Lecture Notes in Artificial Intelligence 10534

Subseries of Lecture Notes in Computer Science

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

Michelangelo Ceci · Jaakko Hollmén
Ljupčo Todorovski · Celine Vens
Sašo Džeroski (Eds.)

Machine Learning and Knowledge Discovery in Databases

European Conference, ECML PKDD 2017
Skopje, Macedonia, September 18–22, 2017
Proceedings, Part I

Springer

Editors
Michelangelo Ceci
Università degli Studi di Bari Aldo Moro
Bari
Italy

Jaakko Hollmén
Aalto University School of Science
Espoo
Finland

Ljupčo Todorovski
University of Ljubljana
Ljubljana
Slovenia

Celine Vens
KU Leuven Kulak
Kortrijk
Belgium

Sašo Džeroski
Jožef Stefan Institute
Ljubljana
Slovenia

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-71248-2 ISBN 978-3-319-71249-9 (eBook)
https://doi.org/10.1007/978-3-319-71249-9

Library of Congress Control Number: 2017961799

LNCS Sublibrary: SL7 – Artificial Intelligence

Printed on acid-free paper

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

Preface

We are delighted to introduce the proceedings of the 2017 edition of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2017). This year the conference was held in Skopje, Macedonia, during September 18–22, 2017. ECML PKDD is an annual conference that provides an international forum for the discussion of the latest high-quality research results in all areas related to machine learning and knowledge discovery in databases as well as innovative application domains. This event is the premier European machine learning and data mining conference and builds upon a very successful series of 18 ECML, ten PKDD (until 2007, when they merged), and nine ECML PKDD (from 2008). Therefore, this was the tenth edition of ECML PKDD as a single conference.

The scientific program was very rich and consisted of technical presentations of accepted papers, plenary talks by distinguished keynote speakers, workshops, and tutorials. Accepted papers were organized in five different tracks:

- The conference track, featuring regular conference papers, published in the proceedings
- The journal track, featuring papers that satisfy the quality criteria of journal papers and at the same time lend themselves to conference talks (these papers are published separately in the journals *Machine Learning* and *Data Mining and Knowledge Discovery*)
- The applied data science track (formerly industrial track), aiming to bring 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 demo track, presenting innovative prototype implementations or mature systems that use machine learning techniques and knowledge discovery processes in a real setting
- The nectar track, offering conference attendees a compact overview of recent scientific advances at the frontier of machine learning and data mining with other disciplines, as published in related conferences and journals.

In addition, the PhD Forum provided a friendly environment for junior PhD students to exchange ideas and experiences with peers in an interactive atmosphere and to get constructive feedback from senior researchers. This year we also introduced the EU Projects Forum with the purpose of disseminating EU projects and their results to the targeted scientific audience of the conference participants. Moreover, two discovery challenges, six tutorials, and 17 co-located workshops were organized on related research topics.

Following the successful experience of last year, we stimulated the practices of reproducible research (RR). Authors were encouraged to adhere to such practices by making available data and software tools for reproducing the results reported in their

papers. In total, 57% of the accepted papers in the conference have accompanying software and/or data and are flagged as RR papers on the conference website. In the proceedings, each RR paper provides links to such additional material. This year, we took a further step toward data publishing and facilitated the authors in making data and software available on a public repository in a dedicated area branded with the conference name (e.g., figshare, where there is an ECML PKDD area).

The response to our call for papers was very good. We received 364 papers for the main conference track, of which 104 were accepted, yielding an acceptance rate of about 28%. This allowed us to define a very rich program with 101 presentations in the main conference track: The remaining three accepted papers were not presented at the conference and are not included in the proceedings. The program also included six plenary keynotes by invited speakers: Inderjit Dhillon (University of Texas at Austin and Amazon), Alex Graves (Google DeepMind), Frank Hutter (University of Freiburg), Pierre-Philippe Mathieu (ESA/ESRIN), John Quackenbush (Dana-Farber Cancer Institute and Harvard TH Chan School of Public Health), and Cordelia Schmid (Inria).

This year, ECML PKDD attracted over 600 participants from 49 countries. It also attracted substantial attention from industry and end users, both through sponsorship and submission/participation at the applied data science track. This is confirmation that the ECML PKDD community is healthy, strong, and continuously growing.

Many people worked hard together as a superb dedicated team to ensure the successful organization of this conference: the general chairs (Michelangelo Ceci and Sašo Džeroski), the PC chairs (Michelangelo Ceci, Jaakko Hollmén, Ljupčo Todorovski, Celine Vens), the journal track chairs (Kurt Driessens, Dragi Kocev, Marko Robnik-Šikonja, Myra Spiliopoulou), the applied data science chairs (Yasemin Altun, Kamalika Das, Taneli Mielikäinen), the nectar track chairs (Donato Malerba, Jerzy Stefanowski), the demonstration track chairs (Jesse Read, Marinka Žitnik), the workshops and tutorials chairs (Nathalie Japkowicz, Panče Panov), the EU projects forum chairs (Petra Kralj Novak, Nada Lavrač), the PhD forum chairs (Tomislav Šmuc, Bernard Ženko), the Awards Committee (Peter Flach, Rosa Meo, Indrė Žliobaitė), the discovery challenge chair (Dino Ienco), the production and public relations chairs (Dragi Kocev, Nikola Simidjievski), the local organizers (Ivica Dimitrovski, Tina Anžič, Mili Bauer, Gjorgji Madjarov), the sponsorship chairs (Albert Bifet, Panče Panov), the proceedings chairs (Jurica Levatić, Gianvito Pio), the area chairs (listed in this book), the Program Committee members (listed in this book), and members of the MLJ and DMKD Guest Editorial Boards. They made tremendous effort in guaranteeing the quality of the reviewing process and, therefore, the scientific quality of the conference, which is certainly beneficial for the whole community. Our sincere thanks go to all of them.

We would also like to thank the Cankarjev Dom congress agency and the student volunteers. Thanks to Springer for their continuous support and Microsoft for allowing us to use their CMT software for conference management. Special thanks to the sponsors (Deutsche Post DHL Group, Google, AGT, ASML, Deloitte, NEC, Siemens, Cambridge University Press, *IEEE/CAA Journal of Automatica Sinica*, Springer, IBM Research, *Data Mining and Knowledge Discovery*, *Machine Learning*, EurAi, and

GrabIT) and the European project MAESTRA (ICT-2013-612944), as well as the ECML PKDD Steering Committee (for their suggestions and advice). Finally, we would like to thank the organizing institutions: the Jožef Stefan Institute (Slovenia), the Ss. Cyril and Methodius University in Skopje (Macedonia), and the University of Bari Aldo Moro (Italy).

September 2017

Michelangelo Ceci
Jaakko Hollmén
Ljupčo Todorovski
Celine Vens
Sašo Džeroski

Organization

ECML PKDD 2017 Organization

Conference Chairs

Michelangelo Ceci	University of Bari Aldo Moro, Italy
Sašo Džeroski	Jožef Stefan Institute, Slovenia

Program Chairs

Michelangelo Ceci	University of Bari Aldo Moro, Italy
Jaakko Hollmén	Aalto University, Finland
Ljupčo Todorovski	University of Ljubljana, Slovenia
Celine Vens	KU Leuven Kulak, Belgium

Journal Track Chairs

Kurt Driessens	Maastricht University, The Netherlands
Dragi Kocev	Jožef Stefan Institute, Slovenia
Marko Robnik-Šikonja	University of Ljubljana, Slovenia
Myra Spiliopoulu	Magdeburg University, Germany

Applied Data Science Track Chairs

Yasemin Altun	Google Research, Switzerland
Kamalika Das	NASA Ames Research Center, USA
Taneli Mielikäinen	Yahoo! USA

Local Organization Chairs

Ivica Dimitrovski	Ss. Cyril and Methodius University, Macedonia
Tina Anžič	Jožef Stefan Institute, Slovenia
Mili Bauer	Jožef Stefan Institute, Slovenia
Gjorgji Madjarov	Ss. Cyril and Methodius University, Macedonia

Workshops and Tutorials Chairs

Nathalie Japkowicz	American University, USA
Panče Panov	Jožef Stefan Institute, Slovenia

Awards Committee

Peter Flach	University of Bristol, UK
Rosa Meo	University of Turin, Italy
Indrė Žliobaitė	University of Helsinki, Finland

Nectar Track Chairs

Donato Malerba	University of Bari Aldo Moro, Italy
Jerzy Stefanowski	Poznan University of Technology, Poland

Demo Track Chairs

Jesse Read	École Polytechnique, France
Marinka Žitnik	Stanford University, USA

PhD Forum Chairs

Tomislav Šmuc	Rudjer Bošković Institute, Croatia
Bernard Ženko	Jožef Stefan Institute, Slovenia

EU Projects Forum Chairs

Petra Kralj Novak	Jožef Stefan Institute, Slovenia
Nada Lavrač	Jožef Stefan Institute, Slovenia

Proceedings Chairs

Jurica Levatić	Jožef Stefan Institute, Slovenia
Gianvito Pio	University of Bari Aldo Moro, Italy

Discovery Challenge Chair

Dino Ienco	IRSTEA - UMR TETIS, France

Sponsorship Chairs

Albert Bifet	Télécom ParisTech, France
Panče Panov	Jožef Stefan Institute, Slovenia

Production and Public Relations Chairs

Dragi Kocev	Jožef Stefan Institute, Slovenia
Nikola Simidjievski	Jožef Stefan Institute, Slovenia

ECML PKDD Steering Committee

Michele Sebag	Université Paris Sud, France
Francesco Bonchi	ISI Foundation, Italy
Albert Bifet	Télécom ParisTech, France
Hendrik Blockeel	KU Leuven, Belgium and Leiden University, The Netherlands
Katharina Morik	University of Dortmund, Germany
Arno Siebes	Utrecht University, The Netherlands
Siegfried Nijssen	LIACS, Leiden University, The Netherlands
Chedy Raïssi	Inria Nancy Grand-Est, France
Rosa Meo	Università di Torino, Italy
Toon Calders	Eindhoven University of Technology, The Netherlands
João Gama	FCUP, University of Porto/LIAAD, INESC Porto L.A., Portugal
Annalisa Appice	University of Bari Aldo Moro, Italy
Indré Žliobaité	University of Helsinki, Finland
Andrea Passerini	University of Trento, Italy
Paolo Frasconi	University of Florence, Italy
Céline Robardet	National Institute of Applied Science in Lyon, France
Jilles Vreeken	Saarland University, Max Planck Institute for Informatics, Germany

Area Chairs

Michael Berthold	Universität Konstanz, Germany
Hendrik Blockeel	KU Leuven, Belgium
Ulf Brefeld	TU Darmstadt, Germany
Toon Calders	Universiteit Antwerpen, Belgium
Bruno Cremilleux	Universite de Caen Normandie, France
Tapio Elomaa	Tampere University of Technology, Finland
Johannes Fuernkranz	TU Darmstadt, Germany
João Gama	FCUP, University of Porto/LIAAD, INESC Porto L.A., Portugal
Alipio Mario Jorge	FCUP, University of Porto/LIAAD, INESC Porto L.A., Portugal
Arno Knobbe	Leiden University, The Netherlands
Stefan Kramer	Johannes Gutenberg University of Mainz, Germany
Ernestina Menasalvas	Universidad Politecnica de Madrid, Spain
Siegfried Nijssen	Leiden University, The Netherlands
Bernhard Pfahringer	University of Waikato, New Zealand
Jesse Read	Ecole Polytechnique, France
Arno Siebes	Universiteit Utrecht, The Netherlands
Carlos Soares	University of Porto, Portugal
Einoshin Suzuki	Kyushu University, Japan
Luis Torgo	University of Porto, Portugal

Matthijs Van Leeuwen	Leiden Institute of Advanced Computer Science, The Netherlands
Indré Žliobaité	University of Helsinki, Finland

Program Committee

Niall Adams
Mohammad Al Hasan
Carlos Alzate
Aijun An
Aris Anagnostopoulos
Fabrizio Angiulli
Annalisa Appice
Ira Assent
Martin Atzmueller
Antonio Bahamonde
Jose Balcazar
Nicola Barberi
Christian Bauckhage
Roberto Bayardo
Martin Becker
Srikanta Bedathur
Jessa Bekker
Vaishak Belle
Andras Benczur
Petr Berka
Michele Berlingerio
Cuissart Bertrand
Marenglen Biba
Silvio Bicciato
Albert Bifet
Paul Blomstedt
Mario Boley
Gianluca Bontempi
Henrik Boström
Marc Boulle
Pavel Brazdil
Dariusz Brzezinski
Rui Camacho
Longbing Cao
Francisco Casacuberta
Peggy Cellier
Loic Cerf
Tania Cerquitelli
Edward Chang
Thierry Charnois
Duen Horng Chau
Keke Chen Wright
Silvia Chiusano
Arthur Choi
Frans Coenen
Roberto Corizzo
Vitor Santos Costa
Boris Cule
Tomaz Curk
James Cussens
Alfredo Cuzzocrea
Claudia d'Amato
Maria Damiani
Stepanova Daria
Jesse Davis
Tijl De Bie
Martine De Cock
Juan del Coz
Anne Denton
Christian Desrosiers
Nicola Di Mauro
Claudia Diamantini
Tom Diethe
Ivica Dimitrovski
Ying Ding
Stephan Doerfel
Carlotta Domeniconi
Frank Dondelinger
Madalina Drugan
Wouter Duivesteijn
Robert Durrant
Ines Dutra
Dora Erdos
Floriana Esposito
Nicola Fanizzi
Fabio Fassetti
Ad Feelders
Stefano Ferilli
Carlos Ferreira
Cesar Ferri
Maurizio Filippone
Asja Fischer
Peter Flach
Eibe Frank
Elisa Fromont
Fabio Fumarola
Esther Galbrun
Patrick Gallinari
Dragan Gamberger
Byron Gao
Paolo Garza
Eric Gaussier
Rainer Gemulla
Konstantinos Georgatzis
Pierre Geurts
Dorota Glowacka
Michael Granitzer
Caglar Gulcehre
Francesco Gullo
Stephan Gunnemann
Tias Guns
Sara Hajian
Maria Halkidi
Jiawei Han
Xiao He
Denis Helic
Daniel Hernandez
Jose Hernandez-Orallo
Thanh Lam Hoang
Arjen Hommersom
Frank Höppner
Tamas Horvath
Andreas Hotho
Yuanhua Huang
Eyke Huellermeier
Dino Ienco
Georgiana Ifrim

Bhattacharya Indrajit
Szymon Jaroszewicz
Klema Jiri
Giuseppe Jurman
Toshihiro Kamishima
Bo Kang
U Kang
Andreas Karwath
George Karypis
Samuel Kaski
Mehdi Kaytoue
Latifur Khan
Frank Klawonn
Dragi Kocev
Levente Kocsis
Yun Sing Koh
Alek Kolcz
Irena Koprinska
Frederic Koriche
Walter Kosters
Lars Kotthoff
Danai Koutra
Georg Krempl
Tomas Krilavicius
Matjaz Kukar
Meelis Kull
Jorma Laaksonen
Nicolas Lachiche
Leo Lahti
Helge Langseth
Thomas Lansdall-Welfare
Christine Largeron
Pedro Larranaga
Silvio Lattanzi
Niklas Lavesson
Nada Lavrac
Florian Lemmerich
Jiuyong Li
Limin Li
Jefrey Lijffijt
Tony Lindgren
Corrado Loglisci
Peter Lucas
Gjorgji Madjarov
Donato Malerba
Giuseppe Manco
Elio Masciari
Andres Masegosa
Wannes Meert
Rosa Meo
Pauli Miettinen
Mehdi Mirza
Dunja Mladenic
Karthika Mohan
Anna Monreale
Joao Moreira
Mohamed Nadif
Ndapa Nakashole
Mirco Nanni
Amedeo Napoli
Sriraam Natarajan
Benjamin Nguyen
Thomas Nielsen
Xia Ning
Kjetil Norvag
Eirini Ntoutsi
Andreas Nuernberger
Francesco Orsini
Nikunj Oza
Pance Panov
Apostolos Papadopoulos
Evangelos Papalexakis
Panagiotis Papapetrou
Ioannis Partalas
Gabriella Pasi
Andrea Passerini
Dino Pedreschi
Jaakko Peltonen
Jing Peng
Ruggero Pensa
Nico Piatkowksi
Andrea Pietracaprina
Gianvito Pio
Susanna Pirttikangas
Marc Plantevit
Pascal Poncelet
Miguel Prada
L. A. Prashanth
Philippe Preux
Kai Puolamaki
Buyue Qian
Chedy Raissi
Jan Ramon
Huzefa Rangwala
Zbigniew Ras
Chotirat Ratanamahatana
Jan Rauch
Chiara Renso
Achim Rettinger
Fabrizio Riguzzi
Matteo Riondato
Celine Robardet
Pedro Rodrigues
Juan Rodriguez
Fabrice Rossi
Celine Rouveirol
Stefan Rueping
Salvatore Ruggieri
Yvan Saeys
Alan Said
Lorenza Saitta
Ansaf Salleb-Aouissi
Claudio Sartori
Christoph Schommer
Matthias Schubert
Konstantinos Sechidis
Giovanni Semeraro
Sohan Seth
Vinay Setty
Junming Shao
Nikola Simidjievski
Sameer Singh
Andrzej Skowron
Dominik Slezak
Kevin Small
Tomislav Smuc
Yangqiu Song
Mauro Sozio
Papadimitriou Spiros
Jerzy Stefanowski
Gerd Stumme
Mahito Sugiyama
Mika Sulkava
Stephen Swift
Sandor Szedmak

Andrea Tagarelli
Domenico Talia
Letizia Tanca
Jovan Tanevski
Nikolaj Tatti
Maguelonne Teisseire
Aika Terada
Georgios Theocharous
Hannu Toivonen
Roberto Trasarti
Volker Tresp
Isaac Triguero
Panagiotis Tsaparas
Karl Tuyls
Niall Twomey
Nikolaos Tziortziotis
Theodoros Tzouramanis
Antti Ukkonen
Jan Van Haaren
Martijn Van Otterlo
Iraklis Varlamis
Julien Velcin
Shankar Vembu
Deepak Venugopal
Vassilios Verykios
Ricardo Vigário
Herna Viktor
Christel Vrain
Jilles Vreeken
Willem Waegeman
Jianyong Wang
Ding Wei
Cheng Weiwei
Zheng Wen
Joerg Wicker
Marco Wiering
Pengtao Xie
Makoto Yamada
Philip Yu
Bianca Zadrozny
Gerson Zaverucha
Filip Zelezny
Bernard Zenko
Junping Zhang
Min-Ling Zhang
Shichao Zhang
Ying Zhao
Mingjun Zhong
Arthur Zimek
Albrecht Zimmermann
Marinka Zitnik

Additional Reviewers

Anes Bendimerad
Asim Karim
Antonis Matakos
Anderson Nascimento
Andrea Pagliarani
Ricardo Ñanculef
Caio Corro
Emanuele Frandi
Emanuele Rabosio
Fábio Pinto
Fernando Martinez-Pluned
Francesca Pratesi
Fangbo Tao
Kata Gabor
Golnoosh Farnadi
Tom Hanika
Heri Ramampiaro
Vasileios Iosifidis
Jeremiah Deng
Johannes Jurgovsky
Kemilly Dearo
Mark Kibanov
Konstantin Ziegler
Ling Luo
Liyuan Liu
Lorenzo Gabrielli
Marcos de Paula
Martin Ringsquandl
Letizia Milli
Christian Poelitz
Riccardo Guidotti
Andreas Schmidt
Shi Zhi
Yujia Shen
Thomas Niebler
Albin Zehe

Sponsors

Gold Sponsors

Deutsche Post DHL Group	http://www.dpdhl.com/
Google	https://research.google.com/

Silver Sponsors

AGT	http://www.agtinternational.com/
ASML	https://www.workingatasml.com/
Deloitte	https://www2.deloitte.com/global/en.html
NEC Europe Ltd.	http://www.neclab.eu/
Siemens	https://www.siemens.com/

Bronze Sponsors

Cambridge University Press	http://www.cambridge.org/wm-ecommerce-web/academic/landingPage/KDD17
IEEE/CAA Journal of Automatica Sinica	http://www.ieee-jas.org/

Awards Sponsors

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

Lanyards Sponsor

KNIME	http://www.knime.org/

Publishing Partner and Sponsor

Springer	http://www.springer.com/gp/

PhD Forum Sponsor

IBM Research	http://researchweb.watson.ibm.com/

Invited Talk Sponsors

EurAi	https://www.eurai.org/
GrabIT	https://www.grabit.mk/

Invited Talks Abstracts

Towards End-to-End Learning and Optimization

Frank Hutter

University of Freiburg

Abstract. Deep learning has recently helped AI systems to achieve human-level performance in several domains, including speech recognition, object classification, and playing several types of games. The major benefit of deep learning is that it enables end-to-end learning of representations of the data on several levels of abstraction. However, the overall network architecture and the learning algorithms' sensitive hyperparameters still need to be set manually by human experts. In this talk, I will discuss extensions of Bayesian optimization for handling this problem effectively, thereby paving the way to fully automated end-to-end learning. I will focus on speeding up Bayesian optimization by reasoning over data subsets and initial learning curves, sometimes resulting in 100-fold speedups in finding good hyperparameter settings. I will also show competition-winning practical systems for automated machine learning (AutoML) and briefly show related applications to the end-to-end optimization of algorithms for solving hard combinatorial problems.

Bio. Frank Hutter is an Emmy Noether Research Group Lead (eq. Asst. Prof.) at the Computer Science Department of the University of Freiburg (Germany). He received his PhD from the University of British Columbia (2009). Frank's main research interests span artificial intelligence, machine learning, combinatorial optimization, and automated algorithm design. He received a doctoral dissertation award from the Canadian Artificial Intelligence Association and, with his coauthors, several best paper awards (including from JAIR and IJCAI) and prizes in international competitions on machine learning, SAT solving, and AI planning. In 2016 he received an ERC Starting Grant for a project on automating deep learning based on Bayesian optimization, Bayesian neural networks, and deep reinforcement learning.

Frontiers in Recurrent Neural Network Research

Alex Graves

Google DeepMind, UK

Abstract. In the last few years, recurrent neural networks (RNNs) have become the Swiss army knife of sequence processing for machine learning. Problems involving long and complex data streams, such as speech recognition, machine translation and reinforcement learning from raw video are now routinely tackled with RNNs. However, significant limitations still exist for such systems, such as their ability to retain large amounts of information in memory, and the challenges of gradient-based training on very long sequences. My talk will review some of the new architectures and training strategies currently being developed to extend the frontiers of this exciting field.

Bio. Alex Graves completed a BSc in Theoretical Physics at the University of Edinburgh, Part III Maths at the University of Cambridge, a PhD in artificial intelligence at IDSIA with Jürgen Schmidhuber, followed by postdocs at the Technical University of Munich and with Geoff Hinton at the University of Toronto. He is now a research scientist at DeepMind. His contributions include the Connectionist Temporal Classification algorithm for sequence labelling (now widely used for commercial speech and handwriting recognition), stochastic gradient variational inference, and the Neural Turing Machine/Differentiable Neural Computer architectures.

Using Networks to Link Genotype to Phenotype

John Quackenbush

Dana-Farber Cancer Institute, Harvard TH Chan School of Public Health

Abstract. We know that genotype influences phenotype, but aside from a few highly penetrant Mendelian disorders, the link between genotype and phenotype is not well understood. We have used gene expression and genetic data to explore gene regulatory networks, to study phenotypic state transitions, and to analyze the connections between genotype and phenotype. I will describe how networks and their structure provide unique insight into how small effect variants influence phenotype.

Bio. John Quackenbush received his PhD in theoretical particle physics from UCLA in 1990. Following a postdoctoral fellowship in experimental high-energy physics, he received an NIH research award to work on the Human Genome Project and helped map chromosome 11 and sequence chromosomes 21 and 4. After four years of working in genomics and computational biology at the Salk Institute and then at Stanford University, John joined The Institute for Genomic Research (TIGR), pioneering microarray expression technologies and analytical methods. In 2005, he joined the Dana-Farber Cancer Institute and the Harvard T. H. Chan School of Public Health, where he uses computational and systems biology methods to explore the complexities of human disease, including cancer. In 2011, he cofounded Genospace, a precision-medicine software company acquired by Hospital Corporation of America in 2017. John's many awards include recognition as a White House Champion of Change for making genomic data useful and widely accessible.

Multi-target Prediction via Low-Rank Embeddings

Inderjit Dhillon

University of Texas at Austin

Abstract. Linear prediction methods, such as linear regression and classification, form the bread-and-butter of modern machine learning. The classical scenario is the presence of data with multiple features and a single target variable. However, there are many recent scenarios where there are multiple target variables. For example, recommender systems, predicting bid words for a web page (where each bid word acts as a target variable), or predicting diseases linked to a gene. In many of these scenarios, the target variables might themselves be associated with features. In these scenarios, bilinear and nonlinear prediction via low-rank embeddings have been shown to be extremely powerful. The low-rank embeddings serve a dual purpose: (i) they enable tractable computation even in the face of millions of data points as well as target variables, and (ii) they exploit correlations among the target variables, even when there are many missing observations. We illustrate our methodology on various modern machine learning problems: recommender systems, multi-label learning and inductive matrix completion, and present results on some standard benchmarks as well as an application that involves prediction of gene-disease associations.

Bio. Inderjit Dhillon is the Gottesman Family Centennial Professor of Computer Science and Mathematics at UT Austin, where he is also the Director of the ICES Center for Big Data Analytics. Currently he is on leave from UT Austin and works as Amazon Fellow at A9/Amazon, where he is developing and deploying state-of-the-art machine learning methods for Amazon search. His main research interests are in big data, machine learning, network analysis, linear algebra and optimization. He received his B. Tech. degree from IIT Bombay, and Ph.D. from UC Berkeley. Inderjit has received several awards, including the ICES Distinguished Research Award, the SIAM Outstanding Paper Prize, the Moncrief Grand Challenge Award, the SIAM Linear Algebra Prize, the University Research Excellence Award, and the NSF Career Award. He has published over 175 journal and conference papers, and has served on the Editorial Board of the Journal of Machine Learning Research, the IEEE Transactions of Pattern Analysis and Machine Intelligence, Foundations and Trends in Machine Learning and the SIAM Journal for Matrix Analysis and Applications. Inderjit is an ACM Fellow, an IEEE Fellow, a SIAM Fellow and an AAAS Fellow.

Enabling a Smarter Planet with Earth Observation

Pierre-Philippe Mathieu

ESA/ESRIN, EO Science, Applications and New Technologies

Abstract. Nowadays, teams of researchers around the world can easily access a wide range of open data across disciplines and remotely process them on the Cloud, combining them with their own data to generate knowledge, develop information products for societal applications, and tackle complex integrative complex problems that could not be addressed a few years ago. Such rapid exchange of digital data is fostering a new world of data-intensive research, characterized by openness, transparency, and scrutiny and traceability of results, access to large volume of complex data, availability of community open tools, unprecedented level of computing power, and new collaboration among researchers and new actors such as citizen scientists. The EO scientific community is now facing the challenge of responding to this new paradigm in science 2.0 in order to make the most of the large volume of complex and diverse data delivered by the new generation of EO missions, and in particular the Sentinels. In this context, ESA is supporting a variety of activities in partnership with research communities to ease the transition and make the most of the data. These include the generation of new open tools and exploitation platforms, exploring new ways to disseminate data, building new partnership with citizen scientists, and training the new generation of data scientists. The talk will give a brief overview of some of ESA activities aiming to facilitate the exploitation of large amounts of data from EO missions in a collaborative, cross-disciplinary, and open way, for uses ranging from science to applications and education.

Bio. Pierre-Philippe Mathieu is an Earth Observation Data Scientist at the European Space Agency in ESRIN (Frascati, Italy). He has spent 20+ years working in the field of environmental and ocean modelling, weather risk management and remote sensing. He has a degree in mechanical engineering and an M.Sc. from the University of Liege (Belgium), a Ph.D. in oceanography from the University of Louvain (Belgium), and a Management degree from the University of Reading Business School (UK).

Automatic Understanding of the Visual World

Cordelia Schmid

Inria

Abstract. One of the central problems of artificial intelligence is machine perception, i.e., the ability to understand the visual world based on input from sensors, such as cameras. Computer vision is the area which analyzes visual input. In this talk, I will present recent progress in visual understanding. It is for the most part due to the design of robust visual representations and learned models capturing the variability of the visual world based on state-of-the-art machine learning techniques, including convolutional neural networks. Progress has resulted in technology for a variety of applications. I will present in particular results for human action recognition.

Bio. Cordelia Schmid holds an M.S. degree in Computer Science from the University of Karlsruhe and a Doctorate, also in Computer Science, from the Institut National Polytechnique de Grenoble (INPG). Her doctoral thesis received the best thesis award from INPG in 1996. Dr. Schmid was a postdoctoral research assistant in the Robotics Research Group of Oxford University in 1996–1997. Since 1997 she has held a permanent research position at INRIA Grenoble Rhone-Alpes, where she is a research director and directs an INRIA team. Dr. Schmid has been an Associate Editor for IEEE PAMI (2001–2005) and for IJCV (2004–2012), editor-in-chief for IJCV (2013—), a program chair of IEEE CVPR 2005 and ECCV 2012 as well as a general chair of IEEE CVPR 2015 and ECCV 2020. In 2006, 2014 and 2016, she was awarded the Longuet-Higgins prize for fundamental contributions in computer vision that have withstood the test of time. She is a fellow of IEEE. She was awarded an ERC advanced grant in 2013, the Humboldt research award in 2015 and the Inria & French Academy of Science Grand Prix in 2016. She was elected to the German National Academy of Sciences, Leopoldina, in 2017.

Abstracts of Journal Track Articles

A Constrained l1 Minimization Approach for Estimating Multiple Sparse Gaussian or Nonparanormal Graphical Models

Beilun Wang, Ritambhara Singh, Yanjun Qi
Machine Learning
http://doi.org/10.1007/s10994-017-5635-7

Identifying context-specific entity networks from aggregated data is an important task, arising often in bioinformatics and neuroimaging applications. Computationally, this task can be formulated as jointly estimating multiple different, but related, sparse undirected graphical models (UGM) from aggregated samples across several contexts. Previous joint-UGM studies have mostly focused on sparse Gaussian graphical models (sGGMs) and can't identify context-specific edge patterns directly. We, therefore, propose a novel approach, SIMULE (detecting Shared and Individual parts of MULtiple graphs Explicitly) to learn multi-UGM via a constrained l1 minimization. SIMULE automatically infers both specific edge patterns that are unique to each context and shared interactions preserved among all the contexts. Through the l1 constrained formulation, this problem is cast as multiple independent subtasks of linear programming that can be solved efficiently in parallel. In addition to Gaussian data, SIMULE can also handle multivariate Nonparanormal data that greatly relaxes the normality assumption that many real-world applications do not follow. We provide a novel theoretical proof showing that SIMULE achieves a consistent result at the rate O (log(Kp)/ntot). On multiple synthetic datasets and two biomedical datasets, SIMULE shows significant improvement over state-of-the-art multi-sGGM and single-UGM baselines (SIMULE implementation and the used datasets @https://github.com/QData/SIMULE).

Adaptive Random Forests for Evolving Data Stream Classification

Heitor M. Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabrício Enembreck, Bernhard Pfharinger, Geoff Holmes, Talel Abdessalem
Machine Learning
http://doi.org/10.1007/s10994-017-5642-8

Random forests is currently one of the most used machine learning algorithms in the non-streaming (batch) setting. This preference is attributable to its high learning performance and low demands with respect to input preparation and hyper-parameter tuning. However, in the challenging context of evolving data streams, there is no random forests algorithm that can be considered state-of-the-art in comparison to bagging and boosting based algorithms. In this work, we present the adaptive random forest (ARF) algorithm for classification of evolving data streams. In contrast to previous attempts of replicating random forests for data stream learning, ARF includes an

effective resampling method and adaptive operators that can cope with different types of concept drifts without complex optimizations for different data sets. We present experiments with a parallel implementation of ARF which has no degradation in terms of classification performance in comparison to a serial implementation, since trees and adaptive operators are independent from one another. Finally, we compare ARF with state-of-the-art algorithms in a traditional test-then-train evaluation and a novel delayed labelling evaluation, and show that ARF is accurate and uses a feasible amount of resources.

An Expressive Dissimilarity Measure for Relational Clustering Using Neighbourhood Trees

Sebastijan Dumančić, Hendrik Blockeel
Machine Learning
http://doi.org/10.1007/s10994-017-5644-6

Clustering is an underspecified task: there are no universal criteria for what makes a good clustering. This is especially true for relational data, where similarity can be based on the features of individuals, the relationships between them, or a mix of both. Existing methods for relational clustering have strong and often implicit biases in this respect. In this paper, we introduce a novel dissimilarity measure for relational data. It is the first approach to incorporate a wide variety of types of similarity, including similarity of attributes, similarity of relational context, and proximity in a hypergraph. We experimentally evaluate the proposed dissimilarity measure on both clustering and classification tasks using data sets of very different types. Considering the quality of the obtained clustering, the experiments demonstrate that (a) using this dissimilarity in standard clustering methods consistently gives good results, whereas other measures work well only on data sets that match their bias; and (b) on most data sets, the novel dissimilarity outperforms even the best among the existing ones. On the classification tasks, the proposed method outperforms the competitors on the majority of data sets, often by a large margin. Moreover, we show that learning the appropriate bias in an unsupervised way is a very challenging task, and that the existing methods offer a marginal gain compared to the proposed similarity method, and can even hurt performance. Finally, we show that the asymptotic complexity of the proposed dissimilarity measure is similar to the existing state-of-the-art approaches. The results confirm that the proposed dissimilarity measure is indeed versatile enough to capture relevant information, regardless of whether that comes from the attributes of vertices, their proximity, or connectedness of vertices, even without parameter tuning.

Constraint-Based Clustering Selection

Toon Van Craenendonck, Hendrik Blockeel
Machine Learning
http://doi.org/10.1007/s10994-017-5643-7

Clustering requires the user to define a distance metric, select a clustering algorithm, and set the hyperparameters of that algorithm. Getting these right, so that a clustering is obtained that meets the users subjective criteria, can be difficult and tedious. Semi-supervised clustering methods make this easier by letting the user provide must-link or cannot-link constraints. These are then used to automatically tune the similarity measure and/or the optimization criterion. In this paper, we investigate a complementary way of using the constraints: they are used to select an unsupervised clustering method and tune its hyperparameters. It turns out that this very simple approach outperforms all existing semi-supervised methods. This implies that choosing the right algorithm and hyperparameter values is more important than modifying an individual algorithm to take constraints into account. In addition, the proposed approach allows for active constraint selection in a more effective manner than other methods.

Cost-Sensitive Label Embedding for Multi-label Classification

Kuan-Hao Huang, Hsuan-Tien Lin
Machine Learning
http://doi.org/10.1007/s10994-017-5659-z

Label embedding (LE) is an important family of multi-label classification algorithms that digest the label information jointly for better performance. Different real-world applications evaluate performance by different cost functions of interest. Current LE algorithms often aim to optimize one specific cost function, but they can suffer from bad performance with respect to other cost functions. In this paper, we resolve the performance issue by proposing a novel cost-sensitive LE algorithm that takes the cost function of interest into account. The proposed algorithm, cost-sensitive label embedding with multidimensional scaling (CLEMS), approximates the cost information with the distances of the embedded vectors by using the classic multidimensional scaling approach for manifold learning. CLEMS is able to deal with both symmetric and asymmetric cost functions, and effectively makes cost-sensitive decisions by nearest-neighbor decoding within the embedded vectors. We derive theoretical results that justify how CLEMS achieves the desired cost-sensitivity. Furthermore, extensive experimental results demonstrate that CLEMS is significantly better than a wide spectrum of existing LE algorithms and state-of-the-art cost-sensitive algorithms across different cost functions.

Efficient Parameter Learning of Bayesian Network Classifiers

Nayyar A. Zaidi, Geoffrey I. Webb, Mark J. Carman,
François Petitjean, Wray Buntine, Mike Hynes, Hans De Sterck
Machine Learning
http://doi.org/10.1007/s10994-016-5619-z

Recent advances have demonstrated substantial benefits from learning with both generative and discriminative parameters. On the one hand, generative approaches address the estimation of the parameters of the joint distribution—P(y, x), which for most network types is very computationally efficient (a notable exception to this are Markov networks) and on the other hand, discriminative approaches address the estimation of the parameters of the posterior distribution—and, are more effective for classification, since they fit P(y—x) directly. However, discriminative approaches are less computationally efficient as the normalization factor in the conditional log-likelihood precludes the derivation of closed-form estimation of parameters. This paper introduces a new discriminative parameter learning method for Bayesian network classifiers that combines in an elegant fashion parameters learned using both generative and discriminative methods. The proposed method is discriminative in nature, but uses estimates of generative probabilities to speed-up the optimization process. A second contribution is to propose a simple framework to characterize the parameter learning task for Bayesian network classifiers. We conduct an extensive set of experiments on 72 standard datasets and demonstrate that our proposed discriminative parameterization provides an efficient alternative to other state-of-the-art parameterizations.

Gaussian Conditional Random Fields Extended for Directed Graphs

Tijana Vujicic, Jesse Glass, Fang Zhou, Zoran Obradovic
Machine Learning
http://doi.org/10.1007/s10994-016-5611-7

For many real-world applications, structured regression is commonly used for predicting output variables that have some internal structure. Gaussian conditional random fields (GCRF) are a widely used type of structured regression model that incorporates the outputs of unstructured predictors and the correlation between objects in order to achieve higher accuracy. However, applications of this model are limited to objects that are symmetrically correlated, while interaction between objects is asymmetric in many cases. In this work we propose a new model, called Directed Gaussian conditional random fields (DirGCRF), which extends GCRF to allow modeling asymmetric relationships (e.g. friendship, influence, love, solidarity, etc.). The DirGCRF models the response variable as a function of both the outputs of unstructured predictors and the asymmetric structure. The effectiveness of the proposed model is characterized on six types of synthetic datasets and four real-world applications where DirGCRF was consistently more accurate than the standard GCRF model and baseline unstructured models.

Generalized Exploration in Policy Search

Herke van Hoof, Daniel Tanneberg, Jan Peters
Machine Learning
http://doi.org/10.1007/s10994-017-5657-1

To learn control policies in unknown environments, learning agents need to explore by trying actions deemed suboptimal. In prior work, such exploration is performed by either perturbing the actions at each time-step independently, or by perturbing policy parameters over an entire episode. Since both of these strategies have certain advantages, a more balanced trade-off could be beneficial. We introduce a unifying view on step-based and episode-based exploration that allows for such balanced trade-offs. This trade-off strategy can be used with various reinforcement learning algorithms. In this paper, we study this generalized exploration strategy in a policy gradient method and in relative entropy policy search. We evaluate the exploration strategy on four dynamical systems and compare the results to the established step-based and episode-based exploration strategies. Our results show that a more balanced trade-off can yield faster learning and better final policies, and illustrate some of the effects that cause these performance differences.

Graph-Based Predictable Feature Analysis

Björn Weghenkel, Asja Fischer, Laurenz Wiskott
Machine Learning
http://doi.org/10.1007/s10994-017-5632-x

We propose graph-based predictable feature analysis (GPFA), a new method for unsupervised learning of predictable features from high-dimensional time series, where high predictability is understood very generically as low variance in the distribution of the next data point given the previous ones. We show how this measure of predictability can be understood in terms of graph embedding as well as how it relates to the information-theoretic measure of predictive information in special cases. We confirm the effectiveness of GPFA on different datasets, comparing it to three existing algorithms with similar objectives—namely slow feature analysis, forecastable component analysis, and predictable feature analysis—to which GPFA shows very competitive results.

Group Online Adaptive Learning

Alon Zweig, Gal Chechik
Machine Learning
http://doi.org/10.1007/s10994-017-5661-5

Sharing information among multiple learning agents can accelerate learning. It could be particularly useful if learners operate in continuously changing environments, because

a learner could benefit from previous experience of another learner to adapt to their new environment. Such group-adaptive learning has numerous applications, from predicting financial time-series, through content recommendation systems, to visual understanding for adaptive autonomous agents. Here we address the problem in the context of online adaptive learning. We formally define the learning settings of Group Online Adaptive Learning and derive an algorithm named Shared Online Adaptive Learning (SOAL) to address it. SOAL avoids explicitly modeling changes or their dynamics, and instead shares information continuously. The key idea is that learners share a common small pool of experts, which they can use in a weighted adaptive way. We define group adaptive regret and prove that SOAL maintains known bounds on the adaptive regret obtained for single adaptive learners. Furthermore, it quickly adapts when learning tasks are related to each other. We demonstrate the benefits of the approach for two domains: vision and text. First, in the visual domain, we study a visual navigation task where a robot learns to navigate based on outdoor video scenes. We show how navigation can improve when knowledge from other robots in related scenes is available. Second, in the text domain, we create a new dataset for the task of assigning submitted papers to relevant editors. This is, inherently, an adaptive learning task due to the dynamic nature of research fields evolving in time. We show how learning to assign editors improves when knowledge from other editors is available. Together, these results demonstrate the benefits for sharing information across learners in concurrently changing environments.

Knowledge Elicitation via Sequential Probabilistic Inference for High-Dimensional Prediction

Pedram Daee, Tomi Peltola, Marta Soare, Samuel Kaski
Machine Learning
http://doi.org/10.1007/s10994-017-5651-7

Prediction in a small-sized sample with a large number of covariates, the "small n, large p" problem, is challenging. This setting is encountered in multiple applications, such as in precision medicine, where obtaining additional data can be extremely costly or even impossible, and extensive research effort has recently been dedicated to finding principled solutions for accurate prediction. However, a valuable source of additional information, domain experts, has not yet been efficiently exploited. We formulate knowledge elicitation generally as a probabilistic inference process, where expert knowledge is sequentially queried to improve predictions. In the specific case of sparse linear regression, where we assume the expert has knowledge about the relevance of the covariates, or of values of the regression coefficients, we propose an algorithm and computational approximation for fast and efficient interaction, which sequentially identifies the most informative features on which to query expert knowledge. Evaluations of the proposed method in experiments with simulated and real users show improved prediction accuracy already with a small effort from the expert.

Learning Constraints in Spreadsheets and Tabular Data

Samuel Kolb, Sergey Paramonov, Tias Guns, Luc De Raedt
Machine Learning
http://doi.org/10.1007/s10994-017-5640-x

Spreadsheets, comma separated value files and other tabular data representations are in wide use today. However, writing, maintaining and identifying good formulas for tabular data and spreadsheets can be time-consuming and error-prone. We investigate the automatic learning of constraints (formulas and relations) in raw tabular data in an unsupervised way. We represent common spreadsheet formulas and relations through predicates and expressions whose arguments must satisfy the inherent properties of the constraint. The challenge is to automatically infer the set of constraints present in the data, without labeled examples or user feedback. We propose a two-stage generate and test method where the first stage uses constraint solving techniques to efficiently reduce the number of candidates, based on the predicate signatures. Our approach takes inspiration from inductive logic programming, constraint learning and constraint satisfaction. We show that we are able to accurately discover constraints in spreadsheets from various sources.

Learning Deep Kernels in the Space of Dot Product Polynomials

Michele Donini, Fabio Aiolli
Machine Learning
http://doi.org/10.1007/s10994-016-5590-8

Recent literature has shown the merits of having deep representations in the context of neural networks. An emerging challenge in kernel learning is the definition of similar deep representations. In this paper, we propose a general methodology to define a hierarchy of base kernels with increasing expressiveness and combine them via multiple kernel learning (MKL) with the aim to generate overall deeper kernels. As a leading example, this methodology is applied to learning the kernel in the space of Dot-Product Polynomials (DPPs), that is a positive combination of homogeneous polynomial kernels (HPKs). We show theoretical properties about the expressiveness of HPKs that make their combination empirically very effective. This can also be seen as learning the coefficients of the Maclaurin expansion of any definite positive dot product kernel thus making our proposed method generally applicable. We empirically show the merits of our approach comparing the effectiveness of the kernel generated by our method against baseline kernels (including homogeneous and non homogeneous polynomials, RBF, etc…) and against another hierarchical approach on several benchmark datasets.

Offline Reinforcement Learning with Task Hierarchies

Devin Schwab, Soumya Ray
Machine Learning
http://doi.org/10.1007/s10994-017-5650-8

In this work, we build upon the observation that offline reinforcement learning (RL) is synergistic with task hierarchies that decompose large Markov decision processes (MDPs). Task hierarchies can allow more efficient sample collection from large MDPs, while offline algorithms can learn better policies than the so-called "recursively optimal" or even hierarchically optimal policies learned by standard hierarchical RL algorithms. To enable this synergy, we study sample collection strategies for offline RL that are consistent with a provided task hierarchy while still providing good exploration of the state-action space. We show that naïve extensions of uniform random sampling do not work well in this case and design a strategy that has provably good convergence properties. We also augment the initial set of samples using additional information from the task hierarchy, such as state abstraction. We use the augmented set of samples to learn a policy offline. Given a capable offline RL algorithm, this policy is then guaranteed to have a value greater than or equal to the value of the hierarchically optimal policy. We evaluate our approach on several domains and show that samples generated using a task hierarchy with a suitable strategy allow significantly more sample-efficient convergence than standard offline RL. Further, our approach also shows more sample-efficient convergence to policies with value greater than or equal to hierarchically optimal policies found through an online hierarchical RL approach.

Preserving Differential Privacy in Convolutional Deep Belief Networks

NhatHai Phan, Xintao Wu, Dejing Dou
Machine Learning
http://doi.org/10.1007/s10994-017-5656-2

The remarkable development of deep learning in medicine and healthcare domain presents obvious privacy issues, when deep neural networks are built on users' personal and highly sensitive data, e.g., clinical records, user profiles, biomedical images, etc. However, only a few scientific studies on preserving privacy in deep learning have been conducted. In this paper, we focus on developing a private convolutional deep belief network (pCDBN), which essentially is a convolutional deep belief network (CDBN) under differential privacy. Our main idea of enforcing ε-differential privacy is to leverage the functional mechanism to perturb the energy-based objective functions of traditional CDBNs, rather than their results. One key contribution of this work is that we propose the use of Chebyshev expansion to derive the approximate polynomial representation of objective functions. Our theoretical analysis shows that we can further derive the sensitivity and error bounds of the approximate polynomial representation. As a result, preserving differential privacy in CDBNs is feasible. We applied our model

in a health social network, i.e., YesiWell data, and in a handwriting digit dataset, i.e., MNIST data, for human behavior prediction, human behavior classification, and handwriting digit recognition tasks. Theoretical analysis and rigorous experimental evaluations show that the pCDBN is highly effective. It significantly outperforms existing solutions.

Robust Regression Using Biased Objectives

Matthew J. Holland, Kazushi Ikeda
Machine Learning
http://doi.org/10.1007/s10994-017-5653-5

For the regression task in a non-parametric setting, designing the objective function to be minimized by the learner is a critical task. In this paper we propose a principled method for constructing and minimizing robust losses, which are resilient to errant observations even under small samples. Existing proposals typically utilize very strong estimates of the true risk, but in doing so require a priori information that is not available in practice. As we abandon direct approximation of the risk, this lets us enjoy substantial gains in stability at a tolerable price in terms of bias, all while circumventing the computational issues of existing procedures. We analyze existence and convergence conditions, provide practical computational routines, and also show empirically that the proposed method realizes superior robustness over wide data classes with no prior knowledge assumptions.

Sparse Probit Linear Mixed Model

Stephan Mandt, Florian Wenzel, Shinichi Nakajima,
John Cunningham, Christoph Lippert, Marius Kloft
Machine Learning
http://doi.org/10.1007/s10994-017-5652-6

Linear mixed models (LMMs) are important tools in statistical genetics. When used for feature selection, they allow to find a sparse set of genetic traits that best predict a continuous phenotype of interest, while simultaneously correcting for various confounding factors such as age, ethnicity and population structure. Formulated as models for linear regression, LMMs have been restricted to continuous phenotypes. We introduce the sparse probit linear mixed model (Probit-LMM), where we generalize the LMM modeling paradigm to binary phenotypes. As a technical challenge, the model no longer possesses a closed-form likelihood function. In this paper, we present a scalable approximate inference algorithm that lets us fit the model to high-dimensional data sets. We show on three real-world examples from different domains that in the setup of binary labels, our algorithm leads to better prediction accuracies and also selects features which show less correlation with the confounding factors.

Varying-Coefficient Models for Geospatial Transfer Learning

Matthias Bussas, Christoph Sawade, Nicolas Kühn, Tobias Scheffer, Niels Landwehr
Machine Learning
http://doi.org/10.1007/s10994-017-5639-3

We study prediction problems in which the conditional distribution of the output given the input varies as a function of task variables which, in our applications, represent space and time. In varying-coefficient models, the coefficients of this conditional are allowed to change smoothly in space and time; the strength of the correlations between neighboring points is determined by the data. This is achieved by placing a Gaussian process (GP) prior on the coefficients. Bayesian inference in varying-coefficient models is generally intractable. We show that with an isotropic GP prior, inference in varying-coefficient models resolves to standard inference for a GP that can be solved efficiently. MAP inference in this model resolves to multitask learning using task and instance kernels. We clarify the relationship between varying-coefficient models and the hierarchical Bayesian multitask model and show that inference for hierarchical Bayesian multitask models can be carried out efficiently using graph-Laplacian kernels. We explore the model empirically for the problems of predicting rent and real-estate prices, and predicting the ground motion during seismic events. We find that varying-coefficient models with GP priors excel at predicting rents and real-estate prices. The ground-motion model predicts seismic hazards in the State of California more accurately than the previous state of the art.

Vine Copulas for Mixed Data: Multi-view Clustering for Mixed Data Beyond Meta-Gaussian Dependencies

Lavanya Sita Tekumalla, Vaibhav Rajan, Chiranjib Bhattacharyya
Machine Learning
http://doi.org/10.1007/s10994-016-5624-2

Copulas enable flexible parameterization of multivariate distributions in terms of constituent marginals and dependence families. Vine copulas, hierarchical collections of bivariate copulas, can model a wide variety of dependencies in multivariate data including asymmetric and tail dependencies which the more widely used Gaussian copulas, used in Meta-Gaussian distributions, cannot. However, current inference algorithms for vines cannot fit data with mixed—a combination of continuous, binary and ordinal—features that are common in many domains. We design a new inference algorithm to fit vines on mixed data thereby extending their use to several applications. We illustrate our algorithm by developing a dependency-seeking multi-view clustering model based on Dirichlet Process mixture of vines that generalizes previous models to arbitrary dependencies as well as to mixed marginals. Empirical results on synthetic and real datasets demonstrate the performance on clustering single-view and multi-view data with asymmetric and tail dependencies and with mixed marginals.

Weightless Neural Networks for Open Set Recognition

Douglas O. Cardoso, João Gama, Felipe M. G. França
Machine Learning
http://doi.org/10.1007/s10994-017-5646-4

Open set recognition is a classification-like task.It is accomplished not only by the identification of observations which belong to targeted classes (i.e., the classes among those represented in the training sample which should be later recognized) but also by the rejection of inputs from other classes in the problem domain. The need for proper handling of elements of classes beyond those of interest is frequently ignored, even in works found in the literature. This leads to the improper development of learning systems, which may obtain misleading results when evaluated in their test beds, consequently failing to keep the performance level while facing some real challenge. The adaptation of a classifier for open set recognition is not always possible: the probabilistic premises most of them are built upon are not valid in a open-set setting. Still, this paper details how this was realized for WiSARD a weightless artificial neural network model. Such achievement was based on an elaborate distance-like computation this model provides and the definition of rejection thresholds during training. The proposed methodology was tested through a collection of experiments, with distinct backgrounds and goals. The results obtained confirm the usefulness of this tool for open set recognition.

Classification of High-Dimensional Evolving Data Streams via a Resource-Efficient Online Ensemble

Tingting Zhai, Yang Gao, Hao Wang, Longbing Cao
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0500-7

A novel online ensemble strategy, ensemble BPegasos(EBPegasos), is proposed to solve the problems simultaneously caused by concept drifting and the curse of dimensionality in classifying high-dimensional evolving data streams, which has not been addressed in the literature. First, EBPegasos uses BPegasos, an online kernelized SVM-based algorithm, as the component classifier to address the scalability and sparsity of high-dimensional data. Second, EBPegasos takes full advantage of the characteristics of BPegasos to cope with various types of concept drifts. Specifically, EBPegasos constructs diverse component classifiers by controlling the budget size of BPegasos; it also equips each component with a drift detector to monitor and evaluate its performance, and modifies the ensemble structure only when large performance degradation occurs. Such conditional structural modification strategy makes EBPegasos strike a good balance between exploiting and forgetting old knowledge. Lastly, we first prove experimentally that EBPegasos is more effective and resource-efficient than the tree ensembles on high-dimensional data. Then comprehensive experiments on synthetic and real-life datasets also show that EBPegasos can cope with various types

of concept drifts significantly better than the state-of-the-art ensemble frameworks when all ensembles use BPegasos as the base learner.

Differentially Private Nearest Neighbor Classification

Mehmet Emre Gursoy, Ali Inan, Mehmet Ercan Nergiz, Yucel Saygin
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0532-z

Instance-based learning, and the k-nearest neighbors algorithm (k-NN) in particular, provide simple yet effective classification algorithms for data mining. Classifiers are often executed on sensitive information such as medical or personal data. Differential privacy has recently emerged as the accepted standard for privacy protection in sensitive data. However, straightforward applications of differential privacy to k-NN classification yield rather inaccurate results. Motivated by this, we develop algorithms to increase the accuracy of private instance-based classification. We first describe the radius neighbors classifier (r-N) and show that its accuracy under differential privacy can be greatly improved by a non-trivial sensitivity analysis. Then, for k-NN classification, we build algorithms that convert k-NN classifiers to r-N classifiers. We experimentally evaluate the accuracy of both classifiers using various datasets. Experiments show that our proposed classifiers significantly outperform baseline private classifiers (i.e., straightforward applications of differential privacy) and executing the classifiers on a dataset published using differential privacy. In addition, the accuracy of our proposed k-NN classifiers are at least comparable to, and in many cases better than, the other differentially private machine learning techniques.

Ensemble-Based Community Detection in Multilayer Networks

Andrea Tagarelli, Alessia Amelio, Francesco Gullo
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0528-8

The problem of community detection in a multilayer network can effectively be addressed by aggregating the community structures separately generated for each network layer, in order to infer a consensus solution for the input network. To this purpose, clustering ensemble methods developed in the data clustering field are naturally of great support. Bringing these methods into a community detection framework would in principle represent a powerful and versatile approach to reach more stable and reliable community structures. Surprisingly, research on consensus community detection is still in its infancy. In this paper, we propose a novel modularity-driven ensemble-based approach to multilayer community detection. A key aspect is that it finds consensus community structures that not only capture prototypical community memberships of nodes, but also preserve the multilayer topology information and optimize the edge connectivity in the consensus via modularity analysis. Empirical

evidence obtained on seven real-world multilayer networks sheds light on the effectiveness and efficiency of our proposed modularity-driven ensemble-based approach, which has shown to outperform state-of-the-art multilayer methods in terms of modularity, silhouette of community memberships, and redundancy assessment criteria, and also in terms of execution times.

Flexible Constrained Sampling with Guarantees for Pattern Mining

Vladimir Dzyuba, Matthijs van Leeuwen, Luc De Raedt
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0501-6

Pattern sampling has been proposed as a potential solution to the infamous pattern explosion. Instead of enumerating all patterns that satisfy the constraints, individual patterns are sampled proportional to a given quality measure. Several sampling algorithms have been proposed, but each of them has its limitations when it comes to (1) flexibility in terms of quality measures and constraints that can be used, and/or (2) guarantees with respect to sampling accuracy. We therefore present Flexics, the first flexible pattern sampler that supports a broad class of quality measures and constraints, while providing strong guarantees regarding sampling accuracy. To achieve this, we leverage the perspective on pattern mining as a constraint satisfaction problem and build upon the latest advances in sampling solutions in SAT as well as existing pattern mining algorithms. Furthermore, the proposed algorithm is applicable to a variety of pattern languages, which allows us to introduce and tackle the novel task of sampling sets of patterns. We introduce and empirically evaluate two variants of Flexics: (1) a generic variant that addresses the well-known itemset sampling task and the novel pattern set sampling task as well as a wide range of expressive constraints within these tasks, and (2) a specialized variant that exploits existing frequent itemset techniques to achieve substantial speed-ups. Experiments show that Flexics is both accurate and efficient, making it a useful tool for pattern-based data exploration.

Identifying Consistent Statements About Numerical Data with Dispersion-Corrected Subgroup Discovery

Mario Boley, Bryan R. Goldsmith, Luca M. Ghiringhelli, Jilles Vreeken
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0520-3

Existing algorithms for subgroup discovery with numerical targets do not optimize the error or target variable dispersion of the groups they find. This often leads to unreliable or inconsistent statements about the data, rendering practical applications, especially in scientific domains, futile. Therefore, we here extend the optimistic estimator framework for optimal subgroup discovery to a new class of objective functions: we show how

tight estimators can be computed efficiently for all functions that are determined by subgroup size (non-decreasing dependence), the subgroup median value, and a dispersion measure around the median (non-increasing dependence). In the important special case when dispersion is measured using the mean absolute deviation from the median, this novel approach yields a linear time algorithm. Empirical evaluation on a wide range of datasets shows that, when used within branch-and-bound search, this approach is highly efficient and indeed discovers subgroups with much smaller errors.

Lagrangian Relaxations for Multiple Network Alignment

Eric Malmi, Sanjay Chawla, Aristides Gionis
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0505-2

We propose a principled approach for the problem of aligning multiple partially overlapping networks. The objective is to map multiple graphs into a single graph while preserving vertex and edge similarities. The problem is inspired by the task of integrating partial views of a family tree (genealogical network) into one unified network, but it also has applications, for example, in social and biological networks. Our approach, called Flan, introduces the idea of generalizing the facility location problem by adding a non-linear term to capture edge similarities and to infer the underlying entity network. The problem is solved using an alternating optimization procedure with a Lagrangian relaxation. Flan has the advantage of being able to leverage prior information on the number of entities, so that when this information is available, Flan is shown to work robustly without the need to use any ground truth data for fine-tuning method parameters. Additionally, we present three multiple-network extensions to an existing state-of-the-art pairwise alignment method called Natalie. Extensive experiments on synthetic, as well as real-world datasets on social networks and genealogical networks, attest to the effectiveness of the proposed approaches which clearly outperform a popular multiple network alignment method called IsoRankN.

Local Community Detection in Multilayer Networks

Roberto Interdonato, Andrea Tagarelli, Dino Ienco,
Arnaud Sallaberry, Pascal Poncelet
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0525-y

The problem of local community detection in graphs refers to the identification of a community that is specific to a query node and relies on limited information about the network structure. Existing approaches for this problem are defined to work in dynamic network scenarios, however they are not designed to deal with complex real-world networks, in which multiple types of connectivity might be considered. In this work, we fill this gap in the literature by introducing the first framework for local community

detection in multilayer networks (ML-LCD). We formalize the ML-LCD optimization problem and provide three definitions of the associated objective function, which correspond to different ways to incorporate within-layer and across-layer topological features. We also exploit our framework to generate multilayer global community structures. We conduct an extensive experimentation using seven real-world multilayer networks, which also includes comparison with state-of-the-art methods for single-layer local community detection and for multilayer global community detection. Results show the significance of our proposed methods in discovering local communities over multiple layers, and also highlight their ability in producing global community structures that are better in modularity than those produced by native global community detection approaches.

Measuring and Moderating Opinion Polarization in Social Networks

Antonis Matakos, Evimaria Terzi, Panayiotis Tsaparas
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0527-9

The polarization of society over controversial social issues has been the subject of study in social sciences for decades (Isenberg in J Personal Soc Psychol 50(6):1141–1151, 1986, Sunstein in J Polit Philos 10(2):175–195, 2002). The widespread usage of online social networks and social media, and the tendency of people to connect and interact with like-minded individuals has only intensified the phenomenon of polarization (Bakshy et al. in Science 348(6239):1130–1132, 2015). In this paper, we consider the problem of measuring and reducing polarization of opinions in a social network. Using a standard opinion formation model (Friedkin and Johnsen in J Math Soc 15(3–4):193–206, 1990), we define the polarization index, which, given a network and the opinions of the individuals in the network, it quantifies the polarization observed in the network. Our measure captures the tendency of opinions to concentrate in network communities, creating echo-chambers. Given this numeric measure of polarization, we then consider the problem of reducing polarization in the network by convincing individuals (e.g., through education, exposure to diverse viewpoints, or incentives) to adopt a more neutral stand towards controversial issues. We formally define the ModerateInternal and ModerateExpressed problems, and we prove that both our problems are NP-hard. By exploiting the linear- algebraic characteristics of the opinion formation model we design polynomial-time algorithms for both problems. Our experiments with real-world datasets demonstrate the validity of our metric, and the efficiency and the effectiveness of our algorithms in practice.

Micro-Review Synthesis for Multi-entity Summarization

Thanh-Son Nguyen, Hady W. Lauw, Panayiotis Tsaparas
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0491-4

Location-based social networks (LBSNs), exemplified by Foursquare, are fast gaining popularity. One important feature of LBSNs is micro-review. Upon check-in at a particular venue, a user may leave a short review (up to 200 characters long), also known as a tip. These tips are an important source of information for others to know more about various aspects of an entity (e.g., restaurant), such as food, waiting time, or service. However, a user is often interested not in one particular entity, but rather in several entities collectively, for instance within a neighborhood or a category. In this paper, we address the problem of summarizing the tips of multiple entities in a collection, by way of synthesizing new micro-reviews that pertain to the collection, rather than to the individual entities per se. We formulate this problem in terms of first finding a representation of the collection, by identifying a number of "aspects" that link common threads across two or more entities within the collection. We express these aspects as dense subgraphs in a graph of sentences derived from the multi-entity corpora. This leads to a formulation of maximal multi-entity quasi-cliques, as well as a heuristic algorithm to find K such quasi-cliques maximizing the coverage over the multi-entity corpora. To synthesize a summary tip for each aspect, we select a small number of sentences from the corresponding quasi-clique, balancing conciseness and representativeness in terms of a facility location problem. Our approach performs well on collections of Foursquare entities based on localities and categories, producing more representative and diverse summaries than the baselines.

MixedTrails: Bayesian Hypothesis Comparison on Heterogeneous Sequential Data

Martin Becker, Florian Lemmerich, Philipp Singer, Markus Strohmaier, Andreas Hotho
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0518-x

Sequential traces of user data are frequently observed online and offline, e.g., as sequences of visited websites or as sequences of locations captured by GPS. However, understanding factors explaining the production of sequence data is a challenging task, especially since the data generation is often not homogeneous. For example, navigation behavior might change in different phases of browsing a website or movement behavior may vary between groups of users. In this work, we tackle this task and propose MixedTrails, a Bayesian approach for comparing the plausibility of hypotheses regarding the generative processes of heterogeneous sequence data. Each hypothesis is derived from existing literature, theory, or intuition and represents a belief about transition probabilities between a set of states that can vary between groups of observed

transitions. For example, when trying to understand human movement in a city and given some data, a hypothesis assuming tourists to be more likely to move towards points of interests than locals can be shown to be more plausible than a hypothesis assuming the opposite. Our approach incorporates such hypotheses as Bayesian priors in a generative mixed transition Markov chain model, and compares their plausibility utilizing Bayes factors. We discuss analytical and approximate inference methods for calculating the marginal likelihoods for Bayes factors, give guidance on interpreting the results, and illustrate our approach with several experiments on synthetic and empirical data from Wikipedia and Flickr. Thus, this work enables a novel kind of analysis for studying sequential data in many application areas.

On Temporal-Constrained Sub-trajectory Cluster Analysis

Nikos Pelekis, Panagiotis Tampakis, Marios Vodas, Christos Doulkeridis, Yannis Theodoridis
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0503-4

Cluster analysis over Moving Object Databases (MODs) is a challenging research topic that has attracted the attention of the mobility data mining community. In this paper, we study the temporal-constrained sub-trajectory cluster analysis problem, where the aim is to discover clusters of sub-trajectories given an ad-hoc, user-specified temporal constraint within the dataset's lifetime. The problem is challenging because: (a) the time window is not known in advance, instead it is specified at query time, and (b) the MOD is continuously updated with new trajectories. Existing solutions first filter the trajectory database according to the temporal constraint, and then apply a clustering algorithm from scratch on the filtered data. However, this approach is extremely inefficient, when considering explorative data analysis where multiple clustering tasks need to be performed over different temporal subsets of the database, while the database is updated with new trajectories. To address this problem, we propose an incremental and scalable solution to the problem, which is built upon a novel indexing structure, called Representative Trajectory Tree (ReTraTree). ReTraTree acts as an effective spatio-temporal partitioning technique; partitions in ReTraTree correspond to groupings of sub-trajectories, which are incrementally maintained and assigned to representative (sub-)trajectories. Due to the proposed organization of sub-trajectories, the problem under study can be efficiently solved as simply as executing a query operator on ReTraTree, while insertion of new trajectories is supported. Our extensive experimental study performed on real and synthetic datasets shows that our approach outperforms a state-of-the-art in-DBMS solution supported by PostgreSQL by orders of magnitude.

Social Regularized von Mises–Fisher Mixture Model for Item Recommendation

Aghiles Salah, Mohamed Nadif
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0499-9

Collaborative filtering (CF) is a widely used technique to guide the users of web applications towards items that might interest them. CF approaches are severely challenged by the characteristics of user-item preference matrices, which are often high dimensional and extremely sparse. Recently, several works have shown that incorporating information from social networks—such as friendship and trust relationships—into traditional CF alleviates the sparsity related issues and yields a better recommendation quality, in most cases. More interestingly, even with comparable performances, social-based CF is more beneficial than traditional CF; the former makes it possible to provide recommendations for cold start users. In this paper, we propose a novel model that leverages information from social networks to improve recommendations. While existing social CF models are based on popular modelling assumptions such as Gaussian or Multinomial, our model builds on the von Mises–Fisher assumption which turns out to be more adequate, than the aforementioned assumptions, for high dimensional sparse data. Setting the estimate of the model parameters under the maximum likelihood approach, we derive a scalable learning algorithm for analyzing data with our model. Empirical results on several real-world datasets provide strong support for the advantages of the proposed model.

The Best Privacy Defense Is a Good Privacy Offense: Obfuscating a Search Engine User's Profile

Jörg Wicker, Stefan Kramer
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-017-0524-z

User privacy on the internet is an important and unsolved problem. So far, no sufficient and comprehensive solution has been proposed that helps a user to protect his or her privacy while using the internet. Data are collected and assembled by numerous service providers. Solutions so far focused on the side of the service providers to store encrypted or transformed data that can be still used for analysis. This has a major flaw, as it relies on the service providers to do this. The user has no chance of actively protecting his or her privacy. In this work, we suggest a new approach, empowering the user to take advantage of the same tool the other side has, namely data mining to produce data which obfuscates the user's profile. We apply this approach to search engine queries and use feedback of the search engines in terms of personalized advertisements in an algorithm similar to reinforcement learning to generate new queries potentially confusing the search engine. We evaluated the approach using a

real-world data set. While evaluation is hard, we achieve results that indicate that it is possible to influence the user's profile that the search engine generates. This shows that it is feasible to defend a user's privacy from a new and more practical perspective.

Tour Recommendation for Groups

Aris Anagnostopoulos, Reem Atassi, Luca Becchetti, Adriano Fazzone, Fabrizio Silvestri
Data Mining and Knowledge Discovery
http://doi.org/10.1007/s10618-016-0477-7

Consider a group of people who are visiting a major touristic city, such as NY, Paris, or Rome. It is reasonable to assume that each member of the group has his or her own interests or preferences about places to visit, which in general may differ from those of other members. Still, people almost always want to hang out together and so the following question naturally arises: What is the best tour that the group could perform together in the city? This problem underpins several challenges, ranging from understanding people's expected attitudes towards potential points of interest, to modeling and providing good and viable solutions. Formulating this problem is challenging because of multiple competing objectives. For example, making the entire group as happy as possible in general conflicts with the objective that no member becomes disappointed. In this paper, we address the algorithmic implications of the above problem, by providing various formulations that take into account the overall group as well as the individual satisfaction and the length of the tour. We then study the computational complexity of these formulations, we provide effective and efficient practical algorithms, and, finally, we evaluate them on datasets constructed from real city data.

Contents – Part I

Ensembles and Meta Learning

Feature Selection and Extraction

Kernel Methods

Networks and Graphs

Neural Networks and Deep Learning

Contents – Part II

Recommendation

Regression

Reinforcement Learning

Subgroup Discovery

Time Series and Streams

Contents – Part III

Applied Data Science Track

Nectar Track