Quantitative Experiments

1 Introduction

The quantitative experiment aims to assess the efficacy of the proposed model in identifying subgroups with significant treatment effects. Based on multiple synthetic datasets and real datasets, we compare the model with various baselines on different metrics.

2 Experiment Setup

2.1 Datasets

We employed synthetic datasets and real-world datasets. Following the settings in [1, 7], we sampled units under the assumption of unconfoundedness, where the covariates are generated from the following distribution:

$$\begin{aligned} &X_1, \cdots, X_i \sim \text{Categorical}(\{A, B, C, D, E\}), \\ &X_{i+1}, \cdots, X_d \sim \text{Normal}(0, 1). \end{aligned} \tag{1}$$

The treatment T is generated according to a Bernoulli distribution, where the probability of $T = 1$ is given by the sigmod function with respect to X. This simulates the non-randomness of treatment assignment in the observational data. Categorical variables are converted to one-hot encoding for calculation. Formally, we have

$$\begin{aligned} f(X) &= \sigma(\langle X, \beta \rangle + \eta), \\ \eta &\sim \text{Uniform}(-1, 1), \\ \beta &\sim \text{Uniform}(0, b)^{|X|}, \\ T &\sim \text{Bernoulli}(f(X)). \end{aligned} \tag{2}$$

The treatment effect TE and the outcome Y is generated by the following formula. An offset is added to Y to ensure that Y is positive. That is,

$$\begin{aligned} TE &= \langle X, \alpha \rangle, \alpha \sim \text{Uniform}(0, 2)^{|X|}, \\ Y &= T \cdot \text{TE} + \langle X, \gamma \rangle + Y_{\text{offset}} + \epsilon, \\ Y_{\text{offset}} &= \max(0, -Y_{\min}), \\ \epsilon &\sim \text{Uniform}(-1, 1), \gamma \sim \text{Uniform}(0, 1)^{|X|}. \end{aligned} \tag{3}$$

We also collected real-world dataset including Twins[1] and IHDP[2]. The detail information of the synthetic data is shown in Table 1.

[1] https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/TWINS

[2] https://search.r-project.org/CRAN/refmans/bartcs/html/ihdp.html

Table 1: Dataset statistics for quantitative experiments.

Dataset	#Units	#Categorical	#Numerical
Syn-1	3000	5	5
Syn-2	3000	5	15
Syn-3	4000	5	25
Syn-4	4000	5	45
Syn-5	4000	5	75
Syn-6	4000	5	95
Twins	23968	3	46
IHDP	747	19	6

2.2 Baselines

We compare the proposed model with two groups of algorithms. The first group is the popular HTE estimation algorithms: (1) Causal Tree (CT) [1]; (2) Causal Forest (CF) [6]; and (3) Causal Rule Ensemble (CRE) [2]. The second group is the rule learning and subgroup discovery algorithms: (1) BRCG [4]; (2) Decision Tree (DT) [3]; (3) Pysubgroup (PYS) [5]. In the first group, CRE can explicitly obtain the antecedent and treatment effect of the subgroup. For CT and CF, it can be considered that the path from the root to the leaf nodes in the tree structure is the antecedent of the causal subgroup. The second group of methods can only get the correlation subgroups. In order to adapt to the causality setting, we add a post-processing step. CATE and variance are calculated on the data covered by each subgroup via eq. (4) and eq. (5).

$$\tau_{\mathcal{S}} = \frac{\sum_{i \in \mathcal{D}_{\mathcal{S}}^{+}} w_i Y_i}{\sum_{i \in \mathcal{D}_{\mathcal{S}}^{+}} w_i} - \frac{\sum_{i \in \mathcal{D}_{\mathcal{S}}^{-}} w_i Y_i}{\sum_{i \in \mathcal{D}_{\mathcal{S}}^{-}} w_i}, \tag{4}$$

where $\mathcal{D}_{\mathcal{S}}$ denotes the covered data, $\mathcal{D}^{+}$ denotes the data that received the treatment ($T = 1$), and $\mathcal{D}^{-}$ denotes the data that did not receive the treatment ($T = 0$), $\mathcal{D}_{\mathcal{S}}^{+} = \{i | i \in \mathcal{D}^{+} \wedge \boldsymbol{\alpha}_{\mathcal{S}}(\mathbf{X}_i) = 1\}$ denotes the units in the treatment group that are covered by the subgroup $\mathcal{S}$, $\mathcal{D}_{\mathcal{S}}^{-} = \{i | i \in \mathcal{D}^{-} \wedge \boldsymbol{\alpha}_{\mathcal{S}}(\mathbf{X}_i) = 1\}$ denotes the units in the control group that are covered by the subgroup $\mathcal{S}$.

$$\begin{aligned} \sigma_{\mathcal{S}}^{2}(0) &= \frac{\sum_{i \in \mathcal{D}_{\mathcal{S}}^{-}} w_i (Y_i - \overline{Y}_w)^2}{\sum_{i \in \mathcal{D}_{\mathcal{S}}^{-}} w_i} \\ \sigma_{\mathcal{S}}^{2}(1) &= \frac{\sum_{i \in \mathcal{D}_{\mathcal{S}}^{+}} w_i (Y_i - \overline{Y}_w)^2}{\sum_{i \in \mathcal{D}_{\mathcal{S}}^{+}} w_i}, \end{aligned} \tag{5}$$

where $\overline{Y}_w$ is the weighted outcome mean.

2.3 Metrics

We evaluate the quality of causal subgroups obtained from different perspectives. First, in order to evaluate the multi-objective optimization of treatment effect and outcome variance, it is proposed that (1) Precision(P) = (the true number of dominating subgroups)/(the number of subgroups in the front discovered by the method). Due to the lack of ground truth for subgroups belonging to the Pareto front, we collected subgroups in the front obtained by all methods and assumed that a subgroup is considered a true dominating subgroup if it is not dominated by any other subgroup. We also considered the interpretability of subgroups, including the metrics (2) #Subgroups(S) = number of subgroups in front, (3) Avg_len(L) = average length of antecedent(i.e.,number of covariates) used to describe the subgroups and (4) Coverage(C%) = The average percent of units in a subgroup to the total number of units.

2.4 Implement detail

We used Bayesian optimization to tune parameters. Specifically, we optimize CT's hyperparameters include cross-validation method cv.option="matching" and pruning factor pru_coef $\in \{0.4, 0.9, 1.5\}$. For CF, its hyperparameter takes the values num.trees $\in \{5, 8, 10\}$, honest version of the CT split.Honest=TRUE and tradeoff between effect and variance split.alpha $\in \{0.2, 0.5, 0.8\}$. The CRE parameters include ntrees $\in \{20, 25\}$, max_depth $\in \{3, 4\}$ and the decay threshold for rules pruning t_decay $\in \{0.025, 0.01, 0.04\}$. For DT, the depth of tree max_depth is fixed as 4. In PYS, the result set has 10 rules, with a maximum rule depth of $\{2, 5\}$, using the subgroup scoring method qf=ps.WRAccQF(). For BRCG, we tune the maximum number of columns generated per iteration K from 8 to 12, and the max rule length is chosen from $\{5, 10\}$.

References

[1] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proc. NAS*, 113(27):7353–7360, 2016.

[2] Falco J. Bargagli-Stoffi, Riccardo Cadei, Kwonsang Lee, and Francesca Dominici. Causal rule ensemble: Interpretable discovery and inference of heterogeneous causal effects. *arXiv preprint arXiv:2009.09036*, 2020.

[3] L. Breiman and Richard A. Olshen. Points of significance: Classification and regression trees. *Nature Methods*, 14:757–758, 2017.

[4] Sanjeeb Dash, Oktay Günlük, and Dennis Wei. Boolean decision rules via column generation. In *Proc. NeurIPS*, 2018.

[5] Florian Lemmerich and Martin Becker. pysubgroup: Easy-to-use subgroup discovery in python. In *Proc. ECML PKDD*, pages 658–662, 2018.

[6] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[7] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. Stable estimation of heterogeneous treatment effects. In *Proc. ICML*, 2023.