An Offline Evaluation Method for Individual Treatment Rules and How to Find Heterogeneous Treatment Effect

Thanaporn Patikorn, Neil T. Heffernan, Jian Zou
100 Institute Rd.
Worcester, MA 01609
{tpatikorn, nth, jzou} @wpi.edu

ABSTRACT

Heterogeneous treatment effects occur when the treatment affects different subgroups of population differently. In this work, we conducted a large scale simulation study to identify the characteristics of treatments that are more likely to have heterogeneous treatment effects, and to estimate how effective the individual treatment rules are compared to the better conditions. We found that heterogeneous treatment effects are rare. When the overall treatment effect is close to zero, we found that individual treatment rule is very likely to be effective. With large positive or negative overall treatment effect, the heterogeneous treatment effect is less likely to occur, and the individual treatment rules are more likely to be ineffective.

Keywords

Heterogeneous Treatment Effect; Individual Treatment Rule; ASSISTments; Randomized Controlled Experiment.

1. INTRODUCTION

Researchers have been using randomized controlled experiments (RCT) to test their interventions. RCTs are considered the gold standard and are widely used in many fields, from healthcare to education. Traditionally, researchers often look for treatment effects across the population. However, in many experiments, the treatment effect differs systematically from one subgroup of the population to another. For example, patients who are allergic to the treatment drugs may react negatively instead of benefiting from the drug. This type of effect is often called heterogeneous treatment effects, as there are different effects for different types of people. Many machine learning methods have been developed to detect heterogeneous treatment effects. For example, [4] introduced the Causal Forest, a decision tree-based method to determine the treatment effect on each subgroup of the population.

In many cases such as [1], it is better to tutor students with lower prior knowledge using step-by-step hints, while it is better to tutor students with high prior knowledge with full problem solutions. In this case, giving personalized tutoring to each student is better than giving the same tutoring to everyone. This type of condition assignment is often called an individual treatment rule or a personalization policy.

In order to evaluate a personalization policy, the most popular method is to deploy the policy in real time and compare the result. However, the on-line method is often costly and sometimes unavailable to the researchers (e.g. because the data have already been collected). As a result, many researchers conduct an offline policy evaluation using past data. In [3], they use the expected outcome of the policy to evaluate their personalization policy. To calculate the expected outcome using past RCT data, we must first find a subset of subjects whose random condition assignments during the RCT matches the personalized condition assignments of the policy. The expected outcome of a personalization policy is the average outcome of this subset across conditions. Comparing two policies using the expected outcome easy and intuitive; if the larger outcome values are better, the policy with larger expected outcome is better. This method is equivalent to policy risk introduced in [2].

The main goals of this work are 1) to find the characteristics of the experiments that are more likely to have heterogeneous treatment effects, and 2) to compare a personalization method, specifically Causal Forest, against assigning every subject to the best conditions to find out how effective a personalization policy can be.

2. METHODOLOGY

In order to gain a better understanding of expected outcome, we investigated how it is calculated in [3]. They first took the subset of the subjects from the RCT whose random condition assignments are the same as the condition assignments given by a personalization policy. For the rest of this paper, we will refer to this subset as the "congruent subset". Then, the expected outcome of the policy is calculated by taking the average outcome values of the congruent subset regardless of conditions. For example, in Table 1, the congruent subset consists of subject 1, 3, 4, and 5, and the expected outcome of the policy is (0.7 + 0.4 + 0.6 +0.7)/4 = 0.6.

2.1 Simulation Study

We conducted a large-scale simulation study to verify the effectiveness of using the congruent subset as an estimate of real outcome values of the policy, and to find types of experiments that are likely to have personalization. We chose simulation study because it allows us to not only calculate the real outcome values of the policy, but also investigate how different settings impact the personalization.

Table 1: an example data to show how congruent subset works

subject	RCT condition	outcome	personalized condition	Is in congruent subset?
1	C	0.7	C	yes
2	T	0.6	C	no
3	C	0.4	C	yes
4	T	0.6	T	yes
5	T	0.7	T	yes
6	C	0.5	T	no

Table 2: Different Distributions for Effect of Conditions

distribution	parameter	values	number of combinations
normal	mean	0, 1, 2, 5, 10	15
	sd	1, 2, 5	
log normal	meanlog	0, 0.5, 1, 2	16
	sdlog	0.25, 0.5, 1, 2	
gamma	shape	0.5, 1, 2, 5, 10	15
	scale	0.5, 1, 2	
total			46

For the simulation study, we focused only on experiments with two conditions. For each condition, we simulated 46 different settings, as shown in Table 2, resulting in 46 * 46 = 2116 different combinations of experiments. We also include lognormal distributions and gamma distributions because real datasets may not always follow normal distributions, for example the mastery speed in [5] resembles lognormal distribution. For each setting, we generated 1000 datasets, each of which has 1000 data points.

Every data set has 3 covariates: one with a positive, negative, and no effect on the outcome. Every covariate value is generated independently for each subject from a normal distribution with mean = 0 and sd = 1. The true effect is generated using the distribution and parameters in Table 2. The observed outcome is

$$\text{observed} = \text{effect} + \text{cov1} * \text{impact1} - \text{cov2} * \text{impact2} + \text{noise}$$

The impacts are from uniform (0,5) and remains constant within experiment. The noise is drawn from a normal (0,1) distribution.

For each personalization policy, we measured 1) if the outcome values of congruent sets are significantly different from the outcome values of actually assigning everyone using personalization policy, and 2) whether the personalization from the Causal Forest is better than the better of the two conditions.

3. RESULTS

From 2,116,000 simulated dataset, we detected the significant difference between the outcome values of the congruent sets and the real personalized outcome values less than 1% of the time, which is far lower than the threshold of 5%, regardless of parameters of the dataset. As for the effectiveness of the Causal Forest, we look at how often the personalization suggested by Causal Forest are better than assigning subjects to the better of the two conditions. We found that personalization is slightly more common when at least one of the distribution is gamma distribution.

Table 3: the Effectiveness of Personalization Suggested by Causal Forest by Overall Observed Treatment Effect

Rounded average observed treatment effect	Causal Forest suggests personalization	Causal Forest's personalization is the most effective
≤ -5	0.03%	15.26%
-4	0.04%	23.19%
-3	0.12%	22.46%
-2	0.41%	44.44%
-1	2.98%	76.43%
0	8.27%	83.56%
1	3.03%	76.26%
2	0.43%	44.79%
3	0.12%	20.76%
4	0.05%	23.35%
≥ 5	0.03%	14.67%

Table 3 shows that when the treatment effect is close to zero, the personalization suggested by the Causal Forest is very effective. Causal Forest policy is better than assigning subjects to the better of the two conditions more than 3/4 of the times when the treatment effects are between -1 and 1. The effectiveness of the personalization quickly drops as the treatment effect is far from zero. It is important to note that the Causal Forest we used in this study has never been optimized and most of parameters we used are default, except the two we specified earlier in the paper.

4. CONCLUSION

This paper has three main contributions. First, we promoted the study of heterogeneous effects and an offline personalization policy evaluation method to the Educational Data Mining. Second, we investigated several different settings of simulated experiments to find the characteristics of the experiments that are more likely to have heterogeneous treatment effects. We found that, generally heterogeneous treatment effects are not common and typically rare when the treatment effects are very large or very small. Third, we investigated the effectiveness of personalization policies given by Causal Forest. We found that the personalization policy is likely to be effective for the experiments with small treatment effects.

5. FUTURE WORK

We plan to investigate different methods for detecting heterogeneous treatment effects on real dataset from ASSISTments to see if we can detect more experiments like [1]. If we can detect such effects, we would be able to improve our system even further, which will improve student learning.

We also plan to compare different methods for detecting heterogeneous treatment effects to see what are the advantages and disadvantages of each model. We also plan to compare these pre-train models to real-time methods like bandits as well. This result will allow us to be able to choose the right tool for the right personalization task.

6. ACKNOWLEDGMENTS

We thank multiple NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736, DGE-1535428 & DRL-1031398), the US Dept. of Ed (IES R305A120125 & R305C100024 and GAANN), and the ONR.

7. REFERENCES

[1] Razzaq, L. M., & Heffernan, N. T. (2009, July). To Tutor or Not to Tutor: That is the Question. In *AIED* (pp. 457-464).

[2] Shalit, U., Johansson, F., & Sontag, D. (2016). Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint arXiv:1606.03976*.

[3] Vickers, A. J., Kattan, M. W., & Sargent, D. J. (2007). Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*, *8*(1), 14.

[4] Wager, S., & Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*.

[5] Xiong, X., Li, S., & Beck, J. E. (2013, May). Will You Get It Right Next Week: Predict Delayed Performance in Enhanced ITS Mastery Cycle. In FLAIRS Conference