Preface

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 viewpoints 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.
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

<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage</th>
<th>Paper Count</th>
<th>Country</th>
<th>Percentage</th>
<th>Paper Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>15.71%</td>
<td>5.04%</td>
<td>USA</td>
<td>10.00%</td>
<td>3.88%</td>
</tr>
<tr>
<td>France</td>
<td>4.29%</td>
<td>2.33%</td>
<td>Korea South</td>
<td>4.29%</td>
<td>1.16%</td>
</tr>
<tr>
<td>Germany</td>
<td>3.57%</td>
<td>1.94%</td>
<td>Iran</td>
<td>3.57%</td>
<td>0.78%</td>
</tr>
<tr>
<td>Japan</td>
<td>3.57%</td>
<td>1.55%</td>
<td>Spain</td>
<td>3.57%</td>
<td>1.16%</td>
</tr>
<tr>
<td>Lithuania</td>
<td>2.86%</td>
<td>0.78%</td>
<td>Cuba</td>
<td>2.86%</td>
<td>0.78%</td>
</tr>
<tr>
<td>Poland</td>
<td>2.14%</td>
<td>0.78%</td>
<td>Canada</td>
<td>2.14%</td>
<td>0.78%</td>
</tr>
<tr>
<td>Australia</td>
<td>2.14%</td>
<td>0.00%</td>
<td>Switzerland</td>
<td>1.43%</td>
<td>0.78%</td>
</tr>
<tr>
<td>Brazil</td>
<td>1.43%</td>
<td>0.39%</td>
<td>Taiwan</td>
<td>1.43%</td>
<td>0.39%</td>
</tr>
<tr>
<td>Pakistan</td>
<td>1.43%</td>
<td>0.00%</td>
<td>Chile</td>
<td>0.71%</td>
<td>0.39%</td>
</tr>
<tr>
<td>Serbia</td>
<td>0.71%</td>
<td>0.39%</td>
<td>Colombia</td>
<td>0.71%</td>
<td>0.39%</td>
</tr>
<tr>
<td>Belgium</td>
<td>0.71%</td>
<td>0.39%</td>
<td>Czech Republic</td>
<td>0.71%</td>
<td>0.39%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>0.71%</td>
<td>0.39%</td>
<td>Ireland</td>
<td>0.71%</td>
<td>0.39%</td>
</tr>
<tr>
<td>Singapore</td>
<td>0.71%</td>
<td>0.00%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1.** Distribution of papers among countries
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
VIII Organization

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

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 Griman 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 Kōiti 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
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palmprint Recognition by Applying Wavelet Subband Representation</td>
<td>628</td>
</tr>
<tr>
<td>and Kernel PCA</td>
<td></td>
</tr>
<tr>
<td><em>Murat Ekinci and Murat Aykut</em></td>
<td></td>
</tr>
<tr>
<td>A Filter-Refinement Scheme for 3D Model Retrieval Based on Sorted</td>
<td>643</td>
</tr>
<tr>
<td>Extended Gaussian Image Histogram</td>
<td></td>
</tr>
<tr>
<td><em>Zhiwen Yu, Shaohong Zhang, Hau-San Wong, and Jiqi Zhang</em></td>
<td></td>
</tr>
<tr>
<td>Fast-Maneuvering Target Seeking Based on Double-Action Q-Learning</td>
<td>653</td>
</tr>
<tr>
<td><em>Daniel C.K. Ngai and Nelson H.C. Yung</em></td>
<td></td>
</tr>
<tr>
<td>Mining Frequent Trajectories of Moving Objects for Location</td>
<td>667</td>
</tr>
<tr>
<td>Prediction</td>
<td></td>
</tr>
<tr>
<td><em>Mikołaj Morzy</em></td>
<td></td>
</tr>
<tr>
<td>Categorizing Evolved CoreWar Warriors Using EM and Attribute Evaluation</td>
<td>681</td>
</tr>
<tr>
<td><em>Doni Pracner, Nenad Tomašev, Miloš Radovanović, and Mirjana Ivanović</em></td>
<td></td>
</tr>
<tr>
<td>Restricted Sequential Floating Search Applied to Object Selection</td>
<td>694</td>
</tr>
<tr>
<td><em>J. Arturo Olvera-López, J. Francisco Martínez-Trinidad, and J. Ariel Carrusco-Ochoa</em></td>
<td></td>
</tr>
<tr>
<td>Color Reduction Using the Combination of the Kohonen Self-Organized Feature Map and the Gustafson-Kessel Fuzzy Algorithm</td>
<td>703</td>
</tr>
<tr>
<td><em>Konstantinos Zagoris, Nikos Papamarkos, and Ioannis Koustoudis</em></td>
<td></td>
</tr>
<tr>
<td>A Hybrid Algorithm Based on Evolution Strategies and Instance-Based Learning, Used in Two-Dimensional Fitting of Brightness Profiles in Galaxy Images</td>
<td>716</td>
</tr>
<tr>
<td><em>Juan Carlos Gomez and Olac Fuentes</em></td>
<td></td>
</tr>
<tr>
<td>Gait Recognition by Applying Multiple Projections and Kernel PCA</td>
<td>727</td>
</tr>
<tr>
<td><em>Murat Ekinci, Murat Aykut, and Eyup Gedikli</em></td>
<td></td>
</tr>
<tr>
<td>Medical, Biological, and Environmental Data Mining</td>
<td></td>
</tr>
<tr>
<td>A Machine Learning Approach to Test Data Generation: A Case Study in Evaluation of Gene Finders</td>
<td>742</td>
</tr>
<tr>
<td><em>Henning Christiansen and Christina Mackeprang Dahmcke</em></td>
<td></td>
</tr>
<tr>
<td>Discovering Plausible Explanations of Carcinogenicity in Chemical Compounds</td>
<td>756</td>
</tr>
<tr>
<td><em>Eva Armengol</em></td>
<td></td>
</tr>
<tr>
<td>One Lead ECG Based Personal Identification with Feature Subspace Ensembles</td>
<td>770</td>
</tr>
<tr>
<td><em>Hugo Silva, Hugo Gamboa, and Ana Fred</em></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Classification of Breast Masses in Mammogram Images Using Ripley’s K Function and Support Vector Machine</td>
<td>784</td>
</tr>
<tr>
<td>Leonardo de Oliveira Martins, Erick Corrêa da Silva, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, and Marcelo Gattass</td>
<td></td>
</tr>
<tr>
<td>Selection of Experts for the Design of Multiple Biometric Systems</td>
<td>795</td>
</tr>
<tr>
<td>Roberto Tronci, Giorgio Giacinto, and Fabio Roli</td>
<td></td>
</tr>
<tr>
<td>Multi-agent System Approach to React to Sudden Environmental Changes</td>
<td>810</td>
</tr>
<tr>
<td>Sarunas Raudys and Antanas Mitasiunas</td>
<td></td>
</tr>
<tr>
<td>Equivalence Learning in Protein Classification</td>
<td>824</td>
</tr>
<tr>
<td>Attila Kertész-Farkas, András Kocsor, and Sándor Pongor</td>
<td></td>
</tr>
<tr>
<td><strong>Text and Document Mining</strong></td>
<td></td>
</tr>
<tr>
<td>Statistical Identification of Key Phrases for Text Classification</td>
<td>838</td>
</tr>
<tr>
<td>Frans Coenen, Paul Leng, Robert Sanderson, and Yanbo J. Wang</td>
<td></td>
</tr>
<tr>
<td>Probabilistic Model for Structured Document Mapping</td>
<td>854</td>
</tr>
<tr>
<td>Guillaume Wisniewski, Francis Maes, Ludovic Denoyer, and Patrick Gallinari</td>
<td></td>
</tr>
<tr>
<td>Application of Fractal Theory for On-Line and Off-Line Farsi Digit Recognition</td>
<td>868</td>
</tr>
<tr>
<td>Saeed Mozaffari, Karim Faez, and Volker Märgner</td>
<td></td>
</tr>
<tr>
<td>Hybrid Learning of Ontology Classes</td>
<td>883</td>
</tr>
<tr>
<td>Jens Lehmann</td>
<td></td>
</tr>
<tr>
<td>Discovering Relations Among Entities from XML Documents</td>
<td>899</td>
</tr>
<tr>
<td>Yangyang Wu, Qing Lei, Wei Luo, and Harou Yokota</td>
<td></td>
</tr>
<tr>
<td><strong>Author Index</strong></td>
<td>911</td>
</tr>
</tbody>
</table>