Connecting the Dots: Predicting student grade sequences from bursty MOOC interactions over time

Tanmay Sinha
ArticuLab
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
tanmays@andrew.cmu.edu

Justine Cassell
ArticuLab
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
justine@cs.cmu.edu

L@S 2015, Mar 14-18, 2015, Vancouver, BC, Canada.
ACM 978-1-4503-3411-2/15/03.
http://dx.doi.org/10.1145/2724660.2728669

Abstract
In this work, we track the interaction of students across multiple Massive Open Online Courses (MOOCs) on edX. Leveraging the "burstiness" factor of three of the most commonly exhibited interaction forms made possible by online learning (i.e, video lecture viewing, coursework access and discussion forum posting), we take on the task of predicting student performance (operationalized as grade) across these courses. Specifically, we utilize the probabilistic framework of Conditional Random Fields (CRF) to formalize the problem of predicting the sequence of grades achieved by a student in different MOOCs, taking into account the contextual dependency of this outcome measure on students' general interaction trend across courses. Based on a comparative analysis of the combination of interaction features, our best CRF model can achieve a precision of 0.581, recall of 0.660 and a weighted F-score of 0.560, outweighing several baseline discriminative classifiers applied at each sequence position. These findings have implications for initiating early instructor intervention, so as to engage students along less active interaction dimensions that could be associated with low grades.

Introduction and Motivation
The potential of MOOCs, despite their impact on education worldwide, is still relatively untapped. There

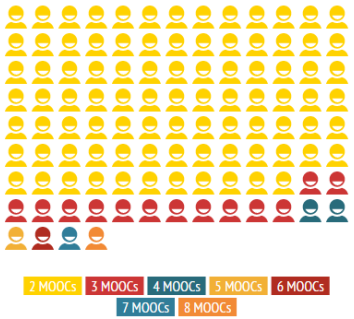

Figure 1: Pictorial Depiction of Data Distribution of Students who took more than 1 course on the edX platform in 2013 ([#MOOCs,#Students]: [2,8906], [3,1259], [4,202], [5,41], [6,7], [7,5], [8,1])

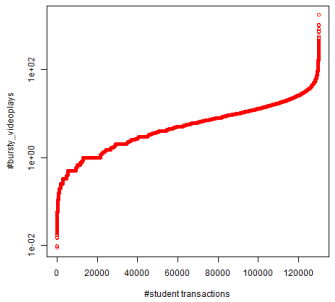

Figure 2: Scatterplot of #student transactions (Linear X axis) vs #burstyvideoplays (Logarithmic Y axis)

has been considerable commentary about gauging achievement in MOOCs and trying to improve retention, while accommodating participation differences and degrees of commitment along the modalities that have been made affordable. One salient assessment measure that indexes performance levels and is conventionally used to incentivize participation is grade. In this paper we consider the problem of "grade sequence prediction" or "grade labeling": to classify a students' grade points scored in multiple MOOC courses into one of the several defined categories. Such a labeling requires contextual information, because the labels are dependent on students' generic online learning strategy across courses.

Prior work predicting student performance from event traces has utilized measures such as motivation for signing up, obtaining help on discussion forums, and course features such as video lecture presentation and course content coverage. However, most of the focus has been on the student's current MOOC. If we were to rely on these analytics alone, it would be hard to capture any inherent student learning styles that might hold across different MOOCs. And we believe that these learning styles, if they match MOOC characteristics, will be a strong contributor to achievement. For example, if a student has been achieving high grades (say, $>$80%) by following a consistent participation pattern of spending more time watching video lectures and posting to the discussion forums, while never accessing chapters related to the coursework, it is legitimate to hypothesize that effective predictions of his grades in future MOOCs can be made based on those behaviors in prior courses. Furthermore, there may be dependencies among the interaction features themselves, for instance, higher video lecture viewing might be related to heavy discussion forum clarifications for a MOOC on specialized topics, while it may not hold for a more introductory level MOOC.

Because including such interdependent interaction features is difficult to do while retaining tractability and making independence assumptions can affect performance, we are interested in estimating the posterior over grade labels achieved by a student across different MOOCs, given the observed interaction. This motivates the use of Conditional Random Fields (CRF), originally proposed by [3] for labeling text sequences. As a contrast to generative modeling approaches that capture the joint probability distribution $p(y, x)$, CRF follow a discriminative modeling approach to capture the conditional probability distribution $p(y|x)$ of label sequences $(y_1, y_2...y_n)$, given the input sequences $(x_1, x_2...x_n)$, without requiring calculation of potentially dependent features in $p(x)$ that are not required for classification.

We focus on the "burstiness" dimension of students' behavior in MOOCs, by which we mean unusually high rates of interaction. For example, some students usually post to MOOC forums once a week for course X, yet suddenly exhibit a posting pattern of 10 posts/day for course Y. Or some students usually watch 3-5 video lectures/course, and then suddenly start watching 20-25 video lectures/course. Such a longitudinal variance, along with the potential to represent "short-term" interest spikes dynamically driven by either a)captivating aspects in a particular MOOC, or b)problems in understanding course content, makes "burstiness" an important dimension to consider. Our interest is in making "long-term" predictions on grade points that a student achieves, by leveraging this aspect of the interaction.

Study Context

Our study context comprises of de-identified data from the first year (Academic Year 2013: Fall 2012, Spring 2013, and Summer 2013) of 13 MITx and HarvardX courses offered on the edX platform [4]. This is the first

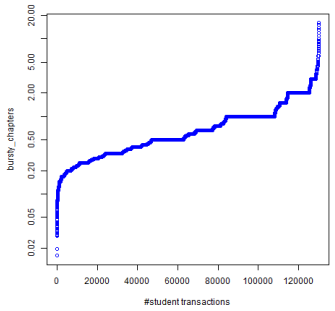

Figure 3: Scatterplot of #student transactions (Linear X axis) vs #burstychapters (Logarithmic Y axis)

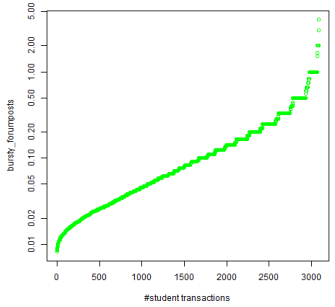

Figure 4: Scatterplot of #student transactions (Linear X axis) vs #burstyforumposts (Logarithmic Y axis)

large scale public MOOC data repository that contains information (aggregated records) about diverse kinds of learning interactions and outcome measures for the same students who took multiple courses on the platform. Our study is confined to $\approx$ 10000 students who took at least 2 or more courses on the edX platform and had some form of interaction with the particular MOOCs (Figure 1).

Methodology

We first operationalize three kinds of "burstiness" indices from students' interactions for all their MOOC courses taken. For all the 3 indices, a higher value denotes more "burstiness" and vice versa. Firstly, bursty number of play video events/clicks **(Bursty #videoplays)** is defined as the total number of play video events within a course, divided by the number of unique days student interacted with the particular course (Figure 2). In an analysis of video lecture viewing behavior across 4 edX courses, prior work [2] has shown that sudden spikes in video play events can be triggered both by type of students accessing the lecture (first time watchers versus re-watchers) and by the type of video lecture itself (such as tutorials versus lectures). Secondly, bursty number of chapters **(Bursty #chapters)** is defined as the total number of chapters (within the Courseware) with which the student interacted, divided by the number of unique days student interacted with the particular course (Figure 3). For edX MOOCs, Courseware tab is the home of the videos lectures, problem sets and exams. In an analysis of navigational differences across the same 4 edX courses, [1] found evidence for students over 40 accessing the digital course textbook 27% more frequently than those under 40 and thus exhibiting more bursty behavior. Thirdly, bursty number of discussion forum posts **(Bursty #forumposts)** is the total number of posts to the discussion forum, divided by the number of unique days student interacted with the particular course (Figure 4). Prior literature suggests that more burstiness on MOOC discussion forums is associated with greater likelihood of attrition [6], which we in turn expect to affect the grade points scored by a student. Next, we discretize the above burstiness indices into four categories based on equal frequency: Low (L), Medium (M), High (H) and Very High (V). Additionally, we also discretize the grade points scored by students in each course into the same 4 categories by equal frequency. To account for variability in course activity levels and course grading criteria, the discretization for each course is performed separately.

We then use a generic sequence of tokens representation for our input to the CRF framework. Figure 5 depicts this representation. For every student, each token represents a)3 feature values (Bursty #videoplays, Bursty #chapters, Bursty #forumposts) for the particular MOOC taken, and b)grade category (tag) for that course which is going to be trained by CRF. Thus, a sequence of tokens represents feature values and grade labels for all MOOCs taken in a sequential order by the student. Additionally, to specify the associations between output grade sequences (denoted by y) and input feature sequences (denoted by x), we utilize the following 4 feature templates for each of the 3 "burstiness" features: a)**1l**: current feature value, previous feature value, next feature value (input ± 1) and independence assumptions among y, b)**2l**: current feature value, previous 2 feature values, next 2 feature values (input ± 2) and independence assumptions among y, c)**1D**: current feature value, previous feature value, next feature value (input ± 1), combinations of previous output token (grade label) and current output token (dependence assumptions among y), d)**2D**: current feature value, previous 2 feature values, next 2 feature values (input ± 2), combinations of previous output token (grade label) and current output token (dependence assumptions among y). The intuition is to a)differentiate both long

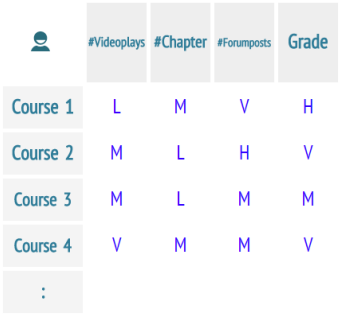

Figure 5: Example depicting "sequence of tokens" representation of input feature sequence along with the grade points sequence for a student who has taken 4 MOOCs. Burstiness indices & Grade points are discretized into four categories based on equal frequency: Low (L), Medium (M), High (H) and Very High (V).

and short term interaction dependencies, and b)accommodate potentially correlated features of the inputs, while performing the training discriminatively.

Results

We do a 70:30 split of the data into training and test sets. The feature cutoff for training the CRF is 5. The output consists a)marginal probabilities for grade label achieved by a student in each course, and b)conditional probability for the grade sequence, given the interaction feature sequence. Evaluation of the grade sequence tagging is done using precision, recall and weighted F-score measure to account for label imbalance. As **baselines**, we also employ Logistic Regression and Sequential Minimal Optimization (SMO with exponent 1) discriminative classifiers that treat the grade sequence labeling problem as a sequence of classification problems, one for each of the labels in the sequence (5 fold cross validation). The results in Table 1 substantiate that the **1I** CRF feature template configuration out-performs these baseline approaches that are myopic about the impact of current decision on later decisions. CRF improves predictability by allowing for a much richer input feature set.

Conclusion

In summary, we showed that "burstiness" factors can make significant predictions on students' grade sequence across different MOOCs. This can be used to facilitate real-time remediation from instructors, by a)combining finer grained features from video viewing, navigational and forum activities [5], b)augmenting such behavioral traces gathered up till specific course time points, with interactions from prior courses, to predict potential grades and tailor course content for students falling in high and low grade clusters, c)filtering these clusters by motivation to precisely gauge the intervention type (for e.g, pointers to applications and challenging learning materials would be a more preferable strategy for students who intended to specialize on certain topics and are likely to score average grade, than those who made an informed commitment for course completion and will score similarly).

Model	Precision	Recall	F-Score
LogReg (baseline)	0.588	0.659	0.558
SMO (baseline)	0.522	0.647	0.521
CRF 1I	**0.581**	**0.660**	**0.560**
CRF 2I	0.568	0.659	0.560
CRF 1D	0.584	0.655	0.551
CRF 2D	0.570	0.656	0.554

Table 1: Evaluation measures on test set using baselines (Logistic Regression & SMO) and 4 different CRF templates.

References

[1] Guo, P. J., and Reinecke, K. Demographic differences in how students navigate through moocs. In *Proceedings of ACM L@S* (2014).

[2] Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., and Miller, R. C. Understanding in video dropouts and interaction peaks inonline lecture videos. In *Proceedings of ACM L@S* (2014), 31–40.

[3] Lafferty, J., McCallum, A., and Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[4] MITx, and HarvardX. Harvardx-mitx person-course academic year 2013 de-identified dataset, version 2.0. In *http://dx.doi.org/10.7910/DVN/26147*, Harvard Dataverse Network [Distributor] V10 [Version] (2014).

[5] Sinha, T., Li, N., Jermann, P., and Dillenbourg, P. Capturing attrition intensifying structural traits from didactic interaction sequences of mooc learners. *Proceedings of EMNLP* (2014).

[6] Yang, D., Sinha, T., Adamson, D., and Rosé, C. P. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the NIPS* (2013).