

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

More information about this series at <http://www.springer.com/series/1244>

Annalisa Appice · Pedro Pereira Rodrigues
Vitor Santos Costa · Carlos Soares
João Gama · Alípio Jorge (Eds.)

Machine Learning and Knowledge Discovery in Databases

European Conference, ECML PKDD 2015
Porto, Portugal, September 7–11, 2015
Proceedings, Part I

Editors

Annalisa Appice
University of Bari Aldo Moro
Bari
Italy

Pedro Pereira Rodrigues
University of Porto
Porto
Portugal

Vitor Santos Costa
University of Porto - CRACS/INESC TEC
Porto
Portugal

Carlos Soares
University of Porto - INESC TEC
Porto
Portugal

João Gama
University of Porto - INESC TEC
Porto
Portugal

Alípio Jorge
University of Porto - INESC TEC
Porto
Portugal

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-23527-1 ISBN 978-3-319-23528-8 (eBook)
DOI 10.1007/978-3-319-23528-8

Library of Congress Control Number: 2015947118

LNCS Sublibrary: SL7 – Artificial Intelligence

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, 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.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

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

Preface

We are delighted to introduce the proceedings of the 2015 edition of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, or ECML PKDD for short. This conference stems from the former ECML and PKDD conferences, the two premier European conferences on, respectively, Machine Learning and Knowledge Discovery in Databases. Originally independent events, the two conferences were organized jointly for the first time in 2001. The synergy between the two led to increasing integration, and eventually the two merged in 2008. Today, ECML PKDD is a world-wide leading scientific event that aims at exploiting the synergies between Machine Learning and Data Mining, focusing on the development and application of methods and tools capable of solving real-life problems.

ECML PKDD 2015 was held in Porto, Portugal, during September 7–11. This was the third time Porto hosted the major European Machine Learning event. In 1991, Porto was host to the fifth EWSL, the precursor of ECML. More recently, in 2005, Porto was host to a very successful ECML PKDD. We were honored that the community chose to again have ECML PKDD 2015 in Porto, just ten years later. The 2015 ECML PKDD was co-located with “Intelligent System Applications to Power Systems”, ISAP 2015, a well-established forum for scientific and technical discussion, aiming at fostering the widespread application of intelligent tools and techniques to the power system network and business. Moreover, it was collocated, for the first time, with the Summer School on “Data Sciences for Big Data.”

ECML PKDD traditionally combines the research-oriented extensive program of the scientific and journal tracks, which aim at being a forum for high quality, novel research in Machine Learning and Data Mining, with the more focused programs of the demo track, dedicated to presenting real systems to the community, the PhD track, which supports young researchers, and the nectar track, dedicated to bringing relevant work to the community. The program further includes an industrial track, which brings together participants from academia, industry, government, and non-governmental organizations in a venue that highlights practical and real-world studies of machine learning, knowledge discovery, and data mining. The industrial track of ECML PKDD 2015 has a separate Program Committee and separate proceedings volume. Moreover, the conference program included a doctoral consortium, three discovery challenges, and various workshops and tutorials.

The research program included five plenary talks by invited speakers, namely, Hendrik Blockeel (University of Leuven and Leiden University), Pedro Domingos (University of Washington), Jure Leskovec (Stanford University), Nataša Milić-Frayling (Microsoft Research), and Dino Pedreschi (Università di Pisa), as well as one ISAP +ECML PKDD joint plenary talk by Chen-Ching Liu (Washington State University). Three invited speakers contributed to the industrial track: Andreas Antrup (Zalando and

University of Edinburgh), Wei Fan (Baidu Big Data Lab), and Hang Li (Noah's Ark Lab, Huawei Technologies).

Three discovery challenges were announced this year. They focused on "MoRe-BikeS: Model Reuse with Bike rental Station data," "On Learning from Taxi GPS Traces," and "Activity Detection Based on Non-GPS Mobility Data," respectively.

Twelve workshops were held, providing an opportunity to discuss current topics in a small and interactive atmosphere: "MetaSel - Meta-learning and Algorithm Selection," "Parallel and Distributed Computing for Knowledge Discovery in Databases," "Interactions between Data Mining and Natural Language Processing," "New Frontiers in Mining Complex Patterns," "Mining Ubiquitous and Social Environments," "Advanced Analytics and Learning on Temporal Data," "Learning Models over Multiple Contexts," "Linked Data for Knowledge Discovery," "Sports Analytics," "BigTargets: Big Multi-target Prediction," "DARE: Data Analytics for Renewable Energy Integration," and "Machine Learning in Life Sciences."

Ten tutorials were included in the conference program, providing a comprehensive introduction to core techniques and areas of interest for the scientific community: "Similarity and Distance Metric Learning with Applications to Computer Vision," "Scalable Learning of Graphical Models," "Meta-learning and Algorithm Selection," "Machine Reading the Web - Beyond Named Entity Recognition and Relation Extraction," "VC-Dimension and Rademacher Averages: From Statistical Learning Theory to Sampling Algorithms," "Making Sense of (Multi-)Relational Data," "Collaborative Filtering with Binary, Positive-Only Data," "Predictive Maintenance," "Eureka! - How to Build Accurate Predictors for Real-Valued Outputs from Simple Methods," and "The Space of Online Learning Problems."

The main track received 380 paper submissions, of which 89 were accepted. Such a high volume of scientific work required a tremendous effort by the Area Chairs, Program Committee members, and many additional reviewers. We managed to collect three highly qualified independent reviews per paper and one additional overall input from one of the Area Chairs. Papers were evaluated on the basis of significance of contribution, novelty, technical quality, scientific, and technological impact, clarity, repeatability, and scholarship. The industrial, demo, and nectar tracks were equally successful, attracting 42, 32, and 29 paper submissions, respectively.

For the third time, the conference used a double submission model: next to the regular conference tracks, papers submitted to the Springer journals Machine Learning (MACH) and Data Mining and Knowledge Discovery (DAMI) were considered for presentation at the conference. These papers were submitted to the ECML PKDD 2015 special issue of the respective journals, and underwent the normal editorial process of these journals. Those papers accepted for one of these journals were assigned a presentation slot at the ECML PKDD 2015 conference. A total of 191 original manuscripts were submitted to the journal track during this year. Some of these papers are still being refereed. Of the fully refereed papers, 10 were accepted in DAMI and 15 in MACH, together with 4+4 papers from last year's call, which were also scheduled for presentation at this conference. Overall, this resulted in a number of 613 submissions (to the scientific track, industrial track and journal track), of which 126 were selected for presentation at the conference, making an overall acceptance rate of about 21%.

Part I and Part II of the proceedings of the ECML PKDD 2015 conference contain the full papers of the contributions presented in the scientific track, the abstracts of the scientific plenary talks, and the abstract of the ISAP+ECML PKDD joint plenary talk. Part III of the proceedings of the ECML PKDD 2015 conference contains the full papers of the contributions presented in the industrial track, short papers describing the demonstrations, the nectar papers, and the abstracts of the industrial plenary talks.

The scientific track program results from continuous collaboration between the scientific tracks and the general chairs. Throughout we had the unfaltering support of the Local Chairs, Carlos Ferreira, Rita Ribeiro, and João Moreira, who managed this event in a thoroughly competent and professional way. We thank the Social Media Chairs, Dunja Mladeníć and Márcia Oliveira, for tweeting the new face of ECML PKDD, and the Publicity Chairs, Ricardo Campos and Carlos Ferreira, for their excellent work in spreading the news. The beautiful design and quick response time of the web site is due to the work of our Web Chairs, Sylwia Bugla, Rita Ribeiro, and João Rodrigues. The beautiful image on all the conference materials is based on the logo designed by Joana Amaral e João Cravo, inspired by Porto landmarks. It has been a pleasure to collaborate with the Journal, Industrial, Demo, Nectar, and PhD Track Chairs. ECML PKDD would not be complete if not for the efforts of the Tutorial Chairs, Fazel Famili, Mykola Pechenizkiy, and Nikolaj Tatti, the Workshop Chairs, Stan Matwin, Bernhard Pfahringer, and Luís Torgo, and the Discovery Challenge Chairs, Michel Ferreira, Hillol Kargupta, Luís Moreira-Matias, and João Moreira. We thank the Awards Committee Chairs, Pavel Brazdil, Sašo Džerosky, Hiroshi Motoda, and Michèle Sebag, for their hard work in selecting papers for awards. A special meta thanks to Pavel: ECML PKDD at Porto is only possible thanks to you. We gratefully acknowledge the work of the Sponsorship Chairs, Albert Bifet and André Carvalho, for their key work. Special thanks go to the Proceedings Chairs, Michelangelo Ceci and Paulo Cortez, for the difficult task of putting these proceedings together. We appreciate the support of Artur Aiguzhinov, Catarina Félix Oliveira, and Mohammad Nozari (U. Porto) for helping to check this front matter. We thank the ECML PKDD Steering Committee for kindly sharing their experience, and particularly the General Steering Committee Chair, Fosca Giannotti. The quality of ECML PKDD is only possible due to the tremendous efforts of the Program Committee; our sincere thanks for all the great work in improving the quality of these proceedings. Throughout, we relied on the exceptional quality of the Area Chairs. Our most sincere thanks for their support, with a special thanks to the members who contributed in difficult personal situations, and to Paulo Azevedo for stepping in when the need was there. Last but not least, we would like to sincerely thank all the authors who submitted their work to the conference.

July 2015

Annalisa Appice
Pedro Pereira Rodrigues
Vítor Santos Costa
Carlos Soares
João Gama
Alípio Jorge

Organization

ECML/PKDD 2015 Organization

Conference Co-chairs

João Gama	University of Porto, INESC TEC, Portugal
Alipío Jorge	University of Porto, INESC TEC, Portugal

Program Co-chairs

Annalisa Appice	University of Bari Aldo Moro, Italy
Pedro Pereira Rodrigues	University of Porto, CINTESIS, INESC TEC, Portugal
Vtor SantosCosta	University of Porto, INESC TEC, Portugal
Carlos Soares	University of Porto, INESC TEC, Portugal

Journal Track Chairs

Concha Bielza	Technical University of Madrid, Spain
João Gama	University of Porto, INESC TEC, Portugal
Alipío Jorge	University of Porto, INESC TEC, Portugal
Indrė Žliobaitė	Aalto University and University of Helsinki, Finland

Industrial Track Chairs

Albert Bifet	Huawei Noah's Ark Lab, China
Michael May	Siemens, Germany
Bianca Zadrozny	IBM Research, Brazil

Local Organization Chairs

Carlos Ferreira	Oporto Polytechnic Institute, INESC TEC, Portugal
João Moreira	University of Porto, INESC TEC, Portugal
Rita Ribeiro	University of Porto, INESC TEC, Portugal

Tutorial Chairs

Fazel Famili	CNRC, France
Mykola Pechenizkiy	TU Eindhoven, The Netherland
Nikolaj Tatti	Aalto University, Finland

Workshop Chairs

Stan Matwin	Dalhousie University, NS, Canada
Bernhard Pfahringer	University of Waikato, New Zealand
Luís Torgo	University of Porto, INESC TEC, Portugal

Awards Committee Chairs

Pavel Brazdil	INESC TEC, Portugal
Sašo Džeroski	Jožef Stefan Institute, Slovenia
Hiroshi Motoda	Osaka University, Japan
Michèle Sebag	Université Paris Sud, France

Nectar Track Chairs

Ricard Gavaldà	UPC, Spain
Dino Pedreschi	Università di Pisa, Italy

Demo Track Chairs

Francesco Bonchi	Yahoo! Labs, Spain
Jaime Cardoso	University of Porto, INESC TEC, Portugal
Myra Spiliopoulou	Otto-von-Guericke University Magdeburg, Germany

PhD Chairs

Jaakko Hollmén	Aalto University, Finland
Panagiotis Papapetrou	Stockholm University, Sweden

Proceedings Chairs

Michelangelo Ceci	University of Bari, Italy
Paulo Cortez	University of Minho, Portugal

Discovery Challenge Chairs

Michel Ferreira	University of Porto, INESC TEC, Geolink, Portugal
Hillol Kargupta	Agnik, MD, USA
Luís Moreira-Matias	NEC Research Labs, Germany
João Moreira	University of Porto, INESC TEC, Portugal

Sponsorship Chairs

Albert Bifet	Huawei Noah's Ark Lab, China
André Carvlho	University of São Paulo, Brazil
Pedro Pereira Rodrigues	University of Porto, Portugal

Publicity Chairs

Ricardo Campos	Polytechnic Institute of Tomar, INESC TEC, Portugal
Carlos Ferreira	Oporto Polytechnic Institute, INESC TEC, Portugal

Social Media Chairs

Dunja Mladeníc	JSI, Slovenia
Márcia Oliveira	University of Porto, INESC TEC, Portugal

Web Chairs

Sylwia Bugla	INESC TEC, Portugal
Rita Ribeiro	University of Porto, INESC TEC, Portugal
João Rodrigues	INESC TEC, Portugal

ECML PKDD Steering Committee

Fosca Giannotti	ISTI-CNR Pisa, Italy
Michèle Sebag	Université Paris Sud, France
Francesco Bonchi	Yahoo! Research, Spain
Hendrik Blockeel	KU Leuven, Belgium and Leiden University, The Netherlands
Katharina Morik	University of Dortmund, Germany
Tobias Scheffer	University of Potsdam, Germany
Arno Siebes	Utrecht University, The Netherlands
Peter Flach	University of Bristol, UK
Tijl De Bie	University of Bristol, UK
Nello Cristianini	University of Bristol, UK
Filip Železný	Czech Technical University in Prague, Czech Republic
Siegfried Nijssen	LIACS, Leiden University, The Netherlands
Kristian Kersting	Technical University of Dortmund, Germany
Rosa Meo	Università di Torino, Italy
Toon Calders	Eindhoven University of Technology, The Netherlands
Chedy Raïssi	INRIA Nancy Grand-Est, France

Area Chairs

Paulo Azevedo	University of Minho
Michael Berthold	Universität Konstanz
Francesco Bonchi	Yahoo Labs Barcelona
Henrik Boström	University of Stockholm
Jean-François Boulicaut	Institut National des Sciences Appliquées de Lyon, LIRIS
Pavel Brazdil	University of Porto
André Carvalho	University of São Paulo
Michelangelo Ceci	Università degli Studi di Bari Aldo Moro

Jesse Davis	Katholieke Universiteit Leuven
Luc De Raedt	Katholieke Universiteit Leuven
Peter Flach	University of Bristol
Johannes Fürnkranz	TU Darmstadt
Thomas Gaertner	Fraunhofer IAIS
Bart Goethals	University of Antwerp
Andreas Hotho	University of Kassel
Eyke Hüllermeier	University of Paderborn
George Karypis	University of Minnesota
Kristian Kersting	Technical University of Dortmund
Arno Knobbe	Universiteit Leiden
Pedro Larrañaga	Technical University of Madrid
Peter Lucas	Radboud University Nijmegen
Donato Malerba	Università degli Studi di Bari Aldo Moro
Stan Matwin	Dalhousie University
Katharina Morik	TU Dortmund
Sri Ram Natarajan	Indiana University
Eugénio Oliveira	University of Porto
Mykola Pechenizkiy	Eindhoven University of Technology
Bernhard Pfahringer	University of Waikato
Michèle Sebag	CNRS
Myra Spiliopoulou	Otto-von-Guericke University Magdeburg
Jerzy Stefanowski	Poznań University of Technology
Luís Torgo	University of Porto
Stefan Wrobel	Fraunhofer IAIS, Germany
Philip Yu	University of Illinois at Chicago

Program Committee

Leman Akoglu	Narayanaswamy	Jerzy Blaszczynski
Mehmet Sabih Aksoy	Balakrishnan	Konstantinos Blekas
Mohammad Al Hasan	Elena Baralis	Mario Boley
Omar Alonso	Daniel Barbará	Gianluca Bontempi
Aijun An	Gustavo Batista	Christian Borgelt
Aris Anagnostopoulos	Christian Bauckhage	José Luís Borges
Marta Arias	Roberto Bayardo	Marc Boullé
Rubén Armañanzas	Vaishak Belle	Ulf Brefeld
Ira Assent	András Benczúr	Róbert Busa-Fekete
Martin Atzmueller	Bettina Berendt	Toon Calders
Chloé-Agathe Azencott	Michele Berlingerio	Rui Camacho
Paulo Azevedo	Indrajit Bhattacharya	Longbing Cao
Antonio Bahamonde	Marenglen Biba	Henrique Lopes Cardoso
James Bailey	Enrico Blanzieri	Francisco Casacuberta

Gladys Castillo
 Loic Cerf
 Tania Cerquitelli
 Edward Chang
 Duen Horng Chau
 Sanjay Chawla
 Keke Chen
 Ling Chen
 Weiwei Cheng
 Silvia Chiusano
 Frans Coenen
 Fabrizio Costa
 Germán Creamer
 Bruno Crémilleux
 Marco Cristo
 Tom Croonenborghs
 Boris Cule
 Tomaž Curk
 James Cussens
 Alfredo Cuzzocrea
 Claudia d'Amato
 Sašo Džeroski
 Maria Damiani
 Jeroen De Knijf
 Gerard de Melo
 Marcílio de Souto
 Kurt DeGrave
 Juan del Coz
 Krzysztof Dembczyński
 François Denis
 Anne Denton
 Mohamed Dermouche
 Christian Desrosiers
 Luigi Di Caro
 Nicola Di Mauro
 Jana Diesner
 Ilica Dimitrovski
 Ying Ding
 Stephan Doerfel
 Anne Driemel
 Chris Drummond
 Brett Drury
 Devdatt Dubhashi
 Wouter Duivesteyn
 Bob Durrant
 Inês Dutra

Tapio Elomaa
 Floriana Esposito
 Roberto Esposito
 Hadi Fanaee-T
 Nicola Fanizzi
 Elaine Faria
 Fabio Fassetti
 Hakan Ferhatosmanoglou
 Stefano Ferilli
 Carlos Ferreira
 Hugo Ferreira
 Cèsar Ferri
 George Fletcher
 Eibe Frank
 Élisabeth Fromont
 Fabio Fumarola
 Mohamed Medhat Gaber
 Fábio Gagliardi Cozman
 Patrick Gallinari
 José A. Gámez
 Jing Gao
 Byron Gao
 Paolo Garza
 Éric Gaussier
 Pierre Geurts
 Fosca Giannotti
 Christophe Giraud-Carrier
 Aris Gkoulalas-Divanis
 Marco Gori
 Pablo Granitto
 Michael Granitzer
 Maria Halkidi
 Jiawei Han
 Daniel Hernández Lobato
 José Hernández-Orallo
 Thanh Lam Hoang
 Frank Hoeppner
 Geoff Holmes
 Arjen Hommersom
 Estevam Hruschka
 Xiaohua Hu
 Minlie Huang
 Dino Ienco
 Iñaki Inza
 Frederik Janssen
 Nathalie Japkowicz

Szymon Jaroszewicz
 Ulf Johansson
 Tobias Jung
 Hachem Kadri
 Theodore Kalamboukis
 Alexandros Kalousis
 U. Kang
 Andreas Karwath
 Hisashi Kashima
 Ioannis Katakis
 Mehdi Kaytoue
 John Keane
 Latifur Khan
 Dragi Koccev
 Levente Kocsis
 Alek Kolcz
 Irena Koprincka
 Jacek Koronacki
 Nitish Korula
 Petr Kosina
 Walter Kosters
 Lars Kottorf
 Georg Krempel
 Artur Krohn-Grimberghe
 Marzena Kryszkiewicz
 Matjaž Kukar
 Meelis Kull
 Sergei Kuznetsov
 Nicolas Lachiche
 Helge Langseth
 Mark Last
 Silvio Lattanzi
 Niklas Lavesson
 Nada Lavrač
 Gianluca Lax
 Gregor Leban
 Sangkyun Lee
 Wang Lee
 Florian Lemmerich
 Philippe Lenca
 Philippe Leray
 Carson Leung
 Lei Li
 Jiuyong Li
 Juanzi Li
 Edo Liberty

Hsuan-Tien Lin	Apostolos Papadopoulos	Alan Said
Shou-de Lin	Panagiotis Papapetrou	Lorenza Saitta
Yan Liu	Ioannis Partalas	Ansaf Salieb-Aouissi
Lei Liu	Andrea Passerini	Jose S. Sanchez
Corrado Loglisci	Dino Pedreschi	Raul Santos-Rodriguez
Eneldo Loza Mencía	Nikos Pelekis	Sam Sarjant
Jose A. Lozano	Jing Peng	Claudio Sartori
Chang-Tien Lu	Yonghong Peng	Yücel Saygin
Panagis Magdalinos	Ruggero Pensa	Erik Schmidt
Giuseppe Manco	Andrea Pietracaprina	Lars Schmidt-Thieme
Yannis Manolopoulos	Fabio Pinelli	Christoph Schommer
Enrique Martinez	Marc Plantevit	Matthias Schubert
Elio Masciari	Pascal Poncelet	Marco Scutari
Florent Masegla	Lubos Popelinsky	Thomas Seidl
Luís Matias	George Potamias	Nazha Selmaoui
Oleksiy Mazhelis	Ronaldo Prati	Giovanni Semeraro
Wannes Meert	Doina Precup	Junming Shao
Wagner Meira	Ricardo Prudêncio	Yun Sing Koh
Ernestina Menasalvas	Kai Puolamäki	Andrzej Skowron
Corrado Mencar	Buyue Qian	Kevin Small
Rosa Meo	Chedy Raïssi	Tomislav Šmuc
Pauli Miettinen	Liva Ralaivola	Yangqiu Song
Dunja Mladenčić	Karthik Raman	Cheng Soon Ong
Anna Monreale	Jan Ramon	Arnaud Soulet
João Moreira	Huzefa Rangwala	Mauro Sozio
Emmanuel Müller	Zbigniew Ras	Alessandro Sperduti
Mohamed Nadif	Chotirat Ann	Eirini Spyropoulou
Mirco Nanni	Ratanamahatana	Steffen Staab
Amedeo Napoli	Jan Rauch	Gregor Stiglic
Houssam Nassif	Soumya Ray	Markus Strohmaier
Benjamin Nguyen	Jesse Read	Enrique Sucar
Thomas Niebler	Steffen Rendle	Mahito Sugiyama
Thomas Nielsen	Achim Rettinger	Johan Suykens
Siegfried Nijssen	Rita Ribeiro	Einoshin Suzuki
Xia Ning	Fabrizio Riguzzi	Panagiotis Symeonidis
Niklas Norén	Céline Robardet	Sándor Szedmák
Kjetil Nørvåg	Marko Robnik-Šikonja	Andrea Tagarelli
Eirini Ntoutsi	Juan Rodriguez	Domenico Talia
Andreas Nürnberger	Irene Rodríguez Luján	Letizia Tanca
Irene Ong	André Rossi	Dacheng Tao
Salvatore Orlando	Fabrice Rossi	Nikolaj Tatti
Gerhard Paaß	Juho Rousu	Maguelonne Teisseire
David Page	Céline Rouveirol	Alexandre Termier
George Paliouras	Salvatore Ruggieri	Evimaria Terzi
Panče Panov	Stefan Rüping	Ljupco Todorovski
Spiros Papadimitriou	Y. van Saeys	Vicenç Torra

Roberto Trasarti	Julien Velcin	Filip Železný
Brigitte Trousse	Shankar Vembu	Bernard Ženko
Panayiotis Tsaparas	Sicco Verwer	Junping Zhang
Vincent Tseng	Vassilios Verykios	Kun Zhang
Grigorios Tsoumakas	Herna Viktor	Lei Zhang
Theodoros Tzouramanis	Ricardo Vilalta	Min-Ling Zhang
Antti Ukkonen	Pavlovic Vladimir	Nan Zhang
Takeaki Uno	Christel Vrain	Shichao Zhang
Athina Vakali	Jilles Vreeken	Zhongfei Zhang
Wil van der Aalst	Willem Waegeman	Liang Zhao
Guy van der Broeck	Byron Wallace	Ying Zhao
Maarten van der Heijden	Fei Wang	Elena Zheleva
Peter van der Putten	Jianyong Wang	Bin Zhou
Matthijs van Leeuwen	Yang Wang	Kenny Zhu
Putten	Takashi Washio	Xiaofeng Zhu
Martijn van Otterlo	Jörg Simon Wicker	Djamel Zighed
Maarten van Someren	Chun-Nam Yu	Arthur Zimek
Joaquin Vanschoren	Jeffrey Yu	Albrecht Zimmermann
Iraklis Varlamis	Jure Zabkar	Blaž Zupan
Raju Vatsavai	Gerson Zaverucha	
Michalis Vazirgiannis	Demetris Zeinalipour	

Additional Reviewers

Greet Baldewijns	Sebastian Kauschke
Jessa Bekker	Jinseok Kim
Nuno Castro	Jan Kralj
Shiyu Chang	Thomas Low
Yu Cheng	Stijn Luca
Paolo Cintia	Rafael Mantovani
Heidar Davoudi	Pasquale Minervini
Thomas Delacroix	Shubhanshu Mishra
Martin Dimkovski	Christos Perentis
Michael Färber	Fábio Pinto
Ricky Fok	Dimitrios Rafailidis
Emanuele Frandi	Giulio Rossetti
Tatiana Gossen	Alexandros Sarafianos
Valerio Grossi	Antonio Vergari
Riccardo Guidotti	Dimtrios Vogiatzis
Ming Jiang	Andreas Zioupos
Nikos Katzouris	

Sponsors

Platinum Sponsors

BNP PARIBAS	http://www.bnpparibas.com/
ONR Global	www.onr.navy.mil/science-technology/onr-global.aspx

Gold Sponsors

Zalando	https://www.zalando.co.uk/
HUAWEI	http://www.huawei.com/en/

Silver Sponsors

Deloitte	http://www2.deloitte.com/
Amazon	http://www.amazon.com/

Bronze Sponsors

Xarevision	http://xarevision.pt/
Farfetch	http://www.farfetch.com/pt/
NOS	http://www.nos.pt/particulares/Pages/home.aspx

Award Sponsor

Machine Learning	http://link.springer.com/journal/10994
Data Mining and	http://link.springer.com/journal/10618
Knowledge	
Discovery Deloitte	http://www2.deloitte.com/

Lanyard Sponsor

KNIME	http://www.knime.org/
-------	---

Invited Talk Sponsors

ECCAI	http://www.eccai.org/
Cliqz	https://cliqz.com/
Technicolor	http://www.technicolor.com/
University of Bari Aldo Moro	http://www.uniba.it/english-version

Additional Supporters

INESCTEC	https://www.inesctec.pt/
University of Porto, Faculdade de Economia	http://sigarra.up.pt/fep/pt/web_page.inicial
Springer	http://www.springer.com/
University of Porto	http://www.up.pt/

Official Carrier

TAP	http://www.flytap.com/
-----	---

Abstracts of Invited Talks

Towards Declarative, Domain-Oriented Data Analysis

Hendrik Blockeel

University of Leuven, Leiden University

Abstract. The need for advanced data analysis now pervades all areas of science, industry and services. A wide variety of theory and techniques from statistics, data mining, and machine learning is available. Addressing a concrete question or problem in a particular application domain requires multiple non-trivial steps: translating the question to a data analysis problem, selecting a suitable approach to solve this problem, correctly applying that approach, and correctly interpreting the results. In this process, specialist knowledge on data analysis needs to be combined with domain expertise. As data analysis becomes ever more advanced, this becomes increasingly difficult. In an ideal world, data analysis would be declarative and domain-oriented: the user should be able to state the question, rather than describing a solution procedure, and the software should decide how to provide an answer. The user then no longer needs to be, or hire, a specialist in data analysis for every step of the knowledge discovery process. This would make data analysis easier, more efficient, and less error-prone. In this talk, I will discuss contemporary research that is bringing the state of the art in data analysis closer to that long-term goal. This includes research on inductive databases, constraint-based data mining, probabilistic-logical modeling, and declarative experimentation.

Bio. Hendrik Blockeel is a professor at the Computer Science department of KU Leuven, Belgium, and part-time associate professor at Leiden University, The Netherlands. His research interests lie mostly in machine learning and data mining. He has made a variety of research contributions in these fields, including work on decision tree learning, inductive logic programming, predictive clustering, probabilistic-logical models, inductive databases, constraint-based data mining, and declarative data analysis. He is an action editor for Machine Learning and serves on the editorial board of several other journals. He has chaired or organized multiple conferences, workshops, and summer schools, including ILP, ECMLPKDD, IDA and ACAI, and he has been vice-chair, area chair, or senior PC member for ECAI, IJCAI, ICML, KDD, ICDM. He was a member of the board of the European Coordinating Committee for Artificial Intelligence from 2004 to 2010, and currently serves as publications chair for the ECMLPKDD steering committee.

Sum-Product Networks: Deep Models with Tractable Inference

Pedro Domingos

University of Washington

Abstract. Big data makes it possible in principle to learn very rich probabilistic models, but inference in them is prohibitively expensive. Since inference is typically a subroutine of learning, in practice learning such models is very hard. Sum-product networks (SPNs) are a new model class that squares this circle by providing maximum flexibility while guaranteeing tractability. In contrast to Bayesian networks and Markov random fields, SPNs can remain tractable even in the absence of conditional independence. SPNs are defined recursively: an SPN is either a univariate distribution, a product of SPNs over disjoint variables, or a weighted sum of SPNs over the same variables. It's easy to show that the partition function, all marginals and all conditional MAP states of an SPN can be computed in time linear in its size. SPNs have most tractable distributions as special cases, including hierarchical mixture models, thin junction trees, and nonrecursive probabilistic context-free grammars. I will present generative and discriminative algorithms for learning SPN weights, and an algorithm for learning SPN structure. SPNs have achieved impressive results in a wide variety of domains, including object recognition, image completion, collaborative filtering, and click prediction. Our algorithms can easily learn SPNs with many layers of latent variables, making them arguably the most powerful type of deep learning to date. (Joint work with Rob Gens and Hoifung Poon.)

Bio. Pedro Domingos is Professor of Computer Science and Engineering at the University of Washington. His research interests are in machine learning, artificial intelligence and data science. He received a PhD in Information and Computer Science from the University of California at Irvine, and is the author or co-author of over 200 technical publications. He is a winner of the SIGKDD Innovation Award, the highest honor in data science. He is a AAAI Fellow, and has received a Sloan Fellowship, an NSF CAREER Award, a Fulbright Scholarship, an IBM Faculty Award, and best paper awards at several leading conferences. He is a member of the editorial board of the Machine Learning journal, co-founder of the International Machine Learning Society, and past associate editor of JAIR. He was program co-chair of KDD-2003 and SRL-2009, and has served on numerous program committees.

Mining Online Networks and Communities

Jure Leskovec

Stanford University

Abstract. The Internet and the Web fundamentally changed how we live our daily lives as well as broadened the scope of computer science. Today the Web is a ‘sensor’ that captures the pulse of humanity and allows us to observe phenomena that were once essentially invisible to us. These phenomena include the social interactions and collective behavior of hundreds of millions of people, recorded at unprecedented levels of scale and resolution. Analyzing this data offers novel algorithmic as well as computational challenges. Moreover, it offers new insights into the design of information systems in the presence of complex social feedback effects, as well as a new perspective on fundamental questions in the social sciences.

Bio. Jure Leskovec is assistant professor of Computer Science at Stanford University. His research focuses on mining large social and information networks. Problems he investigates are motivated by large scale data, the Web and on-line media. This research has won several awards including a Microsoft Research Faculty Fellowship, the Alfred P. Sloan Fellowship and numerous best paper awards. Leskovec received his bachelor’s degree in computer science from University of Ljubljana, Slovenia, and his PhD in machine learning from the Carnegie Mellon University and postdoctoral training at Cornell University. You can follow him on Twitter @jure.

Learning to Acquire Knowledge in a Smart Grid Environment

Chen-Ching Liu

Washington State University

Abstract. In a smart grid, a massive amount of data is collected by millions of sensors and meters on the transmission, distribution, and customers' facilities. There is a strong dependence of the smart grid on the information and communications technology for its monitoring and control. As a result, the cyber systems are also an important source of information. This presentation will be focused on the opportunities and challenges for machine learning and knowledge discovery in a smart grid environment. The application areas of (1) anomaly detection for cyber and physical security, and (2) intelligent remedial control of power grids will be used as examples.

Bio. Chen-Ching Liu is Boeing Distinguished Professor at Washington State University, Pullman, USA. During 1983-2005, he was a Professor of EE at University of Washington, Seattle. Dr. Liu was Palmer Chair Professor at Iowa State University from 2006 to 2008. From 2008-2011, he served as Acting/Deputy Principal of the College of Engineering, Mathematical and Physical Sciences at University College Dublin, Ireland. Professor Liu received an IEEE Third Millennium Medal in 2000 and the Power and Energy Society Outstanding Power Engineering Educator Award in 2004. In 2013, Dr. Liu received a Doctor Honoris Causa from Polytechnic University of Bucharest, Romania. He is a co-founder of the International Council on Intelligent Systems Application to Power Systems (ISAPs) and served as the founding president. He chaired the IEEE Power and Energy Society Technical Committee on Power System Analysis, Computing and Economics. Dr. Liu served on the U.S. National Academies Board on Global Science and Technology. Professor Liu is a Fellow of the IEEE. He was elected a Member of the Washington State Academy of Sciences in 2014.

Untangling the Web's Invisible Net

Nataša Milić-Frayling

Microsoft Research

Abstract. This presentation will shed light on user tracking and behavioural targeting on the Web. Through empirical studies of cookie tracking practices, we will take an alternative view of the display ad business by observing the network of third party trackers that envelopes the Web. The practice begs a question of how to resolve a dissonance between the current consumer tracking practices and the vendors' desire for consumers' loyalty and trustful long-term engagements. It also makes us aware of how computing designs and techniques, inaccessible to individuals, cause imbalance in the knowledge acquisition and enablement, disempowering the end-users.

Bio. As a Principal Researcher at Microsoft Research (MSR) in Cambridge, Nataša is working on the design, prototyping and evaluation of information and communication systems. She is passionate about innovation in personal and social computing and promotes a dialogue between IT industry, consumers, and policy makers on the issues that arise from the adoption and use of technologies. Her current focus is on digital obsolescence, activity based computing, and privacy respecting systems and applications. Nataša is actively involved with a wider community of academics and practitioners through public speaking, collaborative projects, and serving on advisory boards of academic programs and commercial enterprises. She is a Visiting Professor at the UCL and Queen Mary University of London and a member of the ACM Europe Council. She serves on the Advisory Boards for the Course in Entrepreneurship at the University of Cambridge and the Turing Gateway in Mathematics at the Isaac Newton Institute for Mathematical Sciences (INI).

Towards a Digital Time Machine Fueled by Big Data and Social Mining

Dino Pedreschi

University of Pisa

Abstract. My seminar discusses the novel questions that big data and social mining allow to raise and answer, how a new paradigm for scientific exploration, statistics and policy making is emerging, and the major scientific, technological and societal barriers to be overcome to realize this vision. I will focus on concrete projects with telecom providers and official statistics bureau in Italy and France aimed at measuring, quantifying and possibly predicting key demographic and socio-economic indicators based on nation-wide mobile phone data: the population of different categories of city users (residents, commuters, visitors) in urban spaces, the inter-city mobility, the level of well-being and economic development of geographical units at various scales.

Bio. Dino Pedreschi is a Professor of Computer Science at the University of Pisa, and a pioneering scientist in mobility data mining, social network mining and privacy-preserving data mining. He co-leads with Fosca Giannotti the Pisa KDD Lab - Knowledge Discovery and Data Mining Laboratory, a joint research initiative of the University of Pisa and the Information Science and Technology Institute of the Italian National Research Council, one of the earliest research lab centered on data mining. His research focus is on big data analytics and mining and their impact on society. He is a founder of the Business Informatics MSc program at Univ. Pisa, a course targeted at the education of interdisciplinary data scientists. Dino has been a visiting scientist at Barabasi Lab (Center for Complex Network Research) of Northeastern University, Boston (2009-2010), and earlier at the University of Texas at Austin (1989-90), at CWI Amsterdam (1993) and at UCLA (1995). In 2009, Dino received a Google Research Award for his research on privacy-preserving data mining.

Abstracts of Journal Track Articles

A Bayesian Approach for Comparing Cross-Validated Algorithms on Multiple Data Sets

Giorgio Corani and Alessio Benavoli

Machine Learning

DOI: [10.1007/s10994-015-5486-z](https://doi.org/10.1007/s10994-015-5486-z)

We present a Bayesian approach for making statistical inference about the accuracy (or any other score) of two competing algorithms which have been assessed via cross-validation on multiple data sets. The approach is constituted by two pieces. The first is a novel correlated Bayesian t -test for the analysis of the cross-validation results on a single data set which accounts for the correlation due to the overlapping training sets. The second piece merges the posterior probabilities computed by the Bayesian correlated t -test on the different data sets to make inference on multiple data sets. It does so by adopting a Poisson-binomial model. The inferences on multiple data sets account for the different uncertainty of the cross-validation results on the different data sets. It is the first test able to achieve this goal. It is generally more powerful than the signed-rank test if ten runs of cross-validation are performed, as it is anyway generally recommended.

A Decomposition of the Outlier Detection Problem into a Set of Supervised Learning Problems

Heiko Paulheim and Robert Meusel

Machine Learning

DOI: [10.1007/s10994-015-5507-y](https://doi.org/10.1007/s10994-015-5507-y)

Outlier detection methods automatically identify instances that deviate from the majority of the data. In this paper, we propose a novel approach for unsupervised outlier detection, which re-formulates the outlier detection problem in numerical data as a set of supervised regression learning problems. For each attribute, we learn a predictive model which predicts the values of that attribute from the values of all other attributes, and compute the deviations between the predictions and the actual values. From those deviations, we derive both a weight for each attribute, and a final outlier score using those weights. The weights help separating the relevant attributes from the irrelevant ones, and thus make the approach well suitable for discovering outliers otherwise masked in high-dimensional data. An empirical evaluation shows that our approach outperforms existing algorithms, and is particularly robust in datasets with many irrelevant attributes. Furthermore, we show that if a symbolic machine learning method is used to solve the individual learning problems, the approach is also capable of generating concise explanations for the detected outliers.

Assessing the Impact of a Health Intervention via User-Generated Internet Content

Vasileios Lamos, Elad Yom-Tov, Richard Pebody, and Ingemar J. Cox

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0427-9](https://doi.org/10.1007/s10618-015-0427-9)

Assessing the effect of a health-oriented intervention by traditional epidemiological methods is commonly based only on population segments that use healthcare services. Here we introduce a complementary framework for evaluating the impact of a targeted intervention, such as a vaccination campaign against an infectious disease, through a statistical analysis of user-generated content submitted on web platforms. Using supervised learning, we derive a nonlinear regression model for estimating the prevalence of a health event in a population from Internet data. This model is applied to identify control location groups that correlate historically with the areas, where a specific intervention campaign has taken place. We then determine the impact of the intervention by inferring a projection of the disease rates that could have emerged in the absence of a campaign. Our case study focuses on the influenza vaccination program that was launched in England during the 2013/14 season, and our observations consist of millions of geo-located search queries to the Bing search engine and posts on Twitter. The impact estimates derived from the application of the proposed statistical framework support conventional assessments of the campaign.

Beyond Rankings: Comparing Directed Acyclic Graphs

Eric Malmi, Nikolaj Tatti, Aristides Gionis

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0406-1](https://doi.org/10.1007/s10618-015-0406-1)

Defining appropriate distance measures among rankings is a classic area of study which has led to many useful applications. In this paper, we propose a more general abstraction of preference data, namely directed acyclic graphs (DAGs), and introduce a measure for comparing DAGs, given that a vertex correspondence between the DAGs is known. We study the properties of this measure and use it to aggregate and cluster a set of DAGs. We show that these problems are NP-hard and present efficient methods to obtain solutions with approximation guarantees. In addition to preference data, these methods turn out to have other interesting applications, such as the analysis of a collection of information cascades in a network. We test the methods on synthetic and real-world datasets, showing that the methods can be used to, e.g., find a set of influential individuals related to a set of topics in a network or to discover meaningful and occasionally surprising clustering structure.

Clustering Boolean Tensors

Saskia Metzler and Pauli Miettinen

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0420-3](https://doi.org/10.1007/s10618-015-0420-3)

Graphs - such as friendship networks - that evolve over time are an example of data that are naturally represented as binary tensors. Similarly to analysing the adjacency matrix of a graph using a matrix factorization, we can analyse the tensor by factorizing it. Unfortunately, tensor factorizations are computationally hard problems, and in particular, are often significantly harder than their matrix counterparts. In case of Boolean tensor factorizations - where the input tensor and all the factors are required to be binary and we use Boolean algebra - much of that hardness comes from the possibility of overlapping components. Yet, in many applications we are perfectly happy to partition at least one of the modes. For instance, in the aforementioned timeevolving friendship networks, groups of friends might be overlapping, but the time points at which the network was captured are always distinct. In this paper we investigate what consequences this partitioning has on the computational complexity of the Boolean tensor factorizations and present a new algorithm for the resulting clustering problem. This algorithm can alternatively be seen as a particularly regularized clustering algorithm that can handle extremely high-dimensional observations. We analyse our algorithm with the goal of maximizing the similarity and argue that this is more meaningful than minimizing the dissimilarity. As a by-product we obtain a PTAS and an efficient 0.828-approximation algorithm for rank-1 binary factorizations. Our algorithm for Boolean tensor clustering achieves high scalability, high similarity, and good generalization to unseen data with both synthetic and realworld data sets.

Consensus Hashing

Cong Leng and Jian Cheng

Machine Learning

DOI: [10.1007/s10994-015-5496-x](https://doi.org/10.1007/s10994-015-5496-x)

Hashing techniques have been widely used in many machine learning applications because of their efficiency in both computation and storage. Although a variety of hashing methods have been proposed, most of them make some implicit assumptions about the statistical or geometrical structure of data. In fact, few hashing algorithms can adequately handle all kinds of data with different structures. When considering hybrid structure datasets, different hashing algorithms might produce different and possibly inconsistent binary codes. Inspired by the successes of classifier combination and clustering ensembles, in this paper, we present a novel combination strategy for multiple hashing results, named Consensus Hashing (CH). By defining the measure of consensus of two hashing results, we put forward a simple yet effective model to learn

consensus hash functions which generate binary codes consistent with the existing ones. Extensive experiments on several large scale benchmarks demonstrate the overall superiority of the proposed method compared with state-of-the art hashing algorithms.

Convex Relaxations of Penalties for Sparse Correlated Variables With Bounded Total Variation

Eugene Belilovsky, Andreas Argyriou, Gael Varoquaux, Matthew B. Blaschko
Machine Learning

DOI: [10.1007/s10994-015-5511-2](https://doi.org/10.1007/s10994-015-5511-2)

We study the problem of statistical estimation with a signal known to be sparse, spatially contiguous, and containing many highly correlated variables. We take inspiration from the recently introduced k-support norm, which has been successfully applied to sparse prediction problems with correlated features, but lacks any explicit structural constraints commonly found in machine learning and image processing. We address this problem by incorporating a total variation penalty in the k-support framework. We introduce the (k,s) support total variation norm as the tightest convex relaxation of the intersection of a set of sparsity and total variation constraints. We show that this norm leads to an intractable combinatorial graph optimization problem, which we prove to be NP-hard. We then introduce a tractable relaxation with approximation guarantees that scale well for grid structured graphs. We devise several first-order optimization strategies for statistical parameter estimation with the described penalty. We demonstrate the effectiveness of this penalty on classification in the low sample regime, classification with M/EEG neuroimaging data, and image recovery with synthetic and real data background subtracted image recovery tasks. We extensively analyse the application of our penalty on the complex task of identifying predictive regions from low-sample high-dimensional fMRI brain data, we show that our method is particularly useful compared to existing methods in terms of accuracy, interpretability, and stability.

Direct Conditional Probability Density Estimation with Sparse Feature Selection

Motoki Shiga, Voot Tangkaratt, and Masashi Sugiyama
Machine Learning

DOI: [10.1007/s10994-014-5472-x](https://doi.org/10.1007/s10994-014-5472-x)

Regression is a fundamental problem in statistical data analysis, which aims at estimating the conditional mean of output given input. However, regression is not informative enough if the conditional probability density is multi-modal, asymmetric, and heteroscedastic. To overcome this limitation, various estimators of conditional densities themselves have been developed, and a kernel-based approach called

least-squares conditional density estimation (LS-CDE) was demonstrated to be promising. However, LS-CDE still suffers from large estimation error if input contains many irrelevant features. In this paper, we therefore propose an extension of LS-CDE called sparse additive CDE (SA-CDE), which allows automatic feature selection in CDE. SACDE applies kernel LS-CDE to each input feature in an additive manner and penalizes the whole solution by a group-sparse regularizer. We also give a subgradient-based optimization method for SA-CDE training that scales well to high-dimensional large data sets. Through experiments with benchmark and humanoid robot transition datasets, we demonstrate the usefulness of SA-CDE in noisy CDE problems.

DRESS: Dimensionality Reduction for Efficient Sequence Search

*Alexios Kotsifakos, Alexandra Stefan, Vassilis Athitsos, Gautam Das,
and Panagiotis Papapetrou*

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0413-2](https://doi.org/10.1007/s10618-015-0413-2)

Similarity search in large sequence databases is a problem ubiquitous in a wide range of application domains, including searching biological sequences. In this paper we focus on protein and DNA data, and we propose a novel approximate method for speeding up range queries under the edit distance. Our method works in a filter-and-refine manner, and its key novelty is a query-sensitive mapping that transforms the original string space to a new string space of reduced dimensionality. Specifically, it first identifies the most frequent codewords in the query, and then uses these codewords to convert both the query and the database to a more compact representation. This is achieved by replacing every occurrence of each codeword with a new letter and by removing the remaining parts of the strings. Using this new representation, our method identifies a set of candidate matches that are likely to satisfy the range query, and finally refines these candidates in the original space. The main advantage of our method, compared to alternative methods for whole sequence matching under the edit distance, is that it does not require any training to create the mapping, and it can handle large query lengths with negligible losses in accuracy. Our experimental evaluation demonstrates that, for higher range values and large query sizes, our method produces significantly lower costs and runtimes compared to two state-of-the-art competitor methods.

Dynamic Inference of Social Roles in Information Cascade

*Sarvenaz Choobdar, Pedro Ribeiro, Srinivasan Parthasarathy,
and Fernando Silva*

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0402-5](https://doi.org/10.1007/s10618-015-0402-5)

Nodes in complex networks inherently represent different kinds of functional or organizational roles. In the dynamic process of an information cascade, users play different roles in spreading the information: some act as seeds to initiate the process, some limit the propagation and others are in-between. Understanding the roles of users is crucial in modeling the cascades. Previous research mainly focuses on modeling users behavior based upon the dynamic exchange of information with neighbors. We argue however that the structural patterns in the neighborhood of nodes may already contain enough information to infer users' roles, independently from the information flow in itself. To approach this possibility, we examine how network characteristics of users affect their actions in the cascade. We also advocate that temporal information is very important. With this in mind, we propose an unsupervised methodology based on ensemble clustering to classify users into their social roles in a network, using not only their current topological positions, but also considering their history over time. Our experiments on two social networks, Flickr and Digg, show that topological metrics indeed possess discriminatory power and that different structural patterns correspond to different parts in the process. We observe that user commitment in the neighborhood affects considerably the influence score of users. In addition, we discover that the cohesion of neighborhood is important in the blocking behavior of users. With this we can construct topological fingerprints that can help us in identifying social roles, based solely on structural social ties, and independently from nodes activity and how information flows.

Efficient and Effective Community Search

*Nicola Barbieri, Francesco Bonchi, Edoardo Galimberti,
and Francesco Gullo*

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0422-1](https://doi.org/10.1007/s10618-015-0422-1)

Community search is the problem of finding a good community for a given set of query vertices. One of the most studied formulations of community search asks for a connected subgraph that contains all query vertices and maximizes the minimum degree. All existing approaches to min-degree-based community search suffer from limitations concerning efficiency, as they need to visit (large part of) the whole input graph, as well as accuracy, as they output communities quite large and not really cohesive. Moreover, some existing methods lack generality: they handle only single-vertex queries, find communities that are not optimal in terms of minimum degree, and/or require input parameters. In this work we advance the state of the art on

community search by proposing a novel method that overcomes all these limitations: it is in general more efficient and effective—one/two orders of magnitude on average, it can handle multiple query vertices, it yields optimal communities, and it is parameter-free. These properties are confirmed by an extensive experimental analysis performed on various real-world graphs.

Finding the Longest Common Sub-Pattern in Sequences of Temporal Intervals

Orestis Kostakis and Panagiotis Papapetrou

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0404-3](https://doi.org/10.1007/s10618-015-0404-3)

We study the problem of finding the Longest Common Sub-Pattern (LCSP) shared by two sequences of temporal intervals. In particular we are interested in finding the LCSP of the corresponding arrangements. Arrangements of temporal intervals are a powerful way to encode multiple concurrent labeled events that have a time duration. Discovering commonalities among such arrangements is useful for a wide range of scientific fields and applications, as it can be seen by the number and diversity of the datasets we use in our experiments. In this paper, we define the problem of LCSP and prove that it is NP-complete by demonstrating a connection between graphs and arrangements of temporal intervals, which leads to a series of interesting open problems. In addition, we provide an exact algorithm to solve the LCSP problem, and also propose and experiment with three polynomial time and space underapproximation techniques. Finally, we introduce two upper bounds for LCSP and study their suitability for speeding up 1-NN search. Experiments are performed on seven datasets taken from a wide range of real application domains, plus two synthetic datasets.

Generalization Bounds for Learning with Linear, Polygonal, Quadratic and Conic Side Knowledge

Theja Tulabandhula and Cynthia Rudin

Machine Learning

DOI: [10.1007/s10994-014-5478-4](https://doi.org/10.1007/s10994-014-5478-4)

In this paper, we consider a supervised learning setting where side knowledge is provided about the labels of unlabeled examples. The side knowledge has the effect of reducing the hypothesis space, leading to tighter generalization bounds, and thus possibly better generalization. We consider several types of side knowledge, the first leading to linear and polygonal constraints on the hypothesis space, the second leading to quadratic constraints, and the last leading to conic constraints. We show how different types of domain knowledge can lead directly to these kinds of side knowledge.

We prove bounds on complexity measures of the hypothesis space for quadratic and conic side knowledge, and show that these bounds are tight in a specific sense for the quadratic case.

Generalization of Clustering Agreements and Distances for Overlapping Clusters and Network Communities

Reihaneh Rabbany and Osmar R. Zaiane

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0426-x](https://doi.org/10.1007/s10618-015-0426-x)

A measure of distance between two clusterings has important applications, including clustering validation and ensemble clustering. Generally, such distance measure provides navigation through the space of possible clusterings. Mostly used in cluster validation, a normalized clustering distance, a.k.a. agreement measure, compares a given clustering result against the ground-truth clustering. The two widely-used clustering agreement measures are Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). In this paper, we present a generalized clustering distance from which these two measures can be derived. We then use this generalization to construct new measures specific for comparing (dis)agreement of clusterings in networks, a.k.a. communities. Further, we discuss the difficulty of extending the current, contingency based, formulations to overlapping cases, and present an alternative algebraic formulation for these (dis)agreement measures. Unlike the original measures, the new co-membership based formulation is easily extendable for different cases, including overlapping clusters and clusters of inter-related data. These two extensions are, in particular, important in the context of finding communities in complex networks.

Generalized Twin Gaussian Processes Using Sharma-Mittal Divergence

Mohamed Elhoseiny and Ahmed Elgammal

Machine Learning

DOI: [10.1007/s10994-015-5497-9](https://doi.org/10.1007/s10994-015-5497-9)

There has been a growing interest in mutual information measures due to its wide range of applications in Machine Learning and Computer Vision. In this manuscript, we present a generalized structured regression framework based on Sharma-Mittal divergence, a relative entropy measure, firstly addressed in the Machine Learning community, in this work. Sharma-Mittal (SM) divergence is a generalized mutual information measure for the widely used Rényi, Tsallis, Bhattacharyya, and Kullback-Leibler (KL) relative entropies. Specifically, we study Sharma-Mittal divergence as a cost function in the context of the Twin Gaussian Processes, which generalizes over the KL-divergence without computational penalty. We show interesting properties of Sharma-Mittal TGP (SMTGP) through a theoretical analysis,

which covers missing insights in the traditional TGP formulation. However, we generalize this theory based on SM-divergence instead of KL-divergence which is a special case. Experimentally, we evaluated the proposed SMTGP framework on several datasets. The results show that SMTGP reaches better predictions than KL-based TGP (KLTGP), since it offers a bigger class of models through its parameters that we learn from the data.

Half-Space Mass: A Maximally Robust and Efficient Data Depth Method

Bo Chen, Kai Ming Ting, Takashi Washio, and Gholamreza Haffari

Machine Learning

DOI: [10.1007/s10994-015-5524-x](https://doi.org/10.1007/s10994-015-5524-x)

Data depth is a statistical method which models data distribution in terms of centeroutward ranking rather than density or linear ranking. While there are a lot of academic interests, its applications are hampered by the lack of a method which is both robust and efficient. This paper introduces Half-Space Mass which is a significantly improved version of half-space data depth. Half-Space Mass is the only data depth method which is both robust and efficient, as far as we know. We also reveal four theoretical properties of Half-Space Mass: (i) its resultant mass distribution is concave regardless of the underlying density distribution, (ii) its maximum point is unique which can be considered as median, (iii) the median is maximally robust, and (iv) its estimation extends to a higher dimensional space in which the convex hull of the dataset occupies zero volume. We demonstrate the power of Half-Space Mass through its applications in two tasks. In anomaly detection, being a maximally robust location estimator leads directly to a robust anomaly detector that yields a better detection accuracy than halfspace depth; and it runs orders of magnitude faster than L2 depth, an existing maximally robust location estimator. In clustering, the Half-Space Mass version of Kmeans overcomes three weaknesses of K-means.

Improving Classification Performance Through Selective Instance Completion

Amit Dhurandhar and Karthik Sankarnarayanan

Machine Learning

DOI: [10.1007/s10994-015-5500-5](https://doi.org/10.1007/s10994-015-5500-5)

In multiple domains, actively acquiring missing input information at a reasonable cost in order to improve our understanding of the input-output relationships is of increasing importance. This problem has gained prominence in healthcare, public policy making, education, and in the targeted advertising industry which tries to best match people to products. In this paper we tackle an important variant of this problem: Instance Completion, where we want to choose the best k incomplete instances to query from a

much larger universe of $N(>>k)$ incomplete instances so as to learn the most accurate classifier. We propose a principled framework which motivates a generally applicable yet efficient meta-technique for choosing k such instances. Since we cannot know *a priori* the classifier that will result from the completed dataset, i.e. the final classifier, our method chooses the k instances based on a derived upper bound on the expectation of the distance between the next classifier and the final classifier. We additionally derive a sufficient condition for these two solutions to match. We then empirically evaluate the performance of our method relative to the state-of-the-art methods on 4 UCI datasets as well as 3 proprietary e-commerce datasets used in previous studies. In these experiments, we also demonstrate how close we are likely to be to the optimal solution, by quantifying the extent to which our sufficient condition is satisfied. Lastly, we show that our method is easily extensible to the setting where we have a non uniform cost associated with acquiring the missing information.

Incremental Learning of Event Definitions with Inductive Logic Programming

Nikos Katzouris, Alexander Artikis, and Georgios Paliouras

Machine Learning

DOI: [10.1007/s10994-015-5512-1](https://doi.org/10.1007/s10994-015-5512-1)

Event recognition systems rely on knowledge bases of event definitions to infer occurrences of events in time. Using a logical framework for representing and reasoning about events offers direct connections to machine learning, via Inductive Logic Programming (ILP), thus allowing to avoid the tedious and error-prone task of manual knowledge construction. However, learning temporal logical formalisms, which are typically utilized by logic-based event recognition systems is a challenging task, which most ILP systems cannot fully undertake. In addition, event-based data is usually massive and collected at different times and under various circumstances. Ideally, systems that learn from temporal data should be able to operate in an incremental mode, that is, revise prior constructed knowledge in the face of new evidence. In this work we present an incremental method for learning and revising event-based knowledge, in the form of Event Calculus programs. The proposed algorithm relies on abductive-inductive learning and comprises a scalable clause refinement methodology, based on a compressive summarization of clause coverage in a stream of examples. We present an empirical evaluation of our approach on real and synthetic data from activity recognition and city transport applications.

Knowledge Base Completion by Learning Pairwise-Interaction Differentiated Embeddings

Yu Zhao, Sheng Gao, Patrick Gallinari, and Jun Guo

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0430-1](https://doi.org/10.1007/s10618-015-0430-1)

Knowledge base consisting of triple like (subject entity, predicate relation, object entity) is a very important database for knowledge management. It is very useful for humanlike reasoning, query expansion, question answering (Siri) and other related AI tasks. However, knowledge base often suffers from incompleteness due to a large volume of increasing knowledge in the real world and a lack of reasoning capability. In this paper, we propose a Pairwise-interaction Differentiated Embeddings (PIDE) model to embed entities and relations in the knowledge base to low dimensional vector representations and then predict the possible truth of additional facts to extend the knowledge base. In addition, we present a probability-based objective function to improve the model optimization. Finally, we evaluate the model by considering the problem of computing how likely the additional triple is true for the task of knowledge base completion. Experiments on WordNet and Freebase dataset show the excellent performance of our model and algorithm.

Learning from Evolving Video Streams in a Multi-camera Scenario

Samaneh Khoshrou, Jaime dos Santos Cardoso, and Luís Filipe Teixeira

Machine Learning

DOI: [10.1007/s10994-015-5515-y](https://doi.org/10.1007/s10994-015-5515-y)

Nowadays, video surveillance systems are taking the first steps toward automation, in order to ease the burden on human resources as well as to avoid human error. As the underlying data distribution and the number of concepts change over time, the conventional learning algorithms fail to provide reliable solutions for this setting. Herein, we formalize a learning concept suitable for multi-camera video surveillance and propose a learning methodology adapted to that new paradigm. The proposed framework resorts to the universal background model to robustly learn individual object models from small samples and to more effectively detect novel classes. The individual models are incrementally updated in an ensemble based approach, with older models being progressively forgotten. The framework is designed to detect and label new concepts automatically. The system is also designed to exploit active learning strategies, in order to interact wisely with operator, requesting assistance in the most ambiguous to classify observations. The experimental results obtained both on real and synthetic data sets verify the usefulness of the proposed approach.

Learning Relational Dependency Networks in Hybrid Domains

Irma Ravkic, Jan Ramon, and Jesse Davis

Machine Learning

DOI: [10.1007/s10994-015-5483-2](https://doi.org/10.1007/s10994-015-5483-2)

Statistical Relational Learning (SRL) is concerned with developing formalisms for representing and learning from data that exhibit both uncertainty and complex, relational structure. Most of the work in SRL has focused on modeling and learning from data that only contain discrete variables. As many important problems are characterized by the presence of both continuous and discrete variables, there has been a growing interest in developing hybrid SRL formalisms. Most of these formalisms focus on reasoning and representational issues and, in some cases, parameter learning. What has received little attention is learning the structure of a hybrid SRL model from data. In this paper, we fill that gap and make the following contributions. First, we propose Hybrid Relational Dependency Networks (HRDNs), an extension to Relational Dependency Networks that are able to model continuous variables. Second, we propose an algorithm for learning both the structure and parameters of an HRDN from data. Third, we provide an empirical evaluation that demonstrates that explicitly modeling continuous variables results in more accurate learned models than discretizing them prior to learning.

MassExodus: Modeling Evolving Networks in Harsh Environments

Saket Navlakha, Christos Faloutsos, and Ziv Bar-Joseph

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-014-0399-1](https://doi.org/10.1007/s10618-014-0399-1)

Defining appropriate distance measures among rankings is a classic area of study which has led to many useful applications. In this paper, we propose a more general abstraction of preference data, namely directed acyclic graphs (DAGs), and introduce a measure for comparing DAGs, given that a vertex correspondence between the DAGs is known. We study the properties of this measure and use it to aggregate and cluster a set of DAGs. We show that these problems are NP-hard and present efficient methods to obtain solutions with approximation guarantees. In addition to preference data, these methods turn out to have other interesting applications, such as the analysis of a collection of information cascades in a network. We test the methods on synthetic and real-world datasets, showing that the methods can be used to, e.g., find a set of influential individuals related to a set of topics in a network or to discover meaningful and occasionally surprising clustering structure.

Minimum Message Length Estimation of Mixtures of Multivariate Gaussian and von Mises-Fisher Distribution

Parthan Kasarapu and Lloyd Allison

Machine Learning

DOI: [10.1007/s10994-015-5493-0](https://doi.org/10.1007/s10994-015-5493-0)

Mixture modelling involves explaining some observed evidence using a combination of probability distributions. The crux of the problem is the inference of an optimal number of mixture components and their corresponding parameters. This paper discusses unsupervised learning of mixture models using the Bayesian Minimum Message Length (MML) criterion. To demonstrate the effectiveness of search and inference of mixture parameters using the proposed approach, we select two key probability distributions, each handling fundamentally different types of data: the multivariate Gaussian distribution to address mixture modelling of data distributed in Euclidean space, and the multivariate von Mises-Fisher (vMF) distribution to address mixture modelling of directional data distributed on a unit hypersphere. The key contributions of this paper, in addition to the general search and inference methodology, include the derivation of MML expressions for encoding the data using multivariate Gaussian and von Mises-Fisher distributions, and the analytical derivation of the MML estimates of the parameters of the two distributions. Our approach is tested on simulated and real world data sets. For instance, we infer vMF mixtures that concisely explain experimentally determined three dimensional protein conformations, providing an effective null model description of protein structures that is central to many inference problems in structural bioinformatics. The experimental results demonstrate that the performance of our proposed search and inference method along with the encoding schemes improve on the state of the art mixture modelling techniques.

Mining Outlying Aspects on Numeric Data

Lei Duan, Guanting Tang, Jian Pei, James Bailey,

Akiko Campbell, and Changjie Tang

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-014-0398-2](https://doi.org/10.1007/s10618-014-0398-2)

When we are investigating an object in a data set, which itself may or may not be an outlier, can we identify unusual (i.e., outlying) aspects of the object? In this paper, we identify the novel problem of mining outlying aspects on numeric data. Given a query object o in a multidimensional numeric data set O , in which subspace is o most outlying? Technically, we use the rank of the probability density of an object in a subspace to measure the outlyingness of the object in the subspace. A minimal subspace where the query object is ranked the best is an outlying aspect. Computing the outlying aspects of a query object is far from trivial. A naïve method has to calculate the probability densities of all objects and rank them in every subspace, which is very

costly when the dimensionality is high. We systematically develop a heuristic method that is capable of searching data sets with tens of dimensions efficiently. Our empirical study using both real data and synthetic data demonstrates that our method is effective and efficient.

Multiscale Event Detection in Social Media

Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, Pascal Frossard

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0421-2](https://doi.org/10.1007/s10618-015-0421-2)

Event detection has been one of the most important research topics in social media analysis. Most of the traditional approaches detect events based on fixed temporal and spatial resolutions, while in reality events of different scales usually occur simultaneously, namely, they span different intervals in time and space. In this paper, we propose a novel approach towards multiscale event detection using social media data, which takes into account different temporal and spatial scales of events in the data. Specifically, we explore the properties of the wavelet transform, which is a welldeveloped multiscale transform in signal processing, to enable automatic handling of the interaction between temporal and spatial scales. We then propose a novel algorithm to compute a data similarity graph at appropriate scales and detect events of different scales simultaneously by a single graph-based clustering process. Furthermore, we present spatiotemporal statistical analysis of the noisy information present in the data stream, which allows us to define a novel term-filtering procedure for the proposed event detection algorithm and helps us study its behavior using simulated noisy data. Experimental results on both synthetically generated data and real world data collected from Twitter demonstrate the meaningfulness and effectiveness of the proposed approach. Our framework further extends to numerous application domains that involve multiscale and multiresolution data analysis.

Optimised Probabilistic Active Learning (OPAL) for Fast, Non-Myopic, Cost-Sensitive Active Classification

Georg Kreml, Daniel Kottke, and Vincent Lemaire

Machine Learning

DOI: [10.1007/s10994-015-5504-1](https://doi.org/10.1007/s10994-015-5504-1)

In contrast to ever increasing volumes of automatically generated data, human annotation capacities remain limited. Thus, fast active learning approaches that allow the efficient allocation of annotation efforts gain in importance. Furthermore, cost-sensitive applications such as fraud detection pose the additional challenge of differing misclassification costs between classes. Unfortunately, the few existing cost-sensitive active learning approaches rely on time-consuming steps, such as

performing self labelling or tedious evaluations over samples. We propose a fast, non-myopic, and cost-sensitive probabilistic active learning approach for binary classification. Our approach computes the expected reduction in misclassification loss in a labelling candidate's neighbourhood. We derive and use a closed-form solution for this expectation, which considers the possible values of the true posterior of the positive class at the candidate's position, its possible label realisations, and the given labelling budget. The resulting myopic algorithm runs in the same linear asymptotic time as uncertainty sampling, while its non-myopic counterpart requires an additional factor of $O(m \log m)$ in the budget size. The experimental evaluation on several synthetic and real-world data sets shows competitive or better classification performance and runtime, compared to several uncertainty sampling- and error-reduction-based active learning strategies, both in cost-sensitive and cost-insensitive settings.

Poisson Dependency Networks - Gradient Boosted Models for Multivariate Count Data

Fabian Hadiji, Alejandro Molina, Sriraam Natarajan, and Kristian Kersting
Machine Learning

DOI: [10.1007/s10994-015-5506-z](https://doi.org/10.1007/s10994-015-5506-z)

Although count data are increasingly ubiquitous, surprisingly little work has employed probabilistic graphical models for modeling count data. Indeed the univariate case has been well studied, however, in many situations counts influence each other and should not be considered independently. Standard graphical models such as multinomial or Gaussian ones are also often ill-suited, too, since they disregard either the infinite range over the natural numbers or the potentially asymmetric shape of the distribution of count variables. Existing classes of Poisson graphical models can only model negative conditional dependencies or neglect the prediction of counts or do not scale well. To ease the modeling of multivariate count data, we therefore introduce a novel family of Poisson graphical models, called Poisson Dependency Networks (PDNs). A PDN consists of a set of local conditional Poisson distributions, each representing the probability of a single count variable given the others, that naturally facilitates a simple Gibbs sampling inference. In contrast to existing Poisson graphical models, PDNs are non-parametric and trained using functional gradient ascent, i.e., boosting. The particularly simple form of the Poisson distribution allows us to develop the first multiplicative boosting approach: starting from an initial constant value, alternatively a log-linear Poisson model, or a Poisson regression tree, a PDN is represented as products of regression models grown in a stage-wise optimization. We demonstrate on several real world datasets that PDNs can model positive and negative dependencies and scale well while often outperforming state-of-the-art, in particular when using multiplicative updates.

Policy Gradient in Lipschitz Markov Decision Processes

Matteo Pirodda, Marcello Restelli, and Luca Bascetta

Machine Learning

DOI: [10.1007/s10994-015-5484-1](https://doi.org/10.1007/s10994-015-5484-1)

This paper is about the exploitation of Lipschitz continuity properties for Markov Decision Processes (MDPs) to safely speed up policy-gradient algorithms. Starting from assumptions about the Lipschitz continuity of the state-transition model, the reward function, and the policies considered in the learning process, we show that both the expected return of a policy and its gradient are Lipschitz continuous w.r.t. policy parameters. By leveraging such properties, we define policy-parameter updates that guarantee a performance improvement at each iteration. The proposed methods are empirically evaluated and compared to other related approaches using different configurations of three popular control scenarios: the linear quadratic regulator, the mass-spring-damper system and the ship-steering control.

Probabilistic Clustering of Time-Evolving Distance Data

Julia Vogt, Marius Kloft, Stefan Stark, Sudhir S. Raman,

Sandhya Prabhakaran, Volker Roth, and Gunnar Rätsch

Machine Learning

DOI: [10.1007/s10994-015-5516-x](https://doi.org/10.1007/s10994-015-5516-x)

We present a novel probabilistic clustering model for objects that are represented via pairwise distances and observed at different time points. The proposed method utilizes the information given by adjacent time points to find the underlying cluster structure and obtain a smooth cluster evolution. This approach allows the number of objects and clusters to differ at every time point, and no identification on the identities of the objects is needed. Further, the model does not require the number of clusters being specified in advance – they are instead determined automatically using a Dirichlet process prior. We validate our model on synthetic data showing that the proposed method is more accurate than state-of-the-art clustering methods. Finally, we use our dynamic clustering model to analyze and illustrate the evolution of brain cancer patients over time.

Ranking Episodes Using a Partition Model

Nikolaj Tatti

Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0419-9](https://doi.org/10.1007/s10618-015-0419-9)

One of the biggest setbacks in traditional frequent pattern mining is that overwhelmingly many of the discovered patterns are redundant. A prototypical example of such redundancy is a freerider pattern where the pattern contains a true pattern and some

additional noise events. A technique for filtering freerider patterns that has proved to be efficient in ranking itemsets is to use a partition model where a pattern is divided into two subpatterns and the observed support is compared to the expected support under the assumption that these two subpatterns occur independently. In this paper we develop a partition model for episodes, patterns discovered from sequential data. An episode is essentially a set of events, with possible restrictions on the order of events. Unlike with itemset mining, computing the expected support of an episode requires surprisingly sophisticated methods. In order to construct the model, we partition the episode into two subepisodes. We then model how likely the events in each subepisode occur close to each other. If this probability is high—which is often the case if the subepisode has a high support—then we can expect that when one event from a subepisode occurs, then the remaining events occur also close by. This approach increases the expected support of the episode, and if this increase explains the observed support, then we can deem the episode uninteresting. We demonstrate in our experiments that using the partition model can effectively and efficiently reduce the redundancy in episodes.

Regularized Feature Selection in Reinforcement Learning

Dean Stephen Wookey and George Dimitri Konidaris

Machine Learning

DOI: [10.1007/s10994-015-5518-8](https://doi.org/10.1007/s10994-015-5518-8)

We introduce feature regularization during feature selection for value function approximation. Feature regularization introduces a prior into the selection process, improving function approximation accuracy and reducing overfitting. We show that the smoothness prior is effective in the incremental feature selection setting and present closed-form smoothness regularizers for the Fourier and RBF bases. We present two methods for feature regularization which extend the temporal difference orthogonal matching pursuit (OMP-TD) algorithm and demonstrate the effectiveness of the smoothness prior; smooth Tikhonov OMP-TD and smoothness scaled OMP-TD. We compare these methods against OMP-TD, regularized OMP-TD and least squares TD with random projections, across six benchmark domains using two different types of basis functions.

Soft-max Boosting

Matthieu Geist

Machine Learning

DOI: [10.1007/s10994-015-5491-2](https://doi.org/10.1007/s10994-015-5491-2)

The standard multi-class classification risk, based on the binary loss, is rarely directly minimized. This is due to (i) the lack of convexity and (ii) the lack of smoothness (and even continuity). The classic approach consists in minimizing instead a convex

surrogate. In this paper, we propose to replace the usually considered deterministic decision rule by a stochastic one, which allows obtaining a smooth risk (generalizing the expected binary loss, and more generally the cost-sensitive loss). Practically, this (empirical) risk is minimized by performing a gradient descent in the function space linearly spanned by a base learner (a.k.a. boosting). We provide a convergence analysis of the resulting algorithm and experiment it on a bunch of synthetic and real world data sets (with noiseless and noisy domains, compared to convex and non convex boosters).

Tractome: A Visual Data Mining Tool for Brain Connectivity Analysis

*Diana Porro-Munoz, Emanuele Olivetti, Nusrat Sharmin,
Thien Bao Nguyen, Eleftherios Garyfallidis, and Paolo Avesani*
Data Mining and Knowledge Discovery

DOI: [10.1007/s10618-015-0408-z](https://doi.org/10.1007/s10618-015-0408-z)

Diffusion magnetic resonance imaging data allows reconstructing the neural pathways of the white matter of the brain as a set of 3D polylines. This kind of data sets provides a means of study of the anatomical structures within the white matter, in order to detect neurologic diseases and understand the anatomical connectivity of the brain. To the best of our knowledge, there is still not an effective or satisfactory method for automatic processing of these data. Therefore, a manually guided visual exploration of experts is crucial for the purpose. However, because of the large size of these data sets, visual exploration and analysis has also become intractable. In order to make use of the advantages of both manual and automatic analysis, we have developed a new visual data mining tool for the analysis of human brain anatomical connectivity. With such tool, humans and automatic algorithms capabilities are integrated in an interactive data exploration and analysis process. A very important aspect to take into account when designing this tool, was to provide the user with comfortable interaction. For this purpose, we tackle the scalability issue in the different stages of the system, including the automatic algorithm and the visualization and interaction techniques that are used.

Contents – Part I

Research Track

Classification, Regression and Supervised Learning

Data Split Strategies for Evolving Predictive Models	3
<i>Vikas C. Raykar and Amrita Saha</i>	
Discriminative Interpolation for Classification of Functional Data	20
<i>Rana Haber, Anand Rangarajan, and Adrian M. Peter</i>	
Fast Label Embeddings via Randomized Linear Algebra	37
<i>Paul Mineiro and Nikos Karampatziakis</i>	
Maximum Entropy Linear Manifold for Learning Discriminative Low-Dimensional Representation	52
<i>Wojciech Marian Czarnecki, Rafal Jozefowicz, and Jacek Tabor</i>	
Novel Decompositions of Proper Scoring Rules for Classification: Score Adjustment as Precursor to Calibration	68
<i>Meelis Kull and Peter Flach</i>	
Parameter Learning of Bayesian Network Classifiers Under Computational Constraints	86
<i>Sebastian Tschiatschek and Franz Pernkopf</i>	
Predicting Unseen Labels Using Label Hierarchies in Large-Scale Multi-label Learning	102
<i>Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J. Kim, and Johannes Fürnkranz</i>	
Regression with Linear Factored Functions.	119
<i>Wendelin Böhmer and Klaus Obermayer</i>	
Ridge Regression, Hubness, and Zero-Shot Learning	135
<i>Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto</i>	
Solving Prediction Games with Parallel Batch Gradient Descent	152
<i>Michael Großhans and Tobias Scheffer</i>	
Structured Regularizer for Neural Higher-Order Sequence Models.	168
<i>Martin Ratajczak, Sebastian Tschiatschek, and Franz Pernkopf</i>	

Versatile Decision Trees for Learning over Multiple Contexts.	184
<i>Reem Al-Otaibi, Ricardo B.C. Prudêncio, Meelis Kull, and Peter Flach</i>	
When Is Undersampling Effective in Unbalanced Classification Tasks?	200
<i>Andrea Dal Pozzolo, Olivier Caelen, and Gianluca Bontempi</i>	

Clustering and Unsupervised Learning

A Kernel-Learning Approach to Semi-supervised Clustering with Relative Distance Comparisons	219
<i>Ehsan Amid, Aristides Gionis, and Antti Ukkonen</i>	
Bayesian Active Clustering with Pairwise Constraints	235
<i>Yuanli Pei, Li-Ping Liu, and Xiaoli Z. Fern</i>	
ConDist: A Context-Driven Categorical Distance Measure	251
<i>Markus Ring, Florian Otto, Martin Becker, Thomas Niebler, Dieter Landes, and Andreas Hotho</i>	
Discovering Opinion Spammer Groups by Network Footprints	267
<i>Junting Ye and Leman Akoglu</i>	
Gamma Process Poisson Factorization for Joint Modeling of Network and Documents.	283
<i>Ayan Acharya, Dean Teffer, Jette Henderson, Marcus Tyler, Mingyuan Zhou, and Joydeep Ghosh</i>	
Generalization in Unsupervised Learning	300
<i>Karim T. Abou-Moustafa and Dale Schuurmans</i>	
Multiple Incomplete Views Clustering via Weighted Nonnegative Matrix Factorization with $L_{2,1}$ Regularization	318
<i>Weixiang Shao, Lifang He, and Philip S. Yu</i>	
Solving a Hard Cutting Stock Problem by Machine Learning and Optimisation.	335
<i>Steven D. Prestwich, Adejuyigbe O. Fajemisin, Laura Climent, and Barry O’Sullivan</i>	

Data Preprocessing

Markov Blanket Discovery in Positive-Unlabelled and Semi-supervised Data	351
<i>Konstantinos Sechidis and Gavin Brown</i>	
Multi-View Semantic Learning for Data Representation	367
<i>Peng Luo, Jinye Peng, Ziyu Guan, and Jianping Fan</i>	

Unsupervised Feature Analysis with Class Margin Optimization	383
<i>Sen Wang, Feiping Nie, Xiaojun Chang, Lina Yao, Xue Li, and Quan Z. Sheng</i>	

Data Streams and Online Learning

Ageing-Based Multinomial Naive Bayes Classifiers over Opinionated Data Streams	401
<i>Sebastian Wagner, Max Zimmermann, Eirini Ntoutsi, and Myra Spiliopoulou</i>	
Drift Detection Using Stream Volatility	417
<i>David Tse Jung Huang, Yun Sing Koh, Gillian Dobbie, and Albert Bifet</i>	
Early Classification of Time Series as a Non Myopic Sequential Decision Making Problem	433
<i>Asma Dachraoui, Alexis Bondu, and Antoine Cornuéjols</i>	
Ising Bandits with Side Information	448
<i>Shaona Ghosh and Adam Prügel-Bennett</i>	
Refined Algorithms for Infinitely Many-Armed Bandits with Deterministic Rewards.	464
<i>Yahel David and Nahum Shimkin</i>	

Deep Learning

An Empirical Investigation of Minimum Probability Flow Learning Under Different Connectivity Patterns	483
<i>Daniel Jiwoong Im, Ethan Buchman, and Graham W. Taylor</i>	
Difference Target Propagation	498
<i>Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio</i>	
Online Learning of Deep Hybrid Architectures for Semi-supervised Categorization.	516
<i>Alexander G. Ororbia II, David Reitter, Jian Wu, and C. Lee Giles</i>	
Scoring and Classifying with Gated Auto-Encoders	533
<i>Daniel Jiwoong Im and Graham W. Taylor</i>	
Sign Constrained Rectifier Networks with Applications to Pattern Decompositions	546
<i>Senjian An, QiuHong Ke, Mohammed Bennamoun, Farid Boussaid, and Ferdous Sohel</i>	

Aggregation Under Bias: Rényi Divergence Aggregation and its Implementation via Machine Learning Markets	560
<i>Amos J. Storkey, Zhanxing Zhu, and Jinli Hu</i>	

Distance and Metric Learning

Higher Order Fused Regularization for Supervised Learning with Grouped Parameters	577
<i>Koh Takeuchi, Yoshinobu Kawahara, and Tomoharu Iwata</i>	
Joint Semi-supervised Similarity Learning for Linear Classification	594
<i>Maria-Irina Nicolae, Éric Gaussier, Amaury Habrard, and Marc Sebban</i>	
Learning Compact and Effective Distance Metrics with Diversity Regularization.	610
<i>Pengtao Xie</i>	
Scalable Metric Learning for Co-Embedding	625
<i>Farzaneh Mirzazadeh, Martha White, András György, and Dale Schuurmans</i>	

Large Scale Learning and Big Data

Adaptive Stochastic Primal-Dual Coordinate Descent for Separable Saddle Point Problems	645
<i>Zhanxing Zhu and Amos J. Storkey</i>	
Hash Function Learning via Codewords.	659
<i>Yinjie Huang, Michael Georgiopoulos, and Georgios C. Anagnostopoulos</i>	
HierCost: Improving Large Scale Hierarchical Classification with Cost Sensitive Learning.	675
<i>Anveshi Charuvaka and Huzefa Rangwala</i>	
Large Scale Optimization with Proximal Stochastic Newton-Type Gradient Descent	691
<i>Ziqiang Shi and Rujie Liu</i>	
Erratum to: Bayesian Active Clustering with Pairwise Constraints.	E1
<i>Yuanli Pei, Li-Ping Liu, and Xiaoli Z. Fern</i>	
Erratum to: Scalable Metric Learning for Co-Embedding	E3
<i>Farzaneh Mirzazadeh, Martha White, András György, and Dale Schuurmans</i>	
Erratum to: Predicting Unseen Labels Using Label Hierarchies in Large-Scale Multi-label Learning	E6
<i>Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J. Kim, and Johannes Fürnkranz</i>	

Author Index	705
-------------------------------	------------

Contents – Part II

Research Track

Matrix and Tensor Analysis

BoostMF: Boosted Matrix Factorisation for Collaborative Ranking	3
<i>Nipa Chowdhury, Xiongcai Cai, and Cheng Luo</i>	
Convex Factorization Machines	19
<i>Mathieu Blondel, Akinori Fujino, and Naonori Ueda</i>	
Generalized Matrix Factorizations as a Unifying Framework for Pattern Set Mining: Complexity Beyond Blocks	36
<i>Pauli Miettinen</i>	
Scalable Bayesian Non-Negative Tensor Factorization for Massive Count Data.	53
<i>Changwei Hu, Piyush Rai, Changyou Chen, Matthew Harding, and Lawrence Carin</i>	
A Practical Approach to Reduce the Learning Bias Under Covariate Shift . . .	71
<i>Van-Tinh Tran and Alex Aussem</i>	
Hyperparameter Optimization with Factorized Multilayer Perceptrons	87
<i>Nicolas Schilling, Martin Wistuba, Lucas Drumond, and Lars Schmidt-Thieme</i>	
Hyperparameter Search Space Pruning – A New Component for Sequential Model-Based Hyperparameter Optimization	104
<i>Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme</i>	
Multi-Task Learning with Group-Specific Feature Space Sharing	120
<i>Niloofar Yousefi, Michael Georgiopoulos, and Georgios C. Anagnostopoulos</i>	
Opening the Black Box: Revealing Interpretable Sequence Motifs in Kernel-Based Learning Algorithms	137
<i>Marina M.-C. Vidovic, Nico Görnitz, Klaus-Robert Müller, Gunnar Rätsch, and Marius Kloft</i>	

Pattern and Sequence Mining

Fast Generation of Best Interval Patterns for Nonmonotonic Constraints	157
<i>Aleksey Buzmakov, Sergei O. Kuznetsov, and Amedeo Napoli</i>	

Non-Parametric Jensen-Shannon Divergence.	173
<i>Hoang-Vu Nguyen and Jilles Vreeken</i>	
Swap Randomization of Bases of Sequences for Mining Satellite Image Times Series.	190
<i>Nicolas M��ger, Christophe Rigotti, and Catherine Pothier</i>	
The Difference and the Norm Characterising Similarities and Differences Between Databases	206
<i>Kailash Budhathoki and Jilles Vreeken</i>	
Preference Learning and Label Ranking	
Dyad Ranking Using A Bilinear Plackett-Luce Model	227
<i>Dirk Sch��fer and Eyke H��llermeier</i>	
Fast Training of Support Vector Machines for Survival Analysis.	243
<i>Sebastian P��lsterl, Nassir Navab, and Amin Katouzian</i>	
Superset Learning Based on Generalized Loss Minimization.	260
<i>Eyke H��llermeier and Weiwei Cheng</i>	
Probabilistic, Statistical, and Graphical Approaches	
Bayesian Modelling of the Temporal Aspects of Smart Home Activity with Circular Statistics.	279
<i>Tom Diethe, Niall Twomey, and Peter Flach</i>	
Message Scheduling Methods for Belief Propagation.	295
<i>Christian Knoll, Michael Rath, Sebastian Tschatschek, and Franz Pernkopf</i>	
Output-Sensitive Adaptive Metropolis-Hastings for Probabilistic Programs . . .	311
<i>David Tolpin, Jan-Willem van de Meent, Brooks Paige, and Frank Wood</i>	
Planning in Discrete and Continuous Markov Decision Processes by Probabilistic Programming.	327
<i>Davide Nitti, Vaishak Belle, and Luc De Raedt</i>	
Simplifying, Regularizing and Strengthening Sum-Product Network Structure Learning.	343
<i>Antonio Vergari, Nicola Di Mauro, and Floriana Esposito</i>	
Sparse Bayesian Recurrent Neural Networks.	359
<i>Sotirios P. Chatzis</i>	

Structured Prediction of Sequences and Trees Using Infinite Contexts	373
<i>Ehsan Shareghi, Gholamreza Haffari, Trevor Cohn, and Ann Nicholson</i>	
Temporally Coherent Role-Topic Models (TCRTM): Deinterlacing Overlapping Activity Patterns	390
<i>Evgeniy Bart, Bob Price, and John Hanley</i>	
The Blind Leading the Blind: Network-Based Location Estimation Under Uncertainty	406
<i>Eric Malmi, Arno Solin, and Aristides Gionis</i>	
Weighted Rank Correlation: A Flexible Approach Based on Fuzzy Order Relations.	422
<i>Sascha Henzgen and Eyke Hüllermeier</i>	

Rich Data

Concurrent Inference of Topic Models and Distributed Vector Representations.	441
<i>Debakar Shamanta, Sheikh Motahar Naim, Parang Saraf, Naren Ramakrishnan, and M. Shahriar Hossain</i>	
Differentially Private Analysis of Outliers	458
<i>Rina Okada, Kazuto Fukuchi, and Jun Sakuma</i>	
Inferring Unusual Crowd Events from Mobile Phone Call Detail Records . . .	474
<i>Yuxiao Dong, Fabio Pinelli, Yiannis Gkoufas, Zubair Nabi, Francesco Calabrese, and Nitesh V. Chawla</i>	
Learning Pretopological Spaces for Lexical Taxonomy Acquisition	493
<i>Guillaume Cleuziou and Gaël Dias</i>	
Multidimensional Prediction Models When the Resolution Context Changes.	509
<i>Adolfo Martínez-Usó and José Hernández-Orallo</i>	
Semi-Supervised Subspace Co-Projection for Multi-class Heterogeneous Domain Adaptation	525
<i>Min Xiao and Yuhong Guo</i>	
Towards Computation of Novel Ideas from Corpora of Scientific Text	541
<i>Haixia Liu, James Goulding, and Tim Brailsford</i>	

Social and Graphs

Discovering Audience Groups and Group-Specific Influencers	559
<i>Shuyang Lin, Qingbo Hu, Jingyuan Zhang, and Philip S. Yu</i>	

Estimating Potential Customers Anywhere and Anytime Based on Location-Based Social Networks	576
<i>Hsun-Ping Hsieh, Cheng-Te Li, and Shou-De Lin</i>	
Exact Hybrid Covariance Thresholding for Joint Graphical Lasso	593
<i>Qingming Tang, Chao Yang, Jian Peng, and Jinbo Xu</i>	
Fast Inbound Top-K Query for Random Walk with Restart	608
<i>Chao Zhang, Shan Jiang, Yucheng Chen, Yidan Sun, and Jiawei Han</i>	
Finding Community Topics and Membership in Graphs	625
<i>Matt Revelle, Carlotta Domeniconi, Mack Sweeney, and Aditya Johri</i>	
Finding Dense Subgraphs in Relational Graphs	641
<i>Vinay Jethava and Niko Beerenwinkel</i>	
Generalized Modularity for Community Detection	655
<i>Mohadeseh Ganji, Abbas Seifi, Hosein Alizadeh, James Bailey, and Peter J. Stuckey</i>	
Handling Oversampling in Dynamic Networks Using Link Prediction	671
<i>Benjamin Fish and Rajmonda S. Caceres</i>	
Hierarchical Sparse Dictionary Learning	687
<i>Xiao Bian, Xia Ning, and Geoff Jiang</i>	
Latent Factors Meet Homophily in Diffusion Modelling	701
<i>Minh-Duc Luu and Ee-Peng Lim</i>	
Maintaining Sliding-Window Neighborhood Profiles in Interaction Networks	719
<i>Rohit Kumar, Toon Calders, Aristides Gionis, and Nikolaj Tatti</i>	
Response-Guided Community Detection: Application to Climate Index Discovery	736
<i>Gonzalo A. Bello, Michael Angus, Navya Pedemane, Jitendra K. Harlalka, Fredrick H.M. Semazzi, Vipin Kumar, and Nagiza F. Samatova</i>	
Robust Classification of Information Networks by Consistent Graph Learning	752
<i>Shi Zhi, Jiawei Han, and Quanquan Gu</i>	
Author Index	769

Contents – Part III

Industrial Track

Autonomous HVAC Control, a Reinforcement Learning Approach	3
<i>Enda Barrett and Stephen Linder</i>	
Clustering by Intent: A Semi-Supervised Method to Discover Relevant Clusters Incrementally	20
<i>George Forman, Hila Nachlieli, and Renato Keshet</i>	
Country-Scale Exploratory Analysis of Call Detail Records Through the Lens of Data Grid Models	37
<i>Romain Guigourès, Dominique Gay, Marc Boullé, Fabrice Clérot, and Fabrice Rossi</i>	
Early Detection of Fraud Storms in the Cloud	53
<i>Hani Neuirth, Yehuda Finkelstein, Amit Hilbuch, Shai Nahum, Daniel Alon, and Elad Yom-Tov</i>	
Flexible Sliding Windows for Kernel Regression Based Bus Arrival Time Prediction	68
<i>Hoang Thanh Lam and Eric Bouillet</i>	
Learning Detector of Malicious Network Traffic from Weak Labels	85
<i>Vojtech Franc, Michal Sofka, and Karel Bartos</i>	
Online Analysis of High-Volume Data Streams in Astroparticle Physics	100
<i>Christian Bockermann, Kai Brügge, Jens Buss, Alexey Egorov, Katharina Morik, Wolfgang Rhode, and Tim Ruhe</i>	
Robust Representation for Domain Adaptation in Network Security.	116
<i>Karel Bartos and Michal Sofka</i>	
Safe Exploration for Active Learning with Gaussian Processes	133
<i>Jens Schreiter, Duy Nguyen-Tuong, Mona Eberts, Bastian Bischoff, Heiner Markert, and Marc Toussaint</i>	
Semi-Supervised Consensus Clustering for ECG Pathology Classification . . .	150
<i>Helena Aidos, André Lourenço, Diana Batista, Samuel Rota Bulò, and Ana Fred</i>	
Two Step Graph-Based Semi-supervised Learning for Online Auction Fraud Detection	165
<i>Phiradet Bangcharoensap, Hayato Kobayashi, Nobuyuki Shimizu, Satoshi Yamauchi, and Tsuyoshi Murata</i>	

Watch-It-Next: A Contextual TV Recommendation System	180
<i>Michal Aharon, Eshcar Hillel, Amit Kagian, Ronny Lempel, Hayim Makabee, and Raz Nissim</i>	

Nectar Track

Bayesian Hypothesis Testing in Machine Learning	199
<i>Giorgio Corani, Alessio Benavoli, Francesca Mangili, and Marco Zaffalon</i>	
Data-Driven Exploration of Real-Time Geospatial Text Streams	203
<i>Harald Bosch, Robert Krüger, and Dennis Thom</i>	
Discovering Neutrinos Through Data Analytics.	208
<i>Mathis Börner, Wolfgang Rhode, Tim Ruhe, and Katharina Morik</i>	
Listener-Aware Music Recommendation from Sensor and Social Media Data	213
<i>Markus Schedl</i>	
Logic-Based Incremental Process Mining	218
<i>Stefano Ferilli, Domenico Redavid, and Floriana Esposito</i>	
Mobility Mining for Journey Planning in Rome	222
<i>Michele Berlingerio, Veli Bicer, Adi Botea, Stefano Braghin, Nuno Lopes, Riccardo Guidotti, and Francesca Pratesi</i>	
Pattern Structures and Concept Lattices for Data Mining and Knowledge Processing.	227
<i>Mehdi Kaytoue, Victor Codocedo, Aleksey Buzmakov, Jaume Baixeries, Sergei O. Kuznetsov, and Amedeo Napoli</i>	
Privacy Preserving Blocking and Meta-Blocking	232
<i>Alexandros Karakasidis, Georgia Koloniari, and Vassilios S. Verykios</i>	
Social Data Mining and Seasonal Influenza Forecasts: The FluOutlook Platform	237
<i>Qian Zhang, Corrado Gioannini, Daniela Paolotti, Nicola Perra, Daniela Perrotta, Marco Quaggiotto, Michele Tizzoni, and Alessandro Vespignani</i>	
Star Classification Under Data Variability: An Emerging Challenge in Astrominformatics	241
<i>Ricardo Vilalta, Kinjal Dhar Gupta, and Ashish Mahabal</i>	
The Evolution of Social Relationships and Strategies Across the Lifespan . . .	245
<i>Yuxiao Dong, Nitesh V. Chawla, Jie Tang, Yang Yang, and Yang Yang</i>	

Understanding Where Your Classifier Does (Not) Work	250
<i>Wouter Duivesteijn and Julia Thaele</i>	
Visual Analytics Methodology for Scalable and Privacy-Respectful Discovery of Place Semantics from Episodic Mobility Data	254
<i>Natalia Andrienko, Gennady Andrienko, Georg Fuchs, and Piotr Jankowski</i>	
Will This Paper Increase Your <i>h</i> -index?	259
<i>Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla</i>	
Demo Track	
CardioWheel: ECG Biometrics on the Steering Wheel	267
<i>André Lourenço, Ana Priscila Alves, Carlos Carreiras, Rui Policarpo Duarte, and Ana Fred</i>	
Data Patterns Explained with Linked Data	271
<i>Ilaria Tididi, Mathieu d'Aquin, and Enrico Motta</i>	
GAZOUILLE: Detecting and Illustrating Local Events from Geolocalized Social Media Streams	276
<i>Pierre Houdyer, Albrecht Zimmerman, Mehdi Kaytoue, Marc Plantevit, Joseph Mitchell, and Céline Robardet</i>	
HiDER: Query-Driven Entity Resolution for Historical Data.	281
<i>Bijan Ranjbar-Sahraei, Julia Efremova, Hossein Rahmani, Toon Calders, Karl Tuyls, and Gerhard Weiss</i>	
IICE: Web Tool for Automatic Identification of Chemical Entities and Interactions	285
<i>Andre Lamurias, Luka A. Clarke, and Francisco M. Couto</i>	
Interactively Exploring Supply and Demand in the UK Independent Music Scene.	289
<i>Matt McVicar, Cédric Mesnage, Jefrey Lijffijt, and Tijl De Bie</i>	
Monitoring Production Equipment in Remote Locations.	293
<i>Ujval Kamath, Ana Costa e Silva, and Michael O'Connell</i>	
MULTI-NETS: Web-Based Multilayer Network Visualization	298
<i>Matija Piškorec, Borut Sluban, and Tomislav Šmuc</i>	
OMEGA: An Order-Preserving SubMatrix Mining, Indexing and Search Tool	303
<i>Tao Jiang, Zhanhuai Li, Qun Chen, Zhong Wang, Kaiwen Li, and Wei Pan</i>	

Probabilistic Programming in Anglican	308
<i>David Tolpin, Jan-Willem van de Meent, and Frank Wood</i>	
ProbLog2: Probabilistic Logic Programming.	312
<i>Anton Dries, Angelika Kimmig, Wannes Meert, Joris Renkens, Guy Van den Broeck, Jonas Vlasselaer, and Luc De Raedt</i>	
Real Time Detection and Tracking of Spatial Event Clusters	316
<i>Natalia Andrienko, Gennady Andrienko, Georg Fuchs, Salvatore Rinzivillo, and Hans-Dieter Betz</i>	
S&P360: Multidimensional Perspective on Companies from Online Data Sources	320
<i>Michele Berlingiero, Stefano Braghin, Francesco Calabrese, Cody Dunne, Yiannis Gkoufas, Mauro Martino, Jamie Rasmussen, and Steven Ross</i>	
SCAVENGER – A Framework for Efficient Evaluation of Dynamic and Modular Algorithms	325
<i>Andrey Tyukin, Stefan Kramer, and Jörg Wicker</i>	
UrbanHubble: Location Prediction and Geo-Social Analytics in LBSN	329
<i>Roland Assam, Simon Feiden, and Thomas Seidl</i>	
VIPER – Visual Pattern Explorer	333
<i>Matthijs van Leeuwen and Lara Cardinaels</i>	
Visualization Support to Interactive Cluster Analysis	338
<i>Gennady Andrienko and Natalia Andrienko</i>	
Author Index	343