Feature Selection for Discovering Distributional Treatment Effect Modifiers

Yoichi Chikahara\textsuperscript{1,2}, Makoto Yamada\textsuperscript{2}, Hisashi Kashima\textsuperscript{2} \textsuperscript{1}NTT \textsuperscript{2}Kyoto University

Motivation: Elucidate why treatment effects are different
e.g., vaccination, education program e.g., immunity, grades

Many existing methods use a complex ML model to accurately estimate heterogeneous treatment effects across individuals. However, they offer no answer to the following question:

Different individuals have different treatment effects. Why?

We answer this question by solving the feature selection problem:

\begin{itemize}
  \item \textbf{Input}: Observations of features \(X\), treatment \(A\), and outcome \(Y\)
  \item \textbf{Output}: Features related to treatment effect heterogeneity
\end{itemize}

\[ X_2 \quad X_5 \quad X_9 \]
\[ p = 0.002 \quad p = 0.019 \quad p = 0.035 \]

Our Contributions

1. Novel feature importance measure
2. Its computationally efficient estimator
3. Selection algorithm that controls Type I error

Proposed method

\textbf{Our Goal} Detect features whose values affect the functional of joint distribution \(P(Y_0, Y_1|X_m = x)\) (e.g., treatment effect variance)

1. Detecting distributional treatment effect modifiers

   \textbf{Idea}: If the discrepancy between \(P(Y_0|X_m = x)\) and \(P(Y_1|X_m = x)\) depends on \(X_m = x\), then joint distribution also depends on \(X_m = x\).

   \[ P(Y_0|x^*) \quad P(Y_1|x^*) \quad P(Y_0, Y_1|x^*) \]

   Measured by kernel MMD [2]:

   \[ \mathcal{D}_m^2(x) = \mathbb{E}(x) - 2 \mathbb{E}(x|A = 1) \mathbb{E}(x|A = 0) \]

2. Estimating importance measure with IPW and RFFs

   Using inverse probability weighting (IPW), we reformulate \(\mathcal{D}_m^2(x)\) as

   \[ \text{WCMMD}^2 \]

   \[ \mathbb{E}(A|X) \quad \mathbb{E}(A = 0) \quad \mathbb{E}(A = 1) \]

   Empirical estimator:

   \[ \mathcal{D}_m(x) = \frac{1}{n} \sum \left( \frac{w^*(A, x) w(A, x)^2}{n(x)} + \frac{w(A, x)^2 w^*(A, x)}{n(x)} \right) \]

   If \(X_m\) is discrete, \(w_i = \frac{1}{n_i} \quad \text{if } X_m = x_i\); otherwise, \(w_i = \frac{1}{n} \sum \frac{1}{n(x)} \quad \text{if } X_m = x_i\).

   To reduce the computation time, we approximate \(k_y\) with RFFs [3]:

   \[ k_y(y_i, y_j) \approx k(y_i, y_j) = (z(y_i), z(y_j))^T \]

   Estimated feature importance:

   \[ \mathcal{I}_m = \frac{1}{n} \sum \mathcal{D}_m(x_m) - \frac{1}{n} \sum \mathcal{D}_m(x_m) \]

3. Multiple tests with conditional randomization test (CRT)

   We select features by performing multiple hypothesis tests:

   \[ H_{m,0} : \quad I_m = 0 \quad \text{and} \quad H_{m,1} : \quad I_m > 0 \quad (m = 1, ..., d) \]

   To approximately compute the threshold, we employ the CRT [4]:

   \[ P_{\text{ CRT}}(I_m) \quad \text{Original data} \quad X_m \sim \mathcal{L}(X_m|X_m) \]

Traditional mean-based approaches

Using the CATE conditioned on a single feature (i.e., the average treatment effect across individuals with identical attribute \(X_m = x\)):

\[ T_a(x) = \mathbb{E}(Y_1|X_m = x) - \mathbb{E}(Y_0|X_m = x) \]

The existing methods (e.g., [1]) seek treatment effect modifiers:

\textbf{Definition 1} [Rothman et al. [2008]]. Feature \(X_m\) is said to be a treatment effect modifier if there are at least two values of \(X_m, x_m\) and \(x_m\neq x_m\), such that \(\text{CATE} T_a(x)\) takes different values, i.e., \(\text{CATE}(x_m) \neq \text{CATE}(x_m)\).

Weakness: Mean-based methods may overlook important features

\begin{itemize}
  \item Example:
    \begin{tabular}{c|cccc}
      & \(P(Y_0|X = 0)\) & \(P(Y_1|X = 0)\) & \(P(Y_0|X = 1)\) & \(P(Y_1|X = 1)\) \\
      \hline
      Total & 1 & 1 & 1 & 1 \\
      \(x_1 = 0\) & 0 & 0 & 0 & 0 \\
      \(x_1 = 1\) & 0 & 0 & 0 & 0 \\
      \hline
    \end{tabular}
  \item Individuals with \(X = 0\):
    \begin{itemize}
      \item \(Y_1 - Y_0\):
        \begin{itemize}
          \item \(y_1 - y_0 = 1\)
          \item \(y_1 - y_0 = 0\)
        \end{itemize}
    \end{itemize}
  \item Individuals with \(X = 1\):
    \begin{itemize}
      \item \(y_1 - y_0 = 1\)
      \item \(y_1 - y_0 = 0\)
    \end{itemize}
\end{itemize}

How can we detect distributional heterogeneity?

Experimental results

\textbf{Synthetic data} We compare our method with the two baselines:

1. SI-EM [1]: Mean-based approach
2. Naive: Approximate the null distribution via a naive bootstrap

\textbf{Real-world data} We use health record dataset (from NHANES)

| Feature | Adjusted \(p\)-value |
|---------|---------------------|
| Age     | 0.00075 ± 0.00305   |
| Gender  | 0.0046 ± 0.00269    |
| Number of cigarettes smoked | 0.0 ± 0.0 |

\textbf{SI-EM} cannot detect the features related to treatment effect variance

\begin{itemize}
  \item Proposed achieves high TPR while controlling FPR
\end{itemize}

\textbf{Not detected by SI-EM}

[1] Qinyuan Guo, Dylan S. Small, and Ashkan Eslaminia. “Selective inferences for effect modification via the lasso.” Royal Journal of Statistical Society: Series B (Statistical Methodology), 82(4):382–413, 2022.
[2] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. “A kernel two-sample test.” JMLR, 13(7):273–322, 2012.
[3] Ali Rahimi and Benjamin Recht. “Random features for large-scale kernel machines”. In NeurIPS, volume 3, page 5, 2007.
[4] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Li. “Panning for gold: Model-X knockoffs for high dimensional controlled variable selection”. Journal of Royal Statistical Society: Series B (Statistical Methodology), 80(3):551–577, 2018.