Petra Perner (Ed.)

LNAI 4571

Machine Learning and Data Mining in Pattern Recognition

5th International Conference, MLDM 2007
Leipzig, Germany, July 2007
Proceedings

Springer

Petra Perner (Ed.)

Machine Learning and Data Mining in Pattern Recognition

5th International Conference, MLDM 2007
Leipzig, Germany, July 18-20, 2007
Proceedings

Springer

E2007003076

Series Editors

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

Volume Editor

Petra Perner
Institute of Computer Vision and Applied Computer Sciences (IBaI)
Arno-Nitzsche-Str. 43, 04277 Leipzig, Germany
E-mail: pperner@ibai-institut.de

Library of Congress Control Number: 2007930460

CR Subject Classification (1998): I.2, I.5, I.4, F.4.1, H.3

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-73498-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-73498-7 Springer Berlin Heidelberg New York

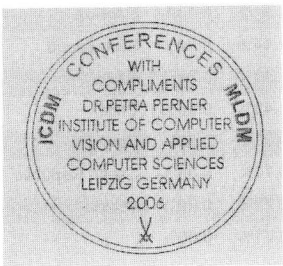
Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12087358 06/3180 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 4571

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Preface



MLDM / ICDM Medaillie
Meissner Porcellan, the "White Gold" of King
August the Strongest of Saxonia

Gottfried Wilhelm von Leibniz, the great mathematician and son of Leipzig, was watching over us during our event in Machine Learning and Data Mining in Pattern Recognition (MLDM 2007). He can be proud of what we have achieved in this area so far. We had a great research program this year.

This was the fifth MLDM in Pattern Recognition event held in Leipzig (www.mldm.de).

Today, there are many international meetings carrying the title machine learning and data mining, whose topics are text mining, knowledge discovery, and applications. This meeting from the very first event has focused on aspects of machine learning and data mining in pattern recognition problems. We planned to reorganize classical and well-established pattern recognition paradigms from the view points of machine learning and data mining. Although it was a challenging program in the late 1990s, the idea has provided new starting points in pattern recognition and has influenced other areas such as cognitive computer vision.

For this edition, the Program Committee received 258 submissions from 37 countries (see Fig. 1).

To handle this high number of papers was a big challenge for the reviewers. Every paper was thoroughly reviewed and all authors received a detailed report on their submitted work.

After the peer-review process, we accepted 66 high-quality papers for oral presentation, which are included in this proceedings book. The topics range from the classical topics within MLDM such as classification, feature selection and extraction, clustering and support-vector machines, frequent and common item set mining and structural data mining.

China	15.71%	5.04%	USA	10.00%	3.88%	England	7.14%	1.94%
France	4.29%	2.33%	Korea South	4.29%	1.16%	Mexico	4.29%	1.16%
Germany	3.57%	1.94%	Iran	3.57%	0.78%	Italy	3.57%	1.94%
Japan	3.57%	1.55%	Spain	3.57%	1.16%	Turkey	2.86%	1.16%
Lithuania	2.86%	0.78%	Cuba	2.86%	0.78%	Greece	2.14%	1.16%
Poland	2.14%	0.78%	Canada	2.14%	0.78%	Portugal	2.14%	0.39%
Australia	2.14%	0.00%	Switzerland	1.43%	0.78%	Sweden	1.43%	0.78%
Brazil	1.43%	0.39%	Taiwan	1.43%	0.39%	India	1.43%	0.00%
Pakistan	1.43%	0.00%	Chile	0.71%	0.39%	Denmark	0.71%	0.39%
Serbia	0.71%	0.39%	Colombia	0.71%	0.39%	Hungary	0.71%	0.39%
Belgium	0.71%	0.39%	Czech Republic	0.71%	0.39%	Russia	0.71%	0.39%
Netherlands	0.71%	0.39%	Ireland	0.71%	0.39%	Belorussia	0.71%	0.00%
Singapore	0.71%	0.00%						

Fig. 1. Distribution of papers among countries

This year we saw new topics in pattern recognition such as transductive inference and association rule mining. The topics of applied research also increased and cover aspects such as mining spam, newsgroups and blogs, intrusion detection and networks, mining marketing data, medical, biological and environmental data mining, text and document mining. We noted with pleasure an increasing number of papers on special aspects of image mining that are the traditional data in pattern recognition.

24 papers have been selected for poster presentation to be published in the MLDM Poster Proceedings Volume. They cover hot topics like text and document mining, image mining, network mining, support vector machines, feature selection, feature maps, prediction and classification, sequence mining, and sampling methods.

We are pleased to announce that we gave out the best paper award for MLDM for the first time this year.

We also established an MLDM/ICDM/MDA Conference Summary Volume for the first time this year that summarizes the vision of three conferences and the paper presentations and also provides a "Who is Who" in machine learning and data mining by giving each author the chance to present himself.

We also thank members of the Institute of Applied Computer Sciences, Leipzig, Germany (www.ibai-institut.de), who handled the conference. We appreciate the help and understanding of the editorial staff at Springer, and in particular Alfred Hofmann, who supported the publication of these proceedings in the LNAI series.

Last, but not least, we wish to thank all the speakers and participants who contributed to the success of the conference. See you in 2009 again.

July 2007

Petra Perner

International Conference on Machine Learning and Data Mining in Pattern Recognition MLDM 2007

Chair

Petra Perner
Institute of Computer Vision and Applied Computer Sciences IBaI Leipzig, Germany

Program Committee

Agnar Aamodt	NTNU, Norway
Max Bramer	University of Portsmouth, UK
Horst Bunke	University of Bern, Switzerland
Krzysztof Cios	University of Colorado, USA
John Debenham	University of Technology, Australia
Christoph F. Eick	University of Houston, USA
Ana Fred	Technical University of Lisbon, Portugal
Giorgio Giacinto	University of Cagliari, Italy
Howard J. Hamilton	University of Regina, Canada
Makato Haraguchi	Hokkaido University Sapporo, Japan
Tin Kam Ho	Bell Laboratories, USA
Atsushi Imiya	Chiba University, Japan
Horace Ip	City University, Hong Kong
Herbert Jahn	Aero Space Center, Germany
Abraham Kandel	University of South Florida, USA
Dimitrios A. Karras	Chalkis Institute of Technology, Greece
Adam Krzyzak	Concordia University, Montreal, Canada
Lukasz Kurgan	University of Alberta, Canada
Longin Jan Latecki	Temple University Philadelphia, USA
Tao Li	Florida International University, USA
Brian Lovell	University of Queensland, Australia
Ryszard Michalski	George Mason University, USA
Mariofanna Milanova	University of Arkansas at Little Rock, USA
Béatrice Pesquet-Popescu	Ecole Nationale des Télécommunications, France
Petia Radeva	Universitat Autonoma de Barcelona, Spain
Fabio Roli	University of Cagliari, Italy
Gabriella Sanniti di Baja	Instituto di Cibernetica, Italy

Linda Shapiro	University of Washington, USA
Sameer Singh	Loughborough University, UK
Arnold Smeulders	University of Amsterdam, The Netherlands
Patrick Wang	Northeastern University, USA
Harry Wechsler	George Mason University, USA
Sholom Weiss	IBM Yorktown Heights, USA
Djemel Ziou	Université de Sherbrooke, Canada

Additional Reviewers

André Lourenço
Ashraf Saad
Chia-Chi Teng
Christian Giusti
Chun-Sheng Chen
Claudia Reisz
David Masip
Gary Yngve
Gero Szepannek
Gian Luca Marcialis
Giorgio Fumera
H.S. Wong
Helge Langseth
Hugo Gamboa
Igino Corona
Ignazio Pillai
Indriyati Atmosukarto
Jordi Vitria
Katarzyna Wilamowska
Luca Didaci
Mineichi Kuido
Natalia Larios Delgado
Oner Ulvi Celepcikay
Orijol Pujol
Rachana Parmar
Rachsuda Jiamthapthaksin
Roberto Perdisci
Roberto Tronci
Sara Rolfe
Vadeerat Rinsurrongkawong
Waclaw Kusnierczyk
Wei Ding
Wojciech Stach
Xingquan Zhu
Yan Liu

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 4597: P. Perner (Ed.), Advances in Data Mining. XI, 353 pages. 2007.

Vol. 4594: R. Bellazzi, A. Abu-Hanna, J. Hunter (Eds.), Artificial Intelligence in Medicine. XVI, 509 pages. 2007.

Vol. 4585: M. Kryszkiewicz, J.F. Peters, H. Rybinski, A. Skowron (Eds.), Rough Sets and Intelligent Systems Paradigms. XIX, 836 pages. 2007.

Vol. 4578: F. Masulli, S. Mitra, G. Pasi (Eds.), Applications of Fuzzy Sets Theory. XVIII, 693 pages. 2007.

Vol. 4573: M. Kauers, M. Kerber, R. Miner, W. Windsteiger (Eds.), Towards Mechanized Mathematical Assistants. XIII, 407 pages. 2007.

Vol. 4571: P. Perner (Ed.), Machine Learning and Data Mining in Pattern Recognition. XIV, 913 pages. 2007.

Vol. 4570: H.G. Okuno, M. Ali (Eds.), New Trends in Applied Artificial Intelligence. XXI, 1194 pages. 2007.

Vol. 4565: D.D. Schmorrow, L.M. Reeves (Eds.), Foundations of Augmented Cognition. XIX, 450 pages. 2007.

Vol. 4562: D. Harris (Ed.), Engineering Psychology and Cognitive Ergonomics. XXIII, 879 pages. 2007.

Vol. 4548: N. Olivetti (Ed.), Automated Reasoning with Analytic Tableaux and Related Methods. X, 245 pages. 2007.

Vol. 4539: N.H. Bshouty, C. Gentile (Eds.), Learning Theory. XII, 634 pages. 2007.

Vol. 4529: P. Melin, O. Castillo, L.T. Aguilar, J. Kacprzyk, W. Pedrycz (Eds.), Foundations of Fuzzy Logic and Soft Computing. XIX, 830 pages. 2007.

Vol. 4511: C. Conati, K. McCoy, G. Paliouras (Eds.), User Modeling 2007. XVI, 487 pages. 2007.

Vol. 4509: Z. Kobti, D. Wu (Eds.), Advances in Artificial Intelligence. XII, 552 pages. 2007.

Vol. 4496: N.T. Nguyen, A. Grzech, R.J. Howlett, L.C. Jain (Eds.), Agent and Multi-Agent Systems: Technologies and Applications. XXI, 1046 pages. 2007.

Vol. 4483: C. Baral, G. Brewka, J. Schlipf (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 327 pages. 2007.

Vol. 4482: A. An, J. Stefanowski, S. Ramanna, C.J. Butz, W. Pedrycz, G. Wang (Eds.), Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. XIV, 585 pages. 2007.

Vol. 4481: J. Yao, P. Lingras, W.-Z. Wu, M. Szczuka, N.J. Cercone, D. Ślęzak (Eds.), Rough Sets and Knowledge Technology. XIV, 576 pages. 2007.

Vol. 4476: V. Gorodetsky, C. Zhang, V.A. Skormin, L. Cao (Eds.), Autonomous Intelligent Systems: Multi-Agents and Data Mining. XIII, 323 pages. 2007.

Vol. 4452: M. Fasli, O. Shehory (Eds.), Agent-Mediated Electronic Commerce. VIII, 249 pages. 2007.

Vol. 4451: T.S. Huang, A. Nijholt, M. Pantic, A. Pentland (Eds.), Artifical Intelligence for Human Computing. XVI, 359 pages. 2007.

Vol. 4438: L. Maicher, A. Sigel, L.M. Garshol (Eds.), Leveraging the Semantics of Topic Maps. X, 257 pages. 2007.

Vol. 4429: R. Lu, J.H. Siekmann, C. Ullrich (Eds.), Cognitive Systems. X, 161 pages. 2007.

Vol. 4426: Z.-H. Zhou, H. Li, Q. Yang (Eds.), Advances in Knowledge Discovery and Data Mining. XXV, 1161 pages. 2007.

Vol. 4411: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), Programming Multi-Agent Systems. XIV, 249 pages. 2007.

Vol. 4410: A. Branco (Ed.), Anaphora: Analysis, Algorithms and Applications. X, 191 pages. 2007.

Vol. 4399: T. Kovacs, X. Llorà, K. Takadama, P.L. Lanzi, W. Stolzmann, S.W. Wilson (Eds.), Learning Classifier Systems. XII, 345 pages. 2007.

Vol. 4390: S.O. Kuznetsov, S. Schmidt (Eds.), Formal Concept Analysis. X, 329 pages. 2007.

Vol. 4389: D. Weyns, H.V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems III. X, 273 pages. 2007.

Vol. 4384: T. Washio, K. Satoh, H. Takeda, A. Inokuchi (Eds.), New Frontiers in Artificial Intelligence. IX, 401 pages. 2007.

Vol. 4371: K. Inoue, K. Satoh, F. Toni (Eds.), Computational Logic in Multi-Agent Systems. X, 315 pages. 2007.

Vol. 4369: M. Umeda, A. Wolf, O. Bartenstein, U. Geske, D. Seipel, O. Takata (Eds.), Declarative Programming for Knowledge Management. X, 229 pages. 2006.

Vol. 4342: H. de Swart, E. Orłowska, G. Schmidt, M. Roubens (Eds.), Theory and Applications of Relational Structures as Knowledge Instruments II. X, 373 pages. 2006.

Vol. 4335: S.A. Brueckner, S. Hassas, M. Jelasity, D. Yamins (Eds.), Engineering Self-Organising Systems. XII, 212 pages. 2007.

Vol. 4334: B. Beckert, R. Hähnle, P.H. Schmitt (Eds.), Verification of Object-Oriented Software. XXIX, 658 pages. 2007.

Vol. 4333: U. Reimer, D. Karagiannis (Eds.), Practical Aspects of Knowledge Management. XII, 338 pages. 2006.

Vol. 4327: M. Baldoni, U. Endriss (Eds.), Declarative Agent Languages and Technologies IV. VIII, 257 pages. 2006.

Vol. 4314: C. Freksa, M. Kohlhase, K. Schill (Eds.), KI 2006: Advances in Artificial Intelligence. XII, 458 pages. 2007.

Vol. 4304: A. Sattar, B.-H. Kang (Eds.), AI 2006: Advances in Artificial Intelligence. XXVII, 1303 pages. 2006.

Vol. 4303: A. Hoffmann, B.-H. Kang, D. Richards, S. Tsumoto (Eds.), Advances in Knowledge Acquisition and Management. XI, 259 pages. 2006.

Vol. 4293: A. Gelbukh, C.A. Reyes-Garcia (Eds.), MICAI 2006: Advances in Artificial Intelligence. XXVIII, 1232 pages. 2006.

Vol. 4289: M. Ackermann, B. Berendt, M. Grobelnik, A. Hotho, D. Mladenič, G. Semeraro, M. Spiliopoulou, G. Stumme, V. Svátek, M. van Someren (Eds.), Semantics, Web and Mining. X, 197 pages. 2006.

Vol. 4285: Y. Matsumoto, R.W. Sproat, K.-F. Wong, M. Zhang (Eds.), Computer Processing of Oriental Languages. XVII, 544 pages. 2006.

Vol. 4274: Q. Huo, B. Ma, E.-S. Chng, H. Li (Eds.), Chinese Spoken Language Processing. XXIV, 805 pages. 2006.

Vol. 4265: L. Todorovski, N. Lavrač, K.P. Jantke (Eds.), Discovery Science. XIV, 384 pages. 2006.

Vol. 4264: J.L. Balcázar, P.M. Long, F. Stephan (Eds.), Algorithmic Learning Theory. XIII, 393 pages. 2006.

Vol. 4259: S. Greco, Y. Hata, S. Hirano, M. Inuiguchi, S. Miyamoto, H.S. Nguyen, R. Słowiński (Eds.), Rough Sets and Current Trends in Computing. XXII, 951 pages. 2006.

Vol. 4253: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. XXXII, 1301 pages. 2006.

Vol. 4252: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. XXXIII, 1335 pages. 2006.

Vol. 4251: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LXVI, 1297 pages. 2006.

Vol. 4248: S. Staab, V. Svátek (Eds.), Managing Knowledge in a World of Networks. XIV, 400 pages. 2006.

Vol. 4246: M. Hermann, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIII, 588 pages. 2006.

Vol. 4223: L. Wang, L. Jiao, G. Shi, X. Li, J. Liu (Eds.), Fuzzy Systems and Knowledge Discovery. XXVIII, 1335 pages. 2006.

Vol. 4213: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Knowledge Discovery in Databases: PKDD 2006. XXII, 660 pages. 2006.

Vol. 4212: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Machine Learning: ECML 2006. XXIII, 851 pages. 2006.

Vol. 4211: P. Vogt, Y. Sugita, E. Tuci, C.L. Nehaniv (Eds.), Symbol Grounding and Beyond. VIII, 237 pages. 2006.

Vol. 4203: F. Esposito, Z.W. Raś, D. Malerba, G. Semeraro (Eds.), Foundations of Intelligent Systems. XVIII, 767 pages. 2006.

Vol. 4201: Y. Sakakibara, S. Kobayashi, K. Sato, T. Nishino, E. Tomita (Eds.), Grammatical Inference: Algorithms and Applications. XII, 359 pages. 2006.

Vol. 4200: I.F.C. Smith (Ed.), Intelligent Computing in Engineering and Architecture. XIII, 692 pages. 2006.

Vol. 4198: O. Nasraoui, O. Zaïane, M. Spiliopoulou, B. Mobasher, B. Masand, P.S. Yu (Eds.), Advances in Web Mining and Web Usage Analysis. IX, 177 pages. 2006.

Vol. 4196: K. Fischer, I.J. Timm, E. André, N. Zhong (Eds.), Multiagent System Technologies. X, 185 pages. 2006.

Vol. 4188: P. Sojka, I. Kopeček, K. Pala (Eds.), Text, Speech and Dialogue. XV, 721 pages. 2006.

Vol. 4183: J. Euzenat, J. Domingue (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 291 pages. 2006.

Vol. 4180: M. Kohlhase, OMDoc – An Open Markup Format for Mathematical Documents [version 1.2]. XIX, 428 pages. 2006.

Vol. 4177: R. Marín, E. Onaindía, A. Bugarín, J. Santos (Eds.), Current Topics in Artificial Intelligence. XV, 482 pages. 2006.

Vol. 4160: M. Fisher, W. van der Hoek, B. Konev, A. Lisitsa (Eds.), Logics in Artificial Intelligence. XII, 516 pages. 2006.

Vol. 4155: O. Stock, M. Schaerf (Eds.), Reasoning, Action and Interaction in AI Theories and Systems. XVIII, 343 pages. 2006.

Vol. 4149: M. Klusch, M. Rovatsos, T.R. Payne (Eds.), Cooperative Information Agents X. XII, 477 pages. 2006.

Vol. 4140: J.S. Sichman, H. Coelho, S.O. Rezende (Eds.), Advances in Artificial Intelligence - IBERAMIA-SBIA 2006. XXIII, 635 pages. 2006.

Vol. 4139: T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (Eds.), Advances in Natural Language Processing. XVI, 771 pages. 2006.

Vol. 4133: J. Gratch, M. Young, R. Aylett, D. Ballin, P. Olivier (Eds.), Intelligent Virtual Agents. XIV, 472 pages. 2006.

Vol. 4130: U. Furbach, N. Shankar (Eds.), Automated Reasoning. XV, 680 pages. 2006.

Vol. 4120: J. Calmet, T. Ida, D. Wang (Eds.), Artificial Intelligence and Symbolic Computation. XIII, 269 pages. 2006.

Vol. 4118: Z. Despotovic, S. Joseph, C. Sartori (Eds.), Agents and Peer-to-Peer Computing. XIV, 173 pages. 2006.

Vol. 4114: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence, Part II. XXVII, 1337 pages. 2006.

Vol. 4108: J.M. Borwein, W.M. Farmer (Eds.), Mathematical Knowledge Management. VIII, 295 pages. 2006.

Vol. 4106: T.R. Roth-Berghofer, M.H. Göker, H.A. Güvenir (Eds.), Advances in Case-Based Reasoning. XIV, 566 pages. 2006.

¥990.00元

Table of Contents

Invited Talk

Data Clustering: User's Dilemma (Abstract) 1
Anil K. Jain

Classification

On Concentration of Discrete Distributions with Applications to Supervised Learning of Classifiers 2
Magnus Ekdahl and Timo Koski

Comparison of a Novel Combined ECOC Strategy with Different Multiclass Algorithms Together with Parameter Optimization Methods 17
Marco Hülsmann and Christoph M. Friedrich

Multi-source Data Modelling: Integrating Related Data to Improve Model Performance 32
Paul R. Trundle, Daniel C. Neagu, and Qasim Chaudhry

An Empirical Comparison of Ideal and Empirical ROC-Based Reject Rules 47
Claudio Marrocco, Mario Molinara, and Francesco Tortorella

Outlier Detection with Kernel Density Functions 61
Longin Jan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac

Generic Probability Density Function Reconstruction for Randomization in Privacy-Preserving Data Mining 76
Vincent Yan Fu Tan and See-Kiong Ng

An Incremental Fuzzy Decision Tree Classification Method for Mining Data Streams 91
Tao Wang, Zhoujun Li, Yuejin Yan, and Huowang Chen

On the Combination of Locally Optimal Pairwise Classifiers 104
Gero Szepannek, Bernd Bischl, and Claus Weihs

Feature Selection, Extraction and Dimensionality Reduction

An Agent-Based Approach to the Multiple-Objective Selection of Reference Vectors 117
Ireneusz Czarnowski and Piotr Jędrzejowicz

On Applying Dimension Reduction for Multi-labeled Problems 131
Moonhwi Lee and Cheong Hee Park

Nonlinear Feature Selection by Relevance Feature Vector Machine 144
Haibin Cheng, Haifeng Chen, Guofei Jiang, and Kenji Yoshihira

Affine Feature Extraction: A Generalization of the Fukunaga-Koontz Transformation 160
Wenbo Cao and Robert Haralick

Clustering

A Bounded Index for Cluster Validity 174
Sandro Saitta, Benny Raphael, and Ian F.C. Smith

Varying Density Spatial Clustering Based on a Hierarchical Tree 188
Xuegang Hu, Dongbo Wang, and Xindong Wu

Kernel MDL to Determine the Number of Clusters 203
Ivan O. Kyrgyzov, Olexiy O. Kyrgyzov, Henri Maître, and Marine Campedel

Critical Scale for Unsupervised Cluster Discovery 218
Tomoya Sakai, Atsushi Imiya, Takuto Komazaki, and Shiomu Hama

Minimum Information Loss Cluster Analysis for Categorical Data 233
Jiří Grim and Jan Hora

A Clustering Algorithm Based on Generalized Stars 248
Airel Pérez Suárez and José E. Medina Pagola

Support Vector Machine

Evolving Committees of Support Vector Machines..... 263
D. Valincius, A. Verikas, M. Bacauskiene, and A. Gelzinis

Choosing the Kernel Parameters for the Directed Acyclic Graph Support Vector Machines 276
Kuo-Ping Wu and Sheng-De Wang

Data Selection Using SASH Trees for Support Vector Machines 286
Chaofan Sun and Ricardo Vilalta

Dynamic Distance-Based Active Learning with SVM 296
Jun Jiang and Horace H.S. Ip

Transductive Inference

Off-Line Learning with Transductive Confidence Machines: An Empirical Evaluation 310
Stijn Vanderlooy, Laurens van der Maaten, and Ida Sprinkhuizen-Kuyper

Transductive Learning from Relational Data 324
Michelangelo Ceci, Annalisa Appice, Nicola Barile, and Donato Malerba

Association Rule Mining

A Novel Rule Ordering Approach in Classification Association Rule Mining 339
Yanbo J. Wang, Qin Xin, and Frans Coenen

Distributed and Shared Memory Algorithm for Parallel Mining of Association Rules 349
J. Hernández Palancar, O. Fraxedas Tormo, J. Festón Cárdenas, and R. Hernández León

Mining Spam, Newsgroups, Blogs

Analyzing the Performance of Spam Filtering Methods When Dimensionality of Input Vector Changes 364
J.R. Méndez, B. Corzo, D. Glez-Peña, F. Fdez-Riverola, and F. Díaz

Blog Mining for the Fortune 500 379
James Geller, Sapankumar Parikh, and Sriram Krishnan

A Link-Based Rank of Postings in Newsgroup 392
Hongbo Liu, Jiahai Yang, Jiaxin Wang, and Yu Zhang

Intrusion Detection and Networks

A Comparative Study of Unsupervised Machine Learning and Data Mining Techniques for Intrusion Detection 404
Reza Sadoddin and Ali A. Ghorbani

Long Tail Attributes of Knowledge Worker Intranet Interactions 419
Peter Géczy, Noriaki Izumi, Shotaro Akaho, and Kôiti Hasida

A Case-Based Approach to Anomaly Intrusion Detection 434
Alessandro Micarelli and Giuseppe Sansonetti

Sensing Attacks in Computers Networks with Hidden Markov Models ... 449
Davide Ariu, Giorgio Giacinto, and Roberto Perdisci

Frequent and Common Item Set Mining

FIDS: Monitoring Frequent Items over Distributed Data Streams 464
Robert Fuller and Mehmed Kantardzic

Mining Maximal Frequent Itemsets in Data Streams Based on FP-Tree 479
Fujiang Ao, Yuejin Yan, Jian Huang, and Kedi Huang

CCIC: Consistent Common Itemsets Classifier 490
Yohji Shidara, Atsuyoshi Nakamura, and Mineichi Kudo

Mining Marketing Data

Development of an Agreement Metric Based Upon the RAND Index for the Evaluation of Dimensionality Reduction Techniques, with Applications to Mapping Customer Data 499
Stephen France and Douglas Carroll

A Sequential Hybrid Forecasting System for Demand Prediction 518
Luis Aburto and Richard Weber

A Unified View of Objective Interestingness Measures 533
Céline Hébert and Bruno Crémilleux

Comparing State-of-the-Art Collaborative Filtering Systems 548
Laurent Candillier, Frank Meyer, and Marc Boullé

Structural Data Mining

Reducing the Dimensionality of Vector Space Embeddings of Graphs ... 563
Kaspar Riesen, Vivian Kilchherr, and Horst Bunke

PE-PUC: A Graph Based PU-Learning Approach for Text Classification 574
Shuang Yu and Chunping Li

Efficient Subsequence Matching Using the Longest Common Subsequence with a Dual Match Index 585
Tae Sik Han, Seung-Kyu Ko, and Jaewoo Kang

A Direct Measure for the Efficacy of Bayesian Network Structures Learned from Data 601
Gary F. Holness

Image Mining

A New Combined Fractal Scale Descriptor for Gait Sequence 616
Li Cui and Hua Li

Palmprint Recognition by Applying Wavelet Subband Representation and Kernel PCA 628
Murat Ekinci and Murat Aykut

A Filter-Refinement Scheme for 3D Model Retrieval Based on Sorted Extended Gaussian Image Histogram 643
Zhiwen Yu, Shaohong Zhang, Hau-San Wong, and Jiqi Zhang

Fast-Maneuvering Target Seeking Based on Double-Action Q-Learning 653
Daniel C.K. Ngai and Nelson H.C. Yung

Mining Frequent Trajectories of Moving Objects for Location Prediction 667
Mikołaj Morzy

Categorizing Evolved CoreWar Warriors Using EM and Attribute Evaluation 681
Doni Pracner, Nenad Tomašev, Miloš Radovanović, and Mirjana Ivanović

Restricted Sequential Floating Search Applied to Object Selection..... 694
J. Arturo Olvera-López, J. Francisco Martínez-Trinidad, and J. Ariel Carrasco-Ochoa

Color Reduction Using the Combination of the Kohonen Self-Organized Feature Map and the Gustafson-Kessel Fuzzy Algorithm 703
Konstantinos Zagoris, Nikos Papamarkos, and Ioannis Koustoudis

A Hybrid Algorithm Based on Evolution Strategies and Instance-Based Learning, Used in Two-Dimensional Fitting of Brightness Profiles in Galaxy Images..... 716
Juan Carlos Gomez and Olac Fuentes

Gait Recognition by Applying Multiple Projections and Kernel PCA ... 727
Murat Ekinci, Murat Aykut, and Eyup Gedikli

Medical, Biological, and Environmental Data Mining

A Machine Learning Approach to Test Data Generation: A Case Study in Evaluation of Gene Finders 742
Henning Christiansen and Christina Mackeprang Dahmcke

Discovering Plausible Explanations of Carcinogenecity in Chemical Compounds 756
Eva Armengol

One Lead ECG Based Personal Identification with Feature Subspace Ensembles 770
Hugo Silva, Hugo Gamboa, and Ana Fred

Classification of Breast Masses in Mammogram Images Using Ripley's K Function and Support Vector Machine 784
Leonardo de Oliveira Martins, Erick Corrêa da Silva, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, and Marcelo Gattass

Selection of Experts for the Design of Multiple Biometric Systems 795
Roberto Tronci, Giorgio Giacinto, and Fabio Roli

Multi-agent System Approach to React to Sudden Environmental Changes 810
Sarunas Raudys and Antanas Mitasiunas

Equivalence Learning in Protein Classification 824
Attila Kertész-Farkas, András Kocsor, and Sándor Pongor

Text and Document Mining

Statistical Identification of Key Phrases for Text Classification 838
Frans Coenen, Paul Leng, Robert Sanderson, and Yanbo J. Wang

Probabilistic Model for Structured Document Mapping 854
Guillaume Wisniewski, Francis Maes, Ludovic Denoyer, and Patrick Gallinari

Application of Fractal Theory for On-Line and Off-Line Farsi Digit Recognition 868
Saeed Mozaffari, Karim Faez, and Volker Märgner

Hybrid Learning of Ontology Classes 883
Jens Lehmann

Discovering Relations Among Entities from XML Documents 899
Yangyang Wu, Qing Lei, Wei Luo, and Harou Yokota

Author Index 911

Data Clustering: User's Dilemma

Anil K. Jain

Department of Computer Science and Engineering
Michigan State University (USA)
http://www.cse.msu.edu/~jain/

Abstract. Data clustering is a long standing research problem in pattern recognition, computer vision, machine learning, and data mining with applications in a number of diverse disciplines. The goal is to partition a set of n d-dimensional points into k clusters, where k may or may not be known. Most clustering techniques require the definition of a similarity measure between patterns, which is not easy to specify in the absence of any prior knowledge about cluster shapes. While a large number of clustering algorithms exist, there is no optimal algorithm. Each clustering algorithm imposes a specific structure on the data and has its own approach for estimating the number of clusters. No single algorithm can adequately handle various cluster shapes and structures that are encountered in practice. Instead of spending our effort in devising yet another clustering algorithm, there is a need to build upon the existing published techniques. In this talk we will address the following problems: (i) clustering via evidence accumulation, (ii) simultaneous clustering and dimensionality reduction, (iii) clustering under pair-wise constraints, and (iv) clustering with relevance feedback. Experimental results show that these approaches are promising in identifying arbitrary shaped clusters in multidimensional data.

P. Perner (Ed.): MLDM 2007, LNAI 4571, p. 1, 2007.