Studies in Computational Intelligence

Volume 615

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/7092

Fabrice Guillet · Bruno Pinaud
Gilles Venturini · Djamel Abdelkader Zighed
Editors

Advances in Knowledge Discovery and Management

Volume 5

Springer

Editors
Fabrice Guillet
LINA (CNRS UMR 6241)
Polytech'Nantes, Nantes University
Nantes
France

Gilles Venturini
Polytech'Tours
François Rabelais Tours University
Tours
France

Bruno Pinaud
LaBRI (CNRS UMR 5800)
University of Bordeaux
Talence
France

Djamel Abdelkader Zighed
ERIC
Lyon 2 University
Bron
France

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-319-23750-3 ISBN 978-3-319-23751-0 (eBook)
DOI 10.1007/978-3-319-23751-0

Library of Congress Control Number: 2015948724

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

The recent and novel research contributions collected in this book are extended and reworked versions of a selection of the five best papers that were originally presented in French at the EGC'2013 Conference held in Toulouse, France, on January 2013 and one paper from the EGC'2014 Conference help in Rennes, France, on January 2014. The five papers from the 2013 edition of the conference have been selected from the 26 papers accepted in long format at the conference. These 26 long papers were themselves the result of a peer and blind review process among the 123 papers initially submitted to the conference in 2013 (acceptance rate of 26 % for long papers). This conference was the 13th edition of this event, which takes place each year and which is now successful and well-known in the French-speaking community. This community was structured in 2003 by the foundation of the International French-speaking EGC society (EGC in French stands for "Extraction et Gestion des Connaissances" and means "Knowledge Discovery and Management", or KDM). This society organizes every year not only its main conference (about 200 attendees) but also workshops and other events with the aim of promoting exchanges between researchers and companies concerned with KDM and its applications in business, administration, industry, or public organizations. For more details about the EGC society, please consult http://www.egc.asso.fr.

Structure of the Book

This book is a collection of representative and novel works done in Data Mining, Knowledge Discovery, Clustering, and Classification. It is intended to be read by all researchers interested in these fields, including Ph.D. or M.Sc. students, and researchers from public or private laboratories. It concerns both theoretical and practical aspects of KDM.

This book has been structured into two parts. The first three chapters are related to novel applications on real datasets of various origins. The second part of this book presents three methodological chapters on the foundations of knowledge extraction and management.

Chapter "A Study of the Spatio-Temporal Correlations in Mobile Calls Networks" proposes an analysis of phone-call detailed records collected during five months in France. MODL, a nonparametric method, is applied to solve two different problems: first, partitioning antennas leading to territory segmentation; and second, discretizing time aiming at determining changes in users' behavior. A set of visualizations, emphasizing the most interesting patterns, eases the analysis and the interpretation of the results. Chapter "Co-Clustering Network-Constrained Trajectory Data" study the problem of clustering moving object trajectories in a road network environment. A bipartite graph representation is used to model the relationships between trajectories and road segments visited. The authors propose three approaches to clustering the vertices of such a graph. Using synthetic data, they demonstrate how the data can be used to gain insight about mobility in road networks such as detecting frequent routes, characterizing road segment roles, etc. The work by Grabar and colleagues presented in Chap. "Medical Discourse and Subjectivity" proposes a contrastive study of corpora from the medical field. The corpora contain documents that are differentiated by their specialization level: documents written by medical experts and by patients. The differentiation features are related to medical notions, uncertainty, emotions, and negation. These features appear to be relevant for the distinction between the types of documents aimed. The authors then discuss the roles played by uncertainty, emotions, and negation in these documents.

Chapter "Relational Concept Analysis for Relational Data Exploration" deals with Relational Concept Analysis (RCA) which is an unsupervised classification method producing a set of connected concept lattices by considering relations between objects from different contexts. While designed to be intuitive to extract knowledge from relational data, dealing with many relations with RCA implies scalability problems. This article presents an adaptation of RCA, tested on environmental data, to explore relations in a guided way in order to increase the performance and the pertinence of the results. In Chap. "Dynamic Recommender System: Using Cluster-Based Biases to Improve the Accuracy of the Predictions", the authors propose a methodology for recommender systems based on Matrix Factorization (MF) that reduces the loss of quality of the recommendations over time. MF is very popular because it gives good scalability at the time of recommending while allowing remarkable prediction accuracy. However, one drawback of MF is that once its model has been generated, it delivers recommendations based on a snapshot of the incoming ratings frozen at the beginning of its generation. To take into account the new ratings, the model has to be computed periodically. The proposed solution to this problem improves the scalability of MF by reducing the frequency of model recomputations. Chapter "Mining (Soft-) Skypatterns Using Constraint Programming" introduces a softness in the skypattern mining problem. Skypatterns enable to express a user-preference point of view w.r.t. a dominance

relation. First, the authors show how softness can provide valuable patterns that would be missed otherwise. Then, thanks to CP, they propose a generic and efficient method to mine (soft-)skypatterns. Finally, the relevance and the effectiveness of the proposed approach through an experimental study is shown.

Nantes — Fabrice Guillet
Bordeaux — Bruno Pinaud
Tours — Gilles Venturini
Lyon — Djamel Abdelkader Zighed
June 2015

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, criticism, and constructive remarks have significantly contributed to the high quality of the selected papers.

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

Nantes — Fabrice Guillet
Bordeaux — Bruno Pinaud
Tours — Gilles Venturini
Lyon — Djamel Abdelkader Zighed
June 2015

Review Committee

All published chapters have been reviewed by two or three referees and at least one non-french speaking referee (two for most papers).

- Fionn Murtagh (Royal Holloway, University of London, UK)
- Luiz Augusto Pizzato (University of Sydney, Australia)
- Sadok Ben Yahia (University of Tunis, Tunisia)
- Francisco de A.T. De Carvalho (Universidade Federal de Pernambuco, Brazil)
- Gilles Falquet (University of Geneva, Switzerland)
- Marc Gelgon (Polytech'Nantes, France)
- Antonio Irpino (Second University of Naples, Italy)
- Lorenza Saitta (University of Torino, Italy)
- Ansaf Salleb-Aouissi (Columbia University, USA)
- Stefan Trausan-Matu (University of Bucharest, Romania)
- Rosanna Verde (University of Naples 2, Italy)
- George Vouros (University of Piraeus, Greece)
- Jef Wijsen (University of Mons-Hainaut, Belgium)

Associated Reviewers

Marc Boulé, Bruno Cremilleux, Sylvie Gibet, Hubert Naacke, Clémentine Nebut, Fabrice Rossi, Cédric Wemmert

Contents

Editors and Contributors

About the Editors

Fabrice Guillet is a CS professor at Polytech'Nantes, the graduate engineering school of University of Nantes, and a member of the "KnOwledge and Decision" team (COD) of the LINA laboratory. He received a Ph.D. degree in CS in 1995 from the "École Nationale Supérieure des Télécommunications de Bretagne", and his Habilitation (HdR) in 2006 from Nantes University. He is a co-founder of the International French-speaking "Extraction et Gestion des Connaissances (EGC)" society. His research interests include knowledge quality and knowledge visualization in the frameworks of Data Mining and Knowledge Management. He has recently co-edited two refereed books of chapter entitled "Quality Measures in Data Mining" and "Statistical Implicative Analysis—Theory and Applications" published by Springer in 2007 and 2008.

Bruno Pinaud received the Ph.D. degree in Computer Science in 2006 from the University of Nantes. He is currently assistant professor at the University of Bordeaux in the Computer Science Department since September 2008. His current research interests are visual data mining, graph rewriting systems, graph visualization, and experimental evaluation in Human Computer Interaction (HCI).

Gilles Venturini is a CS Professor at François Rabelais University of Tours (France). His main research interests concern visual data mining, virtual reality, 3D acquisition, biomimetic algorithms (genetic algorithms, artificial ants). He is co-editor in chief of the French New IT Journal (Revue des Nouvelles Technologies de l'Information) and was recently elected as President of the EGC society.

Djamel Abdelkader Zighed is a CS Professor at the Lyon 2 University. He is the head of the Human Sciences Institute and he was Director of the ERIC Laboratory (University of Lyon). He is also the coordinator of the Erasmus Mundus Master Program on Data Mining and Knowledge Management (DMKM). He is also member of various international and national program committees.

Contributors

Talel Abdessalem is currently a Professor at Telecom ParisTech, holder of the Big Data and Market Insights Chair and head of the IC2 group. His research interests are in large scale data management and mining, recommender systems, web information extraction, large graphs, and social networks analysis.

Florence Le Ber holds a Ph.D. in computer science from Lorraine University (1993). She is currently director of the Research Department at the French National School for Water and Environmental Engineering (ENGEES, Strasbourg University) and a research member of the ICube laboratory, in the BFO team. Her main research interest is on extracting and modeling knowledge from spatio-temporal databases, and on spatio-temporal reasoning, in the framework of agricultural or environmental applications.

Patrice Boizumault is currently professor of computer science at the University of Caen. His research interests are Logic Programming, Metaheuristics, Constraint Programming, and Constraint Satisfaction Problems for Data Mining. Recent works address the resolution of over-constrained problems and constraint relaxation for global constraints as well as for soft constraints for pattern (sets) discovery. Applications concern workforce management (in particular Nurse Rostering Problems) and Chemoinformatics.

Marc Boullé was born in 1965 and graduated from Ecole Polytechnique (France) in 1987 and Sup Telecom Paris in 1989. Currently, he is a Senior Researcher in the data mining research group of Orange Labs. His main research interests include statistical data analysis, data mining, especially data preparation and modeling for large databases. He developed regularized methods for feature preprocessing, feature selection and construction, correlation analysis, model averaging of selective naive Bayes classifiers and regressors.

Pierre Chauveau-Thoumelin is a Ph.D. student in STL UMR8163 Lab at Université Lille 3. He is currently working on linguistic constructions coined with "genre", "type" and "style". His Master internship was dedicated to the study of subjectivity in medical discourse and the difference between specialized and non-specialized languages.

Bruno Crémilleux is currently professor of computer science at the University of Caen. His research interests are in data mining and knowledge discovery in databases with a focus on pattern discovery: pattern (sets) discovery, Constraint Satisfaction Problems and data mining, Natural Language Processing and data mining, preference queries (e.g., skypatterns), unsupervised and supervised methods from several pattern languages (e.g., sequences, graphs). This research work benefits from close collaborations addressing applications in the fields of Chemoinformatics, Biomedical Text Analysis, and Bioinformatics.

Xavier Dolques obtained his Ph.D. in 2010 and is currently a postdoc of computer science at the National School for Water and Environmental Engineering (ENGEES, Strasbourg University) and at the ICube laboratory, in the BFO team. His work is funded by the national agency of research through the project Fresqueau. His main research interest is on Formal Concept Analysis and Relational Concept Analysis applied to software engineering problems, especially in the model driven area, and to data mining problems.

Loïc Dumonet prepared his Master internship in STL UMR8163 Lab at Université Lille 3, France. He worked on the evolution and visualization of emotions and subjectivity in medical discourse. The study was done on contrastive specialized and non-specialized medical corpora.

Mohamed K. El Mahrsi, born in 1984, graduated as a computer engineer from the National School of Computer Science (Tunisia) in 2008 and received a Ph.D. in computer science from Télécom ParisTech (France) in 2013. He currently works as a postdoctoral researcher at the French Institute of Science and Technology for Transport, Development and Networks (France). His main research interests include data mining, exploratory data analysis, data visualization, and their application on mobility data.

Natalia Grabar is a CR1 CNRS researcher in STL UMR8163 Lab at Université Lille 3, France. She obtained her Ph.D. degree from the Université Paris 6 in 2004 in the field of Medical Informatics. Her main area of research is NLP applied to specialized languages with a special interest in terminologies, semantic resources, and information reliability.

Modou Gueye holds a Ph.D. degree from Telecom ParisTech, a leading French engineering school specialized in computer science. He mainly works in designing scalable, but accurate too, recommender systems.

Romain Guigourés was born in 1987 and received a Ph.D. in applied mathematics from the Paris-I Panthéon-Sorbonne University in 2013. He worked for the data mining research group of Orange Labs from 2010 to 2013 and is currently data scientist in the data intelligence department at Zalando. His main research interests include data mining, coclustering, and exploratory data analysis.

Marianne Huchard obtained her Ph.D. in 1992 and is currently Professor of computer science at University of Montpellier since 2004. She is currently Deputy Director of the LIRMM laboratory (Laboratoire d'Informatique, de Robotique, de Micro-Electronique in Montpellier) and she recently served as general chair of the join conferences ECMFA-ECOOP-ECSA 2013 in Montpellier. Her main areas of interest are Formal Concept Analysis (Galois lattice/Concept lattices), in its theoretical aspects as well as in its applications mainly to the domain of software engineering (Model-Driven Engineering, Component-based software engineering and Service-Oriented Architectures).

Alban Lepailleur is currently associate professor in molecular modeling at the University of Caen. He has expertise in applied chemoinformatics and 3D-QSAR methods for the discovery of new ligands with therapeutic potentials. Involved in structure-based and ligand-based virtual screening campaigns, specially in the development of a virtual screening analysis toolkit in collaboration with Discngine. He works in validation in silico methodologies as alternatives to animal experiments for the evaluation of (eco)-toxicity of substances (QSARs models, Expert systems, Read-across).

Samir Loudni is currently associate professor in computer science at the University of Caen. He has expertise in constraint optimization and design of hybrid approaches for solving combinatorial optimization problems. Recent works address the design of generic approaches for data mining using Constraint Programming. He is treasurer of the Executive committee of the French Association for Constraint Programming (AFPC).

Hubert Naacke is an Assistant Professor at University Paris 6. He is the author and co-author of several publications in international conferences and journals, national conferences, and book chapters. A part of his research interests is in large-scale systems.

Clémentine Nebut obtained her Ph.D. in 2004 from Rennes University and is currently Assistant Professor at University of Montpellier since 2006 and member of the MAREL team at the LIRMM laboratory. She is currently co-head of the AIGLE Master formation (software engineering and web speciality) at the University. Her main areas of interest are Model Driven Engineering and Models Refactoring, using artificial intelligence approaches such as Formal Concept Analysis.

Willy Ugarte is currently a Temporary Lecturer and Teaching Assistant at the University of Caen. His research lies on the border between Constraint Programming and Data Mining, with a focus on pattern (sets) discovery with soft constraints (e.g., soft threshold constraints) and optimization (e.g., skypatterns). He has an extensive experience on real application domains such as the discovery of toxicophores and the discovery of mutagenic components in the fields of Chemoinformatics.