Lecture Notes in Artificial Intelligence 4632

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Reda Alhajj Hong Gao Xue Li
Jianzhong Li Osmar R. Zaïane (Eds.)

Advanced Data Mining and Applications

Third International Conference, ADMA 2007
Harbin, China, August 6-8, 2007
Proceedings

Springer

Series Editors

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

Volume Editors

Reda Alhajj
University of Calgary, Computer Science Department
Calgary, AB, Canada
E-mail: alhajj@cpsc.ucalgary.ca

Hong Gao
Jianzhong Li
Harbin Institute of Technology, School of Computer Science and Technology
Harbin, China
E-mail: {honggao, lijzh}@hit.edu.cn

Xue Li
The University of Queensland
School of Information Technology and Electronic Engineering
Queensland, Australia
E-mail: xueli@itee.uq.edu.au

Osmar R. Zaïane
University of Alberta, Department of Computing Science
Edmonton, AB, Canada
E-mail: zaiane@cs.ualberta.ca

Library of Congress Control Number: 2007931453

CR Subject Classification (1998): I.2, H.2.8, H.3-4, K.4.4, J.3, I.4, J.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-73870-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-73870-1 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 2007
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12098095 06/3180 5 4 3 2 1 0

Preface

The Third International Conference on Advanced Data Mining and Applications (ADMA) organized in Harbin, China continued the tradition already established by the first two ADMA conferences in Wuhan in 2005 and Xi'an in 2006. One major goal of ADMA is to create a respectable identity in the data mining research community. This feat has been partially achieved in a very short time despite the young age of the conference, thanks to the rigorous review process insisted upon, the outstanding list of internationally renowned keynote speakers and the excellent program each year. The impact of a conference is measured by the citations the conference papers receive. Some have used this measure to rank conferences. For example, the independent source cs-conference-ranking.org ranks ADMA (0.65) higher than PAKDD (0.64) and PKDD (0.62) as of June 2007, which are well established conferences in data mining. While the ranking itself is questionable because the exact procedure is not disclosed, it is nevertheless an encouraging indicator of recognition for a very young conference such as ADMA.

This year we had the pleasure and honour to host illustrious presenters. Our distinguished keynote speakers were Prof. Jaideep Srivastava from Minnesota, who had an enviable career in both industry and academia, and Prof. Geoff Web from Monash University, who is currently the editor-in-chief of the famous journal on Data Mining and Knowledge Discovery published by Springer Netherlands. Our invited speaker is Prof. Zhi-Hua Zhou, a rising star in the Chinese and international research community. Despite his young career, he has rightfully earned global recognition.

ADMA aims at bringing together researchers and practitioners to focus on advancements in data mining and peculiarities and challenges of real world applications using data mining. The major theme of the conference encompasses the innovative applications of data mining approaches to real-world problems that involve large data sets, incomplete and noisy data, or demand optimal solutions. While researchers are eager to submit their novel ideas and techniques covering the first part of the theme of the conference, *advanced data mining*, getting researchers and practitioners to share their experience with applications of data mining involving different knowledge of data in challenging application domains remains limited. This second part of the conference theme, *applications*, is important, as applications are very inspirational and educational, and the organizers continue to encourage the use of ADMA as a venue for sharing such valuable experiences with the data mining community.

This year ADMA received about 200 online submissions from 20 different countries, making it, yet again, a truly international conference. A rigorous process of pre-screening and review involved 92 well known international program committee members and 3 program co-chairs in addition to numerous external reviewers. This screening process yielded the remarkable papers organized in these proceedings in 44 regular papers and 15 short papers, bearing a total acceptance rate of 29%.

We witnessed this year a significant drop in the number of submissions and many authors chose to withdraw their papers after they were accepted. The main reason for

this phenomenon is the fact that Thompson Scientific decided to take Springer's LNCS series out of the ISI journal citation index as of January 2007 arguing that LNCS publications are not journals. In fact, according to their recently published results for 2006, ISI evaluated a total of 25,576 articles published in all 373 journals they classify as computer science journals, whereas LNCS alone published roughly the same number of articles during the same period of time. However, LNCS together with its subseries LNAI and LNBI (like other proceedings from other publishers) will from 2007 onward be included in ISI's newly established Proceedings Index and, of course, will continue to be included in many other well known and frequently used bibliographic indices such as DBLP or CompuServe. Finally, despite the fact that LNCS is no longer covered by ISI/SCI, the standing and recognition of ADMA among the community remains unchanged.

June 2007

Reda Alhajj
Hong Gao
Xue Li
Jianzhong Li
Osmar R. Zaïane

Organization

ADMA 2007 was organized by Harbin Institute of Technology, China.

Organizers

General Co-chairs

Jianzhong Li	Harbin Institute of Technology, China
Osmar R. Zaïane	University of Alberta, Canada

Program Co-chairs

Reda Alhajj	University of Calgary, Canada
Hong Gao	Harbin Institute of Technology, China
Xue Li	Queensland University, Australia

Local Arrangement Co-chairs

Shengfei Shi	Harbin Institute of Technology, China
Lv Tianyang	Harbin Engineering University, China

Publicity Co-chairs

Haiwei Pan	Harbin Engineering University, China
Gang Li	Deakin Unversity, Australia

Finance Co-chairs

Jizhou Luo	Harbin Institute of Technology, China
Randy Goebel	Informatics Circle of Research Excellence, Canada

Registration Chair

Hongzhi Wang	Harbin Institute of Technology, China

Web Master

Osmar R. Zaïane	University of Alberta, Canada

Program Committee

Adam Krzyzak	Concordia University, Montreal, Canada
Ah-Hwee Tan	Nanyang Technological University, Singapore
Alfredo Cuzzocrea	University of Calabria, Italy
Alipio M. Jorge	University of Porto, Portugal
Andre Ponce Leao	University S.Paulo, Brazil
Andrew Kusiak	University of Iowa, USA
Ashkan Sami	Shiraz University, Iran
Arthur Tay	National University of Singapore, Singapore
Bart Goethals	University of Antwerp, Belgium
Carlos Soares	LIACC/Fac. of Economics, University of Porto, Portugal
Carlos Soares	University of Porto, Portugal
Cesar Rego	University of Mississippi, USA
Christophe Giraud-Carrier	Brigham Young University, USA
Daniel C. Neagu	University of Bradford, UK
Deepak S Padmanabhan	IBM India Research Lab, India
Desheng Dash Wu	University of Toronto, Canada
Dimitrios Katsaros	Aristotle University, Greece
Eamonn Keogh	University of California - Riverside, USA
Ee-Peng Lim	Nanyang Technological University, Singapore
Elena Baralis	Politecnico di Torino, Italy
Ezeife Christie	University of Windsor, Canada
Faruk Polat	Middle East Technical University, Turkey
Fernando Berzal	University of Granada, Spain
Francesco Bonchi	KDD Laboratory – ISTI CNR Pisa, Italy
Frans Coenen	The University of Liverpool, UK
Gang Li	Deakin University, Australia
Giovanni Semeraro	University of Bari, Italy
Giuseppe Manco	National Research Council of Italy, Italy
Grigorios Tsoumakas	Aristotle University, Greece
Guoren Wang	NorthEast University, China
Haiwei Pan	Harbin Engineering University, China
Hassan Abolhassani	Sharif University of Technology, Iran
Heng Tao Shen	University of Queensland, Australia
Jaideep Srivastava	University of Minnesota, USA
James Bailey	University of Melbourne, Australia
Jan Rauch	University of Economics, Prague, Czech Republic
Jean-Gabriel G Ganascia	LIP6 - University of Paris, France
Jeffrey Xu Yu	The Chinese University of Hong Kong, Hong Kong, China
Jeremy Besson	Insa-Lyon, France
Jimmy Huang	York University, Canada
Jing Liu	Xidian University, China
JingTao Yao	University of Regina, Canada
Joao Gama	University of Porto, Portugal
Junbin Gao	The University of New England, Australia

Kay Chen Tan	National University of Singapore, Singapore
Krzysztof Cios	University of Colorado at Denver, USA
Longin Jan Latecki	Temple University Philadelphia, USA
Lotfi A. Zadeh	Berkeley University of California, USA
Luis Torgo	University of Porto, Portugal
Mehmed Kantardzic	University of Louisville, USA
Mehmet Kaya	Firat University, Turkey
Michael Frank Goodchild	University of California, Santa Barbara, USA
Michael R. Berthold	University of Konstanz, Germany
Mohammad El-Hajj	University of Alberta, Canada
Mohammed Zaki	Rensselear Polytechnic Institute, USA
Naren Ramakrishnan	Virginia Tech, USA
Nasrullah Memon	Aalborg University, Denmark
Olfa Nasraoui	University of Louisville, USA
Ozgur Ulusoy	Bilkent University, Turkey
Paul Vitanyi	CWI and University of Amsterdam, The Netherlands
Peter Geczy	National Institute of Advanced Industrial Science and Technology, Japan
Philip S. Yu	IBM T.J. Watson Research Center, USA
Ping Jiang	Bradford University, UK
Raul Giraldez Rojo	University of Seville, Spain
Ricardo Vilalta	University of Houston, USA
Rui Camacho	University of Porto, Portugal
Sarah Zelikovitz	College of Staten Island, NY, USA
Shaobin Huang	Harbin Engineering University, China
Shashi Shekhar	University of Minnesota, USA
Shengfei Shi	Harbin Institute of Technology, China
Shichao Zhang	University of Technology, Sydney, Australia
Shuigeng Zhou	Fudan University, China
Shuliang Wang	Wuhan University, China
Shusaku Tsumoto	Shimane Medical University, Japan
Simsek Sule	University of Missouri-Rolla, USA
Sunil Choenni	University of Twente, Netherlands
Tan Kok Kiong	National University of Singapore, Singapore
Tansel Ozyer	TOBB University, Turkey
Vladimir Gorodetsky	Intelligent System Lab, The Russian Academy of Science, Russia
Wanquan Liu	Curtin University of Technology, Australia
Wei Wang	Fudan University, China
Xiangjun Dong	Shandong Institute of Light Industry, China
Xuemin Lin,	University of New South Wales, Australia
Yang ZHANG	Northwest A&F University, China
Yingshu Li	Georgia State University, USA
Yonghong Peng	University of Bradford, UK
Zbigniew W. Ras	University of North Carolina, USA
Zhanhuai Li	Northwest Polytechnical University, China

Zhaoyang Dong	The University of Queensland, Australia
Zhipeng Xie	Fudan University, China
Zijiang Yang	York University, Canada

Sponsoring Institutions

Harbin Institute of Technology
University of Alberta
Alberta Innovation and Science
Harbin Engineering University

Table of Contents

Invited Talk

Mining Ambiguous Data with Multi-instance Multi-label Representation 1
Zhi-Hua Zhou

Regular Papers

DELAY: A Lazy Approach for Mining Frequent Patterns over High Speed Data Streams 2
Hui Yang, Hongyan Liu, and Jun He

Exploring Content and Linkage Structures for Searching Relevant Web Pages 15
Darren Davis and Eric Jiang

CLBCRA-Approach for Combination of Content-Based and Link-Based Ranking in Web Search 23
Hao-ming Wang and Ye Guo

Rough Sets in Hybrid Soft Computing Systems 35
Renpu Li, Yongsheng Zhao, Fuzeng Zhang, and Lihua Song

Discovering Novel Multistage Attack Strategies 45
Zhitang Li, Aifang Zhang, Dong Li, and Li Wang

Privacy Preserving DBSCAN Algorithm for Clustering 57
K. Anil Kumar and C. Pandu Rangan

A New Multi-level Algorithm Based on Particle Swarm Optimization for Bisecting Graph 69
Lingyu Sun, Ming Leng, and Songnian Yu

A Supervised Subspace Learning Algorithm: Supervised Neighborhood Preserving Embedding 81
Xianhua Zeng and Siwei Luo

A k-Anonymity Clustering Method for Effective Data Privacy Preservation 89
Chuang-Cheng Chiu and Chieh-Yuan Tsai

LSSVM with Fuzzy Pre-processing Model Based Aero Engine Data Mining Technology 100
Xuhui Wang, Shengguo Huang, Li Cao, Dinghao Shi, and Ping Shu

A Coding Hierarchy Computing Based Clustering Algorithm 110
Jing Peng, Chang-jie Tang, Dong-qing Yang, An-long Chen, and Lei Duan

Mining Both Positive and Negative Association Rules from Frequent and Infrequent Itemsets 122
Xiangjun Dong, Zhendong Niu, Xuelin Shi, Xiaodan Zhang, and Donghua Zhu

Survey of Improving Naive Bayes for Classification 134
Liangxiao Jiang, Dianhong Wang, Zhihua Cai, and Xuesong Yan

Privacy Preserving BIRCH Algorithm for Clustering over Arbitrarily Partitioned Databases 146
P. Krishna Prasad and C. Pandu Rangan

Unsupervised Outlier Detection in Sensor Networks Using Aggregation Tree 158
Kejia Zhang, Shengfei Shi, Hong Gao, and Jianzhong Li

Separator: Sifting Hierarchical Heavy Hitters Accurately from Data Streams 170
Yuan Lin and Hongyan Liu

Spatial Fuzzy Clustering Using Varying Coefficients 183
Huaqiang Yuan, Yaxun Wang, Jie Zhang, Wei Tan, Chao Qu, and Wenbin He

Collaborative Target Classification for Image Recognition in Wireless Sensor Networks 191
Xue Wang, Sheng Wang, and Junjie Ma

Dimensionality Reduction for Mass Spectrometry Data 203
Yihui Liu

The Study of Dynamic Aggregation of Relational Attributes on Relational Data Mining 214
Rayner Alfred

Learning Optimal Kernel from Distance Metric in Twin Kernel Embedding for Dimensionality Reduction and Visualization of Fingerprints 227
Yi Guo, Paul W. Kwan, and Junbin Gao

Efficiently Monitoring Nearest Neighbors to a Moving Object 239
Cheqing Jin and Weibin Guo

A Novel Text Classification Approach Based on Enhanced Association Rule 252
Jiangtao Qiu, Changjie Tang, Tao Zeng, Shaojie Qiao, Jie Zuo, Peng Chen, and Jun Zhu

Applications of the Moving Average of nth-Order Difference Algorithm for Time Series Prediction 264
Yang Lan and Daniel Neagu

Inference of Gene Regulatory Network by Bayesian Network Using Metropolis-Hastings Algorithm 276
Khwunta Kirimasthong, Aompilai Manorat, Jeerayut Chaijaruwanich, Sukon Prasitwattanaseree, and Chinae Thammarongtham

A Consensus Recommender for Web Users 287
Murat Göksedef and Şule Gündüz Öğüdücü

Constructing Classification Rules Based on SVR and Its Derivative Characteristics 300
Dexian Zhang, Zhixiao Yang, Yanfeng Fan, and Ziqiang Wang

Hiding Sensitive Associative Classification Rule by Data Reduction 310
Juggapong Natwichai, Maria E. Orlowska, and Xingzhi Sun

AOG-ags Algorithms and Applications 323
Lizhen Wang, Junli Lu, Joan Lu, and Jim Yip

A Framework for Titled Document Categorization with Modified Multinomial Naivebayes Classifier 335
Hang Guo and Lizhu Zhou

Prediction of Protein Subcellular Locations by Combining K-Local Hyperplane Distance Nearest Neighbor 345
Hong Liu, Haodi Feng, and Daming Zhu

A Similarity Retrieval Method in Brain Image Sequence Database 352
Haiwei Pan, Qilong Han, Xiaoqin Xie, Zhang Wei, and Jianzhong Li

A Criterion for Learning the Data-Dependent Kernel for Classification 365
Jun-Bao Li, Shu-Chuan Chu, and Jeng-Shyang Pan

Topic Extraction with AGAPE 377
Julien Velcin and Jean-Gabriel Ganascia

Clustering Massive Text Data Streams by Semantic Smoothing Model 389
Yubao Liu, Jiarong Cai, Jian Yin, and Ada Wai-Chee Fu

GraSeq: A Novel Approximate Mining Approach of Sequential Patterns over Data Stream 401
Haifeng Li and Hong Chen

A Novel Greedy Bayesian Network Structure Learning Algorithm for Limited Data 412
Feng Liu, Fengzhan Tian, and Qiliang Zhu

Optimum Neural Network Construction Via Linear Programming Minimum Sphere Set Covering 422
Xun-Kai Wei, Ying-Hong Li, and Yu-Fei Li

Retracted: How Investigative Data Mining Can Help Intelligence Agencies to Discover Dependence of Nodes in Terrorist Networks 430
Nasrullah Memon, David L. Hicks, and Henrik Legind Larsen

Prediction of Enzyme Class by Using *Reactive Motifs* Generated from Binding and Catalytic Sites 442
Peera Liewlom, Thanawin Rakthanmanon, and Kitsana Waiyamai

Bayesian Network Structure Ensemble Learning 454
Feng Liu, Fengzhan Tian, and Qiliang Zhu

Fusion of Palmprint and Iris for Personal Authentication 466
Xiangqian Wu, David Zhang, Kuanquan Wang, and Ning Qi

Enhanced Graph Based Genealogical Record Linkage 476
Cary Sweet, Tansel Özyer, and Reda Alhajj

A Fuzzy Comprehensive Clustering Method 488
Shuliang Wang and Xinzhou Wang

Short Papers

CACS: A Novel Classification Algorithm Based on Concept Similarity 500
Jing Peng, Dong-qing Yang, Chang-jie Tang, Jing Zhang, and Jian-jun Hu

Data Mining in Tourism Demand Analysis: A Retrospective Analysis . . . 508
Rob Law, Henry Mok, and Carey Goh

Chinese Patent Mining Based on Sememe Statistics and Key-Phrase Extraction 516
Bo Jin, Hong-Fei Teng, Yan-Jun Shi, and Fu-Zheng Qu

Classification of Business Travelers Using SVMs Combined with Kernel Principal Component Analysis 524
Xin Xu, Rob Law, and Tao Wu

Research on the Traffic Matrix Based on Sampling Model 533
Fengjun Shang

A Causal Analysis for the Expenditure Data of Business Travelers 545
Rob Law and Gang Li

A Visual and Interactive Data Exploration Method for Large Data Sets and Clustering 553
David Da Costa and Gilles Venturini

Explorative Data Mining on Stock Data – Experimental Results and Findings 562
Lay-Ki Soon and Sang Ho Lee

Graph Structural Mining in Terrorist Networks 570
Muhammad Akram Shaikh, Jiaxin Wang, Zehong Yang, and Yixu Song

Characterizing Pseudobase and Predicting RNA Secondary Structure with Simple H-Type Pseudoknots Based on Dynamic Programming 578
Oyun-Erdene Namsrai and Keun Ho Ryu

Locally Discriminant Projection with Kernels for Feature Extraction ... 586
Jun-Bao Li, Shu-Chuan Chu, and Jeng-Shyang Pan

A GA-Based Feature Subset Selection and Parameter Optimization of Support Vector Machine for Content-Based Image Retrieval 594
Kwang-Kyu Seo

E-Stream: Evolution-Based Technique for Stream Clustering 605
Komkrit Udommanetanakit, Thanawin Rakthanmanon, and Kitsana Waiyamai

H-BayesClust: A New Hierarchical Clustering Based on Bayesian Networks 616
Morteza Haghir Chehreghani and Hassan Abolhassani

An Improved AdaBoost Algorithm Based on Adaptive Weight Adjusting 625
Lili Cheng, Jianpei Zhang, Jing Yang, and Jun Ma

Erratum

How Investigative Data Mining Can Help Intelligence Agencies to Discover Dependence of Nodes in Terrorist Networks E1
Nasrullah Memon, David L. Hicks, and Henrik Legind Larsen

Author Index 633