Uncovering Hidden Engagement Patterns for Predicting Learner Performance in MOOCs

[1]Arti Ramesh, [1]Dan Goldwasser, [1]Bert Huang, [1]Hal Daumé III, [2]Lise Getoor
[1]University of Maryland, College Park [2]University of California, Santa Cruz
{artir, bert, hal}@cs.umd.edu, goldwas1@umiacs.umd.edu, getoor@soe.ucsc.edu

ABSTRACT

Maintaining and cultivating student engagement is a prerequisite for MOOCs to have broad educational impact. Understanding student engagement as a course progresses helps characterize student learning patterns and can aid in minimizing dropout rates, initiating instructor intervention. In this paper, we construct a probabilistic model connecting student behavior and class performance, formulating student engagement types as latent variables. We show that our model identifies course success indicators that can be used by instructors to initiate interventions and assist students.

Author Keywords

MOOC, learner engagement, probabilistic modeling

ACM Classification Keywords

K.3.1. Computer Uses in Education

INTRODUCTION

Sustaining student engagement is important in both classroom and online courses. Unlike classroom courses, the prevalent method for facilitating student-teacher interaction in MOOCs is to use online forums where students post questions and obtain feedback from the instructor or other students. Absence of direct teacher interaction, large number of students, and their diverse backgrounds make it challenging for MOOC instructors to gauge the level of student engagement and involvement and take appropriate actions.

We develop a data-driven approach for modeling student engagement. Online activities such as interactions with other learners or staff on discussion forums, completion of assignments, and *language* used by the learners in posts serve as useful indicators for gauging engagement. Combining language analysis of forum posts with graph analysis over very large networks of entities (e.g., students, instructors, topics, assignments) to capture domain dynamics is challenging. We propose a model that uses behavioral, structural, and linguistic (polarity and subjectivity of forum posts) aspects to distinguish between forms of student engagement (active and passive). The engagement types are represented as latent variables in our model and are learned from observed data. We then use the latent engagement estimates to predict learners' performance and reason about learners' behavior.

L@S 2014, March 4–5, 2014, Atlanta, Georgia, USA.
ACM 978-1-4503-2669-8/14/03.
http://dx.doi.org/10.1145/2556325.2567857

MODELING LEARNER ENGAGEMENT

We construct the following types of features from learners' interaction with the MOOC website—1) behavioral—constructed from user behavior such as posting in, viewing or voting on discussion forums, lecture views, and quiz completion; 2) linguistic—polarity and subjectivity values of forum-content calculated using Opinionfinder [3]; 3) structural—constructed from forum-interaction; and 4) temporal—features from user activity over time. To model the interactions between these features and learner engagement, we use *probabilistic soft logic* (PSL)[1], which is a system for relational probabilistic modeling. PSL enables us to encode observed features, latent, and target variables as logical predicates and capture domain knowledge by constructing rules over these predicates. PSL interprets these rules in a parameterized probabilistic model and is able to perform efficient inference and parameter fitting using machine learning algorithms. We experiment with predicting two aspects of learner performance—1) whether the learner earned a statement of accomplishment in the course, and 2) whether the learner survived the later part of the course. We refer to these as *learner performance* and *learner survival* models.

Learner Performance Models

We construct two different PSL models for predicting learner performance —1) a direct model (denoted DIRECT) that infers performance from observable features, 2) a latent variable model (LATENT) that infers student engagement as a hidden variable to predict learner performance. We treat learner engagement types—active, passive, and disengaged as latent variables and associate conjunctions of observed features to one or more forms of engagement. We then evaluate the latent formulation by using it to infer learner performance.

Learner Survival Models

In the survival PSL models, we split the course into three phases—*start*, *middle*, and *end*. The phase-splits are chosen according to the number of quizzes and lectures in the courses, with equal distribution of quizzes and lectures in the splits. We use the same features as in the performance models, however the features are computed for the phase(s) of the course in consideration. Here, we predict if each learner *survives* a phase in the course, i.e., whether the learner takes a quiz that immediately follows the split-point. We construct two models for predicting learner survival—a DIRECT model with the features directly implying survival and a LATENT model using engagement as a latent layer. We refer the reader to [2], for more details.

EMPIRICAL EVALUATION

We design our experiments around performance measures—course grades and course attendance, and show how our engagement formulation helps in reliably predicting these measures. We evaluate our models on Coursera MOOC *Surviving Disruptive Technologies*. This seven-week course had 1665 users participating in the forums and 826 users completing the course with a nonzero grade. We use 10-fold cross-validation in our experiments, leaving out 10% of the data for testing and the rest for training, where the model weights are learned.

Learner Performance Results

For the learner performance models, we filter the data to include only learners that attempted one or more quizzes or assignments in the course and earned a non-zero score. We labeled the ones that earned a statement of accomplishment as positive instances (*performance* 1.0) and others as negative (*performance* 0.0). These labels are used as ground truth to train and test the models. From experimental results in Table 1, we observe that the LATENT PSL model performs better at predicting learner performance.

	AUC-PR Pos	AUC-PR Neg.	AUC-ROC	Kendall
DIRECT	0.74	0.54	0.66	0.58
LATENT	0.75	0.57	0.69	0.60

Table 1: Performance of DIRECT and LATENT PSL performance models in Disruptive Technologies course.

	start	middle	end	start-mid	start-end
DIRECT	0.72	0.75	0.89	0.70	0.72
LATENT	0.75	0.80	0.95	0.76	0.82

Table 2: Performance of DIRECT and LATENT PSL survival models for different data-splits (AUC-ROC)

Learner Survival Analysis

Predicting student performance can provide instructors with a powerful tool if these predictions can be made *reliably* before the students disengage and drop out. We model this scenario by training our model over data collected early in the course. In the survival models, we use the subset of learners who earned an overall score greater than 0, and assign binary labels based on activity after our phase-split point. Our experiments in the survival models are aimed at measuring learner health by understanding 1) factors influencing learners' continuous survival, 2) engagement types and movement across types, and 3) phase-splits that are most important for predicting learner survival. Table 2 gives the accuracy values of DIRECT and LATENT models for different phase-splits in the data. The tag *start-mid* refers to data collected by combining phases *start* and *middle*; *start-end* refers to data collected over the entire course. Consistent with previous experiments, LATENT survival model has higher prediction reliability.

Early Prediction

Early prediction scores, described in Table 2 under *start*, *middle*, and *start-mid* tags (i.e., survival prediction using partial data), show that our model makes better predictions (as the data available to our model is closer to the actual decision point).

Results show that monitoring learner activity in the middle phase is most important for predicting whether the learner will survive the length of the course. Our model performs best when using data from the *middle* phase, compared to using data from the *start* phase, and an almost equal accuracy values when compared to *start-mid*. We hypothesize that this is due to the presence of a larger learner population in the *start* that fails to remain engaged. Eliminating data collected from this population helps improve our prediction of learner survival, indicated by an increase in accuracy for *middle*.

Analyzing Engagement Pattern Dynamics

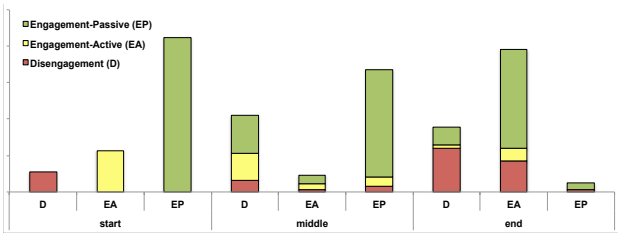

Figure 1: Engagement labels distribution of students who completed the course. Label transitions are captured by coloring bars according to assignments at the previous time point.

We analyze learners' engagement patterns using the engagement values predicted by our model. Learners are classified into one of the engagement types by considering the dominant value of engagement as predicted by the model. Figure 1 shows our engagement values for learners that continued in the course until completion. The labels D, EA and EP refer to *disengagement*, *engagement_active* and *engagement_passive*. We show engagement assignment levels at each time span (*start, middle, end*), and color code the bars according to the previous engagement assignments. It can be observed that the most engaged learners only exhibit passive forms of engagement in the *start* and *middle* phases. While in the *end* phase, learners tend to become more actively engaged.

CONCLUSION

In this work, we formalize, using PSL, our intuition that student engagement can be modeled as a complex interaction of behavioral, linguistic, and social cues. Our results show that our model can construct an interpretation for latent engagement types from data, based on their impact on performance.

REFERENCES

1. Broecheler, M., Mihalkova, L., and Getoor, L. Probabilistic similarity logic. In *Uncertainty in Artificial Intelligence (UAI)* (2010).
2. Ramesh, A., Goldwasser, D., Huang, B., Daume III, H., and Getoor, L. Modeling learner engagement in moocs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education* (2013).
3. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations* (2005).