

Lecture Notes in Artificial Intelligence 2431

Subseries of Lecture Notes in Computer Science
Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer

Berlin

Heidelberg

New York

Barcelona

Hong Kong

London

Milan

Paris

Tokyo

Tapio Elomaa Heikki Mannila
Hannu Toivonen (Eds.)

Principles of Data Mining and Knowledge Discovery

6th European Conference, PKDD 2002
Helsinki, Finland, August 19-23, 2002
Proceedings



Springer

Series Editors

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

Volume Editors

Tapio Elomaa
Heikki Mannila
Hannu Toivonen
University of Helsinki, Department of Computer Science
P.O. Box 26, 00014 Helsinki, Finland
E-mail: {elomaa, heikki.mannila, hannu.toivonen}@cs.helsinki.fi

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Principles of data mining and knowledge discovery : 6th European conference ;
proceedings / PKDD 2002, Helsinki, Finland, August 19 - 23, 2002. Tapio
Elomaa ... (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ;
London ; Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2431 : Lecture notes in
artificial intelligence)
ISBN 3-540-44037-2

CR Subject Classification (1998): I.2, H.2, J.1, H.3, G.3, I.7, F.4.1

ISSN 0302-9743

ISBN 3-540-44037-2 Springer-Verlag 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-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York,
a member of BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2002
Printed in Germany

Typesetting: Camera-ready by author, data conversion by DA-TeX Gerd Blumenstein
Printed on acid-free paper SPIN: 10870106 06/3142 5 4 3 2 1 0

Preface

We are pleased to present the proceedings of the *13th European Conference on Machine Learning* (LNAI 2430) and the *6th European Conference on Principles and Practice of Knowledge Discovery in Databases* (LNAI 2431). These two conferences were colocated in Helsinki, Finland during August 19–23, 2002. ECML and PKDD were held together for the second year in a row, following the success of the colocation in Freiburg in 2001. Machine learning and knowledge discovery are two highly related fields and ECML/PKDD is a unique forum to foster their collaboration.

The benefit of colocation to both the machine learning and data mining communities is most clearly displayed in the common workshop, tutorial, and invited speaker program. Altogether six workshops and six tutorials were organized on Monday and Tuesday. As invited speakers we had the pleasure to have Erkki Oja (Helsinki Univ. of Technology), Dan Roth (Univ. of Illinois, Urbana-Champaign), Bernhard Schölkopf (Max Planck Inst. for Biological Cybernetics, Tübingen), and Padhraic Smyth (Univ. of California, Irvine).

The main events ran from Tuesday until Friday, comprising 41 ECML technical papers and 39 PKDD papers. In total, 218 manuscripts were submitted to these two conferences: 95 to ECML, 70 to PKDD, and 53 as joint submissions. All papers were assigned at least three reviewers from our international program committees. Out of the 80 accepted papers 31 were first accepted conditionally; the revised manuscripts were accepted only after the conditions set by the reviewers had been met.

Our special thanks go to the tutorial chairs Johannes Fürnkranz and Myra Spiliopoulou and the workshop chairs Hendrik Blockeel and Jean-François Boulicaut for putting together an exiting combined tutorial and workshop program. Also the challenge chair Petr Berka deserves our sincerest gratitude. All the members of both program committees are thanked for devoting their expertise to the continued success of ECML and PKDD. The organizing committee chaired by Helena Ahonen-Myka worked hard to make the conferences possible. A special mention has to be given to Oskari Heinonen for designing and maintaining the web pages and Ilkka Koskenniemi for maintaining CyberChair, which was developed by Richard van de Stadt. We thank Alfred Hofmann of Springer-Verlag for cooperation in publishing these proceedings. We gratefully acknowledge the financial support of the Academy of Finland and KDNet.

We thank all the authors for contributing to what in our mind is a most interesting technical program for ECML and PKDD. We trust that the week in late August was most enjoyable for all members of both research communities.

June 2002

Tapio Elomaa
Heikki Mannila
Hannu Toivonen

ECML/PKDD-2002 Organization

Executive Committee

Program Chairs:	Tapio Elomaa (Univ. of Helsinki) Heikki Mannila (Helsinki Inst. for Information Technology and Helsinki Univ. of Technology) Hannu Toivonen (Nokia Research Center and Univ. of Helsinki)
Tutorial Chairs:	Johannes Fürnkranz (Austrian Research Inst. for Artificial Intelligence) Myra Spiliopoulou (Leipzig Graduate School of Management)
Workshop Chairs:	Hendrik Blockeel (Katholieke Universiteit Leuven) Jean-François Boulicaut (INSA Lyon)
Challenge Chair:	Petr Berka (University of Economics, Prague)
Organizing Chair:	Helena Ahonen-Myka (Univ. of Helsinki)
Organizing Committee:	Oskari Heinonen, Ilkka Koskenniemi, Greger Lindén, Pirjo Moen, Matti Nykänen, Anna Pienimäki, Ari Rantanen, Juho Rousu, Marko Salmenkivi (Univ. of Helsinki)

ECML Program Committee

H. Blockeel, Belgium	A. Hyvärinen, Finland
I. Bratko, Slovenia	T. Joachims, USA
P. Brazdil, Portugal	Y. Kodratoff, France
H. Boström, Sweden	I. Kononenko, Slovenia
W. Burgard, Germany	S. Kramer, Germany
N. Cristianini, USA	M. Kubat, USA
J. Cussens, UK	N. Lavrač, Slovenia
L. De Raedt, Germany	C.X. Ling, Canada
M. Dorigo, Belgium	R. López de Màntaras, Spain
S. Džeroski, Slovenia	D. Malerba, Italy
F. Esposito, Italy	S. Matwin, Canada
P. Flach, UK	R. Meir, Israel
J. Fürnkranz, Austria	J. del R. Millán, Switzerland
J. Gama, Portugal	K. Morik, Germany
J.-G. Ganascia, France	H. Motoda, Japan
T. Hofmann, USA	R. Nock, France
L. Holmström, Finland	E. Plaza, Spain

G. Paliouras, Greece	H. Tirri, Finland
J. Rousu, Finland	P. Turney, Canada
L. Saitta, Italy	R. Vilalta, USA
T. Scheffer, Germany	P. Vitányi, The Netherlands
M. Sebag, France	S. Weiss, USA
J. Shawe-Taylor, UK	G. Widmer, Austria
A. Siebes, The Netherlands	R. Wirth, Germany
D. Sleeman, UK	S. Wrobel, Germany
M. van Someren, The Netherlands	Y. Yang, USA
P. Stone, USA	

PKDD Program Committee

H. Ahonen-Myka, Finland	S. Morishita, Japan
E. Baralis, Italy	H. Motoda, Japan
J.-F. Boulicaut, France	G. Nakhaeizadeh, Germany
N. Cercone, Canada	Z.W. Raś, USA
B. Crémilleux, France	J. Rauch, Czech Republic
L. De Raedt, Germany	G. Ritschard, Switzerland
L. Dehaspe, Belgium	M. Sebag, France
S. Džeroski, Slovenia	F. Sebastiani, Italy
M. Ester, Canada	M. Sebban, France
R. Feldman, Israel	B. Seeger, Germany
P. Flach, UK	A. Siebes, The Netherlands
E. Frank, New Zealand	A. Skowron, Poland
A. Freitas, Brazil	M. van Someren, The Netherlands
J. Fürnkranz, Austria	M. Spiliopoulou, Germany
H.J. Hamilton, Canada	N. Spyrtatos, France
J. Han, Canada	E. Suzuki, Japan
R. Hilderman, Canada	A.-H. Tan, Singapore
S.J. Hong, USA	S. Tsumoto, Japan
S. Kaski, Finland	A. Unwin, Germany
D. Keim, USA	J. Wang, USA
J.-U. Kietz, Switzerland	K. Wang, Canada
R. King, UK	L. Wehenkel, Belgium
M. Klemettinen, Finland	D. Wettschereck, Germany
W. Klösgen, Germany	G. Widmer, Austria
Y. Kodratoff, France	R. Wirth, Germany
J.N. Kok, The Netherlands	S. Wrobel, Germany
S. Kramer, Germany	M. Zaki, USA
S. Matwin, Canada	

Additional Reviewers

N. Abe	S. Haustein	L. Peña
F. Aiolfi	J. He	Y. Peng
Y. Altun	K.G. Herbert	J. Petrak
S. de Amo	J. Himberg	V. Phan Luong
A. Appice	J. Hipp	K. Rajaraman
E. Armengol	S. Hoche	T. Reinartz
T.G. Ault	J. Hosking	I. Renz
J. Azé	E. Hüllermeier	C. Rigotti
M.T. Basile	P. Juvan	F. Rioult
A. Bonarini	M. Kääriäinen	M. Robnik-Šikonja
R. Bouckaert	D. Kalles	M. Roche
P. Brockhausen	V. Karkaletsis	B. Rosenfeld
M. Brodie	A. Karwath	S. Rüping
W. Buntine	K. Kersting	M. Salmenkivi
J. Carbonell	J. Kindermann	A.K. Seewald
M. Ceci	R. Klinkenberg	H. Shan
S. Chikkanna-Naik	P. Koistinen	J. Sinkkonen
S. Chiusano	C. Köpf	J. Struyf
R. Cicchetti	R. Kosala	R. Taouil
A. Clare	W. Kusters	J. Taylor
M. Degemmis	M.-A. Krogel	L. Todorovski
J. Demsar	M. Kukar	T. Urbancic
F. De Rosis	L. Lakhal	K. Vasko
N. Di Mauro	G. Lebanon	H. Wang
G. Dorffner	S.D. Lee	Y. Wang
G. Dounias	F. Li	M. Wiering
N. Durand	J.T. Lindgren	S. Wu
P. Erästö	J. Liu	M.M. Yin
T. Erjavec	Y. Liu	F. Zambetta
J. Farrand	M.-C. Ludl	B. Ženko
S. Ferilli	S. Mannor	J. Zhang
P. Floréen	R. Meo	S. Zhang
J. Franke	N. Meuleau	T. Zhang
T. Gaertner	H. Mogg-Schneider	M. Zlochin
P. Gallinari	R. Natarajan	B. Zupan
P. Garza	S. Nijssen	
A. Giacometti	G. Paaß	

Tutorials

Text Mining and Internet Content Filtering

José María Gómez Hidalgo

Formal Concept Analysis

Gerd Stumme

Web Usage Mining for E-business Applications

Myra Spiliopoulou, Bamshad Mobasher, and Bettina Berendt

Inductive Databases and Constraint-Based Mining

Jean-François Boulicaut and Luc De Raedt

An Introduction to Quality Assessment in Data Mining

Michalis Vazirgiannis and M. Halkidi

Privacy, Security, and Data Mining

Chris Clifton

Workshops

Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning

Marko Bohanec, Dunja Mladenić, and Nada Lavrač

Visual Data Mining

Simeon J. Simoff, Monique Noirhomme-Fraiture, and Michael H. Böhlen

Semantic Web Mining

Bettina Berendt, Andreas Hotho, and Gerd Stumme

Mining Official Data

Paula Brito and Donato Malerba

Knowledge Discovery in Inductive Databases

Mika Klemettinen, Rosa Meo, Fosca Giannotti, and Luc De Raedt

Discovery Challenge Workshop

Petr Berka, Jan Rauch, and Shusaku Tsumoto

Table of Contents

Contributed Papers

Optimized Substructure Discovery for Semi-structured Data	1
<i>Kenji Abe, Shinji Kawasoe, Tatsuya Asai, Hiroki Arimura, and Setsuo Arikawa</i>	
Fast Outlier Detection in High Dimensional Spaces	15
<i>Fabrizio Angiulli and Clara Pizzuti</i>	
Data Mining in Schizophrenia Research – Preliminary Analysis	27
<i>Stefan Arnborg, Ingrid Agartz, Håkan Hall, Erik Jönsson, Anna Sillén, and Göran Sedvall</i>	
Fast Algorithms for Mining Emerging Patterns	39
<i>James Bailey, Thomas Manoukian, and Kotagiri Ramamohanarao</i>	
On the Discovery of Weak Periodicities in Large Time Series	51
<i>Christos Berberidis, Ioannis Vlahavas, Walid G. Aref, Mikhail Atallah, and Ahmed K. Elmagarmid</i>	
The Need for Low Bias Algorithms in Classification Learning from Large Data Sets	62
<i>Damien Brain and Geoffrey I. Webb</i>	
Mining All Non-derivable Frequent Itemsets	74
<i>Toon Calders and Bart Goethals</i>	
Iterative Data Squashing for Boosting Based on a Distribution-Sensitive Distance	86
<i>Yuta Choki and Einoshin Suzuki</i>	
Finding Association Rules with Some Very Frequent Attributes	99
<i>Frans Coenen and Paul Leng</i>	
Unsupervised Learning: Self-aggregation in Scaled Principal Component Space	112
<i>Chris Ding, Xiaofeng He, Hongyuan Zha, and Horst Simon</i>	
A Classification Approach for Prediction of Target Events in Temporal Sequences	125
<i>Carlotta Domeniconi, Chang-shing Perng, Ricardo Vilalta, and Sheng Ma</i>	
Privacy-Oriented Data Mining by Proof Checking	138
<i>Amy Felty and Stan Matwin</i>	

XII Table of Contents

Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification	150
<i>George Forman</i>	
Generating Actionable Knowledge by Expert-Guided Subgroup Discovery	163
<i>Dragan Gamberger and Nada Lavrač</i>	
Clustering Transactional Data	175
<i>Fosca Giannotti, Cristian Gozzi, and Giuseppe Manco</i>	
Multiscale Comparison of Temporal Patterns in Time-Series Medical Databases	188
<i>Shoji Hirano and Shusaku Tsumoto</i>	
Association Rules for Expressing Gradual Dependencies	200
<i>Eyke Hüllermeier</i>	
Support Approximations Using Bonferroni-Type Inequalities	212
<i>Szymon Jaroszewicz and Dan A. Simovici</i>	
Using Condensed Representations for Interactive Association Rule Mining	225
<i>Baptiste Jeudy and Jean-François Boulicaut</i>	
Predicting Rare Classes: Comparing Two-Phase Rule Induction to Cost-Sensitive Boosting	237
<i>Mahesh V. Joshi, Ramesh C. Agarwal, and Vipin Kumar</i>	
Dependency Detection in MobiMine and Random Matrices	250
<i>Hillol Kargupta, Krishnamoorthy Sivakumar, and Samiran Ghosh</i>	
Long-Term Learning for Web Search Engines	263
<i>Charles Kemp and Kotagiri Ramamohanarao</i>	
Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database	275
<i>Willi Klösgen and Michael May</i>	
Involving Aggregate Functions in Multi-relational Search	287
<i>Arno J. Knobbe, Arno Siebes, and Bart Marseille</i>	
Information Extraction in Structured Documents Using Tree Automata Induction	299
<i>Raymond Kosala, Jan Van den Bussche, Maurice Bruynooghe, and Hendrik Blockeel</i>	
Algebraic Techniques for Analysis of Large Discrete-Valued Datasets	311
<i>Mehmet Koyutürk, Ananth Grama, and Naren Ramakrishnan</i>	
Geography of Differences between Two Classes of Data	325
<i>Jinyan Li and Limsoon Wong</i>	

Rule Induction for Classification of Gene Expression Array Data	338
<i>Per Lidén, Lars Asker, and Henrik Boström</i>	
Clustering Ontology-Based Metadata in the Semantic Web	348
<i>Alexander Maedche and Valentin Zacharias</i>	
Iteratively Selecting Feature Subsets for Mining from High-Dimensional Databases	361
<i>Hiroshi Mamitsuka</i>	
SVM Classification Using Sequences of Phonemes and Syllables	373
<i>Gerhard Paaß, Edda Leopold, Martha Larson, Jörg Kindermann, and Stefan Eickeler</i>	
A Novel Web Text Mining Method Using the Discrete Cosine Transform	385
<i>Laurence A.F. Park, Marimuthu Palaniswami, and Kotagiri Ramamohanarao</i>	
A Scalable Constant-Memory Sampling Algorithm for Pattern Discovery in Large Databases	397
<i>Tobias Scheffer and Stefan Wrobel</i>	
Answering the Most Correlated N Association Rules Efficiently	410
<i>Jun Sese and Shinichi Morishita</i>	
Mining Hierarchical Decision Rules from Clinical Databases Using Rough Sets and Medical Diagnostic Model	423
<i>Shusaku Tsumoto</i>	
Efficiently Mining Approximate Models of Associations in Evolving Databases	435
<i>Adriano Veloso, Bruno Gusmão, Wagner Meira Jr., Marcio Carvalho, Sriini Parthasarathy, and Mohammed Zaki</i>	
Explaining Predictions from a Neural Network Ensemble One at a Time	449
<i>Robert Wall, Pádraig Cunningham, and Paul Walsh</i>	
Structuring Domain-Specific Text Archives by Deriving a Probabilistic XML DTD	461
<i>Karsten Winkler and Myra Spiliopoulou</i>	
Separability Index in Supervised Learning	475
<i>Djamel A. Zighed, Stéphane Lallich, and Fabrice Muhlenbach</i>	

Invited Papers

Finding Hidden Factors Using Independent Component Analysis	488
<i>Erkki Oja</i>	

XIV Table of Contents

Reasoning with Classifiers	489
<i>Dan Roth</i>	
A Kernel Approach for Learning from Almost Orthogonal Patterns	494
<i>Bernhard Schölkopf, Jason Weston, Eleazar Eskin, Christina Leslie,</i> <i>and William Stafford Noble</i>	
Learning with Mixture Models: Concepts and Applications	512
<i>Padhraic Smyth</i>	
Author Index	513