Fabrice Guillet, Gilbert Ritschard, Djamel Abdelkader Zighed, and Henri Briand (Eds.)

Advances in Knowledge Discovery and Management

Studies in Computational Intelligence, Volume 292

Further volumes of this series can be found on our homepage: springer.com

Vol. 271. Janusz Kacprzyk, Frederick E. Petry, and Adnan Yazici (Eds.)
Uncertainty Approaches for Spatial Data Modeling and Processing, 2009
ISBN 978-3-642-10662-0

Vol. 272. Carlos A. Coello Coello, Clarisse Dhaenens, and Laetitia Jourdan (Eds.)
Advances in Multi-Objective Nature Inspired Computing, 2009
ISBN 978-3-642-11217-1

Vol. 273. Fatos Xhafa, Santi Caballé, Ajith Abraham, Thanasis Daradoumis, and Angel Alejandro Juan Perez (Eds.)
Computational Intelligence for Technology Enhanced Learning, 2010
ISBN 978-3-642-11223-2

Vol. 274. Zbigniew W. Raś and Alicja Wieczorkowska (Eds.)
Advances in Music Information Retrieval, 2010
ISBN 978-3-642-11673-5

Vol. 275. Dilip Kumar Pratihar and Lakhmi C. Jain (Eds.)
Intelligent Autonomous Systems, 2010
ISBN 978-3-642-11675-9

Vol. 276. Jacek Mańdziuk
Knowledge-Free and Learning-Based Methods in Intelligent Game Playing, 2010
ISBN 978-3-642-11677-3

Vol. 277. Filippo Spagnolo and Benedetto Di Paola (Eds.)
European and Chinese Cognitive Styles and their Impact on Teaching Mathematics, 2010
ISBN 978-3-642-11679-7

Vol. 278. Radomir S. Stankovic and Jaakko Astola
From Boolean Logic to Switching Circuits and Automata, 2010
ISBN 978-3-642-11681-0

Vol. 279. Manolis Wallace, Ioannis E. Anagnostopoulos, Phivos Mylonas, and Maria Bielikova (Eds.)
Semantics in Adaptive and Personalized Services, 2010
ISBN 978-3-642-11683-4

Vol. 280. Chang Wen Chen, Zhu Li, and Shiguo Lian (Eds.)
Intelligent Multimedia Communication: Techniques and Applications, 2010
ISBN 978-3-642-11685-8

Vol. 281. Robert Babuska and Frans C.A. Groen (Eds.)
Interactive Collaborative Information Systems, 2010
ISBN 978-3-642-11687-2

Vol. 282. Husrev Taha Sencar, Sergio Velastin, Nikolaos Nikolaidis, and Shiguo Lian (Eds.)
Intelligent Multimedia Analysis for Security Applications, 2010
ISBN 978-3-642-11754-1

Vol. 283. Ngoc Thanh Nguyen, Radoslaw Katarzyniak, and Shyi-Ming Chen (Eds.)
Advances in Intelligent Information and Database Systems, 2010
ISBN 978-3-642-12089-3

Vol. 284. Juan R. González, David Alejandro Pelta, Carlos Cruz, Germán Terrazas, and Natalio Krasnogor (Eds.)
Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), 2010
ISBN 978-3-642-12537-9

Vol. 285. Roberto Cipolla, Sebastiano Battiato, and Giovanni Maria Farinella (Eds.)
Computer Vision, 2010
ISBN 978-3-642-12847-9

Vol. 286. Zeev Volkovich, Alexander Bolshoy, Valery Kirzhner, and Zeev Barzily
Genome Clustering, 2010
ISBN 978-3-642-12951-3

Vol. 287. Dan Schonfeld, Caifeng Shan, Dacheng Tao, and Liang Wang (Eds.)
Video Search and Mining, 2010
ISBN 978-3-642-12899-8

Vol. 288. I-Hsien Ting, Hui-Ju Wu, Tien-Hwa Ho (Eds.)
Mining and Analyzing Social Networks, 2010
ISBN 978-3-642-13421-0

Vol. 289. Anne Håkansson, Ronald Hartung, and Ngoc Thanh Nguyen (Eds.)
Agent and Multi-agent Technology for Internet and Enterprise Systems, 2010
ISBN 978-3-642-13525-5

Vol. 290. Weiliang Xu and John E. Bronlund
Mastication Robots, 2010
ISBN 978-3-540-93902-3

Vol. 291. Shimon Whiteson
Adaptive Representations for Reinforcement Learning, 2010
ISBN 978-3-642-13931-4

Vol. 292. Fabrice Guillet, Gilbert Ritschard, Djamel Abdelkader Zighed, and Henri Briand (Eds.)
Advances in Knowledge Discovery and Management, 2010
ISBN 978-3-642-00579-4

Fabrice Guillet, Gilbert Ritschard,
Djamel Abdelkader Zighed, and Henri Briand (Eds.)

Advances in Knowledge Discovery and Management

Springer

Fabrice Guillet
LINA (CNRS UMR 6241)
Polytechnic School of Nantes University
rue C. Pauc, BP 50609
F-44306 Nantes Cedex 3
France
E-mail: Fabrice.Guillet@univ-nantes.fr

Djamel Abdelkader Zighed
Laboratoire ERIC
Université Lumière Lyon 2
5, avenue Pierre Mendès-France, Bât L.
69600 Bron
France
E-mail: Abdelkader.Zighed@univ-lyon2.fr

Gilbert Ritschard
Université de Genève
Department of Econometrics
Uni-Mail, 40, bd du Pont-d'Arve, room 5232
CH-1211 Geneva 4
Switzerland
E-mail: Gilbert.Ritschard@unige.ch

Henri Briand
LINA (CNRS UMR 6241)
Polytechnic School of Nantes University
rue C. Pauc, BP 50609
F-44306 Nantes Cedex 3
France
E-mail: Henri.Braind@univ-nantes.fr
http://www.polytech.univ-nantes.fr/COD/
?Pages_personnelles:Henri_Briand

ISBN 978-3-642-00579-4 e-ISBN 978-3-642-00580-0

DOI 10.1007/978-3-642-00580-0

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: 2010928588

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

During the last decade, the French-speaking scientific community developed a very strong research activity in the field of Knowledge Discovery and Management (KDM or EGC for "Extraction et Gestion des Connaissances" in French), which is concerned with, among others, Data Mining, Knowledge Discovery, Business Intelligence, Knowledge Engineering and Semantic Web. This emerging research area has also been strongly stimulated by the rapid growth of information systems and the web semantic issues.

The success of the first two French-speaking EGC Conferences in 2001 and 2002 resulted naturally in 2002 in the foundation of the International French-speaking EGC Association[1]. The Association organizes since then regular conferences and workshops with the aim of promoting exchanges between researchers and companies concerned with KDM and its application in business, administration, industry or public organizations.

The recent and novel research contributions collected in this book are extended and reworked versions of a selection of the best papers that were originally presented in French at the EGC 2009 Conference held in Strasbourg, France on January 2009.

Structure of the Book

The volume is organized in four parts.

Part I includes five papers concerned by various aspects of *supervised learning or information retrieval.*

The first paper by Matthias Studer and his colleagues considers complex objects such as state sequences for which we can compute pairwise dissimilarities and proposes an original ANOVA-like approach for measuring the information that a predictor provides about their discrepancy. The technique can be extended in the form of a regression tree for complex objects, which

[1] Association "Extraction et Gestion des Connaissances" (EGC), *www.egc.asso.fr*

is convincingly demonstrated through an application on sequential data describing Swiss occupational trajectories.

Extension to multiple factors and as a tree structured analysis to find out how covariates can explain the object discrepancy. Application to the study of Swiss occupational trajectories.

In the second article, Nicolas Voisine, Marc Boullé and Carine Hue propose a parameter free Bayesian approach to evaluate the overall quality of a decision tree grown from a large data base. This permits to transform the learning problem into an optimization one consisting in searching the tree that optimizes the overall criterion. Extensive experimentation results demonstrate that such optimal trees reach similar predictive performance as state-of-the-art trees, while being much simpler and hence easier to understand.

Thanh-Nghi Do and his colleagues are interested in random forest approaches for very-high-dimensional data with dependencies. They introduce a new oblique decision tree method with SVM-based split functions that work on randomly chosen predictors. Comparative experiments show that the proposed approach makes on very-high-dimensional data clearly better in terms of precision, recall and accuracy than random forests of C4.5.

The contribution of Nguyen-Khang Pham and his associates deals with large scale content-based image retrieval. The authors propose a solution based on Correspondence Analysis (CA) of SIFT local descriptors and introduce an original incremental CA algorithm that scales to huge databases. Response time is further improved by accounting of contextual dissimilarities during the search process. The efficiency of the proposed process is assessed through a series of tests performed on a database of more than 1 million of images.

In the last paper of this first group, Emanuel Aldea and Isabelle Bloch examine structural representations of images for machine learning and image categorization. Resorting to a graph representation in which edges describe spatial relations, they derive metrics between images by means of a graph kernel approach that explicitly accounts for spatial interactions. The authors extend their approach to the case of fuzzy spatial relations and study its behavior in the context of discriminative learning by means of a series of experimentations.

Part II presents five papers concerned with *unsupervised learning* issues.

The first of them by Gilles Hubert and Olivier Teste proposes a new OLAP operator in the context of multidimensional databases that proves very useful to facilitate multigranular analyses. Multigranular analyses aim at looking at the same data at different aggregation levels, which usually supposes to run multiple analyses. The proposed tool permits to switch from one granularity level to the other on the fly during the analysis.

The paper by Sébastien Lefèvre is concerned with image segmentation, an issue that can be seen as a data clustering problem with spatial constraints. Such problems are classically solved by running first a unconstraint clustering method and submitting then the results to additional spatial post-processing. The author proposes here a new solution able to perform image segmentation in a same single round of analysis.

Nistor Grozavu and his associates propose two Self-Organizing-Map-based algorithms for the selection of relevant features through unsupervised weighting. The proposed methods provide also as a byproduct the characterization of clusters. The interest of the methods is demonstrated through a series of experimental results.

The next paper by Guillaume Cleuziou deals with overlapping clustering and presents two extensions of overlapping k-means (OKM). The first one generalizes the k-medoids method to overlapping clustering and proves useful in organizing non metric data from their proximity matrix. The second one, suitable for metric data, is a weighted version of OKM that allows for non-spherical clusters.

The last paper of this second group by Romain Bourqui and his co-authors deals also with overlapping clusters but in a dynamic social network setting. Such networks are modeled as graphs and the aim is to decompose it into similar sets of nodes. The paper provides a very efficient algorithm that can detect major changes in the network as it evolves over time.

Part III includes two papers on *data streaming* and two on *security.*

Nesrine Gabsi and colleagues present a new approach to build historical summaries of data streams. It is based on a combination of sampling and clustering algorithms. The benefit of this combination is empirically demonstrated.

The paper by Lionel Vinceslas and associates is about the mining of sequential patterns in data streams for which it proposes an algorithm that works online using a deterministic finite automaton as a summary structure.

The two next papers are about intrusion detection. Goverdhan Singh and colleagues are concerned by the rate of false alarms in outlier-based intrusion detection systems. They attempt to reduce that rate by looking at the repetition of intrusions from one system to another and propose solutions for separating the outliers from the normal behavioursin a streaming environment and for comparing the outliers of two systems.

Nischal Verma and associates address the problem of intrusion detection on Internet applications and propose a new secure collaborative approach. The main advantage of the proposed method is both to detectnew attacks by using information stored in different sites and to ensure that private data will not be disclosed.

Part IV The last four papers are about *ontologies and semantic.*

The first one by Fayçal Hamdi and his co-authors proposes a two promising ontology-partitioning methods designed to take alignment objective into account in the partitioning process.

The paper by Farid Cerbah is concerned with the automatic construction of rich semantic models or ontologies from relational databases. An important limitation of such automatic processes is that that they most often end up with flat models that simply mirror the definition schemas of the source databases. The paper shows how relevant categorisation patterns can be identified within the data by combining lexical filtering and entropy-based criteria.

The article by Alina Dia Miron and her associates is concerned with formal languages for describing ontologies. It considers OWL Description Logic for which it adapts semantic analysis techniques that permit to exploit individual spatio-temporal annotations to limit the scope of the queries and thus increase efficiency.

The last paper by Alain Lelu and Martine Cadot attempts to find links and anti-links between presence-absence attributes. By mean of a randomization approach the proposed method checks if the co-occurrences in a series of randomized data sets is significantly above (anti-link) or below (link) than the co-occurrences in the original data set. The scope of the method is illustrated on a collection of texts.

Acknowledgments

The editors would like to thank the chapter authors for their insights and contributions to this book.

The editors would also like to acknowledge the members of the review committee and the associated referees for their involvement in the review process of the book. Their in depth reviewing, criticisms and constructive remarks significantly contributed to the high quality of the retained papers.

A special thank goes to Matthias Studer who has efficiently composed and laid out the manuscript.

Finally, we thank Springer and the publishing team, and especially T. Ditzinger and J. Kacprzyk, for their confidence in our project.

Nantes, Geneva, Lyon
February 2010

Fabrice Guillet
Gilbert Ritschard
Djamel Abdelkader Zighed
Henri Briand

Organization

Review Committee

All published chapters have been reviewed by at least 2 referees.

Jiawei Han	Univ. of Illinois, USA - Honorary Member
Tomas Aluja-Banet	UPC, Barcelona, Spain
Nadir Belkhiter	Univ. Laval, Québec, Canada
Samy Bengio	Google Inc., Mountain View California, USA
Younès Bennani	Univ. Paris 13, France
Sadok Ben Yahia	Univ. Tunis, Tunisia
Paula Brito	Univ. of Porto, Portugal
Gilles Falquet	Univ. of Geneva, Switzerland
Georges Gardarin	Univ. of Versailles Saint-Quentin, France
Mohand-Saïd Hacid	Univ. of Lyon I, France
Robert Hilderman	University of Regina, Canada
Yves Kodratoff	LRI, University of Paris-Sud, France
Pascale Kuntz	LINA, University of Nantes, France
Ludovic Lebart	ENST, Paris, France
Philippe Lenca	Enst-Bretagne, Brest, France
Philippe Leray	Univ. of Nantes, France
Monique Noirhomme-Fraiture	FUNDP, Namur, Belgium
Stan Matwin	Univ. of Ottawa, Canada
Pascal Poncelet	LIRMM, Univ. of Montpellier, France
Jan Rauch	University of Prague, Czech Republic
Ansaf Salleb-Aouissi	Columbia Univ., New York, USA
Gilbert Saporta	CNAM, Paris, France
Florence Sédes	Univ. of Toulouse 3, France
Dan Simovici	Univ. of Massachusetts Boston, USA
Gilles Venturini	Univ. of Tours, France
Jeff Wijsen	Univ. of Mons-Hainaut, Belgium

Associated Reviewers

Emanuel Aldea,
Romain Bourqui,
Farid Cerbah,
Guillaume Cleuziou,
Marta Franova,
Céline Fiot,
Nesrine Gabsi,
Nistor Grozavu,
Fayçal Hamdi,
Gilles Hubert,
Alain Lelu,
Sébastien Lefèvre,
Cécile Low-Kam,
Florent Masseglia,
Alina-Dia Miron,
Nguyen-Khang Pham,
Stefano Perabò,
Matthias Studer,
Olivier Teste,
Thanh-Nghi Do,
Lionel Vinceslas,
Nicolas Voisine,

Manuscript Coordinator

Matthias Studer (Univ. of Geneva, Switzerland)

Contents

Part I: Supervised Learning and Information Retrieval

Discrepancy Analysis of Complex Objects Using Dissimilarities . . . 3
Matthias Studer, Gilbert Ritschard, Alexis Gabadinho, Nicolas S. Müller

A Bayes Evaluation Criterion for Decision Trees . . . 21
Nicolas Voisine, Marc Boullé, Carine Hue

Classifying Very-High-Dimensional Data with Random Forests of Oblique Decision Trees . . . 39
Thanh-Nghi Do, Philippe Lenca, Stéphane Lallich, Nguyen-Khang Pham

Intensive Use of Correspondence Analysis for Large Scale Content-Based Image Retrieval . . . 57
Nguyen-Khang Pham, Annie Morin, Patrick Gros, Quyet-Thang Le

Toward a Better Integration of Spatial Relations in Learning with Graphical Models . . . 77
Emanuel Aldea, Isabelle Bloch

Part II: Unsupervised Learning

Multigranular Manipulations for OLAP Querying . . . 97
Gilles Hubert, Olivier Teste

A New Approach for Unsupervised Classification in Image Segmentation . . . 113
Sébastien Lefèvre

Cluster-Dependent Feature Selection through a Weighted Learning Paradigm . . . 133
Nistor Grozavu, Younès Bennani, Mustapha Lebbah

Two Variants of the OKM for Overlapping Clustering 149
Guillaume Cleuziou

A Stable Decomposition Algorithm for Dynamic Social Network Analysis 167
Romain Bourqui, Paolo Simonetto, Fabien Jourdan

Part III: Security and Data Streaming

An Hybrid Data Stream Summarizing Approach by Sampling and Clustering 181
Nesrine Gabsi, Fabrice Clérot, Georges Hébrail

SPAMS: A Novel Incremental Approach for Sequential Pattern Mining in Data Streams 201
Lionel Vinceslas, Jean-Emile Symphor, Alban Mancheron, Pascal Poncelet

Mining Common Outliers for Intrusion Detection 217
Goverdhan Singh, Florent Masseglia, Céline Fiot, Alice Marascu, Pascal Poncelet

Intrusion Detections in Collaborative Organizations by Preserving Privacy 235
Nischal Verma, François Trousset, Pascal Poncelet, Florent Masseglia

Part IV: Ontologies and Semantic

Alignment-Based Partitioning of Large-Scale Ontologies 251
Fayçal Hamdi, Brigitte Safar, Chantal Reynaud, Haïfa Zargayouna

Learning Ontologies with Deep Class Hierarchies by Mining the Content of Relational Databases 271
Farid Cerbah

Semantic Analysis for the Geospatial Semantic Web 287
Alina Dia Miron, Jérôme Gensel, Marlène Villanova-Oliver

Statistically Valid Links and Anti-links Between Words and Between Documents: Applying TourneBool Randomization Test to a Reuters Collection 307
Alain Lelu, Martine Cadot

List of Contributors 325

Author Index 337