Lecture Notes in Computer Science 9385

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

More information about this series at http://www.springer.com/series/7409

Elisa Fromont · Tijl De Bie
Matthijs van Leeuwen (Eds.)

Advances in Intelligent Data Analysis XIV

14th International Symposium, IDA 2015
Saint Etienne, France, October 22–24, 2015
Proceedings

Springer

Editors
Elisa Fromont
Université de Saint Etienne
Saint Etienne
France

Tijl De Bie
Intelligent Systems Lab
University of Bristol
Bristol
UK

Matthijs van Leeuwen
Informatics Section
Katholieke Universiteit Leuven
Leuven
Belgium

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-24464-8 ISBN 978-3-319-24465-5 (eBook)
DOI 10.1007/978-3-319-24465-5

Library of Congress Control Number: 2015948773

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

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

Preface

We are proud to present the proceedings of IDA 2015, the 14th International Symposium on Intelligent Data Analysis, which was held from October 22 to October 24, 2015, in Saint-Etienne, France.

The series started in 1995 and was held biennially until 2009. In 2010, the symposium re-focused to support papers that go beyond established technology and contain genuinely novel and game-changing ideas, while not always being as fully realized as papers accepted at other conferences. To further support this unique focus, IDA 2015 additionally included a so-called "Horizon Track", which contained contributed talks about research that may be too preliminary for archival publication, but with a potentially very high impact. The IDA symposium is an interdisciplinary meeting that solicits contributions on all aspects of intelligent data analysis, including papers on intelligent support for modeling and analyzing data from complex, dynamical systems. Intelligent support for data analysis goes beyond the usual algorithmic offerings in the literature. Papers about established technology were only accepted if the technology was embedded in intelligent data analysis systems, or was applied in novel ways to analyzing and/or modeling complex systems.

The conventional reviewing process, which tends to favor incremental advances on established work, can discourage the kinds of papers that IDA 2015 has published. The reviewing process adopted for IDA addressed this issue explicitly: referees evaluated papers against the stated goals of the symposium, and an informed, thoughtful, positive review written by a Program Committee advisor could outweigh other, negative reviews and result in acceptance of the paper. Indeed, it was noted that this had a notable impact on the selection of papers included in the program. In addition, the new "Horizon Track" allowed researchers to present their most ground breaking research at the symposium without publishing it in the proceedings, stimulating discussions about the most exciting research ideas and visions at an early stage.

We were pleased to have a very strong program. We received 65 submissions in total. In all, 59 papers were submitted to the regular proceedings track, of which 29 were accepted for inclusion in this volume. Six papers were submitted to the Horizon Track, of which three were accepted for presentation at the symposium. The IDA Frontier Prize was awarded to the most visionary contribution. As in previous years, we included a poster and video track for PhD students to promote their work. The best 2-minute video, as decided by the participants of the symposium, was awarded the Video Prize.

We were honored to have three distinguished invited speakers at IDA 2015:

– Tony Veale from University College Dublin, Ireland, talked about "The Shape of Tweets to Come" and how Twitter presents a generative opportunity of another kind to the computationally-minded language researcher, to study how algorithmic models might impose linguistic hypotheses onto large data sources to compose novel and meaningful micro-texts of their own.

- Nick Heard from Imperial College London, UK, talked about "Combining Weak Statistical Evidence in Cyber Security" and how statistical modelling (and the use of p-values) of nodes and edges in a computer network can build up a picture of normal behavior in a system.
- Pascal Van Hentenryck from National ICT (NICTA), Australia, talked about "Evidence-Based Optimization". He showed some case studies in disaster management, energy systems, high-performance computing, and market optimization and presented some emerging architectures for evidence-based optimization.

The conference was held in the buildings of Telecom Saint-Etienne Engineering School in front of the Hubert-Curien Laboratory. All those buildings were part of the new "Manufacture d'armes de Saint-Étienne" (MAS, known for example for the FAMAS assault rifle), built in 1864 and closed in 2001. They are now part of a new university campus dedicated to computer science and physics (and in particular, optics).

We wish to express our gratitude to all authors of submitted papers for their intellectual contributions; to the Program Committee members and the additional reviewers for their effort in reviewing and commenting on the submitted papers; to the members of the IDA Steering Committee for their ongoing guidance and support; and to the Program Committee advisors for their active involvement. Special thanks go to the poster and video chair, Jesse Read; the local chair, Baptiste Jeudy; the publicity chair, Edward Cohen; the sponsorship chair, François Jacquenet; the Frontier Prize chairs, Michael Berthold and Elizabeth Bradley; and the webmaster, Leonor Becerra-Bonache. We gratefully acknowledge those who were involved in the local organization of the symposium: Romain Deville, Rémi Emonet, Damien Fourure, Matthias Gery, Amaury Habrard, Christine Largeron, and Emilie Morvant.

Finally, we are grateful to our sponsors and supporters: KNIME, for funding the IDA Frontier Prize for the most visionary contribution presenting a novel and surprising approach to data analysis; the French Artificial Intelligence Association (AFIA), for funding the IDA Video Prize for the best video presented in the PhD poster and video track; the IT company Eura Nova; Jean Monnet University (UJM); the Artificial Intelligence journal; Télécom Saint-Etienne (for sharing their building); and Springer.

August 2015

Tijl De Bie
Elisa Fromont
Matthijs van Leeuwen

Organization

General Chair

Matthijs van Leeuwen	Leiden University, Netherlands and KU Leuven, Belgium

Program Chairs

Tijl De Bie	Ghent University, Belgium University of Bristol, UK
Elisa Fromont	University of Lyon, St-Etienne, France

Poster and Video Chair

Jesse Read	Aalto University, Finland

Local Chair

Baptiste Jeudy	University of Lyon, St-Etienne, France

Publicity Chair

Edward Cohen	Imperial College London, UK

Sponsorship Chair

François Jacquenet	University of Lyon, St-Etienne, France

Frontier Prize Chairs

Elizabeth Bradley	University of Colorado, USA
Michael Berthold	University of Konstanz, Germany

Advisory Chairs

Joost Kok	Universiteit Leiden, Netherlands
Paul Cohen	University of Arizona, USA
Nada Lavrač	Jožef Stefan Institute, Slovenia

Webmaster

Leonor Becerra-Bonache	University of Lyon, St-Etienne, France

Local Organization Committee

Romain Deville	University of Lyon, INSA de Lyon, France
Rémi Emonet	University of Lyon, St-Etienne, France
Damien Fourure	University of Lyon, St-Etienne, France
Matthias Gery	University of Lyon, St-Etienne, France
Amaury Habrard	University of Lyon, St-Etienne, France
Christine Largeron	University of Lyon, St-Etienne, France
Emilie Morvant	University of Lyon, St-Etienne, France

Program Committee Advisors

Michael Berthold	University of Konstanz, Germany
Hendrik Blockeel	KU Leuven, Belgium
Liz Bradley	University of Colorado, USA
Paul Cohen	University of Arizona, USA
Jaakko Hollmén	Aalto University School of Science, Finland
Frank Klawonn	Ostfalia University of Applied Sciences, Germany
Joost Kok	Leiden University, The Netherlands
Nada Lavrač	Jožef Stefan Institute, Slovenia
Matthijs van Leeuwen	KU Leuven, Belgium
Xiaohui Liu	Brunel University, UK
Arno Siebes	Universiteit Utrecht, Netherlands
Stephen Swift	Brunel University, UK
Hannu Toivonen	University of Helsinki, Finland
Allan Tucker	Brunel University, UK

Program Committee

Fabrizio Angiulli	DEIS, University of Calabria, Italy
Alexandre Aussem	GAMA, Université Lyon 1, France
José L. Balcázar	Universitat Politècnica de Catalunya, Spain
Elena Bellodi	ENDIF-University of Ferrara, Italy
Maria Bielikova	Slovak University of Technology in Bratislava, Slovakia
Christian Borgelt	European Centre for Soft Computing, Spain
Henrik Bostrom	Stockholm University, Sweden
Marc Boullé	Orange Labs, France
Toon Calders	Université Libre de Bruxelles, Belgium
Andre Carvalho	USP-Universidade de São Paulo, Brazil
Loïc Cerf	Universidade Federal de Minas Gerais, Brazil
Edward Cohen	Imperial College London, UK

Bruno Crémilleux	Université de Caen, France
Bernard De Baets	Ghent University, Belgium
José Del Campo-Ávila	Universidad de Málaga, Spain
Wouter Duivesteijn	TU Dortmund, Germany
Saso Dzeroski	Jozef Stefan Institute, Slovenia
Fazel Famili	National Research Council Canada, Institute for Information Technology, Canada
Ad Feelders	Universiteit Utrecht, The Netherlands
Ingrid Fischer	University of Constance, Germany
Douglas Fisher	Vanderbilt University, USA
Johannes Fürnkranz	TU Darmstadt, Germany
Ricard Gavalda	Universitat Politecnica de Catalunya, Spain
Bart Goethals	University of Antwerp, Belgium
Javier Gonzalez	Sheffield University, UK
Kenny Gruchalla	NREL/CU-Boulder, USA
Tias Guns	KU Leuven, Belgium
Lawrence Hall	University of South Florida, USA
Eyke Huellermeier	University of Marburg, USA
Frank Höppner	Ostfalia University of Applied Sciences, Germany
Baptiste Jeudy	Laboratoire Hubert Curien, France
Frank Klawonn	Ostfalia University of Applied Sciences, Germany
Arno Knobbe	LIACS, Netherlands
Rudolf Kruse	University of Magdeburg, Germany
Vincent Lemaire	Orange Labs, France
Xiaohui Liu	Brunel University, UK
Jose A. Lozano	The University of the Basque Country, Spain
George Magoulas	Birkbeck College, UK
Maria-Carolina Monard	USP- Universidade de São Paulo, Brazil
Mohamed Nadif	Paris Descartes University, France
Andreas Nuernberger	Otto-von-Guericke University of Magdeburg, Germany
Panagiotis Papapetrou	Stockholm University, Sweden
Nicos Pavlidis	Lancaster University, UK
Mykola Pechenizkiy	Eindhoven University of Technology, Netherlands
Jose-Maria Pena	Universidad Politécnica de Madrid, Spain
Ruggero G. Pensa	University of Turin, Italy
Marc Plantevit	LIRIS - Université Claude Bernard Lyon 1, France
François Portet	Laboratoire d'Informatique de Grenoble, France
Alexandra Poulovassilis	Birkbeck College, University of London, UK
Miguel A. Prada	Universidad de León, Spain
Ronaldo Prati	Universidade Federal do ABC – UFABC, Brazil
Jesse Read	Aalto University, Finland
Antonio Salmeron	University of Almería, Spain
Vítor Santos Costa	Universidade do Porto, Portugal
Roberta Siciliano	University of Naples Federico II, Italy
Christine Solnon	LIRIS CNRS UMR 5205/INSA Lyon, France
Myra Spiliopoulou	Otto-von-Guericke-Universität Magdeburg, Germany

Maguelonne Teisseire	Cemagref - UMR Tetis, France
Melissa Turcotte	LANL, USA
Antti Ukkonen	Helsinki Institute for Information Technology, Finland
Maarten Van Someren	University of Amsterdam, Netherlands
Veronica Vinciotti	Brunel University, UK
Jilles Vreeken	Max-Planck Institute for Informatics and Saarland University, Germany
Zidong Wang	Brunel University, UK
Leishi Zhang	Middlesex University, UK
Indre Zliobaite	Aalto University, Finland

Additional Reviewers

Braune, Christian
Cerri, Ricardo
Cule, Boris
D'Ambrosio, Antonio
Doell, Christoph
Fassetti, Fabio
Fortino, Vittorio
Freitas, Alex
Gossen, Tatiana
Guy, Michelle
Held, Pascal
Ienco, Dino
Liu, Renhao
Madjarov, Gjorgji
Muelas, Santiago
Nair Benrekia, Yacine Noureddine
Palopoli, Luigi
Pio, Gianvito
Zhang, Jianpeng
Zhou, Mu
Ševcech, Jakub

Invited Talks Abstracts

The Shape of Tweets to Come

Tony Veale

University College Dublin, Ireland
tony.veale@ucd.ie
https://www.csi.ucd.ie/users/tony-veale

Abstract. Twitter has proven itself a rich and varied source of language data for linguistic analysis. For Twitter is more than a popular new platform for social interaction via language; in many ways Twitter constitutes a whole new genre of text, as users adapt to its limitations (140 character *"tweets"*) *and* its novel conventions (e.g. re-tweeting, hashtags). Language researchers can thus harvest Twitter data to study how users convey meaning with affect, and how they achieve stickiness and virality with the texts they compose. But Twitter presents an opportunity of another kind to the computationally-minded language researcher, a *generative* opportunity to study how algorithmic models might impose linguistic hypotheses onto large data sources to compose novel and meaningful micro-texts of their own. This computational turn allows researchers to go beyond merely descriptive models of playful uses of language such as metaphor, sarcasm and irony. It allows researchers to test whether their models embody a sufficiently algorithmic understanding of a phenomenon to facilitate the construction of a fully-automated computational system, one that can generate wholly novel examples that are deemed acceptable to humans. This talk presents and evaluates one such system, a *Twitterbot* named *@MetaphorMagnet* that generates, expresses and shares its own playful insights on Twitter. I shall show how *@MetaphorMagnet*s tweets are inspired by data but shaped by knowledge, and consider how the outputs of this hybrid data/knowledge-driven bot may be usefully anchored in another source of data – the news stream.

Combining Weak Statistical Evidence in Cyber Security

Nicholas A. Heard

Imperial College London, UK
n.heard@imperial.ac.uk
http://wwwf.imperial.ac.uk/~naheard/

Abstract. Cyber attacks on government and industry computer networks are now commonplace and no system can be made invulnerable to intrusion. Instead, much importance is placed on reducing the impact of cyber attacks when they occur, which first means quickly detecting their presence amongst the flow of cyber traffic. However, sophisticated hackers and cyber criminals will act carefully to hide their presence, and so any hard detection rules (signatures) can be circumnavigated. Nonetheless, if an intrusion has a malign purpose, then at least some unusual behaviour will be hidden within the network traffic data. Statistical modelling of nodes and edges in a computer network can build up a picture of normal behaviour in the system. Typical institutional computer networks produce high volume data streams and so, from time to time, surprising but benign behaviour will be observed. The task is to detect the significance of genuine intrusion events against this background. In statistical modelling, p-values are the fundamental quantities for measuring the significance of observed data against a null hypothesis. This talk will review methods of combining p-values to accumulate evidence, investigating their properties in depth. Some new approaches will then be proposed which are better suited for detecting subsets of significant p-values. Finally, the advantages of the proposed approach will be illustrated on a cyber authentication problem, stemming from collaborative work with Los Alamos National Laboratory.

Evidence-Based Optimization

Pascal Van Hentenryck

National ICT (NICTA), Australia
pascal.vanhentenryck@nicta.com.au
http://org.nicta.com.au/people/phentenryck/

Abstract. For the first time in the history of mankind, we are accumulating data sets of unprecedented scale and accuracy about physical infrastructures, natural phenomena, man-made processes, and human behavior. These developments, together with progress in high-performance computing, machine learning, and operations research, offer exciting opportunities for the evidence-based optimization of global systems. This talk reviews some case-studies in disaster management, energy systems, high-performance computing, and market optimization to showcase these unique opportunities and their associated challenges, and presents some emerging architectures for evidence-based optimization.

Horizon Track Abstracts

Towards a Data Science Collaboratory

Joaquin Vanschoren[1], Bernd Bischl[2], Frank Hutter[3], Michele Sebag[4],
Balazs Kegl[4], Matthias Schmid[5], Giulio Napolitano[5],
Katy Wolstencroft[6], Alan R. Williams[7], and Neil Lawrence[8]

[1] Eindhoven University of Technology
j.vanschoren@tue.nl
[2] Ludwig-Maximilians-University Munich
bernd.bischl@stat.uni-muenchen.de
[3] Albert-Ludwigs-Universität Freiburg
fh@informatik.uni-freiburg.de
[4] Université Paris Sud
michele.sebag@lri.fr, balazs.kegl@gmail.com
[5] Universität Bonn
{matthias.schmid,giulio}@imbie.meb.uni-bonn.de
[6] Universiteit Leiden
k.j.wolstencroft@liacs.leidenuniv.nl
[7] University of Manchester
alan.r.williams@manchester.ac.uk
[8] University of Sheffield
n.lawrence@dcs.shef.ac.uk

1 The Fragmented Data Science Ecosystem

Data-driven research requires an ecosystem of people fulfilling many different roles: domain scientists collect and analyze data to study or discover phenomena; data scientists design and evaluate algorithms, and computer scientists implement and maintain these techniques to be used throughout science and industry.

However, there exist large gaps between these communities that slow down the rate of discovery. First, because domain scientists have a more limited view on the state of the art in data science, they are often unsure about the latest and most appropriate techniques. There exist extensive algorithms libraries, but it is often not clear how to optimally use them. Hence, they either spend a lot of time on research and experimentation, or make suboptimal choices. Data scientists, on the other hand, often don't speak the language of domain scientists, hence missing opportunities to work interactively with them and innovate in their respective fields. While there exist many wonderful open data repositories, it is far from obvious how to access, understand and use this data. Hence, much research still uses datasets that are of little scientific or industrial interest today. Knowledge transfer through the literature is inefficient, as findings are spread over millions of papers based on tacit domain-specific knowledge and community-specific jargon. Moreover, while scientific papers are being produced at a tremendous rate, reproducible and readily applicable major discoveries are far fewer [1]. Empirical evaluations of algorithms on known datasets are typically not

organized online, but confined to papers with varying levels of detail, making them virtually impossible to build on. In short, while it is *theoretically* possible for these communities to build on each other's work, in practice there is a lot of friction involved, such as handling myriad data formats, studying source code, emailing authors, and running complex experiments. As a result, many scientists spend vast amounts of time on tasks that others could do in a fraction of that time, that could be done much better using novel/better techniques, or that could be automated altogether.

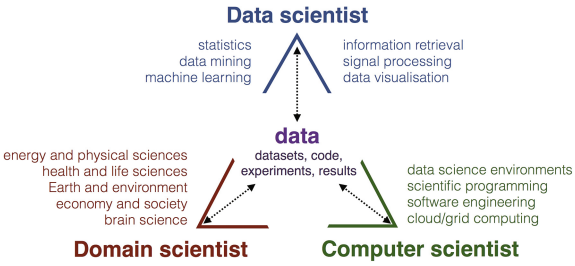

Fig. 1. Roles within the data science ecosystem and the gaps between them.

2 An Online Collaboratory

We propose to create an online *collaboratory*, where data scientists, domain scientists and computer scientists can easily interact and build directly on each other's work, transforming the practice of data science from small-scale local collaborations to massive, real-time, *online* collaborations.

First, we can extract actionable datasets from large scientific databases, identify key software components, and (auto-)annotate them with a practical base vocabulary to create a 'search engine for data science'. This helps scientists find useful tools (e.g. scalable clustering algorithms), and test algorithms on many recent datasets. Next, scientists can challenge the community to solve problems, yielding experiments showing how well each particular solution works. Crucially, experiments and results (e.g., predictive models) should be open and linked to the underlying data sets and workflows, ensuring reproducibility and creating a single, organized body of research: a 'data telescope' that can be wielded by scientists, industry and students alike. The collaboratory should support social networking to protect preliminary research, and track the impact of all contributions (reuse, downloads, altmetrics) to help scientist build their reputation.

The collaboratory can be seamlessly integrated into the tools that scientists already use, to automatically download and upload data, code and experiments [2]. Moreover, it offers unprecedented opportunities to intelligently recommend algorithms or optimize large parameter spaces, thus saving time.

We hope to bring together data scientists, computer scientists, and domain scientists from many domains to see how current tools can be connected, and best practices can be shared. We expect that this *networked* approach to data science will strongly

contribute to speeding up data-driven science in the near future. Indeed, if we are all part of the same 'experimentation system in the sky', we can all become more productive and efficiently learn from each other.

References

1. Ioannidis, J.: Why most published research findings are false. PLoS Med. **2**(8), e214 (2005)
2. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: networked science in machine learning. SIGKDD Explor. **15**(2), 49–60 (2013)

When Learning Indeed Changes the World: Diagnosing Prediction-Induced Drift

Georg Krempl, Dávid Bodnár, Anita Hrubos

Knowledge Management & Discovery, Otto-von-Guericke University,
Universitätsplatz 2, 39106 Magdeburg, Germany
georg.krempl@iti.cs.uni-magdeburg.de
http://kmd.ovgu.de/res/driftmining

Abstract. A fundamental assumption underlying many prediction systems is that they act as an invisible observer without interfering with the evolution of the population under study. More formally, their distribution is assumed to be independent of the system's previous predictions. Nevertheless, this is violated when, for example, the predictor faces intelligent and malevolent adversaries who counteract its classification rules, or when the classification as high-risk might become a self-fulfilling prophecy. The former has received some attention in adversarial machine learning [10, 12] in the context of hardening classifiers against such an adversary, and indications for the latter have been reported for recommender systems [3, 5] and financial applications [9, 13]. However, the problems of *self-defeating* and *self-fulfilling* prophecies in prediction systems have not been studied in an unified framework yet, leaving questions such as how to detect them in drift open.

We address this by presenting a first approach to assess the presence of such *prediction-induced drift* in datasets. Our work complements literature on change detection [11], change and drift mining [2, 6, 7], and concept drift [4, 8], which is based on the assumption that the observed drift is independent of the system's predictions. We illustrate and evaluate our approach on data generated with and without self-defeating and self-fulfilling prophecies. While prediction-induced drift is not limited to classification problems, but might also occur in other fields of machine learning, we focus this initial analysis on the classification task.

Our preliminary results on synthetic datasets are promising but highlight two major challenges: First, while the majority of prediction-induced drifts is correctly detected, detection of self-fulfilling prophecies seems more difficult, requiring further research. Second, this analysis requires knowing the labels that were actually assigned by prediction systems, which are currently not published in real-world benchmark datasets such as the ones in the UCI Machine Learning Repository [1].

Thus, this contribution also aims to motivate the community to collect, to share, and to analyse data with the system's actual predictions in order to allow an assessment of prediction-induced drift in real-world applications.

Keywords: prediction-induced drift, concept drift, dataset shift, population drift, nonstationarity, change mining, drift mining, change detection, self-fulfilling prophecy, self-defeating prophecy, adversarial machine learning.

Acknowledgments. We thank Daniel Kottke, Myra Spiliopoulou, Julia Hempel and Marcus Kamieth from University of Magdeburg, Battista Biggio from University of Cagliari, and Michele Sebag and Marc Schoenauer from INRIA Saclay for the insightful discussions on change mining, adversarial machine learning and ways of assessing prediction-induced drift in data.

References

1. Asuncion, A., Newman, D.J.: UCI machine learning repository (2015). http://archive.ics.uci.edu/ml/
2. Böttcher, M., Höppner, F., Spiliopoulou, M.: On exploiting the power of time in data mining. ACM SIGKDD Explor. Newsl. **10**(2), 3–11 (2008)
3. Fleder, D., Hosanagar, K.: Blockbuster culture's next rise or fall: the impact of recommender systems on sales diversity. Manag. Sci. **55**(5), 697–712 (2009)
4. Gama, J., Zliobaitė, I., Bifet, A., Pechenizkiy, M.: A survey on concept-drift adaptation. under review (2013)
5. Hagen, S.t., Someren, M.v., Hollink, V.: Exploration/exploitation in adaptive recommender systems. In: Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (2003)
6. Hofer, V., Krempl, G.: Drift mining in data: A framework for addressing drift in classification. Comput. Stat. Data Anal. **57**(1), 377–391 (2013)
7. Krempl, G.: Temporal density extrapolation. In: Douzal-Chouakria, A., Vilar, J.A., Marteau, P.F., Maharaj, A., Alonso, A.M., Otranto, E. (eds.) Proceedings of the ECML/PKDD 2015 Workshop on Advanced Analytics and Learning on Temporal Data (AALTD 2015), Porto, Portugal, September 11, 2015. CEUR Workshop Proceedings (2015)
8. Krempl, G., Zliobaitė, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., Stefanowski, J.: Open challenges for data stream mining research. SIGKDD Explor. **16**(1), 1–10 (2014), special Issue on Big Data
9. Larsen, F.: Automatic stock market trading based on technical analysis (2007), masterthesis, Institutt for datateknikk og informasjonsvitenskap, Norges teknisk-naturvitenskapelige universitet
10. Laskov, P., Lippmann, R.: Machine learning in adversarial environments. Mach. Learn. **81**(2), 115–119 (2010)
11. Sebastião, R., Gama, J.: A study on change detection methods. In: Proceedings of the 4th Portuguese Conference on Artificial Intelligence, Lisbon (2009)
12. Tygar, J.D.: Adversarial machine learning. IEEE Internet Comput. **15**(5), 4–6 (2011)
13. Wisniewski, T.P., Lambe, B.: The role of media in the credit crunch: The case of the banking sector. J. Econ. Behav. Organ. **85**, 163–175 (2013)

The Data Problem in Data Mining

Albrecht Zimmermann

INSA de Lyon
albrecht.zimmermann@insa-lyon.fr

Abstract. Computer science is essentially an applied or engineering science, creating tools. In Data Mining, those tools are supposed to help humans understand large amounts of data, and produce actionable insight. In this talk, I argue that for all the progress that has been made in Data Mining, in particular Pattern Mining, we are lacking understanding of key aspects of the performance and results of pattern mining algorithms. I will focus particularly on the difficulty of deriving actionable knowledge from patterns. I trace the lack of progress regarding those questions to a lack of data with varying, controlled properties, and argue that we will need to make a science of digital data generation, and use it to develop guidance to data practitioners.

1 Short-Comings in Evaluation

Data Mining, and in particular Pattern Mining, have been around for about two decades and the work in the field has led to a large number of techniques, which have been applied to pattern domains as diverse as itemsets, attribute-value data, sequences, trees, and graphs, and tasks ranging from finding associations to describing interesting subpopulations, to predicting unseen class labels.

In this talk, I will focus on the unsupervised pattern mining setting, i.e. finding unexpected, interesting and useful patterns that are not related to a variable of interest - nominal or otherwise. As I will argue, the *qualitative* evaluation of proposed techniques, i.e. how "good" the resulting patterns are, has been given short thrift in comparison to *quantitative* evaluation, i.e. how efficiently the output is found.

But also the latter has arguably not been given the attention it deserved. This case has been made convincingly early on by Zheng *et al.* [2], who showed that the evaluations performed in itemset mining up to that point in time had led to an over-fitting on the artificially generated data used. The reported performance did not transfer to real-life data, which showed different characteristics than the artificially generated data. Remarkably enough, the situation has barely improved since then, with quantitative evaluations focused on a small number of data sets, of which typically only few are used in a given evaluation.

The situation is worse for qualitative evaluations, which are rarely performed in the first place. This is understandable since the lack of a target variable corresponds to missing ground truth in the data. But at the same time, it means that even if we knew

how to set parameters appropriately[1], we would not know how found patterns relate to the processes that generated the data. Since pattern mining is supposed to give us insight into those processes, and allow us to act based on found patterns, this is a serious short-coming.

2 Generating Data (and Understanding Pattern Mining)

When there is no ground truth available for real-life data (or when there is little real-life data available in the first place), generating artificial data is a promising alternative. This is not only the case in computer science, where, for instance, the SAT solving community has chosen this direction, but also in "hard sciences" like physics, see for instance [1].

Data generation allows us to both break the bottleneck of too few data sets (or data sets with a too narrow range of characteristics), and to understand how found patterns relate to the processes that generated the data. As Zheng *et al.* showed, however, and others have demonstrated since, approaching this task without forethought and an understanding of the data we aim to generate will lead to unrealistic data sets. Furthermore, limiting ourselves to a narrow selection of generative processes, e.g. generating itemset mining data only by combining itemsets, will restrict the lessons to be learned from matching patterns to processes, and carries the risk of biasing qualitative evaluations.

Fortunately, we do not have to start from scratch. More-or-less successful attempts at data generation have been made, and some infrastructure exists to support this task. Additionally, some researchers have attempted to relate patterns to different processes to evaluate their quality, especially in recent years. Finally, researchers and practitioners in other fields have developed theories of their own that, while necessarily taken with a grain of salt, can be built on to simulate real-life processes. By combining and building on this existing knowledge, we can fill in the current data gaps and start to understand those aspects of pattern mining that escape us so far.

References

1. Cern software development for experiments - Simulation. http://ph-dep-sft.web.cern.ch/project/simulation
2. Zheng, Z., Kohavi, R., Mason, L.: Real world performance of association rule algorithms. In: KDD, pp. 401–406 (2001)

[1] Another area in which there is too little guidance.

Contents