Lecture Notes in Artificial Intelligence 9285

Subseries of Lecture Notes in Computer Science

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

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

Machine Learning and Knowledge Discovery in Databases

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

Springer

Editors
Annalisa Appice
University of Bari Aldo Moro
Bari
Italy

João Gama
University of Porto - INESC TEC
Porto
Portugal

Pedro Pereira Rodrigues
University of Porto
Porto
Portugal

Alípio Jorge
University of Porto - INESC TEC
Porto
Portugal

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

Carlos Soares
University of Porto - INESC TEC
Porto
Portugal

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-23524-0 ISBN 978-3-319-23525-7 (eBook)
DOI 10.1007/978-3-319-23525-7

Library of Congress Control Number: 2015947118

LNCS Sublibrary: SL7 – Artificial Intelligence

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

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

Preface

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

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

Additional Reviewers

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

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 Knowledge Discovery	http://link.springer.com/journal/10618
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 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

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

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 Lampos, Elad Yom-Tov, Richard Pebody, and Ingemar J. Cox
Data Mining and Knowledge Discovery
DOI: 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

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

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

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

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 parameterestimation 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

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

leastsquares 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

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 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

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

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

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

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

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

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 Shama-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

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

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

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 algorithmrelies 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

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

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

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

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

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

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

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 Krempl, Daniel Kottke, and Vincent Lemaire
Machine Learning
DOI: 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

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 facilities 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 Pirotta, Marcello Restelli, and Luca Bascetta
Machine Learning
DOI: 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

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

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

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

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

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 II

Preference Learning and Label Ranking

Probabilistic, Statistical, and Graphical Approaches