Lecture Notes in Artificial Intelligence 11828

Subseries of Lecture Notes in Computer Science

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

Petra Kralj Novak · Tomislav Šmuc ·
Sašo Džeroski (Eds.)

Discovery Science

22nd International Conference, DS 2019
Split, Croatia, October 28–30, 2019
Proceedings

Springer

Editors
Petra Kralj Novak
Jožef Stefan Institute
Ljubljana, Slovenia

Tomislav Šmuc
Rudjer Bošković Institute
Zagreb, Croatia

Sašo Džeroski
Jožef Stefan Institute
Ljubljana, Slovenia

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-030-33777-3 ISBN 978-3-030-33778-0 (eBook)
https://doi.org/10.1007/978-3-030-33778-0

LNCS Sublibrary: SL7 – Artificial Intelligence

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

Preface

The Discovery Science conference presents a unique combination of latest advances in the development and analysis of methods for discovering scientific knowledge, coming from machine learning, data mining, and intelligent data analysis, with their application in various scientific domains.

The 22nd International Conference on Discovery Science (DS 2019) was held in Split, Croatia, during October 28–30, 2019. This was the first time the conference was organized as a stand-alone event. For its first 20 editions, DS was co-located with the International Conference on Algorithmic Learning Theory (ALT). In 2018 it was co-located with the 24th International Symposium on Methodologies for Intelligent Systems (ISMIS 2018).

DS 2019 received 63 international submissions. Each submission was reviewed by at least three Program Committee (PC) members. The PC decided to accept 21 regular papers and 19 short papers. This resulted in an acceptance rate of 33% for regular papers.

The conference included three keynote talks. Marinka Žitnik (Stanford University) contributed a talk titled "Representation Learning as a New Approach to Biomedical Research," Guido Caldarelli (IMT Lucca and ECLT Venice) gave a presentation titled "The Structure of Financial Networks," and Dino Pedreschi (University of Pisa), contributed a talk titled "Data and Algorithmic Bias: Explaining the Network Effect in Opinion Dynamics and the Training Data Bias in Machine Learning." Abstracts of the invited talks with short biographies of the invited speakers are included in these proceedings.

Besides the presentation of regular and short papers in the main program, the conference offered two new sessions. The "PhD Symposium" gave an opportunity to PhD students at an early stage of their studies to participate in the conference by presenting the topics of and early results from their research and discuss their work and experiences with peers, senior researchers and leading experts working on similar problems. The session titled "Late Breaking Contributions" featured poster and spotlight presentations of very recent research results on topics related to Discovery Science.

We are grateful to Springer for their long-term support, which got even stronger this year. Springer publishes the conference proceedings, as well as a regular special issue of the Machine Learning journal on Discovery Science. The latter offers authors a chance of publishing in this prestigious journal significantly extended and reworked versions of their DS conference papers, while being open to all submissions on DS conference topics.

This year, Springer (LNCS and Machine Learning journal), supported the best student paper awards. For DS 2019, the awardees are Anton Björklund, Andreas Henelius, Emilia Oikarinen, Kimmo Kallonen and Kai Puolamäki (for the paper "Sparse Robust Regression for Explaining Classifiers") and Yannik Klein, Michael

Rapp and Eneldo Loza Mencía (for the paper "Efficient Discovery of Expressive Multi-label Rules Using Relaxed Pruning.") We would like to thank the Best Paper Award committee composed of Dragan Gamberger and Toon Calders for their precious and timely evaluations.

On the program side, we would like to thank all the authors of submitted papers, the PC members and the additional reviewers for their efforts in evaluating the submitted papers, as well as the keynote speakers. On the organization side, we would like to thank all the members of the Organizing Committee: Tomislav Lipić, Ana Vidoš, Matija Piškorec and Ratko Mileta, for the smooth preparation and organization of all conference associated activities. We are also grateful to the people behind EasyChair for developing the conference organization system that proved to be an essential tool in the paper submission and evaluation process, as well as in the preparation of the Springer proceedings.

The DS 2019 conference was organized under the auspices of the Rudjer Bošković Institute in Zagreb. The event was also supported by the Project of the Croatian Center for Excellence in Data Science and Advanced Cooperative Systems. Significant support, especially through human resources, was also provided by the Jožef Stefan Institute from Ljubljana. Finally, we are indebted to all conference participants, who contributed to making this exciting event a worthwhile endeavor for all involved.

October 2019

Petra Kralj Novak
Tomislav Šmuc
Sašo Džeroski

Organization

General Chair

Sašo Džeroski	Jožef Stefan Institute, Slovenia

Program Committee Chairs

Petra Kralj Novak	Jožef Stefan Institute, Slovenia
Tomislav Šmuc	Rudjer Bošković Institute, Croatia

PhD Symposium Chair

Tomislav Lipić	Rudjer Bošković Institute, Croatia

Proceedings Chair

Matija Piškorec	Rudjer Bošković Institute, Croatia

Web and Social Media Chairs

Ratko Mileta	Rudjer Bošković Institute, Croatia
Matija Piškorec	Rudjer Bošković Institute, Croatia

Local Arrangements Chair

Ana Vidoš	Rudjer Bošković Institute, Croatia

Program Committee

Annalisa Appice	University of Bari Aldo Moro, Italy
Martin Atzmueller	Tilburg University, The Netherlands
Viktor Bengs	Paderborn University, Germany
Concha Bielza Lozoya	Universidad Politécnica de Madrid, Spain
Albert Bifet	LTCI, Telecom ParisTech, France
Alberto Cano	Virginia Commonwealth University, USA
Michelangelo Ceci	University of Bari Aldo Moro, Italy
Bruno Cremilleux	University of Caen Normandy, France
Claudia d'Amato	University of Bari Aldo Moro, Italy
Nicola Di Mauro	University of Bari Aldo Moro, Italy
Ivica Dimitrovski	Ss. Cyril and Methodius University in Skopje, North Macedonia

Wouter Duivesteijn	Eindhoven University of Technology, The Netherlands
Lina Fahed	IMT Atlantique, France
Hadi Fanaee	University of Oslo, Norway
Nicola Fanizzi	University of Bari Aldo Moro, Italy
Stefano Ferilli	University of Bari Aldo Moro, Italy
Johannes Fürnkranz	Technische Universität Darmstadt, Germany
Mohamed Gaber	Birmingham City University, UK
João Gama	University of Porto, Portugal
Dragan Gamberger	Rudjer Bošković Institute, Croatia
Makoto Haraguchi	Hokkaido University, Japan
Kouichi Hirata	Kyushu Institute of Technology, Japan
Jaakko Hollmén	Aalto University, Finland
Eyke Huellermeier	Paderborn University, Germany
Alípio Jorge	University of Porto, Portugal
Masahiro Kimura	Ryukoku University, Japan
Dragi Kocev	Jožef Stefan Institute, Slovenia
Stefan Kramer	Johannes Gutenberg University Mainz, Germany
Vincenzo Lagani	Ilia State University, Georgia
Pedro Larranaga	University of Madrid, Spain
Nada Lavrač	Jožef Stefan Institute, Slovenia
Jurica Levatić	Institute for Research in Biomedicine, Spain
Tomislav Lipić	Rudjer Bošković Institute, Croatia
Francesca Alessandra Lisi	University of Bari Aldo Moro, Italy
Gjorgji Madjarov	Ss. Cyril and Methodius University in Skopje, North Macedonia
Giuseppe Manco	Institute for High Performance Computing and Networking, Italy
Sanda Martinčić-Ipšić	University of Rijeka, Croatia
Elio Masciari	Institute for High Performance Computing and Networking, Italy
Anna Monreale	University of Pisa, Italy
Siegfried Nijssen	Université Catholique de Louvain, Belgium
Rita P. Ribeiro	University of Porto, Portugal
Panče Panov	Jožef Stefan Institute, Slovenia
Ruggero G. Pensa	University of Torino, Italy
Bernhard Pfahringer	University of Waikato, New Zealand
Gianvito Pio	University of Bari Aldo Moro, Italy
Pascal Poncelet	LIRMM Montpellier, France
Jan Ramon	French Research Institute for Digital Sciences, France
Chedy Raïssi	French Research Institute for Digital Sciences, France
Marko Robnik-Šikonja	University of Ljubljana, Slovenia
Kazumi Saito	University of Shizuoka, Japan
Marina Sokolova	University of Ottawa and Institute for Big Data Analytics, Canada
Jerzy Stefanowski	Poznan University of Technology, Poland

Ljupčo Todorovski	University of Ljubljana, Slovenia
Luis Torgo	Dalhousie University, Canada
Herna Viktor	University of Ottawa, Canada
Albrecht Zimmermann	Université Caen Normandie, France
Blaž Zupan	University of Ljubljana, Slovenia

Additional Reviewers

Ahmadi, Mohsen
Barracchia, Emanuele Pio
Cancela, Brais
Chambers, Lorraine
Fernandes, Sofia
Ghomeshi, Hossein
Guarascio, Massimo
Koptelov, Maksim
Kulikovskikh, Ilona
Oliveira, Mariana
Pasquadibisceglie, Vincenzo
Pisani, Francesco S.
Stepišnik, Tomaž
Tabassum, Shazia
Tornede, Tanja
Wever, Marcel
Zopf, Markus

Keynote Talks

The Structure of Financial Networks

Guido Caldarelli

IMT School for Advanced Studies,
Lucca and European Centre for Living Technology, Venice

Abstract. Financial inter-linkages play an important role in the emergence of financial instabilities and the formulation of systemic risk can greatly benefit from a network approach. In this talk, we focus on the role of linkages along the two dimensions of contagion and liquidity, and we discuss some insights that have recently emerged from network models. With respect to the issue of the determination of the optimal architecture of the financial system, models suggest that regulators have to look at the interplay of network topology, capital requirements, and market liquidity. With respect to the issue of the determination of systemically important financial institutions, the findings indicate that both from the point of view of contagion and from the point of view of liquidity provision, there is more to systemic importance than just size. In particular for contagion, the position of institutions in the network matters and their impact can be computed through stress tests even when there are no defaults in the system.

We present an overview of the use of networks in Finance and Economics. We show how this approach enables us to address important questions as, for example, the stability of financial systems and the systemic risk associated with the functioning of the interbank market. For example with DebtRank, a novel measure of systemic impact inspired by feedback-centrality we are able to measure the nodes that become systemically important at the peak of the crisis. Moreover, a systemic default could have been triggered even by small dispersed shocks. The results suggest that the debate on too-big-to-fail institutions should include the even more serious issue of too-central-to-fail. All these results are new in the field and allow for a better understanding and modelling of different Financial systems.

Keywords: Financial networks · Systemic risk · Interbank market

Short Biography of the Lecturer: Guido Caldarelli is Full Professor in Theoretical Physics at IMT School for Advanced Studies Lucca, and is Research associate at the European Centre for Living Technology, Venice. His main scientific activity is the study of networks, mostly analysis and modelling of financial networks. Author of more than 200 publication on the subject and three books, he is currently the president of the Complex Systems Society. He has been coordinator of the FET IP Project MULTIPLEX: Foundational Research on Multilevel Complex Networks and Systems (2012–2016), the FET OPEN Project FoC: Forecasting Financial Crises (2010–2014), and the FET OPEN Project COSIN: Coevolution and Self Organization in Complex

Networks (2002–2005). Guido Caldarelli received his Ph.D. from SISSA, after which he was a postdoc in the Department of Physics and School of Biology, University of Manchester. He then worked at the Theory of Condensed Matter Group, University of Cambridge. He returned to Italy as a lecturer at National Institute for Condensed Matter (INFM) and later as Primo Ricercatore in the Institute of Complex Systems of the National Research Council of Italy. In this period, he was also the coordinator of the Networks subproject, part of the Complexity Project, for the Fermi Centre. He also spent some terms at University of Fribourg (Switzerland) and in 2006 he has been visiting professor at École Normale Supérieure in Paris. More information and a complete CV are available at: http://www.guidocaldarelli.com.

Data and Algorithmic Bias: Explaining the Network Effect in Opinion Dynamics and the Training Data Bias in Machine Learning

Dino Pedreschi

Università di Pisa, Istituto di Scienza e Tecnologie dell'Informazione,
Consiglio Nazionale delle Ricerche
http://kdd.isti.cnr.it

Abstract. Data science and network science are creating novel means to study the complexity of our societies and to measure, understand and predict social phenomena. My talk gives an overview of recent research at the Knowledge Discovery (KDD) Lab in Pisa within the SoBigData.eu research infrastructure, targeted at explaining the effects of data and algorithmic bias in different domains, using both data-driven and model-driven arguments. First, I introduce a model showing how algorithmic bias instilled in an opinion diffusion process artificially yields increased polarisation, fragmentation and instability in a population. Second, I focus on the urgent open challenge of how to construct meaningful explanations of opaque AI/ML black-box decision systems, introducing the local-to-global framework for the explanation of ML classifiers as a way towards explainable AI. The two cases show how the combination of data-driven and model-driven interdisciplinary research has a huge potential to shed new light on complex phenomena like discrimination and polarisation, as well as to explain how decision making black-boxes, both human and artificial, actually work. I conclude with an account of the open data science paradigm pursued in SoBigData.eu Research Infrastructure and its importance for interdisciplinary data driven science that impacts societal challenges.

Keywords: Explainable AI · Data bias · Algorithmic bias

Short Biography of the Lecturer: Dino Pedreschi is a professor of computer science at the University of Pisa, and a pioneering scientist in data science. He co-leads the Pisa KDD Lab – Knowledge Discovery and Data Mining Laboratory http://kdd.isti.cnr.it, a joint research initiative of the University of Pisa and the Information Science and Technology Institute of the Italian National Research Council. His research focus is on big data analytics and mining and their impact on society. He is a founder of the Business Informatics MSc program at University of Pisa, a course targeted at the education of interdisciplinary data scientists, and of SoBigData.eu, the European H2020 Research Infrastructure "Big Data Analytics and Social Mining Ecosystem" www.sobigdata.eu. Dino has been a visiting scientist at Barabasi Lab (Center for

Complex Network Research) of Northeastern University, Boston, and earlier at the University of Texas at Austin, at CWI Amsterdam and at UCLA. In 2009, Dino received a Google Research Award for his research on privacy-preserving data mining. Dino is a member of the expert group in AI of the Italian Ministry of research and the director of the Data Science PhD program at Scuola Normale Superiore in Pisa. Dino is a co-PI of the 2019 ERC grant XAI – Science and technology for the explanation of AI decision making (PI: Fosca Giannotti).

Representation Learning as a New Approach to Biomedical Research

Marinka Žitnik

Computer Science Department, School of Engineering, Stanford University

Abstract. Large datasets are being generated that can transform science and medicine. New machine learning methods are necessary to unlock these data and open doors for scientific discoveries. In this talk, I will argue that machine learning models should not be trained in the context of one particular dataset. Instead, we should be developing methods that combine data in their broadest sense into knowledge networks, enhance these networks to reduce biases and uncertainty, and then learn and reason over the networks. My talk will focus on two key aspects of this goal: representation learning and network science for knowledge networks. I will show how realizing this goal can set sights on new frontiers beyond classic applications of neural networks on biomedical image and sequence data. I will start by presenting a framework that learns deep models by embedding knowledge networks into compact embedding spaces whose geometry is optimized to reflect network topology, the essence of networks. I will then describe two applications of the framework to drug discovery and medicine. First, the framework allowed us to, for the first time, predict the safety of drug combinations at scale. We embedded a knowledge network of molecular, drug, and patient data at the scale of billions of interactions for all medications in the U.S. Using the embeddings, the approach can predict unwanted side effects for any combination of drugs that patients take, and we can validate predictions in the clinic using real patient data. Second, I will discuss how the framework enabled us to predict what diseases a new drug could treat. I will show how the new approach can make correct predictions for many recently repurposed drugs and can operate even on the hardest, yet critical, diseases for which no good treatments exist. I will conclude with future directions for learning over interaction data and translation of machine learning methods into solutions for biomedical problems.

Keywords: Biomedicine · Representation learning · Network science · Knowledge graphs

Short Biography of the Lecturer: Marinka Žitnik is a postdoctoral scholar in Computer Science at Stanford University. She will join Harvard University as a tenure-track assistant professor in December 2019. Her research investigates machine learning for sciences. Her methods have had a tangible impact in biology, genomics, and drug discovery, and are used by major biomedical institutions, including Baylor College of Medicine, Karolinska Institute, Stanford Medical School, and Massachusetts General Hospital. She received her Ph.D. in Computer Science from University of

Ljubljana while also researching at Imperial College London, University of Toronto, Baylor College of Medicine, and Stanford University. Her work received several best paper, poster, and research awards from the International Society for Computational Biology. She was named a Rising Star in EECS by MIT and also a Next Generation in Biomedicine by The Broad Institute of Harvard and MIT, being the only young scientist who received such recognition in both EECS and Biomedicine. She is also a member of the Chan Zuckerberg Biohub at Stanford.

Contents

Advanced Machine Learning

The CURE for Class Imbalance 3
Colin Bellinger, Paula Branco, and Luis Torgo

Mining a Maximum Weighted Set of Disjoint Submatrices. 18
Vincent Branders, Guillaume Derval, Pierre Schaus, and Pierre Dupont

Dataset Morphing to Analyze the Performance of Collaborative Filtering 29
André Correia, Carlos Soares, and Alípio Jorge

Construction of Histogram with Variable Bin-Width Based on Change Point Detection 40
Takayasu Fushimi, Kiyoto Iwasaki, Seiya Okubo, and Kazumi Saito

A Unified Approach to Biclustering Based on Formal Concept Analysis and Interval Pattern Structure 51
Nyoman Juniarta, Miguel Couceiro, and Amedeo Napoli

A Sampling-Based Approach for Discovering Subspace Clusters. 61
Sandy Moens, Boris Cule, and Bart Goethals

Epistemic Uncertainty Sampling 72
Vu-Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier

Utilizing Hierarchies in Tree-Based Online Structured Output Prediction 87
Aljaž Osojnik, Panče Panov, and Sašo Džeroski

On the Trade-Off Between Consistency and Coverage in Multi-label Rule Learning Heuristics 96
Michael Rapp, Eneldo Loza Mencía, and Johannes Fürnkranz

Hyperparameter Importance for Image Classification by Residual Neural Networks. 112
Abhinav Sharma, Jan N. van Rijn, Frank Hutter, and Andreas Müller

Applications

Cellular Traffic Prediction and Classification: A Comparative Evaluation of LSTM and ARIMA. 129
Amin Azari, Panagiotis Papapetrou, Stojan Denic, and Gunnar Peters

Main Factors Driving the Open Rate of Email Marketing Campaigns 145
Andreia Conceição and João Gama

Enhancing BMI-Based Student Clustering by Considering Fitness as Key Attribute 155
Erik Dovgan, Bojan Leskošek, Gregor Jurak, Gregor Starc, Maroje Sorić, and Mitja Luštrek

Deep Learning Does Not Generalize Well to Recognizing Cats and Dogs in Chinese Paintings 166
Qianqian Gu and Ross King

Temporal Analysis of Adverse Weather Conditions Affecting Wheat Production in Finland 176
Vladimir Kuzmanovski, Mika Sulkava, Taru Palosuo, and Jaakko Hollmén

Predicting Thermal Power Consumption of the Mars Express Satellite with Data Stream Mining 186
Bozhidar Stevanoski, Dragi Kocev, Aljaž Osojnik, Ivica Dimitrovski, and Sašo Džeroski

Data and Knowledge Representation

Parameter-Less Tensor Co-clustering 205
Elena Battaglia and Ruggero G. Pensa

Deep Triplet-Driven Semi-supervised Embedding Clustering 220
Dino Ienco and Ruggero G. Pensa

Neurodegenerative Disease Data Ontology 235
Ana Kostovska, Ilin Tolovski, Fatima Maikore, the Alzheimer's Disease Neuroimaging Initiative, Larisa Soldatova, and Panče Panov

Embedding to Reference t-SNE Space Addresses Batch Effects in Single-Cell Classification 246
Pavlin G. Poličar, Martin Stražar, and Blaž Zupan

Symbolic Graph Embedding Using Frequent Pattern Mining 261
Blaž Škrlj, Nada Lavrač, and Jan Kralj

Feature Importance

Feature Selection for Analogy-Based Learning to Rank 279
Mohsen Ahmadi Fahandar and Eyke Hüllermeier

Ensemble-Based Feature Ranking for Semi-supervised Classification 290
Matej Petković, Sašo Džeroski, and Dragi Kocev

Variance-Based Feature Importance in Neural Networks 306
Cláudio Rebelo de Sá

Interpretable Machine Learning

A Density Estimation Approach for Detecting and Explaining Exceptional Values in Categorical Data 319
Fabrizio Angiulli, Fabio Fassetti, Luigi Palopoli, and Cristina Serrao

A Framework for Human-Centered Exploration of Complex Event Log Graphs 335
Martin Atzmueller, Stefan Bloemheuvel, and Benjamin Kloepper

Sparse Robust Regression for Explaining Classifiers 351
Anton Björklund, Andreas Henelius, Emilia Oikarinen, Kimmo Kallonen, and Kai Puolamäki

Efficient Discovery of Expressive Multi-label Rules Using Relaxed Pruning 367
Yannik Klein, Michael Rapp, and Eneldo Loza Mencía

Networks

Evolving Social Networks Analysis via Tensor Decompositions: From Global Event Detection Towards Local Pattern Discovery and Specification. 385
Sofia Fernandes, Hadi Fanaee-T, and João Gama

Efficient and Accurate Non-exhaustive Pattern-Based Change Detection in Dynamic Networks 396
Angelo Impedovo, Michelangelo Ceci, and Toon Calders

A Combinatorial Multi-Armed Bandit Based Method for Dynamic Consensus Community Detection in Temporal Networks 412
Domenico Mandaglio and Andrea Tagarelli

Resampling-Based Framework for Unbiased Estimator of Node Centrality over Large Complex Network. 428
Kazumi Saito, Kouzou Ohara, Masahiro Kimura, and Hiroshi Motoda

Pattern Discovery

Layered Learning for Early Anomaly Detection: Predicting Critical Health Episodes 445
Vitor Cerqueira, Luis Torgo, and Carlos Soares

Ensemble Clustering for Novelty Detection in Data Streams 460
Kemilly Dearo Garcia, Elaine Ribeiro de Faria, Cláudio Rebelo de Sá, João Mendes-Moreira, Charu C. Aggarwal, André C. P. L. F. de Carvalho, and Joost N. Kok

Mining Patterns in Source Code Using Tree Mining Algorithms 471
Hoang Son Pham, Siegfried Nijssen, Kim Mens, Dario Di Nucci, Tim Molderez, Coen De Roover, Johan Fabry, and Vadim Zaytsev

KnowBots: Discovering Relevant Patterns in Chatbot Dialogues 481
Adriano Rivolli, Catarina Amaral, Luís Guardão, Cláudio Rebelo de Sá, and Carlos Soares

Fast Distance-Based Anomaly Detection in Images Using an Inception-Like Autoencoder 493
Natasa Sarafijanovic-Djukic and Jesse Davis

Time Series

Adaptive Long-Term Ensemble Learning from Multiple High-Dimensional Time-Series 511
Samaneh Khoshrou and Mykola Pechenizkiy

Fourier-Based Parametrization of Convolutional Neural Networks for Robust Time Series Forecasting 522
Sascha Krstanovic and Heiko Paulheim

Integrating LSTMs with Online Density Estimation for the Probabilistic Forecast of Energy Consumption. 533
Julian Vexler and Stefan Kramer

Author Index 545