Predicting Performance in a Small Private Online Course

Wan Han, Ding Jun, Gao Xiaopeng, Yu Qiaoye, Liu Kangxu

School of Computer Science and Engineering
Beihang University
Beijing, China
+86-10-82338059
{wanhan, dingjun, gxp, yuqiaoye, liukangxu}@buaa.edu.cn

ABSTRACT

In this paper, we describe how we build accurate predictive models of students' performance in a SPOC (small private online course). We document a performance prediction methodology from raw logging data based on OpenEdX platform to model analysis. We attempted to predict students' performance of Computer Structure Lab Course (Fall 2016) offering at Beihang University. 28 predictive features extracted for 377 students, and our model achieved an AUC (area under curve) in the range of 0.62-0.83 when predicting one week in advance. This work would help to identify at-risk students in a SPOC.

Keywords

SPOC, student performance prediction, study behavior analysis, educational data mining, at-risk students

1. INTRODUCTION

EdX has designed and built an open-source online learning platform (OpenEdX) for online education. In addition to offering online courses, participating universities are also committed to researching how students learn and how technology can transform learning both on-campus and online throughout the world.

Some researches focus on how to predict students' performance by using study-related data. Stapel, M. [1] presented an ensemble method to predict students' performance, which includes six classification algorithms. Elbadrawy, A. [2] developed multi-regression models based on regression algorithms for predicting, and Ren, Z. [3] designed different kinds of features based on MOOC courses' characters, which improved the performance of their predictor. In addition to study-related data, social behavior data is helpful in predicting [4].

In this paper, we describe the performance prediction problem, and present models we built. A summary of which features played a role in gaining accurate predictions is presented. The most fundamental contribution is the design, development and demonstration of a performance prediction methodology, from raw logging data to model analysis, including data preprocessing, feature engineering, model evaluation and outcome analysis.

2. PREDICTION PROBLEM DEFINITION

Our SPOC was composed of 3 tutorials and 9 projects in Fall 2016, learners studied the tutorials from week 1 to week 6, and we released project 0 at week 7. We found it was important for learners to move on only after they'd mastered the core concept. Students started one project and as they mastered corresponding content, that they need to pass the test in class, and then they could be awarded to the next project.

Here our performance prediction is to predict whether the learner could pass their test at the end of each week according to their study behavior. We define time slices as weekly units. Time slices started the first week in which in class test was offered (week 7), and ended in the 16^{th} week, after the final test had closed.

So we could use the logging data from week 1 to week 6 to predict the learners' performance at week 7. Furthermore, we used 'lead' represents how many weeks in advance to predict performance. We assign the performance label (x_1, 0 for unpassed the test or 1 for passed the test) of the lead week as the predictive problem label. 'Lag' means use how many weeks of historical variables to classify.

3. PREDICTING WEEK PERFORMANCE

We did not use the non-behavioral attribute such as a leaner's age, gender and others. Instead, we used some features that would show different style of learning habits. One type of behavioral variables is based on the learner's interaction with the educational resources, including time spent on resources and problem / homework. As Colin Taylor described in [5], taking the effort to extract complex predictive features that require relative comparison or temporal trends, rather than using the direct covariates of behavior, is one important contributor to successful prediction. For instance, we create an average number of submissions per problem for each learner (x9). Then we compare a learner's x9 value to the distribution for that week. Feature x16 is the percentile over the distribution and x17 is the percent as compared to the max of the distribution. We also extracted features that related to learners' study habits. For instance, feature to describe whether learners begin doing the problem / homework soon after it was released, and features to characterize the learners that submit problem / homework in timely fashion or at last minute fashion.

To build predictive models, we utilize a common approach of flattening the data- assembling the features from different weeks as separate variables.

We first used logistic regression as our binary predictive model. It calculates a weighted average of a set of variables as an input to the logit function. There are different coefficients for the feature values. For the binary classification problem, the output of the logit function becomes the estimated probability of a positive training example.

When applying the logistic regression to learner week performance prediction. We used 28 features to form the feature vectors, and maintained the week performance value as the label.

3.1 Predicting Performance

When evaluating the classifier's performance. A testing set comprised of untrained covariates and labels evaluates the performance of the model as following steps:

The logistic function learned is applied to each data point and the estimated probability of a positive label is produced. And then a decision rule is applied to determine the class label for each probability estimate. Given the estimated labels for each data point and the true labels we calculate the confusion matrix, true positives and false positives and then obtain an operating point on the ROC curve. Then evaluate the area under the curve and report it as the performance of the model on the test data.

We need to present the results for multiple prediction problems for different week simultaneously. Here means for each week during our course, we want to predict the students' week performance using different historical data. The heat map of a lower right triangular matrix is assembled as shown in figure 1.

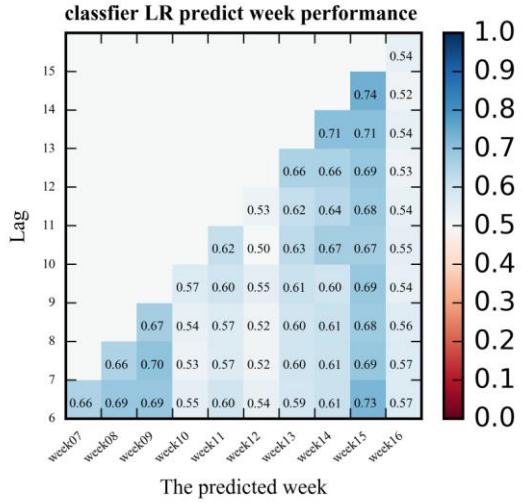

Figure 1. Logistic regression results

The x-axis of figure 1 is the week for which predictions are made in the experiment, while y-axis is the number of the how many week data we use for the prediction (lag). The color shown the area under the curve for the ROC the current model achieved.

We employed cross validation in all of our predictive modelling. Some partitions are used to construct a model, and others are used to evaluate the performance. Considering only 377 samples in our data set, we employed 3-fold cross validation and use the average of the ROC AUC over the folds as evaluation metric.

3.2 Feature Importance

We utilized randomized logistic regression methodology to identify the relative weighting of each features. As shown in figure 2, top features that had the most predictive power include whether learners interact with the resources more time (*max_observed_event_duration*), learners' interaction with the problems (*average_number_of_submissions_percentile*), study habits (*time_first_attempt, problem_finish_time_pre_start24h*, *problem_finish_time_pre_start48h*).

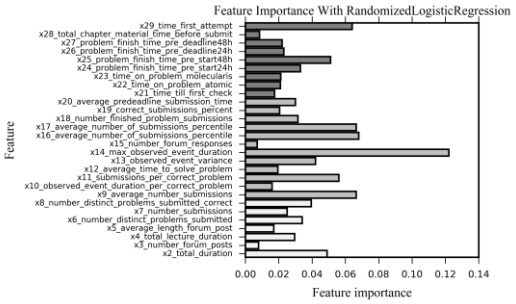

Figure 2. Relative importance of different features across all variants (lag / lead)

4. SUMMARY

We have taken an initial step towards identifying at-risk students in a SPOC, which could help instructors design interventions. Several prediction models are compared, with SVM preferred due to its good performance. The noteworthy accomplishments of our study when compared to other studies including: we extracted variable from the click stream logging data and then generate complex features which explain the learners' study behavior, especially how to describe the learners' study habits. We attributed SVM model to those variables as we achieve AUC in the range of 0.62-0.83 for one week ahead.

In the future, we will collaborate with course instructors to deploy our predictive models. And we will take more attention to why a student is failing, and what strategies make others' success in a SPOC or on-campus course.

5. ACKNOWLEDGMENTS

This research was supported by Teaching Research Funding in Honors College of Beihang University (2017) and Computer Information Specialty Construction Foundation Grant (No.201406025114).

6. REFERENCES

[1] Stapel, M., Zheng, Z., and Pinkwart, N. 2016. An ensemble method to predict student performance in an online math learning environment. In *Proceedings of the 9th International Conference on Educational Data Mining* (June 29 - July 2, 2016, Raleigh, NC, USA), 231-238.

[2] Elbadrawy, A., Studham, S., and Karypis, G. 2014. *Personalized multi-regression models for predicting students' performance in course activities*. Technical Report 14-011. University of Minnesota.

[3] Ren, Z., Rangwala, H., and Johri, A. 2016. Predicting Performance on MOOC Assessments using Multi-Regression Models. *arXiv preprint arXiv:1605.02269*.

[4] Bydžovská, H. 2016. A Comparative Analysis of Techniques for Predicting Student Performance. In *Proceedings of the 9th International Conference on Educational Data Mining* (June 29 - July 2, 2016, Raleigh, NC, USA), 306-311.

[5] Colin Taylor, Kalyan V., and Una-May O., 2014. Likely to stop? Predicting Stopout in Massive Open Online Courses. DOI = http://arxiv.org/pdf/1408.3382v1.pdf.