MOOC Student Dropout: Pattern and Prevention

Yunfan Chen
School of Electronics Engineering and Computer Science
Peking University
Haidian, Beijing, China 100871
chenyunfan@pku.edu.cn

Ming Zhang*
Department of Computer Science
School of Electronics Engineering and Computer Science
Peking University
Haidian, Beijing, China 100871
mzhang@net.pku.edu.cn

ABSTRACT

Massive Open Online Course (MOOC) is a completely new education method which appeared in 2012. MOOC is great for life-long learning and educational resource sharing. Until now, there are more than twenty MOOC platforms online. Some studies pointed out that students, however, can hardly follow a course till the end on MOOCs. For most courses, less than 13% of students could follow. Such a high dropout rate will restrict development of MOOCs in the future. This paper shows an observation of MOOC data. A statistic analysis is applied to students' behavioral data. The result is discussed in the perspective of course design on MOOC platform. A general system for predicting students' dropout is developed. With unsupervised learning, the system can fit on different ongoing courses. A discussion of this system with *Data Structures and Algorithms* course is conducted. We also apply the system to *Introduction to Computing* course to test scalability. Additionally, based on the research above, some suggestions for instructing students on MOOC are given.

CCS CONCEPTS

•**Applied computing** →**Distance learning;** *E-learning; Computer-managed instruction;* •**Information systems** →*Data mining;*

KEYWORDS

MOOC, Dropout, Course Design

1 INTRODUCTION

Since 2012, Massive Open Online Course (MOOC) has been developed rapidly, which gains signifficant popularity among both students and educators, and the year of 2012 is called "The Year of The MOOC" [10]. This is a completely new education method which appeared in 2012. MOOC is great for life-long learning and educational resource sharing. Currently, there are more than twenty MOOC platforms online. Both education institutions and industrial organizations of many nations work hard to develop and promote MOOC platform. Meanwhile, a lot of students register at those MOOC platforms preparing for their career, college life or just for fun.

MOOC is evolutionary for many perspective. First, MOOCs are conducted online. Instructors could collect and analyze data to know students' performances and improve their course designs correspondingly. Second, most MOOC platforms are open to all users. Everyone could access learning material freely on MOOC platforms and improve themselves. Third, there is a distinct community integrated with each course on MOOC platforms, and every student could collaborate with other ones to achieve better learning performance.

Some studies, however, pointed out that students can hardly follow a course till the end on MOOCs. Student dropout is not a newly discovered topic. The research related to this topic could be traced back to 1938 [8]. Compared with traditional education, the MOOC suffers more severe dropout problem. For most courses, less than 13% of students could follow [9]. For instance, MIT course *Circuits and Electronics* on edX attracted 154763 students and merely 7157 students finished the course [2]. Such a high dropout rate will restrict development of MOOCs in the future.

In this paper, we firstly discuss the history and recent studies on the student dropout topic. Then we show an observation of MOOC data. A factor analysis is applied to studentsfi behavioral data. The result is discussed in the perspective of course design on MOOC platforms. After that, we developed a general system for predicting students' dropout. The system consists of a sampler, a classifier, an optional differential system and an optional attenuate system. With unsupervised learning, this system can fit on different on-going courses and find students who are at risk of dropout. A discussion of this system with *Data Structures and Algorithms* course is conducted. And, we also apply the system on *Introduction to Computing* course to test scalability. The scalability of the system is fine as long as an analysis is conducted beforehand to choose optional systems. Based on this result, we classify courses into two categories – high coupling courses and low coupling courses. Additionally, based on the research above, some suggestions for instructing students on MOOC platforms are given.

To summarize, we have made the following contributions.

- In Section 2, we briefly survey and discuss the works related to our topic;
- In Section 3, we do statistical study on the student behavioral data and analyze the pattern of the data. Alone with the analysis, we offer several suggestions on course designs for MOOCs;
- In Section 4, we formally define the student dropout prediction problem, and develop a student dropout prediction system. The system we proposed is an online system

*This author is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM TUR-C '17, May 12-14, 2017, Shanghai, China
© 2017 ACM. ISBN 978-1-4503-4873-7/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3063955.3063959

which can be applied to on-going courses to prevent student dropout. The system is also an unsupervised learning system, which does not require human-powered labels;
- Finally, we conduct an extensive experimental study on a real MOOC platform to demonstrate the effectiveness of our prediction system. The results are discussed in Section 5. We also make several suggestions on course designs based on the experiment results.

2 RELATED WORK

Student dropout is a topic which has drawn a lot of attention since almost a century ago. The first study on student dropout, which could be found, was published in the year of 1938 [8]. The motivation of the study was the dropout rate of US colleges at as high as 45%. Some research tried to model the student dropout behavior by analyzing the learning process [12] or dropout motivations [13]. Other research focused on how orientation for freshmen affects the dropout [1].

In the era of MOOC, high dropout rate also has draw a lot of attention to the topic. Since MOOC offers a platform to collect student data easily, most of dropout research on MOOCs rely on objective data but not self reports. [11] classifies users into positive and negative users, and models user participation to predict learning performances. The study in [2] uses clustering method with data from three MOOCs to predict dropout. With help of social network data and forum data, authors of [14] tried to predict the exactly dropout week with Bayesian based Temporal modeling approaches. An semi-supervised machine learning method of dropout prediction is proposed in [6].

Beside the research focused on dropout itself, some other recent researches pay attention to how to increase student engaging and reduce dropout. In [16], the authors find that social engagement could promote commitment and therefore lower attrition. An comparison experiment performed in [9] reveals that whether there is support offered by a small group of tutor could influence the dropout. A survey and interview study in [5] suggests that restricting accessibility and limiting repeatability of online courses could help student better continue a course study. While our experiment and data statistics reveal a different suggestion. A neural network approach is proposed in [3] to improve MOOC efficiency by personalize learning resources for different students and hopefully to consequently reduce the dropout rate.

Compared to existing work, our work is the first to dig into the relationship between data tendency and dropout. We do not only investigate the data and dropout relation, but also develop a reasonable dropout prediction system based on the observation. Our system is unsupervised and online. We discuss the data from many perspectives with many methods, including statistic study and a data mining system experiment. We combine the discover together to discuss. Some suggestions on course management are also made based on the result of prediction system experiment and data observation. This is a complete investigation of student dropout and behavior pattern relationship.

Table 1: Correlation Analysis

Category	Data	r(Correlation)
View	Visit Course (True or False)	0.300^{**}
	#Pages Viewed	0.390^{**}
	Δ#Pages Viewed	-0.002
Learn	#Videos Watched	0.146^{**}
	Δ#Videos Watched	0.149^{**}
	Ave. #Pauses while Watching	0.149^{**}
	Playback Speed	0.035
Test	#Attempts on Quiz	0.198^{**}
	Δ#Attempts on Quiz	0.021
	#Attempts on Programming	0.357^{**}
	Δ#Attempts on Programming	0.013
Discuss	#Visits of Forum	0.248^{**}
	Δ#Visits of Forum	-0.063^{*}
	#Posts Viewed	0.245^{**}
	Δ#Posts Viewed	-0.039
	#Posts Posted	0.161^{**}
	Δ#Posts Posted	0.031
	#Comments Posted	0.106^{**}
	Δ#Comments Posted	0.009
	#Labels Added	0.042
	Δ#Labels Added	0
	#Labels Deleted	^{b}
	Δ#Labels Deleted	^{b}
	#Likes	0.083^{**}
	Δ#Likes	0.006
	#Dislikes	0.042
	Δ#Dislikes	0.042

*: significance at the $\alpha = 0.05$;
**: significance at the $\alpha = 0.01$;
^{b}: too less data to analyze correlation;
#: number of;
Δ: change of value between two consecutive weeks.

3 STATISTICAL STUDY ON BEHAVIORAL DATA

To analyze the factors which may cause students' dropout, we conduct statistical study on students' behavioral data and the tendency of the data. The result could help us design course on MOOC platform with proper strategy and benefit the dropout prediction system shown in section 4. In this paper, we say a student was dropped out at a specific time if the student remain inactive from that time until the very end week counted towards grade. If a student invokes any event log, we do not consider the student as a dropout one. In this section, if a student remain inactive for more than two weeks consecutively until the very last week, we define the student as a dropout one.

3.1 Data Source

We conduct the statistical study on data of Peking Univeristy *Data Structures and Algorithms* course on Coursera. The course lasts 14

weeks and consists of lecture videos, quizzes, programming assignments and a discussion forum. The final two weeks are optional and not counted toward certification or final grade. In total, there are 13,683 students who registered and visited the course web-page. The course itself is in Chinese, alone with English version of course material and video subtitles. The final grade is calculated as 10% of forum discussion, 30% of quizzes, 20% of programming assignments, 15% of midterm exam (quiz) and 25% of final exam (quiz).

9,088 out of the 13,683 students are from developing countries. 72% of students visited the course from Asia, and 21% of students are from North America. The most of students (59%) are from Mainland China followed by 19% of the students from United States and 5% of Students from India.

For the course, only 1,037 students still visited the course till the last three weeks and merely nearly a hundred students passed the course.

3.2 Correlation Result and Analysis

We conduct correlation analysis on the behavior data to see whether these are correlated with dropout or not. Besides the behavior data, we also take the changing between weeks of the data into account. The results are shown in Table 1.

From the correlation, we can draw following guidelines for course design.

(1) **Course Attendance**: A visiting of the course website can help students keep up with the course progress. A weekly reminder may be helpful for students.
(2) **#Web-page Viewed**: The more web pages a student viewed, the less possible the student will drop out. To encourage students view more course materials, a reminder consists of a table of contents for material would be helpful.
(3) **Video Watching**: To encourage students to watch videos, the links to the videos could be included in the weekly reminder.
(4) **#Video Pause**: The result shows that if a student pauses videos more, the student will keep follow the course better. For online course, students may hardly keep focusing on the course material for a long period of time. So we could design some in-video quizzes to help to divide the video into short pieces.
(5) **#Quiz & #Assignment Attempts**: From the result we could infer whether a student get a satisfied score at first attempt actually does not matter much when we consider dropout. So we could help student follow the course by allowing students to try multiple times for an assignment or a quiz.
(6) **Forum**: Almost all types of forum participation could make a student less likely to drop out. It should be a good strategy for course managers to encourage students to participate in the forum discussion.

From the result we can see that the changing between weeks does not correlated with dropout. This is because that **the** 14 **weeks of the course are independent from each other compared with other courses.** This observation helps us design our prediction system in Section 4.

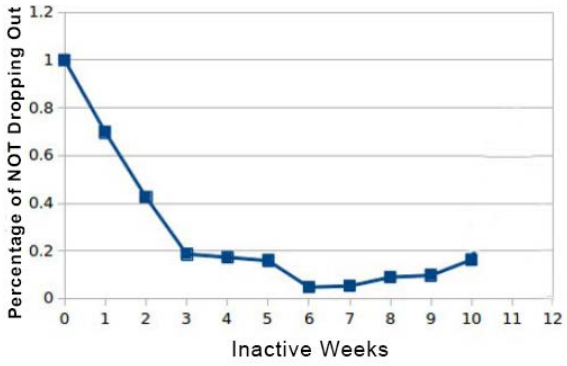

Figure 1: Statistic of Relation Between Inactive and Dropout

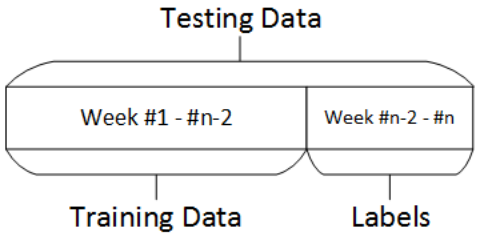

Figure 2: Temporal Workflow

4 DROPOUT PREDICTION SYSTEM

Inspired by the data analysis in Section 3, we design and implement a dropout prediction system. The system is an unsupervised learning system which can effectively predict student dropout. In this section, we first describe the features we used. Due to the fact that we cannot tell whether a student has dropped out or not when a course is still in progress, we then discuss in detail about how to generate label to let the unsupervised learning work. Then we present the general framework of our system, including the basic components and optional components.

4.1 Features And Labels

By discussion in Section 3, we know that the value changing does not matters much for predicting dropout for the course, so we make all those features as optional in our system. We basically use the 15 features present in Table 1 except for the features that start with Δ.

In order to reasonably label the data when course is in progress, we study the relationship between inactive time period and final dropout. The result is shown in Figure 1. Since if a student inactive for more than 3 weeks, the student will relatively highly possible dropout from the course, we label the data with whether there is any activity in most recent three weeks. We label the students who have been inactive for at least three weeks by the time we sample the data as the dropout ones.

That is, we generate our labels with the most recent three weeks' data. Such a strategy makes our system has limitation in dealing with courses less than 4 weeks of history. We test our system from week 4 to the end of course in Section 5 based on this label strategy.

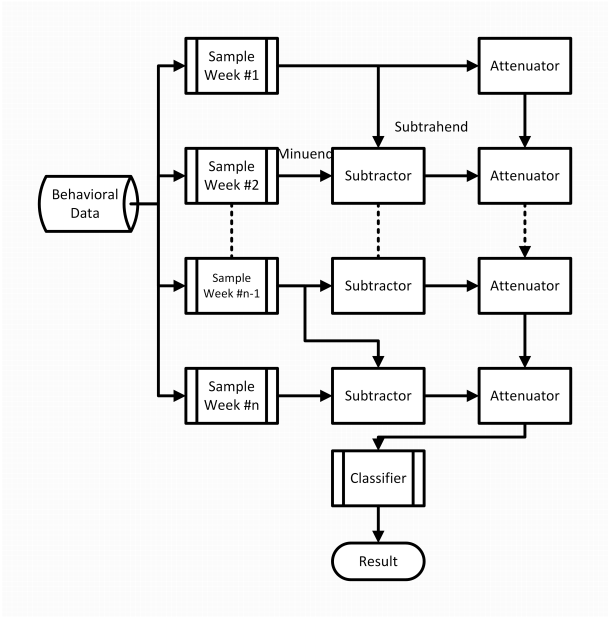

Figure 3: System Workflow

Figure 2 illustrates how we generate label and how we test based on such a strategy.

4.2 Framework

Our system is an online system running on a weekly manner. To get the feature vector for each student, our system samples the data once a week and processes with a subtractor and an attenuator before goes for classification. The work-flow for week n is shown in Figure 3. If the attenuator is not proposed to use, the input of the classifier will be the output of the very last subtractor. If the subtractor is not proposed to use, samples will directly input to the attenuator. If both the subtractor and the attenuator are not proposed to use, the samples (15-dimension feature vectors) directly input the classifier. We discuss the effectiveness and how to select the optional components in Section 5.

The classifier can be any kind of classification method in data mining which takes feature vector as an input and output a binary result. In this paper, we use the **Random Forest** [4] to classify the data.

4.3 Subtractor and Attenuator

Since our discussion in Section 3 reveals the fact that for our course the student behavior change does not correlated with whether a student will dropout or not, we make the subtractor as an optional. Actually, for courses which students need lots of knowledge from previous weeks to learn, the subtractor plays an important role in improving our system accuracy.

We connect the subtractor output to the original feature vector as a new feature vector. Not all features could be subtracted, only the factors shown in Table 1 started with a Δ could be subtracted. Formally speaking, we calculate a new feature vector Y_n from original feature vector X_n by

$$\mathrm{Y}_n = \begin{pmatrix} \mathrm{X}_n \\ \nabla \mathrm{X}_n \end{pmatrix} = \begin{pmatrix} \mathrm{X}_n \\ \mathrm{X}_n - \mathrm{X}_{n-1} \end{pmatrix}. \tag{1}$$

If we choose not to use the subtractor, $\mathrm{Y}_n \equiv \mathrm{X}_n$.

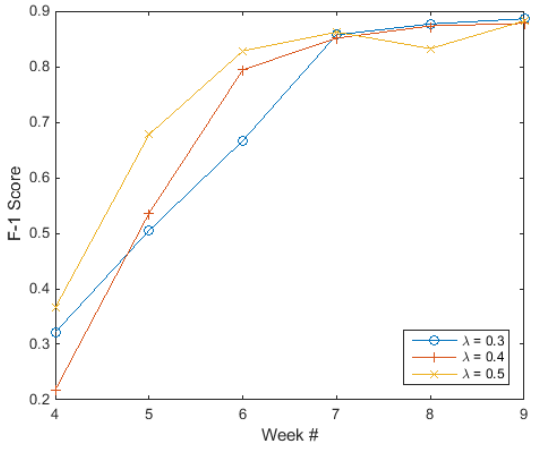

Figure 4: λ Testing for System with Attenuator Only

The attenuator in our system is designed based on the fact that a behavior more recent could tell more about whether a student will dropout in following weeks. We choose exponential decay for our attenuator. The decay constant λ satisfies $0 \le \lambda \le 0.5$. We get a feature vector Z_n output from the attenuator as

$$\begin{aligned} \mathrm{Z}_n &= \lambda^{n-1}\mathrm{Y}_1 + \sum_{k=2}^{n}(1-\lambda)\lambda^{n-k}\mathrm{Y}_k \\ &= \begin{cases} \mathrm{Y}_n & \text{if } n = 1 \\ \lambda \mathrm{Z}_{n-1} + (1-\lambda)\mathrm{Y}_n & \text{else} \end{cases} \end{aligned} \tag{2}$$

The larger λ is, the more history feature remains in the output. If we set $\lambda = 0$, it is exactly the system without the attenuator. If we choose to not use the subtractor, $\mathrm{Z}_n \equiv \mathrm{Y}_n$.

We will see in Section 5 that, for courses with high content coupling, we should choose subtractor but not attenuator. For courses with low content coupling, the attenuator could improve the accuracy of our system but not subtractor.

5 EXPERIMENTAL STUDY

In this section, we first conduct our experiment on course *Data Structures and Algorithms* taught by Peking University on Coursera as described in section 3.1. Then we discuss the result in detail to better understand the system and offer some suggestions on course improvement. Then we select another course – *Introduction to Computing* by the same institution on Coursera to test whether our system could fit other courses as well.

5.1 λ Settings

In order to select a proper λ for the attenuator, we first investigate the system performance under different λ without subtractor. The result is shown in Figure 4. Generally, $\lambda = 0.5$ improve the system

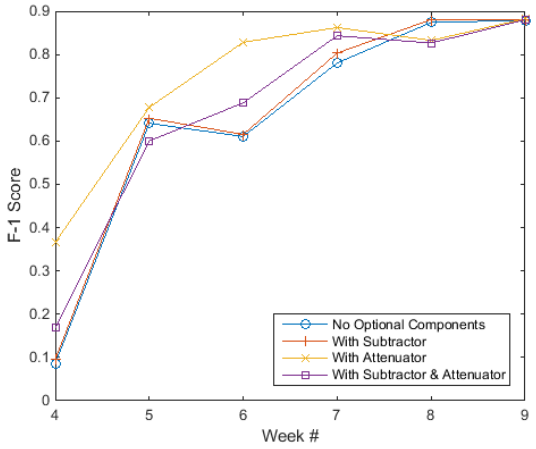

Figure 5: Dropout Prediction on Data Structures and Algorithms

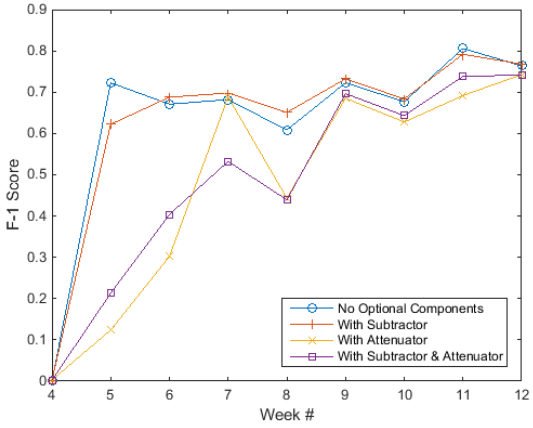

Figure 6: Dropout Prediction on Introduction to Computing

performance best compared to other values. We set $\lambda = 0.5$ for following experiments.

5.2 Result Discussion

There are four different settings for our dropout prediction system – No Optional Components, With Subtractor, With Attenuator, and With All Optional Components. We compared the system F-1 Score under four settings and the results are shown in Figure 5. When $n \leq 7$, a system with attenuator only performs the best. When $n \geq 8$, the four settings have similar performance and the system with subtractor have the best performance. For practical, we'd better apply the system with attenuator only to predict dropout for the course *Data Structures and Algorithms*.

For the first few weeks, our system performance is relatively bad. This is caused by users who registered for the course but did not intend to finish the course. Those users will visit the course material frequently in the first few weeks and cause a lot of noises. Such a noise may be eliminate if we classify user motivation in advance. Motivation proposed in [15] could be applied. The fact that attenuator performs well infers that users' data in several weeks altogether could predict the dropout better. Such a strategy also reduces noises.

Subtractor performs not that good for the *Data Structures and Algorithms* course for the first seven weeks. For each one or two weeks, the course introduces a specific data structure. Such a low relevance between different weeks' course material makes it possible to learn the course at any order. That is, we cannot tell whether a student will dropout simply by the fact that the student visit the course less than before. However, from week 7, the subtractor has a relatively better performance. This is because the course contents require knowledge introduced in previous weeks from week 7 (graph, sort, index, etc.). We call such a kind of course as low coupling course. One week's inactive will not cause dropout from a low coupling course.

We also conduct the same experiment on course *Introduction to Computing*. The course consists of basic computer principles and basic knowledge of C++ programming. The course lasts for 12 weeks and also have quizzes, programming assignments, midterm exam, final exam and forum discussion for grading. Basically, this course has the same setting as the course *Data Structures and Algorithms* but lasts for more weeks. The experiment results are shown in Figure 6.

The system with subtractor only has the best performance among all four settings. We see that the system with attenuator performs relatively bad on *Introduction to Computing* course. The course, not like *Data Structures and Algorithms* course, is high coupling course. That is, learning material requires knowledge introduced weeks before to understand. By analyzing course characteristic in advance could help us choose proper system components. This fact hints us that if a course is high coupling course, one week absence will more likely to cause a student dropout compared with a low coupling course.

We also compare the best system performance among all four settings respectively between the two courses and have the results shown in Figure 7. We see clearly that the system on *Data Structures and Algorithms* course performs better than the system on the other course. This may be caused by the fact that the *Introduction to Computing* course was opened for registration during the whole course period but the *Data Structures and Algorithms* course did not accept new students after the first two weeks.

Since the subtractor and attenuator performances are different under different circumstances, we know that whether different weeks' learning material related with previous learning material matters much on whether student will drop out or not. Therefore, we can help students manage their study by indicating dependencies between course materials.

We see a suddenly F-1 drop of our system when the courses come to week 8 for both courses. That is caused by the midterm exam. For both *Data Structures and Algorithms* course and *Introduction to Computing* course, the midterm exams are conducted at week 8. This fact infers that students may change their behavior pattern when facing to an exam.

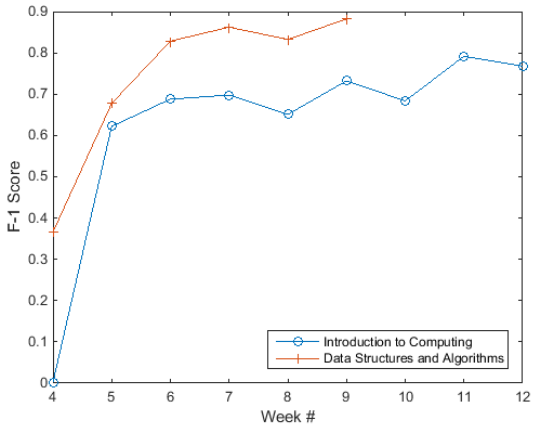

Figure 7: System Comparison between Courses

6 CONCLUSIONS

In this paper, an unsupervised MOOC dropout prediction system is proposed and utilized to predict student dropout based on history data before it happens. Since different features may have different amount of benefit for different courses, we design two optional components based on statistic study of data. Empirical study shows that the dropout prediction system achieves high effectiveness at finding dropout students. The system is inspired by a statistic study of correlations between student behavioral data and dropout. A discussion about the statistic study result is also given in this paper.

Based on statistic study result and the dropout prediction system experiments result, several suggestions are made to help improve course management in the perspective of dropout prevention, such as offering students more chance to try quizzes and assignments, prolong the period of accomplishing graded assignments, encouraging students to participate in forum discussion and designing in-video quizzes to divide each video into short fragments.

Our work is just an initial solution for dropout prediction with student behavioral data. More study on dropout could be conducted to support our system or to develop a new system. Further research on prediction dropout in a short term is a direction, because most of courses last no more than ten weeks and a three-week absence of the prediction system could be an unbearable lose. Students who plan to follow a course till the end have totally different behavioral pattern compared with students who just come for a while. The study of student patterns could be another direction. Some researchers argue that the term "dropout" is a misuse to describe "who failed to complete" [7] and addresses that "success" and "completion" also has different meaning compared with traditional higher education. Therefore, a study on comparing MOOC with traditional higher education could be another direction.

ACKNOWLEDGMENTS

This paper is partially supported by the National Natural Science Foundation of China (NSFC Grant Nos. 61472006 and 91646202) as well as the National Basic Research Program (973 Program No. 2014CB340405).

REFERENCES

[1] Charles A Boudreau and Jeffrey D Kromrey. 1994. A longitudinal study of the retention and academic performance of participants in freshmen orientation course. *Journal of College Student Development* (1994).
[2] Lori Breslow, David E Pritchard, Jennifer DeBoer, Glenda S Stump, Andrew D Ho, and Daniel T Seaton. 2013. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment* 8 (2013).
[3] Brahim Hmedna, Ali El Mezouary, Omar Baz, and Driss Mammass. 2017. Identifying and tracking learning styles in MOOCs: A neural networks approach. *International Journal of Innovation and Applied Studies* 19, 2 (2017), 267.
[4] Tin Kam Ho. 1995. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, Vol. 1. IEEE, 278–282.
[5] Tae-dong Kim, Min-young Yang, Jinhwa Bae, Byoung-a Min, Inseong Lee, and Jinwoo Kim. 2017. Escape from infinite freedom: Effects of constraining user freedom on the prevention of dropout in an online learning context. *Computers in Human Behavior* 66 (2017), 217–231.
[6] Wentao Li, Min Gao, Hua Li, Qingyu Xiong, Junhao Wen, and Zhongfu Wu. 2016. Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. In *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 3130–3137.
[7] Tharindu R Liyanagunawardena, Pat Parslow, and Shirley Williams. 2014. Dropout: MOOC participantsfi perspective. (2014).
[8] John H McNeely and others. 1938. College student mortality. (1938).
[9] Daniel FO Onah, Jane Sinclair, and Russell Boyatt. 2014. Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 Proceedings* (2014), 5825–5834.
[10] Laura Pappano. 2012. The Year of the MOOC. *The New York Times* 2, 12 (2012), 2012.
[11] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé III, and Lise Getoor. 2013. Modeling learner engagement in MOOCs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education*, Vol. 21. 62.
[12] Vincent Tinto. 1975. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research* 45, 1 (1975), 89–125.
[13] Vincent Tinto. 1987. *Leaving college: Rethinking the causes and cures of student attrition*. ERIC.
[14] Wanli Xing, Xin Chen, Jared Stein, and Michael Marcinkowski. 2016. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior* 58 (2016), 119 – 129. DOI: http://dx.doi.org/10.1016/j.chb.2015.12.007
[15] Bin Xu and Dan Yang. 2016. Motivation Classification and Grade Prediction for MOOCs Learners. *Intell. Neuroscience* 2016, Article 4 (Jan. 2016), 1 pages. DOI: http://dx.doi.org/10.1155/2016/2174613
[16] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. 2013. 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. 14.