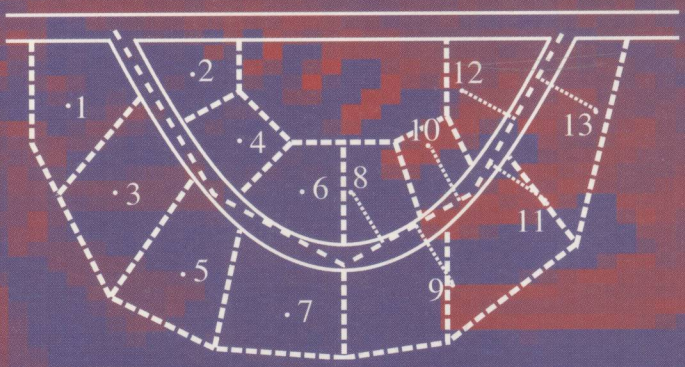
State-of-the-Art Survey

LNAI 3755

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

# Data Mining

## Theory, Methodology, Techniques, and Applications




Springer

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 |



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

Lecture Notes in Artificial Intelligence 3755

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

# Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3874: R. Missaoui, J. Schmidt (Eds.), Formal Concept Analysis. X, 309 pages. 2006.

Vol. 3863: M. Kohlhase (Ed.), Mathematical Knowledge Management. XI, 405 pages. 2006.

Vol. 3848: J.-F. Boulicaut, L. De Raedt, H. Mannila (Eds.), Constraint-Based Mining and Inductive Databases. X, 401 pages. 2006.

Vol. 3847: K.P. Jantke, A. Lunzer, N. Spyratos, Y. Tanaka (Eds.), Federation over the Web. X, 215 pages. 2006.

Vol. 3835: G. Sutcliffe, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIV, 744 pages. 2005.

Vol. 3817: M. Faundez-Zanuy, L. Janer, A. Esposito, A. Satue-Villar, J. Roure, V. Espinosa-Duro (Eds.), Nonlinear Analyses and Algorithms for Speech Processing. XII, 380 pages. 2006.

Vol. 3814: M. Maybury, O. Stock, W. Wahlster (Eds.), Intelligent Technologies for Interactive Entertainment. XV, 342 pages. 2005.

Vol. 3809: S. Zhang, R. Jarvis (Eds.), AI 2005: Advances in Artificial Intelligence. XXVII, 1344 pages. 2005.

Vol. 3808: C. Bento, A. Cardoso, G. Dias (Eds.), Progress in Artificial Intelligence. XVIII, 704 pages. 2005.

Vol. 3802: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), Computational Intelligence and Security, Part II. XLII, 1166 pages. 2005.

Vol. 3801: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), Computational Intelligence and Security, Part I. XLI, 1122 pages. 2005.

Vol. 3789: A. Gelbukh, Á. de Albornoz, H. Terashima-Marín (Eds.), MICAI 2005: Advances in Artificial Intelligence. XXVI, 1198 pages. 2005.

Vol. 3782: K.-D. Althoff, A. Dengel, R. Bergmann, M. Nick, T.R. Roth-Berghofer (Eds.), Professional Knowledge Management. XXIII, 739 pages. 2005.

Vol. 3763: H. Hong, D. Wang (Eds.), Automated Deduction in Geometry. X, 213 pages. 2006.

Vol. 3755: G.J. Williams, S.J. Simoff (Eds.), Data Mining. XI, 331 pages. 2006.

Vol. 3735: A. Hoffmann, H. Motoda, T. Scheffer (Eds.), Discovery Science. XVI, 400 pages. 2005.

Vol. 3734: S. Jain, H.U. Simon, E. Tomita (Eds.), Algorithmic Learning Theory. XII, 490 pages. 2005.

Vol. 3721: A.M. Jorge, L. Torgo, P.B. Brazdil, R. Camacho, J. Gama (Eds.), Knowledge Discovery in Databases: PKDD 2005. XXIII, 719 pages. 2005.

Vol. 3720: J. Gama, R. Camacho, P.B. Brazdil, A.M. Jorge, L. Torgo (Eds.), Machine Learning: ECML 2005. XXIII, 769 pages. 2005.

Vol. 3717: B. Gramlich (Ed.), Frontiers of Combining Systems. X, 321 pages. 2005.

Vol. 3702: B. Beckert (Ed.), Automated Reasoning with Analytic Tableaux and Related Methods. XIII, 343 pages. 2005.

Vol. 3698: U. Furbach (Ed.), KI 2005: Advances in Artificial Intelligence. XIII, 409 pages. 2005.

Vol. 3690: M. Pěchouček, P. Petta, L.Z. Varga (Eds.), Multi-Agent Systems and Applications IV. XVII, 667 pages. 2005.

Vol. 3684: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part IV. LXXIX, 933 pages. 2005.

Vol. 3683: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. LXXX, 1397 pages. 2005.

Vol. 3682: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. LXXIX, 1371 pages. 2005.

Vol. 3681: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LXXX, 1319 pages. 2005.

Vol. 3673: S. Bandini, S. Manzoni (Eds.), AI*IA 2005: Advances in Artificial Intelligence. XIV, 614 pages. 2005.

Vol. 3662: C. Baral, G. Greco, N. Leone, G. Terracina (Eds.), Logic Programming and Nonmonotonic Reasoning. XIII, 454 pages. 2005.

Vol. 3661: T. Panayiotopoulos, J. Gratch, R.S. Aylett, D. Ballin, P. Olivier, T. Rist (Eds.), Intelligent Virtual Agents. XIII, 506 pages. 2005.

Vol. 3658: V. Matoušek, P. Mautner, T. Pavelka (Eds.), Text, Speech and Dialogue. XV, 460 pages. 2005.

Vol. 3651: R. Dale, K.-F. Wong, J. Su, O.Y. Kwong (Eds.), Natural Language Processing – IJCNLP 2005. XXI, 1031 pages. 2005.

Vol. 3642: D. Ślęzak, J. Yao, J.F. Peters, W. Ziarko, X. Hu (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Part II. XXIII, 738 pages. 2005.

Vol. 3641: D. Ślęzak, G. Wang, M. Szczuka, I. Düntsch, Y. Yao (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Part I. XXIV, 742 pages. 2005.

Vol. 3635: J.R. Winkler, M. Niranjan, N.D. Lawrence (Eds.), Deterministic and Statistical Methods in Machine Learning. VIII, 341 pages. 2005.

Vol. 3632: R. Nieuwenhuis (Ed.), Automated Deduction – CADE-20. XIII, 459 pages. 2005.

Vol. 3630: M.S. Capcarrère, A.A. Freitas, P.J. Bentley, C.G. Johnson, J. Timmis (Eds.), Advances in Artificial Life. XIX, 949 pages. 2005.

Vol. 3626: B. Ganter, G. Stumme, R. Wille (Eds.), Formal Concept Analysis. X, 349 pages. 2005.

Vol. 3625: S. Kramer, B. Pfahringer (Eds.), Inductive Logic Programming. XIII, 427 pages. 2005.

Vol. 3620: H. Muñoz-Ávila, F. Ricci (Eds.), Case-Based Reasoning Research and Development. XV, 654 pages. 2005.

Vol. 3614: L. Wang, Y. Jin (Eds.), Fuzzy Systems and Knowledge Discovery, Part II. XLI, 1314 pages. 2005.

Vol. 3613: L. Wang, Y. Jin (Eds.), Fuzzy Systems and Knowledge Discovery, Part I. XLI, 1334 pages. 2005.

Vol. 3607: J.-D. Zucker, L. Saitta (Eds.), Abstraction, Reformulation and Approximation. XII, 376 pages. 2005.

Vol. 3601: G. Moro, S. Bergamaschi, K. Aberer (Eds.), Agents and Peer-to-Peer Computing. XII, 245 pages. 2005.

Vol. 3600: F. Wiedijk (Ed.), The Seventeen Provers of the World. XVI, 159 pages. 2006.

Vol. 3596: F. Dau, M.-L. Mugnier, G. Stumme (Eds.), Conceptual Structures: Common Semantics for Sharing Knowledge. XI, 467 pages. 2005.

Vol. 3593: V. Mařík, R. W. Brennan, M. Pěchouček (Eds.), Holonic and Multi-Agent Systems for Manufacturing. XI, 269 pages. 2005.

Vol. 3587: P. Perner, A. Imiya (Eds.), Machine Learning and Data Mining in Pattern Recognition. XVII, 695 pages. 2005.

Vol. 3584: X. Li, S. Wang, Z.Y. Dong (Eds.), Advanced Data Mining and Applications. XIX, 835 pages. 2005.

Vol. 3581: S. Miksch, J. Hunter, E.T. Keravnou (Eds.), Artificial Intelligence in Medicine. XVII, 547 pages. 2005.

Vol. 3577: R. Falcone, S. Barber, J. Sabater-Mir, M.P. Singh (Eds.), Trusting Agents for Trusting Electronic Societies. VIII, 235 pages. 2005.

Vol. 3575: S. Wermter, G. Palm, M. Elshaw (Eds.), Biomimetic Neural Learning for Intelligent Robots. IX, 383 pages. 2005.

Vol. 3571: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty. XVI, 1028 pages. 2005.

Vol. 3559: P. Auer, R. Meir (Eds.), Learning Theory. XI, 692 pages. 2005.

Vol. 3558: V. Torra, Y. Narukawa, S. Miyamoto (Eds.), Modeling Decisions for Artificial Intelligence. XII, 470 pages. 2005.

Vol. 3554: A.K. Dey, B. Kokinov, D.B. Leake, R. Turner (Eds.), Modeling and Using Context. XIV, 572 pages. 2005.

Vol. 3550: T. Eymann, F. Klügl, W. Lamersdorf, M. Klusch, M.N. Huhns (Eds.), Multiagent System Technologies. XI, 246 pages. 2005.

Vol. 3539: K. Morik, J.-F. Boulicaut, A. Siebes (Eds.), Local Pattern Detection. XI, 233 pages. 2005.

Vol. 3538: L. Ardissono, P. Brna, A. Mitrović (Eds.), User Modeling 2005. XVI, 533 pages. 2005.

Vol. 3533: M. Ali, F. Esposito (Eds.), Innovations in Applied Artificial Intelligence. XX, 858 pages. 2005.

Vol. 3528: P.S. Szczepaniak, J. Kacprzyk, A. Niewiadomski (Eds.), Advances in Web Intelligence. XVII, 513 pages. 2005.

Vol. 3518: T.-B. Ho, D. Cheung, H. Liu (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005.

Vol. 3508: P. Bresciani, P. Giorgini, B. Henderson-Sellers, G. Low, M. Winikoff (Eds.), Agent-Oriented Information Systems II. X, 227 pages. 2005.

Vol. 3505: V. Gorodetsky, J. Liu, V.A. Skormin (Eds.), Autonomous Intelligent Systems: Agents and Data Mining. XIII, 303 pages. 2005.

Vol. 3501: B. Kégl, G. Lapalme (Eds.), Advances in Artificial Intelligence. XV, 458 pages. 2005.

Vol. 3492: P. Blache, E.P. Stabler, J.V. Busquets, R. Moot (Eds.), Logical Aspects of Computational Linguistics. X, 363 pages. 2005.

Vol. 3490: L. Bolc, Z. Michalewicz, T. Nishida (Eds.), Intelligent Media Technology for Communicative Intelligence. X, 259 pages. 2005.

Vol. 3488: M.-S. Hacid, N.V. Murray, Z.W. Raś, S. Tsumoto (Eds.), Foundations of Intelligent Systems. XIII, 700 pages. 2005.

Vol. 3487: J.A. Leite, P. Torroni (Eds.), Computational Logic in Multi-Agent Systems. XII, 281 pages. 2005.

Vol. 3476: J.A. Leite, A. Omicini, P. Torroni, P. Yolum (Eds.), Declarative Agent Languages and Technologies II. XII, 289 pages. 2005.

Vol. 3464: S.A. Brueckner, G.D.M. Serugendo, A. Karageorgos, R. Nagpal (Eds.), Engineering Self-Organising Systems. XIII, 299 pages. 2005.

Vol. 3452: F. Baader, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XI, 562 pages. 2005.

Vol. 3451: M.-P. Gleizes, A. Omicini, F. Zambonelli (Eds.), Engineering Societies in the Agents World V. XIII, 349 pages. 2005.

Vol. 3446: T. Ishida, L. Gasser, H. Nakashima (Eds.), Massively Multi-Agent Systems I. XI, 349 pages. 2005.

Vol. 3445: G. Chollet, A. Esposito, M. Faúndez-Zanuy, M. Marinaro (Eds.), Nonlinear Speech Modeling and Applications. XIII, 433 pages. 2005.

Vol. 3438: H. Christiansen, P.R. Skadhauge, J. Villadsen (Eds.), Constraint Solving and Language Processing. VIII, 205 pages. 2005.

Vol. 3435: P. Faratin, J.A. Rodríguez-Aguilar (Eds.), Agent-Mediated Electronic Commerce VI. XII, 215 pages. 2006.

Vol. 3430: S. Tsumoto, T. Yamaguchi, M. Numao, H. Motoda (Eds.), Active Mining. XII, 349 pages. 2005.

Vol. 3419: B.V. Faltings, A. Petcu, F. Fages, F. Rossi (Eds.), Recent Advances in Constraints. X, 217 pages. 2005.

Vol. 3416: M.H. Böhlen, J. Gamper, W. Polasek, M.A. Wimmer (Eds.), E-Government: Towards Electronic Democracy. XIII, 311 pages. 2005.

Vol. 3415: P. Davidsson, B. Logan, K. Takadama (Eds.), Multi-Agent and Multi-Agent-Based Simulation. X, 265 pages. 2005.

# 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 | 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 |

# Table of Contents

## 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. Lazkano, 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

## 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

# Generality Is Predictive of Prediction Accuracy

Geoffrey I. Webb[1] and Damien Brain[2]

[1] Faculty of Information Technology,
Monash University, Clayton, Vic 3800, Australia
webb@infotech.monash.edu.au
[2] UTelco Systems,
Level 50/120 Collins St Melbourne, Vic 3001, Australia
damien.brain@utelcosystems.com.au

**Abstract.** During knowledge acquisition it frequently occurs that multiple alternative potential rules all appear equally credible. This paper addresses the dearth of formal analysis about how to select between such alternatives. It presents two hypotheses about the expected impact of selecting between classification rules of differing levels of generality in the absence of other evidence about their likely relative performance on unseen data. We argue that the accuracy on unseen data of the more general rule will tend to be closer to that of a default rule for the class than will that of the more specific rule. We also argue that in comparison to the more general rule, the accuracy of the more specific rule on unseen cases will tend to be closer to the accuracy obtained on training data. Experimental evidence is provided in support of these hypotheses. These hypotheses can be useful for selecting between rules in order to achieve specific knowledge acquisition objectives.

## 1 Introduction

In many knowledge acquisition contexts there will be many classification rules that perform equally well on the training data. For example, as illustrated by the version space [1], there will often be alternative rules of differing degrees of generality all of which agree with the training data. However, even when we move away from a situation in which we are expecting to find rules that are strictly consistent with the training data, in other words, when we allow rules to misclassify some training cases, there will often be many rules all of which cover exactly the same training cases. If we are selecting rules to use for some decision making task, we must select between such rules with identical performance on the training data. To do so requires a learning bias [2], a means of selecting between competing hypotheses that utilizes criteria beyond those strictly encapsulated in the training data.

All learning algorithms confront this problem. This is starkly illustrated by the large numbers of rules with very high values for any given interestingness measure that are typically discovered during association rule discovery. Many systems that learn rule sets for the purpose of prediction mask this problem by making arbitrary choices between rules with equivalent performance on the

G.J. Williams and S.J. Simoff (Eds.): Data Mining, LNAI 3755, pp. 1–13, 2006.

training data. This masking of the problem is so successful that many researchers appear oblivious to the problem. Our previous work has clearly identified that it is frequently the case that there exist many variants of the rules typically derived in machine learning, all of which cover exactly the same training data. Indeed, one of our previous systems, The Knowledge Factory [3, 4] provides support for identification and selection between such rule variants.

This paper examines the implications of selecting between such rules on the basis of their relative generality. We contend that learning biases based on relative generality can usefully manipulate the expected performance of classifiers learned from data. The insight that we provide into this issue may assist knowledge engineers make more appropriate selections between alternative rules when those alternatives derive equal support from the available training data.

We present specific hypotheses relating to reasonable expectations about classification error for classification rules. We discuss classification rules of the form $Z \rightarrow y$, which should be interpreted as all cases that satisfy conditions $Z$ belong to class $y$. We are interested in learning rules from data. We allow that evidence about the likely classification performance of a rule might come from many sources, including prior knowledge, but, in the machine learning tradition, are particularly concerned with *empirical* evidence—evidence obtained from the performance of the rule on sample (training) data. We consider the learning context in which a rule $Z \rightarrow y$ is learned from a *training set* $D'=(x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_n, y'_n)$ and is to be applied to a set of previously unseen data called a *test set* $D=(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$. For this enterprise to be successful, $D'$ and $D$ should be drawn from the same or from related distributions. For the purposes of the current paper we assume that $D'$ and $D$ are drawn independently at random from the same distribution and acknowledge that violations of this assumption may affect the effects that we predict.

We utilize the following notation.

- $Z(I)$ represents the set of instances in instance set $I$ covered by condition $Z$.
- $E(Z \rightarrow y, I)$ represents the number of instances in instance set $I$ that $Z \rightarrow y$ misclassifies (the absolute error).
- $\varepsilon(Z \rightarrow y, I)$ represents the proportion of instance set $I$ that $Z \rightarrow y$ misclassifies (the error) $= \frac{E(Z \rightarrow y, I)}{|I|}$.
- $W \gg Z$ denotes that the condition $W$ is a proper generalization of condition $Z$. $W \gg Z$ if and only if the set of descriptions for which $W$ is true is a proper superset of the set of descriptions for which $Z$ is true.
- $NODE(W \rightarrow y, Z \rightarrow y)$ denotes that there is no other distinguishing evidence between $W \rightarrow y$ and $Z \rightarrow y$. This means that there is no available evidence, other than the relative generality of $W$ and $Z$, indicating the likely direction (negative, zero, or positive) of $\varepsilon(W \rightarrow y, D) - \varepsilon(Z \rightarrow y, D)$. In particular, we require that the empirical evidence be identical. In the current research the learning systems have access only to empirical evidence and we assume that $W(D')=Z(D') \rightarrow NODE(W \rightarrow y, Z \rightarrow y)$. Note that $W(D')=Z(D')$ does not preclude $W$ and $Z$ from covering different test cases at classification time and hence having different test set error. We utilize the

notion of *other distinguishing evidence* to allow for the real-world knowledge acquisition context in which evidence other than that contained in the data may be brought to bear upon the rule selection problem.

We present two hypotheses relating to classification rules $W \rightarrow y$ and $Z \rightarrow y$ learned from real-world data such that $W \gg Z$ and $NODE(W \rightarrow y, Z \rightarrow y)$.

1. $Pr(|\varepsilon(W \rightarrow y, D) - \varepsilon(true \rightarrow y, D)| < |\varepsilon(Z \rightarrow y, D) - \varepsilon(true \rightarrow y, D)|) > Pr(|\varepsilon(W \rightarrow y, D) - \varepsilon(true \rightarrow y, D)| > |\varepsilon(Z \rightarrow y, D) - \varepsilon(true \rightarrow y, D)|)$. That is, the error of the more general rule, $W \rightarrow y$, on unseen data will tend to be closer to the proportion of cases in the domain that do not belong to class $y$ than will the error of the more specific rule, $Z \rightarrow y$.
2. $Pr(|\varepsilon(W \rightarrow y, D) - \varepsilon(W \rightarrow y, D')| > |\varepsilon(Z \rightarrow y, D) - \varepsilon(Z \rightarrow y, D')|) > Pr(|\varepsilon(W \rightarrow y, D) - \varepsilon(W \rightarrow y, D')| < |\varepsilon(Z \rightarrow y, D) - \varepsilon(Z \rightarrow y, D')|)$. That is, the error of the more specific rule, $Z \rightarrow y$, on unseen data will tend to be closer to the proportion of negative training cases covered by the two rules[1] than will the error of the more general rule, $W \rightarrow y$.

Another way of stating these two hypotheses is that of two rules with identical empirical and other support,

1. the more general can be expected to exhibit classification error closer to that of a *default rule*, $true \rightarrow y$, or, in other words, of assuming all cases belong to the class, and
2. the more specific can be expected to exhibit classification error closer to that observed on the training data.

It is important to clarify at the outset that we are not claiming that the more general rule will invariably have closer generalization error to the default rule and the more specific rule will invariably have closer generalization error to the observed error on the training data. Rather, we are claiming that relative generality provides a source of evidence that, in the absence of alternative evidence, provides reasonable grounds for believing that each of these effects is more likely than the contrary.

*Observation.* With simple assumptions, hypotheses (1) and (2) can be shown to be trivially true given that $D'$ and $D$ are idd samples from a single finite distribution $\mathcal{D}$.

*Proof.*

1. For any rule $X \rightarrow y$ and test set $D$, $\varepsilon(X \rightarrow y, D) = \varepsilon(X \rightarrow y, X(D))$, as $X \rightarrow y$ only covers instances $X(D)$ of $D$.
2. $\varepsilon(Z \rightarrow y, D) = \frac{E(Z \rightarrow y, Z(D \cap D')) + E(Z \rightarrow y, Z(D - D'))}{|Z(D)|}$
3. $\varepsilon(W \rightarrow y, D) = \frac{E(W \rightarrow y, W(D \cap D')) + E(W \rightarrow y, W(D - D'))}{|W(D)|}$
4. $Z(D) \subseteq W(D)$ because $Z$ is a specialization of $W$.

[1] Recall that both rules have identical empirical support and hence cover the same training cases.

5. $Z(D \cap D') = W(D \cap D')$ because $Z(D') = W(D')$.
6. $Z(D - D') \subseteq W(D - D')$ because $Z(D) \subseteq W(D)$.
7. from 2-6, $E(Z \rightarrow y, Z(D \cap D'))$ is a larger proportion of the error of $Z \rightarrow y$ than is $E(W \rightarrow y, W(D \cap D'))$ of $W \rightarrow y$ and hence performance on $D'$ is a larger component of the performance of $Z \rightarrow y$ and performance on $D - D'$ is a larger component of the performance of $W \rightarrow y$. □

However, in most domains of interest the dimensionality of the instance space will be very high. In consequence, for realistic training and test sets the proportion of the training set that appears in the test set, $\frac{|D \cap D'|}{|D|}$, will be small. Hence this effect will be negligible, as performance on the training set will be a negligible portion of total performance. What we are more interested in is off-training-set error. We contend that the force of these hypotheses will be stronger than accounted for by the difference made by the overlap between training and test sets, and hence that they do apply to off-training-set error. We note, however, that it is trivial to construct no-free-lunch proofs, such as those of Wolpert [5] and Schaffer [6], that this is not, in general, true. Rather, we contend that the hypotheses will in general be true for 'real-world' learning tasks. We justify this contention by recourse to the similarity assumption [7], that in the absence of other information, the greater the similarity between two objects in other respects, the greater the probability of their both belonging to the same class. We believe that most machine learning algorithms depend upon this assumption, and that this assumption is reasonable for real-world knowledge acquisition tasks. Test set cases covered by a more general but not a more specific rule are likely to be less similar to training cases covered by both rules than are test set cases covered by the more specific rule. Hence satisfying the left-hand-side of the more specific rule provides stronger evidence of likely class membership.

A final point that should be noted is that these hypotheses apply to individual classification rules — structures that associate an identified region of an instance space with a single class. However, as will be discussed in more detail below, we believe that the principle is nonetheless highly relevant to 'complete classifiers,' such as decision trees, that assign different regions of the instance space to different classes. This is because each individual region within a 'complete classifier' (such as a decision tree leaf) satisfies our definition of a classification rule, and hence the hypotheses can cast light on the likely consequences of relabeling subregions of the instance space within such a classifier (for example, generalizing one leaf of a decision tree at the expense of another, as proposed elsewhere [8]).

## 2 Evaluation

To evaluate these hypotheses we sought to generate rules of varying generality but identical empirical evidence (no other evidence source being considered in the research), and to test the hypotheses' predictions with respect to these rules.

We wished to provide some evaluation both of whether the predicted effects are general (with respect to rules with the relevant properties selected at random)