Graham J. Williams  Simeon J. Simoff (Eds.)

Data Mining

Theory, Methodology, Techniques, and Applications

Springer
Preface

Data mining has been an area of considerable research and application in Australia and the region for many years. This has resulted in the establishment of a strong tradition of academic and industry scholarship, blended with the pragmatics of practice in the field of data mining and analytics. ID3, See5, RuleQuest.com, MagnumOpus, and WEKA is but a short list of the data mining tools and technologies that have been developed in Australasia. Data mining conferences held in Australia have attracted considerable international interest and involvement.

This book brings together a unique collection of chapters that cover the breadth and depth of data mining today. This volume provides a snapshot of the current state of the art in data mining, presenting it both in terms of technical developments and industry applications. Authors include some of Australia’s leading researchers and practitioners in data mining, together with chapters from regional and international authors.

The collection of chapters is based on works presented at the Australasian Data Mining conference series and industry forums. The original papers were initially reviewed for the workshops, conferences and forums. Presenting authors were provided with substantial feedback, both through this initial review process and through editorial feedback from their presentations. A final international peer review process was conducted to include input from potential users of the research, and in particular analytics experts from industry, looking at the impact of reviewed works.

Many people contribute to an effort such as this, starting with the authors! We thank all authors for their contributions, and particularly for making the effort to address two rounds of reviewer comments. Our workshop and conference reviewers provided the first round of helpful feedback for the presentation of the papers to their respective conferences. The authors from a selection of the best papers were then invited to update their contributions for inclusion in this volume. Each submission was then reviewed by at least another two reviewers from our international panel of experts in data mining.

A considerable amount of effort goes into reviewing papers, and reviewers perform an essential task. Reviewers receive no remuneration for all their efforts, but are happy to provide their time and expertise for the benefit of the whole community. We owe a considerable debt to them all and thank them for their enthusiasm and critical efforts.

Bringing this collection together has been quite an effort. We also acknowledge the support of our respective institutions and colleagues who have contributed in many different ways. In particular, Graham would like to thank Togaware (Data Mining and GNU/Linux consultancy) for their ongoing infrastructural support over the years, and the Australian Taxation Office for its
support of data mining and related local conferences through the participation of its staff. Simeon acknowledges the support of the University of Technology, Sydney. The Australian Research Council’s Research Network on Data Mining and Knowledge Discovery, under the leadership of Professor John Roddick, Flinders University, has also provided support for the associated conferences, in particular providing financial support to assist student participation in the conferences. Professor Geoffrey Webb, Monash University, has played a supportive role in the development of data mining in Australia and the AusDM series of conferences, and continues to contribute extensively to the conference series.

The book is divided into two parts: (i) state-of-art research and (ii) state-of-art industry applications. The chapters are further grouped around common sub-themes. We are sure you will find that the book provides an interesting and broad update on current research and development in data mining.

November 2005

Graham Williams and Simeon Simoff
Many colleagues have contributed to the success of the series of data mining workshops and conferences over the years. We list here the primary reviewers who now make up the International Panel of Expert Reviewers.

AusDM Conference Chairs

Simeon J. Simoff, University of Technology, Sydney, Australia
Graham J. Williams, Australian National University, Canberra

PAKDD Industry Chair

Graham J. Williams, Australian National University, Canberra

International Panel of Expert Reviewers

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mihael Ankerst</td>
<td>Boeing Corp., USA</td>
</tr>
<tr>
<td>Michael Bain</td>
<td>University of New South Wales, Australia</td>
</tr>
<tr>
<td>Rohan Baxter</td>
<td>Australian Taxation Office</td>
</tr>
<tr>
<td>Helmut Berger</td>
<td>University of Technology, Sydney, Australia</td>
</tr>
<tr>
<td>Michael Bohlen</td>
<td>Free University Bolzano-Bozen, Italy</td>
</tr>
<tr>
<td>Jie Chen</td>
<td>CSIRO, Canberra, Australia</td>
</tr>
<tr>
<td>Peter Christen</td>
<td>Australian National University</td>
</tr>
<tr>
<td>Thanh-Nghi Do</td>
<td>Can Tho University, Vietnam</td>
</tr>
<tr>
<td>Vladimir Estivill-Castro</td>
<td>Giffith University, Australia</td>
</tr>
<tr>
<td>Hongjian Fan</td>
<td>University of Melbourne, Australia</td>
</tr>
<tr>
<td>Eibe Frank</td>
<td>Waikato University, New Zealand</td>
</tr>
<tr>
<td>Mohamed Medhat Gaber</td>
<td>Monash University, Australia</td>
</tr>
<tr>
<td>Raj Gopalan</td>
<td>Curtin University, Australia</td>
</tr>
<tr>
<td>Warwick Graco</td>
<td>Australian Taxation Office</td>
</tr>
<tr>
<td>Lifang Gu</td>
<td>Australian Taxation Office</td>
</tr>
<tr>
<td>Hongxing He</td>
<td>CSIRO, Canberra, Australia</td>
</tr>
<tr>
<td>Robert Hilderman</td>
<td>University of Regina, Canada</td>
</tr>
<tr>
<td>Joshua Zhexue Huang</td>
<td>University of Hong Kong, China</td>
</tr>
<tr>
<td>Huidong Jin</td>
<td>CSIRO, Canberra, Australia</td>
</tr>
<tr>
<td>Paul Kennedy</td>
<td>University of Technology, Sydney, Australia</td>
</tr>
<tr>
<td>Weiqiang Lin</td>
<td>Australian Taxation Office</td>
</tr>
<tr>
<td>John Maindonald</td>
<td>Australian National University</td>
</tr>
<tr>
<td>Mark Norrie</td>
<td>Teradata, NCR, Australia</td>
</tr>
<tr>
<td>Peter O’Hanlon</td>
<td>Westpac, Australia</td>
</tr>
</tbody>
</table>
Mehmet Orgun  Macquarie University, Australia
Tom Osborn  Wunderman, NUIX Pty Ltd, Australia
Robert Pearson  Health Insurance Commission, Australia
Francois Poulet  ESIEA-Pole ECD, Laval, France
John Roddick  Flinders University, Australia
Greg Saunders  University of Ballarat, Australia
David Skillicorn  Queen’s University, Canada
Geoffrey Webb  Monash University, Australia
John Yearwood  University of Ballarat, Australia
Osmar Zaiane  University of Alberta, Canada
# Part 1: State-of-the-Art in Research

## Methodological Advances

**Generality Is Predictive of Prediction Accuracy**  
*Geoffrey I. Webb, Damien Brain* .......................................................... 1

**Visualisation and Exploration of Scientific Data Using Graphs**  
*Ben Raymond, Lee Belbin* ................................................................. 14

**A Case-Based Data Mining Platform**  
*Xingwen Wang, Joshua Zhexue Huang* .................................................. 28

**Consolidated Trees: An Analysis of Structural Convergence**  
*Jesús M. Pérez, Javier Muguerza, Olatz Arbelaitz, Ibai Gurrutxaga, José I. Martín* .......................................................... 39

**K Nearest Neighbor Edition to Guide Classification Tree Learning: Motivation and Experimental Results**  
*J.M. Martínez-Otzeta, B. Sierra, E. Lazzano, A. Astigarraga* ............. 53

**Efficiently Identifying Exploratory Rules’ Significance**  
*Shiying Huang, Geoffrey I. Webb* ....................................................... 64

**Mining Value-Based Item Packages – An Integer Programming Approach**  
*N.R. Achuthan, Raj P. Gopalan, Amit Rudra* ......................................... 78

**Decision Theoretic Fusion Framework for Actionability Using Data Mining on an Embedded System**  
*Heungkyu Lee, Sunmee Kang, Hanseok Ko* ........................................... 90

**Use of Data Mining in System Development Life Cycle**  
*Richi Nayak, Tian Qiu* .............................................................. 105

**Mining MOUCLAS Patterns and Jumping MOUCLAS Patterns to Construct Classifiers**  
*Yalei Hao, Gerald Quirchmayr, Markus Stumptner* ......................... 118
# Table of Contents

## Data Linkage

A Probabilistic Geocoding System Utilising a Parcel Based Address File  
*Peter Christen, Alan Willmore, Tim Churches* ........................................ 130

Decision Models for Record Linkage  
*Lifang Gu, Rohan Baxter* ................................................................. 146

## Text Mining

Intelligent Document Filter for the Internet  
*Deepani B. Guruge, Russel J. Stonier* .................................................. 161

Informing the Curious Negotiator: Automatic News Extraction from the Internet  
*Debbie Zhang, Simeon J. Simoff* ......................................................... 176

Text Mining for Insurance Claim Cost Prediction  
*Inna Kolyshkina, Marcel van Rooyen* .................................................. 192

## Temporal and Sequence Mining

An Application of Time-Changing Feature Selection  
*Yihao Zhang, Mehmet A. Orgun, Weiqiang Lin, Warwick Graco* ........... 203

A Data Mining Approach to Analyze the Effect of Cognitive Style and Subjective Emotion on the Accuracy of Time-Series Forecasting  
*Hung Kook Park, Byoungho Song, Hyeon-Joong Yoo, Dae Woong Rhee, Kang Ryoung Park, Juno Chang* .................. 218

A Multi-level Framework for the Analysis of Sequential Data  
*Carl H. Mooney, Denise de Vries, John F. Roddick* ............................. 229

## Part 2: State-of-the-Art in Applications

### Health

Hierarchical Hidden Markov Models: An Application to Health Insurance Data  
*Ah Chung Tsoi, Shu Zhang, Markus Hagenbuchner* .............................. 244
Identifying Risk Groups Associated with Colorectal Cancer
   Jie Chen, Hongxing He, Huidong Jin, Damien McAullay,
   Graham Williams, Chris Kelman ................................. 260

Mining Quantitative Association Rules in Protein Sequences
   Nitin Gupta, Nitin Mangal, Kamal Tiwari, Pabitra Mitra .......... 273

Mining X-Ray Images of SARS Patients
   Xuanyang Xie, Xi Li, Shouhong Wan, Yuchang Gong ............. 282

Finance and Retail

The Scamseek Project – Text Mining for Financial Scams on the Internet
   Jon Patrick .......................................................... 295

A Data Mining Approach for Branch and ATM Site Evaluation
   Simon C.K. Shiu, James N.K. Liu, Jennie L.C. Lam, Bo Feng .... 303

The Effectiveness of Positive Data Sharing in Controlling the Growth
of Indebtedness in Hong Kong Credit Card Industry
   Vincent To-Yee Ng, Wai Tak Yim, Stephen Chi-Fai Chan ........ 319

Author Index .......................................................... 331