Observing Personalizations in Learning: Identifying Heterogeneous Treatment Effects Using Causal Trees

Biao Yin
Worcester Polytechnic Institute
Worcester, MA
byin@wpi.edu

Thanaporn Patikorn
Worcester Polytechnic Institute
Worcester, MA
tpatikorn@wpi.edu

Anthony F. Botelho
Worcester Polytechnic Institute
Worcester, MA
abotelho@wpi.edu

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, MA
nth@wpi.edu

ABSTRACT
The incorporation of computer-based platforms in the classroom has introduced the ability to conduct numerous randomized control trials at scale with student-level randomization. Such systems are able to collect vast amounts of data on each student while completing work in the classroom and at home. It is often the case, however, that the effects of these trials are reported across all students, ignoring the potential for personalized learning. Personalized learning, or the observation of heterogeneous treatment effects, considers that the effects of a studied learning intervention may differ for individual students; while an intervention may work well for low-performing students, for example, it may have no effect for higher performing students. Personalized learning can lead to better instructional practices that maximizes the learning benefits for each individual student, and with the use of computer-based platforms, such individualized instruction is made feasible at large scales. In this work we use a causal decision tree to observe treatment effects in 9 experiments run in the ASSISTments online learning platform.

Author Keywords
Personalization; Heterogeneous Treatment Effects; Randomized Controlled Trials; Causal Tree

L@S 2017, April 20-21, 2017, Cambridge, MA, USA

DOI: http://dx.doi.org/10.1145/3051457.3054009

INTRODUCTION
Many teachers employ computer-based learning platforms to assign content and track students' learning progress. While these systems provide many useful tools to teachers, their usage, or more specifically that of their students, provides the ideal setting to run randomized control trials aimed at identifying what interventions best help students learn. While the results of these trials, often described in terms of the effectiveness of treatment over a control condition, are useful to describe the effectiveness of interventions across all students, it is often the case that there is variance within each condition; the treatment may work very well for an identified subgroup of students and exhibit no effects on another subgroup. It is also possible that no significant effect of treatment is found in whole, but such an effect may exist specifically for a smaller subgroup. Identifying what works for whom is the goal of personalized learning, such that, knowing information about a student, one can apply the best instructional practices that maximize student learning.

This concept of personalization is often referred to in terms of heterogeneous treatment effects. Essentially this is the idea that treatment effects are not the same across all students, and can be more accurately measured by identifying the subgroups where the effect is present and testing for significance. This identification of subgroups could be formed in a variety of ways. While many clustering algorithms attempt to find groups of similar students, the method that has been applied to this problem in the past has instead utilized a form of decision tree, splitting students to maximize differences between the subgroups on some observed outcome metric. One such tree-based method, referred to as Causal Trees [1], has been introduced specifically to identify these heterogeneities, and employs a methodology to split students as well as measure differences in treatment effects.

While not fully explored in the field of education and learning analytics, the measurement of heterogeneous treatment effects has long existed in other fields. Commercial companies such

as Microsoft [3][4], Ebay [2], and many others who often choose not to publish run large quantities of trials at scale on their respective websites and platforms. Often is the goal of these trials is to identify variations in design, layout, and incorporation of advertisement that optimizes an outcome metric of interest; more importantly, it is often the goal to identify who responds best to the tested variation in order to maximize indirectly, if not explicitly, revenue. Such companies understand that effects of treatment can exhibit, if perhaps rarely, significance in the presence of heterogeneous subgroups, so why has this not been widely implemented in the field of education?

The answer to this question in the past has been in regard to scalability. It is difficult in traditional classroom environments to implement individualized instruction practices. As the teacher usually addresses multiple students, using instructional practices that benefit the greatest number of students is more practical simply due to the infeasibility of one-to-one instruction. This changes, however, with the incorporation of computer-based instructional practices in the classroom and for homework. In this case, individualized aid can be given to students in addition to the in-class instruction. We expect that the combination of these two strategies will increase student learning by collecting and utilizing information about student performance.

Online learning platforms and intelligent tutoring systems already collect a breadth of student features as problems are attempted, and such information has been used in a range of predictive and performance measuring tasks. Similarly, such information can be used to better analyse the results of randomized control trials within these systems. Students who have experienced the experiment can be grouped based on features collected before condition assignment, after which treatment effects can be measured within each grouping. Identifying a significant effect within a subgroup can be then implemented for personalized instruction for future students. Future students beginning an assignment could be assessed using the same set of prior performance metrics to identify if one method of instruction is more beneficial than another, and then be given the better strategy.

The goal of this paper is to provide a preliminary analysis to demonstrate that heterogenous treatment effects exist in the field of learning. Beyond this, we show how a Causal Tree (CT) can be implemented to observe this effect.

METHODOLOGY

In order to explore the concept of heterogeneous treatment effects, we observe two approaches. The first explores the use of just one student feature to observe a qualitative interaction exemplifying this effect, while the second applies the more sophisticated method of Causal Trees [1]. This work utilizes a unique dataset consisting of student information from 22 randomized controlled trials run in the ASSISTments online learning platform.

Dataset

The ASSISTments online learning platform is a free web-based platform utilized by a large user-base of teachers and students. The system, based primarily in math content, allows teachers to assign several types of assignments for classwork and homework, reporting on student performance and learning progress. Students are given immediate feedback on each problem, and are also presented with several forms of instructional aid including hints, that provide a useful message, and scaffolded questions that break the problem into smaller steps. The platform has been the subject of a recent study within the state of Maine [5], demonstrating significant learning gains for students using the platform.

The dataset used [7] in this work is unique in that it provides student information collected within the platform, comprising 22 randomized controlled experiments. These experiments were run in assignment types known as "skill builders" in which students are given problems until a threshold of understanding is reached; within ASSISTments, this threshold is traditionally 3 consecutive correct responses. Reaching this threshold denotes sufficient performance and completion of the assignment. In addition to this experimental data, information of the students prior to condition assignment is also provided in the form of problem-level log data providing a breadth of student information at fine levels of granularity.

Initial Approach

A preliminary analysis is performed to introduce a simple method of identifying heterogeneity within one of the experiments in our dataset. This selected experiment compared a treatment group that presented students with video-based hints to a control group that used text-based hints. The experiment was run within an assignment entitled the "Composition of Functions." This analysis observes differences in completion rates of the experimental assignment as an outcome measure.

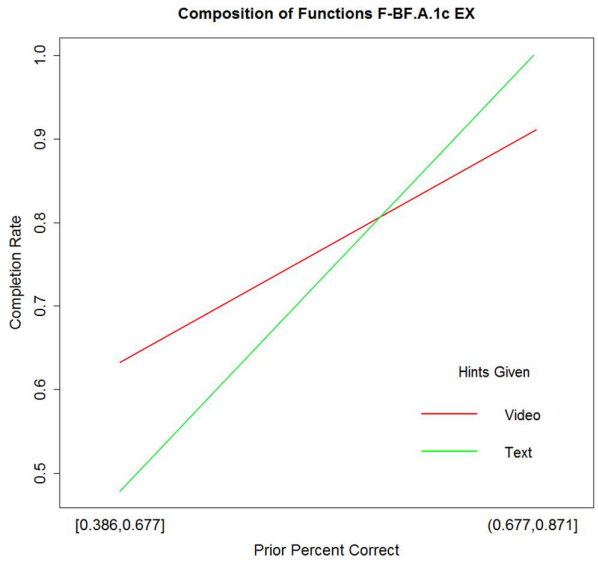

Figure 1. Qualitative interaction found using a median split of prior percent correctness.

After filtering observations in the dataset with missing values, two simple methods are applied to illustrate the heterogeneous effect. Each of these observes the student feature of prior percent correct, representing the average number of correct

responses in all data prior to the experiment for each student. The data is filtered to remove missing values, leaving 110 observations. The first method, depicted in Figure 1 uses a median split of this metric to place students into one of two bins after which the completion rate is averaged within each bin. The second method, illustrated in Figure 2 uses a logistic regression including a term for condition, the prior percent correct covariate, and an interaction term of these two.

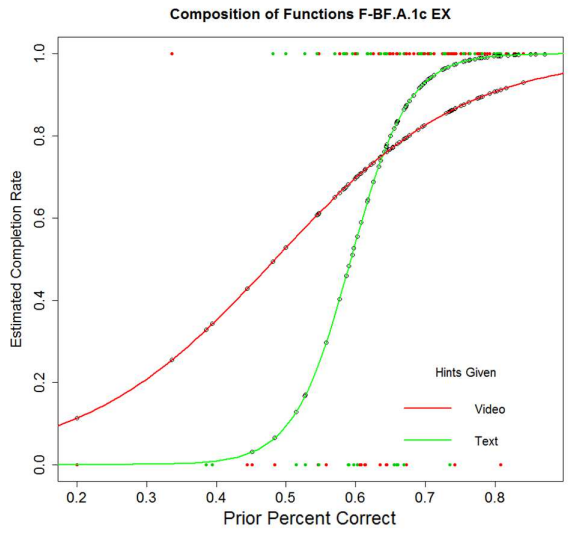

Figure 2. Qualitative interaction found using a logistic regression.

Each of these show an interaction indicative of heterogeneity. As is depicted in the second method, using the continuous representation of prior percent correct, the effect is not equally large for all types of students. This suggests that a more sophisticated method may be able to identify other effects using multiple features rather than just one to group students.

Causal Trees

Expanding upon the preliminary work depicted in the previous section, we also apply Causal Trees [1] to our dataset to identify and measure other such treatment effects. The technique is applied to 9 experiments of the existing 22 due to their similarity of intervention design. Each compares a variation of providing video versus text hints in different contexts. This subset is used for this comparison and as this is presented as a demonstration of the method's utility. Due to the small treatment effects found in the training data, it is expected that heterogeneity in these particular experiments within our dataset is rare.

First, we randomly split the data into two equal subsets; the first is to be used for model training, while the second will be used as a test set. Using the training data, we build a single causal tree across all experiments; as each experiment observes a similar intervention, the results will indicate more generalized effects of video- versus test-based hints. As input into the tree, we observe multiple covariates such as locality of the school, Guessed Gender (as data is anonymized, gender is inferred or left as unknown), Z-Scored Mastery Speed (a performance metric denoting the number of problems needed to complete a skill builder), Prior Percent Completion (average number of completed assignments prior to the experiment), Prior Percent Correct, Prior Homework Percent Completion, Z-Scored HW Mastery Speed, and also the experiment identifier. Using the test set, students are grouped using into leafs using the constructed tree. Within each leaf, students are split by condition and compared to identify treatment effects.

RESULTS

Observing the analysis using the causal tree[1], the method of measuring differences between condition at each leaf can vary. In our analysis, we perform a Fisher test for odds ratios[6] with 95% confidence intervals. This odds ratio analysis provides a relative measure of effect of treatment as compared to the control condition. Again, completion of the experiment is our outcome metric of comparison. To interpret the ratio, a value less than 1 indicates higher effects in the experimental condition, while values greater than one indicate a higher effect in the control condition; values equal to one indicate that there is no effect. This analysis provides significance values and confidence intervals to evaluate the believability of the effect.

Breslow-Day X^2	22.71
p-value of Common Odds Ratio	0.01

Table 1. Breslow-Day test for heterogeneity between nodes in regard to odds ratio.

The causal tree odds ratio analysis is reported in Table 2. These results, generated from the test set applied to the constructed model from training, show two leafs with significant treatment effects. These leafs represent interactions of prior homework completion and prior percent correctness values of students as discovered by the tree. The number of students found to complete, as opposed to those who did not complete, the experimental assignment are also reported in Table 2.

The method is shown to identify heterogeneous effects, but one further step can be taken to ensure that the technique is splitting students such that there is heterogeneity between the leafs themselves. Using a Breslow-Day (B-D) test, we can validate that the causal tree is finding heterogeneous groups of students. The results of this test, reported in Table 1, indicate that there is significant heterogeneity between the leaf nodes.

DISCUSSION

The results of the causal tree analysis suggest that some heterogeneous effects do exist across all experiments. Based on this general analysis, however, it would appear that no qualitative interactions emerge from the results. It is found in the first two leafs that the text-hints condition is significantly better than the video-hint condition, while no other significant effects are found.

As this is merely an exemplary usage of such a model to identify areas for personalization, several limitations of our approach may be affecting our results. For example, a single

[1]The resulting tree can be viewed at http://tiny.cc/LaS_causal_tree

Leaf	Control N Complete	Control N Incomplete	Treatment N Complete	Treatment N Incomplete	Odds Ratio	Lower CI	Upper CI
1	33	18	23	3	2.45*	1.05	5.87
2	19	5	17	21	4.58*	1.30	19.06
3	54	4	65	1	2.07	0.56	9.54
4	51	17	48	8	0.50	0.17	1.37
5	42	15	39	5	0.36	0.09	1.18
6	100	8	78	10	1.60	0.54	4.90
7	28	6	32	7	1.02	0.26	4.15
8	24	1	17	3	4.11	0.30	231.08
9	34	18	33	13	0.75	0.29	1.91
10	25	8	28	11	1.22	0.12	4.12
11	35	27	48	27	0.73	0.35	1.54

Table 2. The number of students for each leaf in the causal tree and odds ratio analysis metrics. *Significance based on the confidence intervals.

tree is used to describe all 9 related experiments for sake of reducing the complexity of our presentation of the approach, but it may be more beneficial to train a separate tree on each experiment. Likewise, in place of using a single tree per experiment, it may also be beneficial to explore the usage of random causal forests [1] to explore a wider range of covariate combinations.

CONTRIBUTION

This paper stresses the need to apply focus toward improving personalizations in learning. As explored in both a simple analysis, and a more sophisticated approach using a causal tree, heterogeneous effects are observable in our dataset. This work is, to our knowledge, among the first to apply causal trees in an education context and thus acts as a pilot study to observe the efficacy of such an application.

In addition to the exploration of these effects, we introduce the application of a very simple statistical test to validate the splitting criterion of the causal tree. The B-D test validates that there is heterogeneity between leafs in terms of the effects of treatment as measured by the odds ratio.

In light of these findings, it becomes apparent that current instructional approaches could greatly benefit from the incorporation of such information. This is particularly the case in computer-based systems that often collect breadths of student information, the potential of which is perhaps not yet fully realized.

ACKNOWLEDGMENTS

We thank multiple current NSF grants (IIS-1636782, 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.

REFERENCES

1. Susan Athey and Guido Imbens. 2015. Recursive partitioning for heterogeneous causal effects. *arXiv preprint arXiv:1504.01132* (2015).
2. Thomas Blake, Chris Nosko, and Steven Tadelis. 2015. Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. *Econometrica* 83, 1 (2015), 155–174.
3. Alex Deng, Pengchuan Zhang, Shouyuan Chen, Dong Woo Kim, and Jiannan Lu. 2016. Concise Summarization of Heterogeneous Treatment Effect Using Total Variation Regularized Regression. *arXiv preprint arXiv:1610.03917* (2016).
4. Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1168–1176.
5. Jeremy Roschelle, Mingyu Feng, Robert F. Murphy, and Craig A. Mason. 2016. Online Mathematics Homework Increases Student Achievement. *AERA Open* 2, 4 (2016). DOI:http://dx.doi.org/10.1177/2332858416673968
6. Nova Scotia. 2010. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry* 19 (2010), 227.
7. Douglas Selent, Thanaporn Patikorn, and Neil Heffernan. 2016. ASSISTments Dataset from Multiple Randomized Controlled Experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 181–184.