

Lecture Notes in Artificial Intelligence 7376

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

Petra Perner (Ed.)

Machine Learning and Data Mining in Pattern Recognition

8th International Conference, MLDM 2012
Berlin, Germany, July 13-20, 2012
Proceedings

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editor

Petra Perner
Institute of Computer Vision
and Applied Computer Sciences, IBaI
Kohlenstr. 2, 04107 Leipzig, Germany
E-mail: pperner@ibai-institut.de

ISSN 0302-9743
ISBN 978-3-642-31536-7
DOI 10.1007/978-3-642-31537-4

e-ISSN 1611-3349
e-ISBN 978-3-642-31537-4

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012940654

CR Subject Classification (1998): I.2, F.4, I.4, I.5, H.3, H.2.8

LNCS Sublibrary: SL 7 – Artificial Intelligence

© Springer-Verlag Berlin Heidelberg 2012

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.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The eighth event of the International Conference on Machine Learning and Data Mining (MLDM) was held in Berlin (www.mldm.de) under the umbrella of the World Congress on “The Frontiers in Intelligent Data and Signal Analysis, DSA 2012”.

For this edition the Program Committee received 212 submissions. After the peer-review process, we accepted 71 high-quality papers for oral presentation, from which 51 are included in this proceedings book. The topics range from theoretical topics for classification, clustering, association rule and pattern mining to specific data mining methods for the different multimedia data types such as image mining, text mining, video mining and Web mining. Extended versions of selected papers will appear in the *International Journal Transactions on Machine Learning and Data Mining* (www.ibai-publishing.org/journal/mldm).

Eight papers were selected for poster presentations and are published in the *MLDM Poster Proceedings* by *ibai-publishing* (www.ibai-publishing.org).

A tutorial on Data Mining, a tutorial on Case-Based Reasoning, a tutorial on Intelligent Image Interpretation and Computer Vision in Medicine, Biotechnology, Chemistry and Food Industry and a tutorial on Standardization in Immunofluorescence were held before the conference.

We were pleased to give out the best paper award for the fourth time this year (www.mldm.de). The final decision was made by the Best Paper Award Committee based on the presentation by the authors and the discussion with the auditorium. The ceremony took place at the end of the conference. This prize is sponsored by ibai solutions (www.ibai-solutions.de), one of the leading companies in data mining for marketing, Web mining and e-commerce.

The conference was rounded up by an outlook of new challenging topics in machine learning and data mining before the Best Paper Award ceremony.

We would like to thank the members of the Institute of Applied Computer Sciences, Leipzig, Germany (www.ibai-institut.de), who handled the conference as secretariat. We appreciate the help and understanding of the editorial staff at Springer Verlag, 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 2013 in New York to the next World Congress on “The Frontiers in Intelligent Data and Signal Analysis, DSA2013” (www.worldcongressdsa.com) will be held in New York, in 2013, combining under its roof the following three events: International Conferences Machine Learning and Data Mining (MLDM), the Industrial Conference on Data Mining (ICDM), and the International Conference on Mass Data Analysis of Signals and Images in Medicine, Biotechnology, Chemistry and Food Industry (MDA).

International Conference on Machine Learning and Data Mining, MLDM 2012

Program Chair

Petra Perner

IBaI Leipzig, Germany

Program Committee

| | |
|-------------------------|--|
| Agnar Aamodt | NTNU, Norway |
| Jacky Baltes | University of Manitoba, Canada |
| Christoph F. Eick | University of Houston, USA |
| Ana Fred | Technical University of Lisbon, Portugal |
| Giorgio Giacinto | University of Cagliari, Italy |
| Makato Haraguchi | Hokkaido University Sapporo, Japan |
| Robert J. Hilderman | University of Regina, Canada |
| Eyke Hüllermeier | University of Marsburg, Germany |
| Atsushi Imiya | Chiba University, Japan |
| Abraham Kandel | University of South Florida, USA |
| Dimitrios A. Karras | Chalkis Institute of Technology, Greece |
| Adam Krzyzak | Concordia University, Montreal, Canada |
| Brian Lovell | University of Queensland, Australia |
| Mariofanna Milanova | University of Arkansas at Little Rock, USA |
| Thang V. Pham | University of Amsterdam, The Netherlands |
| Maria da Graca Pimentel | Universidade de Sao Paulo, Brazil |
| Petia Radeva | Universitat Autònoma de Barcelona, Spain |
| Michael Richter | University of Calgary, Canada |
| Fabio Roli | University of Cagliari, Italy |
| Linda Shapiro | University of Washington, USA |
| Sameer Singh | Loughborough University, UK |
| Harald Steck | Bell Laboratoris, USA |
| Francesco Tortorella | Università degli Studi di Cassino, Italy |
| Patrick Wang | Northeastern University, USA |

Additional Reviewers

| | |
|---------------------------|-------------------------------|
| Pål Sætrom (Paal Saetrom) | NTNU, Norway |
| Gleb Sizov | NTNU, Norway |
| Theoharis Theoharis | NTNU, Norway |
| Luigi Atzori | University of Cagliari, Italy |
| Davide Ariu | University of Cagliari, Italy |
| Giuliano Armano | University of Cagliari, Italy |

| | |
|-----------------|--|
| Battista Biggio | University of Cagliari, Italy |
| Igino Corona | University of Cagliari, Italy |
| Luca Didaci | University of Cagliari, Italy |
| Giorgio Fumera | University of Cagliari, Italy |
| Danilo Pani | University of Cagliari, Italy |
| Ignazio Pillai | University of Cagliari, Italy |
| Luca Piras | University of Cagliari, Italy |
| Riccardo Satta | University of Cagliari, Italy |
| Roberto Tronci | University of Cagliari, Italy |
| Eloisa Vargiu | Barcelona Digital Technolgic Centre, Spain |

Table of Contents

Theory

| | |
|---|-----|
| Bayesian Approach to the Concept Drift in the Pattern Recognition Problems | 1 |
| <i>Pavel Turkov, Olga Krasotkina, and Vadim Mottl</i> | |
| Transductive Relational Classification in the Co-training Paradigm | 11 |
| <i>Michelangelo Ceci, Annalisa Appice, Herna L. Viktor, Donato Malerba, Eric Paquet, and Hongyu Guo</i> | |
| Generalized Nonlinear Classification Model Based on Cross-Oriented Choquet Integral | 26 |
| <i>Rong Yang and Zhenyuan Wang</i> | |
| A General Lp-norm Support Vector Machine via Mixed 0-1 Programming | 40 |
| <i>Hai Thanh Nguyen and Katrin Franke</i> | |
| Reduction of Distance Computations in Selection of Pivot Elements for Balanced GHT Structure | 50 |
| <i>László Kovács</i> | |
| Hot Deck Methods for Imputing Missing Data: The Effects of Limiting Donor Usage | 63 |
| <i>Dieter William Joenssen and Udo Bankhofer</i> | |
| BINER: BINary Search Based Efficient Regression | 76 |
| <i>Saket Bharambe, Harshit Dubey, and Vikram Pudi</i> | |
| A New Approach for Association Rule Mining and Bi-clustering Using Formal Concept Analysis | 86 |
| <i>Kartick Chandra Mondal, Nicolas Pasquier, Anirban Mukhopadhyay, Ujjwal Maulik, and Sanghamitra Bandhopadhyay</i> | |
| Top- <i>N</i> Minimization Approach for Indicative Correlation Change Mining | 102 |
| <i>Aixiang Li, Makoto Haraguchi, and Yoshiaki Okubo</i> | |

Theory: Evaluation of Models and Performance Evaluation Methods

| | |
|---|-----|
| Selecting Classification Algorithms with Active Testing | 117 |
| <i>Rui Leite, Pavel Brazdil, and Joaquin Vanschoren</i> | |

| | |
|--|-----|
| Comparing Logistic Regression, Neural Networks, C5.0 and M5' Classification Techniques | 132 |
| <i>Amit Thombre</i> | |
| Unsupervised Grammar Inference Using the Minimum Description Length Principle | 141 |
| <i>Upendra Sapkota, Barrett R. Bryant, and Alan Sprague</i> | |
| How Many Trees in a Random Forest? | 154 |
| <i>Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas</i> | |

Theory: Learning

| | |
|--|-----|
| Constructing Target Concept in Multiple Instance Learning Using Maximum Partial Entropy | 169 |
| <i>Tao Xu, David Chiu, and Iker Gondra</i> | |
| A New Learning Structure Heuristic of Bayesian Networks from Data ... | 183 |
| <i>Henri Bouhamed, Afif Masmoudi, Thierry Lecroq, and Ahmed Rebaï</i> | |
| Discriminant Subspace Learning Based on Support Vectors Machines ... | 198 |
| <i>Nikolaos Pitelis and Anastasios Tefas</i> | |
| A New Learning Strategy of General BAMs | 213 |
| <i>Hoa Thi Nong and The Duy Bui</i> | |
| Proximity-Graph Instance-Based Learning, Support Vector Machines, and High Dimensionality: An Empirical Comparison | 222 |
| <i>Godfried T. Toussaint and Constantin Berzan</i> | |

Theory: Clustering

| | |
|---|-----|
| Semi Supervised Clustering: A Pareto Approach | 237 |
| <i>Javid Ebrahimi and Mohammad Saniee Abadeh</i> | |
| Semi-supervised Clustering: A Case Study | 252 |
| <i>Andreia Silva and Cláudia Antunes</i> | |
| SOSTream: Self Organizing Density-Based Clustering over Data Stream | 264 |
| <i>Charlie Isaksson, Margaret H. Dunham, and Michael Hahsler</i> | |
| Clustering Data Stream by a Sub-window Approach Using DCA | 279 |
| <i>Minh Thuy Ta, Hoai An Le Thi, and Lydia Boudjeloud-Assala</i> | |
| Improvement of K-means Clustering Using Patents Metadata | 293 |
| <i>Mihai Vlase, Dan Munteanu, and Adrian Istrate</i> | |

WebMining

| | |
|---|-----|
| Content Independent Metadata Production as a Machine Learning Problem | 306 |
| <i>Sahar Changuel and Nicolas Labroche</i> | |
| Discovering K Web User Groups with Specific Aspect Interests | 321 |
| <i>Jianfeng Si, Qing Li, Tieyun Qian, and Xiaotie Deng</i> | |

Image Mining

| | |
|--|-----|
| An Algorithm for the Automatic Estimation of Image Orientation | 336 |
| <i>Mariusz Borawski and Dariusz Frejlichowski</i> | |
| Multi-label Image Annotation Based on Neighbor Pair Correlation Chain | 345 |
| <i>Guang Jiang, Xi Liu, and Zhongzhi Shi</i> | |
| Enhancing Image Retrieval by an Exploration-Exploitation Approach ... | 355 |
| <i>Luca Piras, Giorgio Giacinto, and Roberto Paredes</i> | |
| Finding Correlations between 3-D Surfaces: A Study in Asymmetric Incremental Sheet Forming | 366 |
| <i>M. Sulaiman Khan, Frans Coenen, Clare Dixon, and Subhieh El-Salhi</i> | |

Data Mining in Biometry and Security

| | |
|---|-----|
| Combination of Physiological and Behavioral Biometric for Human Identification | 380 |
| <i>Emdad Hossain and Girija Chetty</i> | |
| Detecting Actions by Integrating Sequential Symbolic and Sub-symbolic Information in Human Activity Recognition | 394 |
| <i>Michael Glodek, Friedhelm Schwenker, and Günther Palm</i> | |
| Computer Recognition of Facial Expressions of Emotion | 405 |
| <i>Ewa Pigłkowska and Jerzy Martyna</i> | |

Data Mining in Medicine

| | |
|---|-----|
| Outcome Prediction for Patients with Severe Traumatic Brain Injury Using Permutation Entropy Analysis of Electronic Vital Signs Data | 415 |
| <i>Konstantinos Kalpakis, Shiming Yang, Peter F. Hu, Colin F. Mackenzie, Lynn G. Stansbury, Deborah M. Stein, and Thomas M. Scalea</i> | |

| | |
|--|-----|
| EEG Signals Classification Using a Hybrid Method Based on Negative Selection and Particle Swarm Optimization | 427 |
| <i>Nasser Omer Ba-Karait, Siti Mariyam Shamsuddin, and Rubita Sudirman</i> | |

Data Mining in Environment and Water Quality Detection

| | |
|--|-----|
| DAGSVM vs. DAGKNN: An Experimental Case Study with Benthic Macroinvertebrate Dataset | 439 |
| <i>Henry Joutsijoki and Martti Juhola</i> | |

Image Mining in Medicine

| | |
|--|-----|
| Lung Nodules Classification in CT Images Using Shannon and Simpson Diversity Indices and SVM | 454 |
| <i>Leonardo Barros Nascimento, Anselmo Cardoso de Paiva, and Aristófanés Corrêa Silva</i> | |

| | |
|--|-----|
| Comparative Analysis of Feature Selection Methods for Blood Cell Recognition in Leukemia | 467 |
| <i>Tomasz Staroszczyk, Stanisław Osowski, and Tomasz Markiewicz</i> | |

| | |
|---|-----|
| Classification of Breast Tissues in Mammographic Images in Mass and Non-mass Using McIntosh's Diversity Index and SVM | 482 |
| <i>Péterson Moraes de Sousa Carvalho, Anselmo Cardoso de Paiva, and Aristófanés Corrêa Silva</i> | |

Text Mining

| | |
|---|-----|
| A Semi-Automated Approach to Building Text Summarisation Classifiers | 495 |
| <i>Matias Garcia-Constantino, Frans Coenen, P.-J. Noble, Alan Radford, and Christian Setzkorn</i> | |

| | |
|--|-----|
| A Pattern Recognition System for Malicious PDF Files Detection | 510 |
| <i>Davide Maiorca, Giorgio Giacinto, and Iginio Corona</i> | |

| | |
|--|-----|
| Text Categorization Using an Ensemble Classifier Based on a Mean Co-association Matrix | 525 |
| <i>Luís Moreira-Matias, João Mendes-Moreira, João Gama, and Pavel Brazdil</i> | |

| | |
|---|-----|
| A Pattern Discovery Model for Effective Text Mining | 540 |
| <i>Luepol Pipanmaekaporn and Yuefeng Li</i> | |

| | |
|--|-----|
| Investigating Usage of Text Segmentation and Inter-passage Similarities to Improve Text Document Clustering | 555 |
| <i>Shashank Paliwal and Vikram Pudi</i> | |

Data Mining in Network

| | |
|--|-----|
| Mining Ranking Models from Dynamic Network Data | 566 |
| <i>Lucrezia Macchia, Michelangelo Ceci, and Donato Malerba</i> | |
| Machine Learning-Based Classification of Encrypted Internet Traffic | 578 |
| <i>Talieh Seyed Tabatabaei, Mostafa Adel, Fakhri Karray, and Mohamed Kamel</i> | |
| Application of Bagging, Boosting and Stacking to Intrusion Detection | 593 |
| <i>Iwan Syarif, Ed Zaluska, Adam Prugel-Bennett, and Gary Wills</i> | |

Applications of Data Mining in Process Automation, Organisation Change Management, Telecommunication and Post Services

| | |
|--|-----|
| Classification of Elementary Stamp Shapes by Means of Reduced Point Distance Histogram Representation | 603 |
| <i>Paweł Forczmański and Dariusz Frejlichowski</i> | |
| A Multiclassifier Approach for Drill Wear Prediction | 617 |
| <i>Alberto Diez and Alberto Carrascal</i> | |
| Measuring the Dynamic Relatedness between Chinese Entities Orienting to News Corpus | 631 |
| <i>Zhishu Wang, Jing Yang, and Xin Lin</i> | |
| Prediction of Telephone User Attributes Based on Network Neighborhood Information | 645 |
| <i>Carlos Herrera-Yagüe and Pedro J. Zufiria</i> | |

Data Mining in Biology

| | |
|---|-----|
| A Hybrid Approach to Increase the Performance of Protein Folding Recognition Using Support Vector Machines | 660 |
| <i>Lavneet Singh, Girija Chetty, and Dharmendra Sharma</i> | |
| Author Index | 669 |