Advances in Integrative Causal Analysis

by Ioannis Tsamardinos

Scientific practice typically involves studying a system over a series of studies and data collection, each time trying to unravel a different aspect. In each study, the scientist may take measurements under different experimental conditions and measure different sets of quantities (variables). The result is a collection of heterogeneous data sets coming from different distributions. Even so, these are generated by the *same causal mechanism*. The general idea in Integrative Causal Analysis (INCA) is to identify the set of causal models that simultaneously fit (are consistent) with all sources of data and prior knowledge and reason with this set of models. Integrative Causal Analysis allows more discoveries than what is possible by independent analysis of datasets. In this talk, we'll present advances in this direction that lead to algorithms that can handle more types of heterogeneity, and aim at increasing efficiency or robustness of discoveries. Specifically, we'll present (a) general INCA algorithms for causal discovery from heterogeneous data [1], (b) algorithms for converting the results of tests to posterior probabilities and allow conflict resolution and identification of the confidence network regions [1, 3], (d) proof-of-concept applications and massive evaluation on real data of the main concepts [6], (c) extensions that can deal with prior causal knowledge [4, 5], and (d) extensions that handle case-control data [2].

Specifically, in recent work [1] we have introduced COmbINE, a novel algorithm for causal discovery from multiple datasets over different variable sets and different (hard) interventions. COmbINE and particularly his predecessors [7] have introduced the conversion of the inverse problem (causal discovery) to a SAT instance as a means to induce causal models. Conversion to SAT opens new directions to causal discovery since it is a powerful, general technique that allows the encoding of different types of problems. INCA problems typically admit an exponential number of solution networks; their enumeration is impractical. We have introduce query-based causal discovery where the user poses a query about the presence or absence of a causal feature of interest, e.g., the presence of a specific causal path, an edge, or a direction. Enumeration of all solutions is avoided and reasoning algorithms focus instead on proving the feature has to be present in all solution, in no solution, or in some solutions. While the SAT-based approach to causality allows novel solutions to INCA problems, it has the disadvantage that it cannot deal with inconsistent and contradicting statistical evidence. To avoid this problem a conflict resolution technique is necessary. In recent work [1, 3], we show how to efficiently and relatively accurately estimate the *posterior* probabilities of conditional dependencies and independencies. These probabilities can rank the constraints stemming from the statistical tests on the data and lead to conflict resolution algorithms. In addition, they can be used to identify regions of the network of high-confidence [3].

While arguably theoretically interesting, INCA algorithms need to prove themselves to practice. In related recent work [6] we show how to test some of these ideas on real data to make specific (statistical) predictions about the presence of an association between two variables never jointly measured, based on a causal analysis of the union of variables measured in two different datasets. This work provides statistical evidence that causal-based modeling and discovery is useful for designing algorithms for non-trivial inferences. The above algorithms and ideas can handle datasets with overlapping variables under different (hard) interventions. We discuss two directions that allow such algorithms to consider prior causal knowledge about the present or absence of causal paths and statistical associations [4, 5] and datasets sampled with the case-control design [2], i.e., with known selection bias. Integrative Causal Analysis seems promising in the co-analysis of heterogeneous data and knowledge. This initial body of work offers ideas and directions for extensions that will be robust, practical, general, and accommodate various types of heterogeneity.

[1] Sofia Triantafillou, Ioannis Tsamardinos, "**Constraint-based Causal Discovery from Multiple Interventions over Overlapping Variable Sets**", (to appear) Journal of Machine Learning Research

[2] Giorgos Borboudakis, Ioannis Tsamardinos, "**Bayesian Network Learning with Discrete Case-Control Data**", (to appear) Uncertainty in Artificial Intelligence (UAI), 2015

[3] Sofia Triantafillou, Ioannis Tsamardinos, and Anna Roumpelaki, "**Learning Neighborhoods of High Confidence in Constraint-Based Causal Discovery**", the Seventh European Workshop on Probabilistic Graphical Models (PGM), 2014

[4] Giorgos Borboudakis, Ioannis Tsamardinos, "**Scoring and Searching over Bayesian Networks with Informative, Causal and Associative Priors**", Uncertainty in Artificial Intelligence (UAI) 2013

[5] Giorgos Borboudakis, Ioannis Tsamardinos, "**Incorporating Causal Prior Knowledge as Path-Constraints in Bayesian Networks and Maximal Ancestral Graphs**", International Conference in Machine Learning (ICML), 2012

[6] Ioannis Tsamardinos, Sofia Triantafillou, Vincenzo Lagani, "**Towards Integrative Causal Analysis of Heterogeneous Datasets and Studies**", Journal of Machine Learning Research 13(Apr):1097−1157, 2012

[7] Sofia Triantafillou, Ioannis Tsamardinos, Ioannis Tollis, "**Learning Causal Structure from Overlapping Variable Sets**", in Y.W. Teh and M. Titterington (Eds.), Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, JMLR: W&CP 9, pp 860-867, 2010, Chia Laguna, Sardinia, Italy, May 13-15, 2010