Advances in Social Science, Education and Humanities Research (ASSEHR), volume 93
2017 3rd Conference on Education and Teaching in Colleges and Universities (CETCU 2017)

An analysis of learners' learning behavior in Massive Open Online Course

Qin jiwei
Xinjiang University, Urumqi

Jia zhenhong
Xinjiang University, Urumqi

Li hongtao
The Xinjiang Uygur Autonomous Region Party School of CPC, Urumqi

Pan Long
Xinjiang University, Urumqi

***Abstract*—In this paper we access to data of 2,954 learner's learning behavior from Xinjiang University online electives are part of the massive online platform named "Erya". Based on the data, we construct learner's learning model, acquire rules of learning and predict learner's academic record. In order to meet learner's personal needs, the results are used to help educator to develop teaching design and strategies, meanwhile, to optimize online platform and design resources.**

Keywords— massive open online; learning behaviors; learner's academic

I. INTRODUCTION

With the rapid development of the computer and network education technology, massive open online learning is becoming more and more popular. The instructional information resource becomes unparalleledly abundant. Massive open online as a novel, open and flexible way of learning is used to meet the needs of anytime, anywhere learning of lifelong learners. A growing number of people are interested in massive open online. Learners choose autonomously their own courses they are interested, meanwhile, learning behavior log of learner is recorded in massive open online. Because of the differences of learning styles, designing and teaching method could not satisfy learner's demand for personalized. Therefore, it is urgent to analyze the learner's behavior log, to discover the implicit information of the data and learning rule, to promote personalized learning in massive open online.

We acquire and analyze the learning behavior log from Xinjiang University online elective course on the massive open online platform named "ErYa". Furthermore, we explore learning rule, predict learning score and help educator to cognize learner's difference, pertinently make the effective learning strategies which can improve learning efficiency and study quality. Following the platform for optimization, developers can improve their application based on meeting the learning behavior.

II. LITERATURE REVIEW

Massive Open Online Course (MOOC) have been applied in distance education since 2008, and developed rapidly in 2012. In February 2012, the professor of Stanford University establishes Udacity which provides free courses to let more people to receive higher education and employment enhancement. In April 2012, Coursera provide free top University Courses for students from all over the world. In May 2012, MIT and harvard get together to introduce of Interactive online learning named Edx. Udemy implements function of the course in the open education web site in 2012. In December 2012, large-scale network open course platform is named Futurelearn come form 12 universities to show a more formal learning experience, because of curriculum design with social learning.

Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. Based on learning analytics, Clow [1] have proposed "funnel effect" of keep-learning. Rayyan et al [10] statistically analyzed learners' behavior and retention rates. Kizilcec [2] analyzed learners' keep-learning trajectories by using unsupervised clustering. Yang [3] studied the participation and persistence of learning from the perspective of social behavior. Ramesh [4] used stochastic theory to propose a framework that affects the learners' implicit variables of online learning and predict the completion rate for the study. Jiang et al [5] collected the data of learning behavior of the six Chinese classes in Peking University, and divided the learners into five categories. Liu et al [6] analyzed difference of learners' learning behavior form the learners' types, gender, educational background, age, curriculum and so

on. The study was based on the learning data of 13 courses on edX platform. The method of data mining is used to analyze the learners' learning model [7], constructing analysis model and process [8] of network problem learning behavior, meanwhile, emphasizing that modeling is an important basis for learning behavioral analysis, otherwise, providing also theoretical guidance and practical methods [9] for the analysis of learning behavior data. The results of this study are mainly focused on the learning rate of learners' learning retention, the analysis of learners' characteristics, the learners' participation and the passing rate of learners in massive learning. There is no doubt that these results have effectively promoted the development of massive open online.

In view of the above results, this paper describes the learning behavior relate to performance in the massive online learning platform. For the first time, we analyzes the correlation between learners of different disciplines and the types of courses, and according to the learner's academic attributes to predict academic performance in the paper.

III. ANALYSIS OF LEARNER BEHAVIOR DATA

A. Data Sources

We collected learners' behavior of elective courses of Xinjiang University form March to July in 2016.The elective course is based on the ErYa Online Education Platform, which uses an interactive teaching model. Among them, the online teaching mainly consist of watching, discussing, visiting and testing of video resources, however, curriculum, practical operations and other tasks as a focus in underline learning. 2,954 undergraduates took 91 general courses as elective course and the study behavior data for each course was stored in an Excel file. This document contains 10 worksheets to store learner's learning behaviors including Video-On-Demand, discussion, performance, the number of accesses.

B. Data analysis

1) Learning situation and academic performance

By analyzing the relationship between students learning situation and academic performance, we can predict learner's academic performance on basis of existing learning behavior data. And all data is used to analyze students' learning situation, for example, Video-On-Demand, discussion, performance and so on.

(1) Data processing

Analyzing student's situation found that the table of the comprehensive complement situation contains the number of discussion items and visits, the percentage of task completion and video viewing. However, some courses do not have assignment, this paper selects a number of discussion and access, the percentage of task completion and video viewing to describe learners' learning situation.

Number of discussion(*x1*), number of access(*x2*), task completion percentage (*x3*), the average percentage of video view (*x4*), comprehensive performance (*Y1*), test scores (*Y2*). Selecting the above variable to obtain data for each student in these variables. Through the data collection, we know the comprehensive results, the number of discussion and visits, the task completion percentage of the 2,954 students, meanwhile, the average percentage of watching video of 2,112 students and the chapter test scores of 2,028 students. Finally, we will get the learning and achievement situations of 1,960 students.

(2) Data analysis

Partial least squares regression method is a new multivariate statistical data analysis method. It mainly studies the regression modeling of multi-dependent variables on multi-independent variables. In particular, when the variables are highly linearly related, using the partial least squares regression method is more effective. In addition, partial least squares regression and principal component analysis can extract the largest information that reflects the data variation, but the principal component analysis only considers an independent variable matrix, while the partial least squares method has a "response" matrix with respective function. Therefore, spearman rank correlation test is used to obtain the relationship between learning and academic performance, and using partial least squares regression to predict academic performance.

We apply the spearman rank correlation test method to the correlation, as shown in TABLE Ⅰ, TABLE Ⅱ. The *p* value of the *x1*, *x2*, *x3*, *x4* and *y1*, *y2* correlations was less than the significance level of 0.05 by the spearman rank correlation test. So *x1*, *x2*, *x3*, *x4* correlate with *y1* and *y2*, according to coefficient of rank correlation, *x1*, *x2*, *x3*, *x4* have positive correlation with *y1* and *y2*.

TABLE I. SPEARMAN TEST

	x1	***x2***	***x3***	***x4***	***y1***	***y2***
x1	1.00000	0.34318	0.34045	0.32466	0.32733	0.30199
		<.0001	<.0001	<.0001	<.0001	<.0001
x2	0.34318	1.00000	0.85148	0.80597	0.83381	0.74078
	<.0001		<.0001	<.0001	<.0001	<.0001
x3	0.34045	0.85148	1.00000	0.88151	0.89926	0.88646
	<.0001	<.0001		<.0001	<.0001	<.0001
x4	0.32466	0.80597	0.88151	1.00000	0.80194	0.75596
	<.0001	<.0001	<.0001		<.0001	<.0001
y1	0.32733	0.83381	0.89926	0.80194	1.00000	0.88745
	<.0001	<.0001	<.0001	<.0001		<.0001
y2	0.30199	0.74078	0.88646	0.75596	0.88745	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	

TABLE II. CORRELATION BETWEEN VARIABLES

variable	r_s	***t* test *p* value**	**Convergence statement**
Number of discussions and results	0.3273	<.0001	correlation
Number of visits and comprehensive results	0.8338	<.0001	correlation
Task completion percentage and overall performance	0.8993	<.0001	correlation
Video viewing average completion and comprehensive performance	0.8019	<.0001	correlation
Discussion and test scores	0.30199	<.0001	correlation
Discussion and test scores	0.74078	<.0001	correlation
Task completion percentage and test score	0.88646	<.0001	correlation
Average score of video viewing and test scores	0.75596	<.0001	correlation

The above spearman correlation test shows that there is a correlation between the variables, the use of partial least squares regression model to predict student performance. First, the PRESS synthesis obtained by leaving a cross validation method is the smallest (as shown in TABLE Ⅲ), and the number of components to be determined is 4. The cumulative contribution of the four principal components to the dependent variable (as shown in TABLE Ⅳ), where the cumulative contribution to *y1* is more than 95%, the cumulative contribution to *y2* is only 4%. The partial least squares regression model of the number of discussions, the number of visits, the percentage of task completion, and the average percentage of the video watched respectively with comprehensive performance, chapter test scores, the following formula (1) and (2).

TABLE III. CROSS VALIDATION

Cross Validation for the Number of Extracted Factors	
Number of Extracted Factors	***Root mean Press***
0	1.00051
1	0.726568
2	0.713446
3	0.711455
4	0.709761
Minimum root mean PRESS	0.7098
Minimizing number of factors	4

TABLE IV. CUMULATIVE CONTRIBUTION OF PRINCIPAL COMPONENT

	y1	*y2*	**Current**	**Total**
x1	91.7942	2.8377	47.3159	47.3159
x2	94.9081	3.5812	1.9287	49.2447
x3	95.1033	4.0290	0.3215	49.5661
x4	95.6860	4.0445	0.2991	49.8653

$$y_1=0.0002904343x_1+0.0635480806x_2+0.7742743892x_3+0.1661460964x_4 \quad (1)$$

$$y_2=0.0087721138x_1-0.0946432645x_2+0.1919985774x_3+0.058549803x_4 \quad (2)$$

Further, we analyze accuracy and fitting effect of model. Partial least squares regression model under the different principal component corresponding its RMSEP (keep the prediction irregularity radix from a cross verification method) and the abscissa is the component number of each model. As can be seen from Fig.1, two dependent variables' corresponding component number is 4 that the mean square error root is minimal. Otherwise, the four principal components involved in modeling are correct. The ordinate is the predicted value of each dependent variable, and the abscissa is the actual predicted value in the Fig.2. The corresponding points of *y1* are distributed on the main diagonal, indicating that the prediction is very good. However, *y2* corresponds to the scatter plot is almost a vertical line, indicating that the effect of the model fit is not good.

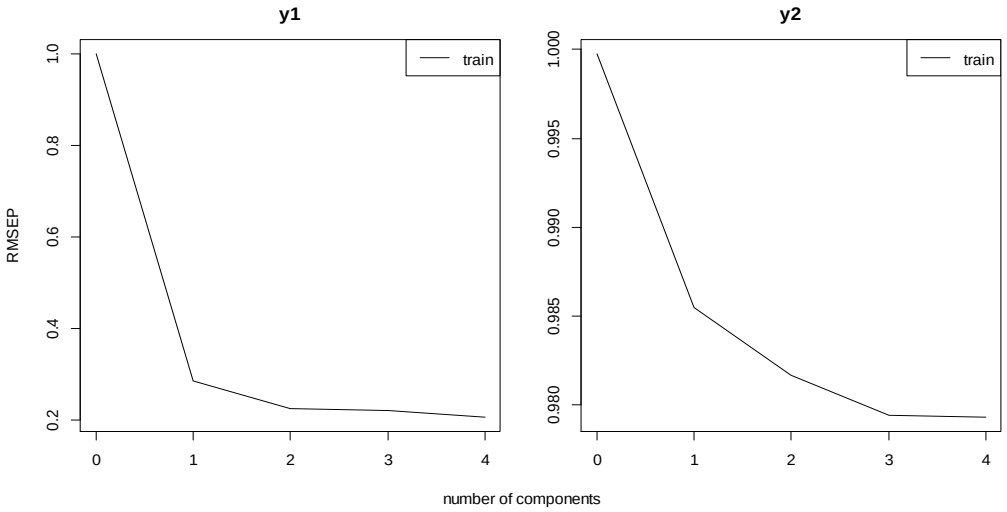

Fig.1. The mean square prediction error.

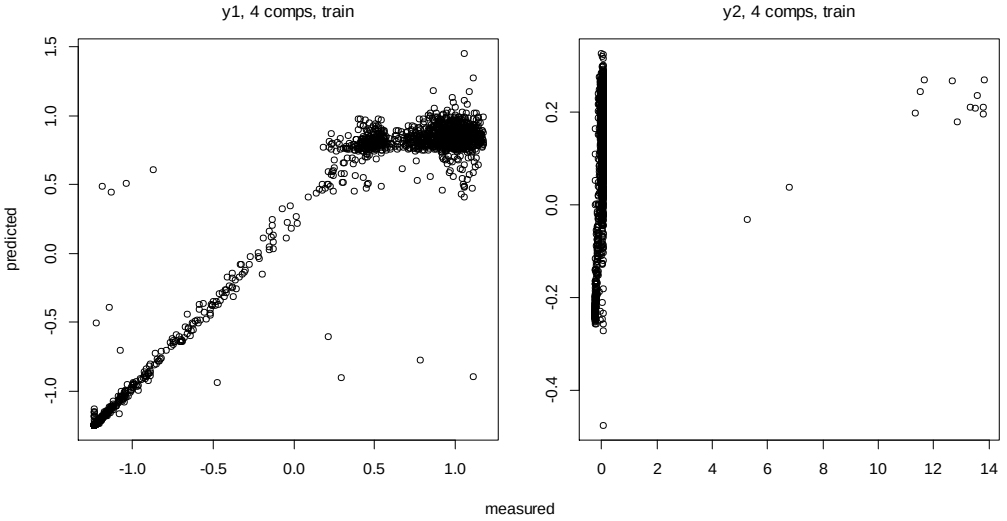

Fig.2. Model fitting effect diagram.

(3) Analysis of results

According to the results of student learning and learning achievement, there are positive correlations among the number of discussions, the number of visits, the percentage of task completion, the average percentage of video viewing and the results of comprehensive and chapter test scores, and the positive correlation between the number of visits, the percentage of task completion, the average percentage of video viewing, the comprehensive score and the score of the test scores are relatively strong, and the relationship between the number of discussions and the comprehensive score, the score of the chapter test is weak. It shows that the number of discussion has little effect on the overall score and the score of the chapter. On the whole, the relationship between the four independent variables and the comprehensive score is greater than the intensity of the relationship between the four variables and the chapter test scores. Therefore, it is suggested that in order to improve students' academic performance, teachers need to pay attention to the task of the learner and the situation of watching the video. At the same time, it is proposed that platform designers optimize the learning situation of students related data tables to improve the proportion of online learning activities in the process of evaluation, to stimulate the initiative of the learners to participate in online learning.

2) Learning situation

The above analysis shows that the number of learners, task completion, video viewing and learner's performance are positively correlated in the learner's learning situation.

(1) Data processing

The above analysis shows that access number, task completion, video viewing and learner's performance are positively correlated in the learner's learning situation. For statistical analysis, we count the number of visits, tasks completed and video views as the number of page visits. The data of the access of all the learners are processed to get the number of pages to be accessed by the students every day. Part of the data is shown in the table below.

(2) Data analysis

We take the time for the abscissa, learn the total number of pages for the vertical plot to draw a line chart, as shown in Fig.3. We can see that students have the most visits in the period from 14 May to 13 June. However, the number of visits was less during 14 March to 13 May and the number of visits was almost zero after 14 June.

Fig.3. The number of access.

(3) Results analysis

According to the results of the students' access to the learning page, the number of access is increasing in the test period, while the number of visits is relatively small in before the examination and after the exam. Therefore, teachers should take measures to increase supervision of elective courses' students, teachers need to design a reasonable and effective learning activity, and regularly inspect students' learning situation of online learning.

3) Academic performance

Study the rules of study of liberal arts, science and engineering students. First of all, we need to analyze whether there is a correlation between liberal arts, science and engineering students in elective courses, excluding the different situation of students' score due to difference of courses. Second logistic regression analysis is used to calculate the excellent rate of students in arts, science and engineering to obtain rule of students' performance.

(1) Data processing

The data of all the 91 courses will be numbered, the 2,954 students of selecting these courses will be divided into the liberal arts, science and engineering. Among them, there are people of 1,283 in engineering, 754 in science, and 917 liberal arts. Comprehensive score of 90 points as a threshold is greater than or equal to 90 divided into "excellent", or "not good". Some of the data are as follows.

(2) Data analysis

Logistic regression analysis is a kind of generalized linear regression analysis model, which is commonly used in data mining. It can statistically estimate each independent variable influence on dependent variable. Logistic regression analysis was used to analyze the relationship between the excellent probability of comprehensive performance and the attributes of students (liberal arts, science and engineering).

We have got the result in TABLE Ⅴ by analyzing the curriculum and students' attributes. The typical correlation coefficient between curriculum and student attributes is 0.01246606, the coefficient of the course is 0.0007467674; the coefficient of liberal arts, science and engineering is 0.02154326. From the typical correlation coefficient can be seen the correlation is small between the curriculum and student attributes (liberal arts, science and engineering).

TABLE V. CORRELATION ANALYSIS

Cor(Typical correlation coefficient)	**Xcoef(The typical load of data *X*)**	**Ycoef(The typical load of data *Y*)**
0.01246606	0.0007467674	0.02154326

Suppose $x = 1$ as engineering, $x = 2$ represents science, $x = 3$ represents arts, $y = 1$ indicates excellent results, $y = 0$ represents the results are not good. Logistic regression model was used to analyze the probability model of $y = 1$. That is, the probability model with excellent results. From the test results (TABLE Ⅵ), we can see the overall effect of the model is better, when R-Square = 0.9847. However, $p < 0.0001$, the model as a whole is significant. TABLE Ⅶ is the result of the Logistic regression model and several descriptive statistics, OR value of the parameter, and 95% confidence interval, the relative excellent rate of engineering, science and liberal arts students. Similarly, the probability model of $y = 1$. That is, the excellent probability model, as shown in TABLE Ⅷ, R-Square = 0.9847, the model fits the study data well. $P<0.0001$, it indicates that the model has significance. As can be seen from TABLE Ⅸ, engineering, liberal arts and science students are relatively good rate.

TABLE VI. FITTING RESULT TEST OF MODEL 1

Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	3,143.129	3,122.056
SC	3,142.921	3,121.431
-2 Log L	3,141.129	3,116.056
R-Square 0.9847		Max-rescaled R-Square0.9847

TABLE VII. RELATIVE EXCELLENCE RATE

Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Engineering	0.908	0.746	1.105
science	1.512	1.210	1.890

TABLE VIII. FITTING RESULT TEST OF MODEL 2

Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	3,143.129	3,122.056
SC	3,142.921	3,121.431
-2 Log L	3,141.129	3,116.056
R-Square 0.9847		Max-rescaled R-Square 0.9847

TABLE IX. RELATIVE EXCELLENCE RATE

Effect	Point Estimate	95% Wald Confidence Limits	
Engineering	0.601	0.489	0.737
liberal arts	0.661	0.529	0.827

(3) Analysis of results

The above results show that the relatively excellent ratio between the engineering students and the liberal arts students is *OR* = 0.908. The outstanding achievement rate of engineering students is about 90.8% of liberal arts students. The relatively excellent rate between science students and liberal arts students achievement is *OR*=1.512 (*p*=0.0003), the outstanding achievement rate of science students is about 151.2% of liberal arts students. The relatively excellent rate between science students and engineering students achievement is *OR*=0.601 (p<0.0001), the outstanding achievement rate of engineering students is about 60.1% of science students. The relatively excellent rate between science students and liberal arts students achievement is *OR*=0.661 (*p*=0.0003), the outstanding achievement rate of liberal arts students is about 66.1% of science students. In summary, the probability of outstanding science students is greater than the liberal arts students. The probability of outstanding liberal arts students is greater than the engineering students'.

IV. SUMMARY AND PROSPECT

In this paper, I collected the learning behavior data of 91 courses in the general elective course platform of 2,954 students from the Xinjiang University in March 2016 - July 2016. Based on the research results of learner's learning behavior data on large-scale online learning, we depict the learning behavior data of students and predict the student's academic on the more flank. The results of the analysis show that the student's learning situation (video viewing, visiting, discussing, completing the task) is positively related to the academic performance. Therefore, teachers take measures to urge students to complete the task, to ensure the frequency of video viewing of Course video, design a reasonable and effective learning activities, regular check the participation of students online learning, strengthen students learning on offline, the establishment of teachers and students of the mutual evaluation mechanism. At the same time, it is recommended that the platform designer optimize the student learning situation related data sheet, improve the proportion of online learning activities in process evaluation, and motivate learners to participate in online learning.

The next step will continue to analyze the reasons for the differences in the performance of science and engineering students, and improve the effect of online learning for engineering students.

ACKNOWLEDGMENT

The research was supported in part by the National Science Foundation of China under Grant Nos.61402392, the China Postdoctoral Science Foundation Nos.2016M592867, CERNET Innovation Project Nos.NGII20160510 and the Doctoral Scientific Research Foundation of Xinjiang University under Grant Nos.BS150263.

REFERENCES

[1] Clow D. MOOCs and the funnel of participation[C]. Proceedings of the Third International Conference on Learning Analytics and Knowledge. ACM, 2013: 185-189.

[2] Kizilcec R F, Piech C, Schneider E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses[C]. Proceedings of the third international conference on learning analytics and knowledge. ACM, 2013: 170-179.

[3] Ramesh A, Goldwasser D, Huang B, et al. Learning latent engagement patterns of students in online courses[C]. Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

[4] Coffrin C, Corrin L, de Barba P, et al. Visualizing patterns of student engagement and performance in MOOCs[C]. Proceedins of the Fourth International Conference on Learning Analytics And Knowledge. ACM, 2014: 83-92.

[5] Jiang Zhouxuan, Zhang Yan, and Li Xiaoming. Learing Behavior Analysis and Prediction Based on MOOC Data[J]. Journal of Computer Research and Development, 2015,52(3):614-628.

[6] LIU San-ya, LIU Zhi, GAO Ju, SUN Jian-wen .Analysis of the Differences of Learners' Learning Behavior in the Context of MOOCs[J], E-education Research, 2016,10:57-63.

[7] Hu Yiling, Gu Xiaoqing, Zhao Chun. Analysis, modeling and mining of online learning behavior[J], Open Education Research,2014,20(2):102-109.

[8] Han Jianhua, Jiang Qiang, et al. Personalized learning model and application effectiveness evaluation in intelligent tutoring environment[J], E-education Research, 2016,7:66-73.

[9] Zheng Qinhua, Sun Hongtao, et al. Modeling and application of online learning evaluation based on learning analysis [J], E-education Research, 2016,12:40-45.

[10] Rayyan ,S. , Seaton,D.T., Belcher ,J., et al. Participation and Performance in 8.02x Electricity and Magnetism: the First Physics MOOC from MITx [A].2013 Proceedings of Physics Education Research Conference [C]. Portland: Physics Education Research Conference 2013.