A MOOCs Dropout Rate Prediction Method Based on Sliding Window Model (Postprint)

Authors: Lu Xiaohang, Wang Shengqing, Junjie Huang, Chen Wenguang, Yan Zengwang

Date: 2017-11-08T00:00:00+00:00

Abstract

【Objective】Using course data from Peking University's programs on the Coursera platform, this study investigates student dropout behavior to predict dropout points and behaviors, aiming to improve MOOC teaching quality and methods. 【Method】Based on the course data, 19 features were extracted, and a sliding window model was constructed using machine learning algorithms to dynamically predict learner dropout rates. 【Results】The model achieves high prediction accuracy, generally above 90%, with stable performance. Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) network methods yield better modeling results. 【Limitations】The course data features relatively high enrollment numbers, does not consider data sparsity issues in other courses, and the model' s portability requires further investigation. 【Conclusion】Using sliding window models can help MOOC instructors and designers dynamically track learner dropout behavior with high accuracy, enabling instructors to adjust courses through rapid feedback and reduce dropout rates.

Full Text

A Dropout Prediction Method for MOOCs Based on Sliding Window Model

Lu Xiaohang[1], Wang Shengqing[2], Huang Junjie[1], Chen Wenguang[1], Yan Zengwang[1]

[1](Department of Information Management, Peking University, Beijing 100871, China)

[2](Center of Faculty Development, Peking University, Beijing 100871, China)

Abstract

[**Objective**] This study investigates student dropout behavior using data from Peking University's courses on the Coursera platform, aiming to predict dropout points and behaviors to improve MOOC quality and teaching methods. [**Methods**] Based on course data, we extracted 19 features and constructed a sliding window model using machine learning algorithms to dynamically predict learner dropout rates. [**Results**] The model achieved high prediction accuracy, generally above 90%, with stable performance. Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) methods demonstrated superior modeling effectiveness. [**Limitations**] The course data had relatively large enrollment numbers, and the model' s applicability to courses with sparse data was not considered; the model' s portability requires further investigation. [**Conclusions**] The sliding window model enables MOOC instructors and designers to dynamically track learner dropout behavior with high accuracy, facilitating rapid feedback for course adjustments and dropout rate reduction.

Keywords: MOOC; Dropout Point; Dropout Rate; Sliding Window Model; Dropout Prediction

---

Since their emergence in the United States in 2011, MOOCs have experienced significant annual growth in both course offerings and user enrollment worldwide, exerting expanding influence and impact on higher education ecosystems, university teaching methods, management systems, and professional training. Numerous universities and institutions domestically and internationally have continuously experimented with and implemented MOOC initiatives. Typically, each MOOC includes modules such as lecture videos, discussion forums, wikis, quizzes, assignments, and examinations. As online teaching platforms, MOOCs record and preserve the most original teaching and learning data for each course, providing great convenience for data-driven learning behavior analysis research and attracting numerous researchers globally to conduct MOOC-related data analysis studies. Simultaneously, as an emerging and continuously evolving field, MOOCs involve many research questions, with learner behavior analysis representing a crucial research direction. Despite their advantages of openness and free access, MOOCs lack face-to-face interaction between teachers and students, creating many problems absent in traditional teaching, most notably the extremely high dropout rates. According to relevant research data, the completion rate for most MOOC courses domestically and internationally is less than 13% [1]. How to improve course pass rates, predict and intervene in dropout behavior, analyze dropout reasons, and enhance course quality and online teaching methods are issues of great concern to MOOC instructors and designers.

This paper uses data from multiple MOOC courses offered by Peking University across three semesters—Fall 2013, Spring 2014, and Fall 2014—on the Coursera platform as the sample dataset. Focusing on learner dropout issues, we conduct statistical analysis on the characteristics of dropout timing and start timing, and propose a sliding window model to dynamically predict the overall dropout rate of course students, helping instructors improve course quality, communicate with potential dropouts in a timely manner, provide assistance and feedback, and ultimately increase course completion rates.

Corresponding Author: Wang Shengqing, ORCID: 0000-0002-7164-073X, E-mail: wangsq@pku.edu.cn.

This research is supported by the Key Project of the Online Education Research Center of the Ministry of Education (All-Tong Education) "Empirical Research on Online Teaching Organization Methods for MOOCs" (Project No.: 2016ZD301).

2 Related Research on MOOC Dropout Rates

Many scholars domestically and internationally have studied when MOOC learners leave courses and drop out. Existing research analysis data can be divided into two major categories: forum data analysis and clickstream data analysis. This paper selects several typical dropout rate studies for analysis. Overall, although numerous studies exist on dropout rates, no standardized, widely accepted research methodology has emerged, with most research remaining exploratory, using different models and methods to improve prediction accuracy.

Amnueypornsakul et al. [1] used learners'clickstream data to predict whether students would drop out. Each learner's weekly learning behaviors were formed into sequences, such as representing a learner' s complete behavior in a course: viewing course wiki, viewing course wiki, taking tests, taking tests, viewing course wiki, submitting tests. Three types of learners were defined: active, abandoned (i.e., no learning behavior), and inactive (learning behavior sequence with fewer than 2 elements). When defining "dropout," three scenarios were considered: classifying inactive learners as dropouts, classifying inactive learners as active learners, and classifying inactive learners as dropouts with 0.5 probability and as active learners with 0.5 probability. When building models using SVM, two situations were also considered: excluding inactive users for prediction model construction, and including inactive users for prediction model construction. In total, six models were constructed. The baseline for each model was predicting dropout in the next week when a learner' s weekly learning behavior sequence was less than or equal to 1. Results showed that excluding inactive users significantly improved accuracy; if including inactive users, defining inactive learners as not belonging to dropouts yielded higher accuracy, but still lower than [2].

Sinha et al. [2] used video click and forum data to construct learners' activity sequences, seeking footprint sequences that could represent students' positive or negative participation in courses. They first constructed learners'weekly activity sequences, from which n-gram sequences, video viewing activity sequences, and forum interaction sequences were extracted to explore which sequences could predict learner dropout and which could keep learners enthusiastically engaged in MOOC learning. Experiments were conducted from two perspectives: how behavior within one week affects dropout in the next week, and how accumulated learning behavior since the first week affects dropout in the next week. Additionally, most dropout students started attending classes a few weeks after the course began, with sparser activity graphs. Two possible explanations exist: these dropout students had specific information needs and stopped attending after obtaining the required information; or later joiners dropped out due to difficulty catching up with previous materials and assignments.

Taylor et al. [3] conducted extensive experiments using different machine learning methods for dropout prediction, including logistic regression, support vector machines, deep belief networks, decision trees, and hidden Markov models. Dropout was defined as learners no longer submitting any assignments or tests, and 14 weeks of learners' learning behavior data were selected for training and testing. Learners were divided into four categories: passive participation, wiki editing participation, forum editing participation, and active participation (both wiki and forum editing), with separate modeling for each category. The researchers proposed a lead-and-lag prediction pattern, i.e., given week i, using data from the first i weeks to predict the remaining 14–i weeks. The study explored the weights of various features and which features could predict at the beginning of learning whether learners would persist until the end of the course. Results showed that prediction accuracy was highest for the passive participation group, while due to insufficient data, prediction accuracy was lower for wiki editing and active participation groups. However, the wiki editing feature could well reflect whether learners would persist until the end of the course. With sufficient data, the prediction accuracy of various modeling methods showed little difference. Moreover, for predictions in a given week, the most recent four weeks of data were more predictive. Regarding predictive features, results showed that features familiar to MOOC participants could provide good predictive power, features related to assignment and test submissions had high predictive value, post length was more predictive than post count, and features related to collaborative social interaction such as wiki and forum were very important in prediction.

Additionally, some related studies are also enlightening. Kloft et al. [4] used clickstream data and machine learning algorithms to predict dropout behavior, and during prediction, tested the dropout predictive power of each feature vector; results showed poor prediction effect in the first 8 weeks, with weekly prediction effect improving afterward. The analysis attributed this to insufficient data in the first few weeks and suggested adding forum data for prediction. Sharkey et al. [5] detailed the iterative process of using machine learning techniques to predict dropout and identified predictive features and their relative weights through research. Results showed that machine learning models were above average in predicting dropout, and predictive factors also used social network analysis methods to analyze learners' behavior sequence graphs. Results showed that dropout students had fewer nodes, edges, strongly connected components, and self-loops, and dropout behavior was more affected by recent weeks' learning behavior. As expected, these were all variables that could show whether students were enthusiastically participating. Yang et al. [6] showed that social factors (forums) indeed affected dropout and provided insights for MOOC designers.

Through the above research, we find that student clickstream data analysis is currently the main direction of dropout rate research, and building more optimized data analysis models based on clickstream data is the key research direction of this paper.

3 Modeling and Experiments

We selected log data from 5 MOOC courses offered across three semesters—Fall 2013, Spring 2014, and Fall 2014—at Peking University as the analysis objects. Through research and analysis, we found that MOOC courses mainly follow two patterns: a single course offered repeatedly across multiple semesters with identical content each semester; or a single course divided into two semesters, where the first semester's course serves as the foundation for the second semester' s course. These patterns served as the principles and basis for course selection. Based on these two patterns, we selected Bioinformatics (offered for 3 semesters) and Social Survey and Research Methods (Part 1) and (Part 2) (offered for 2 semesters), totaling 5 courses. On one hand, these 5 courses satisfied these two course patterns; on the other hand, these 5 courses had many learners and large volumes of learning behavior data, facilitating research.

3.1 Data Sources

To gain a more intuitive understanding of MOOC dropout situations, this paper statistics the registration numbers, numbers with recorded grades, numbers with grades greater than 0, numbers with final grades greater than 60, and the pass rate for each of these 5 MOOC courses, as shown in Table 1 .

Table 2 shows the overlap rate statistics for learners with academic performance and forum behavior. Taking the course methodologysocial-001 as an example, the overlap rate between SET1 and SET2 is approximately 92.8%, meaning that about 92.8% of learners who ultimately passed the course participated in the course forum. Therefore, this paper considers that forum participation behavior has a significant predictive-indicator effect on whether learners will persist in learning, so student forum participation data is added to the prediction model.

3.3 Feature List

Since there are numerous courses with different start times, this paper defines each learner' s start time as the time when their first video click data appears. To determine the learning end time while ensuring sufficient data for prediction in the final weeks to maintain good prediction accuracy, we statistics the time of the last learning behavior of learners who obtained grades in these 5 courses and select the end time of the top 80% of learners who obtained grades as the course end time.

Through research and analysis, the feature data items extracted in this paper mainly include 19 items, as shown in Table 3 .

3.4 Learning Cycle

The learning cycle refers to the start and end time of learners' study. To unify standards, it is necessary to define learning start and end times. As shown in Figure 1 [Figure 1: see original paper], in the 2013 Social Survey and Research Methods (Part 1) course, among learners who obtained grades, the earliest end time for learning behavior was week 9, the latest was week 19, and 80% of learners had finished learning by week 14. Therefore, week 14 is set as the end time for the 2013 Social Survey and Research Methods (Part 1) course to ensure prediction accuracy at the end of the course. Other courses are similar.

3.5 Online Learner Numbers

Online learner numbers refer to the number of people with learning behavior each week. Taking these 5 courses as examples for statistics, the results are shown in Figure 2 [Figure 2: see original paper]. The number of learners in these 5 courses all shows a pattern of rapid increase in the early stage, followed by a sharp decline that transitions to a gradual decline. Moreover, the number of learners still studying in the final stage is far lower than in the initial peak, reflecting to some extent the universality of the MOOC dropout phenomenon.

3.6 Start Points, Dropout Points, and Dropout

We conduct more in-depth statistical analysis of each individual' s start time and dropout time. Since MOOCs allow learners to start and end learning at any time during the course, the number of weeks each person stays in a MOOC varies greatly. This paper defines whether a learner has learning behavior in a given week as 1 or 0 (with learning behavior defined as 1), allowing us to obtain a feature sequence of whether each learner appears each week. If a learner enters the course in week 3, their feature sequence is 0-0-1-……. The week when 1 first appears is defined as the start point; if a learner no longer appears after a certain week, i.e., the sequence is ···1-0-0···-0, the next week after that week is defined as the dropout point. We also consider that from the dropout point onward, the student has dropped out of the course, i.e., the learner no longer appears from the dropout point onward. The research focuses on observing the number of dropouts at each dropout point and the relationship between start points and dropout points.

By observing the weekly dropout numbers in these 5 courses, we find that two weeks after the course begins, the number of dropouts reaches a peak; near the end of the course, dropout numbers rise again; dropout numbers in the middle of the course are relatively stable, as shown in Figure 3 [Figure 3: see original paper]. The possible reasons are that after learning the course for a few weeks, learners gain some understanding of the course and decide whether to drop out based on whether the course suits them; at the end of the course, people who don' t care about the final exam or obtaining grades may choose to drop out, and some learners may drop out because they believe they cannot obtain a certificate. The stable dropout numbers in the middle stage may reflect the normal flow of learners in and out, and also indicate that learners who persist to the middle stage have a lower probability of dropping out.

Figure 4 [Figure 4: see original paper] examines the relationship between start points and dropout points. In Figure 4, the vertical axis represents the weekly dropout rate (i.e., the ratio of dropout numbers to the total number of people who started in a given week), and the horizontal axis represents time, i.e., the current week number. The lines from left to right represent starting in week *n*. By comparing commonalities in the charts, we find: the later the start time, the higher the dropout rate after one week of learning; the earlier the start time, the higher the rate of persistence to later weeks. This indicates that starting earlier makes it more likely to persist to the end, while starting later increases the likelihood of dropping out in the next week. This can be explained by referencing Sinha et al. [2]: one possibility is that later starters struggle to keep up with previous materials and course content, while another possibility is that later starters are more likely to come specifically seeking certain information and leave after obtaining it.

3.7 Construction of the Sliding Window Model

Based on the above analysis and definitions, we focus on discussing the construction of the sliding window model to predict whether learners will drop out. This model treats the entire learning cycle as a continuous sequence, using feature vectors from previous weeks to predict whether learners will participate in the course in future weeks. As shown in Figure 5 [Figure 5: see original paper], the sliding window model consists of two windows: the first window has length *w1*, i.e., the *w1* weeks before the current week. If the current week is week *n*, the window includes weeks *n-w1* to *n-1*. The second window has length *w2*, i.e., including the current week and the following *w2* weeks. If the current week is week *n*, the window includes week *n* to week *n+w2-1*. This model uses 19-dimensional feature vector clickstream data from a window of length *w1* to predict labels for the following *w2* weeks, with the label (Label) being the classification target—dropout or not dropout. Through window sliding, the model predicts whether learners will drop out for each week of the course. Therefore, dropout refers to whether learners have learning behavior during the *w2*-week period; absence of behavior is considered dropout. Since this model only focuses on learners' feature vectors within the current window without connecting all learning behaviors from beginning to end, it does not focus on individual learners' dropout and when they drop out, but only on the overall situation of whether students have learning behavior, i.e., how many people will drop out within the current window.

Since most learners in MOOC courses do not persist to the end, the baseline is defined as predicting that all learners will not appear in the course next week, and then comparing the improvement of each model relative to the baseline. Baseline prediction would cause many errors, but it is also the simplest and most direct prediction. This prediction method also has many applications in reality, such as instructors continuously sending mass emails to encourage learners to continue learning, even if the learner will not persist in the next week. Introducing machine learning methods can improve this strategy. For the prediction model, we adopt Logistic Regression (LR) [7], Support Vector Machine (SVM) [8], Multi-Layer Perceptron (MLP) [9], and Long Short-Term Memory (LSTM) [10] as classifiers.

We conducted sliding window model experiments on the aforementioned 5 courses. When selecting the value of *w1*, if *w1* is too short, e.g., *w1*=1 (predicting the next week using only the previous week), the model is too simple and rough; if *w1* is too long, it becomes overly redundant. Based on previous research, we selected *w1*=3, *w2*=1, and performed 5-fold cross-validation on the model. The data was divided into 5 parts, with 1 part selected as test data and the remaining 4 parts as training data. The experiment was repeated 5 times, and the average results are shown in Figure 6 [Figure 6: see original paper]. Through analysis, the following characteristics are observed:

(1) Baseline accuracy is generally high. This is because many people drop out each week, with dropout rates typically at 70% under normal circumstances, so baseline accuracy is high.

(2) Dropout rates peak at the beginning and end of courses. Many people leave early because learners try out courses at the beginning and abandon them if they are unsuitable; near the end of the course, many people also give up due to course pressure. Appropriate intervention and encouragement at these times may significantly reduce dropout rates.

(3) Generally, machine learning methods have good prediction effects. Among different machine learning methods, logistic regression represents the basic prediction capability of machine learning models. Compared with simply assuming learners cannot persist, machine learning can better identify learners who can persist. However, its prediction capability is limited and requires more feature data to achieve better results.

(4) LSTM and SVM perform better. Compared with multi-layer perceptron and logistic regression, these two methods have better prediction capabilities and stable performance, unaffected by data volume.

Further expanding the posterior window to predict performance over the next few weeks, as shown in Figure 7 [Figure 7: see original paper], we set *w2*=3. This results in 8 scenarios: 000, 001, 010, 100, 011, 101, 110, 111, corresponding to different classifications 0-7. For the posterior window scenarios, this paper uses Baseline1 to represent the prediction that no learning behavior will occur in the next 3 weeks, i.e., 000; Baseline2 represents the prediction that learning behavior will occur for 3 consecutive weeks, i.e., 111.

Experiments revealed the following characteristics: among learners' learning behavior in the next three weeks, the proportion of 000 is relatively high, the proportion of 111 is relatively low, and the proportion of 111 is lowest at both the beginning and end of the course. That is, in the beginning and end stages of the course, the proportion of learners who study continuously for multiple weeks is very low. Additionally, when predicting whether students will drop out over multiple subsequent weeks, SVM' s simplification to a multi-classification problem is effective because the proportion of 000 is relatively high. Although LSTM and SVM differ methodologically, both maintain high accuracy in their results.

The sliding window model initially solves the problem of monitoring and predicting different stages of courses, and finds that the beginning and end stages of MOOC courses are the most challenging stages for learners and likely to cause dropout. The reasons for dropping out at these stages are related to the characteristics of MOOCs themselves. Learners may initially just explore the course and choose to leave after discovering it is unsuitable, which explains the high early dropout rate. Therefore, if the course itself can stimulate learners' interest and maintain care and help for learners in the early stage, it will be beneficial for the course to enter a relatively stable period. At the end of the semester, due to final exams, many learners choose to leave the course. Therefore, if some strategies can be adopted at the end of the semester to encourage learners to complete the final assessment, such as review courses, it will help reduce dropout rates at the end of the semester and encourage them to ultimately obtain grades. Although this study only selected data from 5 courses, the course name is a variable in experiments and can be modified to any course. Therefore, applying this model to other MOOC courses for dropout rate prediction, such as University Chemistry, Introduction to Computing, Criminal Law, and People and Networks courses, still proves effective.

This paper starts with Peking University's MOOC course data, cleans it, and extracts 19 features based on previous research and data characteristics to study the course data. Based on analyzing changes in online learner numbers and characteristics of learner start points and dropout points, we construct a sliding window model that can effectively predict the overall student dropout situation in a MOOC course. We find that MOOC course dropout is severe, and learners who start later have a higher probability of dropping out. Results show that machine learning methods and the sliding window model have high accuracy in predicting student dropout, helping instructors track courses and predict dropout students, and grasp classroom progress. This model can help designers and instructors adjust courses through rapid feedback during the course to reduce dropout rates. Future research will focus on similarities and differences in dropout patterns across different courses, observing and analyzing similarities and differences in learner dropout motivations across courses, and predicting whether learners can obtain grades. Further experiments will be conducted on more data to improve the model, provide different feedback suggestions for different types of courses, assist MOOC instructors in better designing their courses, help teachers and students communicate more effectively and timely, and intervene with potential dropouts to help learners study more effectively, ultimately reducing MOOC dropout rates and improving online teaching quality and effectiveness.

References

[1] Amnueypornsakul B, Bhat S, Chinprutthiwong P. Predicting Attrition Along the Way: The UIUC Model [C]// Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, Doha, Qatar. Association for Computational Linguistics, 2014: 55-59.

[2] Sinha T, Jermann P, Li N, et al. Your Click Decides Your Fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions [OL]. arXiv Preprint. arXiv:1407.7131, 2014.

[3] Taylor C, Veeramachaneni K, O' Reilly U M. Likely to Stop? Predicting Stopout in Massive Open Online Courses [OL]. arXiv Preprint. arXiv: 1408.3382, 2014.

[4] Kloft M, Stiehler F, Zheng Z, et al. Predicting MOOC Dropout over Weeks Using Machine Learning Methods [C]// Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, Doha, Qatar. Association for Computational Linguistics, 2014.

[5] Sharkey M, Sanders R. A Process for Predicting MOOC Attrition [C]// Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, Doha, Qatar. Association for Computational Linguistics, 2014.

[6] Yang D, Sinha T, Adamson D, et al. "Turn on, Tune in, Drop out" : Anticipating Student Dropouts in Massive Open Online Courses [C]// Proceedings of the 2013 NIPS Data-driven Education Workshop. 2013: 11-14.

[7] Lipsitz S R. Categorical Data Analysis [J]. Statistics in Medicine, 1992, 13(11): 1791-1792.

[8] Cortes C, Vapnik V. Support-Vector Networks [J]. Machine Learning, 1995, 20(3): 273-297.

[9] Rosenblatt F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms [J]. American Journal of Psychology, 1962, 7(3): 218-219.

[10] Hochreiter S, Schmidhuber J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8): 1-32.

Author Contributions

Lu Xiaohang: Partial data statistics, partial manuscript writing, and final editing;
Wang Shengqing: Overall research design and final manuscript review;
Huang Junjie: Data processing, modeling, and experiments;
Chen Wenguang: Data analysis method design and optimization;
Yan Zengwang: Partial data statistics and partial manuscript writing.

Conflict of Interest Statement: All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors, E-mail: 1300016603@pku.edu.cn.

[1] Lu Xiaohang. methodologysocial-001.zip. methodologysocial-001 data.
[2] Lu Xiaohang. methodologysocial2-001.zip. methodologysocial2-001 data.
[3] Lu Xiaohang. pkubioinfo-001.zip. pkubioinfo-001 data.
[4] Lu Xiaohang. pkubioinfo-002.zip. pkubioinfo-002 data.
[5] Lu Xiaohang. pkubioinfo-003.zip. pkubioinfo-003 data.
[6] Lu Xiaohang. pkuic-001.zip. pkuic001 data.
[7] Lu Xiaohang. peopleandnetworks-001.zip. peopleandnetworks-001 data.
[8] Lu Xiaohang. chemistry-002.zip. chemistry-002 data.
[9] Lu Xiaohang. criminal law-001.zip. criminal law-001 data.

Received: 2017-02-27
Revised: 2017-04-06

Predicting Dropout Rates of MOOCs with Sliding Window Model

Lu Xiaohang[1], Wang Shengqing[2], Huang Junjie[1], Chen Wenguang[1], Yan Zengwang[1]
[1](Department of Information Management, Peking University, Beijing 100871, China)
[2](Center of Faculty Development, Peking University, Beijing 100871, China)

Abstract: [Objective] This paper aims to improve the MOOCs curriculum quality and pedagogy by analyzing the dropout behaviors with data from the MOOC of Peking University on Coursera. [Methods] We extracted 19 major features from the logs and then constructed a siding window model to predict the dropout rates. [Results] The precision of the proposed model was maintained above 90%. The SVM and LSTM methods further improved the performance of the proposed model. [Limitations] The new method needs to be examined with smaller sized courses. [Conclusions] Predicting dropout rates could help us improve the course quality effectively.

Keywords: MOOC; Dropout Point; Dropout Rates; Sliding Window Model; Dropout Prediction