Counterfactual Cross-Validation

Yuta Saito
Tokyo Institute of Technology
saito.y.bj@m.titech.ac.jp

Shota Yasui
CyberAgent, Inc.
yasui_shota@cyberagent.co.jp

ABSTRACT

There is intense attention in applying machine learning to make causal inference in fields such as marketing, economics, and education. In particular, individual-level treatment effect (ITE) prediction has important applications such as personalized recommendation and precision medicine. Most of the previous papers study prediction methods of the ITE. On the other hand, how to evaluate given ITE predictors from observable data has not yet been thoroughly investigated despite that it plays the critical part of causal inference. In this paper, we propose a method to effectively select the best ITE predictors from a set of candidates using only observational validation set.

ACM Reference Format:
Yuta Saito and Shota Yasui. 2019. Counterfactual Cross-Validation. In *Proceedings of the ACM Recsys Workshop on Reinforcement and Robust Estimators for Recommendation (REVEAL'19).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Predicting Individual-level Treatment Effects (ITE) of treatments is essential to optimize the metric of interest in various domains. For example, recommendation systems wish to recommend items having positive causal effects on users preferences to maximize user experiences.

Most of the previous studies propose machine-learning based ITE prediction methods and achieve promising results on some benchmark datasets. On the other hand, the evaluation of these prediction methods is another essential step to conduct valid model selection and hyperparameter tuning. However, the evaluation problem of ITE prediction models has not much yet studied in spite of its importance.

In this paper, we propose a model validation procedure called *CounterFactual Cross-Validation (CF-CV)* that accurately ranks the performance of ITE predictors with high confidence using only factual validation set. Our proposed evaluation procedure satisfies desirable theoretical properties to be used for model selection or hyperparameter tuning of ITE predictors. Moreover, experimental results show that our proposed method effectively ranks candidate ITE predictors and select a better predictor among a set of candidates. Besides, the model selection performance of our method is stable, and this is critical because we never know the true performance of any ITE predictors due to the existence of counterfactuals.

2 PROBLEM FORMULATION

We denote $X \in \mathcal{X}$ as the feature vector and $T \in \{0, 1\}$ as a binary treatment assignment indicator. Here, we follow the potential outcome framework [4] and assume that there exist two potential outcomes denoted as $\left(Y^{(0)}, Y^{(1)}\right) \in \mathcal{Y} \times \mathcal{Y}$ for each individual. $Y^{(0)}$ is a potential outcome associated with $T = 0$, on the other hand, $Y^{(1)}$ is associated with $T = 1$.

Now let us formally define the **Individual-level Treatment Effect (ITE)** for an individual with a feature vector $x \in \mathcal{X}$ as:

$$\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \,|\, X = x\right] \tag{1}$$

Next, we define the conditional probability of treatment assignment as $e(x) = \mathbb{P}(T = 1 \,|\, X = x)$. This parameter is called **propensity score** in causal inference and widely used to estimate treatment effects from observational data [4]. Throughout this paper, we make the standard assumptions in causal inference including *Unconfoundedness*, *Overlap*, and *Consistency* [4].

In previous studies [2, 5], the evaluation of an ITE predictor $\hat{\tau}(\cdot)$ is formulated as accurately estimating the following metric from observational validation dataset $\mathcal{V} = \{X_i, T_i, Y_i^{obs}\}$ as:

$$\mathcal{R}_{true}(\hat{\tau}) = \mathbb{E}_X\left[(\tau(X) - \hat{\tau}(X))^2\right] \tag{2}$$

Here, $\mathcal{R}_{true}$ is the true performance metric of an ITE predictor $\hat{\tau}(\cdot)$. In this paper, we aim to construct a performance estimator $\widehat{\mathcal{R}}(\hat{\tau})$ satisfying the following condition:

$$\mathcal{R}_{true}(\hat{\tau}) \leq \mathcal{R}_{true}(\hat{\tau}') \Rightarrow \widehat{\mathcal{R}}(\hat{\tau}) \leq \widehat{\mathcal{R}}(\hat{\tau}'), \; \forall \, \hat{\tau}, \hat{\tau}' \in \mathcal{M}. \tag{3}$$

where $\mathcal{M} = \{\hat{\tau}_1, ..., \hat{\tau}_{|\mathcal{M}|}\}$ is a set of candidate ITE predictors.

An estimator satisfying Eq. (3) gives accurate **ranking** of candidate predictors by the true metric values, and one can select the best model among $\mathcal{M}$.

3 METHOD

To achieve our goal, we consider the following feasible estimator of the performance metric:

$$\widehat{\mathcal{R}}(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}(X_i) - \hat{\tau}(X_i))^2 \tag{4}$$

where $\tilde{\tau}(\cdot)$ is called **the oracle** and is constructed from $\mathcal{V}$. Under our formulation, we aim to answer the question: *What is the best plug-in oracle to rank the performance of given candidate ITE predictors from observational validation dataset?*

3.1 Proposed Oracle

Here we define our proposed oracle inspired by the doubly robust estimator used to estimate average causal effects of treatments [1].

REVEAL'19, September 20th, 2019, Copenhagen, Denmark

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Definition 3.1. Let $f(\cdot,\cdot): \mathcal{X}\times\mathcal{T} \rightarrow \mathcal{Y}$ be a hypothesis predicting potential outcomes and is defined as $f(x,t) = h(\Phi(x), t)$. Then, the doubly robust oracle for a given data (X, T, Y^{obs}) is defined as:

$$\tilde{\tau}(X,T,Y^{obs}) = \frac{T}{e(X)}\left(Y^{obs} - f(X,1)\right) - \frac{1-T}{1-e(X)}\left(Y^{obs} - f(X,0)\right) + \left(f(X,1) - f(X,0)\right) \tag{5}$$

where, the loss function to derive a hypothesis h and a representation function Φ is:

$$h, \Phi = \min_{h,\Phi} \frac{1}{n}\sum_{i=1}^{n} w^{(t)}(x_i)\cdot L\left(h\left(\Phi\left(x_i\right), t_i\right), y_i\right) + \lambda\cdot\Omega(h) + \alpha\cdot \mathrm{IPM_G}\left(\{\Phi\left(x_i\right)\}_{i:t_i=0}, \{\Phi\left(x_i\right)\}_{i:t_i=1}\right) \tag{6}$$

where, $L(\cdot,\cdot)$ is squared loss, $\Omega(h)$ is the regularization term for model complexity, and $\mathrm{IPM_G}\left(\{\Phi\left(x_i\right)\}_{i:t_i=0}, \{\Phi\left(x_i\right)\}_{i:t_i=1}\right)$ is called Integral Probability Metric (IPM) measuring distance between two distributions [6, 7]. Finally, $w^{(t)}(x) = \frac{t(1-2e(x))+e(x)^2}{e(x)(1-e(x))}$ is a weighting function depending on the propensity score.

3.2 Summary of Theoretical Results

Here, we state critical theoretical results for our proposed oracle and the resulting performance estimator [1].

Current Theoretical Results

1. The performance estimator based on our oracle in Eq. (5) preserves the difference between the true metric values; the predictor having the smallest expected value of our performance estimator among candidate predictors also has the smallest value of $\mathcal{R}_{true}$ among them.
2. Our oracle minimizes the upper bound of the finite sample uncertainty term in the empirical version of the performance estimator in Eq. (4).

Thus, our oracle is desirable because the performance estimator using our oracle is expected to preserve the difference of the true performance metric and minimizes the upper bound of the finite sample uncertainty; one can expect to select the best ITE predictor among a set of candidates with high confidence.

4 EARLY EXPERIMENTAL RESULT

Here we report the early experimental results on a semi-synthetic dataset comparing the model selection performance of our CF-CV with previous baselines.

Datasets and Setup: We used IHDP[2] dataset provided by [3]. This is a semi-synthetic dataset containing 747 children with 25 features. The detailed description of this dataset can be found in Section 5.1 of [6]. We follow the experimental procedure in [5]; each metric evaluated and ranked pre-trained ITE predictors using only observational validation set, and the performance of each metric was evaluated by the estimated performance of the candidate predictors. We conducted the experimental procedure over 30 realizations with 35/35/30 train/validation/test splits.

Candidate Predictors: We constructed a set of candidate predictors $\mathcal{M}$ by combining five machine learning algorithms[3] and five meta-learners implemented in *EconML* package[4]. Thus, $|\mathcal{M}| = 25$.

Baselines: We compared CF-CV with the following baseline metrics. (1) IPW validation: This metric is proposed in Section 4.2 of [2] and Section 2.3 of [5]. (2) Plug-in validation: It uses predicted values by an arbitrary ITE prediction model as $\tilde{\tau}(\cdot)$ in Eq. (4). We used Counterfactual Regression [6] for $\tilde{\tau}(\cdot)$ to ensure fair comparison. (3) τ-risk: This metric is proposed in [5]. We used Gradient Boosting Regressor to estimate observed outcomes (Y^{obs}).

Results: Table1 reports the averaged and the worst-case performance over 30 realizations. Rank Correlation is the Spearman rank correlation between the ranking by the true performance and the estimated metric values. Relative Root Mean Squared Error (RMSE)[5] is the true performance of the selected model in each metric relative to the best one in $\mathcal{M}$. The results show the effective model selection performance of our CF-CV. Moreover, ours significantly outperformed with respect to the worst-case performance, and this empirically suggests the stability of our metric.

Table 1: Experimental results on IHDP over 30 realizations.

	Rank Correlation		Relative RMSE	
	Avg (± SE)	Worst	Avg (± SE)	Worst
IPW	0.224 (±0.073)	−0.659	2.027 (±0.242)	7.779
τ-risk	−0.399 (±0.051)	−0.797	3.408 (±0.250)	8.884
Plug-in	0.887 (±0.021)	0.385	1.123 (±0.039)	1.841
CF-CV	**0.929 (±0.008)**	**0.830**	**1.040 (±0.019)**	**1.515**

5 CONCLUSION

This paper studies the evaluation problem of ITE prediction models. We proposed the *counterfactual cross-validation* procedure, ensuring accurate model selection with high confidence. The experimental results on IHDP dataset show the effectiveness of our metric.

REFERENCE

[1] Heejung Bang and James M Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 4 (2005), 962–973.

[2] Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*. 1–13.

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

[4] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

[5] Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. 2018. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146* (2018).

[6] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3076–3085.

[7] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. 2012. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics* 6 (2012), 1550–1599.

[1] Formal description of the theoretical results and proofs will be provided in the full version of the paper.

[2] the Infant Health Development Program.

[3] Decision Tree, Random Forest, Gradient Boosting Tree, Ridge Regressor, and Support Vector Regressor with RBF kernel.

[4] https://github.com/microsoft/EconML/blob/master/econml/meta learners.py

[5] We used Relative RMSE because potential outcomes of IHDP dataset have different scales among realizations.