Predicting Attrition Along the Way: The UIUC Model

Bussaba Amnueypornsakul, Suma Bhat and Phakpoom Chinprutthiwong
University of Illinois,
Urbana-Champaign, USA
{amnueyp1,spbhat2,chinpru2}@illinois.edu

Abstract

Discussion forum and clickstream are two primary data streams that enable mining of student behavior in a massively open online course. A student's participation in the discussion forum gives direct access to the opinions and concerns of the student. However, the low participation (5-10%) in discussion forums, prompts the modeling of user behavior based on clickstream information. Here we study a predictive model for learner attrition on a given week using information mined just from the clickstream. Features that are related to the quiz attempt/submission and those that capture interaction with various course components are found to be reasonable predictors of attrition in a given week.

1 Introduction

As an emerging area that promises new horizons in the landscape resulting from the merger of technology and pedagogy massively open online courses (MOOCs) offer unprecedented avenues for analyzing many aspects of learning at scales not imagine before. The concept though in its incipient stages offers a fertile ground for analyzing learner characteristics that span demographies, learning styles, and motivating factors. At the same time, their asynchronous and impersonal approach to learning and teaching, gives rise to several challenges, one of which is student retention.

In the absence of a personal communication between the teacher and the student in such a scenario, it becomes imperative to be able to understand class dynamics based on the course logs that are available. This serves the efforts of the instructor to better attend to the needs of the class at large. One such analysis is to be able to predict if a student will drop out or continue his/her participation in the course which is the shared task of the EMNLP 2014 Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses (Rose and Siemens, 2014).

Our approach is to model student attrition as being a function of interaction with various course components.

2 Related Works

The task of predicting student behavior has been the topic of several recent studies. In this context course logs have been analyzed with an effort to predict students? behavior. The available studies can be classified based on the type of course data that has been used for the analysis as those using discussion forum data and those using clickstream data.

Studies using only discussion forum to understand user-behavior rely only on available discussion forum posts as their source of information. In this context, in (Rosé et al., 2014) it was observed that students' forum activity in the first week can reasonably predict the likelihood of users dropping out. Taking a sentiment analysis approach, Wen et al. (Wen et al., 2014b) observed a correlation between user sentiments expressed via forum posts and their chance of dropping out. Motivation being a crucial aspect for a successful online learning experience, (Wen et al., 2014a) employs computational linguistic models to measure learner motivation and cognitive engagement from the text of forum posts and observe that participation in discussion forums is a strong indicator of student commitment.

Even though discussion forum serves as a rich source of information that offers insights into many aspects of student behavior, it has been observed that a very small percentage of students (5-10%) actually participate in the discussion forum. As an alternate data trace of student interaction with the course material, the clickstream

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 55–59,
October 25-29, 2014, Doha, Qatar.

data of users contains a wider range of information affording other perspectives of student behavior. This is the theme of studies such as (Guo and Reinecke, 2014), which is focused on the navigation behavior of various demographic groups, (Kizilcec et al., 2013) which seeks to understand how students engage with the course, (Ramesh et al., 2014), that attempts to understand student disengagement and their learning patterns towards minimizing dropout rate and (Stephens-Martinez et al., 2014) which seeks to model motivations of users by mining clickstream data.

In this study, the task is to predict if a user will stay in the course or drop out using information from forum posts and clickstream information. Our approach is to use only clickstream information and is motivated by key insights such as interaction with the various course components and quiz attempt/submission.

3 Data

Data from one MOOC with approximately 30K students was distributed as training data. This included discussion post information and clickstream information of the students with completely anonymized user ids. Of this a subset of 6583 users was considered the held-out dataset on which we report the performance of the model.

3.1 Preprocessing Stage

Since participants (posters) in the discussion forum constitute a very small minority of the users in a course (between 5-10% as observed in prior studies), we mine the clickstream information for course-interaction. From the clickstream we extract the following information to indicate involvement in the course.

- Total watch time: From the video view information the amount of time watched is calculated by taking the summation of the difference between the time of the last event a user interacts with a video and the initial time a user starts the same video. If a user is idle for longer than 50 minutes, we add the difference between the current time before the user goes idle and the time the user initially interacts with the video to the total time. The new initial time will be after the user goes active again. Then we repeat the process until there is no more viewing action in the clickstream for that user.
- Number of quiz attempts;
- Number of quiz submissions;
- Number of times a user visits the discussion forum;
- Number of times a user posts: The number of times a user posts in a forum is counted. This count includes whether the user starts a thread, posts, or comments.
- Action sequence: We define an *action sequence* of a given user as being the sequence of course-related activity in a given week for a given user. It captures the user's interaction with the various components of a course in chronological order, such as seeking information on the course-wiki, watching a lecture video, posting in the discussion forum. The activities are, p = forum post, a = quiz attempt, s = quiz submit, l = lecture page view, d = lecture download, f = forum view, w = wiki page visited, t = learning tool page visited, o = play video. As an example, the action sequence of a user **wwaaws** in a given week indicates that the user began the course-activity with a visit to the course wiki, followed by another visit to the wiki, then attempted the quiz two successive times and finally submitted the quiz.

Each of the items listed above, captures important aspects of interaction with the course serving as an index of attrition; the more a user interacts with the course in a given week, the less the chances are of dropping out in that week.

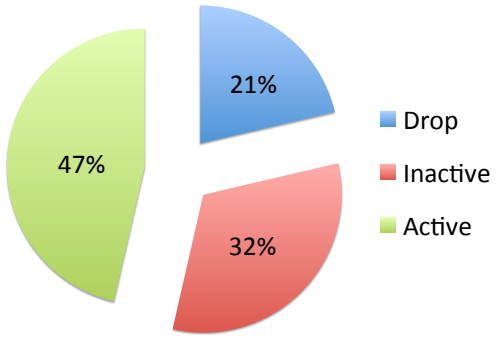

Figure 1: Percentage of each type of users

An exploratory analysis of the data reveals that there are three classes of users based on their interaction with the course components as revealed by the clickstream activity. More specifically, with respect to the length of their action sequence, the 3 classes are:

1. **Active**: This class is the majority class represented by 47% of the users in the course. The users actively interact with more than one component of the course and their enrollment status shows that they did not drop.

2. **Drop**: This is the class represented by a relative minority of the users (21%). The users hardly interact with the course and from their enrollment status they have dropped.

3. **Inactive**: This class of students, represented by 32% of the course, shares commonalities with the first two classes. Whereas their enrollment status indicates that they have not dropped (similar to the **Active** group), their clickstream information shows that their level of course activity is similar to that of the **Drop** class (as evidenced by the length of their action sequence. We define a user to be *inactive* if the action sequence is less than 2 and the user is still enrolled in the course.

The distribution of the three classes of users in the training data is shown in Figure 1. This key observation of the presence of three classes of users prompts us to consider three models to predict user attrition on any given week since we only predict whether a user dropped or not.

1. Mode 1 (Mod1): Inactive users are modeled to be users that dropped because of their similar activity pattern;

2. Mode 2 (Mod2): Inactive users are modeled as **Active** users because they did not formally drop out;

3. Mode 3 (Mod3): Inactive users are modeled as **Drop** with a probability of 0.5 and **Active** with a probability of 0.5. This is because they share status attributes with **Active** and interaction attributes with **Drop**.

4 Features

We use two classes of features to represent user-behavior in a course and summarize them as follows.

- Quiz related: The features in this class are: whether a user submitted the quiz (binary), whether a user attempted the quiz (binary), whether a user attempted but did not submit the quiz (binary). The intuition behind this set of features is that in general quiz-related activity denotes a more committed student with a higher level of involvement with the course. This set is also intended to capture three levels of commitment, ranging from only an attempt at the lowest level, attempting but not submitting at a medium level, to submitting the quiz being the highest level.

- Activity related: The features in this category are derived from the action sequence of the user during that week and they are:

  1. Length of the action sequence (numeric);
  2. The number of times each activity (p, a, s, l, d, f, w, o, or t) occurred (numeric);
  3. The number of wiki page visits/length of the action sequence (numeric).

  The features essentially capture the degree of involvement as a whole and the extent of interaction with each component.

5 Experiments

5.1 Models

We consider two input data distributions of the training data: a) a **specific** case, where the inactive users are excluded. In this case, the model is trained only on users that are either active or those that have dropped. b) a **general** case, where the inactive users are included as is. In both cases, the testing data has the inactive users included, but are either modeled as Mode 1, 2 or 3. This results in 6 models {specific, general} x {Mode1, Mode2, Mode3}.

We train an SVM for each model and observe that an *rbf* kernel achieves the best accuracy among the kernel choices. We use the scikit implementation of SVM (Pedregosa et al., 2011). The parameter γ was tuned to maximize accuracy via 5 fold cross validation on the entire training set. We observe that the performance of Mode 3 was much lower than that of Modes 1 and 2 and thus exclude it from the results.

The tuned models were finally evaluated for accuracy, precision, recall, F-measure and Cohen's

κ on the held-out dataset.

5.2 Experimental Results

	Mode 1		Mode 2	
	Specific	General	Specific	General
Baseline	46.42%	46.42%	78.66%	78.66%
Accuracy	**91.31**%	85.34%	78.48%	**78.56**%

Table 1: Accuracy of the models after parameter tuning.

We compare the accuracy of the tuned models with a simple baseline which classifies a user, who, during a given week, submits the quiz and has an action sequence length more than 1 as one who will not drop. The baseline accuracy is 46.42% for Mode 1 and 78.66% for Mode 2. We observe that modeling the inactive user as one who drops performs significantly better than the baseline, whereas modeling the inactive user as one who stays, does not improve the baseline. This is summarized in Table 1.

Of these models we chose two of the best performing models and evaluate them on the held-out data. The chosen models were: Model 1 = (specific,Mode1) and Model 2 = (general,Mode2). The resulting tuned Model 1 (inactive = drop) had $\gamma = 0.1$ and Model 2 (inactive = stay) had a $\gamma = 0.3$ and C as the default value.

	Model 1	Model 2
Accuracy	50.98%	80.40%
Cohen's Kappa	-0.06	0.065
P	0.167	0.482
R	0.371	0.058
F	0.228	0.104

Table 2: Accuracy, Cohen's kappa, Precision (P), Recall (R) and F-measure (F) scores for the models on the held-out data.

The performance (accuracy, Cohen's κ, precision, recall and F-measure scores of the two models on the held-out data are shown in Table 2. The final model submitted for evaluation on the test set is Model 2. It was more general since its training data included the inactive users as well. However, the skew in the data distribution is even larger for this model.

We highlight some important observations based on the result.

- Model 2, which is trained to be more general and has the inactive users included, but operates in Mode 2 (regards inactive users as active) has a better accuracy compared to Model 1, which is trained by excluding the inactive users, but operates in Mode 1 (regards inactive users as drop).
- In terms of the κ score, Model 2 shows some agreement, but Model 1 shows no agreement.
- The increased accuracy of Model 2 comes at the expense of reduced recall. This suggests that Model 2 has more false negatives compared to Model 1 on the held-out set.
- Even with reduced recall, Model 2 is more precise than Model 1. This implies that Model 1 tends to infer a larger fraction of false positives compared to Model 2.

6 Discussion

6.1 Data Imbalance

The impact of class imbalance on the SVM classifier is well-known to result in the majority class being well represented compared to the minority class (Longadge and Dongre, 2013). In our modeling with different input data distributions as in the *specific* case (Model 1), where we exclude inactive users, the data imbalance could have significantly affected the performance. This is because, the class of active users is more than double the size of the class of users who dropped.

Our attempt to counter the effect of the minority class by oversampling, resulted in no improvement in performance. In future explorations, other efforts to counter the data imbalance may be helpful.

6.2 Parameter tuning

The models studied here were tuned to maximize accuracy. In the future, models that are tuned to maximize Cohen's κ may be worth exploring.

6.3 Ablation Analysis

	Quiz Related	Activity Related
Model 1	80.48%	50.95%
Model 2	80.48%	80.41%

Table 3: Accuracy and kappa scores for the models by removing the corresponding set of features.

Table 3 summarizes the results of the ablation study conducted for each model by removing each class of features. For **Model 1**, the activity-related features constitute the most important set of features as seen by the drop in accuracy resulting from its omission. For **Model 2**, however, both sets of features have nearly the same effect.

References

Philip J. Guo and Katharina Reinecke. 2014. Demographic differences in how students navigate through moocs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 21–30, New York, NY, USA. ACM.

René F. Kizilcec, Chris Piech, and Emily Schneider. 2013. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, pages 170–179, New York, NY, USA. ACM.

R. Longadge and S. Dongre. 2013. Class Imbalance Problem in Data Mining Review. *ArXiv e-prints*, May.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. 2014. Uncovering hidden engagement patterns for predicting learner performance in moocs. In *ACM Conference on Learning at Scale*, Annual Conference Series. ACM, ACM Press.

Carolyn Rose and George Siemens. 2014. Shared task on prediction of dropout over time in massively open online courses. In *Proceedings of the 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*.

Carolyn Penstein Rosé, Ryan Carlson, Diyi Yang, Miaomiao Wen, Lauren Resnick, Pam Goldman, and Jennifer Sherer. 2014. Social factors that contribute to attrition in moocs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 197–198, New York, NY, USA. ACM.

Kristin Stephens-Martinez, Marti A. Hearst, and Armando Fox. 2014. Monitoring moocs: Which information sources do instructors value? In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 79–88, New York, NY, USA. ACM.

Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. 2014a. Linguistic reflections of student engagement in massive open online courses. In *ICWSM*.

Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. 2014b. Sentiment analysis in mooc discussion forums: What does it tell us? In *the 7th International Conference on Educational Data Mining*.