Counterfactual Cross-Validation

Yuta Saito[1], Shota Yasui[2]

1.Tokyo Institute of Technology 2.CyberAgent, Inc.

Tokyo Tech CyberAgent AI Lab

Overview

- The model evaluation of causal effect predictors using observational data has not yet fully investigated despite of its practical importance.
- We develop a method that allows one to select the best model or set of hyper-parameters from many candidates.
- In both model selection and parameter tuning experiments, our proposed approach stably leads to a better causal inference model or set of hyper-parameters than existing metrics.

Problem Setup and Notation

X :feature vector

T :treatment indicator

Following Rubin causal model, we assume there exist two potential outcomes $(Y^{(0)}, Y^{(1)})$ associated with each treatment

One can observe only one of them: $Y^{obs} = TY^{(1)} + (1-T)Y^{(0)}$

Then the **Individual Treatment Effect (ITE)** is: $\tau(X) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X\right]$

Conventional objective is to estimate the following true performance of a predictor $\hat{\tau}(\cdot)$ using only observational validation set $\mathscr{V} = \{X_i, T_i, Y_i^{obs}\}$

Ground Truth Performance: $R_{true} = \mathbb{E}_X\left[\left(\tau(X) - \hat{\tau}(X)\right)^2\right]$

However, in model evaluation, **we only need to know the rank order of the true value of** R_{true} **for candidate predictors.**

Thus, we consider the following new objective:

$$R_{true}(\hat{\tau}) \le R_{true}(\hat{\tau}') \Rightarrow \hat{R}(\hat{\tau}) \le \hat{R}(\hat{\tau}'), \forall \hat{\tau}, \hat{\tau}' \in M$$

where $\hat{R}$ is a performance estimator and M is a set of candidate predictors

- A performance estimator that well satisfies the objective above can accurately **rank** the causal model performance.
- One can select the best model among M with the estimator, even though the true performance of each model R_{true} is remain unknown.

Proposed Performance Estimator

We propose to use the following form of performance estimator with a **doubly robust-style oracle function** $\tilde{\tau}_{DR}(\cdot)$

$$\hat{R}(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\tilde{\tau}_{DR}\left(X_i, T_i, Y_i^{\text{obs}}\right)}_{\text{constructed from a given validation set}} - \hat{\tau}\left(X_i\right)\right)^2$$

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

where $e(X) = \mathbb{P}(T=1 \mid X)$ is the propensity score and

the function $f(x,t) = h(\Phi(x), t)$ is obtained by the following loss function

$$h, \Phi = \min_{h,\Phi} \frac{1}{n}\sum_{i=1}^{n}\left(h\left(\Phi\left(x_i\right), t_i\right) - y_i^{obs}\right)^2 + \alpha \cdot \mathrm{IPM}_G\left(\left\{\Phi\left(x_i\right)\right\}_{i:t_i=0}, \left\{\Phi\left(x_i\right)\right\}_{i:t_i=1}\right)$$

IPM is a distance measure between two distributions

Theoretical Results

Our proposed performance estimator has the following theoretical properties:

1. **The proposed estimator preserves the true ranking of candidate predictors in expectation, i.e.,**

$$R_{true}(\hat{\tau}) \le R_{true}(\hat{\tau}') \Rightarrow \mathbb{E}\left[\hat{R}(\hat{\tau})\right] \le \mathbb{E}\left[\hat{R}(\hat{\tau}')\right], \forall \hat{\tau}, \hat{\tau}' \in M$$

2. **The proposed estimator minimizes the upper bound of the finite sample uncertainty term in model selection.**

Ours is guaranteed to conduct accurate model selection with high confidence

Experimental Results

We conducted model selection and hyper-parameter tuning experiments using a well-known semi-synthetic dataset (the IHDP dataset).

• Model Selection

Procedure:

1. Randomly split the dataset into train/validation/test sets
2. Train 25 candidates ITE predictors on a training set.
3. Rank 25 candidates by each metric on a validation set.
4. The true performances of candidates are measured using a test set.

Performance measures:

Rank Correlation: Spearman rank correlation between metric values and ground truth performances of candidate predictors.

Relative RMSE: the true performance of the selected model in each metric relative to the best one.

$$\text{Relative RMSE} = \frac{R_{true}(\hat{\tau}^*)}{\min_{\hat{\tau}\in M} R_{true}(\hat{\tau})}, \hat{\tau}^* = \arg\min_{\hat{\tau}\in M} \hat{R}(\hat{\tau})$$

Results: **Our metric selects better ITE predictors!**

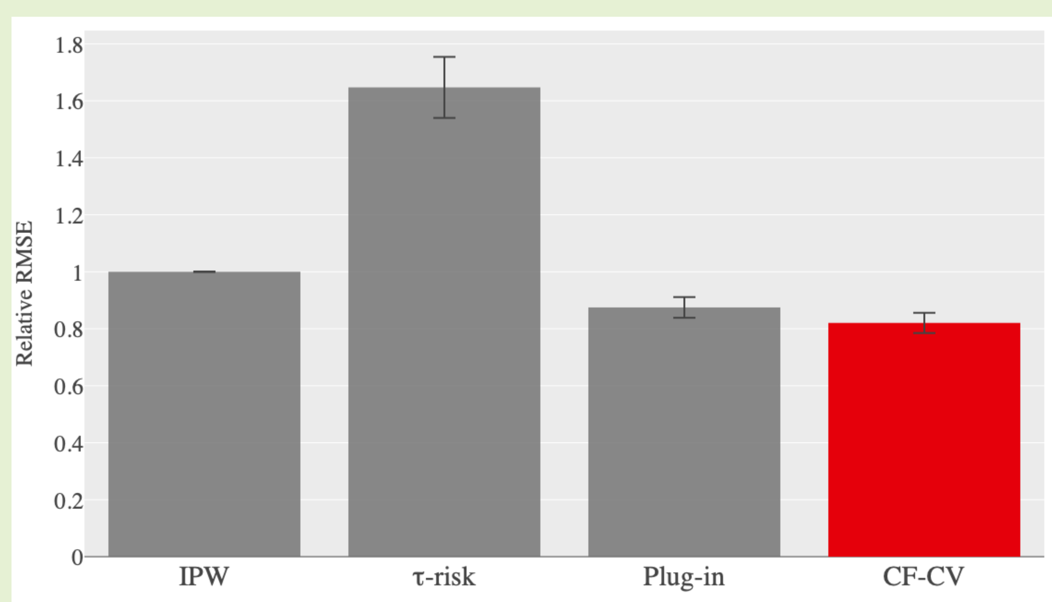
	Rank Correlation		Relative RMSE	
	Avg	Worst-Case	Avg	Worst-Case
IPW	0.224	-0.659	2.027	7.779
τ-risk	-0.399	-0.797	3.408	8.884
Plug-in	0.887	0.385	1.123	1.841
CF-CV (ours)	0.929	0.830	1.040	1.515

• Hyper-parameter Tuning

Procedure:

1. Randomly split the dataset into train/validation/test sets
2. Tune a set of hyper-parameters of a ITE prediction model* using each metric on training and validation sets
3. The true performances of tuned models by each metric are measured using a test set.

Results: **Our metric selects better sets of hyper-parameters!**



* We used a combination of Domain Adaptation Learning implemented in *EconML* and Gradient Boosting Regressor.