Predicting Student Examinee Rate in Massive Open Online Courses

Wei Lu, Tongtong Wang, Min Jiao, Xiaoying Zhang, Shan Wang, Xiaoyong Du, and Hong Chen✉

Key Laboratory of Data Engineering and Knowledge Engineering(Renmin University of China), MOE and School of Information, Renmin University of China
{lu-wei,wttrucer,shingle,xyzruc,swang,duyong,chong}@ruc.edu.cn

Abstract. Over the past few years, massive open online courses (a.b.a MOOCs) has rapidly emerged and popularized as a new style of education paradigm. Despite various features and benefits offered by MOOCs, however, unlike traditional classroom-style education, students enrolled in MOOCs often show a wide variety of motivations, and only quite a small percentage of them participate in the final examinations. To figure out the underlying reasons, in this paper, we make two key contributions. First, we find that being an examinee for a learner is almost a necessary condition of earning a certificate and hence investigation of the examinee rate prediction is of great importance. Second, after conducting extensive investigation of participants' operation behaviours, we carefully select a set of features that are closely reflect participants' learning behaviours. We apply existing commonly used classifiers over three online courses, generously provided by China University MOOC platform, to evaluate the effectiveness of the used features. Based on our experiments, we find there does not exist a single classifier that is able to dominate others in all cases, and in many cases, SVN performs the best.

Keywords: MOOC, machine learning methods, examinee rate, prediction

1 Introduction

MOOCs target at unlimited participation of learners and provide an open access to the courses via the web. Since first introduced in 2008, MOOCs have rapidly emerged as a popular online learning paradigm and a wide variety of MOOC providers, including Coursera[3], Udacity[6], edX[5], XuetangX[7], China University MOOC[2], offer high-quality courses from the world's best universities, institutions and enterprises to learners everywhere.

MOOCs are proposed as an important complement but rather than a replacement to the traditional classroom-style education. From the perspective of learners, MOOCs not only provide traditional course materials lectures notes, readings, question-and-answer drills, and quizzes, but also provide videos, and online/offiline discussion forum zones to support community interactions among

students, lecturers, and teaching assistants [22, 23, 27, 26, 25]. More importantly, many MOOC providers offer certificates or statements of completion as long as students have successfully completed an online course, and in some cases, a verified certificate can help provide an important reference for job hunting and further education exploring.

Despite various benefits provided by MOOCs, as observed in the majority of current MOOC providers [8, 11, 18, 28], one of the critical issues related to MOOCs is their low completion rate. That is, only quite a small percentage of learners finally earn the certificates. A recent research conducted over 29 online courses in Coursera shows that the averaged completion rate for the majority of MOOCs is less than 7% [1]. The reason of incurred low completion rate is that, unlike traditional classroom-style education, students enrolled in MOOCs often show a wide variety of motivations, and not all the students are well motivated to earn certificates. In this paper, we find that, although the final examination score only occupies 20% of total mark, being a examinee (participate in the final examination) is almost a necessary condition for a student to earn a certificate. Hence, to raise the completion rate for the learners of earning the certificates and help students improve the learning performance, it is necessary for us to analyze behaviours of the examinees by raising the examinee rate.

Predicting the examinee rate by analyzing the learners' behaviours is challenging. First, the motivations of students engaging in an online course are quite diverse. A student who does not engage in the final examination could either indicate poor learning attitudes, or give up the attendance of the final examination but with positive learning attitudes, or forget the attendance of the final examination due to some unknown reason, or mean that he or she already knows the knowledge point well and is not well motivated to earned a certificate. Due to the above reasons, the data for examinee rate predication is quite noisy. Second, MOOC providers maintain various fact table data and behavioural data, but only a very small fraction of them might be relevant to examinee rate predication. Consequently, the input data of predicting the examinee rate could be sparse. Third, there exist a wide spectrum of machine learning techniques, including naive Bayes classifier, decision tree, logistic regression [13], support vector machine [10], etc., among which selecting the best one that could accurately predict whether a student is an examinee is required to study.

Collaborating with China Higher Education Press [1], we open up an online course named as "Introduction to Database Systems" [2] in the China University MOOC platform and attract more than 20,000 students to engage in our course. As a parter, we are provided by the platform with students' behaviour data of three online courses. As the privacy concerns, we omit their specific names. In this paper, we take these online courses as the object of study, and explore the analysis over the examinees' behaviour data. Our contributions are listed below:

– We find that being an examinee for a learner is almost a necessary condition of earning a certificate and hence, we investigate the problem of examinee

[1] http://www.hep.edu.cn/

[2] http://www.icourse163.org/course/RUC-488001#/info

rate prediction, i.e., verifying whether a student will participate in the final examination of the course. To the best of our knowledge, we are the first group to investigate the examinee rate prediction in this research filed. Our research is of great importance to improve the learning performance and provide references for the teaching effectiveness.
– Collaborating with and generously supported by China University MOOC platform, we conduct extensive investigation of participants' operation behaviours, and carefully select a set of features that are closely reflect participants' learning behaviours. We apply existing commonly used classifiers over three online courses to evaluate the effectiveness of the used features. The experimental results show that there does not exist a single classifier dominating others in all cases, and in many cases, SVN performs the best.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 presents the feature selection. Section 4 briefly introduces the machine learning methods used as the comparison. Section 5 reports the experimental results and Section 6 concludes the paper.

2 Related work

MOOCs have open up a new era of education and attracted a great deal of research interest in MOOC data analysis. Some top conferences launch special workshops or competitions for MOOC data analysis, such as the first data driven education workshop together with the twenty-seventh annual conference on neural information processing systems (NIPS) [4] and the SIGKDD Cup 2015 Workshop for MOOC dropout prediction [8]. For these works, many of them focus on collecting the statistics of the online courses and designing more effective assessment methodologies. For example, Nesterko et al. collect geographic data of analyzing 18 online courses in HarvardX [17]. As an indispensable link of the teach-learn-assess cycle in education, the assessment of students' performance is of great importance. For this reason, peer-assessment methods have been proposed in MOOCs to evaluate the works (especially for subjective questions) of students engaged in the courses [15, 20]. Recently, there is an increasing research interest in MOOC dropout prediction. A great deal of them focus on the prediction using contextual information like posts or click-stream data. Yang et al. try to model students' participation patterns by analyzing their posting behaviour in the discussion forum zones. This approach is restrictive in real applications because posts are just partial features of students' behaviour data. Similar work is proposed in [19] based on students' behaviours in discussion forum. Kloft, et al. target to model the relation between the dropout and the students' most active time. Kim et al. aim to model students' participation patterns by analyzing the students' video click-stream activities, such as skipping, zooming, pausing, playing. Another set of these research work aim to investigate more effective machine learning methods to model the dropout prediction. By extracting the features, [9, 14] apply SVM, [24] applies logistic regression, and [21] applies decision tree

Table 1. Statistics of three online courses

Attributes	**Course A**	**Course B**	**Course C**
Number of enrolled students	71,753	22,145	28,413
Number of students with certificates	371	176	7,790
Number of examinees	625	239	9,211
Number of students engaged in examination view	1,880	1,530	15,780
Number of students with quiz scores	2,855	753	7,413
Number of students engaged in quiz view	3,328	2,684	14,825
Number of students engaged in forum view	1,715	1,223	3,995
Number of students with posts in forum	1,074	304	1,530
Number of students engaged in page views	9,478	6,997	8,214

Table 2. Symbols and definitions

Event	**Definition**
E_1	event that s participates the final examination.
E_2	event that s completes the course A (B or C) successfully with a certificate.
E_3	event that s views the examination page of the course A (B or C).
E_4	event that s views the quiz page of the course A (B or C).
E_5	event that s has quiz score of course A (B or C).
E_6	event that s views a discussion forum page of the course A (B or C).
E_7	event that s has a post in discussion forum zones of the course A (B or C).
E_8	event that s views a video page of the course A (B or C).

to the extracted features to predict the active participation of a student. Besides using a single method, hybrid machine learning methods are used to predict the dropout [16]. Mi et al. take the drop prediction as the sequence labeling problem and hence apply the recurrent neural network (RNN) modeling method [12]. Although a large number of prior research has been conducted, the difference between our work and the state-of-the-art approaches is two-fold. First, we investigate the examinee rate prediction problem while they focus on the dropout prediction problem or others. Second, as we collaborate with the MOOC provider and they provide us with fairly complete student-behaviour data, in this way, we are able to capture more features of contextual information.

3 Feature Selection

China University MOOC platform provides us with three data sets, each of which involves in an individual online course. Without loss of generality, we refer to above three data sets as course A, course B, course C. We collect the statistics of used online courses which are shown in Table 1. We omit the description of these statistics as they are self-explained.

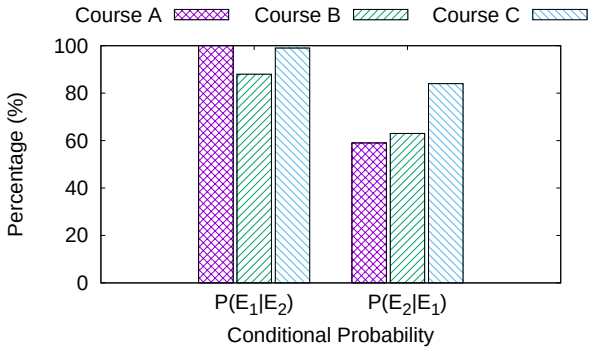

Fig. 1. Interrelationship between examinees and the students with certificates.

3.1 Preliminary Investigation

For ease of illustration, we list the definitions and symbols below for a student s in Table 2 that are used throughout the remainder of the paper.

We first study the necessity for a student to participate in the final examination in order to earn a certificate. We evaluate $P(E_1|E_2)$ and $P(E_2|E_1)$, which are formalized below.

$$P(E_1|E_2) = \frac{\text{Number of examinees with certificates}}{\text{Number of students with certificates}} \tag{1}$$

$$P(E_2|E_1) = \frac{\text{Number of examinees with certificates}}{\text{Number of examinees}} \tag{2}$$

Figure 1 shows $P(E_1|E_2)$ and $P(E_2|E_1)$ for all used online courses. On one hand, as we can see from the values of $P(E_1|E_2)$, every student with the certificate almost participates in the final examination, i.e., although the final examination only occupies 20% of total mark, it almost becomes a necessary condition to earn the certificate. On the other hand, more than 59% of examinees in the end earn the certificates. Apparently, study of examinee rate prediction to improve the completion rate of in MOOCs is of great importance.

We then study $P(E_1|E_3)$, the conditional probability of E_1 given E_3 and present the formula below. We omit $P(E_3|E_1)$ that always equals to 1 since an examinee always view the examination page.

$$P(E_1|E_3) = \frac{\text{Number of examinees}}{\text{Number of students with examination page view}} \tag{3}$$

Figure 2 plots values of $P(E_1|E_3)$ for three used online courses. An interesting observation is that for Course C, $P(E_1|E_3)$ is 72%, i,e., 72% of the students who view the examination pages will participate in the final examination while that for Course A and Course B is just 22% and 15%. The reason of resulting in low $P(E_1|E_3)$, as we guess, is that the examination is too difficult to complete for students or there are too many questions in Course A and Course B.

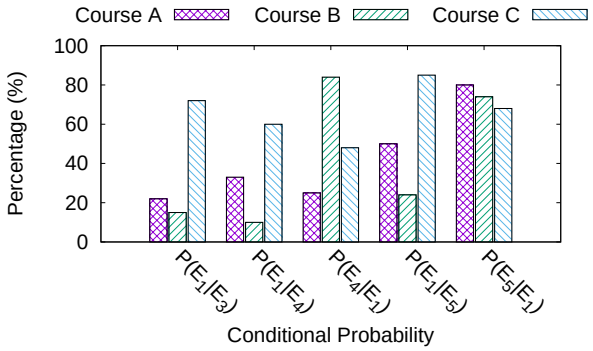

Fig. 2. Interrelationship between examinees and examination page view, quiz page view, quiz scores.

This observation could help the lecturer design the difficulty of the examination properly.

Similarly, we study $P(E_1|E_4)$, $P(E_1|E_5)$, $P(E_1|E_6)$, $P(E_1|E_7)$, $P(E_1|E_8)$ and $P(E_4|E_1)$, $P(E_5|E_1)$, $P(E_6|E_1)$, $P(E_7|E_1)$, $P(E_8|E_1)$, which are formalized from Equation 4 to Equation 13.

$$P(E_1|E_4) = \frac{\text{Number of examinees with quiz page view}}{\text{Number of students with quiz page view}} \tag{4}$$

$$P(E_4|E_1) = \frac{\text{Number of examinees with quiz page view}}{\text{Number of examinees}} \tag{5}$$

$$P(E_1|E_5) = \frac{\text{Number of examinees with quiz scores}}{\text{Number of students with quiz scores}} \tag{6}$$

$$P(E_5|E_1) = \frac{\text{Number of examinees with quiz scores}}{\text{Number of examinees}} \tag{7}$$

$$P(E_1|E_6) = \frac{\text{Number of examinees with forum page view}}{\text{Number of students with forum page view}} \tag{8}$$

$$P(E_6|E_1) = \frac{\text{Number of examinees with forum page view}}{\text{Number of examinees}} \tag{9}$$

$$P(E_1|E_7) = \frac{\text{Number of examinees with posts}}{\text{Number of students that have posts}} \tag{10}$$

$$P(E_7|E_1) = \frac{\text{Number of examinees with posts}}{\text{Number of examinees}} \tag{11}$$

$$P(E_1|E_8) = \frac{\text{Number of examinees with video page view}}{\text{Number of students with video page view}} \tag{12}$$

$$P(E_8|E_1) = \frac{\text{Number of examinees with video page view}}{\text{Number of examinees}} \tag{13}$$

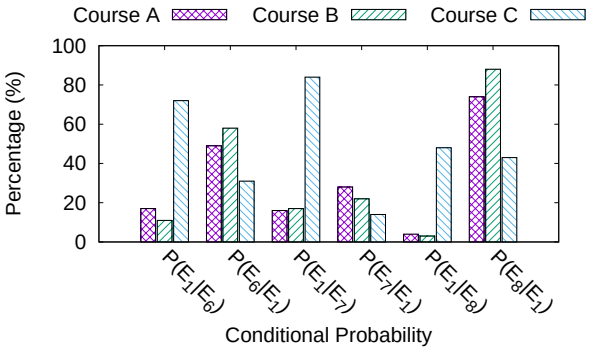

Fig. 3. Interrelationship between examinees and forum page view, number of posts, video page view.

Values of the above conditional probabilities engaged in three used online courses are plotted in Figure 2 and Figure 3. Interestingly, at most 60% of students (from Course C) who view the quiz pages finally participate in the final examination, and only up to 48% of examinees in Course A and C view the quiz pages. Although quiz scores show much higher correlated with the examinee likelihood, the conditional probability for Course A and Course B is still low. Similar observation of interrelationship is found between examinees and forum page view, number of posts, video page view. We omit the details in order to avoid repeatedly elaboration. To summarize, using individual features to predict whether a student is an examinee is far from satisfactory. Although examination view shows a fairly strong correlation with examinee, on one hand, it still leads to a large number of false positives; on the other hand, getting the behaviour of examination view is somehow difficult and one of our objectives in the paper is to motivate students, who do not view examination pages, to participate in the final examination. Therefore, we study of the problem of applying existing prediction models on a set of carefully selected features to predict the examinee rate.

3.2 Features of the Prediction Model

Following the above analysis, we use the features that are closely related to the examinee rate prediction in MOOCs and list them with details in Table 3. Our feature selection is conducted from three perspectives. First, from the perspective of diverse MOOC contents, our selected features (Feature 4, 5, 8–12) cover participants' behaviours over various course materials, including video, online discussion forum, quizzes, examinations. Second, feature 1,2,3,6,9 reflect the engagement of participants' behaviours from the perspective of learning attitude. Third, feature 12 captures the daily class performance from the perspective of learning ability.

Table 3. Used Features of the Prediction Model

ID	Features	Description
1	Number of sessions	Session is a unit of measurement in web analytics, capturing either a user's actions within a particular time period, or a user's actions in completing a particular task. Session helps capture the time period from logging in to logging out the learning platform. A larger number of sessions reflects a higher engagement of user behaviours.
2	Number of requests	Total number of requests including forum views, examination views, video views, etc.
3	Number of active days	We refer to a day on which the user has a request record as an active day.
4	Number of video views	Total number of accessing the video page.
5	Number of video views per session	Averaged number of accessing the video page per session
6	The time gap between video view and corresponding chapter release	Typically, online courses in MOOC platforms periodically release the chapters. A student can access the chapters at any time after they are released. We collect the averaged gap between the time of viewing the video page and the release time of the video. Apparently, a smaller time gap reflects a higher engagement and more active attitude of user behaviours.
7	Number of forum views	Total number of accessing the pages in the course discussion forum zone
8	Number of posts	Total number of posts published in discussion forum zone
9	Number of examination view	Total number of accessing the examination page
10	Number of quiz view	Total number of accessing the quiz pages
11	Quiz score	The score of the quiz

4 The Prediction Models

We define each course with n students engaged and each student is characterized as p features. Therefore, each student is viewed as p-dimensional vector, which is formalized below:

$$X = \{x_1, x_2, ..., x_p\} \in \mathbb{R}^{n \times p}$$

The examinee prediction of X is:

$$Y = f(X) \in \mathbb{R}^n, y \in \{0, 1\}$$

where $y = 1$ represents the attendance of the final examination while $y = 0$ represents the absence of the final explanation.

Naive Bayes Classifier. Naive Bayes is a conditional probability model. In our case, naive Bayes classifier essentially assigns the possibility $P(y|x_1, x_2, \ldots, x_p)$ of a student $\{x_1, x_2, \ldots, x_p\}$ to be a class y ($\in \{0, 1\}$). Consider $P(y|x_1, x_2, \ldots, x_p) =$

$\frac{P(y,x_1,x_2,\ldots,x_p)}{P(x_1,x_2,\ldots,x_p)}$, where the numerator is equivalent to the joint probability model of X and class y. Together with the assumption in naive Bayes classifier that features of objects are pairwise independent, a Bayes classifier, is the function listed below:

$$\begin{cases} 1, & p(y=1) \bullet \prod_{i=1}^{n} P(x_i|y=1) > p(y=0) \bullet \prod_{i=1}^{n} P(x_i|y=0) \\ 0, & \text{otherwise;} \end{cases} \tag{14}$$

Decision Tree. Decision Tree is commonly used as a predictive model, in which each leaf represents a class label associated with a probability, and each internal node is labeled with a certain input feature. Given a student $X = x_1, x_2, \ldots, x_p$, it is able to predict the class label of a y based on several input variables, i.e., $x_1, x_2, \ldots, x_p$ by traversing the decision tree.

Binomial Logistic Regression. The binomial logistic model is used to estimate the probability of a binary classification based on one or more independent features. Logistic regression is based on logistic function with parameter W and b for input X as follows:

$$f(X) = \frac{e^{W \bullet X + b}}{1 + e^{W \bullet X + b}}$$

SVM. Supported vector machine (a.b.a. SVM) is a non-probabilistic binary linear classifier. SVM targets to separate the positive and negative samples using a hyperplane that maximizes the margin of distances from the nearest training-data point of positive and negative objects to the hyperplane. Generally, the SVM classifier targets to solve the following optimization problem to find the above hyperplane: minimize $||w||$ subject to

$$y(w \bullet X - b) \geq 1$$

All the above prediction models are provided by a popular machine learning toolkit Scikit-Learn[3]. All parameters of used prediction models are set by default.

5 Experiments

The input three data sets are divided into two parts, training set and test set, and the ratio of the size of training set to that of test set is 8:2. We use the training set to learn the parameters of each classifiers, and the test set to evaluate the performance of the classifiers, respectively. The metrics of evaluating the performance of classifiers are listed as follows:

- **Accuracy ($\mathcal{A}$).** The accuracy is defined as the probability of correctly verifying whether a participant is an examinee or not an examinee. Formally, it is quantified as:

$$\mathcal{A} = P(f(X) = y|y) \tag{15}$$

[3] http://scikit-learn.org/stable/

Table 4. Student examinee rate prediction

Course	Classifier	Accuracy	Precision	Recall	F-score
A	Bayes	0.903	0.28	0.72	0.41
	DecisionTree	0.973	0.79	0.57	0.66
	Logistic Regression	0.972	0.78	0.55	0.64
	SVM	0.969	0.83	0.72	0.74
B	Bayes	0.958	0.31	0.64	0.41
	DecisionTree	0.99	0.83	0.78	0.81
	Logistic Regression	0.98	0.83	0.66	0.74
	SVM	0.991	0.87	0.74	0.8
C	Bayes	0.67	0.72	0.56	0.63
	DecisionTree	0.852	0.82	0.9	0.86
	Logistic Regression	0.803	0.84	0.73	0.78
	SVM	0.824	0.87	0.76	0.81

– **Precision ($\mathcal{P}$).** The precision is defined as the probability of correctly verifying whether a participant is an examinee. Formally, it is quantified as:

$$\mathcal{P} = P(f(X) = 1 | y = 1) \tag{16}$$

– **Recall ($\mathcal{R}$).** The recall is defined as the fraction of examinees that are retrieved. Formally, it is quantified as:

$$\mathcal{R} = \frac{|\mathsf{examinees} \cap \mathsf{retrieved\ examinees}|}{\mathsf{number\ of\ examinees}} \tag{17}$$

– **F-score.** F-score is defined as the harmonic mean of precision and recall, shown below:

$$\mathsf{F\text{-}score} = 2 \bullet \frac{\mathcal{P} \bullet \mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{18}$$

The experimental results are shown in Table 4. Generally speaking, the accuracy of the classifiers over three data sets is quite high (up to 99%) since the majority of the participants are not examinees. Except naive Bayes, the other three classifiers show very similar accuracy for each course. The precision of applying Naive Bayes classifier is much lower than that of the other classifiers, and SVM achieves the best precision. Decision tree takes the best F-scores over Course B and Course C while naive Bayes and SVM take the best recall over Course A. Based on the performance in precision and recall, SVM takes the best F-score over Course A and decision tree[4] takes the best F-score over Course B and Course C. To summarize, we find, (1) there does not exist a single prediction model that can outperform others in all cases, and (2) in terms of precision, SVM takes the best, and (3) in terms of recall and F-score, SVM and decision tree take the best.

[4] The performance of SVM is fairly close to that of decision tree in Course B and Course C.

6 Conclusion

In this paper, we find that being an examinee for a learner is almost a necessary condition of earning a certificate and hence, the problem of studying student examinee rate prediction is of great importance to improve the learning performance and provide references for the teaching effectiveness. By conducting extensive investigation of participants operation behaviours over three online courses, we carefully select a set of features that cover participants learning contents, learning attitudes and learning abilities. We apply existing commonly used prediction models, and report our findings in the experiment evaluation.

Although we use a wide spectrum of features to do the prediction, the recall and F-score are not as high as we imagined. The reason is that the granularity of the features is still too coarse. For example, the behaviours of participants that watch videos could be quite different. A good participant plays a video with many forward jumps to make a better understanding of the content while a passive participant may just close the video immediately after he plays the video. Our current features cannot differentiate these two guys. To address this issue, investigation of selecting features with finer-granularity will be our future work. Besides, deep learning has become a research hot spot recently and study of applying deep learning to improve the performance of the prediction could be another our future work.

Acknowledgements. Hong Chen is the corresponding author of this paper. The work in this paper was in part supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China under No. 297615121721, No. 297616331721, No. 2015-ms007, and No. 15XNLF09. Our experimental environment is in part supported by the National Virtual Experimental Teaching Center on Big Data Aided Comprehensive Training for Liberal Arts and Social Science, Renmin University of China.

References

1. Academic & university news — times higher education (the). https://www.timeshighereducation.com/news/mooc-completion-rates-below-7/2003710.article/.
2. China university mooc. http://www.icourse163.org/.
3. Coursera. https://www.coursera.org.
4. Data driven education workshop. https://nips.cc/Conferences/2013/.
5. edx. https://www.edx.org/.
6. Udacity. https://cn.udacity.com/.
7. Xuetangx. http://www.xuetangx.com/.
8. K. C. 2015. Mooc dropout prediction. https://biendata.com/competition/kddcup2015/.
9. B. Amnueypornsakul, S. Bhat, and P. Chinprutthiwong. Predicting attrition along the way: The uiuc model. In *Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*, 2014.

10. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2010.
11. T. F. Encyclopedia. Massive open online course. `https://en.wikipedia.org/w/index.php?title=Massive_open_online_course&oldid=694372484/`.
12. M. Fei and D. Yeung. Temporal models for predicting student dropout in massive open online courses. In *IEEE International Conference on Data Mining Workshop*, pages 256–263, 2015.
13. D. Freedman, editor. *Statistical Models Theory and Practice*. Cambridge University Press, 2009.
14. M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting mooc dropout over weeks using machine learning methods. In *Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*, 2014.
15. O. Luaces, J. Díez, A. Alonso-Betanzos, A. T. Lora, and A. Bahamonde. A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments. *Knowl.-Based Syst.*, 85:322–328, 2015.
16. L. M. B. Manhães, S. M. S. da Cruz, and G. Zimbrão. WAVE: an architecture for predicting dropout in undergraduate courses using EDM. In *Symposium on Applied Computing*, pages 243–247, 2014.
17. S. O. Nesterko, D. T. Seaton, J. Reich, J. McIntyre, Q. Han, I. L. Chuang, and A. D. Ho. Due dates in moocs: does stricter mean better? In *First (2014) ACM Conference on Learning @ Scale, L@S 2014, Atlanta, GA, USA, March 4-5, 2014*, pages 193–194, 2014.
18. J. Qiu, J. Tang, T. X. Liu, J. Gong, C. Zhang, Q. Zhang, and Y. Xue. Modeling and predicting learning behavior in moocs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 93–102, 2016.
19. A. Ramesh, D. Goldwasser, B. Huang, H. D. III, and L. Getoor. Learning latent engagement patterns of students in online courses. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 1272–1278, 2014.
20. N. B. Shah, J. Bradley, A. Parekh, M. J. Wainwright, and K. Ramchandran. A case for ordinal peer-evaluation in moocs. In *Neural Information Processing Systems (NIPS): Workshop on Data Driven Education*, 2013.
21. M. Sharkey and R. Sanders. A process for predicting mooc attrition, 2014.
22. J. She, Y. Tong, and L. Chen. Utility-aware social event-participant planning. In *SIGMOD 2015*, pages 1629–1643, 2015.
23. J. She, Y. Tong, L. Chen, and C. C. Cao. Conflict-aware event-participant arrangement and its variant for online setting. *IEEE Trans. Knowl. Data Eng.*, 28(9):2281–2295, 2016.
24. C. Taylor, K. Veeramachaneni, and U. O'Reilly. Likely to stop? predicting stopout in massive open online courses. *CoRR*, abs/1408.3382, 2014.
25. Y. Tong, J. She, B. Ding, L. Chen, T. Wo, and K. Xu. Online minimum matching in real-time spatial data: Experiments and analysis. *PVLDB*, 9(12):1053–1064, 2016.
26. Y. Tong, J. She, B. Ding, L. Wang, and L. Chen. Online mobile micro-task allocation in spatial crowdsourcing. In *ICDE 2016*, pages 49–60, 2016.
27. Y. Tong, J. She, and R. Meng. Bottleneck-aware arrangement over event-based social networks: the max-min approach. *World Wide Web*, 19(6):1151–1177, 2016.
28. J. Zhuoxuan, Z. Yan, and L. Xiaoming. Learning behavior analysis and prediction based on mooc data. *Journal of Computer Research and Development*, 52(3):614–628, 2015.