Doubly Robust Prediction and Evaluation Methods Improve Uplift Modeling for Observational Data

So-net Media Networks

Yuta Saito, Hayato Sakata, and Kazuhide Nakata

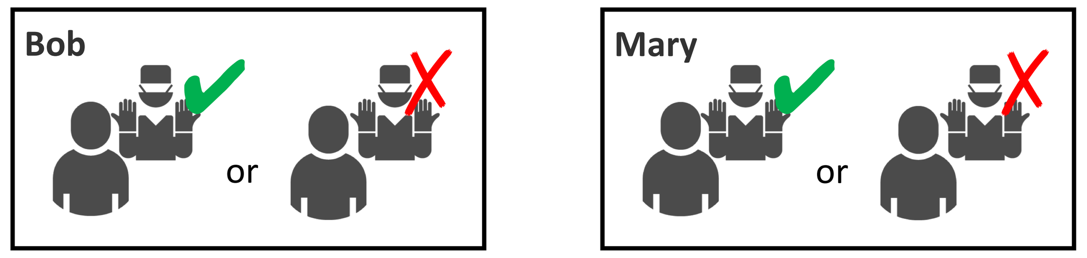
yuta_saito@so-netmedia.jp, hayato_sakata@so-netmedia.jp, nakata.k.ac@m.titech.ac.jp

Tokyo Tech

Motivation

- Achieving optimal treatment assignments
  ex) Medical Treatment, Advertisement, Coupon Distribution

Should We Treat Them ?

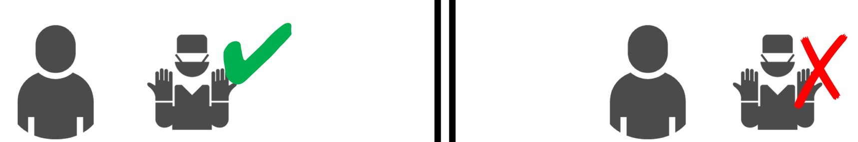


If we know the **optimal treatment of each individual**,
We could achieve **the best possible future** (highest survival rates)

Problem Setting

- Uplift Modeling tries to find an optimal treatment by analyzing the **causal effect** using **Potential Outcomes**

$Y_{Bob}^{(1)}$ Survival Rate **With** Treatment	$Y_{Bob}^{(0)}$ Survival Rate **Without** Treatment



Goal: Predict the **Individual Treatment Effect (ITE)**

$$\tau_{Bob} = Y_{Bob}^{(1)} - Y_{Bob}^{(0)}$$

Causal Effect of Treatment on Bob; Survival Rate With Treatment; Survival Rate Without Treatment

We have 2 options for gathering training and test data

RCT

randomized treatments for data gathering

Pros: Treatments and Features are Independent

Cons: Cost and time ineffective

Observational

historical log data depending on past policies

Pros: Cost and time effective

Cons: Treatment assignments **depend on past policies**

In this work, we focused on **observational data**, which is generally available and we can extend the applications

Related Work

- **Transformed Outcome (*TO*)** as proxy ITE [Athey+ 2015]

$$Y_i^{TO} = \frac{W_i}{e_i} Y_i^{obs} - \frac{1 - W_i}{1 - e_i} Y_i^{obs}$$

  - Y_i^{obs} is the observed outcome
  - $W_i \in \{0, 1\}$ is the treatment assignment indicator
  - $e_i = \mathbb{P}(W_i = 1 \mid X_i)$ is the **true propensity score**
- *TO* is an **unbiased estimator** for the ITE [Athey+ 2015]

$$\mathbb{E}[Y_i^{TO} \mid X_i] = \tau_i$$

Unbiasedness of the *TO* is desirable, but...

- **True Propensity Score** is often missing and *TO* can be biased with an estimated propensity score
- **Variance of *TO*** has yet to be analyzed thus *TO* can be inaccurate proxy ITE

Proposed Techniques

- ***Doubly Robust Estimation***
  Incorporate Potential Outcome Models into *TO*

$$Y_i^{DR} = \frac{W_i}{e_i}(Y_i^{obs} - \hat{\mu}_i^{(1)}) - \frac{1 - W_i}{1 - e_i}(Y_i^{obs} - \hat{\mu}_i^{(0)}) + (\hat{\mu}_i^{(1)} - \hat{\mu}_i^{(0)})$$

$\hat{\mu}_i^{(1)}, \hat{\mu}_i^{(0)}$ are predicted values of $Y_i^{(1)}, Y_i^{(0)}$

We assume the condition below holds in theoretical analyses

$$\underbrace{\left|\Delta_i^{(k)}\right| = \left|\mu_i^{(k)} - \hat{\mu}_i^{(k)}\right|}_{\text{Potential Outcome Estimation Bias}} < \underbrace{\left|\mu_i^{(k)} - 0\right|}_{\text{Bias of Zero Prediction}} = \underbrace{\left|\mu_i^{(k)}\right|}_{\text{Expectation of Potential Outcome}} \quad \forall k \in \{0, 1\}$$

- Bias Analysis

Expectations of Potential Outcomes

$$Bias(Y_i^{TO} \mid X_i) = |\delta_i^{(1)}| \left(\mu_i^{(1)} + \frac{\hat{e}_i}{1 - \hat{e}_i} \mu_i^{(0)}\right)$$

$$\vee \qquad \vee \qquad \vee$$

$$\underbrace{Bias(Y_i^{DR} \mid X_i)}_{\text{Our Method}} = |\delta_i^{(1)}| \left(\Delta_i^{(1)} + \frac{\hat{e}_i}{1 - \hat{e}_i} \Delta_i^{(0)}\right)$$

Potential Outcome Estimation Biases

- Variance Analysis

Expectations of Potential Outcomes

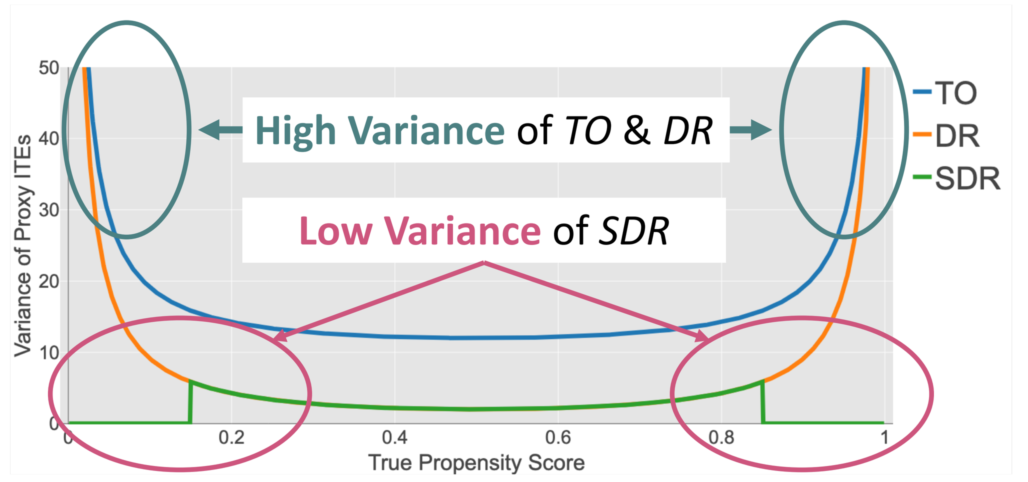
$$\mathbb{V}(Y_i^{TO} \mid X_i) = \frac{1 - e_i}{e_i} \left(\mu_i^{(1)}\right)^2 \left(1 - \delta_i^{(1)}\right)^2 + \frac{e_i}{1 - e_i} \left(\mu_i^{(0)}\right)^2 \left(1 - \delta_i^{(0)}\right)^2 + C$$

$$\vee \qquad \vee \qquad \vee$$

$$\underbrace{\mathbb{V}(Y_i^{DR} \mid X_i)}_{\text{Our Method}} = \frac{1 - e_i}{e_i} \left(\Delta_i^{(1)}\right)^2 \left(1 - \delta_i^{(1)}\right)^2 + \frac{e_i}{1 - e_i} \left(\Delta_i^{(0)}\right)^2 \left(1 - \delta_i^{(0)}\right)^2 + C$$

Potential Outcome Estimation Biases

- ***Switching Technique***
  Substitute extreme propensity scores



Proposed Proxy: Switch Doubly Robust Outcome

$$Y_i^{SDR}(\gamma) = \begin{cases} \underbrace{\hat{\mu}_i^{(1)} - \hat{\mu}_i^{(0)}}_{\text{Low Variance}} & \text{for Extreme Propensity Score } (W_i = 1 \,\&\, e_i < \gamma \text{ or } W_i = 0 \,\&\, 1 - \gamma < e_i) \\ \underbrace{Y_i^{DR}}_{\text{High Variance}} & \text{Otherwise.} \end{cases}$$

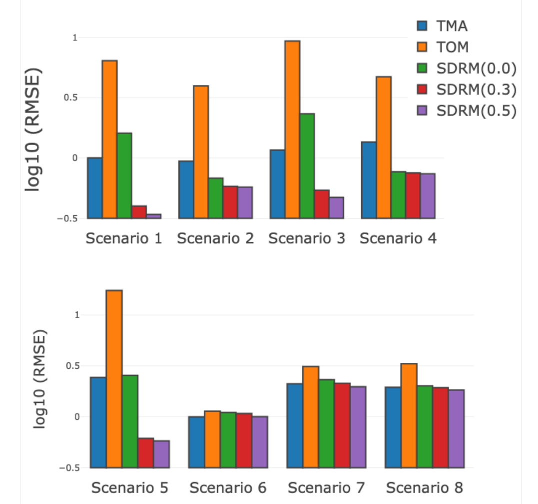
Synthetic Experiment

Setup

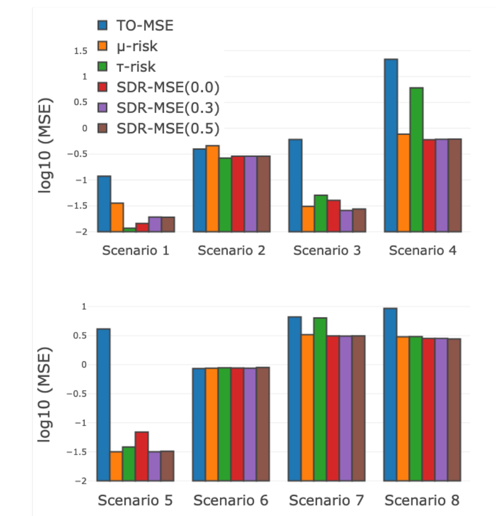
- Used **8 data generating processes** from [Powers+ 2017]
- For prediction methods:
  - Compared TMA, TOM, **SDRM ($\gamma = 0.0, 0.3, 0.5$)** by ITE prediction performance
- For evaluation metrics:
  - Compared μ-risk, τ-risk, TO-MSE, **SDR-MSE ($\gamma = 0.0, 0.3, 0.5$)** by model selection performance

Results

Prediction



Evaluation



- Our prediction method (SDRM) demonstrated the best prediction accuracies in all scenarios
- 0.5 is the optimal value for hyper-parameter γ

- Our evaluation method (SDR-MSE) demonstrated the stable performance across the scenarios
- The effect of varying γ is relatively small but a positive value is better than zero

Real-World Experiment

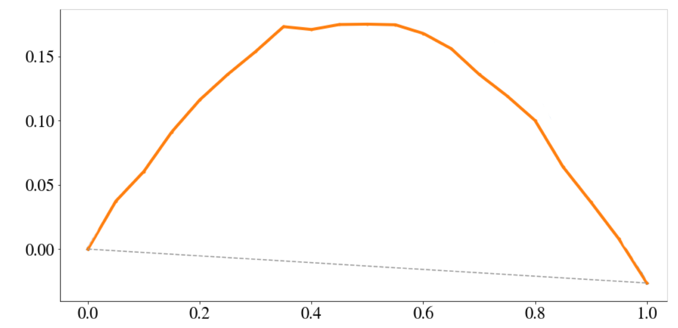
Setup

- **Right Heart Catheterization (RHC)** data
  - well-known public dataset
  - 5,735 critically ill patients
  - **Average Treatment Effect** of RHC was found to be **negative**

Uplift Curve

- a widely used metric in Uplift modeling

y-axis: Difference of Survival rates between the treated and the controlled



x-axis: Ratio of Treated Individuals

Results

- Ours found **20% of positively affected patients**