

# Lecture Notes in Artificial Intelligence 3755

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

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

# Data Mining

Theory, Methodology, Techniques,  
and Applications



Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA  
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Graham J. Williams  
Togaware Data Mining  
Canberra, Australia  
E-mail: graham.williams@togaware.com

Simeon J. Simoff  
University of Technology, Faculty of Information Technology  
Sydney Broadway PO Box 123, NSW 2007, Australia  
E-mail: simeon@it.uts.edu.au

Library of Congress Control Number: 2006920576

CR Subject Classification (1998): I.2, H.2.8, H.2-3, D.3.3, F.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN	0302-9743
ISBN-10	3-540-32547-6 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-32547-5 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 11677437 06/3142 5 4 3 2 1 0

# 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

# Organization

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

Mihael Ankerst	Boeing Corp., USA
Michael Bain	University of New South Wales, Australia
Rohan Baxter	Australian Taxation Office
Helmut Berger	University of Technology, Sydney, Australia
Michael Bohlen	Free University Bolzano-Bozen, Italy
Jie Chen	CSIRO, Canberra, Australia
Peter Christen	Australian National University
Thanh-Nghi Do	Can Tho University, Vietnam
Vladimir Estivill-Castro	Giffith University, Australia
Hongjian Fan	University of Melbourne, Australia
Eibe Frank	Waikato University, New Zealand
Mohamed Medhat Gaber	Monash University, Australia
Raj Gopalan	Curtin University, Australia
Warwick Graco	Australian Taxation Office
Lifang Gu	Australian Taxation Office
Hongxing He	CSIRO, Canberra, Australia
Robert Hilderman	University of Regina, Canada
Joshua Zhexue Huang	University of Hong Kong, China
Huidong Jin	CSIRO, Canberra, Australia
Paul Kennedy	University of Technology, Sydney, Australia
Weiqiang Lin	Australian Taxation Office
John Maindonald	Australian National University
Mark Norrie	Teradata, NCR, Australia
Peter O'Hanlon	Westpac, Australia

Mehmet Orgun  
Tom Osborn  
Robert Pearson  
Francois Poulet  
John Roddick  
Greg Saunders  
David Skillicorn  
Geoffrey Webb  
John Yearwood  
Osmar Zaiane

Macquarie University, Australia  
Wunderman, NUIX Pty Ltd, Australia  
Health Insurance Commission, Australia  
ESIEA-Pole ECD, Laval, France  
Flinders University, Australia  
University of Ballarat, Australia  
Queen's University, Canada  
Monash University, Australia  
University of Ballarat, Australia  
University of Alberta, Canada

# Table of Contents

## Part 1: State-of-the-Art in Research

### Methodological Advances

Generality Is Predictive of Prediction Accuracy <i>Geoffrey I. Webb, Damien Brain</i> .....	1
Visualisation and Exploration of Scientific Data Using Graphs <i>Ben Raymond, Lee Belbin</i> .....	14
A Case-Based Data Mining Platform <i>Xingwen Wang, Joshua Zhexue Huang</i> .....	28
Consolidated Trees: An Analysis of Structural Convergence <i>Jesús M. Pérez, Javier Muguerza, Olatz Arbelaiz, Ibai Gurrutxaga, José I. Martín</i> .....	39
K Nearest Neighbor Edition to Guide Classification Tree Learning: Motivation and Experimental Results <i>J.M. Martínez-Otzeta, B. Sierra, E. Lazkano, A. Astigarraga</i> .....	53
Efficiently Identifying Exploratory Rules' Significance <i>Shiying Huang, Geoffrey I. Webb</i> .....	64
Mining Value-Based Item Packages – An Integer Programming Approach <i>N.R. Achuthan, Raj P. Gopalan, Amit Rudra</i> .....	78
Decision Theoretic Fusion Framework for Actionability Using Data Mining on an Embedded System <i>Heungkyu Lee, Sunmee Kang, Hanseok Ko</i> .....	90
Use of Data Mining in System Development Life Cycle <i>Richi Nayak, Tian Qiu</i> .....	105
Mining MOUCLAS Patterns and Jumping MOUCLAS Patterns to Construct Classifiers <i>Yalei Hao, Gerald Quirchmayr, Markus Stumptner</i> .....	118



## Data Linkage

A Probabilistic Geocoding System Utilising a Parcel Based Address File <i>Peter Christen, Alan Willmore, Tim Churches</i> .....	130
--	-----

Decision Models for Record Linkage <i>Lifang Gu, Rohan Baxter</i> .....	146
--	-----

## Text Mining

Intelligent Document Filter for the Internet <i>Deepani B. Guruge, Russel J. Stonier</i> .....	161
---	-----

Informing the Curious Negotiator: Automatic News Extraction from the Internet <i>Debbie Zhang, Simeon J. Simoff</i> .....	176
---	-----

Text Mining for Insurance Claim Cost Prediction <i>Inna Kolyshkina, Marcel van Rooyen</i> .....	192
--	-----

## Temporal and Sequence Mining

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

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

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

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

### Health

Hierarchical Hidden Markov Models: An Application to Health Insurance Data <i>Ah Chung Tsoi, Shu Zhang, Markus Hagenbuchner</i> .....	244
---	-----

Identifying Risk Groups Associated with Colorectal Cancer <i>Jie Chen, Hongxing He, Huidong Jin, Damien McAullay, Graham Williams, Chris Kelman</i> .....	260
Mining Quantitative Association Rules in Protein Sequences <i>Nitin Gupta, Nitin Mangal, Kamal Tiwari, Pabitra Mitra</i> .....	273
Mining X-Ray Images of SARS Patients <i>Xuanyang Xie, Xi Li, Shouhong Wan, Yuchang Gong</i> .....	282
 <b>Finance and Retail</b>	
The Scamseek Project – Text Mining for Financial Scams on the Internet <i>Jon Patrick</i> .....	295
A Data Mining Approach for Branch and ATM Site Evaluation <i>Simon C.K. Shiu, James N.K. Liu, Jennie L.C. Lam, Bo Feng</i> .....	303
The Effectiveness of Positive Data Sharing in Controlling the Growth of Indebtedness in Hong Kong Credit Card Industry <i>Vincent To-Yee Ng, Wai Tak Yim, Stephen Chi-Fai Chan</i> .....	319
<b>Author Index</b> .....	331