ASSISTments Dataset for a Data Mining Competition to Improve Personalized Learning

Thanaporn Patikorn, Douglas Selent, Neil Heffernan, Biao Yin, Anthony Botelho
{tpatikorn, dselent, nth, byin, abotelho}@wpi.edu
Worcester Polytechnic Institute
100 Institute Road, Worcester, MA, 01609

Abstract

Online learning platforms often act as a medium to run and apply large-scale experimentation to improve student learning due to the large number of users and ability to record information at granularities ranging from the smallest action to performance over extended timeframes. ASSISTments is one such platform that expands on this opportunity, providing the means for external researchers to propose and run randomized controlled experiments with the many students utilizing the learning tool for both homework and classwork. The resulting data collected on such students participating in the experiments, combined with the large sums of data collected on assignments prior to beginning the experiment, culminates in a unique opportunity to progress personalization in learning. While experimentation often observes the utility of different learning interventions and methodologies, it is often the case that something that works well for one group of students may not work as well for another group, introducing our usage of the term personalization; identifying what intervention, described by experimental condition, is best for each student can lead to better instructional practices to benefit student learning.

This paper proposes an open data mining competition using the student data collected from ASSISTments, spanning 22 experiments and the large number of assignments students worked on prior to each experiment. The goal of this competition is to use the student information prior to each experiment to predict the condition that would be most beneficial to the student. The information gained from such predictive models can be used to personalize student learning in the future. This paper further discusses the details of this proposed competition, including information regarding the dataset and validation methods. This work also provides a preliminary analysis, using several student covariates, to predict the probability that a student will complete the assignment for each condition of the experiments and assign the student to a condition with higher probability of completion.

Introduction

Many intelligent tutoring systems attempt to use machine learning techniques to improve student learning. While these systems generate large quantities of feature-rich data pertaining to student classwork and homework each day, only a small portion of the data is available to the educational data mining community. ASSISTments researchers realize that such data should be made available to any researchers who are interested in utilizing such information, because any findings can be used to improve the tutoring systems and then benefit the student. We believe this is the ultimate goal of intelligent tutoring systems. ASSISTments has worked towards this goal and provides a gateway, known as ASSISTments TestBed, that provides the means for external researchers to propose experiments on ASSISTments users, as well as a means of

1 The details of ASSISTments TestBed can be found here www.assistmentstestbed.org
collecting, organizing, and retrieving collected data for analysis. While many prefer to use simple statistical methods such as ANOVA and regressions, we are also interested in what information can be obtained using modern statistical and machine learning methods.

While several new interventions and experimentations that aim to increase student learning are published every year, very few of those publications make their data publically available for the educational data mining community. The main objective of this paper is to excite the educational data mining community by proposing data mining competition where both the dataset and competition goals are unique.

In this paper, we provide a large dataset for action-level, problem-level, and student-level data of all students who participated in selected 22 randomized controlled A/B experiments run in the ASSISTments online learning platform. We intend to make this data publically available to any researchers so they can use statistical methods and data mining techniques to broaden what we know about student learning.

Bloom has shown that a smaller student-teacher ratio can have a large effect on student learning (Bloom, 1984). Applications of this research are now manifest in many MOOCs, where there are a large number of learners and personalization through online tutoring systems is desired to mimic a small class size. Intuitively, we know that interventions may impact various students differently. For instance, Razzaq & Heffernan found that students with high prior knowledge benefit more from full solutions, while students with low prior knowledge benefit more from tutored problem solving (Razzaq, L. & Heffernan, N., 2009). Several other works such that (Chaplot et al, 2016), (Reddy et al, 2016), and (Lee et al, 2016) also uses time-step data to better personalize learning, in either problem or problem set levels.

In this paper, we combine our prior work on running randomized controlled experiments and making the data publically available with the prior literature on the positive effects on personalized learning. We leverage several of our previously run A/B experiments in order to generate a large dataset to use in our proposed data mining competition for improving the benefits of personalized learning. The ultimate goal of this paper is to inspire the educational data mining community to devote much-needed research toward personalizing learning.

**Dataset**

In the ASSISTments online learning platform, a “problem set” is a hierarchical ordering of problems and problem sections that can be structured in a variety of different ways. A student assignment is an instantiation of a given problem set. All of the problem sets included in the dataset of 22 randomized controlled experiments are considered “Skill Builders,” where students will continue to receive similar problems until they answer a specific number of problems (usually three) correctly in a row. Although all of the experiment data comes from students who worked on Skill Builder assignments, the prior data is not exclusive to a specific type of assignment. This dataset provides more detailed information for the experiment data presented in (Selent, D., Patikorn, T. & Heffernan, N. T., 2016).

The dataset for the competition contains logs from a total of 14,283 unique students who attempted 3,095,530 problems from 22 A/B experiments run in ASSISTments. In addition to the data collected during the experiments, we also include all action- and problem-level information of the students before participating in any of the A/B experiments. We believe that the action-level, representing the finest granularity of data recorded, and problem-level information, representing logged data on each problem attempted, will allow researchers to learn more about students in a way that will ultimately lead to personalized learning. For example, Baker et al.
created a detector that allows researchers to measure students’ affect as either concentrating, bored, frustrated, or confused as they progress through the problems (Baker, R.S., et. al 2010).

We split our data into three sets, training, validation, and test; we further refer to the combination of the validation and test set as the holdout set. For each student, a number K is drawn from a binomial distribution with \( p = 0.5 \) and \( N = \) the number of experiments (out of 22) that the student participated in. The first K participations, in chronological order (of the 22 experiments) of the student, will be put into the training set while the rest will be put into the holdout set. This is done to ensure that, 50% of the data is in each set, and that all of the data in the holdout set is chronologically after the data in the training set to prevent the scenario of predicting past data from future data. Then, each student’s participation is randomly assigned to the validation set and the test set, regardless of students and problem sets.

Participants in the competition will receive all student data in the training set, including labels indicating experimental condition and student data prior to each experiment. The validation set will be used to calculate a score for the public leaderboard to provide an immediate indication as to how well a submitted model performs on the unseen validation set. The test set will be used to calculate the final competition score in order to determine a winner. To clarify, the holdout set contains only student covariates, e.g. student ID and prior percent correct, and it is the goal of the participants to, from provided prior student information, assign each student to a condition to maximize the likelihood of that student completing the experiments. The information about which data point in the holdout set belongs to either the validation set or the test set will be hidden from the participants.

Data Format

This paper provides the dataset in two formats. The first format is a large CSV (comma separated values) file containing all information from student level features such as inferred gender, date of birth, teacher id, school id, as well as problem features such as the text of the problem, prerequisite and post-requisite skills associated with the problem, and finally demographic information on the school and district. There are over one hundred features provided in this file. All the data is anonymized so that specific students or schools cannot be identified. The data is in the form of one-row-per-problem-per student, meaning that each row represents a problem attempt for a given student. This file covers all problems that students have attempted both before and during the A/B experiments.

The second format consists of a set of CSV files for each experiment containing student data for the given experiment. These data files originated from the Assessment of Learning Infrastructure (ALI), which is a system that provides data from ASSISTments to researchers in several formats (Ostrow et al, 2016). A set of files for an experiment is made up of the following five files: Covariate, action level, problem level, student level, and student level with problem level\(^2\).

Note that only the covariate file will be given out for students who are in the holdout set. Also note that some information is duplicated across the set of files. The same information may be represented in different formats to make it easier to use for different type of analyses.

\(^2\) Details on each of the files are described here: \url{http://www.assistmentstestbed.org/the-data/ali-s-raw-data}
Data Mining Competition

We propose a data mining competition on a dataset collected using ASSISTments comprised of student performance prior to and including experimentation run in the platform. The goal of this competition is to find better ways to personalize learning by using information of each individual student to determine which experimental condition would be most beneficial\(^3\). For this year’s competition, we want to know how to best assign students to conditions for the greatest learning benefit.

We use the predicted completion rate as the model performance measure. We define the predicted completion rate as the completion rate of a subset of students whose suggested condition assignment is the same as actual condition assignment. The completion rate is the number representing the percent of the students who complete the assignment. This method allows us to simulate real-time condition assignments with past data. Since students were actually assigned to conditions at random when the student started the problem set, one could see this method as averaging the completion count of roughly half the data.

<table>
<thead>
<tr>
<th>student id</th>
<th>prior percent correct</th>
<th>Completion (hidden)</th>
<th>actual condition (hidden)</th>
<th>suggested condition 1 (submitted)</th>
<th>suggested condition 2 (submitted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>1</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>0</td>
<td>E</td>
<td>C</td>
<td>E</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>0</td>
<td>C</td>
<td>E</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
<td>0</td>
<td>C</td>
<td>C</td>
<td>E</td>
</tr>
</tbody>
</table>

Table 1: A mock-up data of the validation set (column 1 and 2) and other information used to compute the predicted completion rate of a given set of suggested condition assignment.

In Table 1, the predicted completion rate of ‘suggested condition assignment 1’ is computed with 2 students (1, 4), resulting in predicted completion rate of 0.50. On the other hand, the predicted completion rate of ‘suggested condition assignment 2’ is computed with 3 students (1, 2, 3), resulting in predicted completion rate of 0.33. This example also serves as a reminder that the goal is not to predict the actual condition, but to predict which condition should be assigned to a student so that he/she is more likely to complete the problem set.

Preliminary Result

Several preliminary analyses have been conducted to illustrate the feasibility of personalization within the dataset. Similarly aiming to maximize completion rates, a random assignment model was created to be used as a baseline. With random assignment, the average completion rates across all experiments on the validation set is 0.81 with a standard deviation of 0.39. This can be interpreted as 81% of the students assigned to conditions at random will complete the experimental assignment. This result agrees with the average completion rate of the entire validation set of 0.809.

Our first comparison to this baseline attempts to apply the result of (Razzaq, L. & Heffernan, N. 2009). We discretized students’ prior percent correct into two categorical values

\(^3\) The dates and details of the competition will be announced later in 2016 at the following website: https://sites.google.com/site/assistmentskdd2016/
using the within-problem set average as a cut-off. For each of the resulting categories of students, we computed the average completion rate of each condition using the training set. For each student in the validation set, we chose the condition with higher average completion rate, given the prior percent correct of that student. With this method, the average completion rates across all experiments on the validation set is 0.84 with a standard deviation of 0.37. This implies that 84% of the students assigned to conditions by this method will complete the problem sets, which is 3% higher than the baseline model. The result of an unpaired t-test of unequal variances (Welch’s t-test) is shown as: Model 1: random (Mean = 0.81, SD = 0.39), Model 2: prior percent correct (Mean = 0.84, SD = 0.37) df = 6130.65, p = 0.00144 < 0.01

A second preliminary analysis explores personalization within a selected experiment to further illustrate the personalization effect. The selected experiment, comparing video-based hints to text-based hints within the problem set entitled “Composition of Functions,” exhibits the qualitative interaction on which personalization is based. This simple analysis observes differences in completion rates within the training set for students with different values of prior percent correct. After filtering out observations with missing values, a logistic regression is applied to the remaining 110 observations.

The results of this simple analysis find that all terms of the regression are significant under a 95% confidence interval, suggesting that the interaction effect exists. The model itself, again run only within the training set, yields an AUC of 0.83 and R-Square of 0.4 using a 10-fold cross validation. To visualize this effect, similar to our first comparative analysis, we exemplify two methods of discretizing the continuous variable of prior percent correct.

In agreement with our logistic regression model, it is found that students with higher prior percent correct tend to have higher completion rates in the video hint condition, while students exhibiting lower prior percent correct appear to have higher completion rates in the text hint condition within the training set.

Contributions, Future Work, and Conclusions

This paper offers several contributions and offers an opportunity to significantly progress personalization in learning. Our first contribution is the unique dataset consisting of data from 22 randomized controlled A/B experiments and prior data on those participating students. The Pittsburgh Science of Learning Center’s DataShop has a repository consisting of data from many different experiments however they differ in their structures, formatting, and many include differing sets of covariates across experiments. To our knowledge there are no open datasets that have the same student covariates across several experiments with the same experimental design. Therefore we believe our dataset is unique in providing a means for researchers to perform methods of analysis that are not possible with other existing datasets. We propose a data mining competition using our dataset to search for features of personalized student learning. We believe this competition will be exciting to the educational data mining community as well as provide an unprecedented opportunity to search for features to help improve personalizing student learning in online educational platforms. We also provide a baseline model and preliminary analyses to accompany this data and competition proposal.

For immediate future work we intend to create the infrastructure for competitions to submit their predictions online for evaluation and implement the public leaderboard. With this infrastructure we can run more competitions focusing on the features that are important in this competition to further investigate the underlying student behavior. We can also examine other performance measures such as student completion, to observe other potential meaningful effects.
Citations


