

Score Prediction Model of MOOCs Learners Based on Neural Network

<https://doi.org/10.3991/ijet.v13i10.9461>

Yuan Zhang

Qinhuangdao Institute of Technology, Qinhuangdao, China

Wenbo Jiang^(✉)

Hebei University of Environmental Engineering, Qinhuangdao, China
305898103@qq.com

Abstract—Through analyzing the behavior data of MOOCs learners, a MOOCs learner's score prediction model is constructed based on clustering algorithm and neural network in this paper. By using this model, we can find out the neglected information and hidden learning rules in the MOOCs learning process. The model can provide personalized guidance for each user and improve learning efficiency. The model can provide personalized service to help learners form personalized learning strategies, and it also can alert learners with low grades and risk of dropping out.

Keywords—massive open online course; score prediction model; data mining; clustering algorithm; neural network

1 Introduction

Since 2012, Massive open online courses (MOOCs) have broken forth and spread widely among various universities in the world, it has an extremely important impact on the teaching of higher education in the world [1-2]. Coursera, edX and Udacity three platforms have been introduced by many world-renowned universities to open up high-quality online educational resources and services to users worldwide [3].

The biggest characteristic of MOOCs is that it has a large number of learners and a variety of learner groups. At the same time, learners' knowledge background and learning motivation are varied. In the MOOCs environment, the learner's learning behavior will be recorded in a variety of data. Therefore, it is necessary to study and analyze the learning behavior data of MOOCs learners [4].

Although MOOCs has many advantages, there are many disadvantages, such as, lack of interaction between learners and teachers, high suffixing rate and so on. To solve the above problem, some scholars analyzed the data of MOOCs learners' learning behavior [5-7], and they find methods to overcome the limitations of MOOCs, and some scholars used the analysis results of learning behavior to improve the learning effects and methods in MOOCs [8-10].

In the research of the emergence and development of MOOCs, Rai et al. proposed the application of social software to MOOC platform from the perspective of "learner center" to achieve the horizontal expansion and gradient development of MOOCs [11]. Fu et al. used visual technology to study various literature key words related to MOOCs. Through the methods of word analysis and social network analysis, they revealed the hot spots and development trend of MOOCs [12].

In the analysis of the characteristics of MOOCs, Kennedy analyzed the typical characteristics of MOOCs from four aspects: scale, openness, networking and innovation [13]. Hughes et al. selected seven typical MOOCs, which were analyzed from six aspects, such as platform positioning, curriculum organization and teaching methods [14].

On the innovation of MOOCs and the impact on traditional teaching, Zeng et al. elaborate the challenge of MOOCs to the traditional education [15]. Lai et al. proposed a mixed teaching mode suitable for university teaching based on MOOCs [16].

On the curriculum and platform analysis of the MOOCs platform, Kim used the evaluation standard of web-based teaching platform, and elaborated the usefulness of MOOCs in exploring the learner group's conscious learning [17]. After analyzing the background of MOOCs, Liu et al. compared curriculum resources, learning activities and learning evaluation based on the characteristics of the curriculum design on MOOCs platform [18].

In the analysis of the MOOCs data, Jiang et al. proposed the factor analysis model based on intrinsic motivation, basic psychological demand factors and the design factors of MOOCs [19]. Zhuo et al. take a course in UOOC as an example to collect all the data of the course, and study the user's learning behavior based on big data analysis [20]. Mou et al. took six MOOC courses as an example to analyze the learners' behavioral, they explored the learning behavior of the learner at the course level [21].

2 Analysis of Learner Behavior in MOOCs

The most important problem of learner behavior analysis is to get accurate, reliable and comprehensive learner data. A large number of structured and unstructured data will be generated through interaction between learners and instructors, or between learners and learning resources.

There are two kinds of data sources, one is to collect learner behavior data by questionnaires and the other one is to share open data sets published by some international organizations. In this paper, Canvas dataset was chosen as the source dataset for MOOCs learner behavior data analysis. Data in Canvas open dataset is related to MOOC courses and learner learning behavior. Learners' behavior in the course may help researchers to better describe and understand the learning situation and learning behavior in MOOCs. The Canvas dataset has a large number of attributes to record a learner's behavior in a particular course. The specific feature attributes are shown in Table 1.

There are 325198 records in the Canvas dataset, and 89213 records of them are missing part of the information. This part of data will be cleaned out when we analyzed.

Age distribution of learners is shown in Table 2.

Table 1. Feature attributes of Canvas Dataset

No.	Feature Attribute	Description
1	course_id	Course ID
2	discipline	The discipline of the course
3	User_id	User ID
4	viewed	The number of interactions in the course
5	explored	Percentage of learning content
6	grade	Learner's achievement
7	age	Learner's age
8	LoE	Level of education
9	start_time	The time of learner for the beginning of the course
10	last_event_time	The last time of the user interaction
11	nday_events	The number of days of interactive learning with the course
12	forum_posts	The number of posts in the forum
13	course_length	The length of the course
14	ncontent	Number of course chapters
15	reason	The reasons for the participation of the course
16	expected_hours_week	The number of hours expected for a week in the course

Table 2. Age Distribution of Learners

No.	Age distribution interval	Number	Description
1	{18-30}	15672	Number of learners between 18 and 30 years of age
2	{31-45}	17821	Number of learners between 31 and 45 years of age
3	{more than 45}	6092	Number of learners more than 45 years of age
4	unknown	196400	The number of learners with an empty property

As shown Table 2, the number of learners aged 18 to 30 is the largest. The number of learners aged over 45 is the least. It shows that MOOCs learners are younger and more familiar with the Internet platform, and they have strong thirst for knowledge and learning goals.

The distribution of educational level of MOOCs learners can explain some characteristics of the MOOC user groups. As shown in Figure 1, MOOCs learners mainly focus on bachelor's degree and master's degree. It shows that the learners attracted by MOOCs are basically receiving good higher education.

Most MOOCs learners can't spend a lot of learning time. For this purpose, the Canvas platform had investigated the appropriate value of learners' expectation of spending time each week on learning. From Figure 2 we can see that 39% learners tend to spend 2 to 4 hours per week for learning, and 33% learners tend to spend 1 to 2 hours per week for learning. From the data point of view, the learners are expected to spend less time on the MOOCs platform, which means that the course arrangement should be as short as possible to meet the needs of the learners.

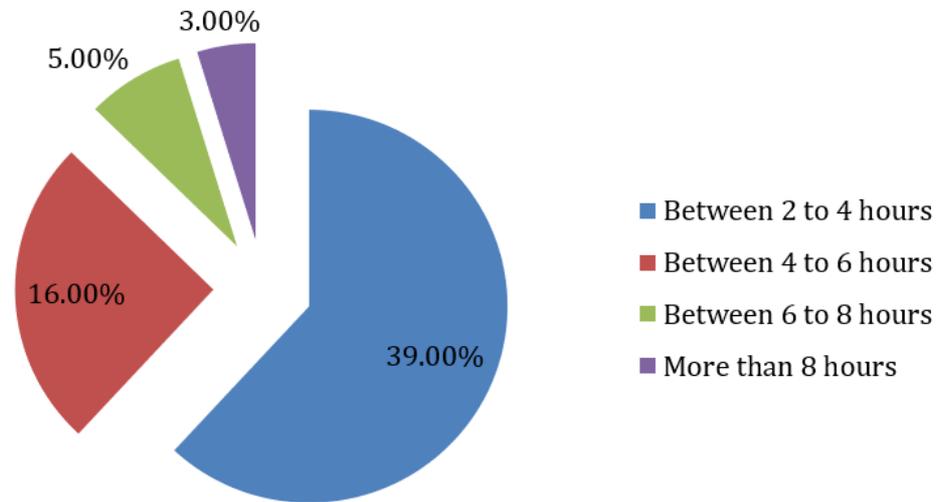


Fig. 1. Education Level of Learners

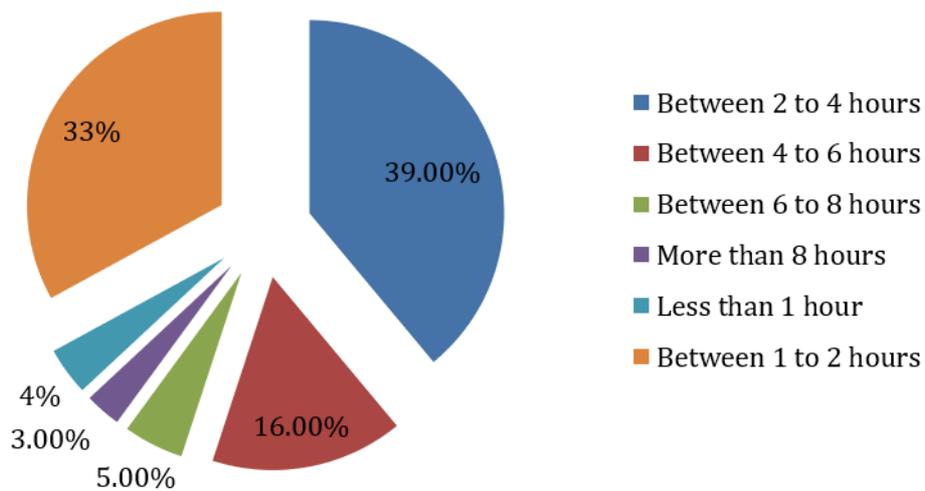


Fig. 2. Weekly Expected Hours of Learners

In this paper, we assume that learners are divided into three categories: Active learners, Passive learners, and Negative learners. Active users are active in learning activities, actively speak in the forum, actively finish their homework and learn more than 50% of the course content. Passive learners, who use traditional learning methods, only watch videos, browse courseware and finish homework. Negative learners have passive learning, few learning activities or less learning behavior, and lack of autonomous learning ability. According to analysis results, there are only 9856 learners in the Canvas dataset of active users, there 79827 passive learners, and the rest are negative learners.

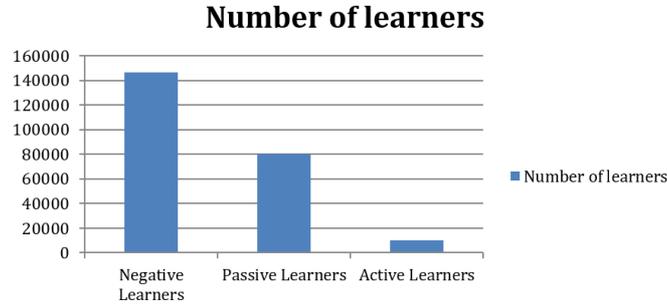


Fig. 3. Learner classification

The analysis of factors that affect learners’ grades is an important part of MOOCs analysis. In order to make a more intuitive analysis of the factors that affects the MOOCs performance, the following typical feature attributes are selected as reference for analysis: course interaction times, number of interactive days in the course, number of course chapters, number of posts in the forum and the length of course. The relationship between the five characteristics and scores is analyzed, and scatter plots of the scores and the five characteristics are plotted respectively.

As shown in Figure 4, in terms of courses interaction times, the average level is low except for few high values. Moreover, there is not much difference in courses interaction times corresponding to different grades.

From Figure 5, with the increment of the number of interactive days in the course, the score becomes higher.

The number of course chapters is distributed from 0 to 100, and the results are dense at both ends. There are obvious linear rules in Figure 6.

As shown in Figure 7, there was a clear positive correlation between the number of posts on the forum and the scores.

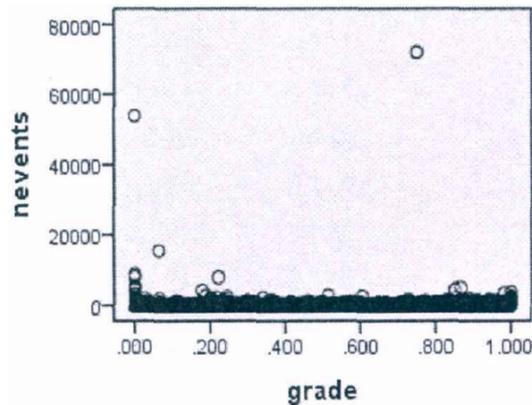


Fig. 4. A scatter plot of score and course interaction times

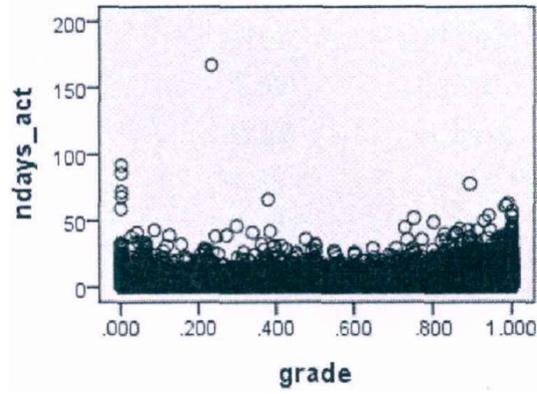


Fig. 5. A scatter plot of score and number of interactive days in the course

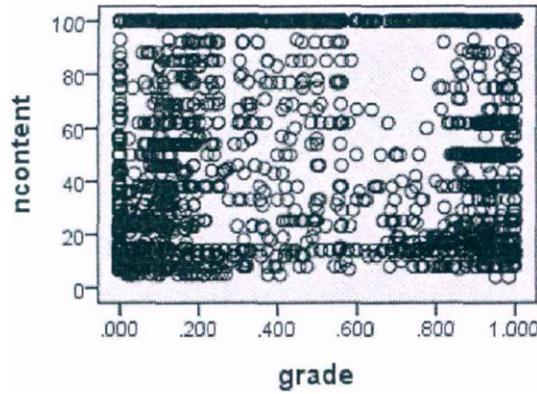


Fig. 6. A scatter plot of score and number of course chapters

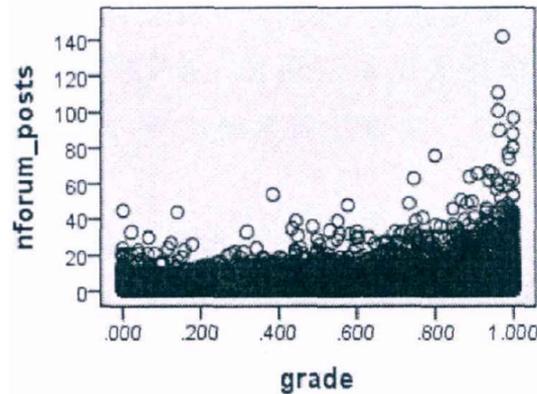


Fig. 7. A scatter plot of score and number of posts in the forum

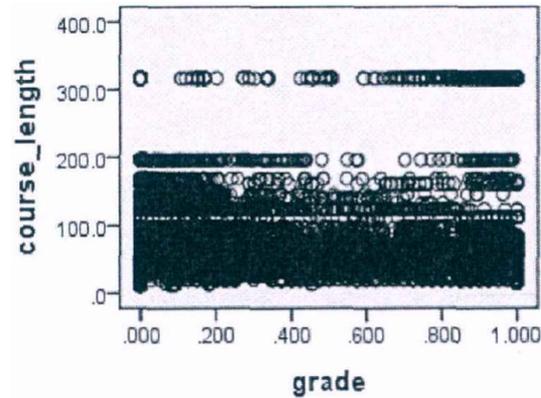


Fig. 8. A scatter plot of score and the length of course

There is no obvious trend in the length of the course. However, unlike the number of chapters, the length of course is evenly distributed, and there is no polarization.

3 MOOCs Score Prediction Model Based on Neural Network

Aiming at the various learning behaviors and records contained in the learning process of MOOCs, a score prediction model is constructed to provide learners with strong prediction data support. In the process of learner learning behavior, there are uncertain behavior changes. Therefore, in this paper we choose RBF neural network as a method to construct a score prediction model.

The model takes the learner behavior data as the training data and selects the behavior characteristics of the processed earners as the input of the model, and predicts and outputs the prediction results of MOOCs learners. For constructing a score prediction model, we need to qualitatively analyze the independent variables of the function according to the learner's behavior characteristics, and construct a one-to-one correspondence function relation.

In this paper, five independent variables are selected as the input variables of the score prediction model, and the scores of the learners are taken as the output variables of the prediction model, then the input and output relation of the score prediction model can be defined as follow:

$$y = f(x_1, x_2, x_3, x_4, x_5) \quad (1)$$

where y represents score, x_1 represents course completion ratio, x_2 represents the number of posts in the forum, x_3 represents homework completion ratio, x_4 represents course interaction times, x_5 represents the number of interactive days in the course.

In this paper, RBF neural network is used to construct learner score prediction model. RBF neural network is a single hidden layer feed-forward network. It is divided into three layers. The first level is the input layer. The function is to input the n input to the m hidden layer. The input number is the number of input variables in the prediction

model. The second level is the hidden layer, and the transformation function of the layer selects the radial basis function. The third level is the output layer, which can adjust the linear weight in response to the input mode.

The mathematical model of RBF neural network with threshold is defined as follows:

$$y_i = \sum_{k=1}^M w_{ij} h(\|x_i - c_k\|) + \theta_j \quad (2)$$

Where $h(\|x_i - c_k\|)$ represents radial basis function, $\|\cdot\|$ is Euclidean norm, it represents the distance between x_i and c_k , $x_i \in R^n$ is the i -th input of the neural network, c_k is the center of the k -th node of the hidden layer, θ_j is the threshold of the j -th output node.

The structure of MOOCs score prediction model based on RBF neural network is shown in Figure 9.

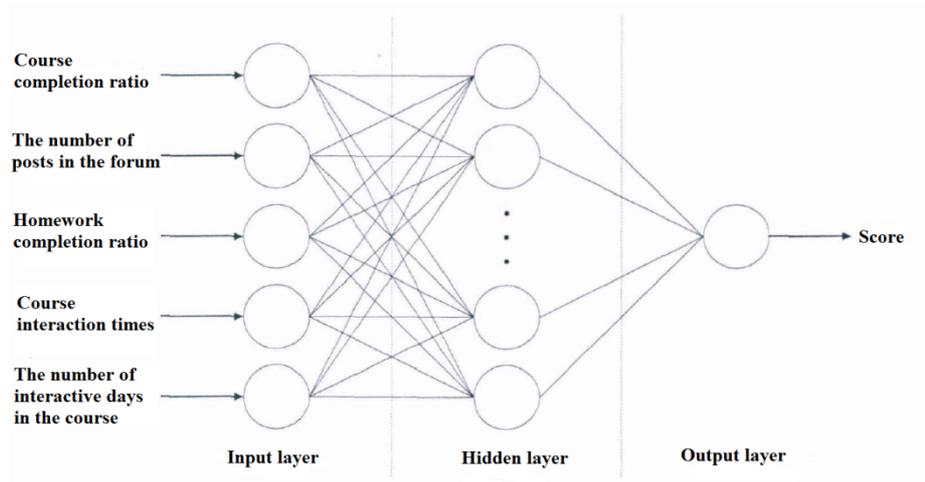


Fig. 9. The structure of score prediction model

In this paper, Gauss function is chosen as the radial basis function of RBF neural network, the Gauss function is defined as follows:

$$h(v) = e^{(-v^2/2\sigma^2)} \quad (3)$$

On the basis of the use of Gauss's function, the output of the y -th node of the hidden layer is:

$$H_j = \exp\left(-\frac{\|x-c_j\|^2}{2\sigma_j^2}\right) \quad (4)$$

The linear mapping is realized by $H_j \rightarrow Y$ in the output layer, Y represents the output of the output layer node, then the output of the j -th node in output layer is:

$$Y = \sum_{j=1}^k w_j H_j + \theta \tag{5}$$

where k is the number of hidden layer nodes, w_j is the weight of the j -th hidden layer node to the output layer node. H_j is the output of the j -th hidden layer node. θ is the threshold of the output layer node.

4 Simulation Experiment and Result Analysis

According to the analysis and feature attributes of Canvas dataset, the score prediction model based on the K-means feature selection is used to predict the learner's score in simulation experiment. In experiment, we extract 3100 groups of pre-processed data to construct prediction model, 3000 groups of data are acted as training dataset and 100 groups of data are acted as test dataset.

The simulation results of RBF neural network score prediction model based on K-means feature selection is shown in Figure 10.

The simulation results of RBF neural network score prediction model based on random centers is shown in Figure 11.

By observing the time efficiency of the two prediction models, it is found that the time efficiency of the score prediction model based on K-means feature selection is higher. It shows that the convergence accuracy of the score prediction model proposed in this paper is better than that of the RBF neural network prediction model based on random centers.

In order to further use 100 groups of test data to verify the validity of the constructed score prediction model. In this paper, a comparative experiment is designed to compare users' actual performance and prediction results. The results show that the model proposed in this paper has a high accuracy in predicting the score of MOOCs learners.

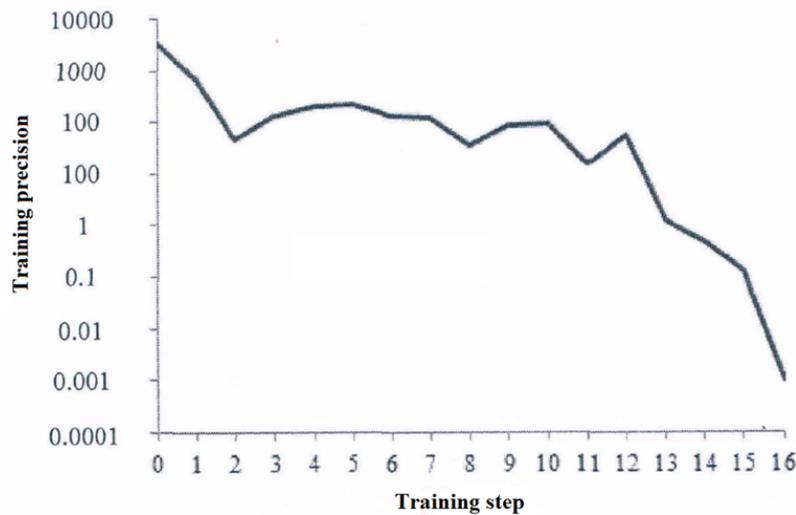


Fig. 10. RBF neural network score prediction model based K-means feature selection

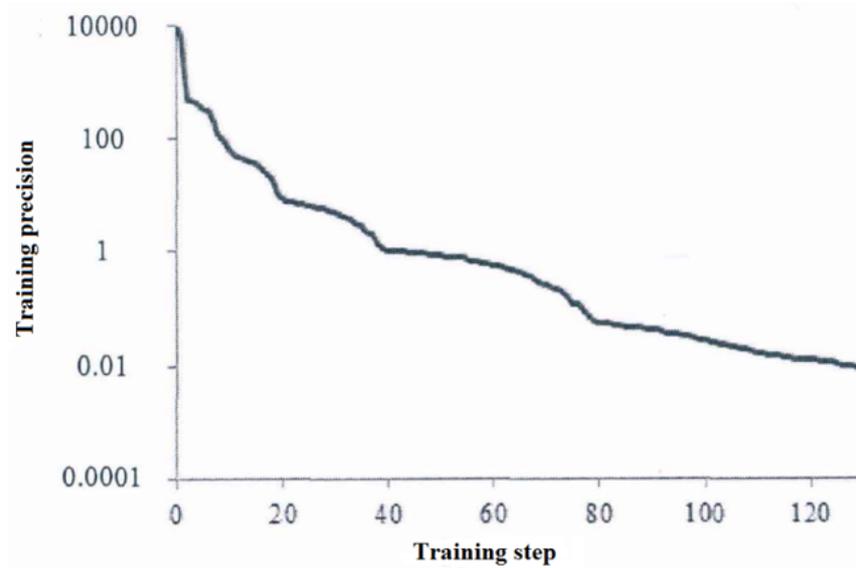


Fig. 11.RBF neural network score prediction model based random centers

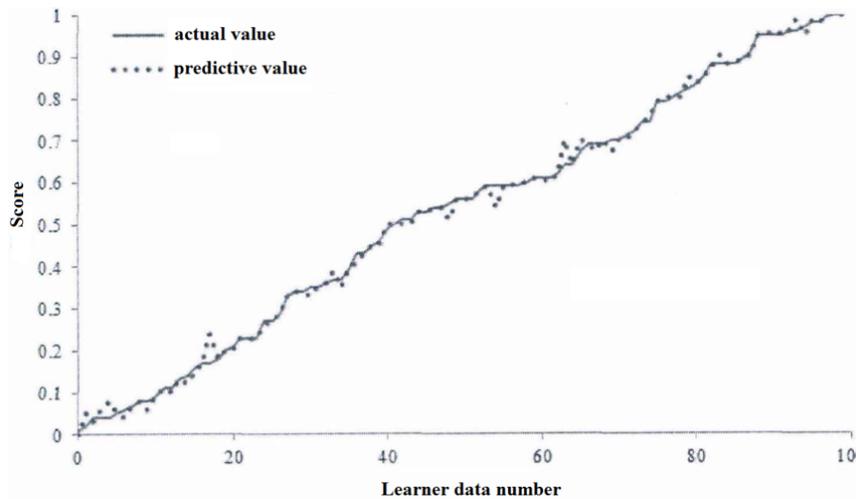


Fig. 12.Comparison of score prediction model based on K-means feature selection

Based on the above experimental analysis, we compares the RBF neural network prediction model with random center points to verify that the proposed score prediction model based on K-Means feature selection has higher accuracy, and it is more accurate in predicting the score of the MOOCs learners.

5 Conclusions

Aiming at learning behavior data for MOOCs learners, we use K-means algorithm to extract the features of learners, and the central point of clustering is used as the central point of neural network, an RBF neural network score prediction model based on K-means feature selection is constructed. By setting the parameters of the prediction model, the prediction model can be updated dynamically to achieve more accurate prediction results.

6 References

- [1] Johnston, T.C. (2014). What Makes a MOOC? Massive Open Online Courses (MOOCs) Compared to Mainstream Online University Courses. *Journal of Learning in Higher Education*, 10: 17-23.
- [2] Hashmi, A.R. (2014). Rapid Growth of Massive Open Online Courses (MOOCs) and the Market for University Graduates. *Social Science Electronic Publishing*, 1-21.
- [3] Gao, J.M., Han, J., Liu, J., Wang, N.N., Dong, S.W. (2014). The Comparison and Analysis of MOOCs Platform---Take edX, Coursera and Udacity platform as an example. *Software Guide*, 12: 178-180.
- [4] Cheng, H.N.H., Liu, Z., Sun, J. (2017). Unfolding online learning behavioral patterns and their temporal changes of college students in SPOCs, *Interactive Learning Environments*, 25(2): 1-13. <https://doi.org/10.1080/10494820.2016.1276082>
- [5] Jiang, Z., Zhang, Y., Li, X. (2015). Learning behavior analysis and prediction based on MOOC data. *Journal of Computer Research & Development*, (3): 614-628.
- [6] Brinton, C.G., Buccapatnam, S., Chiang, M. (2015). Mining MOOC Clickstreams: On the Relationship Between Learner Behavior and Performance. *Computer Science*, 64(14): 1-1.
- [7] Yang, T.Y., Brinton, C.G., Joe-Wong, C. (2017). Behavior-Based Grade Prediction for MOOCs Via Time Series Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(5): 716-728. <https://doi.org/10.1109/JSTSP.2017.2700227>
- [8] Brinton, C.G., Buccapatnam, S., Chiang, M., Poor, H.V. (2015). Mining MOOC Clickstreams: On the Relationship Between Learner Behavior and Performance. *Computer Science*, 64(14): 1-10.
- [9] Brinton, C.G., Buccapatnam, S., Chiang, M., Poor, H.V. (2016). Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance. *IEEE Transactions on Signal Processing*, 64(14): 3677-3692. <https://doi.org/10.1109/TSP.2016.2546228>
- [10] Chen, Q., Chen, Y., Liu, D., Shi, C., Wu, Y. (2016). PeakVizor: Visual Analytics of Peaks in Video Clickstreams from Massive Open Online Courses. *IEEE Transactions on Visualization & Computer Graphics*, 22(10): 2315-2330. <https://doi.org/10.1109/TVCG.2015.2505305>
- [11] Rai, L., Deng, C. (2016). Influencing Factors of Success and Failure in MOOC and General Analysis of Learner Behavior, 6(4): 262-268.
- [12] Fu, S., Zhao, J., Cui, W., Qu, H. (2016). Visual Analysis of MOOC Forums with iForum. *IEEE Transactions on Visualization & Computer Graphics*, 23(1): 201-210. <https://doi.org/10.1109/TVCG.2016.2598444>
- [13] Kennedy, J. (2014). Characteristics of Massive Open Online Courses (MOOCs): A Research Review, 2009-2012. *Journal of Interactive Online Learning*, 13(1): 1-16.

- [14] Hughes, G., Dobbins, C. (2015). The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs). *Research & Practice in Technology Enhanced Learning*, 10(1): 10-21. <https://doi.org/10.1186/s41039-015-0007-z>
- [15] Zeng, X.J. (2014). The Rise of MOOC in US Universities and Its Challenge to Traditional Higher Education. *Comparative Education Review*, 32-40.
- [16] Lai, J.H. (2017). Research of MOOC Mixed Teaching Mode in the Course of Fundamentals of Programming. *Experiment Science & Technology*, 15(5): 77-81.
- [17] Kim, J.Y. (2015). A Study of Perceptual Typologies on Computer Based Assessment (CBA): Instructor and Student Perspectives. *Journal of Educational Technology & Society*, 18(2): 80-96.
- [18] Liu, L.G., Huang, J., Yao, Q., Jia, H.L., Zhou, Y.F. (2010). Research on the Characteristics of Communications Networks Curriculum Construction and the Orientation of Intellectual Training. *Journal of Zhejiang Sci-Tech University*, 13(5): 211-220.
- [19] Jiang, X., Zhuo, J., Chen, H.C. (2015). Empirical study on the important function of mobile interactive model for MOOC. *Global Journal of Computer Sciences*, 5(1): 24-29. <https://doi.org/10.18844/gjcs.v5i1.29>
- [20] Zhuo, C., Dong, X. (2017). An Applicable Way of Teaching Quality Evaluation Based on MOOC Platform. *International Journal of Emerging Technologies in Learning*, 12(3): 57-67. <https://doi.org/10.3991/ijet.v12i03.6421>
- [21] Mou, Z.J., Wu, F.T. (2017). The Exploration of Learning Outcome Prediction Indicators and Analysis of Learning Group Characteristics for MOOC. *Modern Distance Education Research*, (3): 58-66.

7 Authors

Yuan Zhang received her B.Sc. degree in 2009 from social work in Taiyuan University of Science and Technology; M.Sc. degree in 2012 from sociology in the University of Science and Technology Beijing. Her main research interests include teaching management and social security.

Wenbo Jiang received his degree of Master of Fine Arts in 2012 from Yanshan University. After graduation, he has been a lecturer at Hebei University of Environmental Engineering. His main research interests include public art and art education.

Article submitted 17 June 2018. Resubmitted 23 July 2018. Final acceptance 12 August 2018. Final version published as submitted by the authors.