# What Decides the Dropout in MOOCs?

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

[1] Department of Information Management, Peking University, Beijing 10086, China
chenwg@pku.edu.cn,
[2] Teaching Development Center, Peking University, Beijing 10086, China
wangsq@pku.edu.cn

**Abstract.** Based on the datasets from the MOOCs of Peking University running on the Coursera platform, we extract 19 major features of tune in after analyzing the log structure. To begin with, we focus on the characteristics of start and dropout point of learners through the statistics of their start time and dropout time. Then we construct two models. First, several approaches of machine learning are used to build a sliding window model for predicting the dropout probabilities in a certain course. Second, SVM is used to build the model for predicting whether a student can get a score at the end of the course. For instructors and designers of MOOCs, dynamically tracking the records of the dropouts could be helpful to improve the course quality in order to reduce the dropout rate.

**Keywords:** MOOCs, dropout rate, sliding window model, dropout prediction

## 1 Introduction

Big impacts and challenges on teaching methods, learning methods, student credit affairs management and so on in higher education are brought by MOOCs. MOOCs are also changing the lifelong professional training system gradually. In typical MOOCs platforms, not only could the learners access lecture videos, assignments and examinations, but they can also use collaborative learning approaches such as online forums[6]. A lot of universities and institutions at home and abroad are exploring and researching MOOCs in different ways. For example, the US Department of Education[2] has begun to focus on how to improve student performance using online learning systems. They use automated data mining and analysis techniques to monitor teaching quality, to test teaching improvement, to help teachers understand the performance of students. After that, teachers can help students adapt to teaching individualized. Therefore, the excavation and analysis of teaching data can lead to an important direction for the future development of education.

MOOCs, despite of being open, free and many other advantages, lacks of face to face interaction coupled with such high student teacher ratios[4], leading to new problems that traditional teaching hasn't encountered. The most serious problem is the surprisingly high dropout rate. According to a previous study, most MOOCs courses have a passing rate less than 13%[1]. Thus, the teachers and designers of MOOCs are deeply concerned about how to improve the MOOCs' passing rate and the course quality.

In this paper, a total of 32 MOOCs data of Peking University are collected from Coursera platform in Autumn 2013, Spring to Autumn 2014, as the basic datasets. To analyze the dropout problem of MOOCs learners, the characteristics of dropout time and start time are analyzed statistically. Besides, we propose a sliding window model to dynamically predict the overall dropout rate of the learners in the course and a model to predict whether a student can get a score at the end. It can help teachers improve the quality of the course, communicate with the potential dropout in time, provide timely help and feedback, and increase the passing rate.

## 2 Related Work

Many scholars at home and abroad have studied on when the learners of MOOCs will drop out of the course. There are two kinds of analysis data in the current research: the data of the forum and the data of the clickstream. Here, several typical dropout studies are analyzed and presented.

Amnueypornsakul et al.[1] used learners' clickstream data to predict whether or not a student would drop out of the course. Researchers formed a sequence of weekly learning behaviors for each learner. Then the researchers defined three learners: active, inactive, drop. The results showed that the accuracy rate was significantly improved when the inactive learners were excluded from the model construction, and when including inactive learners, the accuracy rate of modeling inactive learners as active learners was relative higher, but lower than the baseline.

Sinha et al.[12] leveraged combined data of video clickstream and forum to form the action sequence to seek traits that were predictive of decreasing engagement over time. The results showed that dropout behavior was more affected by learning behavior of recent weeks. And most of the dropout students started classes a few weeks after the beginning of the course. There were two possible explanations. One was that these dropout students have needs in specific information, and they ceased to attend classes after they met their needs. Or, the students who joined later had to give up due to the excessive material and work to keep up with the course.

Taylor et al.[13]used different machine learning approaches to predict the dropout students, including logistic regression, support vector machines, deep belief networks, decision trees, and hidden Markov models. The researchers defined dropout as a learner no longer submitting any assignments and tests, and screened 14 weeks' data of learning behavior as training and testing. Researchers divided learners into four categories: passive collaborator, wiki contributor, forum contributor, fully collaborator, and constructed the model for the four types of learners respectively. The researchers proposed a lead and lag prediction model, which predicted the rest $14 - i$ weeks using the previous week's data when given a week $i$. Therefore, there were a total of 91 forecast tasks. The results showed that the prediction accuracy of the passive collaborator was the highest, and the accuracy of wiki contributor and fully collaborator was low due to the lack of data. However, the feature of editing wiki was better to reflect whether a learner will persist to the end of the course. Unless the amount of data is insufficient, the prediction accuracy of the various methods of building the model is similar. Moreover, for a given week forecast, the data of the most recent four weeks was more predictive. In terms of predictive features, the results showed that those who were familiar with MOOCs can present better predictive features.

In addition, some other relevant studies also provide some inspiring ideas. Kloft et al.[6] used click stream data and machine learning algorithms to predict dropout behavior. In the prediction process, a predictive test was performed for each feature. The results showed that the prediction accuracy is not high in the first eight weeks due to the insufficiency of data, followed by weekly prediction accuracy increasing, and suggested to involve the forum data in prediction. Sharkey et al.[11]described in detail the iterative process of using machine learning techniques to predict dropouts and derive predictive features as well as their relative weights. The results showed that the prediction accuracy of the model using machine learning algorithms was above average, and the predictive features, as people expected, showed whether students were enthusiastic to participate in the course. Another study by Yang et al.[14] showed that social factors do have an effect on dropout, and gave MOOCs designers inspiration to design social engagement activities that can enhance students' participation to prevent dropout behavior.

## 3 Data Cleaning and Learning Metrics

In this paper, a total of 32 sessions-based MOOC courses data were collected from Coursera platform in Autumn 2013, Spring to Autumn 2014, as sample datasets. Through the preliminary analysis and screening of the course data, we find that:

1. the three semester courses of bioinformatics are independent to each other and the contents of the three semesters are exactly the same;
2. the course of social survey and research method is divided into two consecutive courses each occupying one semester with associated contents.

These two modes are highly representative in the MOOCs course.Therefore, we mainly use the data of three semesters of bioinformatics, pkubioinfo-001(BIO01), pkubioinfo-002(BIO02), pkubioinfo-003(BIO03)and two semesters of social survey and research methods,methodologysocial-001(MS11), methodologysocial2-001(MS21), a total of five courses.

To understand the phenomenon of MOOCs dropout in an instant way, we compute the number of enrollment, the number of students whose score were greater than 0, the number of students passing the course, and the passing rate of each course (as shown in Table 1).

**Table 1.** Course information

| Couse ID | Registers | Records | Score>0 | Score>60 | Passing Rate |
|---|---|---|---|---|---|
| MS21 | 3566 | 3184 | 371 | 185 | 0.051879 |
| MS11 | 7836 | 6051 | 6051 | 255 | 0.032542 |
| BIO02 | 16714 | 15790 | 1268 | 510 | 0.030513 |
| BIO01 | 18367 | 18367 | 1620 | 520 | 0.028312 |
| BIO03 | 16958 | 16072 | 909 | 360 | 0.021229 |

### 3.1 Feature Extraction

Usually, every MOOCs course includes course videos, quiz, forums, and other learning modules. In order to acquire the accuracy of prediction, multiple learning module data were extracted. In the process of obtaining specific data, we extracted the key words from the URL in the log file to identify the learning modules used by the learners and to obtain the learning behavior data for a certain time.

Learners participate in the course by watching videos, taking part in forums, conducting quiz and so on. The participatory process has two important characteristics:

1. Course progress is based on weekly units. A learner can complete the week's learning tasks at any time within a week. In fact, the tracing of learning behavior can be discussed in the context of smaller granularity.
2. The learning behavior is one-direction-oriented mostly, bi-directional occasionally. Most of the learning behaviors,such as clicking and browsing, aim at absorbing knowledge. The interactive activities are relatively rare. This is also the reason why a new form of peer review is introduced to many current courses to enhance bi-directional participation.

Interestingly, disagreement among researchers lies in whether forum data can be introduced as features. Amnueypornsakul and other researchers believe that only 5% -10% of the students would participate in the forum, implying that most learners do not have any forum behavior data. For this majority of learners, it is not appropriate to use forum data to make predictions, and therefore they decided not to use forum data. Yet, we carry out the correlation test between the participation of the forum and whether the learner gets score $\geq 60$.

The correlation test formula is shown below:

$$R = \frac{SET1 \cap SET2}{SET2} \tag{1}$$

We define SET1 as the set of users who participate in forum at least once, and SET2 as the users who get a score $\geq 60$. We find that SET1 and SET2 have high coincidence rate in each course, as shown in Table 2.

**Table 2.** The correlation test

| CourseID | Forum-Participate | Score>60 | Forum-participate Score >60 | Ratio |
|---|---|---|---|---|
| BIO01 | 2645 | 580 | 511 | 0.881034 |
| BIO02 | 1425 | 508 | 395 | 0.777559 |
| BIO03 | 1523 | 358 | 316 | 0.882682 |
| MS11 | 1165 | 290 | 269 | 0.927586 |
| MS21 | 326 | 203 | 153 | 0.753695 |

In the methodologysocial-001 course, the coincidence rate of SET1 and SET2 is 92.8%. That is to say, 92.8% of the learners who have had score $\geq 60$ in this course were involved in the course forum. Therefore, this paper argues that the participatory behavior of the forums has a significant effect on the learners' adherence to learning, hence the data of the students' participation in the forum are added to the forecasting model.

The final feature list we extract is as shown in Table 3.

**Table 3.** Feature list

| Clickstream | | Assignment and test | | Forum | |
|---|---|---|---|---|---|
| Field | Data type | Field | Data type | Field | Data type |
| page_view | Int | try_hw | Int | view_forum | Int |
| page_view_quiz | Int | try_quiz | Int | thread_forum | Int |
| page_view_forum | Int | try_lec | Int | post_thread | Int |
| page_view_lecture | Int | | | post_comments | Int |
| page_view_wiki | Int | | | Upvote | Int |
| video_view_times | Int | | | Downvote | Int |
| video_pause_times | Int | | | add_tag | Int |
| video_pause_speed | Float | | | del_tag | Int |

### 3.2 Learning Cycle

Learning cycle refers to the learners' start and end time. In order to unify the standard, the start and end of the learning time needs a unified standard definition, and the duration needs to be divided in weeks. Because of the large number of courses and the difference of starting time, we define the starting time of each course as the time when the first video click data appears. To determine the end of the course and to maintain the accuracy of the prediction, we count the time of the last learning behavior of the learners in these five courses, and choose the time point when the first 80% students got a score.

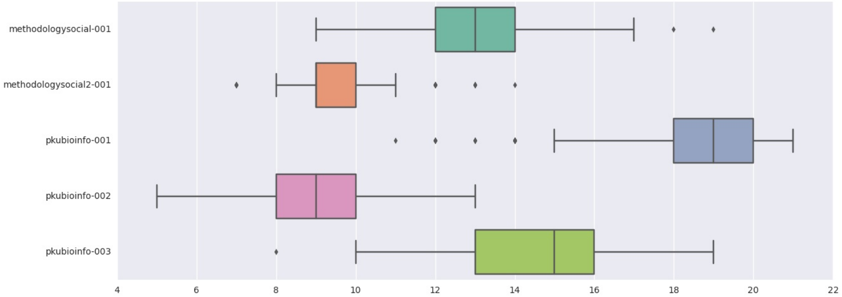

**Fig. 1.** The learning cycles of different courses

As shown in Fig.1, among the learners of 2013 Bioinformatics 001 who got a score, the earliest learning behavior ended in the 9th week and the last 19th week. 80% learners ended the study in the 14th week. Accordingly, the 14th week would be regarded as the end time of 2013 bioinformatics 001. Other courses share the same principle.

### 3.3 The Starting Point and Dropout Point

We do a more in-depth statistical analysis to each persons start time and dropout time. We define whether or not a learner participates in the course in a given week as 1 or 0, then we get a sequence about whether the learner participates in the course every week. Then, we define the point when the number 1 first comes up as the starting point, which means when the learner begins the course. We define the week after the number 1 last comes up as the dropout point, from which the learner no longer comes up.

Then we examine the relationship between the starting point and the dropout point. We find:

1. The later a learner begins, the higher the rate of dropping out after learning a week.
2. The earlier a learner begins, the higher the rate of persisting to the end of the course.

This shows that the earlier the start time, the more likely to stick to the end. The later you start, the more likely to drop out of the course in the following week. Here we can refer to the explanation in [12]. One possibility is that people who start late are difficult to follow due to excessive material, and the other may be people who start late are more likely to seek specific information and no longer learn after acquisition.[12]

# 4 Modeling

## 4.1 The Model of Predicting Dropout

**The Sliding Window Model** Based on the above analysis, the following discussion focuses on the construction of the sliding window model to predict whether a student will drop out or not.

The model views the entire learning cycle as a continuous sequence. According to the learning behavioral features of the weeks before, the model will predict whether or not learners will participate in the course in the next few weeks.

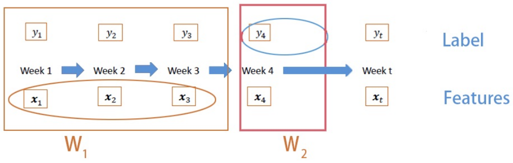

**Fig. 2.** The diagram of the sliding window model.The first window length is $w_1$, the length of the weeks before current week $n$. The second window length is $w_2$, the length of the weeks thereafter current week $n$.[7]

As shown in Figure 2, the model uses the learning behavioral features in the first window to predict the label of the second window. The definition of dropout here is that in the second window the learner does not have learning behavior. The model does not focus on individual learner dropping out, and mainly focuses on the overall situation of the learners' behavior. That is, in the current window, how many learners drop out.

As is mentioned above, most learners in MOOCs will not stick to the end of a course. Therefore, the definition of baseline in this paper is to predict that all learners will not appear in the course of the next week, and then compare the improvement of each model with respect to the baseline. This prediction actually product a large number of applications in reality, such as teachers will send a series of emails to encourage learners to follow during the course even though the learner keeps learning. This paper introduces the method of machine learning to improve this strategy. For the prediction model, we use Logistic Regression (LR), Support Vector Machine (SVM), MLP, LSTM.

**Results** Experiments on the sliding window model were carried out in the five courses mentioned above. If the value of $w_1$ is too small, such as $w_1 = 1$, that is, data of one week is used to predict the current week, the model would be rough and simple. If too large, it would be redundant. To make a compromise, according to previous studies, we do the experiments on $w_1 = 3, w_2 = 1$, using a 5-fold cross validation.Take $w_1 = 3$ as an example, other results($w_1 = 2, w_1 = 4$) work well. The results are shown below (Figure 3)

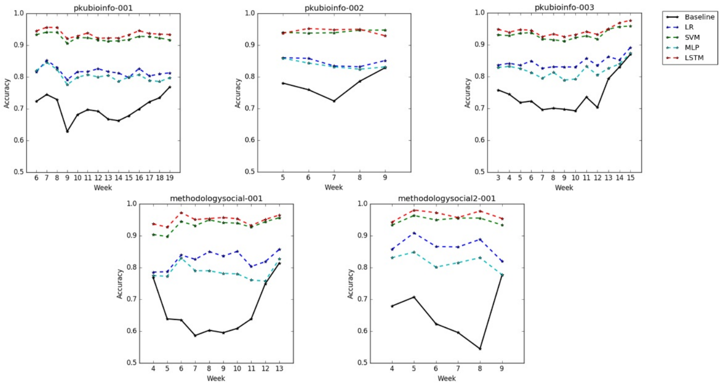

**Fig. 3.** Prediction accuracy of five courses($w_1 = 3, w_2 = 1$)

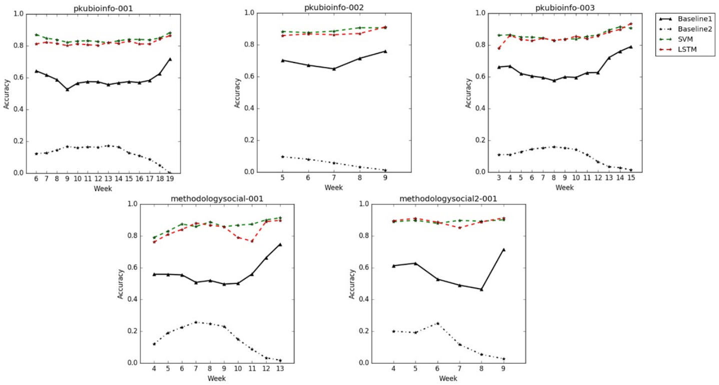

**Fig. 4.** Prediction accuracy of five courses($w_1 = 3, w_2 = 3$)

The experiment result shows that:

1. The prediction accuracy of baseline is generally high. This is due to the large number of dropout every week with a dropout rate of 70% in general
2. Drop-out rate reaches at a peak at the start and end of the course. There were many people who leave early for the course was not suitable for them.
3. In general, the machine learning methods have high prediction accuracy. In different machine learning methods, logistic regression represents the predictive effect in the basic case. In contrast to baseline, machine learning is better at identifying learners who can persist in learning. However, its prediction ability is limited due to the greater need for data.

4. LSTM and SVM have better prediction effect. Compared to multi-layer perceptron and logistic regression, these two methods, the result comes out of which could hardly affected by the data volume, have better prediction ability and can be more stable.

Moreover, we expand the post-window to predict the performance over the next few weeks. If $w_2 = 3$, then there will be $000, 001, 010, 100, 011, 101, 110, 111$, a total of 8 kinds of situations, corresponding to different categories 0 to 7. In the case of the posterior window, baseline1 is used to predict the absence of learning behavior in the next 3 weeks, which is 000; baseline2 represents the predicted behavior for 3 consecutive learning weeks, which is 111.

We find that in the next three weeks, the ratio of 000 is relatively high, and 111 is relatively low. At the beginning and the end of the course, 111 reaches the lowest ratio, which means that only a small percentage of learners performed continuous week learning. In addition, SVM is effective to the multi-classifying. More to say, although there are some differences between LSTM and SVM, the results come of both maintained on a high accuracy.

**Application** The sliding window model solves the problem of monitoring and forecasting the course at different stages, and finds out that the beginning and ending of a MOOC course are the most challenging stage for the learner. In the early stage, they know less about the curriculum and choose to leave after discovering that the course is not suitable for themselves. This is one of the reasons why the early dropout rate is high. Therefore, if the course itself stimulates the learners' interest and keeps the learners concerned in the early stage, it will be more likely for the course to enter a relatively stable period. At the end of the semester, due to the final exam, there will be a lot of learners choose to leave the course. In fact, if the teachers take some strategies to encourage learners to complete the final assessment at the end of the semester, such as courses review, it may help to reduce dropout and encourage students to achieve their final grades and get a certificate.

### 4.2 The Model of Predicting Getting a Score

This model needs to find a course stage as early as possible from the beginning of the course to accurately predict whether the student can get a score. Here, it is defined as the course prediction point. It means after the feature extraction of the learning behavior data before the prediction point, we can judge whether a learner will get a score at the end of the course. This model uses SVM method and the extracted feature data mentioned above. Previous studies have shown that the proportion of learners who can get a score is low, at 5%, showing a positive and negative imbalance in the data. Therefore, AUC and ROC are used to evaluate the model. The classifier with larger AUC value is more effective. It is generally believed that when AUC is greater than 0.9, the classifier has better classification effect. The results are shown below(Figure 5,6,7,8,9).

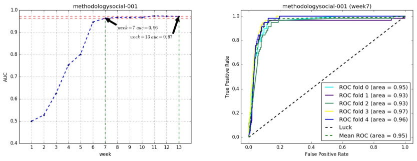

**Fig. 5.** AUC and ROC value of MS11

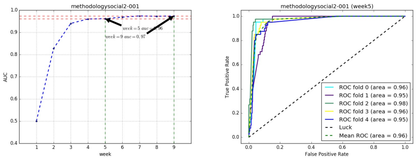

**Fig. 6.** AUC and ROC value of MS21

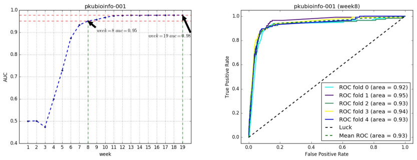

**Fig. 7.** AUC and ROC value of BIO01

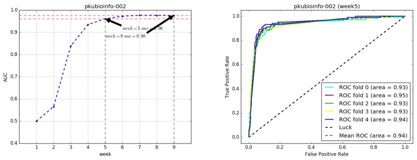

**Fig. 8.** AUC and ROC of BIO02

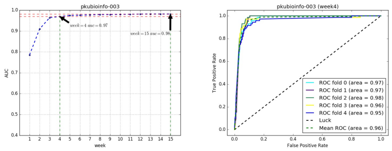

**Fig. 9.** AUC and ROC of BIO03

The time when the AUC reaches 0.9 is inconstant in different courses. In fact, after the prediction point, the model tends to be stable, and the AUC value does not increase significantly. The results of 5-fold cross-validation also show that the model is relatively stable and efficient to predict whether a learner can get a score at the prediction point. The importance of the prediction point is reflected in two ways. On the one hand, the prediction point reflects the number of weeks of learning behavior data needed to classify the classifier. The results show that most of the prediction points are in the middle of the course, indicating that whether a learner can get a score has been determined in the middle of the course. On the other hand, it shows that the course can distinguish the different learners mainly through the first half of the learning stages. The classifier functions better in prediction since the middle of the course.

Further, we discuss whether the classifier can perform well in different courses (Figure 10). In the model of serial courses (MS11 and MS21), the prediction effect is great with AUC = 0.96. This shows that in the serial courses, the classifier can be generalized on account of the similar learning pattern. But it still needs a certain length of learning week data to predict. In the case of using data of only 2 weeks, the AUC decreased to 0.8, which means the predictive ability decreases.

In the case of the same course(Figure 11), data from bioinformatics-001 and bioinformatics-003 were used for model test; The AUC is 0.97, which means the prediction effect is great. It indicates that in the same course, the learning pattern is fixed. The classifier can easily identify whether a learner can get a score at the end of the course. Likewise, with a shorter length of data, the predictive power decreases.

In different courses(Figure 12), with the use of a longer length of learning weeks' data, the AUC value is relatively high, indicating that even in the MOOCs learning, whether a learner can adhere to the medium term basically determines whether a learner can get a score at the end. While with short length of learning data training, the result of the performance is not satisfactory. The AUC value is only 0.7. The result shows that there are still differences between the learning patterns of different courses, and the model cannot be easily reused. The model of the corresponding course needs to be retrained. This is related to the content of the course, the instructor, the nature of the course, etc.

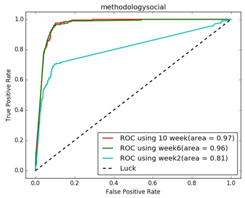

**Fig. 10.** ROC for MS11 model to predict MS21

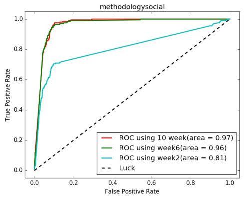

**Fig. 11.** ROC for BIO01 model to predict BIO03

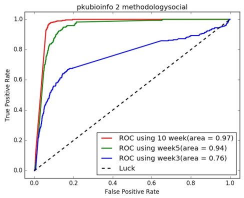

**Fig. 12.** ROC for BIO01 model to predict MS11

## 5 Conclusion

Based on the MOOC curriculum data of Peking University, we extract 19 features according to the past research and the data characteristic. We first did some basic calculation about the dropout students and summarized the characteristics of leaners' starting point and dropout point. We find that the MOOCs course dropout is high, and the learners who start later are more likely to drop out of school. Then, two predictive models are constructed. Model 1 focuses on the overall learner dropout of the course, while model 2 focuses on predicting whether a learner will get a score. The results of Model 1 show that the machine learning and sliding window model can predict the loss of students with high accuracy, and can help teachers track the curriculum, predict the loss of students, grasp the progress of the class. Model 2 results show that the classifier for predicting whether a learner can get a score achieve high accuracy, and whether a learner can adhere to the middle of the course basically determines their probability to get a score. Predictive results of both models can help designers and faculty adjust curriculums from quick feedback to reduce dropout rates.

In the future, we will focus on the similarities and differences of the dropout patterns among different courses and analyze the similarities and differences of learners' dropping motivations among different courses (even in different universities). Then we will do further experiments on a larger dataset, improve the model and give different classes different feedback advice to assist teachers to design their own courses better and to help learners more effective in learning, thus reduce the dropout rate.

**Acknowledgments** This paper is supported by Peking University Education Foundation(2015ZD05).

## References

1. Amnueypornsakul, B., Bhat, S., Chinprutthiwong, P.: Predicting attrition along the way: The uiuc model. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 55–59 (2014)
2. Bienkowski, M., Feng, M., Means, B.: Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. US Department of Education, Office of Educational Technology pp. 1–57 (2012)
3. Elouazizi, N.: Point-of-view mining and cognitive presence in moocs: A (computational) linguistics perspective. EMNLP 2014 p. 32 (2014)
4. Guo, P.J., Reinecke, K.: Demographic differences in how students navigate through moocs. In: Proceedings of the first ACM conference on Learning@ scale conference. pp. 21–30. ACM (2014)
5. Jordan, K.: Mooc completion rates: The data. Availabe at: http://www. katyjordan. com/MOOCproject. html.[Accessed: 27/08/2014] (2013)
6. Kloft, M., Stiehler, F., Zheng, Z., Pinkwart, N.: Predicting mooc dropout over weeks using machine learning methods. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 60–65 (2014)
7. Mi, F.: Machine learning models for some learning analytics issues in massive open online courses. Ph.D. thesis, The Hong Kong University of Science and Technology (2015)
8. Mining, T.E.D.: Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In: Proceedings of conference on advanced technology for education (2012)
9. Moon, S., Potdar, S., Martin, L.: Identifying student leaders from mooc discussion forums through language influence. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 15–20 (2014)
10. Rosé, C.P., Siemens, G.: Shared task on prediction of dropout over time in massively open online courses. In: Proc. of EMNLP. vol. 14, p. 39 (2014)
11. Sharkey, M., Sanders, R.: A process for predicting mooc attrition. In: Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs. pp. 50–54 (2014)
12. Sinha, T., Jermann, P., Li, N., Dillenbourg, P.: Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. arXiv preprint arXiv:1407.7131 (2014)
13. Taylor, C., Veeramachaneni, K., O'Reilly, U.M.: Likely to stop? predicting stopout in massive open online courses. arXiv preprint arXiv:1408.3382 (2014)
14. Yang, D., Sinha, T., Adamson, D., Rosé, C.P.: Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In: Proceedings of the 2013 NIPS Data-driven education workshop. vol. 11, p. 14 (2013)
15. Yang, D., Wen, M., Rose, C.: Towards identifying the resolvability of threads in moocs. EMNLP 2014 p. 21 (2014)