Behavior Prediction in MOOCs using Higher Granularity Temporal Information

Cheng Ye
Vanderbilt University
Cheng.Ye@vanderbilt.edu

Douglas H. Fisher
Vanderbilt University
douglas.h.fisher@vanderbilt.edu

John S. Kinnebrew
Vanderbilt University
john.s.kinnebrew@vanderbilt.edu

Gayathri Narasimham
Vanderbilt University
gayathri.narasimham@vanderbilt.edu

Gautam Biswas
Vanderbilt University
gautam.biswas@vanderbilt.edu

Katherine A. Brady
Vanderbilt University
katherine.a.brady@vanderbilt.edu

Brent J. Evans
Vanderbilt University
b.evans@vanderbilt.edu

L@S 2015, Mar 14-18, 2015, Vancouver, BC, Canada
ACM 978-1-4503-3411-2/15/03.
http://dx.doi.org/10.1145/2724660.2728687

Abstract
In this paper, we present early research evaluating the predictive power of a variety of temporal features across student subpopulations with distinctive behaviors at the beginning of the course. Initial results illustrate that these features predict important differences across the subpopulations and over time in the courses. Ultimately, these results have implications for effectively targeting adaptive scaffolding tailored to the particular intentions and goals of subpopulations in MOOCs.

Author Keywords
MOOCs, Behavior Prediction, Machine Learning

Introduction
The popularity and large initial enrollments in Massive Online Open Courses (MOOCs), driven by their wide accessibility, relative openness, and the reputation of the instructors and institutions offering these courses, has created significant interest in analyzing students' learning behaviors in these systems. However, a common pattern across the spectrum of MOOCs has been the large number of no-shows, early dropouts, and the generally low completion rates [2]. However, a certificate of completion may not be the goal of many students taking a MOOC. Rather, a variety of distinct MOOC subpopulations have been identified by their behavior [1, 4].

The common trend of low completion rates and the recognition of subpopulations with different behavior trends in MOOCs suggests an overall research question that motivates the analysis and initial results presented in this paper: *What behavior features are useful for accurate early prediction of student dropout and performance in different MOOC subpopulations*? In the longer term, our hope is that early dropout predictions tailored to specific subpopulations will provide a framework for targeting adaptive scaffolding mechanisms in MOOCs that could provide individualized guidance and small group support for achieving a variety of goals.

A number of researchers have explored factors related to low completion rates and dropout prediction in MOOCs. For example, Halawa *et al.* [3] attempted to predict dropout with a variety of lecture-viewing behavior features; specifically they looked at: video-skip, assignment-skip, lag (e.g., whether the student was viewing week 1 videos after week 3 videos had been released), and assignment performance. In this paper, we use similar lecture behavior features to those in [3] (e.g., number of lecture video views for current and previous weeks) as part of a baseline feature set, but show that the introduction of finer-grained and temporal features boosts early predictive performance in certain subpopulations.

Methodology, Tools, and Data

In this paper, we present initial results from the analysis of two MOOCs offered on the Coursera platform. These courses have an overall structure common to many MOOCs, including a series of short video lectures released on a weekly basis, weekly multiple-choice quizzes that are related to the lecture materials, and larger peer-graded assignments. However, in order to provide a preliminary evaluation of the generality of our approach and analysis, we chose courses from two different domains: a course on principles in a programming domain and a course on data management and administration in the clinical research domain.

The first course, "Pattern-Oriented Software Architectures for Concurrent and Networked Software" (POSA), is a ten-week Coursera MOOC covering programming principles for networking and other software dealing with concurrent operations. The second course, "Data Management for Clinical Research" (DM), is a seven-week Coursera MOOC focused on (clinical) research data collection and management. Both courses offered certificates of completion and distinction based on weekly quizzes and peer-graded assignments.

Initial analysis of student behavior in the first week of the course suggested two distinct subpopulations among the "active" students who at least watched a lecture in week 1: (1) students who watched at least one lecture in week 1 in each MOOC but did not take the quiz, *i.e.*, the *lecture only* group; and (2) students who watched at least one lecture and took the summative multiple-choice quiz that week, *i.e.*, the *lecture and quiz* group. Analyzing students' subsequent behavior with respect to lecture-watching and quiz-taking, indicated that most students in the lecture-and-quiz group continued watching lectures and taking the weekly quizzes (unless they dropped out of the course), and most students in the lecture-only group continued watching most lectures but not taking quizzes (unless they dropped out of the course).

The specific categories of features we defined for this analysis were:
Baseline lecture quantity features provide a baseline for lecture-watching features that are defined by the combinations along 3 dimensions: (1) aggregation (total

number of lecture accesses or number of unique lectures accessed), (2) access type (viewing online or downloading), and (3) course segment (grouping lectures by whether they are designated for previous weeks, the current week, or subsequent weeks), resulting in individual features, such as "total number of lectures designated for the current week that were viewed online".
Lecture temporal features are aggregated across the current week's lectures accessed by the student: (1) average first access time of lectures (as the offset between the time the lecture became available and the time the student first accessed the lecture), (2) average first lecture-embedded quiz answer time (as the offset between the time the lecture became available and the time the student first answered an embedded quiz question in the lecture), and (3) difference between average first lecture-embedded quiz answer time for the current week versus the previous week.
Lecture-quiz quantity features describe the frequency of lecture-watching activity with respect to answering lecture-embedded quiz questions: (1) the total number of lecture-embedded quiz questions answered, (2) average number of times a lecture was accessed before the first lecture-embedded quiz question was answered, and (3) average number of times the lecture was accessed after the first lecture-embedded quiz question was answered.

In order to evaluate the predictive potential of different feature sets early in the MOOCs, we assigned one of four labels to each student in the course that indicated whether they ultimately dropped out in week 1, week 2, week 3, or not dropped out by the end of week 3 (i.e., the student dropped out in a later week or continued through the end of the course). We used a soft threshold for dropout, identifying a student as having dropped out in a given week if they accessed fewer than 10% of the remaining lectures in the course and performed no further assessment activities (i.e., summative weekly quizzes or submission of peer-graded assignments). Random forests were used for classification as they consistently performed better than, or on par with, other classifiers (logistic regression, support vector machines, and decision trees) tested in preliminary analysis.

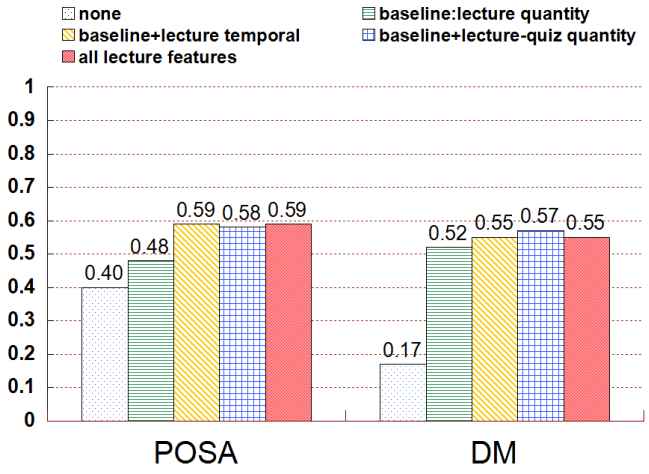

Figure 1: Dropout prediction F1 score in lecture-only group

Results

To better understand how useful these features were in early dropout prediction across different subpopulations, Figures 1 and 2 show the results of this analysis applied separately to the lecture-only and the lecture-and-quiz groups. In the lecture-only group, the addition of either *lecture temporal* or *lecture-quiz quantity* features improved predictive performance over the baseline, however, little or no additional predictive power appears to be gained by using both additional feature sets. This suggests that while temporal lecture-access patterns and

lecture-quiz patterns may each be used for prediction, the underlying factors likely to result in dropout may tend to manifest in *both* the timing of lecture-watching activities and the quantity of lecture-embedded question answering.

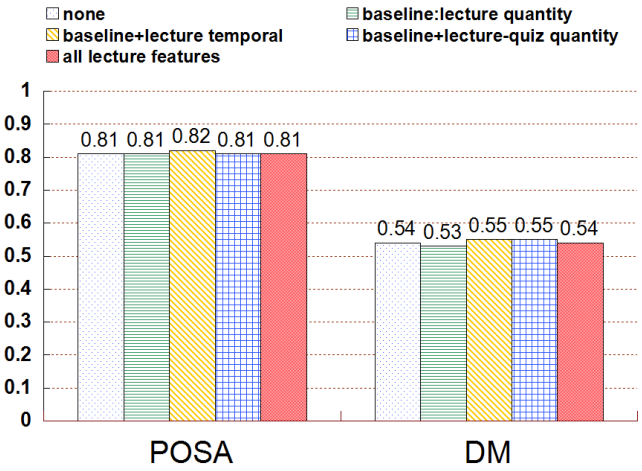

Figure 2: Dropout prediction F1 score in lecture-and-quiz group

Discussion and Conclusion

Initial results suggest that in the lecture-only subpopulation, sufficiently fine-grained lecture behavior features could be very useful in targeting immediate scaffolding for students who are dis-engaging with the course as early as the first week. However, all lecture behavior features analyzed were of little to no use in early prediction of dropout for the lecture-and-quiz subpopulation. Overall, the results illustrate the importance of identifying and analyzing different subpopulations for targeting scaffolding in the diverse populations of MOOCs.

In future work, we will extend these results by identifying specific rules based on the presented and further fine-grained behavior features to predict dropout in different subpopulations. We will improve the definition of subpopulations by running unsupervised learning methods on the extended feature sets to determine finer-grained groups that provide contextual information for developing specific scaffolds. In addition to helping us better understand which specific behaviors relate to dropout and performance, this will be vital for targeting adaptive scaffolding to support students. In the longer term, we hope to employ refined versions of these predictive models to identify students for targeted scaffolding in a variety of subpopulations to support them in achieving their individual objectives in taking a MOOC.

References

[1] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., and Seaton, D. Studying learning in the worldwide classroom: Research into edxs first mooc. *Research & Practice in Assessment 8* (2013), 13–25.

[2] Clow, D. Moocs and the funnel of participation. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, ACM (2013), 185–189.

[3] Halawa, S., Greene, D., and Mitchell, J. Dropout prediction in moocs using learner activity features. In *Proceedings of the European MOOC Summit (EMOOCs 2014)* (Lausanne, Switzerland, 2014).

[4] Kizilcec, R. F., Piech, C., and Schneider, E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge*, ACM (2013), 170–179.