ISBN 978-981-11-7861-0
Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering
(WCSE 2018)
Bangkok, 28-30 June, 2018, pp. 323-329
doi: 10.18178/wcse.2018.06.058

Relief Approach for Predicting Learner Performance on MOOC

Cheng Ma [+]

College of Computer Engineering in Bengbu University of China

Abstract. The advent of Massive Online Open Courses(MOOC) has led to the availability of large educational datasets collected from researchers. The problem of predicting learner performance on MOOC has received much attention. Considering improving the predictive effect of the learner performance on MOOC, we propose Relief approach to select feature of learners. Based on the online data of edX platform, we divide the characters of learners into three categories, use Relief algorithm to select seven important features, and adopt several classical supervised machine learning methods to build the model for learner performance prediction. The experiments show that learner performance is mainly determined by two kinds of characters of learner type and learner behavior. The Logistic Regression algorithm and Support Vector Machine algorithm is demonstrated that they have high accuracy in predicting learner performance by comparison of the evaluation metrics.

Keywords: Data Mining, MOOC, Prediction

1. Introduction

With the coming of big data and the establishment of the policy about education information, the education environment has transformed from tradition to openness at present. Massive Open Online Course is the typical representative in the large-scale open online course, which has received much attention and rapid development.

MOOC was first introduced in 2006 and emerged as a popular mode of learning in 2012. In 2012, MOOC was conducted by Harvard and MIT. In September 2013, the first-class universities in China announced joining into the team of MOOC, such as Tsinghua, Peking University, Fudan and build the MOOC platform in China. It provides more online resource for a large number of learners. At present, besides three big MOOC platforms on the internet, such as Coursera, Udacity and edX, there are many popular online learning platforms in China. Compared with the traditional remote courses, MOOC provides interactive user forums to support community interactions among students, professors, and teaching assistants. MOOC has many advantages, which is free, easily accessible, completely online courses and so on, and has been widely accepted by the society.

2. Related Work

In recent years, more and more scholars have begun to focus on the development of MOOC. In December 2012, [1] introduced the conception of MOOC in Science journal. In March 2013, [2] detailed the development, status and trends of MOOC in Nature journal, which cause more researchers' attention and participation, [3] analyzed the first batch of all courses on edX platform systematically, [4-6] predicted learners' learning behavior and studied the withdrawal rate of learners through modeling learners, [7] builded the model through history data of students on the MOODLE course, and adopted the neural network and support vector machines to predict whether a student can complete the course successfully, [8] used logic regression to predict whether students can complete online courses by analyzing curriculum records of

[+] Corresponding author. Tel.: +86(0)13205526117; fax: +86-0552-3177310.
E-mail address: bbxymc@126.com.

students, [9] utilized classification and clustering algorithms to predict students' performance based on online course.

In summary, the relevant work has been focused on modeling learners' behavior and predicting students' performance, such as whether students can successfully complete online courses and analyze the losing learners. To the best of our knowledge, few scholars have studied the problem whether they can eventually get a certificate or not. Therefore, we take the open data of the MOOC platform as the research object. Firstly, we classify all the learners' features. Secondly, we select the learners' important influencing characteristics by Relief algorithm. Finally, we adopt several classical supervised machine learning methods to model learners and compare the evaluation index of the prediction results. The details of our research framework are shown in Figure 1.

3. Preliminary

3.1. Data Description

We take the online data of edX platform as the research data, which include data on 16 courses set up by Harvard University and MIT on the edX platform for the 2012-2013 academic year. The dataset consists of 641138 lines and 20 entries. Each record describes the learning information for each student to register a course. The items include LoE, YoB, Gender, Registered, Viewed, Explored, Certified, Nevents, Ndays_act, Nplay_video, Nchapters, Nforum_posts, Grade and so on. Firstly, we divided the characters of learners into following three categories, i.e., the learners' own characters, the learners' type characters and the learners' behavior characters. Next, we select 13 characters to build the prediction model. Table 1 is the description of 13 characters in this research.

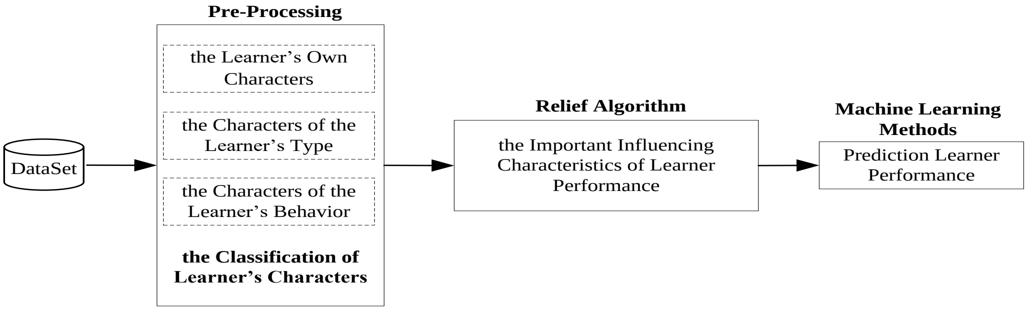

Pre-Processing
the Learner's Own Characters
the Characters of the Learner's Type
the Characters of the Learner's Behavior
the Classification of Learner's Characters
DataSet
Relief Algorithm
the Important Influencing Characteristics of Learner Performance
Machine Learning Methods
Prediction Learner Performance

Fig. 1: Research framework.

Table 1: Learners' features description

ID	Feature Name	State	Type
1	LoE	the highest degree	the learners' own characters
2	YoB	date of birth	the learners' own characters
3	Gender	sex	the learners' own characters
4	Final_cc_cname_DI	country	the learners' own characters
5	Registered	registered learner	the learners' type characters
6	Viewed	general learner	the learners' type characters
7	Explored	positive learner	the learners' type characters
8	Nevents	time of registration course	the learners' behavior characters
9	Ndays_act	number of course visit	the learners' behavior characters
10	Nplay_video	number of video	the learners' behavior characters

11	Nchapters	number of learning chapters	the learners' behavior characters
12	Nforum_posts	number of posting of the forum	the learners' behavior characters
13	Grade	grade	the learners' behavior characters

3.2. Analysis of the Learners’ Own Characters

The basic information of the learners' own characters mainly includes LoE, YoB, and Gender. As found in Figure 2, most MOOC learners have a bachelor’s degree from the education background mostly, and the degree above doctorate and below secondary school only take a small portion. The proportion of male is large and the proportion of female is small from a gender perspective.

Compared with other countries, the degree distribution of English learners is relatively balanced. Learners having secondary school degree in Brazil and India are more than other countries, while the most learners in France and Spain have the master's degree or above. In order to combine learners' educational background, we conducted cluster analysis of countries and found four distinct clusters, as shown in Figure 3.

- Class 1: England, Columbia, Greece, Mexico, Morocco
- Class 2: Brazil, India, Canada, Pakistan, Australia, USA
- Class 3: China, Japan, Nigeria, Egypt, Indonesia, Philippines
- Class 4: France, Spain, Germany, Russia, Ukraine, Portugal, Poland

In class 1, the distribution of learners having the degree of master, bachelor and secondary school is relatively balanced. In class 3, the most learners have bachelor’s degree, while the most learners have master’s degree in class 4.

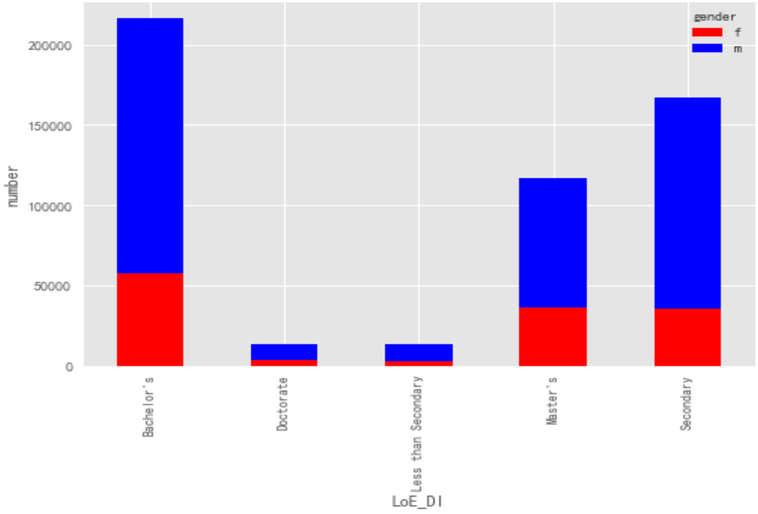


Fig. 2: Distribution of learners’ educational background and gender information.

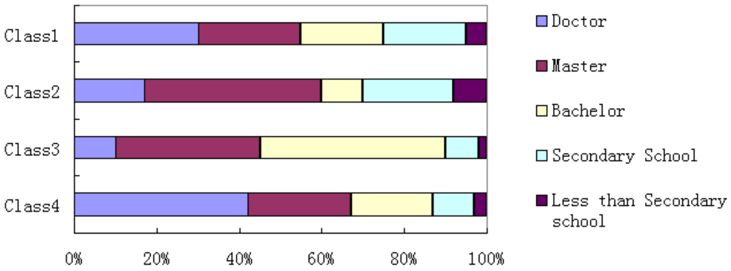


Fig. 3: National clustering based on educational background information.

From Figure 4, the number of below 20 years old is 4032, the number of between 20 and 40 years old is 557419, the number of between 40 and 60 years old is 69667, and the number of above 60 years old is 10020, the average age of the learners is 27 years old. So the proportion of the learners between 20 and 40 years old is larger, and the proportion under the age of 20 is the smallest.

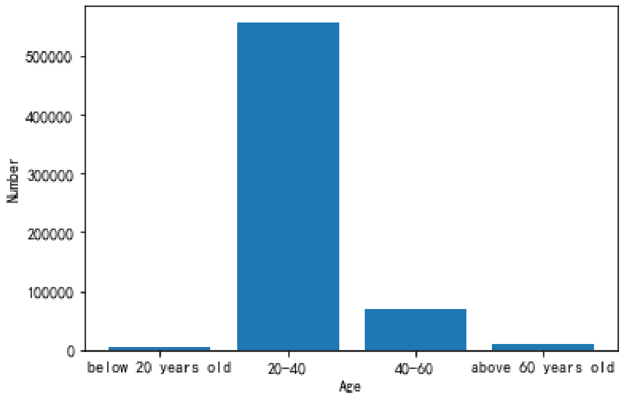
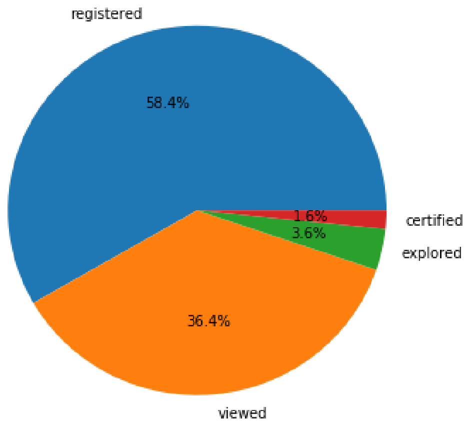


Fig. 4: Distribution of learners' age.

3.3. Analysis of the Learners' Type Characters

The learners' type characters include four basic information, registered learners, general learners, positive learners, and certificate holders. As shown in Figure 5, the proportions of them are 58.4%, 36.4%, 3.6% and 1.6% respectively. It can be seen that after the registration of the MOOC platform, most learners are mainly general learners, and Only access a small number of courseware. Fewer learners have access more than half of the courseware and few learners get the certificate. Few learners get the certificate.



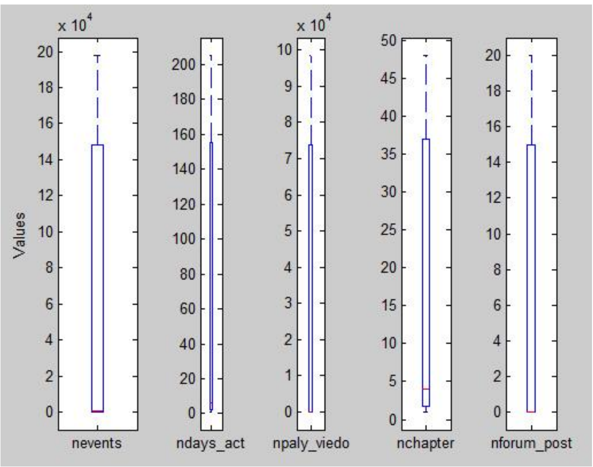


Fig. 5: Analysis of the learners' type characters.

Fig. 6: Analysis of the learners' behavior characters.

3.4. Analysis of the Learners' Behavior Characters

The learners' behavior characters mainly include the following basic information, such as Nevents, Ndays_act, Nplay_video, Nchapters, Nforum_posts. As shown in Figure 6, it reflects the interaction between the learners and the MOOC platform. We find from the data sets by intuitionistic analysis that the learner has a certificate, which is not only related to the course grade, but also to other characters. Learners who have the same grades do not always get certificates, but also relates to other characters of learners. The learner has a score of 0.6 and above, the probability of obtaining a certificate is higher. The learner has a score between 0.5 and 0.59, who is likely to get a certificate with lower probability. While learners with grades below 0.5 have not obtained a certificate.

4. Learner Performance Prediction

According to the content of the third section, we first select 12 learners' characters that may directly affect the learners' performance, Next, in order to obtain accurate prediction results, we adopt Relief Algorithm to select feature of learners on edX platform.

4.1. Relief Algorithm

Relief is an algorithm developed by Kira and Rendell in 1992 that takes a filter-method approach to feature selection that is notably sensitive to feature interactions. It was originally designed for application to binary classification problems with discrete or numerical features. Relief calculates a feature score for each feature which can then be applied to rank and select top scoring features for feature selection. Alternatively, these scores may be applied as feature weights to guide modeling.

4.2. Feature Selecting on Learners

Feature selection plays a critical role in data mining, so we plan to adopt feature selection for learners' characters. We use the Relief Algorithm to calculate the weight of the 13 learners’ characters and remove the characters less than the threshold value of 1E-07. Because the sample is selected in running randomly, which will lead different result. We run the main program 20 times and sort the average of 13 characters importance in descending order. As shown in table 2, there are five learners’ feature importance less than the threshold, such as Registered, YoB, Gender, Nplay_video, LoE. Therefore, we reserve the first 7 learners' characters to train model in the following experiment.

Table 2: Ranking of learners' characters importance.

Rank	Feature Name	Importance	Rank	Feature Name	Importance
1	Grade	5.24E-06	8	LoE	9.52E-08
2	Explored	4.28E-06	9	Final_cc_cname_DI	8.46E-08
3	Nchapters	2.25E-06	10	Nplay_video	7.94E-08
4	Nevents	1.63E-06	11	Gender	6.01E-08
5	Ndays_act	6.78E-07	12	YoB	4.16E-08
6	Viewed	1.72E-07	13	Registered	0
7	Nforum_posts	1.01E-07			

4.3. Experimental Results

For analyzing the effectiveness of predicting learner performance, the four standard evaluation metrics are employed. We use Accuracy, Precision, Recall, and F-Score as the evaluation metrics. we adopt 10-fold cross-validation on edX dataset. There are 641138 records in the dataset(including 17687 positive data and 623451 negative data), which are divided into training set and test set.

The baseline methods we used are as follows: Logistic Regression(LR), Support Vector Machine(SVM), Naive Bayes(NB), K nearest Neighborhood(KNN) and Bayes Network(BN). By employing 12 learner characters and 7 learner characters extracted for training respectively, we can predict whether a student can obtain the certificate successfully. For better illustration, the prediction results are shown in table 3.

Table 3: Comparision of prediction results

Classifier	Accuracy	Precision	Recall	F-score
LR-12features	99.8431%	0.9680	0.9760	0.9720
LR-7features	99.8470%	0.9680	0.9760	0.9720
SVM-12features	99.8439%	0.9640	0.9800	0.9720
SVM-7features	99.8443%	0.9630	0.9810	0.9720
NB-12features	98.1988%	0.6050	1.0000	0.7540

NB-7features	98.2536%	0.6120	1.0000	0.7600
KNN-12features	99.8290%	0.9660	0.9690	0.9680
KNN-7features	99.8291%	0.9680	0.9700	0.9690
BN-12features	99.0311%	0.7440	0.9890	0.8490
BN-7features	99.0420%	0.7460	0.9890	0.8510

After using the Relief algorithm for feature selection, the predictive effect of each classifier was improved, while the accuracy of LR and SVM algorithm is higher. Because it is associated with the purpose of the two loss functions, both of them increase the weight having great influence on the classification, and decrease the weight having a small influence on the classification at the same time. We first adopt the Relief method to compare the weight of the learners’ characters and make the classification of the two methods more efficient. LR algorithm has no changes in the precision, recall rate and F index after feature selection, because of having a good resistance to the noise data. The accuracy of NB and BN classifier is low because the existence of Bayes theory requires strong independent hypothesis, but this assumption is often not established in the actual situation. KNN classifier generated the error of classification which is lower than BN or NB method and had the better character of fitting partial sample.

5. Conclusion

In this paper, the learners' features of edX platform have been divided into three categories. We analyzed each kind of data, sorted the characteristics of the learners' learning performance by Relief algorithm, and selected the first 7 features to train several typical classification models for predicting the performance of the learners. Through the experiment, it is found that the two classification models of logistic regression and support vector machine have high prediction accuracy and short training time. At the same time, it is found that the learners' learning performance is mainly determined by the two types of characteristics of the learner type and the learner's behavior. That is, as long as the learners have a firm learning goal, a good learning attitude and a strong learning perseverance, he can achieve better learning effect.

6. Acknowledgments

This research was supported by grants from the Key Projects for Natural Science Research of Anhui Province(Grant No. KJ2017A568). Finally, I wish to thank my friends and families who helped me to accomplish this paper.

7. References

[1] Stein L A. Casting a wider net. Science, 2012, 338(6113): 1422-1423.

[2] Waldrop M M. Online learning: Campus 2.0. Nature. 2013, 495(7440): 160-163.

[3] Ho A D, Reich J, Nesterko S O, et al. HarvardX and MITx: The first year of open online courses, HarvardX and MITx Working Paper NO.1. Cambridge: Harvard University and MIT, 2014.

[4] Yang D, Sinha T, Adamson D, et al. ”Turn on, tune in, drop out”: Anticipating student dropouts in massive open online courses[C/OL]. Neural Information Processing Systems Workshop on Data Driven Education,2013.

[5] Balakrishnan G. Predicting student retention in massive open online courses using hidden markov models, UCB/EECS-2013-109[R/OL]. Berkeley: University of California at Berkeley, 2013.

[6] Kizilcec R, Piech C, Schneider E. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. Proc of the 3rd Int Conf on Learning Analytics and Knowledge.New York: ACM, 2013: 170-179.

[7] Lara JA, Lizcano D, Mart ńez MA, Pazos J, Riera T.A system for knowledge discovery in e-learning environments within the European higher education area—Application to student data from open university of madrid, UDIMA. Computers & Education, 2014, 72: 23-36.

[8] Hachey AC, Wladis CW, Conway KM. Do prior online course outcomes provide more information than G.P.A.

alone in predicting subsequent online course grades and retention? An observational study at an urban community college. Computers & Education, 2014, 72: 59-67.

[9] Romero C, Lopez MI, Luna JM, Ventura S. Predicting students' final performance from participation in on-line discussion forums. Computers & Education, 2013, 68: 458-472.