Continual Causal Inference: Challenges and Opportunities

Anonymous submission

Introduction

A further understanding of cause and effect within observational data is critical across many domains, such as economics, health care, public policy, web mining, online advertising, and marketing campaigns (Yao et al. 2021). Although significant advances have been made to overcome the challenges in causal effect estimation with observational data, such as missing counterfactual outcomes and selection bias between treatment and control groups, the existing methods mainly focus on source-specific and stationary observational data. Such learning strategies assume that all observational data are already available during the training phase and from only one source (Chu, Rathbun, and Li 2020).

Along with the fast-growing segments of industrial applications, this assumption is unsubstantial in practice. Taking Alipay as an example, which is one of the world's largest mobile payment platforms and offers financial services to billion-scale users, a tremendous amount of data containing much privacy-related information is produced daily and collected from different cities or countries. In conclusion, the following two points are summed up. The first one is based on the characteristics of observational data, which are incrementally available from non-stationary data distributions. For instance, the number of electronic financial records for one marketing campaign is growing every day, or the electronic financial records for one marketing campaign may be collected from different cities or even other countries. This characteristic implies that one cannot have access to all observational data at a one-time point and from one single source. The second reason is based on the realistic consideration of accessibility. For example, when new observational data are available, if we want to refine the model previously trained by original data, maybe the original training data are no longer accessible due to a variety of reasons, e.g., legacy data may be unrecorded, proprietary, the sensitivity of financial data, too large to store, or subject to privacy constraint of personal information (Zhang et al. 2020). This practical concern of accessibility is ubiquitous in various academic and industrial applications. That's what it boiled down to: in the era of big data, we face new challenges in causal inference with observational data, i.e., the **extensibility** for incrementally available observational data, the **adaptability** for extra domain adaptation problem except for the imbalance between treatment and control groups, and the **accessibility** for an enormous amount of data.

In this position paper, we formally define the problem of continual treatment effect estimation, describe its research challenges, and then present possible solutions to this problem. Moreover, we will discuss future research directions on this topic.

Problem Definition

Suppose that the observational data contain n units collected from d different domains and the d-th dataset D_d contains the data $\{(x, y, t)|x \in X, y \in Y, t \in T\}$ collected from d-th domain, which contains n_d units. Let X denote all observed variables, Y denote the outcomes in the observational data, and T be a binary variable. Let $D_{1:d} = \{D_1, D_2, ..., D_d\}$ be the set of combination of d dataset, separately collected from d different domains. For d datasets $\{D_1, D_2, ..., D_d\}$, they have the commonly observed variables, but due to the fact that they are collected from different domains, they have different distributions with respect to X, Y, and T in each dataset. Each unit in the observational data received one of two treatments. Let t_i denote the treatment assignment for unit i; $i = 1, ..., n$. For binary treatments, $t_i = 1$ is for the treatment group and $t_i = 0$ for the control group. The outcome for unit i is denoted by y_t^i when treatment t is applied to unit i. For observational data, only one of the potential outcomes is observed. The observed outcome is called the factual outcome, and the remaining unobserved potential outcomes are called counterfactual outcomes.

This task can follow the potential outcome framework for estimating treatment effects (Rubin 1974; Splawa-Neyman, Dabrowska, and Speed 1990). The individual treatment effect (ITE) for unit i is the difference between the potential treated and control outcomes, and is defined as $\text{ITE}_i = y_1^i - y_0^i$. The average treatment effect (ATE) is the difference between the mean potential treated and control outcomes, which is defined as $\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}(y_1^i - y_0^i)$. The success of the potential outcome framework is based on the following assumptions (Imbens and Rubin 2015), which ensure that the treatment effect can be identified. *Stable Unit Treatment Value Assumption (SUTVA)*: The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. *Consistency*: The potential outcome of treatment

t is equal to the observed outcome if the actual treatment received is t. *Positivity*: For any value of x, treatment assignment is not deterministic, i.e.,$P(T = t|X = x) > 0$, for all t and x. *Ignorability*: Given covariates, treatment assignment is independent of the potential outcomes, i.e., $(y_1, y_0) \perp\!\!\!\perp t|x$.

Our goal is to develop a novel continual causal inference framework to estimate the causal effect for all available data, including new data D_d and the previous data $D_{1:(d-1)}$, without having access to previous data $D_{1:(d-1)}$.

Research Challenges

Existing causal effect inference methods, however, are unable to deal with the aforementioned new challenges, i.e., extensibility, adaptability, and accessibility. Although it is possible to adapt existing causal inference methods to cater to these issues, these adjusted methods still have inevitable defects. Three straightforward adaptation strategies are described as follows: (1) If we directly apply the model previously trained based on original data to new observational data, the performance on new task will be very poor due to the domain shift issues among different data sources; (2) Suppose we utilize newly available data to re-train the previously learned model for adapting changes in the data distribution. In that case, old knowledge will be completely or partially overwritten by the new one, which can result in severe performance degradation on old tasks. This is the well-known *catastrophic forgetting* problem (McCloskey and Cohen 1989; French 1999); (3) To overcome the catastrophic forgetting problem, we may rely on the storage of old data and combine the old and new data together, and then retrain the model from scratch. However, this strategy is memory inefficient and time-consuming, and it brings practical concerns such as copyright or privacy issues when storing data for a long time (Samet, Miri, and Granger 2013). Any of these three strategies, in combination with the existing causal effect inference methods, are deficient.

Potential Solution

To address the continual treatment effect estimation problem, we propose a **C**ontinual Causal **E**ffect **R**epresentation **L**earning framework (CERL) for estimating causal effect with incrementally available observational data. Instead of having access to all previous observational data, we only store a limited subset of feature representations learned from previous data. Combining selective and balanced representation learning, feature representation distillation, and feature transformation, our framework preserves the knowledge learned from previous data and updates the knowledge by leveraging new data so that it can achieve the continual causal effect estimation for incrementally new data without compromising the estimation capability for previous data.

Framework Overview. To estimate the incrementally available observational data, the framework of CERL is mainly composed of two components: (1) the baseline causal effect learning model is only for the first available observational data, and thus we don't need to consider the domain shift issue among different data sources. This component is equivalent to the traditional causal effect estimation problem; (2) the continual causal effect learning model is for the sequentially available observational data, where we need to handle more complex issues, such as knowledge transfer, catastrophic forgetting, global representation balance, and memory constraint.

Baseline Causal Effect Learning Model. We first train the baseline causal effect learning model for the initial observational dataset and then bring in subsequent datasets. The task on the initial dataset can be converted to a traditional causal effect estimation problem. Owing to the success of deep learning for counterfactual inference (Shalit, Johansson, and Sontag 2017), we propose to learn the selective and balanced feature representations for units in treatment and control groups, and then infer the potential outcomes based on learned representation space.

Sustainability of Model Learning. To avoid catastrophic forgetting when learning new data, we propose to preserve a subset of lower-dimensional feature representations rather than all original covariates. We also can adjust the number of preserved feature representations according to the memory constraint.

Continual Causal Effect Learning. We have stored memory and the baseline model. To continually estimate the causal effect for incrementally available observational data, we incorporate feature representation distillation and feature representation transformation to estimate the causal effect for all seen data based on a balanced global feature representation space.

Research Opportunities

Although significant advances have been made to overcome the challenges in causal effect estimation from an academic perspective, industrial applications based on observational data are always more complicated and harder. Unlike source-specific and stationary observational data, most real-world data are incrementally available and from non-stationary data distributions. Significantly, we also face the realistic consideration of accessibility. This work is the first attempt to investigate the continual lifelong causal effect inference problem and propose the corresponding evaluation criteria. However, constructing the comprehensive analytical tools and the theoretical framework derived from this brand-new problem requires non-trivial efforts. Specifically, there are four potential directions for continual causal inference: (1) The basic assumptions for traditional causal effect estimation may not be completely applicable. New assumptions may be supplemented, or previous assumptions need to be relaxed. (2) There exists a natural connection with continual domain adaptation among different times or domains ("continual" causal inference) and between treatment and control groups (continual "causal inference"). (3) Compared to traditional causal effect estimation tasks based on a small amount of medical data, the continual causal inference method will face big data computing or cloud computing due to its objective task. (4) With the increasing public concern over privacy leakage in data, federated learning, which collaboratively trains the machine learning model without directly sharing the raw data among the data holders, may become a potential solution for continual causal inference.

References

Chu, Z.; Rathbun, S.; and Li, S. 2020. Continual Lifelong Causal Effect Inference with Real World Evidence.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Samet, S.; Miri, A.; and Granger, E. 2013. Incremental learning of privacy-preserving Bayesian networks. *Applied Soft Computing*, 13(8): 3657–3667.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085. PMLR.

Splawa-Neyman, J.; Dabrowska, D. M.; and Speed, T. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 465–472.

Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; and Zhang, A. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5): 1–46.

Zhang, J.; Zhang, J.; Ghosh, S.; Li, D.; Tasci, S.; Heck, L.; Zhang, H.; and Kuo, C.-C. J. 2020. Class-incremental learning via deep model consolidation. In *The IEEE Winter Conference on Applications of Computer Vision*, 1131–1140.