Studies in Computational Intelligence

Volume 834

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

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 world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are submitted to indexing to Web of Science, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink.

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

Bruno Pinaud · Fabrice Guillet ·
Fabien Gandon · Christine Largeron
Editors

Advances in Knowledge Discovery and Management

Volume 8

Springer

Editors
Bruno Pinaud
University of Bordeaux
Bordeaux, France

Fabien Gandon
University of Côte d'Azur
Inria
Sophia Antipolis, France

Fabrice Guillet
Polytechnic School of the University of Nantes
University of Nantes
Nantes, France

Christine Largeron
CNRS, Hubert Curien Laboratory
University of Lyon, Université Jean Monnet Saint-Etienne
Saint-Étienne, France

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-030-18128-4 ISBN 978-3-030-18129-1 (eBook)
https://doi.org/10.1007/978-3-030-18129-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Review Committee

All published chapters have been reviewed by two or three referees and at least one not native French speaker referee.

- Valerio Basile (University of Turin, Italy)
- Paula Brito (University of Porto, Portugal)
- Francisco de A. T. De Carvalho (Univ. Federal de Pernambuco, Brazil)
- Carlos Ferreira (LIAAD INESC Porto LA, Portugal)
- Antonio Irpino (Second University of Naples, Italy)
- Daniel Lemire (LICEF Research Center, University of Québec, Canada)
- Paulo Maio (GECAD—Knowledge Engineering and Decision Support Research Group, Portugal)
- Fionn Murtagh (University of Huddersfield, UK)
- Dan Simovici (University of Massachusetts Boston, USA)
- Stefan Trausan-Matu (Univ. Politehnica of Bucharest, Romania)
- Jef Wijsen (Univ. of Mons-Hainaut, Belgium)

Associated Reviewers

Nicolas Béchet, Agnès Braud, Bruno Cremilleux, Roland Kotto Kombi, Florence Le Ber, Stéphane Loiseau, Jerry Lonlac, Sofian Maabout, Arnaud Martin, François Meunier, Suzanne Pinson, Gildas Tagny Ngompe.

Preface

At a time when there is much talk about artificial intelligence and data, the challenges of knowledge discovery, representation and management appear as major topics for both research and development. Indeed, the processing and integration of data from various sources constantly raises new needs in terms of methods and tools for acquiring data, classifying them, integrating them, representing them, storing them, indexing them, processing them, visualizing them, interacting with them, and, in fine, to transform them into useful knowledge.

This book is a collection of height novel scientific contributions addressing several of these challenges. These articles are extended versions of a selection of the best papers that were initially presented at the French-speaking conferences EGC'2017 and EGC'2018 held, respectively, in Grenoble (France, January 22–27, 2017) and Paris (France, January 22–26, 2018). These extended versions have been accepted after an additional peer-review process among papers already accepted in long format at the conference. Concerning the conference, the long papers selection was also the result of a double-blind peer-review process among the hundreds of papers initially submitted to each edition of the conference (acceptance rate for long papers is about 25%). These conferences were the 17th and 18th editions 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 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.

The height chapters resulting from this selection process have been grouped into four sections, each one containing two related chapters:

Chapters “Model Based Co-clustering of Mixed Numerical and Binary Data” and “Co-clustering Based Exploratory Analysis of Mixed-Type Data Tables” are dedicated to co-clustering that aims at simultaneously clustering the set of instances and the set of variables of a data table. In chapter "Model Based Co-clustering of Mixed Numerical and Binary Data", an extended version of the Latent Block Model is introduced for co-clustering mixed data containing numerical and binary variables by combining Gaussian mixture models with Bernoulli mixture models. To solve the same task, another approach is presented in chapter "Co-clustering Based Exploratory Analysis of Mixed-Type Data Tables". This alternative solution is based on a discretization of all variables into a given number of bins, followed by a classical co-clustering to estimate the joint density between the set of instances and the set of variables. This strategy enables to detect the underlying correlations between the variables while performing a clustering of the instances.

Chapters "Automatically Selecting Complementary Vector Representations for Semantic Textual Similarity" and "Detecting Sections and Entities in Court Decisions Using HMM and CRF Graphical Models" study textual data. Chapter "Automatically Selecting Complementary Vector Representations for Semantic Textual Similarity" presents a method that aims to combine different sentence-based vector representations in order to improve the computation of semantic similarity values. The method's main difficulty lies in the selection of the most complementary representations. The proposed optimization method is assessed on the dataset of the 2016 SemEval evaluation campaign. Chapter "Detecting Sections and Entities in Court Decisions Using HMM and CRF Graphical Models" describes the problems of document sectioning and entity detection in the case of court decisions. The authors suggest a two-stage architecture using some handcrafted features in the graphical probabilistic models HMM and CRF. The impact of some designing aspects is also discussed through different experimental results.

Chapters "Discriminant Chronicle Mining" and "A Semantic-Based Approach for Landscape Identification" deal with patterns extraction or identification. In chapter "Discriminant Chronicle Mining", authors focus on temporal patterns called chronicles which are considered and extracted from labeled sequences of timestamped events by combining pattern mining and machine learning. The proposed method is evaluated on a real case study, which consists in analyzing care pathways to answer a pharmaco-epidemiological question. Chapter "A Semantic-Based Approach for Landscape Identification" focuses on landscape automatic identification in a satellite image. The study of landscapes natural and artificial as well, and their evolution over time is one approach of addressing major social, economic, and environmental challenges. Acquiring new knowledge is very demanding especially within the context of satellite images. As a consequence, the authors propose a mixed-knowledge based strategy to both successfully extract appropriate landscapes and organize knowledge through ontologies for further dissemination.

Finally, both chapter "Measuring the Expertise of Workers for Crowdsourcing Applications" and chapter "Trust Assessment for the Security of Information Systems" address human and social dimensions through indicator definitions. In chapter "Measuring the Expertise of Workers for Crowdsourcing Applications", a metric, based on the theory of belief functions, is introduced to evaluate the expertise of workers for crowdsourcing applications. Chapter "Trust Assessment for the Security of Information Systems" presents trust measures for security in information systems. Indeed, modern information systems are supplied by various sensors and communicating devices. However, security concerns about these devices raise the question to what extent one can have trust in their pieces of information as well as in the whole system. For this purpose, new trust measures are introduced in this paper and then tested using simulations conducted in the framework of ship navigation system.

Bordeaux, France	Bruno Pinaud
Nantes, France	Fabrice Guillet
Sophia Antipolis, France	Fabien Gandon
Saint-Étienne, France	Christine Largeron

Contents

Part I Clustering

Part II Textual Data

Part III Spatial and Temporal Dimension

Part IV Human and Social Dimension

Editors and Contributors

About the Editors

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 HCI (Human–Computer Interaction).

Fabrice Guillet is a Full Professor in CS at Polytech'Nantes, the graduate engineering school of University of Nantes, France, and a member of the "Data User Knowledge" team (DUKe) of the LS2N 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 and President 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 Science and Knowledge Management. He has 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.

Fabien Gandon is a Research Director in Informatics and Computer Science at Inria and Leader of the joint Wimmics team at the Sophia Antipolis Research Center (UCA, Inria, CNRS, I3S). His professional interests include Web, Semantic Web, Social Web, Ontologies, Knowledge Engineering and Modelling, Mobility, Privacy, Context-Awareness, Semantic Social Network/Semantic Analysis of Social Network, Intraweb, and Distributed Artificial Intelligence. He is representative of Inria at the World-Wide Web Consortium (W3C), Director of the joint research laboratory QWANT-Inria, responsible for the research convention between the Ministry of Culture and Inria, and Vice-head of research for Inria Sophia Antipolis—Méditerranée.

Christine Largeron is a Full Professor in Computer Science. She received a Ph.D. in Computer Science from Claude Bernard University (Lyon, France) in 1991. She is Professor at Jean Monnet University (France) since 2006 and, she is the Head of the Data Mining and Information Retrieval group of the Hubert Curien Laboratory. Her research interests focus on machine learning, data mining, information retrieval, text mining, social mining, and network analysis. She has published more than 100 papers in refereed international conferences and journals and she regularly acts as PC member of several conferences and as co-organizer on a number of international workshops and conferences. She was PC chair of EGC'2018.

Contributors

Ouassim Ait-Elhara is the Lead Data Scientist at Octopeek, a French company specialized in Big Data and Data Science. He has a Ph.D. degree in Artificial Intelligence from the University of Paris-Saclay. His research interests include the application of Machine Learning algorithms to real-world problems.

Aichetou Bouchareb is a Ph.D. student at Université Paris 1 Panthéon-Sorbonne and research and development engineer with the "Profiling and Data-Mining" research group of Orange Labs under the supervision of Mr. Boullé, Mr. Rossi, and Mr. Clérot. Her main interests are machine learning and data mining, especially visualization and modeling of data sets for knowledge extraction.

Marc Boullé is currently 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, and model averaging of selective naive Bayes classifiers and regressors.

Fabrice Clérot is Senior Researcher in data mining and Head of the "Profiling and Data-Mining" research group of Orange Labs.

Gouenou Coatrieux is Full Professor within the Département Image et Traitement de l'Information (ITI) of IMT—Atlantique in France. He conducts his research activities in the Laboratoire de Traitement de l'Information Médicale (LaTIM Inserm U1101). Since several years, he is working on the protection of medical multimedia data, in particular, of images, with an approach essentially based on watermarking.

Benjamin Costé is Ph.D. student at Chair of Naval Cyber Defense located in Brest, France. His research is focused on new detection methods based on trust assessment of information systems.

Yann Dauxais is a postdoctoral fellow in the KU Leuven DTAI group and Ph.D. in Computer Science from the University of Rennes. His fields of expertise are pattern mining and machine learning. During his Ph.D., he designed a new algorithm to extract discriminant temporal patterns from sequential data, and applied this theoretical outcome to the field of care pathway analytics.

Eric Delaître is a Researcher at the French National Research Institute for Sustainable Development (IRD). He is a specialist in remote sensing (optics) and GIS, applied on the terrestrial surfaces of tropical and equatorial regions. He develops processing algorithms to extract information from Earth observation satellites, to improve the management of natural resources in different areas of the world (Amazonia, North Africa, Madagascar).

Laurent Demagistri is an engineer at the French Research Institute for Sustainable Development (IRD) in scientific computing and geomatics, specialized in Remote Sensing and Spatial Data Infrastructures. His main activities consist in the development of workflows for the processing of satellite images and value-added database management products. His areas of expertise cover data registration and calibration, bio-geophysical parameters, and object or process detection.

Bich-Liên Doan is a Professor in Computer Science at CentraleSupélec and a member of the LRI (University Paris-Saclay and CNRS). She is specialized in information retrieval, and her current research is related to contextual and more particularly personalized information retrieval and recommender systems. She is interested in new models from NLP, deep learning, semiotics, and quantum mechanics that can help in representing information.

Jean-Christophe Dubois is Associate Professor at University of Rennes 1. His Ph. D. degree, received in 1998 from the University of Nancy 1, in the LORIA laboratory, concerned human–machine dialogue and picture archive interrogation in natural speaking. He worked at University of Angers for 4 years before joining the IUT of Lannion in 2002. His current research activity deals with belief functions, data fusion, and digital accessibility.

Laetitia Gros joined the research and development entity of Orange (Orange Labs) in 2003 as a research engineer on perception and quality of experience of audio-visual technologies and service, after a Master of Science degree of Acoustic, Signal Processing, and Informatics Applied to Music, at IRCAM (Institute of Research and Coordination of Acoustic/Music) and a Ph.D. degree on psychoacoustics. Her work mainly concerns methodologies to understand and assess the user experience related to audio and audio-visual technologies/services.

David Gross-Amblard is a Professor at Rennes University. He is the co-head of the IRISA/DRUID team and develops original research in the field of crowdsourcing, including declarative management of participants and complex task workflows. He is the President of the Advisory Board of the French BDA association (Advanced database association) and the coordinator of the Headwork ANR project.

Thomas Guyet is an Assistant Professor at AGROCAMPUS-OUEST and he is doing his research in the Inria/IRISA LACODAM Team. The research interests of Thomas Guyet range from cognitive foundations to the practical application of discovering spatial and temporal patterns in semantically complex datasets. He develops research in large range of artificial intelligence domains including (sequential) pattern mining, knowledge discovery, and declarative programming (answer set programming).

André Happe is a Research Engineer at the Brest University Hospital and is strongly involved in the development of a platform for digital pharmaco-epidemiology funded by the national agency of drug safety (PEPS project). He is a specialist in medical informatics and data management for health.

Sébastien Harispe is Ph.D. in Computer Science, Researcher in Artificial Intelligence (AI), and Associate Professor at Institut Mines Télécom. He is a problem-solving enthusiast studying both theoretical and practical aspects of complex AI problems, e.g., Approximate Reasoning, Natural Language Understanding. In addition to his theoretical work, he is also actively collaborating with companies and startups for solving real-world problems using cutting edge research technologies and theoretical tools.

Julien Hay is a Ph.D. student at CentraleSupéc and a member of the LRI (University of Paris-Saclay and CNRS). He obtained a master's degree in software development and AI. His work focuses on the application of NLP and machine learning techniques on text data from social networks. He analyzes what users read and write in order to enhance user profiling in recommender systems.

Mouloud Kharoune is Associate Professor at University of Rennes 1. He obtained his Ph.D. degree in 1988 and since has been assigned to the IUT Lannion. He worked in the field of human/machine dialogue. Currently, he is interested in the theory of belief functions and their use in the domain of social networks and crowdsourcing.

Anne–Elisabeth Laques is a Researcher at the French Research Institute for Sustainable Development (IRD). She is a geographer and landscape specialist. She is particularly interested in measuring the spatial footprints of human activity in tropical environments. In this context, she produces spatialized indicators on socio-environmental dynamics, particularly from satellite images.

Yolande Le Gall is Associate Professor at University of Rennes 1. She received a Ph.D. degree in 1994, on speech processing and training process from the University of Nancy 1, in the LORIA laboratory. Her current research focuses on belief functions for digital accessibility and crowdsourcing.

Arnaud Martin was born in France, in 1974. He received the Master and Ph.D. degrees from the University of Rennes 1, Rennes, France, respectively, in 1998 and 2001, and his Habilitation à Diriger des Recherches from University of Occidental Brittany, in 2009. Since 2010, he has been a Full Professor at the University of Rennes 1. His research interests mainly focus on the theory of belief functions and artificial intelligence for social networks and crowdsourcing.

Zoltan Miklos is an Associate Professor at University of Rennes 1. Before taking this position in 2012, he used to work as a postdoctoral researcher at EPFL. He earned a D.Phil. degree from University of Oxford in 2008. His research interests include questions in data management and in artificial intelligence. Zoltan is a senior member of the ACM.

Jacky Montmain is Full Professor at the École des Mines d'Alès, Nîmes, France. His research area is related to knowledge and preference representation in decision-making. He was a Research Engineer and a Senior Expert at the French Atomic Energy Commission from 1991 to 2005 where his work was focused on decisions in model-based diagnosis and industrial supervision issues. His current points of interests include the application of artificial intelligence and operations research techniques to knowledge representation and multi-criteria fuzzy approaches to decision-making.

Isabelle Mougenot is an Associate Professor in Computer Science at the University of Montpellier, France. She is currently involved in the ESPACE-DEV research unit that is conducting some interdisciplinary projects on joint aspects of computer science, and applied mathematics for environmental studies. Specifically, her research interests include distributed database systems, metadata modeling, ontologies, and knowledge integration.

Philippe Muller is an Associate Professor of Computer Science at the University of Toulouse. His research interests are mainly in Natural Language Processing, and include Discourse and Dialogue analysis, and Computational Semantics.

Stéphanse Mussard is Full Professor of economics at the University of Nîmes and CHROME research fellow. He is specialized in statistics, econometrics, and machine learning.

Gildas Tagny Ngompé is a Ph.D. student at IMT Mines Alès, working on the design and application of natural language processing methods to extract information from a corpus of French court decisions in order to build a legal knowledge base.

Hosna Ouni had her first engineering diploma at Tunisia Polytechnical school in 2016. She had an internship at University of Rennes where she worked mainly in crowdsourcing and belief functions. She had second a double diploma: master and MBA in international management at ESCE international school in Paris from 2016 to 2018; in parallel, she worked in software asset management at Société Générale. She is currently a consultant in Cash Management at BNPParibas.

Fabrice Popineau is a Professor of Computer Science at CentraleSupélec and a full member of the "Laboratoire de Recherche en Informatique" (UMR8623 of the Paris-Saclay University and the CNRS). For the past 15 years or so, his research has focused on the contributions of artificial intelligence to the personalization of the user experience on web platforms. He is particularly interested in personalized recommendation in the context of social networks and also for online educational platforms.

Cyril Ray is Associate Professor in Computer Science at Naval Academy Research Institute (IRENav) in France. His current research is oriented to the modeling and design of location-based services applied to human mobility, maritime, and urban transportation systems. This work includes integration of location-acquisition technologies and real-time tracking of moving objects, modeling of heterogeneous and large spatiotemporal datasets, movement data processing, modeling of context-aware systems, and traffic simulation and prediction.

Fabrice Rossi is Professor of applied mathematics at University Paris 1 Panthéon Sorbonne. He is a member of the SAMM laboratory. He leads a research team on statistical learning, statistics and networks, with nine permanent researchers and seven Ph.D. students. He specializes in exploratory data analysis with a special interest in graph data, change detection, and visual data exploration. More generally, his research covers numerous important themes of machine learning including large-scale data processing, feature selection, learning theory, and clustering. He works frequently with researchers from other fields, especially from the humanities, including archaeology, history, and sociology. In 2017, he was guest editor of a special issue on humanities and statistics of the main French statistics journal. He has (co)-authored more than 150 articles in journals and conference proceedings.

Emmanuel Roux is a Researcher at the French Research Institute for Sustainable Development (IRD), in the ESPACE-DEV Unit. His research area is Data Science, favoring approaches related to exploratory data analysis and automatic learning to obtain data representations, rule sets, and models with high explanatory power.

Anne Toulet is a Researcher in Computer Science at LIRMM, University of Montpellier, France. Her research topics are related to data management, ontology design, knowledge representation, ontology metadata, and semantic web. She works on several projects on areas such as agronomy, biodiversity, environment, or Earth observation.

Tim Van de Cruys is a Researcher at CNRS & IRIT, Toulouse. His research is within natural language processing, with a focus on the unsupervised extraction of semantics from text.

Guillaume Zambrano is Assistant Professor at University of Nîmes (France), and member of research unit EA7352 CHROME. His research interests include quantitative legal prediction, statistical analysis of case law, and artificial intelligence applied to the Study of Law.