Michael R. Berthold
John Shawe-Taylor
Nada Lavrač (Eds.)

Advances in Intelligent Data Analysis VII

LNCS 4723

7th International Symposium
on Intelligent Data Analysis, IDA 2007
Ljubljana, Slovenia, September 2007, Proceedings

IDA 07

Michael R. Berthold John Shawe-Taylor
Nada Lavrač (Eds.)

Advances in Intelligent Data Analysis VII

7th International Symposium
on Intelligent Data Analysis, IDA 2007
Ljubljana, Slovenia, September 6-8, 2007
Proceedings

Springer

Volume Editors

Michael R. Berthold
University of Konstanz, Department of Computer and Information Science
Box M 712, 78457 Konstanz, Germany
E-mail: michael.berthold@uni-konstanz.de

John Shawe-Taylor
University College London
Centre for Computational Statistics and Machine Learning
Department of Computer Science, Gower Street, London WC1E 6BT, UK
E-mail: jst@cs.ucl.ac.uk

Nada Lavrač
Jožef Stefan Institute, Department of Knowledge Technologies
Jamova 39, 1000 Ljubljana, Slovenia
E-mail: nada.lavrac@ijs.si

Library of Congress Control Number: 2007934038

CR Subject Classification (1998): H.3, I.2, G.3, I.5.1, I.4.5, J.2, J.1, J.3

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-74824-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-74824-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12120369 06/3180 5 4 3 2 1 0

Lecture Notes in Computer Science 4723

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Preface

We are proud to present the proceedings of the seventh biennial conference in the Intelligent Data Analysis series. The conference took place in Ljubljana, Slovenia, September 6-8, 2007. IDA continues to expand its scope, quality and size. It started as a small side-symposium as part of a larger conference in 1995 in Baden-Baden (Germany). It quickly attracted more interest in both submissions and attendance as it moved to London (1997) and then Amsterdam (1999). The next three meetings were held in Lisbon (2001), Berlin (2003) and then Madrid in 2005. The improving quality of the submissions has enabled the organizers to assemble programs of ever-increasing consistency and quality. This year we made a rigorous selection of 33 papers out of almost 100 submissions. The resulting oral presentations were then scheduled in a single-track, two-and-a-half-day conference program, summarized in the book that you have before you.

In accordance with the stated IDA goal of "bringing together researchers from diverse disciplines," we believe we have achieved an excellent balance of presentations from the more theoretical – both statistical and machine learning – to the more application-oriented areas that illustrate how these techniques can be used in practice. For example, the proceedings include papers with theoretical contributions dealing with statistical approaches to sequence alignment as well as papers addressing practical problems in the areas of text classification and medical data analysis. It is reassuring to see that IDA continues to bring such diverse areas together, thus helping to cross-fertilize these fields.

Organizing a conference such as IDA is not possible without the assistance and support of many individuals. We are particularly grateful to Tina Anžič and Heather Fyson, who worked tirelessly behind the scenes. Ingrid Fischer worked with Richard van de Stadt to make sure the proceedings were finished flawlessly in time for the meeting. But, crucially, putting together a program of assured quality was only possible through the detailed refereeing of the members of the Program Committee, many of whom also submitted papers and attended the conference.

September 2007

Michael R. Berthold
John Shawe-Taylor
Nada Lavrač

Conference Organization

General Chair	Nada Lavrač Jožef Stefan Institute Ljubljana, Slovenia
Program Chairs	Michael R. Berthold University of Konstanz Konstanz, Germany
	John Shawe-Taylor University College London London, UK
Conference Chair	Mitja Jermol Jožef Stefan Institute Ljubljana, Slovenia
Local Chair	Tina Anžič Jožef Stefan Institute Ljubljana, Slovenia
Registration Chair	Špela Sitar Jožef Stefan Institute Ljubljana, Slovenia
Publication Chair	Ingrid Fischer University of Konstanz Konstanz, Germany
Publicity Chairs	Marjana Plukavec Jožef Stefan Institute Ljubljana, Slovenia
	Alan Tucker Brunel University London, UK
Webmaster	Peter Burger University of Konstanz Konstanz, Germany
Tutorial Chair	Frank Klawonn Fachhochschule Braunschweig/Wolfenbüttel Wolfenbüttel, Germany

Student Grant and Awards Chair	Joost N. Kok Leiden University Leiden, The Netherlands
Finance Chair	Heather Fyson University of Konstanz Konstanz, Germany

Program Committee

Niall Adams, Imperial College London, UK
Martin Atzmüller, University of Würzburg, Germany
Bettina Berendt, Humboldt University Berlin, Germany
Daniel Berrar, University of Ulster, Northern Ireland, UK
Guillaume Beslon, INSA-Lyon, LIRIS, France
Prasad Bhanu, Florida A&M University, USA
Christian Borgelt, European Centre for Soft Computing, Spain
Jean-François Boulicaut, INSA-Lyon, LIRIS UMR CNRS, France
Pavel Brazdil, University of Porto, Portugal
Michael Böhlen, Free University of Bozen-Bolzano, Italy
Klemens Böhm, Universität Karlsruhe, Germany
Luis de Campos, Universidad de Granada, Spain
Bob Clark, Tripos, Inc., USA
Fabio Crestani, University of Lugano, Switzerland
Oscar Cubo-Medina, Laboratorio de SSOO (DATSI), Spain
Luc De Raedt, Katholieke Universiteit Leuven, Belgium
Giuseppe Di Fatta, The University of Reading, UK
Werner Dubitzky, University of Ulster, UK
Sašo Džeroski, Jožef Stefan Institute, Slovenia
André Elisseeff, IBM Research, Switzerland
Fazel Famili, National Research Council of Canada, Canada
Jason Farquhar, Max Planck Institute for Biological Cybernetics, Germany
Ad Feelders, Universiteit Utrecht, The Netherlands
Fridtjof Feldbusch, Universität Karlsruhe, Germany
Ingrid Fischer, University of Konstanz, Germany
Douglas Fisher, Vanderbilt University, USA
Alex Freitas, University of Kent, UK
Elisa Fromont, Katholieke Universiteit Leuven, Belgium
Elisabeth Gassiat, Université Paris-Sud 11, France
Fosca Giannotti, KDDLAB, ISTI-CNR, Pisa, Italy
Santiago González, Universidad Politécnica de Madrid, Spain
Marko Grobelnik, Jožef Stefan Institute, Ljubljana, Slovenia
Gabriela Guimaraes, CENTRIA UNL/FCT, Portugal
Larry Hall, University of South Florida, USA
Pilar Herrero, Universidad Politécnica de Madrid, Spain

Alexander Hinneburg, Martin Luther University Halle-Wittenberg, Germany
Frank Höppner, University of Applied Sciences BS/WF, Germany
Daniel Keim, University of Konstanz, Germany
Frank Klawonn, University of Applied Sciences BW/WF, Germany
Joost N. Kok, Leiden University, The Netherlands
Paul Krause, University of Surrey, UK
Rudolf Kruse, University of Magdeburg, Germany
Matti Kääriäinen, University of Helsinki, Finland
Antonio LaTorre, Universidad Politécnica de Madrid, Spain
Pedro Larrañaga, University of the Basque Country, Spain
Wee Sun Lee, National University of Singapore, Singapore
Hans-J. Lenz, Instute of Statistics and Econometrics, Germany
Xiaohui Liu, Brunel University, United Kingdom
Sofian Maabout, LaBRI Université Bordeaux 1, France
Trevor Martin, University of Bristol, United Kingdom
Dunja Mladenic, Jožef Stefan Institute, Ljubljana, Slovenia
Igor Mozetič, Jožef Stefan Institute, Ljubljana, Slovenia
Susana Nascimento, Universidade Nova de Lisboa, Portugal
Olfa Nasraoui, University of Louisville, USA
Detlef Nauck, BT Intelligent Systems Research Centre, United Kingdom
Andreas Nürnberger, University of Magdeburg, Germany
Tim Oates, CSEE University of Maryland, USA
Francesco d'Ovidio, University of Bari, Italy
Simon Parsons, City University of New York, USA
Kristiaan Pelckmans, Katholieke Universiteit Leuven, Belgium
José-Maria Peña, Technical University of Madrid, Spain
Tuan Pham, James Cook University, Australia
Marco Ramoni, Harvard Medical School, United States
Celine Robardet, LIRIS/INSA-Lyon, France
Víctor Robles Forcada, Universidad Politécnica de Madrid, Spain
Roberta Siciliano, University of Naples, Italy
Arno Siebes, Universiteit Utrecht, Netherlands
Rosaria Silipo, Spoken Translation Inc., Germany
Myra Spiliopoulou, Otto-von-Guericke-University, Magdeburg, Germany
Martin Spott, BTexact Technologies, United Kingdom
Fay Sudweeks, Murdoch University, Australia
Luis Talavera, Universitat Politécnica de Catalunya, Spain
Ljupčo Todorovski, Jožef Stefan Institute, Ljubljana, Slovenia
Hannu Toivonen, University of Helsinki, Finland
Antony Unwin, Augsburg University, Germany
Reinhard Viertl, Vienna University of Technology, Austria
Stefan Wrobel, Fraunhofer IAIS & University of Bonn, Germany
Hong Yan, City University of Hong Kong, Hong Kong, PR China
Daniel Yeung, IEEE SMC Research Institute, Hong Kong
Mohammed Zaki, Rensselaer Polytechnic Institute, USA

Referees

Maurizio Atzori
Robert Banfield
Mario Boley
Janez Brank
Bjoern Bringmann
Patrick Pak-Kei Chan
Giuseppe Delvecchio
Frank Eichinger
Patrik Hoyer
Aneta Ivanovska
Sandra Lima
Laura Madejón López
Wing W. Y. Ng
Siegfried Nijssen
Daniela Oelke
Bostjan Pajntar
Pance Panov
Aleksandar Peckov
Jean-Philippe Pellet
Friedrich Pukelsheim
Joern Schneidewind
Petteri Sevon
Larry Shoemaker
Ivica Slavkov
Matthias Steinbrecher
Dennis Wegener
Hartmut Ziegler

Lecture Notes in Computer Science

Sublibrary 3: Information Systems and Application, incl. Internet/Web and HCI

For information about Vols. 1– 4239
please contact your bookseller or Springer

Vol. 4723: M.R. Berthold, J. Shawe-Taylor, N. Lavrač (Eds.), Advances in Intelligent Data Analysis VII. XIV, 380 pages. 2007.

Vol. 4658: T. Enokido, L. Barolli, M. Takizawa (Eds.), Network-Based Information Systems. XIII, 544 pages. 2007.

Vol. 4656: M.A. Wimmer, J. Scholl, Å. Grönlund (Eds.), Electronic Government. XIV, 450 pages. 2007.

Vol. 4655: G. Psaila, R. Wagner (Eds.), E-Commerce and Web Technologies. VII, 229 pages. 2007.

Vol. 4654: I.Y. Song, J. Eder, T.M. Nguyen (Eds.), Data Warehousing and Knowledge Discovery. XVI, 482 pages. 2007.

Vol. 4653: R. Wagner, N. Revell, G. Pernul (Eds.), Database and Expert Systems Applications. XXII, 907 pages. 2007.

Vol. 4636: G. Antoniou, U. Aßmann, C. Baroglio, S. Decker, N. Henze, P.-L. Patranjan, R. Tolksdorf (Eds.), Reasoning Web. IX, 345 pages. 2007.

Vol. 4611: J. Indulska, J. Ma, L.T. Yang, T. Ungerer, J. Cao (Eds.), Ubiquitous Intelligence and Computing. XXIII, 1257 pages. 2007.

Vol. 4607: L. Baresi, P. Fraternali, G.-J. Houben (Eds.), Web Engineering. XVI, 576 pages. 2007.

Vol. 4606: A. Pras, M. van Sinderen (Eds.), Dependable and Adaptable Networks and Services. XIV, 149 pages. 2007.

Vol. 4605: D. Papadias, D. Zhang, G. Kollios (Eds.), Advances in Spatial and Temporal Databases. X, 479 pages. 2007.

Vol. 4602: S. Barker, G.-J. Ahn (Eds.), Data and Applications Security XXI. X, 291 pages. 2007.

Vol. 4592: Z. Kedad, N. Lammari, E. Métais, F. Meziane, Y. Rezgui (Eds.), Natural Language Processing and Information Systems. XIV, 442 pages. 2007.

Vol. 4587: R. Cooper, J. Kennedy (Eds.), Data Management. XIII, 259 pages. 2007.

Vol. 4577: N. Sebe, Y. Liu, Y.-t. Zhuang, T.S. Huang (Eds.), Multimedia Content Analysis and Mining. XIII, 513 pages. 2007.

Vol. 4568: T. Ishida, S. R. Fussell, P. T. J. M. Vossen (Eds.), Intercultural Collaboration. XIII, 395 pages. 2007.

Vol. 4566: M.J. Dainoff (Ed.), Ergonomics and Health Aspects of Work with Computers. XVIII, 390 pages. 2007.

Vol. 4564: D. Schuler (Ed.), Online Communities and Social Computing. XVII, 520 pages. 2007.

Vol. 4563: R. Shumaker (Ed.), Virtual Reality. XXII, 762 pages. 2007.

Vol. 4561: V.G. Duffy (Ed.), Digital Human Modeling. XXIII, 1068 pages. 2007.

Vol. 4560: N. Aykin (Ed.), Usability and Internationalization, Part II. XVIII, 576 pages. 2007.

Vol. 4559: N. Aykin (Ed.), Usability and Internationalization, Part I. XVIII, 661 pages. 2007.

Vol. 4558: M.J. Smith, G. Salvendy (Eds.), Human Interface and the Management of Information, Part II. XXIII, 1162 pages. 2007.

Vol. 4557: M.J. Smith, G. Salvendy (Eds.), Human Interface and the Management of Information, Part I. XXII, 1030 pages. 2007.

Vol. 4541: T. Okadome, T. Yamazaki, M. Makhtari (Eds.), Pervasive Computing for Quality of Life Enhancement. IX, 248 pages. 2007.

Vol. 4537: K.C.-C. Chang, W. Wang, L. Chen, C.A. Ellis, C.-H. Hsu, A.C. Tsoi, H. Wang (Eds.), Advances in Web and Network Technologies, and Information Management. XXIII, 707 pages. 2007.

Vol. 4531: J. Indulska, K. Raymond (Eds.), Distributed Applications and Interoperable Systems. XI, 337 pages. 2007.

Vol. 4526: M. Malek, M. Reitenspieß, A. van Moorsel (Eds.), Service Availability. X, 155 pages. 2007.

Vol. 4524: M. Marchiori, J.Z. Pan, C.d.S. Marie (Eds.), Web Reasoning and Rule Systems. XI, 382 pages. 2007.

Vol. 4519: E. Franconi, M. Kifer, W. May (Eds.), The Semantic Web: Research and Applications. XVIII, 830 pages. 2007.

Vol. 4518: N. Fuhr, M. Lalmas, A. Trotman (Eds.), Comparative Evaluation of XML Information Retrieval Systems. XII, 554 pages. 2007.

Vol. 4508: M.-Y. Kao, X.-Y. Li (Eds.), Algorithmic Aspects in Information and Management. VIII, 428 pages. 2007.

Vol. 4506: D. Zeng, I. Gotham, K. Komatsu, C. Lynch, M. Thurmond, D. Madigan, B. Lober, J. Kvach, H. Chen (Eds.), Intelligence and Security Informatics: Biosurveillance. XI, 234 pages. 2007.

Vol. 4505: G. Dong, X. Lin, W. Wang, Y. Yang, J.X. Yu (Eds.), Advances in Data and Web Management. XXII, 896 pages. 2007.

Vol. 4504: J. Huang, R. Kowalczyk, Z. Maamar, D. Martin, I. Müller, S. Stoutenburg, K.P. Sycara (Eds.), Service-Oriented Computing: Agents, Semantics, and Engineering. X, 175 pages. 2007.

Vol. 4500: N.A. Streitz, A. Kameas, I. Mavrommati (Eds.), The Disappearing Computer. XVIII, 304 pages. 2007.

Vol. 4495: J. Krogstie, A. Opdahl, G. Sindre (Eds.), Advanced Information Systems Engineering. XVI, 606 pages. 2007.

Vol. 4480: A. LaMarca, M. Langheinrich, K.N. Truong (Eds.), Pervasive Computing. XIII, 369 pages. 2007.

Vol. 4471: P. Cesar, K. Chorianopoulos, J.F. Jensen (Eds.), Interactive TV: A Shared Experience. XIII, 236 pages. 2007.

Vol. 4469: K.-c. Hui, Z. Pan, R.C.-k. Chung, C.C.L. Wang, X. Jin, S. Göbel, E.C.-L. Li (Eds.), Technologies for E-Learning and Digital Entertainment. XVIII, 974 pages. 2007.

Vol. 4443: R. Kotagiri, P.R. Krishna, M. Mohania, E. Nantajeewarawat (Eds.), Advances in Databases: Concepts, Systems and Applications. XXI, 1126 pages. 2007.

Vol. 4439: W. Abramowicz (Ed.), Business Information Systems. XV, 654 pages. 2007.

Vol. 4430: C.C. Yang, D. Zeng, M. Chau, K. Chang, Q. Yang, X. Cheng, J. Wang, F.-Y. Wang, H. Chen (Eds.), Intelligence and Security Informatics. XII, 330 pages. 2007.

Vol. 4425: G. Amati, C. Carpineto, G. Romano (Eds.), Advances in Information Retrieval. XIX, 759 pages. 2007.

Vol. 4412: F. Stajano, H.J. Kim, J.-S. Chae, S.-D. Kim (Eds.), Ubiquitous Convergence Technology. XI, 302 pages. 2007.

Vol. 4402: W. Shen, J. Luo, Z. Lin, J.-P.A. Barthès, Q. Hao (Eds.), Computer Supported Cooperative Work in Design III. XV, 763 pages. 2007.

Vol. 4398: S. Marchand-Maillet, E. Bruno, A. Nürnberger, M. Detyniecki (Eds.), Adaptive Multimedia Retrieval: User, Context, and Feedback. XI, 269 pages. 2007.

Vol. 4397: C. Stephanidis, M. Pieper (Eds.), Universal Access in Ambient Intelligence Environments. XV, 467 pages. 2007.

Vol. 4380: S. Spaccapietra, P. Atzeni, F. Fages, M.-S. Hacid, M. Kifer, J. Mylopoulos, B. Pernici, P. Shvaiko, J. Trujillo, I. Zaihrayeu (Eds.), Journal on Data Semantics VIII. XV, 219 pages. 2007.

Vol. 4365: C. Bussler, M. Castellanos, U. Dayal, S. Navathe (Eds.), Business Intelligence for the Real-Time Enterprises. IX, 157 pages. 2007.

Vol. 4353: T. Schwentick, D. Suciu (Eds.), Database Theory – ICDT 2007. XI, 419 pages. 2006.

Vol. 4352: T.-J. Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua, L.-T. Chia (Eds.), Advances in Multimedia Modeling, Part II. XVIII, 743 pages. 2006.

Vol. 4351: T.-J. Cham, J. Cai, C. Dorai, D. Rajan, T.-S. Chua, L.-T. Chia (Eds.), Advances in Multimedia Modeling, Part I. XIX, 797 pages. 2006.

Vol. 4328: D. Penkler, M. Reitenspiess, F. Tam (Eds.), Service Availability. X, 289 pages. 2006.

Vol. 4321: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), The Adaptive Web. XII, 763 pages. 2007.

Vol. 4317: S.K. Madria, K.T. Claypool, R. Kannan, P. Uppuluri, M.M. Gore (Eds.), Distributed Computing and Internet Technology. XIX, 466 pages. 2006.

Vol. 4312: S. Sugimoto, J. Hunter, A. Rauber, A. Morishima (Eds.), Digital Libraries: Achievements, Challenges and Opportunities. XVIII, 571 pages. 2006.

Vol. 4306: Y. Avrithis, Y. Kompatsiaris, S. Staab, N.E. O'Connor (Eds.), Semantic Multimedia. XII, 241 pages. 2006.

Vol. 4302: J. Domingo-Ferrer, L. Franconi (Eds.), Privacy in Statistical Databases. XI, 383 pages. 2006.

Vol. 4299: S. Renals, S. Bengio, J.G. Fiscus (Eds.), Machine Learning for Multimodal Interaction. XII, 470 pages. 2006.

Vol. 4295: J.D. Carswell, T. Tezuka (Eds.), Web and Wireless Geographical Information Systems. XI, 269 pages. 2006.

Vol. 4286: P.G. Spirakis, M. Mavronicolas, S.C. Kontogiannis (Eds.), Internet and Network Economics. XI, 401 pages. 2006.

Vol. 4282: Z. Pan, A. Cheok, M. Haller, R.W.H. Lau, H. Saito, R. Liang (Eds.), Advances in Artificial Reality and Tele-Existence. XXIII, 1347 pages. 2006.

Vol. 4278: R. Meersman, Z. Tari, P. Herrero (Eds.), On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Part II. XLV, 1004 pages. 2006.

Vol. 4277: R. Meersman, Z. Tari, P. Herrero (Eds.), On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Part I. XLV, 1009 pages. 2006.

Vol. 4276: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE, Part II. XXXII, 752 pages. 2006.

Vol. 4275: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE, Part I. XXXI, 1115 pages. 2006.

Vol. 4273: I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. Aroyo (Eds.), The Semantic Web - ISWC 2006. XXIV, 1001 pages. 2006.

Vol. 4270: H. Zha, Z. Pan, H. Thwaites, A.C. Addison, M. Forte (Eds.), Interactive Technologies and Sociotechnical Systems. XVI, 547 pages. 2006.

Vol. 4261: Y.-t. Zhuang, S.-Q. Yang, Y. Rui, Q. He (Eds.), Advances in Multimedia Information Processing - PCM 2006. XXII, 1040 pages. 2006.

Vol. 4256: L. Feng, G. Wang, C. Zeng, R. Huang (Eds.), Web Information Systems – WISE 2006 Workshops. XIV, 320 pages. 2006.

Vol. 4255: K. Aberer, Z. Peng, E.A. Rundensteiner, Y. Zhang, X. Li (Eds.), Web Information Systems – WISE 2006. XIV, 563 pages. 2006.

Vol. 4254: T. Grust, H. Höpfner, A. Illarramendi, S. Jablonski, M. Mesiti, S. Müller, P.-L. Patranjan, K.-U. Sattler, M. Spiliopoulou, J. Wijsen (Eds.), Current Trends in Database Technology – EDBT 2006. XXXI, 932 pages. 2006.

Vol. 4244: S. Spaccapietra (Ed.), Journal on Data Semantics VII. XI, 267 pages. 2006.

Vol. 4243: T. Yakhno, E.J. Neuhold (Eds.), Advances in Information Systems. XIII, 420 pages. 2006.

Table of Contents

Ranking

Trees

Sequence/ Time Series Analysis

Knowledge Discovery

Visualization

Text Mining

Bioinformatics

Applications

Compact and Understandable Descriptions of Mixtures of Bernoulli Distributions

Jaakko Hollmén and Jarkko Tikka

Helsinki Institute of Information Technology – HIIT
Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FI-02015 TKK, Espoo, Finland
Jaakko.Hollmen@tkk.fi, tikka@mail.cis.hut.fi

Abstract. Finite mixture models can be used in estimating complex, unknown probability distributions and also in clustering data. The parameters of the models form a complex representation and are not suitable for interpretation purposes as such. In this paper, we present a methodology to describe the finite mixture of multivariate Bernoulli distributions with a compact and understandable description. First, we cluster the data with the mixture model and subsequently extract the maximal frequent itemsets from the cluster-specific data sets. The mixture model is used to model the data set globally and the frequent itemsets model the marginal distributions of the partitioned data locally. We present the results in understandable terms that reflect the domain properties of the data. In our application of analyzing DNA copy number amplifications, the descriptions of amplification patterns are represented in nomenclature used in literature to report amplification patterns and generally used by domain experts in biology and medicine.

1 Introduction

In data analysis, the model should absorb the essentials about the data measured from a phenomenon and abstract away the irrelevant details about a particular data set. Parsimonious representations aim at particularly compact and simplified models. These kind of models offer an appealing basis for understanding and describing a phenomenon of interest. Previously, we have investigated parsimonious model representations in ecology [12], where we predicted nutrient concentrations in coniferous trees with a sparse regression methodology. In a time series prediction context, we have proposed a fast input selection method for long-term prediction [14] using a filter strategy. This strategy selects a possibly non-contiguous set of autoregressive variables with linear techniques and builds more complex non-linear prediction models using only the selected variables. In our experience, the parsimonious models are highly desired by domain experts, for instance, in biology, medicine, and ecology. In the models mentioned above, we have included roughly ten percent of the variables (in fact, parameters) compared to the models represented by all the parameters (full models). The sparse models still produce as accurate predictions as the full models. Another line of

M.R. Berthold, J. Shawe-Taylor, and N. Lavrač (Eds.): IDA 2007, LNCS 4723, pp. 1–12, 2007.

research where an attempt is made to concisely describe a data set is reported in [15]. We have presented a tool for automatically generating data survey reports for the modeler to be aware of the properties of the data set. While technically slightly different, the spirit still remains the same as in the current work: the focus is on describing the cluster structure and the contents of the clusters. The aim of the current paper is to present a way to summarize a finite mixture model for 0-1 data concisely and with a simple, domain-compatible representation.

Our research has been motivated by work in analyzing DNA copy number amplifications represented as 0-1 data by profiling [9] and by mixture modeling [13]. The mixture modeling approach offers an elegant way to model DNA amplification patterns in a probabilistic framework. However, the mixture models are summarized by arrays of numerical probability values that are hard to grasp. Therefore, we investigate how to describe the essential properties of the mixture models through the parameters of the models, or alternatively through the clustered data sets. Our proposed solution is based on the maximal frequent itemsets which are extracted from the clustered data sets. The descriptions are represented in the style of the descriptions used in literature to report amplification patterns and generally used by domain experts.

The rest of the paper is organized as follows: Sect. 2 describes the DNA copy number amplification data and our previous research in this context. Sect. 3 describes mixture models in the analysis of 0-1 data and the partitioning scheme for dividing the data in to cluster-specific data sets. The main topic of the paper — how to describe the mixture model for 0-1 data in a compact and understandable fashion — is explained in Sect. 4. Experiments are reported in Sect. 5 and the paper is summarized in Sect. 6. The nomenclature for the chromosome regions, which is used in the experimental part of the paper, is described in Appendix A.

2 DNA Copy Number Amplification Database

We have analyzed the database of DNA copy number amplifications collected with a bibliomics survey from 838 journal articles covering a publication period of ten years from 1992 until 2002 (for details, see [9]). DNA copy number amplifications are localized chromosomal aberrations that increase the number of copies of a chromosomal region from two to at least five. In the database, the DNA copy number amplifications are recorded for $N = 4590$ cancer patients in $d = 393$ chromosomal regions covering the whole human genome, and the observed data are the presence ($x_{ij} = 1$) or the absence ($x_{ij} = 0$) of DNA copy number amplifications for the patient i in the chromosomal region j, where $i = 1, \ldots, 4590$ and $j = 1, \ldots, 393$. For the case including only chromosome 1 presented later in the paper, the dimensions of the data are $N = 446$ and $d = 28$. The nomenclature for the chromosome regions used later in this paper is briefly described in Appendix A. In our previous work, we have analyzed a large 0-1 database of DNA copy number amplification patterns in human neoplasms [9]. We characterized the genome-wide data with cancer-specific amplification profiles with a probabilistic interpretation and