Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Early Career Track

Bridging Causality and Learning: How Do They Benefit from Each Other?

Mingming Gong
School of Mathematics and Statistics, University of Melbourne, Australia
mingming.gong@unimelb.edu.au

Abstract

Modern machine learning techniques can discover complicated statistical dependencies between random variables, usually in the form a statistical model, and make use of these dependencies to perform predictions on future observations. However, many real problems involve causal inference, which aims to infer how the data generating system should behave under changing conditions. To perform causal inference, we need not only statistical dependencies but also causal structures to determine the system's behavior under external interventions. In this paper, I will be focusing on two essential problems that bridge causality and learning and investigate how they can benefit from each other. On the one hand, since conducting randomized controlled experiments for causal structure discovery is often expensive or infeasible, it would be valuable to investigate how we can explore modern machine learning algorithms to search for causal structures from observational data. On the other hand, since causal structure provides information about the distribution change properties, it can be used as a fundamental tool to tackle a major challenge for machine learning: the capability of generalization to new distributions and prediction in nonstationary environment.

1 Introduction

The main goal of statistics and machine learning is to discover associations between random variables, and these dependencies can be used to perform prediction on future observations. Statistics has a long tradition in dealing with co-occurring events. For example, in medical statistics, scientists might want to find abnormal cell morphology from red blood cells that co-occur with a disease. Discovering this dependence can lead to the prediction of the occurrence of the disease based on a blood test. Built on statistical inference, machine learning exploits the vast computational power of modern computing platforms and develops efficient algorithms to solve complex real-world problems. During the last decade, machine learning has made spectacular progress, approaching or even surpassing human performance in complex tasks such as visual recognition [He *et al.*2015, Szegedy *et al.*2015], speech recognition [Hannun *et al.*2014], and computer gaming [Silver *et al.*2016].

In many machine learning problems, it is sufficient to model the dependencies between random variables in a probability distribution if the underlying data generating system is stable. However, many studies in scientific research and decision making aim to answer questions involving prediction under *interventions* outside the data-generating system [Pearl2000, Spirtes *et al.*2001]. For example, epidemiologists may wonder whether life expectancy will change if they advise people to change their diets. To predict the consequences of interventions, we need to understand the underlying mechanisms that generate the data. Causal relationships among variables provide coarse descriptions of the mechanisms, at a level sufficient to predict the system's behavior under interventions [Pearl and others2009]. The causal relationships encode asymmetric information between variables: by intervening on the cause we can control its effect, but not vice versa.

There are three major problems in causal inference: 1) mathematical formulation of causality, 2) identify and estimate causal effects, and 3) causal structure discovery. Mathematical formulation and representation of causation are fundamental to automated causal inference. Neyman–Rubin causal model [Rubin2005], Granger causality model [Granger1969], and the structural causal model (SCM) [Pearl2000] are three well known causal models that formulate causality from different perspectives. In particular, the structural causal model unifies other models and provides a coherent mathematical foundation for causal analysis. Identifying and estimating the effects of interventions (or causal effects) given a causal graph and empirical data, is our final goal of causal inference [Tian and Pearl2002, Shpitser and Pearl2006, Jaber *et al.*2019]. Discovering the causal structure is of practical importance because the causal structure is not always known in real applications.

In this paper, I will be focusing on two essential problems that bridge causality and learning and investigate how they can benefit from each other. On the one hand, since causality can be considered as directional statistical dependencies, causal inference naturally builds on the statistical associations and thus, many modern machine learning tech-

niques can be explored for causal analysis. One typical example would be causal discovery from observational data. This problem can be viewed as a model selection problem that aims to search for the optimal causal graph explaining the data, leading to score-based causal discovery methods [Chickering2002]. As such, we can develop more advanced causal discovery approaches by exploiting modern learning techniques such as kernel methods and reinforcement learning [Huang *et al.*2018, Zhu and Chen2019]. Along this direction, we have developed new approaches for causal discovery from time series by exploring identifiable unsupervised learning methods that provide theoretical guarantees of the uniqueness of the learned causal graph [Gong* *et al.*2015, Gong *et al.*2017, Geiger *et al.*2015, Huang *et al.*2019, Chenwei *et al.*2019].

On the other hand, since causal structure provides information about the distribution changing properties, it can be used as a fundamental tool to tackle a major challenge for machine learning: the capability of generalization to new distributions and prediction in nonstationary environment. Motivated by the 'independence' postulate, which states that P_{cause} and $P_{effect|cause}$ change independently, we have performed a series of studies on domain adaptation from a causal perspective [Zhang *et al.*2015, Gong *et al.*2016, Gong *et al.*2018, Zhang *et al.*2020]. In particular, we use causal models to represent the relationship between the features and labels and consider possible situations where different modules of the causal model change with the domain. We show that causal structures provide a powerful tool to estimate how the distribution changes across domains and develop efficient algorithms for adaptive learning. For other types of learning problems that can benefit from causality, please refer to a recent paper on causality for machine learning [Schölkopf2019].

In the following two sections, I will briefly introduce our representative research works on learning causal structures from low-resolution time series and causal adaptive learning.

2 Learning Causal Relations from Low-resolution Time Series

Traditional approaches to causal inference from time-series data rely on the assumption that cause precedes effect, and in particular, Granger causality [Granger1969] is a widely applied causal discovery method for time series. However, Granger causality may be very different from true causality when the measurements are obtained at a coarser timescale (data resolution) than the underlying causal time scale (causal resolution). Recovering causal relations at the causal resolution from the lower-resolution data is difficult due to the information loss in the missing observations.

For Granger causal analysis in the linear case, one fits the following VAR model [Sims1980] on the data:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{e}_t, \tag{1}$$

where $\mathbf{x}_t = (x_{1t}, x_{2t}, ..., x_{nt})^\intercal$ is the vector of the observed data, $\mathbf{e}_t = (e_{1t}, ..., e_{nt})^\intercal$ is the temporally and contemporaneously independent noise process, and $\mathbf{A}$ contains the temporal causal relations. Suppose that due to the low resolution of the data, there is an observation every k time steps. That is, the low-resolution observations $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, , ..., \tilde{\mathbf{x}}_t)$ are $(\mathbf{x}_1, \mathbf{x}_{1+k}, ..., \mathbf{x}_{1+(t-1)k})$; here we have assumed that the first sampled point is $\mathbf{x}_1$. We then have

$$\begin{aligned}\tilde{\mathbf{x}}_{t+1} &= \mathbf{x}_{1+tk} = \mathbf{A}\mathbf{x}_{1+tk-1} + \mathbf{e}_{1+tk} \\ &= \mathbf{A}(\mathbf{A}\mathbf{x}_{1+tk-2} + \mathbf{e}_{1+tk-1}) + \mathbf{e}_{1+tk} \\ &= ... \\ &= \mathbf{A}^k\tilde{\mathbf{x}}_t + \sum_{l=0}^{k-1}\mathbf{A}^l\mathbf{e}_{1+tk-l}. \end{aligned} \tag{2}$$

We denote by $\vec{\mathbf{e}}_{t+1}$ the noise term, i.e., $\vec{\mathbf{e}}_{t+1} = \sum_{l=0}^{k-1}\mathbf{A}^l\mathbf{e}_{1+tk-l}$. We call $(\mathbf{A}, \mathbf{e}_t, k)$ the representation of the kth order subsampled time series $\tilde{\mathbf{x}}_t$. $\mathbf{A}$ is called the causal transition matrix. Then we provide solutions to the following two problems.

- *Identifiability of causal relations from low-resolution data*
  Due to information loss in the low-resolution data, estimating causal relations suffers from the identifiability problem: under the commonly adopted Gaussianity assumption of the data, the solutions are generally not unique. This non-identifiability problem has existed for decades, and many researchers believed that identifying causal relations from low-resolution data is hopeless. We proved that, however, if the noise terms are non-Gaussian, the underlying model for the high-resolution data is identifiable from low-resolution data under mild conditions [Gong* *et al.*2015, Gong *et al.*2017].
- *Estimation of causal relations by unsupervised learning models*
  We proposed practical algorithms for causal discovery from low-resolution data by transforming our model into non-Gaussian latent variable models. Then we can develop efficient learning procedures by variational inference [Gong* *et al.*2015] or adversarial learning [Chenwei *et al.*2019].

3 Causal Adaptive Learning

A common assumption in standard supervised learning is that the training and test set follow the same distribution. However, this standard assumption is often violated in many real-world applications due to intrinsic nonstationarity of the environment or inevitable sample selection bias. We propose causal domain adaptation methods for adaptive learning under distribution changes by injecting causal knowledge in the learning process.

Unsupervised domain adaptation aims to construct the classifier for the target/test domain, on which only the features $\mathbf{x}^\tau = (x_k^\tau)_{k=1}^m$ are available, by making proper use of the data in s source/training domains $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = (\mathbf{x}_k^{(i)}, y_k^{(i)})_{k=1}^{m_i}$, $i = 1, ..., s$. Traditional graphical models provide a compact, yet flexible, way to decompose the joint distribution of as a product of simpler, lower-dimensional factors [Pearl1994, Koller and Friedman2009], as a consequence of conditional independence relations between the variables. In particular, if the conditional independence relations can

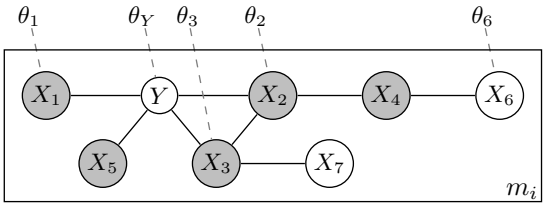

Figure 1: An augmented DAG over Y and X_i. See main text for its interpretation.

be presented by a Directed Acyclic Graph (DAG), then the joint distribution can be factorized as the product of factors, each of which corresponds to the conditional distribution of a variable given its parents. For our purpose, we need encode not only conditional independence relations between the variables, but also whether the conditional distributions change across domains. To this end, we propose an augmented DAG as a flexible yet compact way to describe how a joint distribution changes across domains, assuming that the distributions in all domains can be represented by such a graph (in other words, this graph encodes the property of the mother distribution behind the domain-specific distributions). It is an augmented graph in the sense that it is over not only features X_i and Y, but also external latent variables $\boldsymbol{\theta}$.

Figure 1 gives an example of such a graph. Nodes in gray are in the Markov Blanket (MB) of Y. The $\boldsymbol{\theta}$ variables are mutually independent, and take the same value across all data points within each domain and may take different values across domains. They indicate the property of distribution shift–for any variable with a $\boldsymbol{\theta}$ variable directly into it, its conditional distribution given its parents (implied by the DAG over X_i and Y) depends on the corresponding $\boldsymbol{\theta}$ variable, and hence may change across domains. In other words, the distributions across domains differ only in the values of the $\boldsymbol{\theta}$ variables. Once their values are given, the domain-specific joint distribution is given by $P(\mathbf{X}, Y \mid \boldsymbol{\theta})$, which can be factorized according to the augmented DAG. By making use of such graphical representations, we can formulate domain adaptation as an inference problem on graphical models and reconstruct the joint distribution $P(\mathbf{X}, Y)$ in the target domain by transferring supervision information from source domains.

4 Conclusion

In this paper, I briefly introduced our recent research on causality and machine learning. On the one hand, we explore modern machine learning techniques to learn causal structures from observational data. On the other hand, we explore causal knowledge for adaptive learning in nonstationary environment.

References

[Chenwei *et al.*, 2019] DING Chenwei, Mingming Gong, Kun Zhang, and Dacheng Tao. Likelihood-free overcomplete ica and applications in causal discovery. In *Advances in Neural Information Processing Systems*, pages 6880–6890, 2019.

[Chickering, 2002] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[Geiger *et al.*, 2015] Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, and Dominik Janzing. Causal inference by identification of vector autoregressive processes with hidden components. In *International Conference on Machine Learning*, pages 1917–1925, 2015.

[Gong* *et al.*, 2015] M. Gong*, K. Zhang*, D. Tao, P. Geiger, and B. Schölkopf. Discovering temporal causal relations from subsampled data. In *Proc. 32th International Conference on Machine Learning (ICML 2015)*, 2015.

[Gong *et al.*, 2016] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *Proceedings of the 33nd International Conference on Machine Learning (ICML 2016)*, volume 48, pages 2839–2848, 2016.

[Gong *et al.*, 2017] Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. Causal discovery from temporally aggregated time series. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2017. NIH Public Access, 2017.

[Gong *et al.*, 2018] Mingming Gong, Kun Zhang, Biwei Huang, Clark Glymour, Dacheng Tao, and Kayhan Batmanghelich. Causal generative domain adaptation networks. *arXiv preprint arXiv:1804.04333*, 2018.

[Granger, 1969] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[Hannun *et al.*, 2014] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[Huang *et al.*, 2018] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

[Huang *et al.*, 2019] Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. *Proceedings of machine learning research*, 97:2901, 2019.

[Jaber *et al.*, 2019] Amin Jaber, Jiji Zhang, and Elias Bareinboim. Identification of conditional causal effects under markov equivalence. In *Advances in Neural Information Processing Systems*, pages 11512–11520, 2019.

[Koller and Friedman, 2009] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

[Pearl and others, 2009] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[Pearl, 1994] Judea Pearl. A probabilistic calculus of actions. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 454–462. Morgan Kaufmann Publishers Inc., 1994.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

[Rubin, 2005] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[Schölkopf, 2019] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[Shpitser and Pearl, 2006] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference, AAAI-06/IAAI-06*, pages 1219–1226, 2006.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Sims, 1980] C. A. Sims. Macroeconomics and reality. *Econometrica*, 48:1–48, 1980.

[Spirtes *et al.*, 2001] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2001.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[Tian and Pearl, 2002] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

[Zhang *et al.*, 2015] K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3150–3157. AAAI Press, 2015.

[Zhang *et al.*, 2020] Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, and Clark Glymour. Domain adaptation as a problem of inference on graphical models. *arXiv preprint arXiv:2002.03278*, 2020.

[Zhu and Chen, 2019] Shengyu Zhu and Zhitang Chen. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.