Guest Editorial: Special Issue on Causal Discovery 2017

Lin Liu¹ · Jiuyong Li¹ · Kun Zhang² · Emre Kıcıman³ · Negar Kiyavash⁴

© Springer International Publishing AG, part of Springer Nature 2018

Causal discovery from data is emerging as a new topic of interest in the data mining community. In two consecutive years of 2016 and 2017, the Causal Discovery Workshops held in conjunction with the KDD conference (ACM SIGKDD International Conference on Knowledge Discovery and Data Mining) have attracted great attention from KDD participants and have provided researchers in the data mining and machine learning fields opportunities to present and exchange their ideas and achievements in causal discovery. Complementary to the KDD Causal Discovery Workshops, the JDSA Special Issues on Causal Discovery (2016 and 2017) have served as a platform specialized in causal discovery, for publishing selected research work presented at the workshops, as well as papers directly submitted to the special issues.

In recent decades, research in causal discovery has gained notable progress. However, there are many challenges for discovering causal relationships from real-world data. For example, the data may not be complete or clean and often contain different types of variables. The papers included in this special issue address such practical problems.

Datasets with mixed types of variables are common in many real-world applications. The paper “Scoring Bayesian

Networks of Mixed Variables”, by Andrews, Ramsey and Cooper, tackles the problem of causal structure learning with mixed data when using a score-based approach. The paper is focused on the scalability of the learning in the presence of both continuous and discrete variables, by proposing two novel and scalable scoring functions. A structure prior is also introduced to improve the efficiency of learning large networks. Additionally, authors of the paper show how the scoring functions may be readily adapted as conditional independence tests for constraint-based Bayesian network learning algorithms.

In their paper, “Constraint-based Causal Discovery with Mixed Data”, Tsagris, Borboudakis, Lagani and Tsamardinos also study mixed data types, but in the context of constraint-based approach to causal structure discovery. The conditional independence tests derived in the paper can be directly applied in existing constraint-based structure learning algorithms to mixed data types, including those for learning the structure of a Bayesian network (such as PC) and those for learning a maximal ancestral graphs (such as FCI), respectively.

The paper “Comparison of Strategies for Scalable Causal Discovery of Latent Variable Models from Mixed Data”, by Raghu, Ramsey, Morris, Manatakis, Spirtes, Chrysanthis, Glymour and Benos, continues to investigate constraint-based causal discovery methods with mixed data, by comparing the accuracy and efficiency of different strategies in large, mixed datasets with latent confounders. The experiments show that two extensions of the FCI algorithm: a maximum probability search procedure for more accurate identification of causal orientations and a strategy for quick elimination of unlikely adjacencies in order to achieve scalability to high dimensional data, significantly outperform the state of the art in the field by achieving both accurate edge orientations and tractable running time in simulation experiments on datasets with up to 500 variables. The efficacy of the best performing approach is also demonstrated with a real-world biomedical dataset of HIV-infected individuals.

Missing values are another frequently encountered practical issue for causal discovery. In the case where a dataset

✉ Lin Liu
lin.liu@unisa.edu.au
Jiuyong Li
jiuyong.li@unisa.edu.au
Kun Zhang
kunz1@cmu.edu
Emre Kıcıman
emrek@microsoft.com
Negar Kiyavash
kiyavash@illinois.edu

¹ School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, Australia

² Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, USA

³ Microsoft Research, Redmond, WA, USA

⁴ University of Illinois at Urbana-Champaign, Urbana, IL, USA

contains variables with values missing not at random, a commonly used approach is to remove the samples with any missing values completely. However, such a list-wise deletion may not efficiently utilize the information in the dataset as samples with only a small number of missing values are discarded. Motivated by this fact, in the paper, “Fast Causal Inference with Non-Random Missingness by Test-Wise Deletion”, Strobl, Visweswaran and Spirtes propose to apply test-wise deletion to save samples for constraint-based structure learning methods, by deleting samples only among the variables required for each conditional independence test used in the searches. The paper shows that test-wise deletion is sound under the justifiable assumption that the missingness mechanisms do not causally affect each other in the underlying causal structure, and the causal discovery methods perform better when paired with the test-wise deletion, compared to the list-wise deletion and missing value imputation.

We hope the selected papers published in this special issue form a small but focused collection of the current research in tackling some of the essential practical issues in causal

discovery from data, which in turn can stimulate readers to develop more reliable methods to solve more real life problems.

To conclude this guest editorial, we would like to thank everyone who has contributed to this special issue. Firstly, we would like to thank all authors for submitting their papers to the 2017 KDD Causal Discovery Workshop and this special issue. We would also like to thank Professor Longbing Cao, editor-in-chief of JDSA for all his support and help throughout the whole process of preparing this special issue. Finally, we would like to express our great gratitude to all reviewers, who have worked hard to provide high-quality reviews within the very tight review timeline: Ruichu Cai, Philipp Geiger, Adam Glynn, Mingming Gong, Samantha Kleinberg, Thuc Le, Tsai-Ching Lu, Wolfgang Mayer, Amit Sharma, Shohei Shimizu, Ricardo Silva, Petar Stojanov, Eric Strobl, Sofia Triantafillou, Ioannis Tsamardinos, Kui Yu, and Qingyuan Zhao.