Counterfactual Reasoning in Observational Studies

Negar Hassanpour
University of Alberta, Alberta Machine Intelligence Institute
Edmonton, Alberta, Canada
hassanpo@ualberta.ca

Introduction

To identify the appropriate action to take, an intelligent agent must infer the causal effects of every possible action choices. A prominent example is precision medicine, that attempts to identify which medical procedure ($t \in \mathcal{T}$) will benefit each individual patient (x) the most. This requires answering counterfactual questions (Rubin 1974; Pearl 2009) such as: *"Would this patient have lived longer, had she received an alternative treatment?"*. There are several reasons why this is more challenging than conventional supervised machine learning:

First, for the i^{th} patient, the training data only contains the observed outcome $y_i^{t_i}$ of the administered treatment t_i but never the outcome(s) of the alternative treatment(s) $\neg t_i \in \mathcal{T} \setminus \{t_i\}$ – *i.e.*, counterfactual outcome(s). This challenge can be mitigated if we are allowed to perform experimentation (on-line exploration), or have access to a randomized controlled trial (RCT) dataset (Pearl 2009). In many cases, however, conducting an experiment or RCT is expensive, impractical, or even infeasible. As a result, we are forced to approximate causal effects from data that is available: off-line datasets collected through observational studies. Such datasets, however, often exhibit *sample selection bias* (Imbens and Rubin 2015). That is, the treatment t assignment procedure depends on some or all of the attributes x of the individual – *i.e.*, $\Pr(t \mid x) \neq \Pr(t)$. This becomes detrimental to the accuracy and confidence of counterfactuals prediction when the same covariates that determine treatment also [partially] determine the outcome y. Existence of such confounders is the **second** challenge with causal inference.

There are two categories of performance measures for evaluating causal inference methods. **Population**-based measures help find the one treatment that works [somewhat] well for the entire population. **Individual**-based measures, in contrast, focus on identifying the best treatment option personalized to each individual x. An example of individual-based measure – that requires estimating outcomes of all possible treatments – is Precision in Estimation of Heterogeneous Effect (Hill 2011): $\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{e}_i - \hat{\text{e}}_i)^2}$, where $\text{e}_i = y_i^1 - y_i^0$ is the true Individual Treatment Effect (ITE) for individual i, $\hat{\text{e}}_i = \hat{y}_i^1 - \hat{y}_i^0$ is the estimated ITE, and n is the sample size.

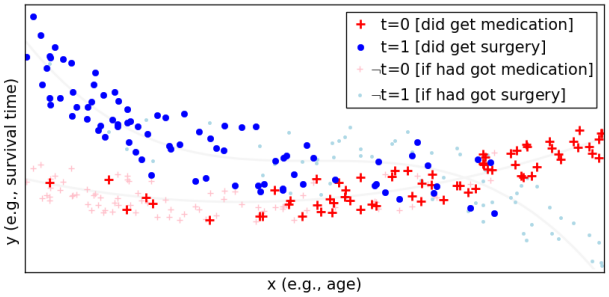

Figure 1: An example observational dataset (best viewed in color). Here, to treat heart disease, a doctor prescribes surgery $(t=1)$ to younger patients (•) and medication $(t=0)$ to older ones (+) – hence we have sample selection bias. The unobservable counterfactuals are illustrated by small and faint • (for $\neg t=1$) and + (for $\neg t=0$).

Research Questions

In my PhD, I will explore ways to address the above-mentioned challenges associated with causal effect estimation; with a focus on devising methods that enhance performance according to the individual-based measures. Specifically, my Research Questions (RQs) are the following:

1. The first challenge makes it impossible to properly evaluate the proposed methods with real-world observational datasets, since the ground truth for counterfactuals are in fact unobservable. Therefore, we require algorithms that can generate realistic synthetic observational datasets that exhibit various degrees of sample selection bias.

The remaining RQs are related to the second challenge:

2. Learning a representation space (Bengio et al. 2013) Φ, shared between treatment arms, is a good strategy to reduce sample selection bias. This is effective if distributions of the transformed instances $\Phi(x)$ belonging to every treatment arm are similar – making the dataset close to an RCT. Reasonably, it should be possible to further alleviate the bias by incorporating appropriate weighting schemes.
3. Employ generative models to create new virtual samples that can fill in the gap in order to help better predicting

the counterfactuals. Methods such as Variational Auto-Encoder (Kingma and Welling 2014) and Generative Adverserial Network (Goodfellow et al. 2014)

4. Many observational datasets include censored instances.[1] These instances should not be ignored as they might embed important information regarding the effectiveness of a certain treatment (*e.g.*, a highly effective medication might make follow-up unlikely). We need survival prediction methods that can exploit censored data for achieving a higher performance in causal effect estimation.

5. Devise methods that accommodate counterfactual regression when more than two treatment options are available – *i.e.*, (i) categorical (*e.g.*, $\mathcal{T} = \{\text{bypass}, \text{stent}, \text{medication}\}$ for curing heart disease), (ii) multiple-binary (*e.g.*, $\mathcal{T} = \{0, 1\}^k$ – *i.e.*, combination[2] of a subset of medications for controlling depression), or (iii) $\mathcal{T} \subseteq \mathbb{R}$ (*e.g.*, the right dosage of insulin for a diabetic patient).

Related Work

Learning treatment effects from observational studies is closely related to "off-policy learning in contextual bandits" and "learning from logged bandit feedback" (Swaminathan and Joachims 2015). A common family of statistical methods use weighting to handle this sample selection bias. An example is Inverse Propensity Scoring (IPS), which tries to balance the source (observed) and target (counterfactual) distributions. Such methods however neglect the fact that controlling for the covariates that only determine the treatment (not outcome)[3] can have a negative impact on the accuracy of ITE predictions. Because the weights calculated according to such factors are irrelevant and perhaps detrimental to predicting accurate outcomes (both observed and counterfactual).

Another family of methods use representation learning to find a representation space Φ that reduces the sample selection bias between treatment arms. For example, Atan et al. (2018) learn Φ using an auto-encoder that tries to minimize the cross entropy loss between $\Pr(\, t \,)$ and $\Pr(\, t \mid \Phi(\, x \,) \,)$. However, by training an auto-encoder, they force their representation space to be able to reproduce *all* the covariates in x from Φ – some of which might have had no effect on determining the observed outcome. In another work, Shalit et al. (2017) learn Φ such that $\Pr(\, x \mid t = 0 \,)$ and $\Pr(\, x \mid t = 1 \,)$ are as close to each other as possible, provided that $\Phi(\, x \,)$ retains enough information that a learned regression model $h^t(\, \Phi \,)$ (with t being the administrated treatment) can generalize well on the observed outcomes. Neither of these methods, however, employ a context-aware weighting scheme to further alleviate the sample selection bias.

Research Plan

I have already addressed **RQ#1** in (Hassanpour and Greiner 2018) and **RQ#2** whose paper is in preparation; planned to be submitted to the IJCAI 2019. Table 1 summarized the tentative timeline for the remainder of my PhD program.

Table 1: My PhD tentative timeline

TIME	MILESTONE
2019 Feb.	Publish results for **RQ#3** at one of NeurIPS
2019 Apr.	Defend Proposal (Candidacy Exam)
2019 Sep.	Publish results for **RQ#4** at one of AAAI / AISTAT
2020 Feb.	Publish results for **RQ#5** at one of IJCAI / ICML / UAI / NeurIPS
2020 May	Write up my PhD Dissertation
2020 Aug.	Defend my PhD Dissertation

References

Atan, O.; Jordon, J.; and van der Schaar, M. 2018. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *AAAI*, 2071–2078.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE TPAMI* 35(8):1798–1828.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*, 2672–2680.

Hassanpour, N., and Greiner, R. 2018. A novel evaluation methodology for assessing off-policy learning methods in contextual bandits. In *Canadian AI*, 31–44.

Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1):217–240.

Imbens, G. W., and Rubin, D. B. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Pearl, J. 2009. *Causality*. Cambridge University Press.

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

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, 3076–3085.

Swaminathan, A., and Joachims, T. 2015. Batch learning from logged bandit feedback through counterfactual risk minimization. *JMLR* 16.

[1] That is, when some of the outcomes (time to *event* – *e.g.*, death) are only partially known (*e.g.*, we know the i^{th} patient survived 5 months but do not know if she died a day later, a year, or ...), either because the respective patients have lost to follow-up, or the data collection period has ended without the *event* being occurred.

[2] However, the algorithm should be mindful of drug interactions; since some combinations might neutralize the effect of the treatment or worse, be detrimental to the patient's health.

[3] For example, a doctor might be less likely to prescribe an expensive treatment to poor patients (thus, imposing sample selection bias in the data); although, outcomes of the possible treatments are not particularly dependent on the patients' wealth status.