(2014). Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information. *Journal of Learning Analytics*, *1* (3), 169–172.

Early Prediction of Student Dropout and Performance in MOOCs using Higher Granularity Temporal Information

Cheng Ye and Gautam Biswas
Vanderbilt University, USA
cheng.ye@vanderbilt.edu

ABSTRACT: Our project is motivated by the early dropout and low completion rate problem in MOOCs. We have extended traditional features for MOOC analysis with richer and higher granularity information to make more accurate predictions of dropout and performance. The results show that finer-grained temporal information increases the predictive power in the early phases of the Pattern-Oriented Software Architectures (POSA) MOOC offered in summer 2013 by Vanderbilt University. As a next step, we plan to develop unsupervised learning methods with our extended feature set to define profiles that can be used for effective scaffolding and feedback.

KEYWORDS: MOOCs, feature engineering, performance, dropout, prediction

1. MOTIVATION

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 them has created significant interest in analyzing and characterizing student learning behaviours to support scaffolding in these systems (Brown, 2013; Matkin, 2013). A common pattern in MOOCs has been the large number of no shows, early dropout, and low completion rates (Clow, 2013; Hill, 2013). For example, in the POSA course offered by Vanderbilt University, there were 31,053 enrollees: only 6,953 students watched a video lecture in week 1; of these, only 1,699 students also took a quiz that week. At the end of the course, 1,051 students got a pass certificate and 592 obtained a distinction grade. These numbers motivate our research question: Can we derive accurate and reliable early predictors of student dropout and performance in MOOC environments? Early dropout predictions will provide a framework for developing scaffolding mechanisms in MOOCs that provide individualized guidance and small-group support, which should significantly increase retention rates. For example, avoiding the video-quizzes embedded in the lectures is a good indicator of early dropout. For students who tend to ignore the embedded quizzes, therefore, we may briefly explain the importance of formative assessments, provide feedback that links the quiz questions to the topic-related material in the video, and provide more explanation for the alternatives provided for the answer.

A number of studies have been conducted on student retention in traditional settings, e.g., dropout analysis of university freshmen (Dekker, Pechenizkiy, & Vleeshouwers, 2009). MOOCs, in their current form, represent a different form of learning and target a greater diversity of learners as compared to traditional courses. Recent studies have explored factors related to student dropout in MOOCs. Some researchers have successfully used the number and frequency of forum posts, and subsequent communication (Yang et al., 2013) to measure student interest and retention in MOOCs. Ramesh et al. (2013) extended forum post analysis to include linguistic and structural features from forum interactions. Kloft et al. (2014) extracted weekly aggregated behaviour features, such as number of video views, number of active days and non-behaviour features, such as country, for analysis of dropout rates. Halawa, Greene, & Mitchell (2014) selected four features as predictors: video-skip, assignment-skip, lag, and assignment performance. Others have introduced temporal features, e.g., “lastQuiz,” “lastLecture” (Ramesh et al., 2013), “lag” (Halawa et al., 2014), and “pre-deadline submission time” (Taylor, Veeramachaneni, & O'Reilly, 2014). Whereas some of these methods predict student dropout accurately one week in advance (e.g., Taylor et al., 2014; Kloft et al., 2014), others have predicted students’ final performance using only week 1 data (Jiang et al., 2014; Taylor et al., 2014). In our work, we extend the feature space to include data at more specific temporal granularities with the goal of making more accurate predictions of dropout and performance early in a MOOC offering.

2. RESULTS AND ANALYSIS

We analyzed data from video lectures, weekly quizzes, and peer assessments from the ten-week POSA course. Two discrete variables were selected as the prediction targets: 1) the dropout week (when a student watched less than 10% of the remaining lectures and stopped submitting assignments); 2) the final grade (normal certificate or distinction) upon completing the course. We included traditional features for defining student behaviour, such as video lecture downloads, taking weekly quizzes, and solving peer assessments, along with the following additional features: 1) number of lecture views and video quiz attempts by week; and 2) temporal information such as when a lecture was viewed or an assessment started during the week. We performed a Pearson’s r correlation between individual features and the student’s final performance (normal completion vs. distinction) and found that our higher granularity features, e.g., the number of video quizzes taken per week and when a lecture video was first accessed, increased accuracy in predicting dropout and final performance over earlier studies. Similarly, the time when students started peer-graded assessments were a good early predictor of their dropout rate and performance. Once scores on a peer assessment were available, they became the best indicators of performance.

Our initial analysis of week 1 data indicated two dominant groups: 1) students who watched lectures and took the assigned quizzes (1,699 students) and 2) students who only watched lectures (6,953 students). This grouping was a strong indicator of dropouts: 60% of the lecture-only group dropped out by week 4 whereas only 20% of the quiz takers dropped out. Further, the average grade for the lecture-only students was 3.2%. This number was 66% for the other students. Overall, analysis showed that more precise temporal features and more quantitative information improved early prediction accuracies and false alarm rates as compared to using only assessment score features.

In terms of contributions to learning analytics, we are studying feature selection and feature extraction methods to make better predictions with large data sets derived from MOOCs. We have gone beyond traditional methods and demonstrated the advantages of extracting more fine-grained features for analysis and early prediction. As a next step, we will use this more fine-grained data with unsupervised learning methods to determine if we can better profile students in terms of likelihood for dropping out and reasons for dropping out. More contextual data may also be helpful for such analyses.

REFERENCES

Brown, A. (2013). MOOCs make their move. *The Bent*, *104*(2), 13−17. http://www.tbp.org/pubs/Features/Sp13Brown.pdf

Clow, D. (2013). MOOCs and the funnel of participation. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK ’13), 8–12 April 2013, Leuven, Belgium (pp. 185−189). New York: ACM.

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting student drop out: A case study. International Working Group on Educational Data Mining. http://files.eric.ed.gov/fulltext/ED539082.pdf

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *Proceedings of the European MOOC Summit* (EMOOCs 2014). https://oerknowledgecloud.org/sites/oerknowledgecloud.org/files/In_depth_37_1 (1).pdf

Hill, P. (2013. Emerging student patterns in MOOCs: A (revised) graphical view. Retrieved from http://mfeldstein.com.

Jiang, S., Warschauer, M., Williams, A. E., O’Dowd, D., & Schenke, K. (2014). Predicting MOOC performance with week 1 behavior. *Proceedings of the 7th International Conference on Educational Data Mining*. http://educationaldatamining.org/EDM2014/uploads/procs2014/short papers/273_EDM-2014-Short.pdf

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014) Predicting MOOC dropout over weeks using machine learning methods. *Modeling Large Scale Social Interaction in Massively Open Online Courses Workshop* (EMNLP 2014). http://www2.informatik.hu-berlin.de/~kloftmar/publications/emnlp_mooc.pdf

Matkin, G. W. (2013). Massive open online courses: Looking ahead by looking back. *Continuing Higher Education Review, 77,* 49. https://www.unx.uci.edu/pdfs/dean/matkin_2014_moocs.pdf

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013). Modeling learner engagement in MOOCs using probabilistic soft logic. *NIPS Workshop on Data Driven Education*. http://linqs.cs.umd.edu/basilic/web/Publications/2013/ramesh:nipsws13/ramesh-nipsws13.pdf

Taylor, C., Veeramachaneni, K., & O'Reilly, U. M. (2014). Likely to stop? Predicting stopout in massive open online courses. arXiv preprint arXiv:1408.3382. http://arxiv.org/pdf/1408.3382v1.pdf

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. *Proceedings of the 2013 NIPS Data-Driven Education Workshop*. http://lytics.stanford.edu/datadriveneducation/papers/yangetal.pdf