input data attributes using linear regression and scatterplots. Also we examine the effect on the learning task of using longer periods of data as well as brief details about other investigations. In section 6 we provide a summary of the results of the work and, finally, in section 7 we outline what we think are interesting further directions for study.

2 The Problem Definition

The aim of this paper is to implement a neural model to test Fama and French's theory of size and value effect in influencing stock performance. The theory suggests that firms with low market capitalization (i.e. small firms) and firms with high book to market ratios tend to perform better in the future. This investigation also tests the validity of the EMH where excess returns due to the size and value effect can only be gained through taking on extra risk.

2.1 Efficient Market Hypothesis

The efficient market hypothesis states that at any given time, stock prices fully reflect all available information of the asset's value. All past stock information can be reflected in the current stock price where this price changes only with the availability or release of new information. The present value of the stock is determined by discounting the expected future cash–flows (or dividends) of the stock by using all information investors have available to them (Kingdon 1997).

Using the present value model, the value of a stock can computed as

$$\text{PresentValue} = \text{CF}_1/(1+r)+\ldots+\text{CF}_i/(1+r)^i+\ldots+\infty \quad (1)$$

where CF represents the cash flow for the period, and r represents the required rate of return for the period. The term $\text{CF}_i/(1+r)^i$ represents the return for year i. The present value is the total sum of all the future values.

This means that the main determinant of a stock price lies in r. Generally speaking, the riskier the company, the higher the rate of return demanded by investors and the higher r (Brailsford & Heaney 1998). For instance, if a firm is risky, investors may only want to buy the stock if a return of 23% is guaranteed. Hence if the expected annual dividend for the rest of the company's life is \$0.50, then the expected value of the share is (\$0.50/1.23) + (0.5/1.232) + (0.5/1.233) + ... + ∞ = \$2.17. So if new information about a firm's future cash flow indicates that the current market price is lower than the stock expected value (e.g. \$2), then according to the EMH, investors will see that the share is undervalued resulting in purchase of more quantities of the stock, thus raising it to its fair value (i.e. back to \$2.17).

2.2 Three Factor Model

Using the three factor model, Fama and French (1993) argue that the rate of return r for a portfolio of stocks is determined by three attributes: (i) its return in relation to the market; (ii) its size; and (iii) its book to market ratio. These are known as the Capital Asset Pricing Model (CAPM), the size effect and the value effect respectively. The size effect indicates that portfolios of firms with low market capitalization (smaller firms) will perform better than the average market return in the long run, while the value effect suggests that portfolios of firms with high book to market ratios will also perform higher than average. According to Fama and French's three factor model, portfolios of firms with higher book to market ratios and low market capitalization tend to perform well (better than the market) as they tend to be more risky. To account for the extra risk, they will require a higher rate of return r from the stock in the future. This is modeled by the equation

$$r = \text{CAPM} + b_s \times \text{SMB} + b_\sigma \times \text{HML} + \alpha. \quad (2)$$

where CAPM refers to a model used by investors to determine the rate of return for valuing a portfolio of stocks. CAPM is defined as

$$\text{CAPM} = r_f + \beta(E[r_m] - r_f) \quad (3)$$

where r_f is the rate of the risk free asset (generally the government bond rate), $E[r_m]$ is the expected return of the market and β is the sensitivity of the stock to the market. In this study, r_f and $E[r_m]$ are constant for all companies at particular time periods, allowing us to use β to be an effective proxy for CAPM.

The Three Factor Model extends CAPM by adding two other factors namely SMB (small minus big) and HML (high minus low). Both these factors reflect the excess return that stocks of smaller companies and stocks with high book to market ratios are capable of delivering. The coefficients b_s and b_σ show the relative scale of the factors in relation to the portfolio, with $b_s = 1$ representing a portfolio having small capitalization and $b_s = 0$ representing a portfolio with large capitalization. Similarly, b_σ shows how high the firm's book to market ratio is compared to the market.

In summary, the Three Factor Model predicts a firm's future expected return on the basis of its return in relation to the market, its size and its book to market ratio.

3 Data

We examine evidence of the following in affecting the performance of companies:

1. CAPM model of returns;
2. size effect; and
3. value effect.

However, as the Three Factor Model is used to examine *portfolios* of stocks rather than individual stocks and due to the difficulty of obtaining information on portfolios of stocks with the inherent inconsistencies of the available data, we will proxy data items for the factors in equation (2). The proxy data items for individual companies (Beta, Market Capitalization and Book to Market Ratio respectively) are used instead of the actual CAPM, size effect and value effect. Hence the neural network will examine the effect that Beta, Market Capitalization, and Book to Market Ratio have on the future return of each individual firm. Additionally, another data item "Standard Deviation" will be used as a proxy and control factor for volatility.

Financial data for a group of companies listed on the Australian Stock Exchange (ASX) for the years 2000–2004 were obtained from Aspect Huntley [1] and

[1] See `http://www.aspectfinancial.com.au/af/aerhome?xtm-licensee=aer` and `http://www.aspectfinancial.com.au/af/finhome?xtm-licensee=finanalysis` for details.

from quarterly reports from the Australian Graduate School of Management (AGSM). To ensure the consistency of returns over the time period, only companies which reported returns in June by Aspect Huntley and companies with no missing data were used. From the 1315 companies reported by the AGSM, 346 were used in this study. Despite the filtering, the final dataset still represents a broad cross section of the ASX.

Performance of these companies is divided into three categories: high performing, medium performing and low performing. High performing companies are those with market return over the calculated period falling into the top third of the group, with the medium comprising of the next third and low performing companies in the bottom third.

Data for Beta, Market Capitalization and Standard Deviation were obtained from the AGSM while Book to Market Ratio and the category rankings came from Aspect Huntley.

Next we describe the individual input factors.

Market Capitalization of a firm is the value of the total amount of stock it has outstanding. It can be calculated by multiplying its current share price by the number of stocks it has on issue. The Market Capitalization is a useful indicator of the size of a company, and has been used to evaluate the effectiveness of Fama and French's size effect in determining returns (Fama & French 1995).

The Book to Market Ratio is the historical (or accounting value of a firm) divided by its Market Capitalization. It will be used to determine whether a stock is over or undervalued. The book value represents the net assets of a firm calculated by subtracting its current assets from its current liabilities.

Standard Deviation measures volatility or risk of an investment by showing the average amount by which it deviates from the mean. Generally speaking, the higher the standard deviation of a stock the higher its risk. It will be used as the control factor for risk to test the size effect and value effect. It is understood from Fama and French's work that the smaller companies and value stocks encounter greater risk, and hence it is necessary to determine how much excess return is attributed to higher risk and how much is attributed to the size factor.

4 Neural Network Model

A fully connected feed forward multilayer perceptron (MLP) was trained using backpropagation and momentum. Multilayer perceptrons were chosen for this study due to their known ability to act as universal approximators (Haykin 1999) and hence are suitable for this type of non-linear problem.

Preliminary investigations compared performance of the MLP with Support Vector Machines (SVM) and Naive Bayes classifiers on the datasets. Performance by the SVM was not markedly better than with the MLP so we chose to continue with the MLP in the investigations.

All networks in this study had four inputs, three output neurons and one hidden layer. The four inputs correspond to the input factors: Beta, Market Capitalization, Book to Market Ratio and Standard Deviation respectively. The three output neurons correspond to the category ranking classes: low, medium and high performing respectively. Our networks contained between three and eight neurons in the hidden layer. All neurons in the hidden and output layers used the sigmoid transfer function.

All data was scaled into $[0, 1]$ before presentation to the network. Additionally, Market Capitalization was first transformed with the logarithmic function before scaling so as to compress the range of possible values due to the enormous variation in the values.

The output class for a particular input pattern was the one associated with the output neuron having the highest value.

5 Experiments

5.1 Initial Results

Initially we compared two methods of training the MLP: (i) stopping training using a holdout set; and (ii) training for a fixed number of epochs with estimation of test error using 10–fold cross validation (Witten & Frank 2005).

The dataset of companies was divided randomly into three datasets for training with the holdout set method. Seventy percent of the data was presented to the MLP for training. The next 15% was used to determine when to stop training the MLP. When the error on this set started to increase training was stopped. The final 15% was used for quoting the accuracy of the MLP in the classification task. The other method of training the MLP used a 10–fold cross validation scheme with all available data.

Networks were trained on data from 2000 to 2001. Results are shown in Table 1. The results show that the best accuracy arose using the holdout method with 7 neurons in the hidden layer. This accuracy is much greater than random choice of the company's class of performance. While this is an improvement on the random model of 33%, it is not a high number. These initial results over the 2000–2001 period do not strongly support the value and size effects due to the limited predictability using the input variables.

We next investigate potential reasons for the relatively poor classification accuracy.

5.2 Investigating relationships between the input variables and the output

Linear regression analysis is used to investigate the degree to which the input variables affect the output results. In particular, we are interested in which input variables are the strongest predictors of the output class. Table 2 shows a regression of inputs to the real–valued output (i.e. the actual performance rather than the class). There is only a very weak relation between the output and the input attributes as evidenced by an R–squared value of 0.058. The only factor which has a strong linear relation with the output variable is the Standard Deviation, as indicated by its $\rho < .05$ within a 95% confidence interval. The negative coefficient for the Standard Deviation demonstrates that the higher the volatility (an indicator of risk) of the company, the higher the likelihood of a poor return in the future. Beta, Market Capitalization and Book to Market Ratio do not show any statistically significant linear relationship with performance.

Table 2: Results from linear regression of inputs to performance. 2000–2001 data.

Attribute	Coefficients	P–value
Intercept	30.91	0.13
Market Capitalisation	11.10	0.67
Beta	-37.88	0.25
Standard deviation	-61.49	0.02
Book to market ratio	58.70	0.08

Figures 1 and 2 show scatter plots of the input attributes for the 346 companies for the 2000 to 2001 period. All values on the axis have been scaled between 0 and 1. The plots show that there is a high

Table 1: Results of training neural networks, 2000–2001 data.

Hidden Neurons	Parameters	Accuracy	Mean Absolute Error
	Training using the holdout method		
3	Learning rate 0.1, momentum 0.5	47.83%	0.38604
4	Learning rate 0.2, momentum 0.5	45.45%	0.36732
5	Learning rate 0.2, momentum 0.4	50.00%	0.38785
6	Learning rate 0.1, momentum 0.7	45.65%	0.37204
7	**Learning rate 0.1, momentum 0.7**	**58.70%**	**0.34210**
8	Learning rate 0.1, momentum 0.7	47.83%	0.39517
	Training with fixed 500 epochs and 10–fold crossvalidation		
3	Learning rate: 0.3, momentum 0.2	50.00%	-

degree of overlap between the classes and that they cannot be easily differentiated. This is one reason why the MLP has difficulty with this problem. Figure 1 plots Beta against Market Capitalisation and shows that the dataset is unevenly skewed to smaller companies. This is expected due to market's value dominated by a small number of very large companies. Although most of data for the different categories is overlapping, a slight pattern emerges for firms with a Market Capitalisation greater than 0.5. There are very few low performing firms in this region and more high performing than mid performing suggesting that larger firms perform better than smaller firms and contradicting the size effect.

5.3 Longer periods

The data used up to this point has been from the period June 2000 to June 2001 which coincides with the period directly after the market crash caused by technology stocks. This period was a time of considerable restructure in the stock market globally. The size and value effect are known to be more apparent after a longer period of time (Arnott 2005). It is hence necessary to investigate longer periods of time.

Table 3 summarizes test accuracies for MLPs trained on different periods. All networks were trained for 500 epochs (learning rate: 2, momentum: 3) and had a topology of three hidden neurons. The table shows, as expected, that MLPs trained with data over a longer period were generally more accurate. This suggests that the size and value effect is evident in the dataset, however require the dataset to be analyzed from longer periods of time in order for it to emerge more clearly.

Figure 3 plots Standard Deviation against the Book to Market Ratio for the period 2000–2003. Data points for each class of company cluster reasonably clearly into three distinct areas. This suggests that the higher the Standard Deviation, the lower the future returns for the firm implying that firms which are more volatile tend to perform poorly in the long term, and reinforcing our findings from the linear regression. The value effect can also be observed in Fig. 3 with the low performing firms clustered towards the lower end of the Y–axis (book to market ratio) while high performing firms are spread towards the higher areas of the graph. This suggest that underpriced firms (with high Book to Market Ratios) will revert back to their true value over the three year period and generally exceed the market's performance.

Figure 4 plots Beta against Market Capitalisation. It is interesting to compare this graph with Fig. 1 which plots the same values over the shorter period. In Fig. 1 large firms (with Market Capitalization greater than 0.5) tend to be the best performers, however after three years the majority of firms with large Market Capitalisation actually becomes mid performing firms. This shows that the large companies which may have performed well in 2000–2001 but after another two years their performance fell to a medium level. This anomaly may have been caused by investors favoring well established large companies over the more risky technology stocks after the crash of 2000, pushing their stock price up, but then abandoning them later when investors realized they were overpriced. This observation does not entirely reinforce the size effect as there is a mix of high and low performing smaller firms, suggesting that the size of a firm does not have a direct bearing on its future performance. Even though only the data from 2000-2003 has been presented here, analysis of the other periods also yielded similar findings.

5.4 Other Investigations

Whilst investigating the relationship between Market Capitalisation and risk, a MLP was trained to determine the current Market Capitalisation from the Beta and Standard Deviation. Using this analysis a correlation was discovered between the size of a company and it's risk. Specifically, that larger companies had a lower Standard Deviation and had a positive correlation with Beta. This reinforces Fama and French's suggestions indicated earlier that smaller firms were underpriced due to their inherent riskiness.

Investigations of classifying companies into five performance classes rather three resulted in a 10–fold cross validation scheme yielded accuracy of 34.50% compared to random selection of 20% over the 2002–2004 period. The confusion matrix (Table 4) shows that the neural network has difficulty differentiating between the high performers and the low performers. It was observed that high performing stocks tended to be either very large or very small firms.

Table 4: Confusion matrix of classification into five performance classes rather than three.

Very High	High	Medium	Low	Very Low	classified as ←
48	5	1	3	23	Very High
31	2	7	8	23	High
9	10	9	11	13	Medium
15	2	7	14	28	Low
18	3	0	7	45	Very Low

Other investigations which were conducted included the removal of Technology firms from the dataset (as classified by the Global Industry Classification Standard), in order to control for the influence that the technology crash of 2000 had on the results. It was observed that the network found it more difficult to predict the performance of the remaining companies after the removal. Elimination of outliers from the data set also did not yield any improvement in the results.

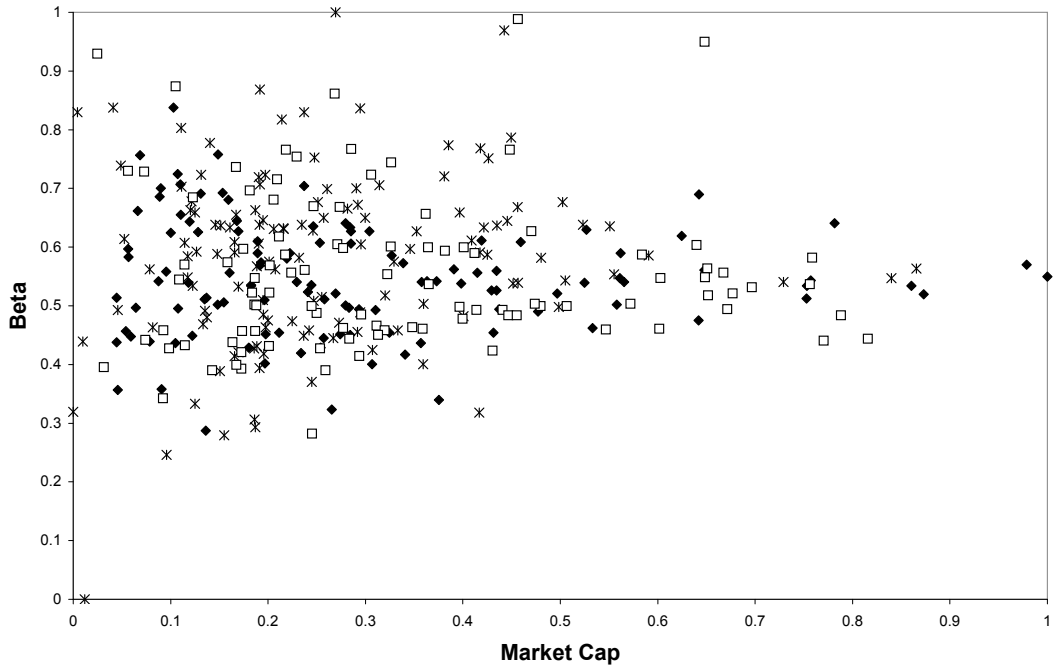


Figure 1: Scatter plots of attributes for 2000–2001 data: Beta vs Market Capitalisation. × = low performing companies, ♦ = medium performing companies, □ = high performing companies.

Table 3: Accuracy of MLP trained on different periods of data. Accuracy is the 10–fold cross validation value.

Period from	2000				2001			2002	
Period to	2001	2002	2003	2004	2002	2003	2004	2003	2004
Accuracy (%)	50.00	51.16	58.55	56.52	45.09	52.75	54.49	59.36	57.02

6 Conclusion

This study examined evidence for the size and value effect by building MLP models to predict the class of performance of ASX–listed companies over the period 2000–2004.

We were able to classify 58.70% of the companies correctly on 2000–2001 data with an MLP built using the holdout method. This network had 7 neurons in the hidden layer. By contrast, the best network we were able construct with 3 neurons in the hidden layer training for 500 epochs had an accuracy estimated by 10–fold cross validation of 50.00%.

Investigation of the 2000–2001 dataset with linear regression suggested that only the Standard Deviation attribute showed any significant relationship with the performance of the firms. Standard Deviation was inversely proportional to the future return of the company suggesting that investment into volatile companies led to poor investment return. Scatter plots of the attributes did not show distinct areas for different performance classes.

However, analysis of longer time periods, specifically 2000–2003, showed evidence of the value effect with the accuracies of 3–neuron hidden layer MLPs considerably higher than that of the 2000–2001 time periods. There is only weak evidence for the size effect, with small firms tending to be both high performers as well as poor performers. It was, however, discovered that the high performing large firms from 2000 had fallen to medium performance over a longer time period.

In terms of the three factor model, the only factor which was clearly observed was the value effect as proxied by the Book to Market Ratio. There was a relationship between Book to Market Ratio and the future return on a stock over the long term, implying that underpriced firms will eventually revert to their fair value. The findings show evidence against the EMH as existence of the value effect represents an anomoly in the expected behavior of stocks. The outperformance of high book to market firms also cannot be simply attributed to higher risk, with our findings indicating that higher Standard Deviation leading to lower returns in the future. Although this study did not *directly* examine the three factor model as individual stocks were examined rather than portfolios and also proxied for values, it does offer new insights into the size and value effect on the Australian stock market.

Our results support Gaunt's findings in (2004) that there was a value effect, however, we were unable to find strong evidence for the size effect. Our findings also support the study by Albanis and Batchelor (2000) in that nonlinear methods resulted in more accurate models than linear models.

7 Future Work

Further work with other financial measures to predict future investment return, such as the price to earnings ratio or the debt to equity ratio, would yield interesting insights into the EMH.

Also, given the superior performance of the seven hidden neuron MLPs over those with three hidden neurons, it would also be interesting to extend the additional modeling of longer time periods and finer granularity performance classes to these networks.

Moreover, extending Gaunt's study to more recent data with portfolios of stocks rather than individual companies may yield additional insights into the size and value effects.

Also, it would be interesting to compare models built with the MLP with an SVM. As mentioned

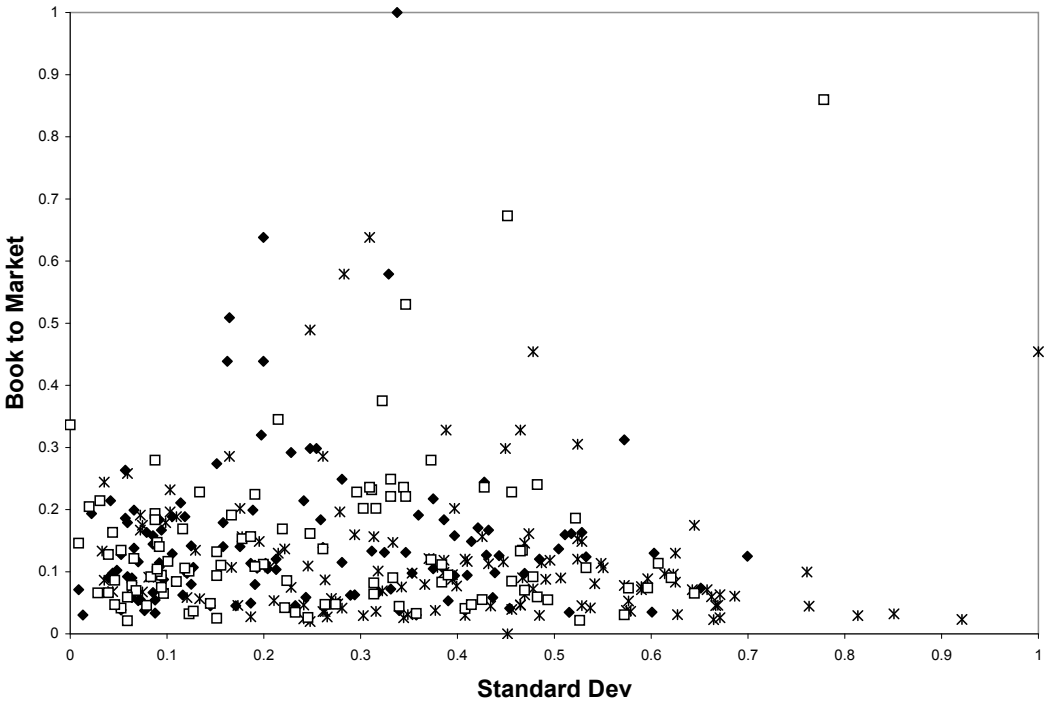


Figure 2: Scatter plots of attributes for 2000–2001 data: Book to market ratio vs standard deviation. × = low performing companies, ♦ = medium performing companies, □ = high performing companies.

above, we conducted initial experiments comparing performance of an MLP and an SVM in this task and found that there was not significant differences in the results. However, more careful modelling, particularly using lessons learnt in this work, may allow more accurate SVMs to be constructed.

References

Albanis, G. T. & Batchelor, R. A. (2000), Five classification algorithms to predict high performance stocks, *in* C. Dunis, ed., 'Advances in quantitative asset management', Kluwer Academic Publishers, pp. 295–318.

Arnott, R. D. (2005), 'Disentangling size and value', *Financial analysts journal* pp. 12–15. CFA Institute.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

Brailsford, T. & Heaney, R. (1998), *Investment: Concepts and applications in Australia*, Harcourt, London.

Fama, E. E. & French, K. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* (33), 3–56.

Fama, E. E. & French, K. (1995), 'Size and book to market factors in earnings and returns', *Journal of Finance* (50), 131–155.

Fama, E. F. (1965), 'Random walks in stock market prices', Paper No. 16 in the series of Selected Papers from the Graduate School of Business, University of Chicago, 1965. Reprinted in the Financial Analysts Journal (Sept–Oct 1965), The Analyst Journal, London (1966), The Institutional Investor, 1968.

Gaunt, C. (2004), 'Size and book to market effects and the Fama French three factor asset pricing model: evidence from the Australian stockmarket', *Accounting and Finance* **44**, 27–44.

Halliwell, J., Heaney, J. & Sawicki, J. (1999), 'Size and book to market effects in Australian share markets: a time series analysis', *Accounting Research Journal* **12**, 122–137.

Haykin, S. (1999), *Neural networks: a comprehensive foundation*, 2nd edition edn, Prentice–Hall.

Kingdon, J. (1997), *Intelligent Systems and Financial Forecasting*, Springer–Verlag Telos.

Reed, R. D. & Marks II, R. J. (1999), *Neural Smithing*, MIT Press, Cambridge, Massachusetts.

Witten, I. H. & Frank, E. (2005), *Data mining: Practical machine learning tools and techniques*, second edn, Elsevier.

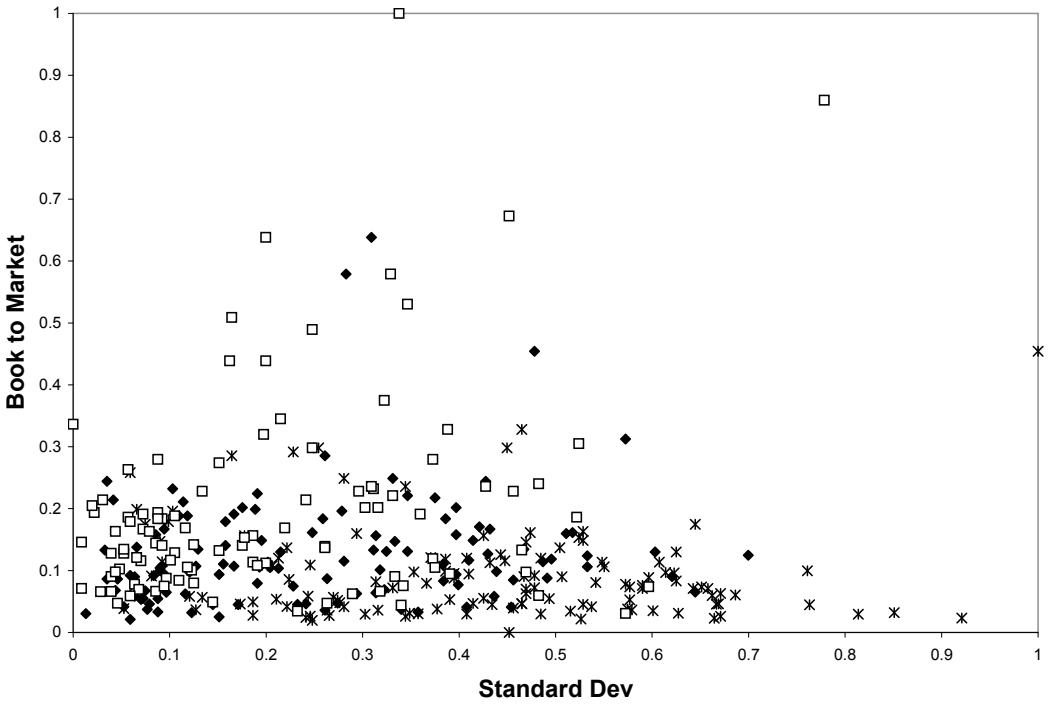


Figure 3: Scatter plots of attributes for 2000–2003 data: Standard deviation vs book to market ratio. × = low performing companies, ♦ = medium performing companies, □ = high performing companies.

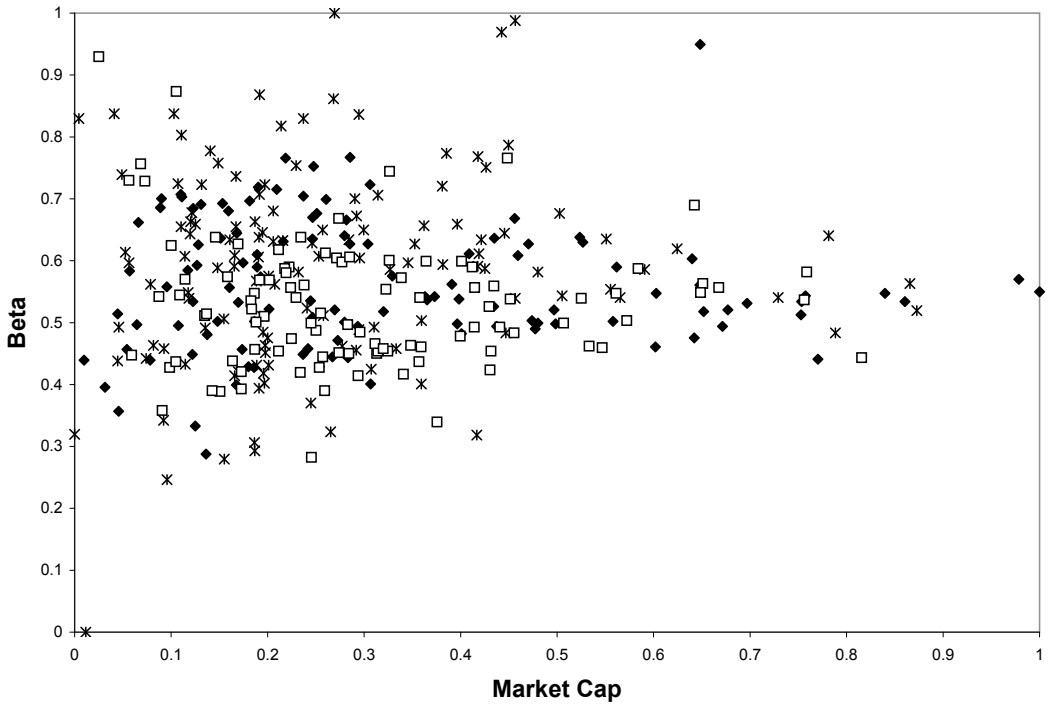


Figure 4: Scatter plots of attributes for 2000–2003 data: Market capitalisation vs beta. × = low performing companies, ♦ = medium performing companies, □ = high performing companies.

Extraction of Flat and Nested Data Records from Web Pages

Siddu P Algur[1] and P S Hiremath[2]
[1]Dept. of Info. Sc. & Engg., SDM College of Engg & Tech, Dharwad,Karnataka ,India
siddu_p_algur @hotmail.com
[2]Dept. of Computer Science,Gulbarga University, Gulbarga,Karnataka ,India
hiremathps@yahoo.co.in

Abstract

This paper deals with studies the problem of identification and extraction of flat and nested data records from a given web page. With the explosive growth of information sources available on the World Wide Web, it has become increasingly difficult to identify the relevant pieces of information, since web pages are often cluttered with irrelevant content like advertisements, navigation-panels, copyright notices etc., surrounding the main content of the web page. Hence, it is useful to mine such data regions and data records in order to extract information from such web pages to provide value-added services. Currently available automatic techniques to mine data regions and data records from web pages are still unsatisfactory because of their poor performance. In this paper, we propose a new method to identify and extract the data records from the web pages automatically. Given a page, the proposed technique first identifies the data region based on the visual clue information. It then extracts each record from the data region and identifies it whether it is a flat or nested record based on visual information – the area covered by and the number of data items present in each record. The experimental results show that the proposed technique is effective and better than existing techniques.

Keywords: Web mining, Web data regions, Web data records

1. Introduction

Many companies manage their business and publish their products and services on the Web. Collection and organization of this dynamic information can produce the data for many value-added applications. In order to collate and compare the prices and features of products available from the various Web sites, we need tools to extract attribute descriptions of each product (called data object) within a specific region (called data region) in a pages.

As illustrated in Fig. 1, there are many irrelevant components intertwined with the descriptions of data objects in web pages. These items include advertisement bar, product category, search panel, navigator bar, and copyright statement. In many web pages, there are

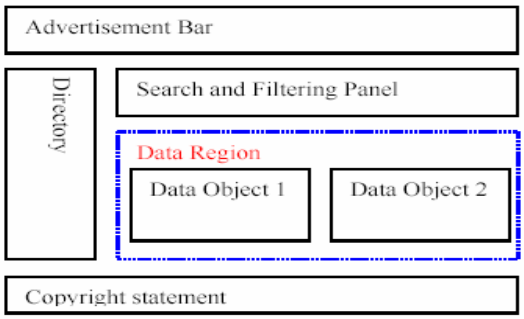


Fig 1: A schematic view of a webpage

normally more than one data object intertwined together in a data region. Furthermore, the raw source of the web page for depicting the objects might be non-contiguous. So it is difficult to discover the attributes for each object.

In real applications, what the users want from complex web pages is the description of individual data object derived from the partitioning of data region. There are several approaches by Hammer, Garcia Molina, Cho, and Crespo (1997), Kushmerick (2000), Chang and Lui (2001), Crescenzi, Mecca, and Merialdo (2001), Zhao, Meng, Wu and Raghavan (2005) proposed in the literature to address the problem of web data extraction, which are called wrapper generation.

The first approach by Hammer, Garcia Molina, Cho, and Crespo (1997) is to manually write an extraction program for each web site based on observed format patterns of the site. This manual approach is very labor intensive and time consuming and thus does not scale to a large number of sites.

The second approach Kushmerick (2000) is wrapper induction or wrapper learning, which is currently the main technique . Wrapper learning works as follows: The user first manually labels a set of trained pages. A learning system then generates rules from the training pages. The resulting rules are then applied to extract target items from web pages. These methods either require prior syntactic knowledge or substantial manual efforts. An example of wrapper induction systems is WEIN by Baeza Yates (1989).

The third approach Chang and Lui (2001) is the automatic approach. Since structured data objects on the web are normally database records retrieved from underlying web databases and displayed in web pages with some fixed templates, automatic methods aim to find

patterns/grammars from the web pages and then use them to extract data. Examples of automatic systems – IEPAD by Chang and Lui (2001), ROADRUNNER by Crescenzi, Mecca, and Merialdo (2001).

The fourth approach is MDR by Liu, Grossman, and Zhai (2003) which basically exploits the regularities in the HTML tag structure directly. It is often very difficult to derive accurate wrappers entirely based on HTML tags. The MDR algorithm makes use of the HTML tag tree of the web page to extract data records from the page. However, erroneous tags in the HTML source pages may result in building of incorrect trees, which in turn makes it impossible to extract data records correctly. MDR has several other limitations which will be discussed in the latter half of this paper. DEPTA by Zhai, and Liu (2005) uses visual information (locations on the screen at which the tags are rendered) to infer the structural relationship among tags and to construct a tag tree. But this method of constructing a tag tree has the limitation that, the tag tree can be built correctly only as far as the browser is able to render the page correctly. The computation time for constructing the tag tree is also an overhead. Further, this method also fails to identify some of the data records.

NET by Benchalli, Hiremath, Siddu, and Renuka (2005) extracts data from web pages that contain a set of flat or nested data records automatically in two steps. This approach also depends on building of tag tree and post order traversal of the tag tree to identify data records at different levels.

We propose a novel and more effective method to extract data records from web pages that contain a set of flat or nested data records automatically. Our method is called **ENDR (E**xtraction of Flat and **N**ested **D**ata **R**ecords from Web Pages). The experimental results show that the proposed technique is more effective than existing techniques substantially.

2. Related Work

Extraction the regularly structured flat or nested data records from a web page is an important problem. So far, some attempts have been made to deal with the problem. For automatic extraction, in Crescenzi, Mecca, and Merialdo (2001), Zhao, Meng, Wu and Raghavan (2005), Lerman, Getoor, Minton, and Knoblock (2004), it is proposed to find patterns or grammars from multiple pages containing similar data records. They require an initial set of pages containing similar data records which is, however, a limitation. In Lerman, Getoor, Minton and Knoblock, (2004), it proposes a method that tries to explore the detail information pages behind the current page to segment the data records. The need for such detail pages is a drawback because many data records do not have such pages or perhaps such pages are hard to find. In Chang and Lui (2001), string matching method is studied. However, it could not find nested data records. A similar method is proposed in Wang, Lochovsky (2003). Liu, Grossman, and Zhai (2003) and Zhao, Meng, Wu and Raghavan (2005), it some algorithms are proposed to identify data records, but they do not extract data items from the data records and do not handle nested data records. DEPTA by Zhai and Liu (2005) is able to align and extract data items from the data records but does not handle nested data records.

The NET by Benchalli, Hiremath, Siddu and Renuka (2005) (Nested data Extraction using Tree matching) works in two main steps:

(i) Building a tag tree of the page: Due to numerous tags and unbalanced tags in the HTML code of the page, building a correct tag tree is a complex task. A Visual based method is used to deal with this problem.

(ii) Identifying data records and extracting data from them: The algorithm performs a post order traversal of tag tree to identify data records at different levels. This ensures that nested data records are found. The tree edit distance algorithm and visual clues are used to perform these task.

Though the technique is able to extract the flat or nested data records, construction of tag tree and its post order traversal is consider to be an overhead.

The above automatic methods are inaccurate, tag dependant, incorporate time-consuming tag tree construction, and are based many assumptions which do not always hold good for all web pages. The proposed method does not make such assumptions and can scale well for almost all web pages. It is also independent of the type of tags and dispenses with the time-consuming tag tree construction procedure.

3. Data Region Extraction

We now start to present our proposed technique .This section focuses on the extraction of Data Region from the web page. The extraction of the data records and extraction of the data items in the data records will be the topic of the next section. Since this step is an improvement of our previous technique VSAP by Benchalli, Hiremath, Siddu, and Renuka (2005), we give a brief overview of the VSAP algorithm.

3.1. The Basic Idea of VSAP

An effective method to mine the data region in a web page automatically is the VSAP by Benchalli, Hiremath, Siddu and Renuka (2005). The visual information (i.e., the locations on the screen at which tags are rendered) helps the system to identify gaps that separate data records, and, thus, helps to segment data records correctly, because the gap within a data record (if any) is typically smaller than that in between data records. Also, by the visual structure analysis of the web pages, it can be observed that the relevant data region seems to occupy the major central portion of the web page.

The VSAP technique works as follows in two steps:

Step 1) Determination of the co-ordinates of all bounding rectangles in the web page

The first step of the VSAP technique determines the co-ordinates of all the bounding rectangles in the web page. The rendering engine of the browser produces the boundary coordinates. A bounding rectangle is constructed by obtaining the co-ordinate of the top-left corner of the tag, the height and the width of that tag. The left and top co-ordinates of the tag are obtained from the

offsetLeft and offsetTop properties of the HTMLObjectElement.

Fig 2 A sample web page of a product related website

Step 2) Data Region Identification

The second step of the VSAP technique is to identify the data regions of the web page. There are 3 steps involved in identifying the data region:

a) ***Identify the largest rectangle.***

Based on the height and width of bounding rectangles obtained in the previous step, the area of the bounding rectangles of each of the children of the BODY tag are determined. Then the largest rectangle amongst these bounding rectangles is found. The reason for doing this is due to the observation that the largest bounding rectangle will always contain the most relevant data in that web page. Thus, by determining the largest rectangle, a superset of the data region is obtained.

b) Identify the container within the largest rectangle.

Once the largest rectangle is obtained, a set of all the bounding rectangles whose area is more than half the area of the largest rectangle is formed. The rationale behind this is that the most important data of a web page must occupy a significant portion of the web page. Then the bounding rectangle having the smallest area in this set is found. The reason for determining the smallest rectangle within this set is that the smallest rectangle will only contain data records. Thus a *container* is obtained, which contains the data region and some irrelevant data.

c) Identify the data region containing the data records within this container

To filter the irrelevant data from the container, a *filter is used*. The filter determines the average height of children within the container. Those children whose heights are less than the average height are identified as irrelevant data and are filtered off. The outcome of the filter is a data region. The data regions of the web page in Fig 2 are shown in the Fig 3.

4. The Proposed Technique

We propose a more effective method to extract flat or nested data records from a given web page automatically. The method is called **ENDR (E**xtraction of Flat and **N**ested **D**ata **R**ecords from Web Pages). Before presenting the method, we discuss three observations about data records in web pages, which simplify the extraction task. These observations were made in Benchalli, Hiremath, Siddu, and Renuka (2005).

Fig 3. Filtered Data Region

(a) A group of data records, that contains the descriptions of a set of similar observations, is typically presented in contiguous region of a page.

(b) The area covered by rectangle that bounds the data region is more than the area covered by rectangles bounding other regions. eg., advertisements and links.

(c) The height of irrelevant data records within a collection of data records is less than the average height of relevant data records within that region.

The experimental results show that these observations are true.

Definition 1: A *flat data* record is defined as a collection of data items that together represents a single meaningful entity.

eg., the product having single size, look, price etc.,

Definition 2: A *nested data record* is defined as one that provides multiple description of the same entity.

eg., the same type of products but different sizes, looks, prices etc.,

The Fig.4 illustrates an example, which is a segment of a web page that shows flat and nested data records.

The system model of the **ENDR (E**xtraction of Flat and **N**ested **D**ata **R**ecords from Web Pages) technique is shown in Fig.5.
When a web page having description of products is given to VSAP, it identifies and extracts the data region. All the noises of a given web page are eliminated using filter. The filtered data region corresponding to the figure is shown in Fig.6.

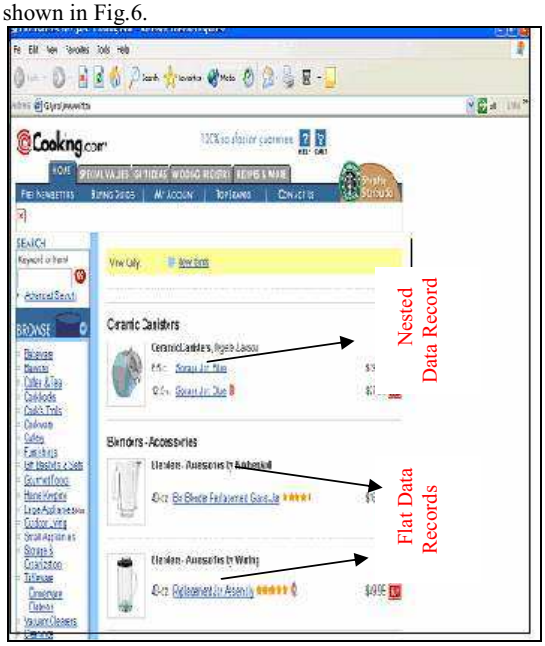


Fig. 4 An example of flat and nested data records

The filtered data region is given as the input to our system which extracts the flat and nested data records from the given data region and extracts data fields from the identified records.

4.1 Extraction of data records

Extraction of data records is based on visual clues. In the first step of the proposed technique, we determine the height of all the data records. This approach uses the MSHTML parsing and rendering engine that gives the height of each data record. The height of the data record is obtained from the offsetHeight property of the HTMLObjectElement. Next, the average height of the records is calculated. The average height of all the records provides the approximate height of each record The height of each data record is compared with the average height. If the height of the child is greater than or equal to the average height, then the data record is extracted.

The procedure Extract Data Record extracts the flat and nested records from given data region. It is as follows.

```
Procedure
ExtractDataRecord(dataRegion)
{
  THeight=0
        For each child of dataRegion
  BEGIN
     THeight += height of the bounding
                rectangle of child
  END
AHeight = THeight/no of children of
dataRegion
        For each child of dataRegion
BEGIN
    If height of child's bounding
       rectangle > AHeight
       BEGIN
         dataRecord=child
       END
END
}
```

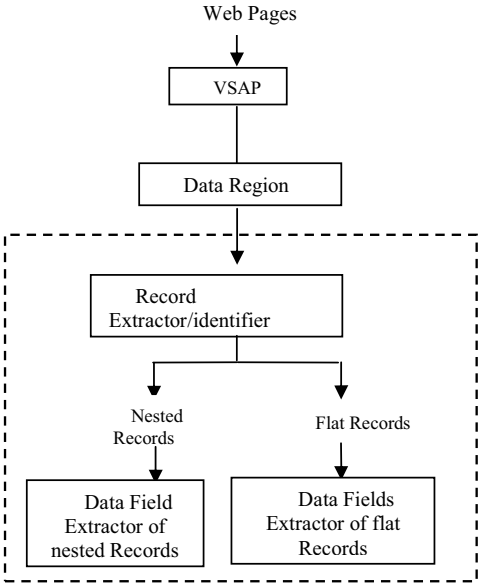


Fig. 5 System Model

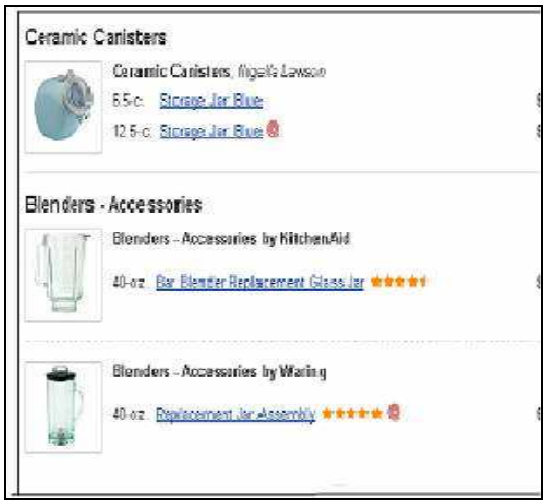


Fig. 6 Filtered data region

The Fig. 7 shows extracted data records from the data region shown in the Fig.6.

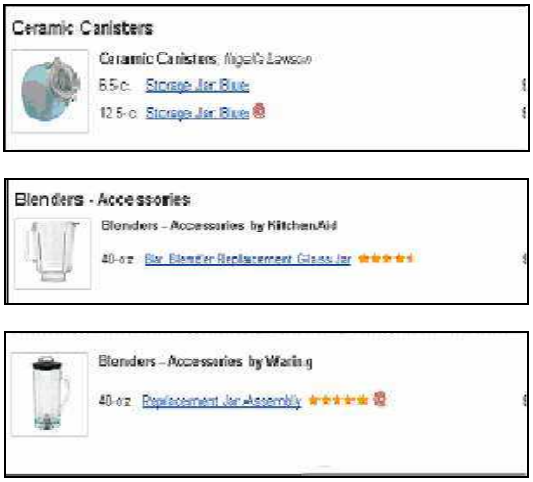


Fig .7 Extracted data records

4.2 Identification of data records

Identification of data records, as flat or nested, is essential in order to simplify the task of extracting the data items, which is very useful for various applications as mentioned earlier.

This technique determines the data fields for each data record within the data region. Various tags such as <TD>, <TR>, <A>, <IMG>, represent the data fields. By counting these tags as they are encountered, the number of fields is obtained. The flat record gives description of a single entity, whereas the nested data record gives multiple description of a single entity, so the data fields in flat records are less as compared to that of nested records. Experimental observations have shown that the number of fields in the nested data records is atleast 40% (approx) more than that of the flat records. The number of fields in the first record is compared with the number of fields in the next record. If the number of fields is more than 40%, then it is a nested record else it is a flat record. Suppose a condition is encountered where the number of fields is equal then in both cases. Then the record is compared with the third record and so on until the condition is satisfied.

The procedure IdentifyNestedData identifies whether the record is flat or nested based on the number of data items present in the data record. It is as follows:

```
Procedure
IdentifyNestedData(dataRecord[I],
dataRecord[I+1])
  {
      noofField[I]=0
  For I 1 to no of records
    BEGIN
       noofFields [I]=
           noofFields[I]+noofFields
             in the record[I]
    END
  DO
      For I 1 to no of records
     BEGIN
  For dataRecord [I], dataRecord[I+1]
       IF the no of fields in the
      [I+1]th record>=40% of the no of
           fields in the [I]th record
        Then [I+1]th record is a
              nested data record
         ELSE
           The [I]th record is a nested
            data record
      END
  WHILE (EOF)
   }
```

Fig. 8
(a) Identified nested data record, No. of data fields=12
(b) Identified flat data record, No. of data fields = 7

The Fig. 8 shows the identified nested and flat data records. In Fig.8 (a), the number of data fields is 12 and in Fig.8 (b) the number of data fields is 7. The number of data fields in Fig 8(a) is 58.3% more than the number of data fields in Fig 8(b).

5. Empirical evaluation and experimental results

In this section, we evaluate the proposed **ENDR** (**E**xtraction of Flat and **N**ested **D**ata **R**ecords from Web Pages) technique. We compare it with the state-of-the-art existing system NET by Bing and Yanhong (2005). We do not compare it with DEPTA by Zhai, and Liu (2005) here as it is shown that NET is better than DEPTA. For flat nested data records, the proposed method performs very well. The experimental results are given in Table 1.

Column 1 lists the site of each test page. Due to the space limitations, we have not listed all the URL's considered for experimentation. We have not considered many erroneous pages for testing because such pages are relatively rare and quite difficult to find.

Column 2 and 4 give the number of data items extracted wrongly (Wr) by NET and proposed method from each page respectively. In x/y, x is the number of extracted results that are incorrect and y is the number of results that are not extracted. Columns 3 and 5 give the numbers

of correct (Corr) data items extracted by NET and proposed method from each page respectively. Here, in x/y, x is the number of correct items extracted and y is number of items in the page. From the table, we observe that, for flat and nested data records, the proposed method performs better than the other. The precision and recalls are computed based on extraction performed on all test pages.

URL	NET		ENDR	
	Wr.	Corr.	Wr.	Corr.
Without Nesting				
http://www.bookpool.com	0/0	15/15	0/0	15/15
http://www.amazon.com	0/0	22/22	0/0	22/22
http://www.shopping.com	0/0	20/20	0/0	20/20
http://www.barnesandnobles.com	0/0	10/10	0/0	10/10
http://www.cooking.com	0/0	28/29	0/0	29/29
http://tigerdirect.com	0/0	12/14	0/0	13/14
http://www.kmart.com	0/0	70/70	0/0	70/70
Recall	**97.15%**		**98.99%**	
Precision	**99.3%**		**98.92%**	
With Nesting				
http://www.amazon.com	0/0	22/25	0/0	25/25
http://kmart.com	1/0	42/43	0/0	43/43
http://www.cooking.com	1/0	62/63	0/0	62/63
Recall	**98.63%**		**100%**	
Precision	**99%**		**100%**	

Table 1: Experimental Results

6. Conclusion

In this paper, we have proposed a more effective technique to perform the automatic extraction of flat and nested data records from the web pages. Given a web page, the proposed method first identifies correct data region based on visual clue information. It counts the number of the data items in each record and then identifies the record as either flat or nested. Although the problem has been studied by several researchers, existing techniques are either inaccurate or make many strong assumptions. The proposed method is a pure visual clue based extraction of flat and nested data records. Experimental results show that the method performs data extraction more effectively.

7. References

A. Arasu, H. Garcia-Molina, (2003): Extracting structured data from web pages, ACM SIGMOD 2003,

Baeza Yates(1989): R. Algorithms for string matching: A survey. ACM SIGIR Forum, 23(3-4):34—58.

Benchalli, Hiremath, Siddu and Renuka 2005): "Mining Data Regions from Web Pages" , COMAD2005b, DEC.

Bing Liu , Kevin chen-chuan chang(2002): Editorial: Special issue on web content mining, WWW 02.

Bing Liu and Yanhong Zhai(2005): "NET - A System for Extracting Web Data from Flat and Nested Data Records." Proceedings of 6th International Conference on Web Information Systems Engineering (WISE-05).

Cai, D., Yu, S., Wen, J.-R. and Ma, W.-Y. (2003). Extracting Content Structure for Web Pages based on Visual Representation, Asia Pacific Web Conference (APWeb 2003), pp. 406417.

Chang, C-H., Lui, S-L(2001): IEPAD Information Extraction Based on Pattern Discovery. WWW-01.

Crescenzi, V., Mecca, G. and Merialdo, P, (2001): ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites. VLDB-01.

D. Buttler, L. Liu, C. Pu. (2001): A Fully Automated Object Extraction System for the World Wide Web. International Conference on Distributed Computing Systems (ICDCS 2001).

D. Embley, Y. Jiang, and Y. K . Ng(1999): Record-boundary discovery in Web documents. ACM SIGMOD Conference.

Eying, H. Zhang, (2001): HTML Page Analysis based on Visual Cues. 6th International Conference on Document Analysis and Recognition.

H. Zhao, W. Meng, Z. Wu, Raghavan (2005): Clement Yu. Fully Automatic Wrapper Generation For Search Engines, International WWW conference 2005, May 10-14, Japan. ACM 1-59593-046-9/05/005

J. Hammer, H. Garcia Molina, J. Cho, and A. Crespo (1997): Extracting semi-structured information from the web.In Proc.of the Workshop on the Management of Semi-structured Data.

J. Wang, F. H Lochovsky (2003): Data Extraction and Label Assignment for Web Databases.WWW conference,.

Kushmerick, N(2000): Wrapper Induction Efficiency and Expressiveness. Artificial Intelligence, 118:15-68, Clustering-based Approach to Integrating Source Query]

Lerman, K., Getoor L., Minton, S. and Knoblock, C(2004): Using the Structure of Web Sites for Automatic Segmentation of Tables. SIGMOD'04.

Liu, B., Grossman, R. and Zhai, Y(2003): Mining Data Records in Web Pages. KDD-03.

Zhai, Y., Liu, B(2005): Web Data Extraction Based on Partial Tree Alignment , WWW-05, 2005, May 10-14, , Chiba, Japan. ACM 1-59593-046-9/05/00

Tracking the Changes of Dynamic Web Pages in the Existence of URL Rewriting

Ping-Jer Yeh **Jie-Tsung Li** **Shyan-Ming Yuan**

Department of Computer Science
National Chiao Tung University
1001 Ta Hsueh Road, Hsinchu 300, Taiwan
Email: {pjyeh, jtli, smyuan}@cis.nctu.edu.tw

Abstract

Crawlers in a knowledge management system need to collect and archive documents from websites, and also track the change status of these documents. However, the existence of URL rewriting mechanism raises a page tracking problem since the URLs of a pair of dynamic page instances obtained during different sessions will no longer be the same. This paper proposes a series of algorithms in a bottom-up manner to find the corresponding pairs of dynamic page instances, and then to judge the change status of them. Experiments showed that the performance was very good and the outcome was 100% accurate.

Keywords: URL rewriting, crawler, HTTP session, string matching.

1 Introduction

A knowledge management system (KMS) needs to collect and archive numerical and textual data from a variety of sources. It also tracks the status of the data so that users can be notified whenever any update has occurred since last time. Nowadays a growing number of data is exposed to the Internet, particularly the Web, e.g., newspapers, technical reports, business intelligence, patent databases, discussion forums, and stock prices. Consequently, the data acquisition module in the KMS (often called the crawler) has to consider the unique characteristics of Web pages, particularly dynamic pages.

This section introduces the challenges faced by crawlers when they try to track dynamic Web pages.

1.1 HTTP Session

Since HTTP is a stateless protocol (Fielding, Gettys, Mogul, Frystyk, Masinter, Leach & Berners-Lee 1999), there are some workaround mechanisms trying to simulate a stateful connection (or more formally, *session*) on top of HTTP. Among them, the *cookie* mechanism (Kristol & Montulli 2000, Kristol 2001) is a popular way to enable a series of stateful request-response interactions between Web clients and servers. It is very easy to use, but it is also very prone to abuse, snoop, and attack (Sit & Fu 2001). Therefore, a safer way to use cookies is proposed. That is, to encode and transmit only the session identifier (abbreviated as *session-id* or *sid*) during the lifetime of the session (Hallam-Baker 1996).

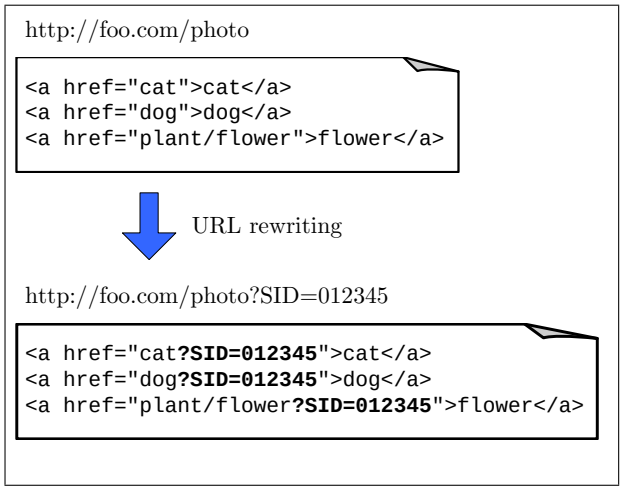


Figure 1: An example illustrating the URL rewriting effect

The use of cookie + session-id combination is now a de facto or default session mechanism in major Web servers and server-side script engines. For example, Java servlets/JSP generates a `JSESSIONID` cookie to carry the session-id (Coward & Yoshida 2004). By default the cookie is named `ASP.NET_SessionId` in ASP.NET, `PHPSESSID` in PHP, and `_session_id` in Ruby on Rails.

Despite the simplicity and popularity, the bad reputation of cookies has made many power users disable the cookie function of Web browsers entirely, at the risk of disabling the subsequent session tracking at the same time, though. To solve this anti-cookie dilemma, another transparent way to place and transmit the session-id is devised. When a new session starts, the server generates a corresponding session-id and then, before sending the dynamic page back to the client, inserts the session-id as a part of relevant URLs rather than places it in the cookies. This so-called *URL rewriting* step is applied on the fly to all subsequent URL links in the dynamic pages without the page author's awareness. Take Figure 1 for example. In the beginning a user requests the dynamic page `http://foo.com/photo`. If the session-id is `012345`, the server will rewrite all relevant links by inserting into them some formatted string like `SID=012345`. From the user's point of view, all links in the dynamic page have such session-id string embedded, and therefore all subsequent navigation through the page will naturally carry the same session-id information. In this way the server knows which session to track for.

Different server-side script engines rewrite URLs in quite different ways. Take again the "cat" link in Figure 1 for example. In Java servlets/JSP the link would be rewritten as

`http://foo.com/photo/cat;jsessionid=`*sid* (Coward & Yoshida 2004). By default the link would become `http://foo.com/photo/(`*sid*`)/cat` in ASP.NET's cookieless mode, `http://foo.com/photo/cat?PHPSESSID=`*sid* in PHP's transparent sid mode.

1.2 Challenges of Changing URLs

The use of URL rewriting raises a problem for crawlers. For a crawler to determine whether a page has been changed since last time, it has to be sure that it can access and identify the same page source, albeit some parts of the page content may not be identical. For example, if the content of page p was first p_{v1} and later changed to p_{v2}, the crawler has to identify that both p_{v1} and p_{v2} are two instances/versions of the same p. With the existence of URL rewriting, however, instances p_{v1} and p_{v2} are retrieved through different session-ids, and so do their URLs $l(p_{v1})$ and $l(p_{v2})$. The crawler would, therefore, have difficulty in identifying that $l(p_{v1})$ and $l(p_{v2})$ eventually correspond to the same p at different time.

The purpose of this paper is to propose a set of algorithms in a bottom-up manner to find the corresponding pairs of dynamic Web page instances obtained during different sessions, and then to judge which pages remain the same, and which pages have been updated or removed entirely since last time.

2 Problem Formulation

To put it more formally, the highest level algorithm in this paper is as follows. Given two pools of rewritten URLs L and L' collected recursively from the same starting point (base URL b) during two different sessions, we try to find

1. every l_i and l'_j pair that corresponds to the same page source p_i, where $l_i \in L$ and $l'_j \in L'$;
2. the list of page links L'_{upd} whose page contents have been changed since last time;
3. the list of page links L'_{sam} whose page contents remain unchanged;
4. the list of page links L'_{new} whose pages never appear last time; and
5. the list of page links L'_{err} whose page contents cannot be accessed this time.

In order to achieve the first goal, we also need to find in advance the session-ids s embedded in L and s' in L'.

For brevity, the following symbols are used throughout the whole paper.

- b: base URL.
- l: URL link; converted to the absolute form for simplicity.
- L: the set of links l_i where $i \in \{1..n\}$.
- L and L': two set of links obtained from the same b in distinct sessions, which can happen simultaneously or at different time.
- p: Web page.
- p_{v1}, p_{v2}: instances/versions of p obtained during two distinct sessions.
- s: session-id.
- s and s': two instances of session-ids obtained from the same b in distinct sessions, which can happen simultaneously or at different time.

3 Finding the Session-Id in the URL Set

This section discusses how to find the session-id string in the URL set L collected recursively from the starting point b during one session. First we need to identify some important properties of session-ids that are useful in designing the algorithms. After that we introduce a series of subroutines to enumerate the session-id candidates, and finally design an algorithm to determine the exact session-id among these candidates.

3.1 Properties of Session-Ids

A dynamic Web page usually has not only static links that just point to some locations with no parameters but also dynamic links. Every dynamic link may consist of the location, parameters, and the session-id if URL rewriting is enabled. It follows that the average length of static links is usually shorter than that of dynamic links within the same website.

When a session-id is encoded and embedded within the URL, its valid character set is inevitablly constrained by that of the URL. According to the URL specifications (Berners-Lee, Masinter & McCahill 1994, Berners-Lee, Fielding & Masinter 2005), some special characters (e.g., '`;`', '`/`', '`?`', '`=`', '`&`', '`%`', and '`#`') cannot be used directly within the ⟨*scheme-specific-part*⟩; the same restriction applies to embedded session-ids as well. Consequently we can say that the next character following the session-id must be either a special character or an end-of-string (EOS) mark.

Although Web servers have much freedom to implement the session-ids, for security reasons most of them use some cryptographically secure one-way functions to construct the session-ids (Hallam-Baker 1996, Gutterman & Malkhi 2005). It follows that all session-ids on the same website normally have a fixed string length (e.g., 128 bits), that they seldom collide, and that their string contents are usually unlikely to appear elsewhere.

The URL rewriting mechanism also has another feature by nature. Since the session-id s remains identical during the same session, all dynamic pages' links in L should carry the s with themselves. It means that the s must be a substring of all dynamic links in L.

To sum up, session-ids have the following properties which are crucial to designing algorithms:

Property 1. *The session-id remains unchanged during the same session.*

Property 2. *Session-ids on the same website normally have a fixed length, even in distinct sessions.*

Property 3. *Session-ids on the same website seldom collide. It is also very unlikely to see more than one occurrence of the same session-id string literal within the body of the URL when URL rewriting mechanism is activated.*

Property 4. *When URL rewriting mechanism is activated, the session-id is embedded in the URL, and the next character following the session-id must be either a special character or an EOS mark.*

Property 5. *When URL rewriting mechanism is activated, the session-id must be one of the common substrings among all dynamically-generated URLs.*

Property 6. *When URL rewriting mechanism is activated, dynamic links are usually longer than the static ones within the same website.*

1. `http://www.amazon.com/exec/obidos/subst/home/home.html/ref=three_tab_gw/002-9355727-0611208`
2. `http://www.amazon.com/exec/obidos/tg/browse/-/229220/ref=gw_subnav_gft/002-9355727-0611208?%5Fencoding=UTF8`
3. `http://www.amazon.com/exec/obidos/tg/stores/static/-/gateway/international-gateway/ref=gw_subnav_in/002-9355727-0611208?%5Fencoding=UTF8`
4. `http://www.amazon.com/exec/obidos/tg/new-for-you/new-releases/-/main/ref=gw_subnav_nr/002-9355727-0611208?%5Fencoding=UTF8`

Figure 2: An excerpt of URL rewriting results taken from Amazon.com; full data is shown in Figure 6

3.2 Determining the Fixed Length of Session-Ids

The session-id must be one of the common substrings among all dynamically-generated URLs (see Property 5), but not all common substrings are the session-id. Take Figure 2 for example. At first glance many common substrings can be found, even if special characters (considering Property 4) are excluded:

- `exec`
- `obidos`
- `ref`
- `002-9355727-0611208`

But which one is the correct session-id? It seems that they all satisfy the properties listed in Section 3.1. It also seems difficult to generalize a set of syntactic rules to extract the correct session-id out of them.

Let's look at the session-id problem from yet another angle. The syntactic rules that synthesize the rewritten URLs may be complex and subject to changes, but the mechanism that generates the session-id itself is seldom subject to changes. Therefore, it seems easier to base our problem-solving strategy on the algorithmic nature of the session-id.

To find the right common substring as the session-id, first we need to determine the length of the valid session-id. Since the length on the same website normally remains the same (see Property 2), its value can be determined either by human inspection or by an easy-to-program yet effective heuristic. We can write a program to make two distinct but concurrent connections to the same website. The pages fetched in the two nearly-concurrent sessions would have almost identical layouts, topology, and contents; except for the links L and L' (or more precisely, session-ids s and s'). A simple diff-like comparison (Myers 1986) between L and L' would reveal the fixed length of session-ids on this website.

The probe process needs to be done only once for each website unless the website changes its configuration, for example, server version, script engine, or session-key generation algorithm.

3.3 Finding the Common Substrings of the Given Length Between Two Strings

Before digging into the details of discovering the fixed-length common substrings among a set of links L, let's first discuss the basic subroutine of finding them between just two links.

The main idea of FIND-CS(a, b, k) algorithm in Figure 3 is discussed as follows. Since the session-id is a common substring in both strings a and b,

Algorithm: FIND-CS(a, b, k)
Input: a, b: two strings to be scanned
k: length of substring(s) to be found
Output: R: the set of all k-length common substrings between a and b

```
Begin:
1    R ← {}
2    C ← the set of URL special characters
3    ▷ for each substring of length k in b
4    for i ← 1 to length(b) − k + 1 do
5        t ← substring b[i..i + k − 1]
6        if t contains no c ∈ C
           and b[i + 1] ∈ C ∪ {EOS} then
7            ▷ try to find t in a
8            STRING-MATCHING(a, t)
9            if found then
10               R ← R ∪ {t}
11           end if
12       end if
13   end for
14   return R
```

Figure 3: Algorithm for finding the common substrings of the given length between two strings

we can find all session-id candidates R by extracting every well-formed substring t_i from b and then trying to see if t_i is also in a. Every t_i has to, of course, be selected according to the constraint mentioned in Property 4. Its length k is obtained in the way stated previously in Section 3.2, and is specified as input to the FIND-CS algorithm.

Line 8 invokes some string-matching algorithm to see if t_i is also in a. Well-known algorithms such as the Knuth-Morris-Pratt (KMP) algorithm and the Boyer-Moore algorithm can be used here (Cormen, Leiserson, Rivest & Stein 2001).

3.4 Finding the Exact Session-Id

Just by knowing the exact length of session-ids (see Section 3.2) will not eliminate entirely the problem of finding the exact session-id strings. Take Figure 4 for example. There are two very long common substrings: the length of "`1PS2V6KKYBZ34F3RK1PJ`" is 20 and "`002-9355727-0611208`" is 19. Even if we are sure that the length of session-ids to be found is 19, and therefore "`1PS2V6KKYBZ34F3RK1P`" (notice that the tailing character '`J`' is excluded) is ruled out by Property 4, FIND-CS$(a, b, 19)$ still discovers more than one candidate:

- `002-9355727-0611208`
- `PS2V6KKYBZ34F3RK1PJ` (notice that the beginning character '`1`' is excluded)

Just by looking at the two links a and b is not enough; it gives us little information about which candidate is the session-id. All we need is to look at the whole set of links. The complete algorithm for finding session-ids in the whole set is listed in Figure 5. Let's explain the rationale behind the algorithm.

A dynamic Web page typically has both static and dynamic-generated links in it. Obviously sesson-ids can exist only in the dynamic ones. Therefore we need to seperate the dynamic links from the static ones and focus on the former.

However, there is no general rules to distinguish both types of links literally. Two properties can help us solve this problem. Dynamic links tend to be longer than the static ones (according to Property 6),

```
a. http://www.amazon.com/gp/amabot/?pf_rd_url=http%3A%2F%2Fwww
   .amazon.com%3A80%2Fgp%2Fredirect.html%2Fref%3Dgf_gw_wine%2F
   002-9355727-0611208%3Flocation%3Dhttp%3A%2F%2Fwww.wine.com%
   2Fpromos%2Famazonwine.asp%253Fan%253Damazon%2526s%253Damazo
   n%2526cid%253Damazon%255Fgateway%255Ffpbox%26token%3D21513C
   47B07C95040C8597937CAB64718E97425C&pf_rd_p=163187901&pf_rd_
   s=left-nav&pf_rd_t=101&pf_rd_i=507846&pf_rd_m=ATVPDKIKX0DER
   &pf_rd_r=1PS2V6KKYBZ34F3RK1PJ
b. http://www.amazon.com/gp/amabot/?pf_rd_url=http%3A%2F%2Fwww
   .amazon.com%3A80%2Fgp%2Fredirect.html%2Fref%3Damb_link_6494
   02_2%2F002-9355727-0611208%3Flocation%3Dhttp%3A%2F%2Fwww.am
   azon.com%2Fgp%2Fsearch.html%2F%253Fplatform%253Dgurupa%2526
   url%253Dnode%253D11055981%2526keywords%253DT3%26token%3DD2E
   7DF9A889DD105A02CC33159540A6D52DE4D8B&pf_rd_p=160346101&pf_
   rd_s=right-4&pf_rd_t=101&pf_rd_i=507846&pf_rd_m=ATVPDKIKX0D
   ER&pf_rd_r=1PS2V6KKYBZ34F3RK1PJ
```

Figure 4: Another excerpt of URL rewriting results taken from Amazon.com; full data is shown in Figure 6

Algorithm: FIND-SID(L, k)
Input: L: pool of links
k: length of session-id string to be found
Output: session-id string of length k
Begin:
1 ▷ find session-id candidates V
2 $M \leftarrow$ clone of L
3 **do**
4 $t \leftarrow$ longest string in M
5 $u \leftarrow$ second longest string in M
6 remove t and u from M
7 $V \leftarrow$ FIND-CS(t, u, k)
8 **while** $V = \{\}$
9 ▷ calculate the number of occurrences for each $v \in V$ in M
10 D: associative array mapping from string to integer
11 **for each** $v \in V$ **do**
12 $D[v] \leftarrow 0$
13 **for each** $m \in M$ **do**
14 **if** v is a substring of m **then**
15 $D[v] \leftarrow D[v] + 1$
16 **end if**
17 **end for**
18 **end for**
19 ▷ identify the right session-id among candidates V
20 $s \leftarrow \arg\max_v(D[v])$
21 **return** s

Figure 5: Algorithm for finding the exact session-id

and therefore Lines 4–5 pick the longest links t and u as a guess. Moreover, the session-id is one of the common substrings of these dynamic links (according to Property 5), and therefore Line 7 calls the FIND-CS algorithm to enumerate the candidates V. If no candidate is found, implying that our previous guess was wrong, the surrounding loop in Lines 3–8 is responsible for making a second guess (or more, if necessary).

At present V may also contain some non-session-id strings by chance, as Figure 4 has showed. A property can help us distinguish the real session-id from the others. According to Property 5, the session-id string should appear in *all* dynamic links, implying that the most frequent candidate is more likely to be the real session-id. Therefore Lines 10–18 try to find the frequency of each candidate in the whole links (albeit not in the *dynamic* links), and Line 20 picks the most frequent one as the session-id.

3.5 A Complete Example

Let's use a complete example to demonstrate the whole scenario of finding the session-id from the start. The example was taken from Amazon.com in the middle of March, 2006. The base URL b was `http://www.amazon.com` and the traversal depth was 0. Seventeen links were obtained and shown in Figure 6.

Let L denote the whole set of 17 links $l_1..l_{17}$. By the probe process mentioned in Section 3.2, the length of session-id is found to be 19 on this website. Lines 4–5 in FIND-SID($L, 19$) would first pick the longest two links: l_9 and l_{10}. l_9 and l_{10} equal to a and b in Figure 2, and therefore FIND-CS($l_9, l_{10}, 19$) in Line 7 outputs the same results mentioned in the first paragraph of Section 3.4: `002-9355727-0611208` and `PS2V6KKYBZ34F3RK1PJ`. Afterwards Lines 10–18 calculate the frequency of them as 14 and 5, respectively. Consequently the most frequent one, `002-9355727-0611208`, is decided as the session-id.

3.6 Discussion

With these algorithms, we can easily find the session-id among a collection of URLs. But the mechanism bases on some assumptions and therefore has some limitations.

Many aspects of these algorithms rely on the probability that Property 6 holds. If dynamic links are shorter than the static ones on some websites, Lines 4–5 of the FIND-SID algorithm would be in the wrong direction. In addition, if the number of dynamic links is too small, Line 13 would cover too

1. http://www.amazon.com/exec/obidos/subst/home/home.html/ref=three_tab_gw/002-9355727-0611208
2. http://www.amazon.com/exec/obidos/tg/browse/-/229220/ref=gw_subnav_gft/002-9355727-0611208?%5Fencoding=UTF8
3. http://www.amazon.com/exec/obidos/tg/stores/static/-/gateway/international-gateway/ref=gw_subnav_in/002-9355727-0611208?%5Fencoding=UTF8
4. http://www.amazon.com/exec/obidos/tg/new-for-you/new-releases/-/main/ref=gw_subnav_nr/002-9355727-0611208?%5Fencoding=UTF8
5. http://www.amazon.com/exec/obidos/tg/new-for-you/top-sellers/-/main/ref=gw_subnav_ts/002-9355727-0611208?%5Fencoding/=UTF8
6. http://www.amazon.com/gp/amabot/?pf_rd_url=%2Fgp%2Fbrowse.html%2Fref%3Dgw_br_el_feat%2F002-9355727-0611208%3F%255Fencoding%3DUTF8%26node%3D172282&pf_rd_p=163187901&pf_rd_s=leftnav&pf_rd_t=101&pf_rd_i=507846&pf_rd_m=ATVPDKIKX0DER&pf_rd_r=1PS2V6KKYBZ34F3RK1PJ
7. http://www.amazon.com/gp/amabot/?pf_rd_url=%2Fgp%2Fbrowse.html%2Fref%3Dgw_br_gf%2F002-9355727-0611208%3F%255Fencoding%3DUTF8%26node%3D3370831&pf_rd_p=163187901&pf_rd_s=left-nav&pf_rd_t=101&pf_rd_i=507846&pf_rd_m=ATVPDKIKX0DER&pf_rd_r=1PS2V6KKYBZ34F3RK1PJ
8. http://www.amazon.com/gp/amabot/?pf_rd_url=%2Fgp%2Fbrowse.html%2Fref%3Dgw_br_jw%2F002-9355727-0611208%3F%255Fencoding%3DUTF8%26node%3D3367581&pf_rd_p=163187901&pf_rd_s=left-nav&pf_rd_t=101&pf_rd_i=507846&pf_rd_m=ATVPDKIKX0DER&pf_rd_r=1PS2V6KKYBZ34F3RK1PJ
9. http://www.amazon.com/gp/amabot/?pf_rd_url=http%3A%2F%2Fwww.amazon.com%3A80%2Fgp%2Fredirect.html%2Fref%3Dgf_gw_wine%2F002-9355727-0611208%3Flocation%3Dhttp%3A%2F%2Fwww.wine.com%2Fpromos%2Famazonwine.asp%253Fan%253Damazon%2526s%253Damazon%2526cid%253Damazon%255Fgateway%255Ffpbox%26token%3D21513C47B07C95040C8597937CAB64718E97425C&pf_rd_p=163187901&pf_rd_s=left-nav&pf_rd_t=101&pf_rd_i=507846&pf_rd_m=ATVPDKIKX0DER&pf_rd_r=1PS2V6KKYBZ34F3RK1PJ
10. http://www.amazon.com/gp/amabot/?pf_rd_url=http%3A%2F%2Fwww.amazon.com%3A80%2Fgp%2Fredirect.html%2Fref%3Damb_link_649402_2%2F002-9355727-0611208%3Flocation%3Dhttp%3A%2F%2Fwww.amazon.com%2Fgp%2Fsearch.html%2F%253Fplatform%253Dgurupa%2526url%253Dnode%253D11055981%2526keywords%253DT3%26token%3DD2E7DF9A889DD105A02CC33159540A6D52DE4D8B&pf_rd_p=160346101&pf_rd_s=right-4&pf_rd_t=101&pf_rd_i=507846&pf_rd_m=ATVPDKIKX0DER&pf_rd_r=1PS2V6KKYBZ34F3RK1PJ
11. http://www.amazon.com/exec/obidos/redirect-to-external-url/002-9355727-0611208?%5Fencoding=UTF8&path=http%3A%2F%2Fwww.amazon.co.uk%2Fexec%2Fobidos%2Fredirect-home%3Ftag%3Dintl-usstoreh-ukhome-21%26site%3Damazon
12. http://www.amazon.com/exec/obidos/redirect-to-external-url/002-9355727-0611208?%5Fencoding=UTF8&path=http%3A%2F%2Fwww.amazon.de%2Fexec%2Fobidos%2Fredirect-home%3Ftag%3Dintl-usstoreh-dehome-21%26site%3Dhome
13. http://www.amazon.com/exec/obidos/redirect-to-external-url/002-9355727-0611208?%5Fencoding=UTF8&path=http%3A%2F%2Fwww.amazon.co.jp%2Fexec%2Fobidos%2Fredirect-home%3Ftag%3Dintl-usstoreh-jphome-22%26site%3Damazon
14. http://www.amazon.com/exec/obidos/redirect-to-external-url/002-9355727-0611208?%5Fencoding=UTF8&path=http%3A%2F%2Fwww.amazon.fr%2Fexec%2Fobidos%2Fredirect-home%3Fsite%3Damazon%26tag%3Dusfr-storeh-footer-21
15. http://ec1.images-amazon.com/images/G/01/marketing/visa/amzn_visa-save-30-today.gif
16. http://ec1.images-amazon.com/images/G/01/gateway/partner-logos/target_logo._V58636053_.gif
17. http://images.amazon.com/images/P/6305428050.01._PE30_SCTZZZZZZZ_.jpg

Figure 6: A full example of URL rewriting results taken from Amazon.com with the traversal depth of 0

many irrelevant links and consequently contribute too many counts to the wrong candidates.

A possible workaround to these limitations is to increase the breadth and depth of traversal in the beginning. In this way, every single wrong candidate may have less chance to have higher frequency than the real session-id does. More evaluations are to be done in Section 5.1.

4 Tracking the States of Web Pages

4.1 Algorithm

Based upon previous subroutines, we are ready to track the change states of Web pages in the existence of URL rewriting. The tracking algorithm is shown in Figure 7. Inside it, two subroutines (CRAWL and GET-NEWEST-DOCUMENT) are left as implementation details.

In the beginning the crawler's repository has a record containing a base URL b and a set of pages (referred to by L) traversed and collected from b last time. The traversal breadth and depth are determined according to some pre-given values or inclusion/exclusion rules.

Now we would like to know if these pages have any change since last time. At first Line 4 crawls the documents starting from b in the same way that L was obtained last time.

To compare the change states between pages referred to by L and L', their session-ids have to be unified to the same ground in advance. It is obvious that if pages from L still exist in the meantime, their links would remain identical except for the session-id part. Therefore Lines 8–10 replace the old s in L with the new s'. The replacement is safe because session-id strings are meant to be unique; they seldom appear elsewhere in the URL string (see Property 3). Normally we do not need to worry about the risk of corrupting the non-session-id parts.

When the unification is done, the relation between two sets L and L' can be illustrated in Figure 8. The case for L'_{new} is so trivial that Line 11 sets it as $L' - L$ accordingly.

The case for $L - L'$ is a little tricky. It may due to access error (Lines 15–17); it may also due to the orphan phenomenon. A page is called an orphan if it existed before and still exists at present, but for some reasons the links to it disappear so that it is not recursively reachable via b anymore. The contents of

Algorithm: TRACK(b, L, k)
Input: b: base URL
L: pool of links collected last time
k: length of session-id string
Output: L'_{sam}: current links; content unchanged since last time
L'_{upd}: current links; content updated since last time
L'_{new}: current links; never appeared last time
L'_{err}: current links; fail to access this time
Begin:
1 $L'_{\text{sam}} \leftarrow \{\}$
2 $L'_{\text{upd}} \leftarrow \{\}$
3 $L'_{\text{err}} \leftarrow \{\}$
4 $L' \leftarrow$ CRAWL(b)
5 ▷ update the old session-id with the new one
6 $s' \leftarrow$ FIND-SID(L', k)
7 $s \leftarrow$ FIND-SID(L, k)
8 **for each** $l \in L$ **do**
9 replace the substring s in l with s', if any
10 **end for**
11 $L'_{\text{new}} \leftarrow L' - L$
12 ▷ check: orphan or error?
13 $O \leftarrow \{\}$ ▷ orphan
14 **for each** $l \in L - L'$ **do**
15 $p_{\text{v2}} \leftarrow$ GET-NEWEST-DOCUMENT(l)
16 **if** p_{v2} is null **then**
17 $L'_{\text{err}} \leftarrow L'_{\text{err}} \cup \{l\}$
18 **else**
19 $O \leftarrow O \cup \{l\}$
20 **end if**
21 **end for**
22 ▷ check: same or updated?
23 **for each** $l' \in (L \cap L') \cup O$ **do**
24 $p_{\text{v1}} \leftarrow$ previous version of l' page content in the repository
25 $p_{\text{v2}} \leftarrow$ current version of l' page content in the repository
26 **if** $p_{\text{v1}} = p_{\text{v2}}$ **then**
27 $L'_{\text{sam}} \leftarrow L'_{\text{sam}} \cup \{l'\}$
28 **else**
29 $L'_{\text{upd}} \leftarrow L'_{\text{upd}} \cup \{l'\}$
30 update p_{v1} with p_{v2} in the repository
31 **end if**
32 **end for**
33 **return** L'_{sam}, L'_{upd}, L'_{new}, L'_{err}

Figure 7: Algorithm for tracking the states of Web pages

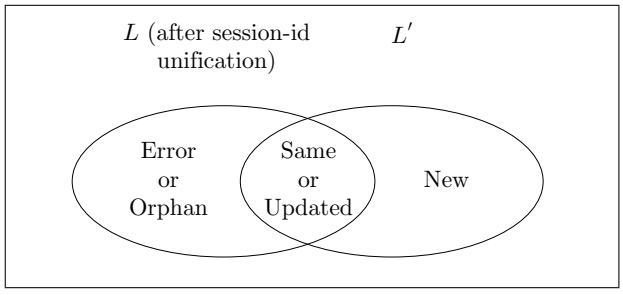


Figure 8: Relation between two sets of links L and L' (after session-id unification)

such orphans may, of course, remain the same or have been updated since last time. Consequently Lines 13–21 try to identify them for Lines 23–32 to process.

The main loop in Lines 23–32 is obvious. It tries to validate every link in $(L \cap L)$ plus orphans to decide whether the page has been updated or not.

4.2 Implementation Hints

There are some other things that are important for implementing the TRACK algorithm.

If there are a large proportion of static pages in the collection, the CRAWL subroutine in Line 4 can be optimized further. When it tries to fetch a document from a remote website, the `If-Modified-Since` request header (Fielding et al. 1999) can be used to avoid retransmission of the unchanged content.

Web servers and server-side script engines have a timeout value for sessions. For example, by default the timeout is 60 minutes in Tomcat (an open-source Java servlets/JSP container), 20 minutes in ASP.NET, and 24 minutes in PHP. But the timeout value is reconfigurable, especially on e-commerce websites. Therefore implementation of TRACK should consider this issue and should not let the session idle

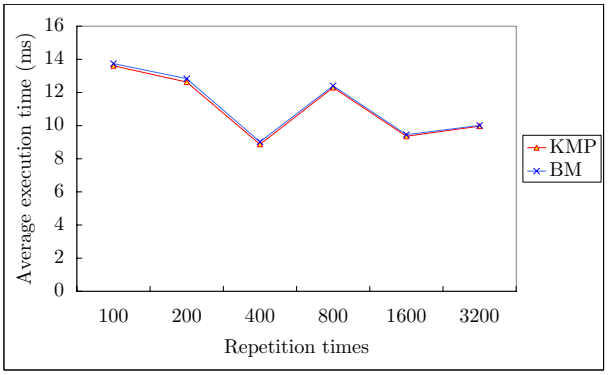


Figure 9: Performance of FIND-SID algorithm on Amazon.com. The experiment was done on Pentium 4 2.8 GHz running Mandriva Linux 2006 and JDK 1.4.2_11-b06. In the figure, KMP stands for the Knuth-Morris-Pratt algorithm, and BM stands for the Boyer-Moore algorithm.

too long between Line 4 and the first invocation of Line 15.

Roughly speaking, the idle time would be proportional to the execution time for FIND-SID since the time complexity for Line 6 is identical to that for Line 7, and the time complexity for Lines 8–10 is a little lower. It seems a good place to use a real-world example to evaluate the performance of FIND-SID. The experiment was conducted by repeating the one mentioned in Section 3.5 several times. Both Knuth-Morris-Pratt and Boyer-Moore algorithms were employed for comparison. The time spent in network transmission was excluded from the result.

As can be seen in Figure 9, even for such a website with very long dynamic links and session-ids, the FIND-SID algorithm took about 10 milliseconds—much shorter than typical session idle time, we think.

If performance pursuers still worry about the session idle time, it can be reduced to about 2/3 by moving Line 7 to any location before Line 4. Line 11 can also be moved to any location after Lines 13–21 or Lines 23–32 since it has no side effect on the latter two blocks. Finally, the loop in Lines 14–21 can be executed on another concurrent thread to be triggered by Line 9.

5 Experimental Results

To validate the effectiveness of these algorithms, we conducted a set of experiments in a bottom-up manner.

5.1 Correctness of FIND-SID Algorithm

To test the correctness of FIND-SID algorithm, we first try to investigate the number of candidates found by FIND-CS algorithm, and also the recall of the finding. Finally we investigate the correctness of FIND-SID algorithm.

The experiments were conducted on 4 websites, with the traversal depth of 1. Each was repeated 100 times.

- A: `http://www.amazon.com/` and the length of session-id = 19.
- B: `http://www.amazon.co.jp/` and the length of session-id = 19.
- C: `http://webcaspar.nsf.gov/` and the length of session-id = 32.

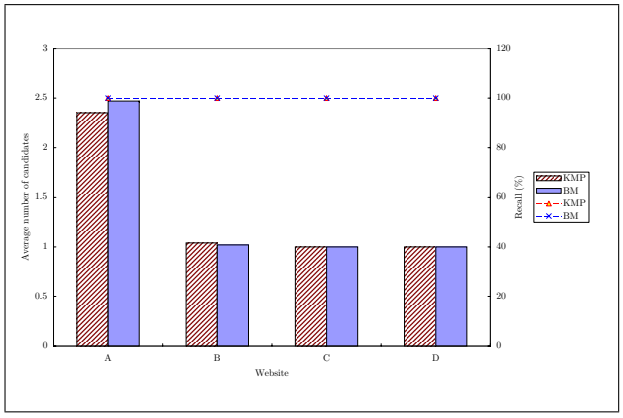


Figure 10: Correctness of FIND-CS algorithm on websites A, B, C, and D, as mentioned in Section 5.1. The bar chart and the y-axis on the left-hand side show the average number of candidates found, while the line graph and the y-axis on the right-hand side show the average recall of session-ids.

- D: `http://www.jesishpettsburgh.org/` and the length of session-id = 32.

First, Figure 10 shows that the average numbers of candidates were very small, and the recalls were 100% all the time. The outcome was very accurate.

The next step is to evaluate the correctness of FIND-SID. The outcome was all 100% correct, so the figure is omitted.

5.2 Correctness of TRACK Algorithm

The experiments were conducted on 2 portfolios of Web pages. The tracking step was done within a half hour after the first crawling step. It was to avoid too few intersections and too many missing pages since these websites changed their contents very frequently.

- A: `http://www.amazon.com/` with depth = 1.
- B: composed of 3 subtasks:
  - B_1: `http://tw.news.yahoo.com/finance/` with depth = 2,
  - B_2: `http://tw.news.yahoo.com/technology/` with depth = 2,
  - B_3: search `http://tw.news.yahoo.com/` with keyword "google".

Bitwise comparison between L and L' pairs showed that the outcome in Table 1 was 100% accurate. Close inspection of the pairs revealed further that varying advertisements and embedded JavaScript codes contributed to a smaller L'_{sam}. If we want a more tolerant L'_{sam}, i.e., not necessarily bitwise identical, the "equality" test in Line 26 of the TRACK algorithm should be relaxed to allow for content-based filtering and customization.

6 Conclusion and Future Work

The URL rewriting mechanism is a popular way to enable sessions between Web clients and servers. However, the mechanism and the lack of regularity in the URLs produced have raised problems for knowledge management systems to track the changes of Web pages. This paper has devised a series of algorithms in a bottom-up manner, and also demonstrated that they are very effective and efficient in tracking dynamic Web pages in the existence of URL rewriting.

Table 1: Experimental results of the TRACK algorithm

	Portfolio A	Portfolio B
First crawling	2006-06-06 20:31	2006-06-07 14:26
L:	340	139
Tracking	2006-06-06 20:55	2006-06-07 14:40
L'_{sam}:	76	19
L'_{upd}:	264	120
L'_{new}:	54	3
L'_{err}:	0	0

This paper has focused mainly on the issue raised by the URL rewriting mechanism. In the future, we will move on to other practical issues when implementing the crawler module of a KMS. For example, it is important to investigate more proper ways to define and handle the "equality" of dynamic pages from the user's point of view. To do this, first we plan to incorporate the idea of AJAX-based screen-scraping toolkit so that users can specify the regions of interest in a more intuitive way. Second, we plan to provide content-based filtering so that users can customize their own equality test.

Another important issue is to investigate a sophisticated but also intuitive way for users to specify and refine a series of steps required to navigate into a specific subset of the target website. To do this, the same technique of AJAX-based screen scraping can also be helpful. We think that this approach is more finer-grained and easier-to-use than traditional command-line argument or HTML screen-scraping approaches.

7 Acknowledgments

This work was mainly supported by National Science Council grant NSC95-2752-E-009-PAE: advanced technologies and applications for next generation information networks, and partially supported under grant NSC94-2213-E-009-026 and NSC94-2520-S-009-004.

References

Berners-Lee, T., Fielding, R. & Masinter, L. (2005), Uniform Resource Identifier (URI): Generic syntax, RFC 3986, Internet Engineering Task Force.

Berners-Lee, T., Masinter, L. & McCahill, M. (1994), Uniform Resource Locators (URL), RFC 1738, Internet Engineering Task Force.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2001), *Introduction to Algorithms*, second edn, The MIT Press.

Coward, D. & Yoshida, Y. (2004), Java servlet specification version 2.4, JSR 154, Sun Microsystems.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. & Berners-Lee, T. (1999), Hypertext Transfer Protocol – HTTP/1.1, RFC 2616, Internet Engineering Task Force.

Gutterman, Z. & Malkhi, D. (2005), 'Hold your sessions: An attack on Java session-id generation', *Lecture Notes in Computer Science* **3376**, 44–57.

Hallam-Baker, P. M. (1996), Session identification URI, W3C working draft, World Wide Web Consortium.

Kristol, D. M. (2001), 'HTTP cookies: Standards, privacy, and politics', *ACM Transactions on Internet Technology* **1**(2), 151–198.

Kristol, D. & Montulli, L. (2000), HTTP state management mechanism, RFC 2695, Internet Engineering Task Force.

Myers, E. W. (1986), 'An $O(ND)$ difference algorithm and its variations', *Algorithmica* **1**(1), 251–266.

Sit, E. & Fu, K. (2001), 'Inside risks: Web cookies: not just a privacy risk', *Communications of the ACM* **44**(9), 120.

A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses

Faten Khalil, Jiuyong Li and Hua Wang

Department of Mathematics & Computing
University of Southern Queensland
Toowoomba, Australia, 4350,
Email: {khalil,jiuyong and wang}@usq.edu.au

Abstract

The importance of predicting Web users' behaviour and their next movement has been recognised and discussed by many researchers lately. Association rules and Markov models are the most commonly used approaches for this type of prediction. Association rules tend to generate many rules, which result in contradictory predictions for a user session. Low order Markov models do not use enough user browsing history and therefore, lack accuracy, whereas, high order Markov models incur high state space complexity. This paper proposes a novel approach that integrates both association rules and low order Markov models in order to achieve higher accuracy with low state space complexity. A low order Markov model provides high coverage with low state space complexity, and association rules help achieve better accuracy.

Keywords: Association rules, Markov models, prediction.

1 Introduction

The need to predict the next Web page to be accessed by the user is apparent in most Web applications today whether they are search engines or e-commerce solutions or mere marketing sites. Web applications today are driven to provide a more personalized experience for their users. Therefore, it is extremely important to form some kind of interaction with Web users and always be one step ahead of them when it comes to predicting next accessed pages. For instance, knowing the user browsing history on the site grants us valuable information as to which one of the most frequently accessed pages will be accessed next. Also, it provides us with extra information like the type of user we are dealing with and the users preferences as well. There are various ways that can help us make such a prediction, but the most common approaches are Markov models and association rules. Each of the approaches used for this purpose has its own weaknesses when it comes to accuracy, coverage and performance. Lower order Markov models lack accuracy because of the limitation in covering enough browsing history; whereas higher order Markov models usually result in higher state space complexity. On the other hand, association rules have the problem of identifying the one correct prediction out of the many rules that lead to a large number of predictions (Mobasher, Dai, Luo & Nakagawa 2001, Yang, Li, & Wang 2004). This paper proposes an improved approach, based on a combination of Markov models and association rules that results in better prediction accuracy and more coverage. We use low order Markov models to predict multiple pages to be visited by a user and then we apply association rules to predict the next page to be accessed by the user based on long history data.

This project was partially supported by Australian Research Council Discovery Grant DP0559090.

1.1 Related Work

The importance of Web usage mining has led to a number of research papers in the area. However, most of these papers were hindered by some kind of limitations. For instance, many of the papers proposed using association rules or Markov models for next page prediction, however, none of these papers have addressed the use of a combination of both methodologies. Some of the papers that proposed the use of association rules for better predicting the next page to be accessed by the user are (Mobasher et al. 2001, Spiliopoulou, Faulstich & Winkler 1999, Yong, Zhanhuai & Yang 2005); whereas, other papers like (Cadez, Heckerman, Meek, Smyth & White 2000, Deshpande & Karypis 2004, Dongshan & Junyi 2002, Garafalakis, Pastogi, Seshadri & Shim 1999, Gunduz & Ozsu 2003, Jespersen, Pedersen & Thorhauge 2003) covered Markov models.
Mobasher *et al.* (2001) were confronted with the problem of providing user personalisation at an early stage of the Web session. They proposed the use of collaborative filtering approaches like the k-Nearest Neighbour (KNN) approach. However, some problems were identified like scalability and efficiency. KNN requires that neighbourhood identification be performed online. This is not feasible most of the time because of the large amount of data. Another problem is the effectiveness in terms of coverage and precision. Low coverage is caused by larger user histories and low precision is due to the sparsity of Web data. The authors then proposed a solution that gives better results than the KNN approach in terms of scalability and effectiveness. They recommended an approach that uses association rules techniques that are based on storing the most frequent items used in a data structure and using an algorithm to identify the most suitable items to be used with online recommendations. The main problem associated with association rules in general is scalability due to the large number of itemsets. However, when the authors proposed a method that includes increasing the window size, it caused scalability problems as well as lower

coverage. On the other hand, using multiple support thresholds resulted in better coverage but it did not improve on accuracy.
Yang *et al.* (2004) have studied five different representations of association rules which are: Subset rules, Subsequence rules, Latest subsequence rules, Substring rules and Latest substring rules. As a result of the experiments performed by the authors concerning the precision of these five association rules representations using different selection methods, the latest substring rules were proven to have the highest precision with decreased number of rules.
On the other hand, Yong *et al.* (2005) explored sequential association rules further and they proposed a new sequential association rule model for Web document prediction based on the comparison of different types of sequential association rules according to sequence constrains and temporal constrains. They proved through means of experimentation that both sequence constrains and temporal constrains affect the precision of Web document prediction and that temporal constrains have more influence than sequence constrains.
Numerous papers dealt with the topic of Markov Model as a method to solve the prediction problem with higher coverage, better accuracy and performance than association rules. For example, Deshpande *et al.* (2004) addressed the reduced accuracy problem of the low-order Markov Models. They proposed an all-kth order model instead. They solved the state space complexity problem of the all-kth order model by pruning some of the states according to frequency, confidence and error representations. This proposed solution to the state space complexity of the all-kth order model may not be feasible in some instances, especially when it comes to very large data sets. It requires a lot of time and effort to build the all-kth order models and prune the pages according to the three criteria. It also involves a great deal of calculations (different types of thresholds for different pruning methods.)
Dongshan *et al.* (2002) proposed the use of a hybrid-order tree like Markov Model (HTMM) in order to solve the problems associated with traditional Markov Models especially the state space complexity and low coverage. They identified the suitability of HTMM with predicting the next pages to be accessed by the user and caching such pages in order to improve Web pre-fetching. HTMM combines two methods: a tree-like Markov model method and a hybrid order method. The k-order Tree-like Markov model is a tree constructed using a sequence of visited Web pages accessed by the user. Each node of the tree conforms to a visited page URL and a count that records the number of times the page was visited. The height of the tree is k + 2 where k is the order of the Markov model and the width of the tree is no more than the number of sequences of the visited pages. The tree-like Markov model results in low coverage that results in low accuracy. As a solution, the authors proposed training varying order Markov models and combining those models together for prediction. They used two methods for combining the models: accuracy voting and blending. To evaluate the results of these methods, the authors used Web server log files of an educational site and after cleaning and preprocessing the log data, they came up with the following results: When it comes to precision and accuracy, both HTMM methods showed better results than traditional Markov models. Also, when it comes to time associated with building the models and giving prediction, the HTMM methods showed better results than traditional Markov models. However, with prediction time, HTMM methods and traditional methods showed similar results. These results are apparent with HTMM in general. However, when it comes to building the tree, it is based on all-kth order model and it has the same complexity as the all-kth order model. This places a great limitation on the approach as a whole.
Related work was presented by Gunduz *et al.* (2003)where they proposed a new model that takes into consideration the time spent on the page as well as the sequence of visiting pages in a Web session. First, pages are clustered according to their similarities. Then, a click-stream tree is used to generate recommendations. This approach is rather complicated and data has to go in various stages before prediction takes place. Other researchers that went through similar work, like Mobasher *et al.* (2000) and Sarukkai (2000), did not take Web data complexity into consideration. Of course, more complex data would lead to higher storage space and runtime overhead.
Kim *et al.* (2004) presented a combination of association rules, Markov models, sequential association rules and clustering. This paper presented the use of four Web personalisation models in order to improve on their performance especially when it comes to precision and recall. The authors argued that both association rules and sequential association rules techniques can use All-Kth order model to increase coverage but this produced less precision.

1.2 Organisation of the Paper

This paper is organised as follows. In section 2, we cover Web access prediction using Markov model and association rules. In section 3, we introduce our proposed solution. In section 4 we analyse the data and produce the experiments results. Section 5 concludes our work.

2 Related Technologies

2.1 Markov Model

Markov models are becoming very commonly used in the identification of the next page to be accessed by the Web site user based on the sequence of previously accessed pages (Deshpande et al. 2004).
Let $P = \{p1, p2, \ldots, pm\}$ be a set of pages in a Web site. Let W be a user session including a sequence of pages visited by the user in a visit. Assuming that the user has visited l pages, then prob($pi|W$) is the probability that the user visits pages pi next. Page p_{l+1} the user will visit next is estimated by:

$$\begin{aligned} P_{l+1} &= \text{argmax}_{p \in \mathbb{P}}\{P(P_{l+1} = p|W)\} \\ &= \text{argmax}_{p \in \mathbb{P}}\{P(P_{l+1} = p|p_l, p_{l-1}, \ldots, p_1)\} \end{aligned} \quad (1)$$

This probability, $prob(pi|W)$, is estimated by using all W sequences of all users in history (or training data), denoted by W. Naturally, the longer l and the larger W, the more accurate $prob(pi|W)$. However, it is infeasible to have very long l and large W and it leads to unnecessary complexity. Therefore, to overcome this problem, a more feasible probability is estimated by assuming that the sequence of the Web pages visited by users follows a Markov process. The Markov process imposed a limit on the number of previously accessed pages k. In other words, the probability of visiting a page pi does not depend on all the pages in the Web session, but only on a small set of k preceding pages, where $k << l$.
The equation becomes:

$$P_{l+1} = \text{argmax}_{p \in \mathbb{P}}\{P(P_{l+1} = p|p_l, p_{l-1}, \ldots, p_{l-(k-1)})\} \quad (2)$$

where k denotes the number of the preceding pages and it identifies the order of the Markov model. The resulting model of this equation is called the Kth-Order Markov model. Of course, the Markov model starts calculating the highest probability of the last page visited because during a Web session, the user can only link the page he is currently visiting to the next one. The example is similar to Desphpandes Figure 1 (Deshpande et al. 2004):
Let S_j^k be a state containing k pages, $S_j^k = \langle p_{l-(k-1)}, p_{l-(k-2)}, \ldots, p_l \rangle$. The probability of $P\left(p_i|S_j^k\right)$ is estimated as follows from a history (training) data set.

$$P\left(p_i|S_j^k\right) = \frac{\text{Frequency}\left(\langle S_j^k, p_i\rangle\right)}{\text{Frequency}\left(S_j^k\right)} \quad . \qquad (3)$$

This formula calculates the conditional probability as the ratio of the frequency of the sequence occurring in the training set to the frequency of the page occurring directly after the sequence.
The fundamental assumption of predictions based on Markov models is that the next state is dependent on the previous k states. The longer the k is, the more accurate the predictions are. However, longer k causes the following two problems: The coverage of model is limited and leaves many states uncovered; and the complexity of the model becomes unmanageable. Therefore, the following are three modified Markov models for Predicting Web page access.

1. All kth Markov model: This model is to tackle the problem of low coverage of a high order Markov model. For each test instance, the highest order Markov model that covers the instance is used to predict the instance. For example, if we build an all 4- Markov model including 1-, 2-, 3-, and 4-, for a test instance, we try to use 4-Markov model to make prediction. If the 4-markov model does not contain the corresponding states, we then use the 3-markov model, and so forth (Pitkow & Pirolli 1999).

2. Frequency pruned Markov model: Though all-kth order Markov models result in low coverage, they exacerbate the problem of complexity since the states of all Markov models are added up. Note that many states have low statistically predictive reliability since their occurrence frequencies are very low. The removal of these low frequency states affects the accuracy of a Markov model. However, the number of states of the pruned Markov model will be significantly reduced.

3. Accuracy pruned Markov model: Frequency pruned Markov model does not capture factors that affect the accuracy of states. A high frequent state may not present accurate prediction. When we use a means to estimate the predictive accuracy of states, states with low predictive accuracy can be eliminated. One way to estimate the predictive accuracy using conditional probability is called confidence pruning. Another way to estimate the predictive accuracy is to count (estimated) errors involved, called error pruning.

2.2 Association Rules

Association rule mining is a major pattern discovery technique as proved by Mobasher *et al.* (2000). The original goal of association rule mining is to solve market basket problem. For a data set containing shopping transactions, association rules summarise relationships illustrated by the following example. Customers who buy bread and milk will most likely buy eggs, or, bread and milk $\rightarrow$ eggs. Association rules are mainly defined by two metrics: support and confidence. The applications of association rules are far beyond market basket applications. Let us look at how association rules are used in Web data mining.
Let $P = \{p_1, p_2, , p_m\}$ be a set of pages in a Web site. Let W be a user session including a sequence of pages visited by the user in a visit, and D includes a collection of user sessions. Let A be a subsequence of W, and p_i be a page. We say that W supports A if A is a subsequence of W, and W supports $\langle A, p_i\rangle$ if $\langle A, p_i\rangle$ is a subsequence of W. The support for sequence A is the fraction of sessions supporting A in D, denoted by $\text{supp}(A)$. An implication is $A \rightarrow p_i$. The support of implication $A \rightarrow p_i$ is $\text{supp}(\langle A, p_i\rangle)$, and the confidence of the implication is $\text{supp}(\langle A, P\rangle)/\text{supp}(A)$, denoted by $\text{conf}(A \rightarrow p_i)$. When we use the same terminologies of Markov model, $\text{supp}(\langle A, p_i\rangle) = \text{prob}(\langle A, p_i\rangle)$, and confidence $(A, p_i) = \text{prob}(p_i|A)$. An implication is called an association rule if its support and confidence are not less than some user specified minimum thresholds.
The minimum support requirement dictates the efficiency of association rule mining. One major motivation for using the support factor comes from the fact that we are usually interested only in rules with certain popularity. Support corresponds to statistical significance, and confidence is a measure of the rules strength.
There are four types of sequential association rules presented by Yang *et al.* (2004):

1. Subsequence rules: they represent the sequential association rules where the items are listed in order.

2. Latest subsequence rules: They take into consideration the order of the items and most recent items in the set.

3. Substring rules: They take into consideration the order and the adjacency of the items.

4. Latest substring rules: They take into consideration the order of the items, the most recent items in the set as well as the adjacency of the items.

The immense number of generated rules gives rise to the need of some predictive models that reduce the rule numbers and increase their quality by weeding out the rules that were never applied. Yang *et al.* (2004), introduced the following predictive models:

1. Longest match: This method assumes that longer browsing paths produce higher quality information about the user access pattern. Therefore, in the case where we have more than one rule, all with support above a certain threshold and they match an observed sequence, the rule with the longest length will be chosen for predication purposes and the rest of the rules will be disregarded.

2. Most-confidence matching: This is a very common method where the rule with the highest confidence is chosen amongst the rest of all the applicable rules whose support values are above a certain threshold.

3. Least error matching: This is a method to combine support and confidence, based on the observed error rate and the support of each rule, to form a unified selection measure and to avoid the need to set a minimum support value artificially. The observed error rate is calculated by dividing

the number of incorrect predictions by the number of training instances that support it. The rule with the least error rate is chosen amongst all the other applicable rules.

From a previous study (Yang et al. 2004), the latest substring with the least error matching produces the most accurate models for Web document prediction. In this paper, we will use sequential association rule mining on user transaction data to discover Web page usage patterns. Prediction of the next page to be accessed by the user is performed by matching the discovered patterns against the user sessions. This is usually done online.

3 A framework for integration

The main problem associated with association rules that apply to large data item sets is the discovery of large number of rules and the difficulty in identifying the one rule that leads to the correct prediction. In regards to Markov models, low order Markov models lack web page prediction accuracy because they do not use enough history and high order Markov models have high state space complexity.
There is apparent a direct relationship between Markov models and association rules techniques. According to the Markov model pruning methods presented by Mobasher *et al.* (2004) and association rules selection methods presented by Yang *et al.* (2004), there exists a great resemblance between the two. The substring association rules with most confidence prediction model form a frequency pruned all kth order Markov model, where k is the number of maximum items in the association rules. They also share similar problems. For instance, the number of states (rules) becomes unmanageable when k is large. In contrast, short history is not enough for making accurate predictions.
We propose to use low order all kth Markov models to keep low state complexity and high coverage. The accuracy of low order Markov models is normally not satisfactory. For those Markov states that provide ambiguous predictions, we make use of association rules to sample long history. Association rules help those states to make more accurate predictions. Association rules are complicated as well, but we only use rules to complement Markov states that provide ambiguous predictions. Therefore, this does not add too much complexity to the system. We use the following example to show the idea of the integration.
Consider the set of Web page structure for an online computer shop in Figure 1.
Note that letters are assigned to nodes names in Figure 1 for simplicity purposes. Table 1 examines the following 6 user sessions:

Table 1: User sessions

T1	A,C,G,A,D,H,M,C,F,C,G,R,I,P,H,O,J
T2	A,G,T,A,C,S,G,J,R,A,D,H,M,D,J
T3	A,F,I,B,A,E,D,H,N,P,I,Q,F,J,D,H,N,G,C
T4	A,I,J,B,A,E,C,T,D,H,M,I,Q,G
T5	F,D,H,N,J,A,D,A,E,D,J,R,H,N,G,C,F,G
T6	F,L,S,D,H,N,J,Q,E,I,P,C,I,O,A,D,H,M

Calculating the frequencies of accessed pages, Table 2 lists the pageviews with their frequencies.
A 100% support results in a very large number of rules and is rather cumbersome. Therefore, assuming that the minimum support is 4; B, K, L, O, P, Q, R, S and T are removed from the itemsets. Table 3 lists the user sessions that pass the frequency and support tests.

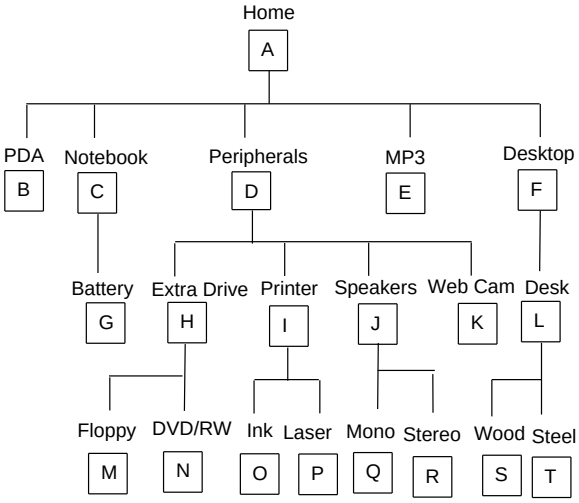


Figure 1: Online computer store Web page structure.

Table 2: Pageviews frequencies

Page	A	B	C	D	E	F	G	H	I	J
Freq	12	2	8	11	4	6	8	10	7	8

Page	K	L	M	N	O	P	Q	R	S	T
Freq	0	1	4	4	3	3	3	3	2	2

Table 3: User sessions after frequency and support pruning

T1	A,C,G,A,D,H,M,C,F,C,G,R,I,H,J
T2	A,G,A,C,G,J,A,D,H,M,D,J
T3	A,F,I,A,E,D,H,N,I,F,J,D,H,N,G,C
T4	A,I,J,A,E,C,D,H,M,I,G
T5	F,D,H,N,J,A,D,A,E,D,J,H,N,G,C,F,G
T6	F,D,H,N,J,E,I,C,I,A,D,H,M

Applying the 2^{nd} order Markov Model to the above training user sessions we notice that the most frequent state is $\langle D, H\rangle$ and it appeared 8 times as follows:

$$P_{l+1} = \text{argmax}\{P(M|H,D)\} = M \text{ OR } N$$

Obviously, this information alone does not provide us with correct prediction of the next page to be accessed by the user as we have high frequencies for both pages, M and N. To break the tie and find out which page would lead to the most accurate prediction, we have to look at previous pages in history. This is where we use subsequence association rules as it shows in Table 4 below.

Table 4: User sessions history

A, C, G, A,	$\langle D, H\rangle$	M
A, G, A, C, G, J, A,	$\langle D, H\rangle$	M
A, F, I, A, E,	$\langle D, H\rangle$	N
I, F, J,	$\langle D, H\rangle$	N
A, I, J, A, E, C,	$\langle D, H\rangle$	M
F,	$\langle D, H\rangle$	N
F,	$\langle D, H\rangle$	N
J, E, I, C, I, A,	$\langle D, H\rangle$	M

Tables 5 and 6 summarise the results of applying subsequence association rules to the training data. Table 5 shows that C → M has the highest confidence of 100%. While Table 6 shows that F → N has the highest confidence of 100%.

Table 5: Confidence of accessing page M using subsequence association rules

A → M	AM/A	4/10	40%
C → M	CM/C	4/4	100%
E → M	EM/E	2/3	67%
F → M	FM/F	0/4	0%
G → M	GM/G	2/3	67%
I → M	IM/I	2/5	40%
J → M	JM/J	3/4	67%

Table 6: Confidence of accessing page N using subsequence association rules

A → N	AN/A	1/10	10%
C → N	CN/C	0/4	0%
E → N	EN/E	1/3	33%
F → N	FN/F	4/4	100%
G → N	GN/G	0/3	0%
I → N	IN/I	2/5	40%
J → N	JN/J	1/4	25%

Using Markov models, we can determine that there is a 50/50 chance that the next page to be accessed by the user after accessing the pages D and H could be either M or N. Whereas subsequence association rules take this result a step further by determining that if the user accesses page C before pages D and H, then there is a 100% confidence that the user will access page M next. Whereas, if the user visits page F before visiting pages D and H, then there is a 100% confidence that the user will access page N next.
Applying this result back to our example, we find that if the user buys a notebook, there is more chance that he/she will buy an external floppy drive. However, if the user buys a desktop, there is more chance that he/she will buy an extra DVD/RW drive. This extra bit of information is very important as knowing user browsing history gives us an added advantage of knowing the browsing habits of our users.
In this paper, we introduced the Integrated Markov and Association Model (IMAM) that inputs a database(D) and a session (s) and outputs the next page(p) that will be accessed by the user with high prediction. IMAM is summarised as follows:

Training:

```
Build a low order Markov model
FOR each state of the Markov model
    IF the prediction is ambiguous
        THEN
        Collect all sessions satisfying
            the state
        Construct association rules to
            resolve ambiguity
        Store the association rules with
            the state
    ENDIF
ENDFOR
```

Test:

```
Find a matching state of the Markov model
      for a test session
IF the matching state provides an non-
      ambiguous prediction
    THEN the prediction is made by the state
    ELSE
    Use its corresponding association
      rules to make prediction
ENDIF
```

In this work, we define an ambiguous prediction as two or more predictive pages that have the same conditional probability by a Markov model. The ambiguous prediction potentially has other definitions, for example, the certainty of a prediction is below a threshold. We did not explore other options in this paper.

4 Experimental Evaluation

4.1 Data Collection and Preprocessing

For our experiments, the first step was to gather a log file from an active Web server. Usually, Web log files are the main source of data for any e-commerce or Web related session analysis (Spiliopoulou et al. 1999). The log file we used as a data source for our experiments is a day's worth of all HTTP requests to the EPA WWW server located at Research Triangle Park, NC. The logs are an ASCII file with one line per request, with the following information: The host making the request, date and time of request, requested page, HTTP reply code and bytes in the reply. The logs were collected for Wednesday, August 30 1995. There were 47,748 total requests, 46,014 GET requests, 1,622 POST requests, 107 HEAD requests and 6 invalid requests. The gathered Web log data had to be cleaned and filtered (Zhao, Bhowmick & Gruenwald 2005, Sarukkai 2000).
Cleaning the data involved removing erroneous and invalid pages. Those included HTTP error codes 400s and 500s, HTTP 1.0 errors, and CGI entries. The total number of valid entries was diminished to 19,121. Then, 302 and 304 HTTP errors that involve requests with no server replies were also removed and the number of entries went down to 14,091. Filtering and cleaning the log files made them ready for further preprocessing and analysis. Pages links are converted to numbers for easy manipulation. Repeated pages are removed because it is uncommon for the same Web page to be accessed more than once and any internal links are irrelevant. Next step was to identify user sessions. Taking a 30-minute timeout into consideration, the number of user sessions amounted to 1,868. Short sessions were then removed and only sessions with at least 5 pages were considered. Distinct Web pages were identified and they amounted to 2,891 pages.
The EPA data was further pre-processed before being used for our analysis purposes. The last page of each session was removed for testing purposes. Also, the frequency of each page visited by the user was calculated. The page access frequency is shown in Figure 2 which reveals that page number 3 is the most frequent page and it was accessed 73 times.

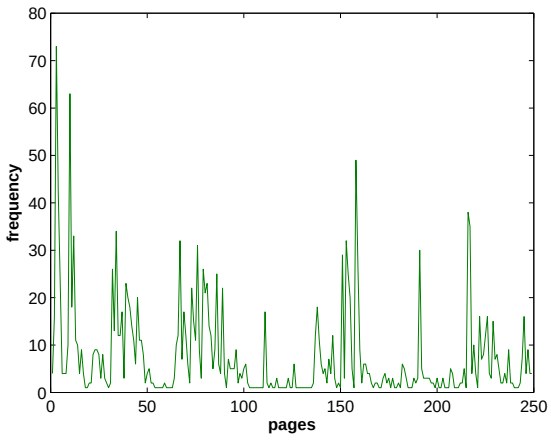


Figure 2: Frequency chart for the most frequent visited pages.

4.2 Experiments Results

Having all data sets processed, filtered and analysed, 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order Markov models were created. Then, all 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} order frequency pruned (Deshpande et al. 2004) Markov model analysis took place considering 4 as the frequency threshold. Prediction results were achieved using the maximum likelihood based on conditional probabilities as stated in equation 3 above. All predictions in the test data that did not exist in the training data sets were assumed incorrect and were given a zero value. All implementations were carried out using MATLAB. Figure 3 below illustrates the difference between Markov model orders and Frequency pruned all-kth Markov model results. The Figure demonstrates that as the order of Markov model increases, precision decreases due to the reduced coverage of the data. Coverage is defined as the ratio of the Web sessions in the test set that have a corresponding state in the training set to the number of Web sessions in the test set (Deshpande et al. 2004). Also, the increase of the frequency pruned Markov model precision is limited due to the elimination of states that could be of importance to the precision process. The frequency threshold parameter used was a fixed parameter of size 4.

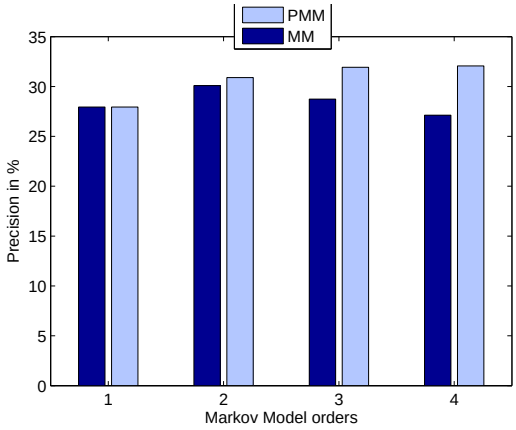


Figure 3: Precision of 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order Markov models and all 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order frequency pruned Markov models.

Table 7 below reveals that the all- 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} order frequency pruned Markov models have considerably less states than the 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} order Markov models.

Table 7: Number of states of Markov model and frequency pruned Markov model orders.

Model	MM States	All-kth FP States
1^{st} order	1945	745
2^{nd} order	39162	9162
3^{rd} order	72524	14977
4^{th} order	101365	17034

The reported accuracies in this section are based on 10-fold cross validation. The data was split into ten equal sets. First, we considered the first nine sets as training data and the last set for test data. Then, the second last set was used for testing and the rest for training. We continued moving the test set upward until the first set was used for testing and the rest for training. The reported accuracy is the average of ten tests.

The 1^{st} order and 2^{nd} order Markov model results cannot be 100% reliable simply because we did not look back into the history of pages accessed by the user. We assumed that the pages visited long before the current page in a Web session do tend to influence the users actions. These previously accessed pages affect the prediction process as they interfere with the user browsing behaviour and are not mere information providers. Performing 3^{rd} and 4^{th} order Markov models techniques solves the problem of examining the users previous browsing behaviour, but it results in an increase in the number of states as it is obvious in Table 7 above that illustrates the number of states generated based upon non empty states. To overcome this shortcoming, we applied subsequence association rules techniques in order to generate the most appropriate rule. Before applying association rules techniques, the most frequent occurrences or the Markov model frequent states are removed.
Since association rules techniques require the determination of a minimum support factor and a confidence factor, we used the experimental data to help determine such factors. We can only consider rules with certain support factor and above a certain confidence threshold.
Figure 4 below shows that the number of generated association rules dramatically decreases with the increase of the minimum support threshold with a fixed 90% confidence factor. Reducing the confidence factor results in an increase in the number of rules generated. This is apparent in Figure 5 where the number of generated rules decreases with the increase of the confidence factor while the support threshold is a fixed 4% value. It is also apparent from Figure 4 and Figure 5 below that the influence of the minimum support factor is much greater on the number of rules than the influence of the confidence factor.

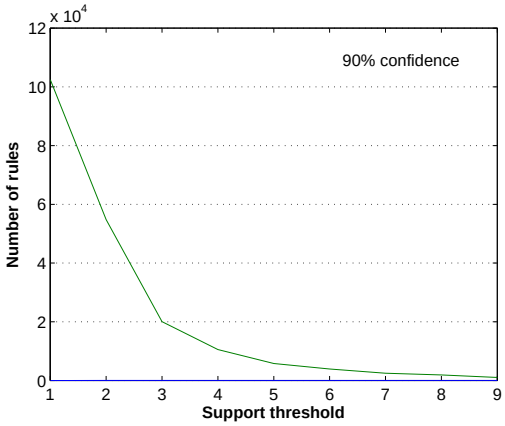


Figure 4: Number of rules generated according to different support threshold values and a fixed confidence factor: 90%.

Referring back to Figure 4 and Figure 5, we considered a minimum support threshold of 4%. The integration model, IMAM, involves calculating association rules techniques prediction accuracy using the longest match precision method. In IMAM, association rules were applied in two cases:

1. When we were unable to make a correct prediction in the case of a 2^{nd} order Markov model because of a tie. In such a case, using association rules techniques to look further back at previously visited pages, we were able to break the tie by looking at the page in history that leads to the most appropriate page for prediction. Looking at Figure 6, using 1^{st} order Markov

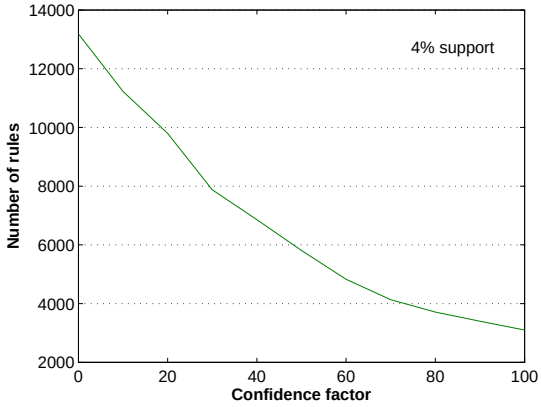


Figure 5: No. of rules generated according to a fixed support threshold: 4%.

model, the most frequently accessed page after EPA-PEST1995Aug23 is EPA-PEST1995Aug17 with 100% probability. Using 2^{nd} order Markov model, the most frequently accessed pages after EPA-PEST1995Aug17 are EPA-PEST1995July and OOPTPubs with 50% probability each. To decide which of the two pages would result in higher prediction precision, we look further back. Using association rules we find out that there is 100% chance that if EPA-PEST1995Aug16pr-373 is accessed before EPA-PEST1995Aug23, EPA-PEST1995July will be accessed next. And, there is 100% chance if PressReleases1995Aug is accessed before EPA-PEST1995Aug23, OOPTPubs will be accessed next. As a result, precision is calculated according to the results of association rules.

The precision of the proposed IMAM model was calculated by adding all successes and dividing the result by the number of states in the test data. According to Figure 7, the proposed IMAM model shows better precision than the 2^{nd} order Markov model (MM) and the frequency pruned all 2^{nd} order Markov model (PMM).

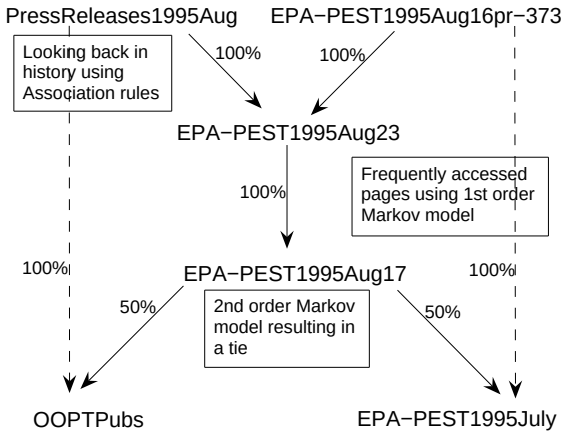


Figure 6: Portion of association rules results.

2. If the test data does not match any of the 2^{nd} order Markov model outcomes, we use the globally generated association rules to look back at previous user browsing history. Users have different browsing experiences, some of them get to the page they request using a shorter path than others depending upon the web site structure and internal links. For example, the same page could be accessed by a user after visiting 5 pages and by another user after visiting 2 pages.

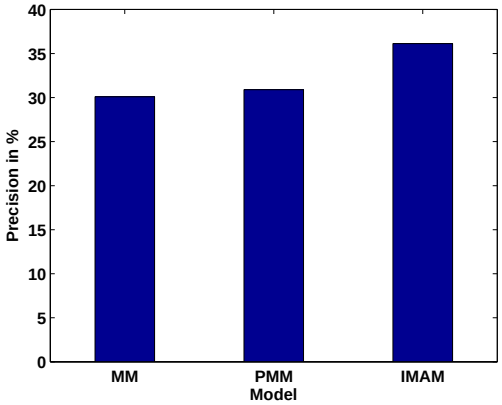


Figure 7: Precision of 2^{nd} order Markov model (MM), Frequency Pruned all 2^{nd} order Markov model (PMM) and IMAM model.

The main problem associated with this approach is that it is dependent on the length of user session of data available. This is usually not a problem when modelling a particular site with long user sessions and therefore, more history. But it becomes more difficult when performing multi-site analysis with shorter user sessions.

5 Conclusion

In this paper, we proposed a method to integrate Markov model and association rules for predicting Web page accesses. The integration is based on a low order Markov model. Sets of subsequence association rules are used to complement the Markov model for resolving ambiguous predictions by using long history data. The integration avoids the complexity of high order Markov model and the limitation of Markov model using short history. This model also reduces the large number of association rules since association rules are only used when ambiguous predictions occur. The experimental results show that the combined model increases the accuracy of the Web page access prediction of Markov model and association rules.

References

Cadez, I., Heckerman, D., Meek, C., Smyth P. & White S. (2000), Visualization of Navigation Patterns on a Web Site Using Model Based Clustering, *in* 'ACM SIGMOD International Conference on Knowledge Discovery and Data Mining', ACM Press, Washington DC, USA, pp. 280-284.

Deshpande, M. & Karypis, G. (2004), 'Models for Predicting Web Page Accesses', *Transactions on Internet Technology* **4**(2), 163–184.

Dongshan, X. & Junyi, S. (2002), 'A New Markov Model for Web Access Prediction', *Computing in Science and Engineering* **4**(6), 34–39.

Garafalakis, M., Rastogi, R., Seshadri, S. & Shim, K. (1999), Data Mining and the Web: Past, Present and Future, *in* 'WIDM Conference', Kansas City, USA, pp. 43-47.

Gunduz, S. & Ozsu, M. (2003), A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior, *in* 'SIGKDD 03 Conference', Washington, DC, USA, pp. 24-27.

Jespersen, S., Pedersen, T. B. & Thorhauge, J. (2003), Evaluating the Markov Assumption for Web Usage Mining, *in* '5th international workshop on WIDM03', New Orleans, USA, pp. 82-89.

Kim, D., Lm, L., Adam, N., Atluri, V., Bieber, M. & Yesha, Y. (2004), A Clickstream-Based Collaborative Filtering Personalization Model: Towards A Better Performance, *in* 'WIDM 04 Conference', Washington DC, USA, pp. 12-13.

Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2000), Discovery of Aggregate Usage Profiles for Web Personalisation, *in* 'WebKDD Workshop 2000', USA, pp. 61-82.

Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2001), Effective Personalization Based on Association Rule Discovery from Web Usage Data, *in* 'WIDM 01, 3^{rd} ACM Workshop on Web Information and Data Management', Atlanta, Georgia, USA, pp. 9-15.

Pei, J., Han, J., Mortazavi-asl, B. & Zhu, H. (2000), Mining access patterns efficiently from Web logs, *in* 'PAKDD conference', USA, pp. 396-407.

Pitkow, J. & Pirolli, P. (1999), Mining longest repeating subsequence to predict world wide Web surfing, *in* 'the 2^{nd} USENIX Symposium on Internet Technologies and Systems', Boulder, CO., pp. 139-150.

Sarukkai, R. (2000), Link Prediction and path analysis using Markov Chains, *in* 'the Ninth International World Wide Web Conference', Amsterdam, pp. 377-386.

Spiliopoulou M., Faulstich L. C. & Winkler K. (1999), A Data Miner analyzing the Navigational Behaviour of Web Users, *in* 'Workshop on Machine Learning in User Modelling of the ACAI99 International Conference', Creta, Greece.

Yang, Q., Li, T. & Wang, K. (2004), 'Building Association-Rule Based Sequential Classifiers for Web-document Prediction', *Journal of Data Mining and Knowledge Discovery* **8**(3), 253–273.

Yong, W., Zhanhuai, L. & Yang, Z. (2005), Mining Sequential Association-Rule for Improving WEB Document Prediction, *in* 'Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA05)', pp. 146-151.

Zhao, Q., Bhowmick, S. S. & Gruenwald, L. (2005), WAM:Miner: In the Search of Web Access Motifs from Historical Web Log Data, *in* 'CIKM05 conference', Germany, pp. 421-428

Modeling Spread of Ideas in Online Social Networks

Muhammad A Ahmad[1] **Ankur Teredesai**[2]

[1]Department of Computer Science and Engineering,
University of Minnesota,
Minnesota, MN 55455, USA
[2]Institute of Technology,
University of Washington,
Tacoma, WA 98402, USA
Email: mahmad@cs.umn.edu, ankurt@u.washington.edu

Abstract

Internet based online social networks collectively facilitate the spread of ideas. Hence, to understand how social networks evolve as a function of time, it is critical to learn the relationship between the information dissemination pathways or flows and the type of ideas being disseminated. We first classify the spread of ideas into two types based on their rate and nature of proliferation; fads and non-fads. A 'fad' refers to an idea that quickly becomes popular in a culture, remains popular for a brief period, and then loses popularity dramatically. We then model the information dissemination pathways for both these types of ideas. Our results indicate that the proliferation of information in a network strongly correlates with the the type of idea, the degree of participation of the nodes, and a node's availability *i.e.,* presence. Further we derived that after reaching a certain saturation point, a fad exhibits periodic spreading behavior implying that a fad rarely completely disappears from a network. We use data from an instant messaging network community to verify the proposed theoretical modeling framework.

Keywords: Simulation, Social Networks, Information Dissemination, Instant Messaging, Memes, Social Network Analysis.

1 Introduction

In the real world, information disseminates because of information asymmetry. In economics, this information asymmetry as discussed by Arrow (1963) is referred to as the situation when one party to a transaction has more or better information than the other party. Hence, when ideas spread, and information disseminates, information asymmetry decreases. Moreover, information asymmetry has recently been noted to be on the decline thanks to the Internet. The Internet facilitates users that are unknowlegeable to acquire heretofore unavailable information very rapidly. Think of how easy it is to get costs of competing insurance policies, various dealer's quotes for the same car, etc. (Levitt S., & Dubner, S., 2005). Thus, information dissemination or the motivation to spread ideas is the central doctrine behind online social networking.

Existing efforts that model information dissemination assume that the underlying social networks are static *i.e.,* the topology of the network remains constant. In a dynamic social network the edges between the nodes are not fixed and can change over the course of time. Thus, the topology reflects the participation of the nodes in the network. The focus of this paper is to develop a novel framework that attempts to explain the phenomenon of information dissemination in dynamic online social networks. Instant Messaging (IM) networks are examples of such dynamic online social networking environments. It is a very popular way of computer-based communication. The task of studying information dissemination in dynamic networks has gained even more importance from a cybersecurity perspective, since it can used to study information flows in terrorist networks. A feature that terrorist organizations share with highly dynamic networks like IM networks is that the unavailability (removal) of several actors does not make much difference in spreading information. For terrorist networks, this is so because the network is structured to minimise the loss of utility and sustain the network inspite of removal of a few actors (Erickson 1981). The information pathways evolve and adapt as a response to anti-terrorist activities (Sageman 2004) with the consequence that the topology of the network does not remain constant.

There have been few prior attempts to propose appropriate modeling solutions for information flow prediction problems in dynamic social networks. Domingos et al mine the network value of customers using a Markov chaining process (Domingos, P., & Richardson, M. 2001). Kempe et al (2003) provide a theoretical perspective to study the spread of influence through a social network. A key difference and advancement in our efforts being that our methods are not constrained by the availability of connectionist information. *i.e.,* we do not presume that the network topology is a known parameter of the modeling algorithm. In fact, we compare the results of our algorithm for a connections-known network and a network where the nodes can randomly become active or passive (real-network). For theoretical grounding, we borrow the notions of susceptibility and transmissibility from epidemiology; and adapt them to our problem domain. We use a real world dataset, consisting of IM status logs, and IM user behavior data to calibrate our theoretical models. We should also emphasize that the current study is exploratory in nature and thus the goal is to find general trends in proliferation of information in social networks instead of making particular (and speculative) predictions.

To provide some background, IM can be defined

Ankur M. Teredesai is a faculty member on leave from Rochester Institute of Technology, Rochester, NY.

as a communications service that enables its users to create a kind of private chat room with another individual in real time over the Internet if both of them are using the same service. Users can initiate chat sessions with other people in their buddylist. IM technology lets users communicate across networks, in remote areas, and in a highly pervasive and ubiquitous manner. Industrial and governmental organizations are very interested in understanding the nature of broad knowledge-sharing networks that exist within their organizations. IM communication is fast becoming a standard platform for such networks (Teredesai, A., Resig, J., Dawara, S., & Homan, C., 2004). To preserve the privacy of IM users, neither the connectionist data consisting of the user's buddylist nor the content of the chat-sessions was collected. Consequently, any modeling has been restricted to user status logs which are made publicly available as traffic data through open standards. Also, this anonymized IM network status-log data is publicly available (imscanwebsite 2006).

This article is further divided into the following sections. Section 2 talks about related work in Social Network Analysis, section 3 provides the theoretical background and introduces key concepts. Section 4 gives the formalization for the general model, section 5 introduces a variant of the current model that can be used to study fads. Finally experiments and conclusion are covered in sections 6 and 7 respectively.

2 Related Work

The field of social network analysis has been gaining tremendous importance in recent years. The scale-free nature of the link distribution in the World-Wide Web indicates that collective phenomena play a previously unsuspected role in the development of the web. While earlier formalization and modeling was on random graphs, Albert et al note that we need to look beyond the traditional random graph models to gain a better understanding of the web's topology in order to design effective strategies for making information widely accessible (Albert, R., Jeong, E., & Barabasi, A-L. 1999). One can expand the question and ask, Is the same true of large scale social networks such as IM networks? Thus, the question of information dissemination is strongly linked to the study of changes in network topology and vice-versa.

Prior to our attempt, Kautz et al attempted to model the dissemination of information in social networks using collaborative filtering techniques and adopted a simulation approach to validate the results (Kautz, H., Selman, B.& Shah, M., 1997). Our current work is along a similar vein, although the setting and the nature of our network puts different constraints in modeling and formalization. Another interesting area where our current research can be extended is to address the problem of discovering connection subgraphs as networks evolve over time. Previously, Faloutsos et al. (2004) provided a fast algorithm for discovery of connection subgraphs for static networks. Given the discrete nature of IM Networks, Faloutsos's fast algorithm can be implemented at each time-step iteratively to determine the change in the connection subgraphs. This can lead to discovery of change points or prominent events within the social network evolution. Our network model is based upon the work by Moore and Newman (2000) which was itself built upon a previous work by Newman and Watts (1998). Watts and Strogatz (1998) developed a small-world model where any two individuals in the network have only a small degree of separation between them. Moore and Newman used a variant of this model to study disease transmission in small-world networks. In this paper we use the Moore and Newmann's model as a basis for our models and experiments.

Karp et al. studied epidemic algorithms for the lazy transmission of updates to distributed copies of a database (Karp, R., Schindelhauer, C., Shenker, S. & Vocking, B., 2000). Kempe et al. explored gossip protocols to study the dynamic behavior of a network in which information is changing continuously over time (Kempe, D., Kleinberg, J., & Demers, A., 2003). These attempts are noteworthy but the scope of our paper is different since we are more interested in models for information dissemination.

3 Theoretical Background

Watts and Strogatz (1998) developed a small-world model where any two individuals in the networks only have a small degree of separation between them. Moore and Newman used a variant of this model to study disease transmission in small-world networks (2000). In this paper we use a variant of their model as a basis for our modeling. Moore and Newman's network can be constructed as follows: Consider a k dimensional lattice. For simplification consider k=1. Each vertex in the network is connected to all its neighboring sites. Any two different are then randomly connected with an edge. The process is continued until a small world network is obtained. The proof that this network is a small world network is beyond the scope of this paper and is discussed by Moore and Newman.

As stated earlier a major difference between an IM network and traditional networks is that in the IM networks any arbitrarily chosen node in the network cannot be always be assumed to be an active participant in the network. This is so because the person corresponding to the node may go offline, thus effectively severing all links with its neighbors for the time period that she is offline. In the case of the spread of infectious diseases this is akin to an infected person being physically absent and thus leaving the social network and then rejoining later. In IM networks this situation happens fast enough and often enough to be noticeable even in small intervals of time. In our network this situation is reflected by IM status changes. The following concepts will be useful in studying information dissemination in dynamic online social networks.

Susceptibility: We define susceptibility σ as the probability that an individual exposed to a meme will be infected by the meme and will herself become a 'carrier'. Given any subject or topic, people with different backgrounds are likely to have varying opinions on the subject. A person usually adopts a meme if it conforms with the world view that the person already holds. It is also possible that memes can change its state during transmission but in the current scenario we assume that the memes are fixed. In the current context, a higher probability for σ denotes that the new piece of information conforms well with views a person already hold. On the other hand a low probability would imply the opposite. To reflect this situation each node in our network is assigned a random value between 0 and 1 for, susceptibility σ.

Transmissibility: We define Transmissibility τ as the probability that whenever two nodes A and B come into contact, and A is already infected, then the meme will be transmitted to B.

In an ideal network (there is only one status *i.e.,* the transmissibility will be perfect for all individuals since everyone will want to share the information with all their neighbors. However, in our model which

more closely mimics a real IM network, transmissibility depends upon the agent being active or inactive.

Status: The status s_t of an IM user at time t is an element of the set $\{s_{online}, s_{idle}, s_{away}, s_{offline}\}$ which can be mapped to active and inactive nodes $\{1, 0\}$.

Origin: The node from which the meme originates will be referred to as the origin.

Active Node: A node at time t is said to be active if the corresponding user is online.

3.1 IM Network as a Graph

Consider a network of IM users represented by V nodes, let the links between the IM users be represented by the edge-set E and let A be the set of users who have adopted the meme.

At the beginning of the experiment one of the nodes, say v_B is chosen at random and is flagged as a carrier. v_B will henceforth be referred to as the origin. The node v_B is then placed in the set A and all its neighbors are selected and the meme is said to be transmitted from v_B to a neighbor v_C if the joint probability $p(C \cap B)$, described in section 4, is greater than a predefined threshold T. The joint probability is given by equation (1):

$$p(C \cap B) = p(C|B).p(C) \quad (1)$$
$$p(C \cap B) = p(B).p(C) \quad (2)$$

For the next iteration and all subsequent iterations the elements of set A are selected and an attempt is made to transmit the meme to their respective neighbors until the network gets saturated with the meme or all subsequent transactions result in a probability that is less than or equal to the threshold.

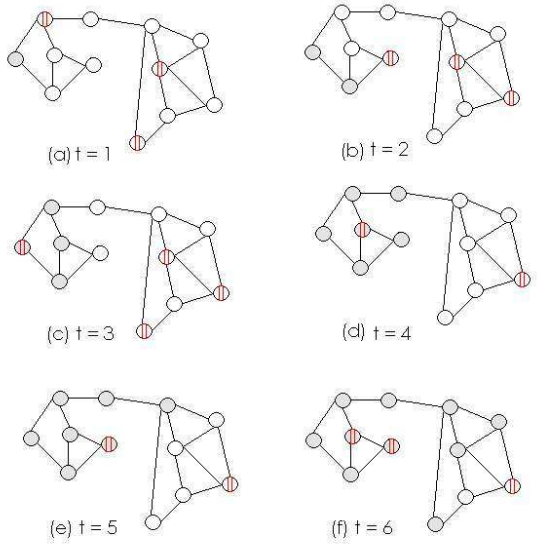


Figure 1: Proliferation of a meme in the network over four iterations. The gray nodes represented are flagged nodes while the nodes with red bars are inactive.

4 The General Information Dissemination Model

Now we consider the question that at time t what would be the number of nodes reachable from the origin? In other words, how many nodes will be infected at time t. This quantity can be termed as **reachability**. The set of such agents would form a subgraph R. Consider the ideal case where transmissibility and susceptibility are perfect *i.e,* equal to one. In this case the number of iterations required to saturate the the network with the meme would be equal to the eccentricity of the origin (initially flagged node). This is so because after t iterations, all the nodes with distance $\leq t$ will be reachable in an ideal network and eccentricity just defines the maximum distance between the origin and any other node in the network. In the best case scenario, where eccentricity is equal to radious, reachability would be equal to the radius of the network while in the worst case it would be equal to the diameter of the network when eccentricity would be equal to the diameter. In the ideal network, at time t, the set R would be equal to A since the probability for the transmission of the meme would be either equal (when T =1) to or greater than the threshold. For the ideal case, after t iterations, A_t can be given as:

$$A_t = \{\ v_t\ :\ v_t\ \in\ V,\ d(v_t, v_0) \leq t\} \quad (3)$$

Let η_i be the set of neighbors of $v_i \in A_t$ such that $\eta_i = \{x : x \notin A_t,\ d(x, v_i) = 1\}$ then A_t for the ideal case can also be given as:

$$A_t = A_{t-1} \cup \left(\bigcap_{i=1}^{k} \eta_i \right) \quad (4)$$

Let us now consider the generalized case where the susceptibility is different for each agent and the transmissibility changes with time. In the case of the IM network the transmissibility is given by the status of the user. Let the paths between node A and B be represented by $\rho(A, B)$ then for the generalized case, if i is the iteration then A_t can be given as:

$$A_t = \left\{ \begin{array}{c} v_t\ :\ v_t\ \in\ V,\ d(v_t, v_0) \leq t, \\ \wedge v_t \in \rho(v_0, v_i),\ T_i \geq T \end{array} \right\} \quad (5)$$

Equation 4 morphologically remains the same with the difference that $\eta_i = \{x : x \notin A_t,\ d(x, v_i) = 1\, s_{(i-1)t} = 1\}$ Consequently the set A_t of users who have adopted the meme at time t will be different for the ideal case and the IM case, specifically: $A_{IM} \subseteq A_{IDEAL}$.

Now we consider the question that given a vertex v_t and the origin, what is the probability that after t iterations v_t will also be infected? To avoid cumbersome notation we introduce the notation α_{it} such that at time t, α_{it} is 1 if $v_{it} \in A_t$, and 0 otherwise. Hence α_{it} represents all the nodes that are in A_t at time t. Since the transmissibility τ and the susceptibility σ are independent of one another, the probability a_{i_t} that at time t node v_i will be infected with the meme can be given as:

$$a_{i_t} = \tau_{i_t}.\sigma_{i_t} \quad (6)$$

The above equation is however an oversimplification since it does not take into account the state of the neighbors of the node or even the rest of the network. This can be remedied as follows. Given any two nodes v_i and v_j, the probability of transmission or infection from v_i and v_j is a conditional probability as defined above and also depends upon the set of paths between v_i and v_j. First consider the simplest case where v_0 and v_1 are neighbors, then the probability that at time t, v_1 has been infected can be determined as:

$$p_{1_t} \quad = \quad max(\alpha_{1_t}, a_{1_t}.a_{0_t}) \quad (7)$$

Now consider the more complex case where two paths from v_2 to v_0 via v_1 and also via v_3 and there are four nodes v_0, v_1, v_2, v_3 then:

$$p_{2_t} = max(\alpha_{2_t}, a_{2_t}.a_{1_t}.a_{0_t}, a_{3_t}.a_{2_t}.a_{0_t}) \quad (8)$$

Now consider the generalized case where there are n paths from any vertex to any other node, then the probability is given by:

$$p_{n_t} = max \left\{ \begin{array}{c} \alpha_{n_t}, \prod_{i=1}^{n} \rho_{1i_t}, \prod_{i=1}^{n} \rho_{2i_t}, \\ \ldots, \prod_{i=1}^{n} \rho_{(n-1)i_t}, \prod_{i=1}^{n} \rho_{ni_t} \end{array} \right\} \quad (9)$$

Equation 9 gives the probability that after t iteration the node v_n will also be a infected. Notice that in case of the IM Network many of the path are invalidated if one or more corresponding users of the network in the path ρ_i are offline when the meme is supposed to reach them. Now we consider the question, how does the connectivity of the network affect the acceptance of the meme in the network? First we enunciate the following definitions:

Articulation Node: is a node whose removal from a graph disconnects the graph.

Separate: A set of nodes (or edges) of a graph G is said to separate two nodes u and v of G if the removal of these elements from G produces a graph that lie in different components.

Menger's Theorem: Let u and v be non-adjacent nodes in a graph G, then the minimum number of nodes that separate u and v (Let M be the set of such nodes) is equal to the maximum number of disjoint paths in G.(Weisstein 2006)

Consider a graph which is defined by the nodes and the corresponding edges that are reachable from the origin v_0 at time t, for the IM Network. After each iteration, the number of cliques or components can increase, decrease or remain constant depending the number of agents who are offline. Alternatively the topology of the network can even become like a disconnected graph if many users go offline. Hence the number of nodes that are reachable from the origin v_0 may change since there could be multiple paths of length t connecting the origin and other nodes. Additionally the user who goes offline is represented by a cut node or a set of such nodes S which could separate many other nodes and thus effectively changing the number of nodes that are reachable from R.

In the context of the IM network, Menger's Theorem implies that the number of disjoint paths between different nodes change as the topology of the network changes. Consequently the probability of transfer directly depends upon the minimum size of set M. This can be illustrated by considering the case where the size of M is small and the IM users corresponding to the elements of M go offline before they are infected by the meme. Thus all the nodes from the uninfected part of the graph that are connected by these nodes to the infected side are effectively disconnected from the infected side of the graph. Even if one has a large graph one can still end up in a similar situation. This can be illustrated as follows: Consider a nodes v_i and suppose that S exists, then for$x : x \in M,\ x \notin A_t\ \forall\ t \geq d(v_0, v_i)$ we get $p_i t = 0$. Hence having a highly connected node (hub) in the u-v path is not sufficient to ensure that most of the nodes are infected in a timely manner. This leads to the conclusion that the time at which the hubs go offline or comes online is equally crucial.

5 The Fads Model

One of the assumptions made in our model is that any node that is successfully infected by a meme will always remain infected. If this assumption is dropped then the afore described model can be modified to study "fads" as well. A fad or a "craze" can be defined as a meme such that the meme no longer has a binary value of 1 or 0 but can take a whole range of values. However the value associated with a fad decreases over the course of time unless the node is re-exposed to it. Two variants of the Fad model are considered: The first case is the one in which just as in the original model a node can infect all of its neighbors. In the second case the chance of a fad spreading decreases if it an node is reexposed to it. Additionally the assumption is made that the decay rate of the fads decrease even when the node is inactive.

When a node is exposed to a fad, the value of the fad is initially set to 1. However the passage of each iteration in which the node is not reexposed to the fad causes the value associated with the fad to decrease by a specified decay rate. It should be noted that a decay rate of 0 is equivalent to the original information dissemination model. If the fad's value for a particular node drops to 0, that node will stop exposing others to the fad. Additionally fads decay even when the user is not online. If d is the decay rate then the value of a fad after i iterations can be given as:

$$f = max(1 - d.(i - k), 0),\ 0 \leq\ k\ \leq\ i \quad (10)$$

where k is the last iteration when the node is reexposed to the meme. The two versions of the model are described below:

Fixed Transmission Case: In the fixed transmission case, the value of a fad does not affect its transmissability. The assumption from the original model that an infected node will always expose each other node it comes in contact with to the fad is maintained.

Weakening Transmission Case: In the weakening transmission case, a node is no longer guaranteed to expose every node with which it communicates to the fad. Transmission is instead a random event whose probability is equal to the fad's value in the currently infected node. Each neighbor node of an infected node has an equal chance of getting infected at each iteration.

6 Experiments

We conducted a series of simulation-based experiments to recreate the process of evolution of a social network *i.e.,* how information disseminates in such networks and how fads get proliferated in such networks.

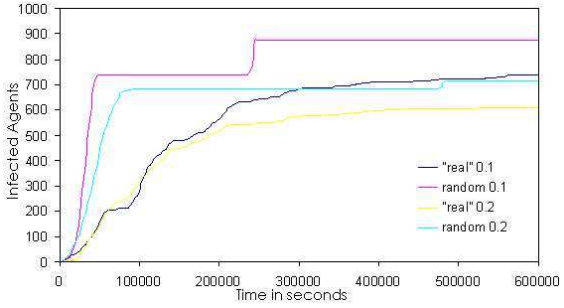


Figure 2: The average results for an actual network for the actual data and random data for a small world network.

6.1 Dataset Overview

The IM status-log data was used as a base for determining the transmissibility of the nodes. Thus an

IM user was randomly assigned to a node and the status of the IM user was used for driving the simulation at each interval. The annonymized dataset used in these experiments was collected by Teredesai et al.(2004) Each user is given a unique identifier in place of their actual screen name.

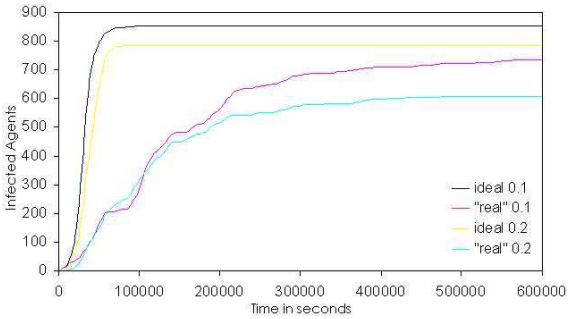


Figure 3: The average results for an ideal and a actual network for a small world network.

6.2 Results for the General Model

In order to compare IM stats in controlled data vs random behavior of agents we conducted additional experiments. A few comparisons of these experiments are described in figure 3 which gives the average results for 100 runs for the ideal network and the actual network for a small world network. It is clear from the figure that the ideal network quickly reaches the saturation point while it takes longer for a actual network to get saturated and the infection is slower. This should not come as a surprise since the situation is analogous to what happens in real life *i.e.,* even though the graph indicates that there is a tipping point, the idea is not readily adopted by everyone in the network. Additional experiments were performed to contrast the results for IM status driven data vs. randomly generated data to see if the results obtained from the previous experiments were not artifacts of the experimental arrangement. Some of the representative results of these experiments are given in Figure 2.

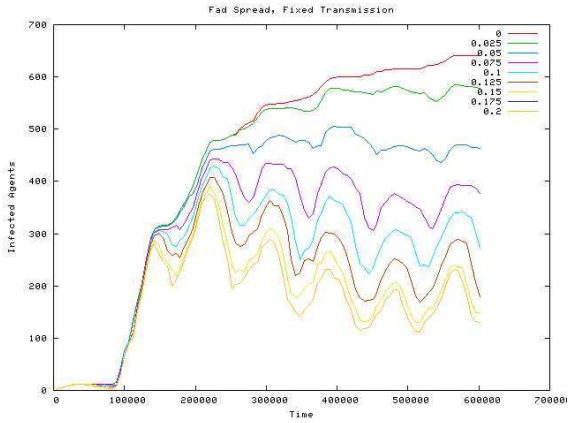


Figure 4: Fixed Transmission Results, varying decay rates.

In is interesting to note that the simulation with randomly generated status data converges earlier as compared to the actual IM Status data. Given this behavior, it is conjected that while the offline/online usage patterns exhibited by IM users are somewhat

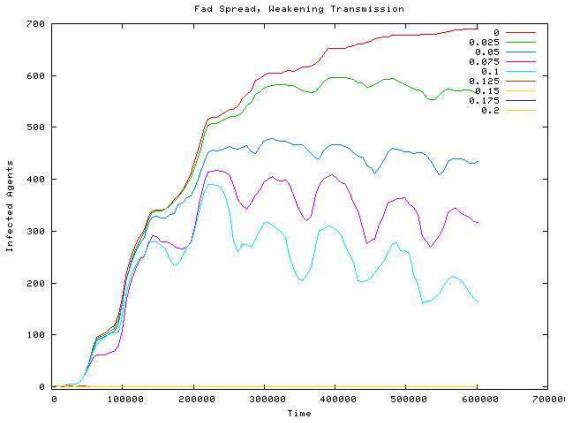


Figure 5: Weakening Transmission Results, varying decay rates.

fixed that leads towards an early saturation, the randomly generated allows those nodes to infect others that would otherwise be offline most of the time. The results also indicate that the status of the IM users is not completely random.

6.3 Results for the Fad Model

A dataset of 996 randomly chosen IM users was used for this series of experiments. Representative results from these experiments and some observations are presented below.

6.3.1 Fixed Transmission Case

The fixed transmission case shows that the count of infected nodes tends to level off at a position lower than the ideal. The leveling-off point comes earlier as the decay rate increases. Higher decay rates also begin to produce sinusoidal-looking curves in the infection count. It could be conjectured that this phenomenon could be mimicking the daily cycle of users going offline during the night. The conjecture is in line with the intuition that fads go in and out of vogue not only depending upon how actively interested people are interested in them but also how much cost they can incur. With very high decay rates, we see that the average infection count begins to slowly drop off after an initial peak. The dataset does not, however, extend far enough in time to allow us to determine if this is another cyclical variation or a permanent loss.

6.3.2 Weakening Transmission Case

It is observed that in the weakening transmission model the network behaves similarly to the original information dissemination model for low decay rates. However the final count of infected nodes is less as compared to the original model if the infection rates are set high. For very high values the initially infected node does not even maintain the fad long enough to infect many of its neighbors thus causing the total count of infected nodes to rapidly fall to 0.

7 Conclusion

In this paper we presented a formal framework to understand information dissemination of fads and non fads in dynamic online social networks. We developed a basic models based on probabilistic interaction between IM users. It was observed that the dissemination of information in a dynamic network depends

upon the level of participation of the nodes in the network. We compared this process for a real world IM Network and an Ideal IM Network where all the users are online most of the time and readily accept an idea when they come across it. It was discovered that not only the connectivity of some of the nodes (hubs) determine how fast the meme is proliferated but also the time-span in which the corresponding person is online or offline. Another important factor in proliferation is the size of the set S as defined in Menger's Theorem. It was also observed that after initial quick proliferation the extent of proliferation of fads is periodical in nature. In conclusion, there is significant scope and need for developing effective models to study the spread of information in online social networks and we formulated two such models for the IM based social networks in this paper.

8 Acknowledgments

We would like to acknowledge Alexander Jarocha-Ernst at the Rochester Institute of Technologyfor suggesting improvements to the fads model and implementing the 'fads' part of the model.

References

Albert, R., Jeong, E., & Barabasi A-L. (1999) Diameter of the World Wide Web Nature 401 130-131,

Carley, K. M., Prietula, Michael J. & Lin, Zhiang (1998) Political Science Convention, April 3-6, Chicago, IL. Design Versus Cognition: The interaction of agent cognition and organizational design on organizational performance. Journal of Artificial Societies and Social Simulation vol. 1, no. 3, http://www.soc.surrey.ac.uk/JASSS/1/3/4.html

Domingos, P., & Richardson, M., (2001) Mining the network value of customers, KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining 57–66 San Francisco, California, http://doi.acm.org/10.1145/502512.502525 ACM Press

Erickson, B., (1981)Secret societies and social structure. Social Forces, 60(1):188210.

Faloutsos C., McCurley K. S., & Tomkins A., (2004) Fast discovery of connection subgraphs, KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, 118–127, Seattle, WA, USA, http://doi.acm.org/10.1145/1014052.1014068,ACM Press.

IMSCAN Website (http://lacuna.rit.edu)

Karp, R., Schindelhauer, C., Shenker, S. & Vocking, B., Randomized (2000) Rumor Spreading. 41st IEEE Symposium on Foundations of Computer Science.

Kautz, H., Selman, B.& Shah, M., (2000) Referral Web: combining social networks and collaborative filtering, Communications of the ACM, v.40 n.3, p.63-65.

Kempe, D., Kleinberg, J., & Tardos, E., (2003) Maximizing the spread of influence through a social network KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.

Kempe, D., Kleinberg, J., & Demers, A.,(2003) Spatial gossip and resource location protocols. Proc. 33rd ACM Symposium on Theory of Computing, 2003 137–146 Washington, D.C. http://doi.acm.org/10.1145/956750.956769 ACM Press

Moore, C., & Newman, M., (2000) Epidemics and Percolation in Small-World Networks, Cristopher , Santa Fe Institute Working Papers 00-01-002

Sageman, M. (2004) Understanding Terror Networks. University of Pennsylvania Press.

Teredesai, A., Resig, J., Dawara, S., & Homan, C., (2004) Extracting Social Networks from Instant Messaging Populations. KDD 2004 Link Discovery Workshop.

Teredesai, A., Resig, J.,(2004) A Framework for Mining Instant Messaging Services. SIAM DM 2004 Workshop on Link Analysis, Counter-Terrorism & Privacy.

Arrow, K. J.(1963) Uncertainty and the Welfare Economics of Medical Care. American Economic Review.

Levitt S., & Dubner, S., (2005) Freakonomics: A Rouge Economist Explores the Hidden Side of Everything. William Morrow/HarperCollins.

Watts, D. J., & Strogatz, S. H., (1998) Collective dynamics ofsmall-world networks, Nature 393, 440442.

Weisstein, E. W., "Menger's n-Arc Theorem." From MathWorld–A Wolfram Web Resource. (http://mathworld.wolfram.com/MengersnArcTheorem.html)

Young, H. P., (2000) The Diffusion of Innovations in Social Networks, Economics Working Paper Archive 437, The Johns Hopkins University,Department of Economics.

Author Index

Recent Volumes in the CRPIT Series

ISSN 1445-1336

Listed below are some of the latest volumes published in the ACS Series *Conferences in Research and Practice in Information Technology*. The full text of most papers (in either PDF or Postscript format) is available at the series website `http://crpit.com`.

Volume 41 - Theory of Computing 2005
Edited by Mike Atkinson, *University of Otago, New Zealand* and Frank Dehne, *Griffith University, Australia*. January, 2005. 1-920-68223-6.
Contains the papers presented at the Eleventh Computing: The Australasian Theory Symposium (CATS2005), Newcastle, NSW, Australia, January/February 2005.

Volume 42 - Computing Education 2005
Edited by Alison L. Young, *UNITEC, New Zealand* and Denise Tolhurst, *University of New South Wales, Australia*. January, 2005. 1-920-68224-4.
Contains the papers presented at the Seventh Australasian Computing Education Conference (ACE2005), Newcastle, NSW, Australia, January/February 2005.

Volume 43 - Conceptual Modelling 2005
Edited by Sven Hartmann, *Massey University, New Zealand* and Markus Stumptner, *University of South Australia*. January, 2005. 1-920-68225-2.
Contains the papers presented at the Second Asia-Pacific Conference on Conceptual Modelling (APCCM2005), Newcastle, NSW, Australia, January/February 2005.

Volume 44 - ACSW Frontiers 2005
Edited by Rajkumar Buyya, *University of Melbourne*, Paul Coddington, *University of Adelaide*, Paul Montague, *Motorola Australia Software Centre*, Rei Safavi-Naini, *University of Wollongong*, Nicholas Sheppard, *University of Wollongong* and Andrew Wendelborn, *University of Adelaide*. January, 2005. 1-920-68226-0.
Contains the papers presented at the Australasian Workshop on Grid Computing and e-Research (AusGrid 2005) and the Third Australasian Information Security Workshop (AISW 2005), Newcastle, NSW, Australia, January/February 2005.

Volume 45 - Information Visualisation 2005
Edited by Seok-Hee Hong *NICTA, Australia*. January, 2005. 1-920-68227-9.
Contains the papers presented at the Asia-Pacific Symposium on Information Visualisation, APVis.au, Sydney, Australia, January 2005.

Volume 46 - ICT in Education
Edited by Graham Low *University of New South Wales, Australia*. October, 2005. 1-920-68228-7.
Contains selected refereed papers presented at the South East Asia Regional Computer Confederation (SEARCC) 2005: ICT Building Bridges Conference, Sydney, Australia, September 2005.

Volume 47 - Safety Critical Systems and Software 2004
Edited by Tony Cant, *University of Queensland*. March, 2005. 1-920-68229-5.
Contains all papers presented at the Ninth Australian Workshop on Safety-Related Programmable Systems, (SCS2004), Brisbane, Australia, October 2004.

Volume 48 - Computer Science 2006
Edited by Vladimir Estivill-Castro, *Griffith University* and Gillian Dobbie, *University of Auckland, New Zealand*. January, 2006. 1-920-68230-9.
Contains the papers presented at the Twenty-Ninth Australasian Computer Science Conference (ACSC2006), Hobart, Tasmania, Australia, January 2006.

Volume 49 - Database Technologies 2006
Edited by Gillian Dobbie, *University of Auckland, New Zealand* and James Bailey, *University of Melbourne*. January, 2006. 1-920-68231-7.
Contains the papers presented at the Seventeenth Australasian Database Conference (ADC2006), Hobart, Tasmania, Australia, January 2006.

Volume 50 - User Interfaces 2006
Edited by Wayne Piekarski, *University of South Australia*. January, 2006. 1-920-68232-5.
Contains the papers presented at the Seventh Australasian User Interface Conference (AUIC2006), Hobart, Tasmania, Australia, January 2006.

Volume 51 - Theory of Computing 2006
Edited by Barry Jay *UTS, Australia* and Joachim Gudmundsson, *NICTA, Australia*. January, 2006. 1-920-68233-3.
Contains the papers presented at the Twelfth Computing: The Australasian Theory Symposium (CATS2006), Hobart, Tasmania, Australia, January 2006.

Volume 52 - Computing Education 2006
Edited by Denise Tolhurst, *University of New South Wales, Australia* and Samuel Mann, *Otago Polytechnic, Otago, New Zealand*. January, 2006. 1-920-68234-1.
Contains the papers presented at the Eighth Australasian Computing Education Conference (ACE2006), Hobart, Tasmania, Australia, January 2006.

Volume 53 - Conceptual Modelling 2006
Edited by Markus Stumptner, *University of South Australia*, Sven Hartmann, *Massey University, New Zealand* and Yasushi Kiyoki*Keio University, Japan*. January, 2006. 1-920-68235-X.
Contains the papers presented at the Third Asia-Pacific Conference on Conceptual Modelling (APCCM2006), Hobart, Tasmania, Australia, January 2006.

Volume 54 - ACSW Frontiers 2006
Edited by Rajkumar Buyya, *University of Melbourne*, Tianchi Ma, *University of Melbourne*, Rei Safavi-Naini, *University of Wollongong*, Chris Steketee, *University of South Australia* and Willy Susilo, *University of Wollongong*. January, 2006. 1-920-68236-8.
Contains the papers presented at the Fourth Australasian Workshop on Grid Computing and e-Research (AusGrid 2006) and the Fourth Australasian Information Security Workshop (AISW 2006), Hobart, Tasmania, Australia, January 2006.

Volume 55 - Safety Critical Systems and Software 2005
Edited by Tony Cant, *University of Queensland*. Late 2005. 1-920-68237-6.
Contains all papers presented at the 10th Australian Workshop on Safety Related Programmable Systems, August 2005, Sydney, Australia.

Volume 56 - Visual Information Processing 2005
Edited by Hong Yan, *City University of Hong Kong*, Jesse Jin, *University of Newcastle, Australia*, Zhiqiang Liu, *City University of Hong Kong* and Daniel Yeung, *Hong Kong Polytechnic University*. Late 2005. 1-920-68238-4.
Contains papers from the Asia-Pacific Workshop on Visual Information Processing (VIP2005), Hong Kong, December 2005.

Volume 57 - Multimodal User Interaction 2005
Edited by Fang Chen and Julien Epps *National ICT Australia*. December, 2005. 1-920-68239-2.
Contains the proceedings of the Multimodal User Interaction Workshop 2005, NICTA-HCSNet, Sydney, Australia, 13-14 September 2005.

Volume 58 - Advances in Ontologies 2005
Edited by Thomas Meyer, *National ICT Australia, Sydney* and Mehmet Orgun *Macquarie University*. December, 2005. 1-920-68240-6.
Contains the proceedings of the Australasian Ontology Workshop (AOW 2005), Sydney, Australia, 6 December 2005.