Predicting MOOC Dropout over Weeks Using Machine Learning Methods

Marius Kloft, Felix Stiehler, Zhilin Zheng, Niels Pinkwart
Department of Computer Science
Humboldt University of Berlin
Berlin, Germany
{kloft, felix.stiehler, zhilin.zheng, pinkwart}@hu-berlin.de

Abstract

With high dropout rates as observed in many current larger-scale online courses, mechanisms that are able to predict student dropout become increasingly important. While this problem is partially solved for students that are active in online forums, this is not yet the case for the more general student population. In this paper, we present an approach that works on click-stream data. Among other features, the machine learning algorithm takes the weekly history of student data into account and thus is able to notice changes in student behavior over time. In the later phases of a course (i.e., once such history data is available), this approach is able to predict dropout significantly better than baseline methods.

1 Introduction

In the past few years, with their dramatically increasing popularity, Massive Open Online Courses (MOOCs) have become a way of online learning used across the world by millions of people. As a result of efforts conducted (sometimes jointly) by academia and industry, many MOOC providers (such as Coursera, Udacity, Edx, or iversity) have emerged, which are able to deliver well-designed online courses to learners. In typical MOOC platforms, learners can not only access lecture videos, assignments and examinations, but can also use collaborative learning features such as online discussion forums. Despite all the MOOC features and benefits, however, one of the critical issues related to MOOCs is their high dropout rate, which puts the efficacy of the learning technology into question. According to the online data provided by Jordan (2014), most MOOCs have completion rates of less than 13%. While discussions are still ongoing as to whether these numbers are actually a problem indicating partial MOOC failures or whether they merely indicate that the community of MOOC learners is diverse and by far not every participant intends to complete a course, researchers and MOOC providers are certainly interested in methods for increasing completion rates. The analysis of MOOC data can be of help here. For instance, a linguistic analysis of the MOOC forum data can discover valuable indicators for predicting dropout of students (Wen et al., 2014). However, only few MOOC students (roughly 5-10%) use the discussion forums (Rose and Siemens, 2014), so that dropout predictors for the remaining 90% would be desirable. In order to get insights into the learning behaviors of this majority of participants, the clickstream data of the MOOC platform usage is the primary source for analysis in addition to the forum data. That is also the motivation of the shared task proposed by the MOOC workshop at the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014) (Rose and Siemens, 2014). Addressing this task, we propose a machine learning method based on support vector machines for predicting dropout between MOOC course weeks in this paper.

The rest of this paper is organized as follows. We begin with the description of the data set and features extracted from the data set. We then describe our prediction model. Next, the prediction results and some experimental findings are presented. Finally, we conclude our work in this paper.

2 Dataset

The dataset we used in this paper was prepared for the shared task launched by the Modeling Large Scale Social Interaction in Massively Open Online Courses Workshop at the Conference on Empirical Methods in Natural Language Processing

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 60–65,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

(EMNLP 2014) (Rose and Siemens, 2014). The data was collected from a psychology MOOC course which was launched in March 2013. The whole course lasted for 12 weeks with 11,607 participants in the beginning week and 3,861 participants staying until the last course week. Overall, 20,828 students participated, with approximately 81.4% lost at last. Note that the data cover the whole life cycle of this online course up to 19 weeks. The original dataset for this task had two types of data: clickstream data and forum data. In this paper, we only make use of clickstream data to train our prediction model and we do not further consider forum data. Obviously, this will lower the prediction quality for the 5% of students that use the forum, but it will hopefully shed light on the utility of the clickstream data for the larger set of all participants. The clickstream data includes 3,475,485 web log records which can be generally classified into two types: the page view log and the lecture video log. In the following section, we will describe attributes extracted from the raw clickstream data which (we believed) could be correlated to drop-out over the 12 course weeks.

2.1 Attributes description

Our model is an attempt to predict the participants' drop-out during the next week (defined as no activity in that week and in any future week) using the data of the current and past weeks. Consequently, all attributes are computed for each participant and for each week. Note that this results in having more data for later course weeks, since the approach allows for comparing a student's current activity with the activity of that student in the past weeks. The complete attributes list is shown in Table 1.

2.2 Attribute Generation

The attributes required for the predictions are extracted by parsing the clickstream file where each line represents a web request. For each line the corresponding Coursera ID is taken from the database containing the forum data and the course week is calculated from the timestamp relative to the start date of the course. Then the request is analysed regarding its type and every present attribute is saved.

After collecting the raw attributes, the data needs to be post-processed. There are 3 kinds of attributes: attributes that need to be summed up, attributes that need to be averaged and attributes

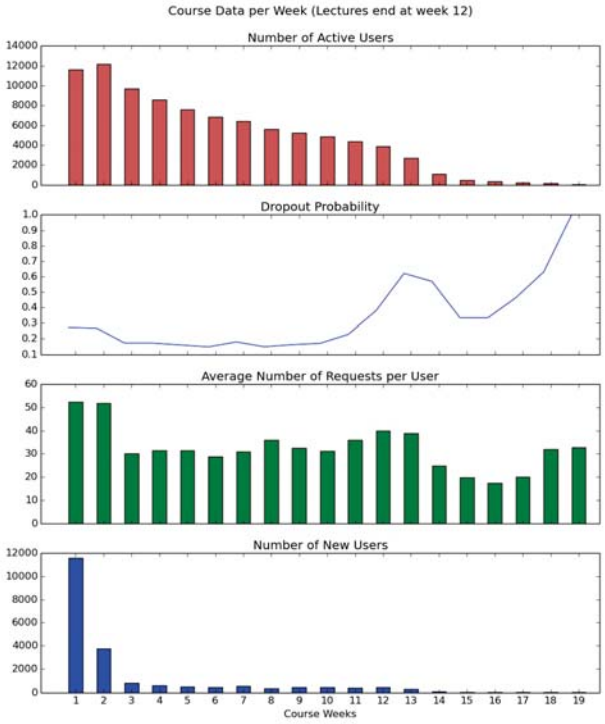

Figure 1: Several basic properties of the analyzed data set.

that need to be decided by majority vote. After the post-processing the data consists of lists of attributes each correlated to a unique tuple consisting of the Coursera ID and the course week number. Invalid attributes are getting replaced with the median of that week. Note that every missing week is getting replaced by the median of the attributes of active users in that week that were also active in the original week.

2.3 A First Glance on the Data Set

We have visualized several basic properties of the data in Figure 1. We observe that the number of active user quickly decreases over time. Furthermore the dropout probability is especially high in the first two weeks, and then of course at the end of the course starting around week 11 and 12.

3 Methodology & Results

In this section we concisely describe the employed feature extraction and selection pipeline, as well as the employed machine learning algorithms. For each week of the course ($i = 1, \ldots, 19$) we computed the dropout label of each of the n_i participants (user ids) being active in that week, based on checking whether there is any activity associated to the same user id in proceeding next week.

ID	Attributes
1	**Number of requests**: total number of requests including page views and video click actions
2	**Number of sessions:** number of sessions is supposed to be a reflection of high engagement, because more sessions indicate more often logging into the learning platform
3	**Number of active days:** we define a day as an active day if the student had at least one session on that day
4	**Number of page views:** the page views include lecture pages, wiki pages, homework pages and forum pages
5	**Number of page views per session:** the average number of pages viewed by each participant per session
6	**Number of video views:** total number of video click actions
7	**Number of video views per session:** average number of video click actions per session
8	**Number of forum views:** number of course discussion forum views
9	**Number of wiki views:** number of course wiki page views
10	**Number of homework page views**
11	**Number of straight-through video plays:** this is a video action attribute. Straight-trough playing video means that the participates played video without any jump (e.g. pause, resume, jump backward and jump forward). Since the lecture videos are the most important learning resource for the learning participants, the video playing should be investigated as other researchers did (Brotherton and Abowd, 2004). In this paper, five video behaviors are taken into account including the number of full plays as well as four others: start-stop during video plays, skip-ahead during video plays, relisten during video plays and the use of low play rate
12	**Number of start-stop during video plays:** start-stop during video plays stands for a lecture video being paused and resumed
13	**Number of skip-ahead during video plays:** skip-ahead means that the participant played a video with a forward jump
14	**Number of relisten during video plays:** relisten means that a backward jump was made as the participant was playing a video
15	**Number of slow play rate use:** this attribute is considered as an indicator of weak understanding of the lecturer's lecture presentation, possibly because of language difficulties or a lack of relevant background knowledge
16	**Most common request time:** our attempt with this attribute is to separate day time learning from night time learning. We define night time from 19:00 to 6:59 in the morning and the other half day as day time
17	**Number of requests from outside of Coursera:** this is to discover how many requests from third-party tools (such as e-mail clients and social networks) to the course were made, which could be an indicator of the participant's social behavior
18	**Number of screen pixels:** the screen pixels is an indicator of the device that the student used. Typically, mobile devices come with fewer pixels
19	**Most active day:** through this attribute, we can investigate if starting late or early could have an impact on dropout
20	**Country:** this information could reflect geographical differences in learning across the world
21	**Operating System**
22	**Browser**

Table 1: Attributes list.

This resulted in label vectors $y_i \in \{-1, 1\}^{n_i}$ for $i = 1, \ldots, 19$, where $+1$ indicates dropout (and thus -1 indicates no dropout). We experimented on the 22 numerical features described in the pre-

vious section. The features with ids 1–19 could be represented a single real number, while all other features had to be embedded into a multidimensional space. For simplicity we thus first focused on features 1–19. For each week i of the course, this results in a matrix $X_i^{\text{preliminary}} \in \mathbb{R}^{19 \times n_i}$, the rows and columns of which correspond to the features and user ids, respectively. We then enriched the matrices by considering also the "history" of the features, that is, for the data of week i, all the features of the previous weeks were appended (as additional rows) to the actual data matrix, resulting in $X_i \in \mathbb{R}^{19i \times n_i}$. We can write this as $X_i = (x_1, \ldots, x_{n_i})$, where x_j is the feature vector of the jth user. Box plots of these features showed that the distribution is highly skewed and non-normal, and furthermore all features are non-negative. We thus tried two standard features transformations: 1. logarithmic transformation 2. box-cox transformation. Subsequent box plots indicated that both lead to fairly non-skewed distributions. The logarithmic transformation is however much faster and lead to better results in later pipeline steps, which is why it was taken for the remaining experiments.

Subsequently, all features were centered and normalized to unit standard deviation. We then performed simple t-tests for each feature and computed also the Fisher score $f_j = \sqrt{\frac{\mu_+ - \mu_-}{\sigma_+^2 + \sigma_-^2}}$, where $\mu_\pm$ and $\sigma_\pm^2$ are the mean and variance of the positive (dropout) and negative class, respectively. Both t-tests and Fisher scores lead to comparable results; however, we have made superior experiences with the Fisher score, which is why we focus on this approach in the following methodology. We found that the video features (id 11–15), the most common request time (id 17), and the most active day feature (id 19) consistently achieved scores very close to zero, which is why they were discarded. The remaining features are shown in Figure 2 (a similar plot was generated using t-tests and found to be consistent with the Fisher scores, but is omitted due to space constraints). The results indicate that features related to a more balanced behaviour pattern over the course of a week (especially the number of sessions and number of active days) were (weakly) predictive of dropout in the beginning of the course. From week 6 to 12 we could also measure a rising importance of the number of wiki page views (id 9) and homework submission page views (id 10). Past week 12 features related to activity in a more general way like the number of requests (id 1) or the number of page views (id 4) became the most predicative.

We proceeded with an exploratory analysis, where we performed a principal component analysis (PCA) for each week, the result is shown in Figure 3. The plot indicates that the users that have dropped out can be better separated from the users that did not drop out when the week id increases. To follow up on this we trained, for each week, a linear support vector machine (SVM) (Cortes and Vapnik, 1995) using the `-s 2` option in LIBLINEAR (Fan et al., 2008), which is one of the fastest solvers to train linear SVMs (Fan et al., 2008). The SVM computes an affine-linear prediction function $f(x) := \langle w, x \rangle + b$, based on maximizing the (soft) margin between positive and negative examples: $(w, b) := \text{argmin}_{w,b} \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(\langle w, x_i \rangle + b)$. Note that this is very similar to regularized logistic regression, which uses the term $1/(1+\exp(-y_i(\langle w, x_i \rangle + b)))$ instead of $\max(0, 1 - y_i(\langle w, x_i \rangle + b)$, but with additional sparsity properties (only a subset of data points are active in the final solution) that make it more robust to outliers. The prediction accuracy was estimated via 5-fold cross validation. The regularization parameter was found to have little influence on the prediction accuracy, which is why it was set to the default value $C = 1$. We compared our SVM to the trivial baseline of a classifier that constantly predicts either -1 or 1; if the dropout probability in week i is denoted by p_i, then the classification accuracy of such a classifier is given by $\text{acc}_{trivial} := \max(p_i, 1 - p_i)$. The result of this experiment is shown in Figure 4. Note that we found it beneficial to use the "history" features, that is the information about the previous weeks only within the weeks 1–12. For the weeks 13–19 we switched the history features off (also the PCA above is computed without the history features). We observe from the figure that for weeks 1–8 we can not predict the dropout well, while then the prediction accuracy steadily increases. Our hypothesis here is that this could result from the more and more history features being available for the later weeks.

4 Conclusion

We proposed a machine learning framework for the prediction of dropout in Massive Open Online Courses solely from clickstream data. At the

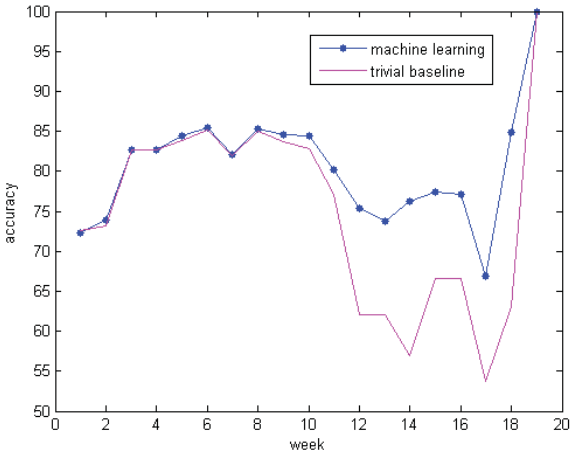

Figure 4: SVM classification accuracies per week. The baseline accuracy is computed as $\max(p_i, 1 - p_i)$, where p_i denotes the weekwise dropout probability.

heart of our approach lies the extraction of numerical features capturing the activity level of users (e.g., number of requests) as well technical features (e.g., number of screen pixels in the employed device/computer). We detected significant signals in the data and achieved an increase in prediction accuracy up to 15% for some weeks of the course. We found the prediction is better at the end of the course, while at the beginning we still detect rather weak signals. While this paper focuses on clickstream data, the approach could in principle also combined with forum data (e.g., using multiple kernel learning (Kloft et al., 2011)), which we would like to tackle in future work. Furthermore, another interesting direction is to explore non-scalar features (e.g., country, OS, browser, etc.) and non-linear support vector machines.

References

Katy Jordan. *MOOC Completion Rates: The Data*. Availabe at: http://www.katyjordan.com/MOOCproject.html. [Accessed: 27/08/2014].

Miaomiao Wen, Diyi Yang and Carolyn P. Rose. *Linguistic Reflections of Student Engagement in Massive Open Online Courses*. ICWSM'14, 2014.

Carolyn Rose and George Siemens. *Shared Task on Prediction of Dropout Over Time in Massively Open Online Courses*. Proceedings of the 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses, Qatar, October 2014.

Jason A. Brotherton and Gregory D. Abowd. *Lessons learned from eClass: Assessing automated capture and access in the classroom*. ACM Transactions on Computer-Human Interaction, Vol. 11, No. 2, pp. 121–155, June 2004.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research (JMLR)*, 9:1871–1874, 2008.

M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011.

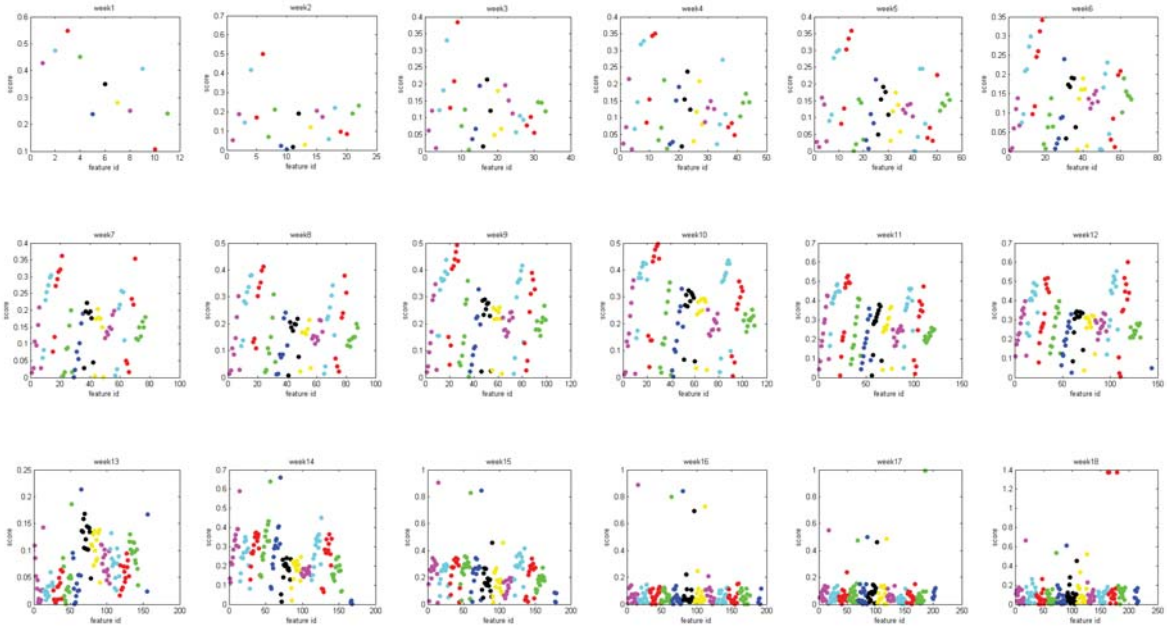

Figure 2: Fisher scores indicate which features are predictive of the dropout. Features are ordered from left to right with increasing ids; i.e., pink indicates the number of requests (feature id 1), cyan the number of sessions (feature id 2), etc. In particular, we observe that features related to a more balanced behaviour pattern such as the number of active days (feature id 3) are the most important ones in the first couple of weeks while more general features like the number of requests rise in importance past week 12.

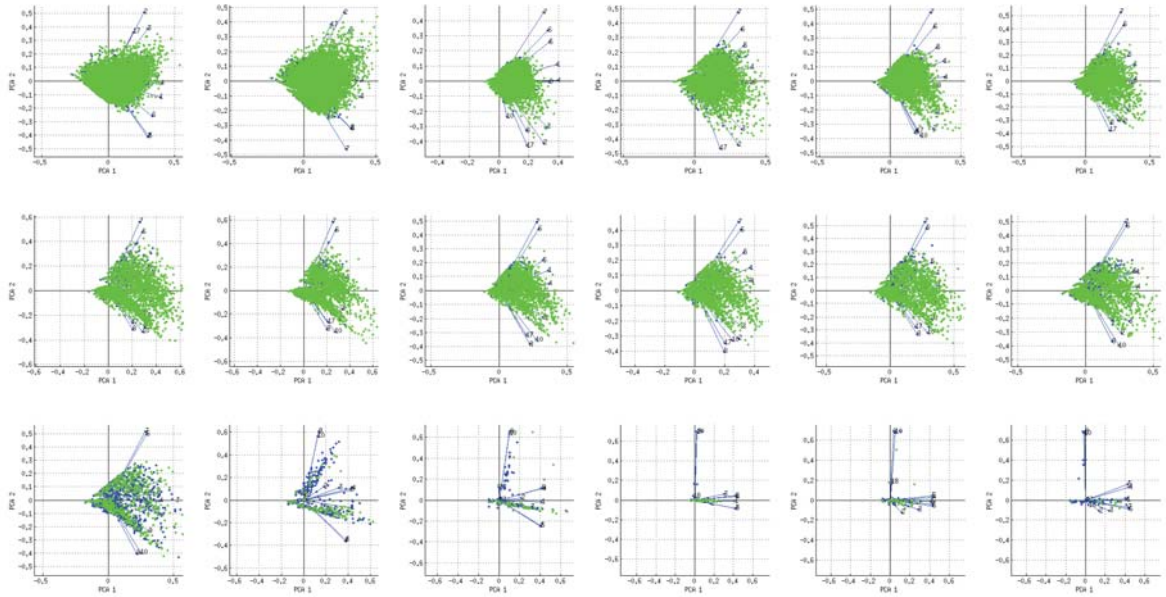

Figure 3: Result of principal component analysis. The data becomes more non-isotropic within the later weeks (from week 13), and can also be separated better.