Study on the Portrait of Online Learners' Personality and Attitude

Tao Xu, Maoyang Zou(✉), Zhongyue Fan, Yuxin Chen, Yiran Zhang, and Pan Min

Chengdu University of Information Technology, Chengdu 610225, China
zoumy@cuit.edu.cn

Abstract. In order to help colleges better understand students' learning personality and attitude, better guide students to learn and improve the quality of teaching. This paper uses K-Means, MiniBatchKMeans, and Birch to analyze students' learning personality and attitude. Compared with the three algorithms, we analyze the clustering results of K-means, divide students' learning personalities into 3 categories: "Active", "Normal", and "Dull", and the attitudes of students are divided into four categories: "Negative and lazy", "Perfunctory and active", "Medium-general", and "Proactive". The function model is fitted by multiple linear regression to predict students' scores.

Keywords: Cluster analysis · Learning analysis · Learner profile

1 Introduction

The development of education has entered a new stage of informatization, and it is transforming from digital education to smart education supported by technologies such as data mining and machine learning [1]. In order to achieve the goal of wisdom education, different management requirements should be adopted for students with different personalities and attitudes, and early warming should be given to students who may fail. To meet these needs, this paper collects MOOC student learning behavior data, makes different researches and comparisons on the personality and attitude of students through different machine learning algorithms. Finally, students' learning personality is divided into 3 categories, and students' learning attitude is divided into 4 categories, and through multiple linear regression fitting function model to predict students' scores.

2 Related Work

In terms of clustering algorithms, S. M. Mostafa [2] et al. used nine clustering algorithms to cluster eight datasets, comparing multiple algorithms by seven performance measures. H Cui [3] et al., who proposed a K-means++− based clustering method for social e-commerce users. It is shown that the proposed method can accurately classify social e-commerce users. There are many research results on the analysis of learner behavior data. such as, José A et al. used learner learning data provided by a regional

X. Sun et al. (Eds.): ICAIS 2022, LNCS 13339, pp. 415–424, 2022.
https://doi.org/10.1007/978-3-031-06788-4_35

MOOC provider in Jordan to explore the differences in learners' behavior and preferences. In the end, it was found that the region attracted younger learners, women, and learners with lower levels of education [4]. Juan Zambrano et al. [5] put forward six measures of student performance in the course based on the data provided by the Massachusetts Institute of Technology in online courses, based on these measurement indicators, the student population was divided into multiple categories, and analyzed the resource usage of middle school students in each class. Zeng Shufang [6] and others analyzed MOOC data to extract learners' learning behavior characteristics. Finally, she used Ward's and K-Means clustering to classify learners, which were mainly divided into three categories: "active learners", "passive learners" and "bystanders". The results show that active learners have a higher completion rate and achieve better final grades. Tang Mingwei analyzed students' study behaviors through big data and gave a hidden Markov model. This model establishes the relationship between classroom behavior and student performance, through which the content of the classroom can be adjusted [7]. Zhang Xiaoying applied the K-Means analysis algorithm to classify and analyze the various behaviors of students at school, establishing mathematical model, applied correlation analysis to explain and predict the behavior of college students [8]. Zhang Liyuan [9] and others used data analysis and machine learning methods to research and analyze student behavior, and found that students' online learning performance is highly correlated with learning behavior. Deng Tianping [10] et al. clustered analysis of student learning data and final exam scores on the MOOC platform. Explore the impact of each learning dimension on the learning effect in different class groups. Tian Chunzi [11] and others used K-Means and DBSCAN to analyze multiple types of data generated by students during school, and compared the two algorithms. In terms of score prediction, M. Zaffard [12] et al. proposed a hybrid feature selection framework to predict student performance. Jin Xiuling [13] optimized the SVM model parameters and established the GA-SVM student performance prediction model. Zhao Xiaoyan [14] based on multi-source data fusion technology, fused various data of college students, including sports, consumption and social behavior, and used support vector machines (SVM) and machine learning (ML) to predict college students' English scores. Tian Yu [15] and others proposed a novel multi-feature neural network model to predict college entrance examination scores, and verified the effectiveness of the algorithm through simulation experiments. Li Longzhen [16] uses decision tree C4.5 to establish a student's score prediction model for research, and its prediction accuracy is about 88%. Ren Ge [17] and others used BP neural network to predict students' score in multiple courses, and the prediction accuracy rate could reach 70%. Yu Tiesuo [18] and others used SVR (Support Vector Regression) to predict performance, and used the prediction results for statistical analysis and early warning.

In this paper, different clustering algorithms are used to analyze the personality and attitude of students' behavior data. Comparing the quality of different clustering methods, the K-means clustering results are analyzed for students' learning personality and attitude. Finally, students' scores and results are obtained by multiple linear regression in this paper.

3 Experimental Process and Result Analysis

This paper mainly focuses on the classification of student's personality and attitude and the prediction of student performance by a course in the online learning platform. The process of classifying students' personality and attitude mainly includes:

1. Data acquisition, data cleaning and preprocessing.
2. Classification by K-Means, MiniBatchKMeans, Birch algorithm.
3. Compare the three algorithms and analyze the clustering results of the algorithms.

The predicting process of multiple linear regression is as follows:

1. Select the attribute that has a relatively obvious linear relationship between student behavior data and student performance as the independent variable.
2. Fit the relationship function between independent variables and performance through multiple linear regression.
3. Analyze the function model.

3.1 Data Processing

The data source in this paper is the student behavior data and student basic data of a course on MOOC platform. We extracted the data related to student behavior, including student ID (Id), name (Name), video views (Video), unit detection times (Unit), document reading times (Document), discussion times (Discussion), number of postings (Message), login Number (Login) and final grade (Score). Then we cleaned the data and mainly deleted the students with missing or abnormal field data, and finally left 1685 student data.

3.2 Student Personality Analysis

We use K-Means, MiniBatchKMeans, Birch for cluster analysis, and use the contour coefficient to compare the quality of the algorithm, where the larger the contour coefficient, the better the clustering effect. The contour coefficient is the SC index, which indicates the degree of aggregation within each cluster and the degree of separation between each cluster after clustering. The smaller the distance between samples in the same class, the larger the sample distance between different classes [19] 16, the larger the value of SC, the better the clustering effect will be. Therefore, SC is often used as a performance index to evaluate the clustering results. We let a_i represent the average distance between sample i and other samples in the cluster, and b_i represents the average separation distance between each cluster. Then we can use the following formula to calculate the contour coefficient SC_i.

$$SC_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{1}$$

First, we select Document and Discussion in the data, and standardize the data. The behavioral data of students selected by us are analyzed for their personality through

K-Means, MiniBatchKMeans, and Birch. In addition, the curve changes of the three algorithms' classification cluster number = 2, 3, 4, 5, 6, 7, 8 and profile coefficient are plotted. In the graph, blue represents K-Means, orange represents Birch, and green represents MiniBatchKMeans (Fig. 1).

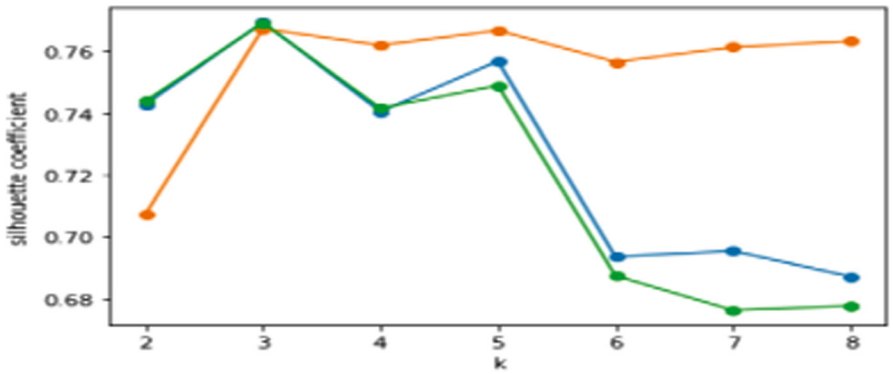

Fig. 1. The curve change of the cluster number and silhouette coefficient of the three algorithms of student personality.

The clustering results show that when the K-Means clustering result is optimal, the students are divided into 3 categories at this time, and the contour coefficient at this time is 0.7695. When the MiniBatchKMeans algorithm is optimal, the students are divided into 3 categories, and the contour coefficient is at this time. It is 0.7692. When the result of Birch algorithm is optimal, the students are divided into 3 categories. At this time, the contour coefficient is 0.7668. The results show that when the three algorithms have the best clustering effect, students are divided into three categories. We number the three

Table 1. K-Means cluster analysis of student personality.

	personality0	personality1	personality2
Accuracy	99.14%	88.24%	93.10%
Recall	98.30%	90.90%	94.67%
F1	98.72%	89.55%	93.88%

Table 2. Birch cluster analysis of student personality.

	Personality0	personality1	personality2
Accuracy	99.06%	75.42%	92.56%
Recall	98.3%	89.90%	90.31%
F1	98.68%	82.03%	91.42%

Table 3. MiniBatchKmeans cluster analysis of student personality

	personality0	personality1	personality2
Accuracy	99.14%	84.76%	92.87%
Recall	98.3%	89.90%	93.76%
F1	98.72%	87.26%	93.31%

personalities as 0, 1, and 2, and analyze the three clustering algorithms, as shown in the following Tables 1, 2 and 3:

3.3 Student Attitude Analysis

The realization of student attitude analysis is similar to personality. First we select Video, Unit, Document, Message, Login from the data set as the original clustering data set, and standardize the data, then use the principal component analysis method to transform the data into 2 dimensions. Principal Component Analysis, a method of processing data [20], converts high-dimensional data containing a large amount of redundant information into a small amount of low-dimensional data, and contains the effective information of the original data. Its basic idea is to find a projection transformation matrix that best represents the main personalityistics of the original data under the constraint of the minimum mean square error [21]. Then K-Means, MiniBatchKMeans, and Birch are used to analyze student attitudes based on our selected student behavior data. The curves of cluster number = 2, 3, 4, 5, 6, 7, 8 and contour coefficient are drawn. In the graph, blue represents K-Means, orange represents Birch, and green represents MiniBatchKMeans (Fig. 2).

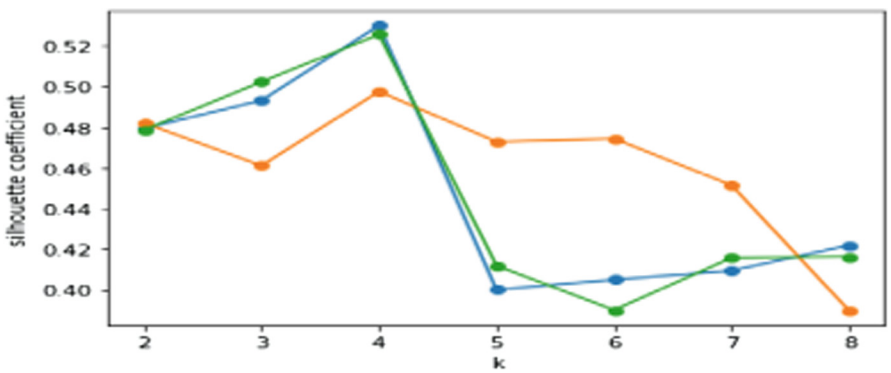

Fig. 2. The curve change of the cluster number and silhouette coefficient of the three algorithms of student attitude.

The clustering results show that when the K-Means clustering result is optimal, the students are divided into 3 categories at this time, and the contour coefficient at this time

Table 4. K-Means cluster analysis of student attitude.

	attitude0	attitude1	attitude2	attitude3
Accuracy	95.99%	92.20%	73.77%	93.15%
Recall	94.29%	82.42%	89.51%	88.31%
F1	95.13%	87.04%	80.88%	90.67%

Table 5. Birch cluster analysis of student attitude

	attitude0	attitude1	attitude2	attitude3
Accuracy	96.06%	85.81%	67.15%	93.65%
Recall	94.07%	75.15%	87.27%	72.83%
F1	95.05%	80.13%	75.90%	81.94%

Table 6. MiniBatchKmeans cluster analysis of student attitude

	attitude0	attitude1	attitude2	attitude3
Accuracy	96.05%	83.25%	54.44%	86.75%
Recall	93.77%	48.18%	87.27%	88.89%
F1	94.90%	61.04%	67.05%	87.81%

is 0.7695. When the MiniBatchKMeans algorithm is optimal, the students are divided into 3 categories, and the contour coefficient is at this time. It is 0.7692. When the result of Birch algorithm is optimal, the students are divided into 3 categories. At this time, the contour coefficient is 0.7668. The results show that when the three algorithms have the best clustering effect, students are divided into three categories. We number the three personalities as 0, 1, and 2, and analyze the three clustering algorithms, as shown in the following Tables 4, 5 and 6:

3.4 Student Performance Prediction

Student performance prediction is to use multiple linear regression to fit a function and then predict the performance. Multiple linear regression analysis forecasting method refers to the establishment of a forecasting model through the correlation analysis of two or more independent variables and dependent variables. When there is a linear relationship between the independent variable and the dependent variable, it is called multiple linear regression analysis [22]. One of the significance test methods of the regression equation is to test by Multi-correlation coefficient. When the result of Multi-correlation coefficient is closer to 1, the better the correlation fitting effect will be [23].

The calculation formula of Multi-correlation coefficient is:

$$R = \sqrt{\frac{\Sigma\left(\hat{y} - \bar{y}\right)^2}{\Sigma(y_i - y)^2}} \tag{2}$$

Through our analysis of each attribute and score, we found that the linear relationship between Video, Unit, Document and Score is relatively high. We calculated the Pearson correlation coefficients between them through SPSS software, which were 0.894, 0.935, and 0.937, respectively. So we finally choose Video, Unit, Document as the independent variables, and the linear relationship between these three and the score is as follows (Fig. 3):

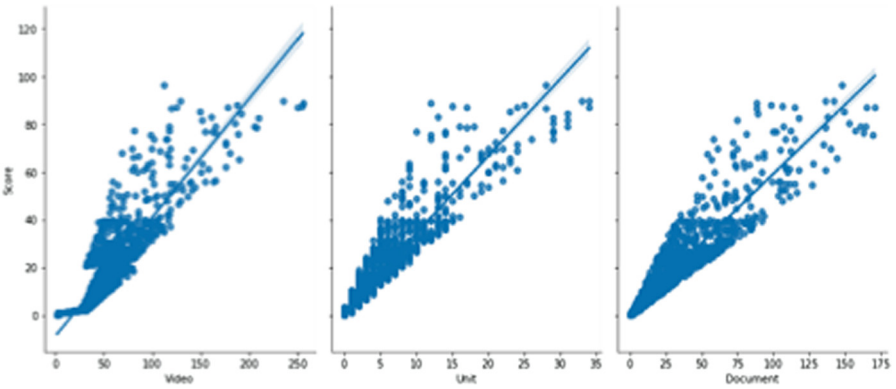

Fig. 3. Linear relationship diagram of video, unit, document and score.

In the second step, we use Video, Unit, Document as independent variables and score as dependent variables, and finally use SPSS software to perform multiple linear regression to obtain the function model: $y = 0.141 * x_1 + 1.335 * x_2 + 0.239 * x_3 - 2.545$.

3.5 Result Analysis

For student personality analysis, comparing the K-Means, Birch, and MiniBatchKMeans algorithms, the three algorithms divide the students' learning personality into three categories when the contour coefficient of the three algorithms is the largest. The three algorithms found the best for personality 0, personality 2 the second, and personality 1 the worst. Among the three algorithms, the Birch algorithm has a lower clustering accuracy than the other two algorithms. We analyze the three personalities through the clustering results of K-means. As shown in Table 7. We can see from the table that the behavioral data values of students in category 1 are relatively small, students in category 2 have the largest value, and students in category 3 are in the middle. We divide students into three categories: "active", "ordinary", and "dull". Among them, the active type is more enthusiastic about things, the normal type is positive and indifferent to things, and the dull type is introverted and indifferent to things.

Table 7. Student personality clustering results.

Type	Message center	Discussion center	Total number
1	−0.501637	−0.596167	1163
2	2.820511	2.681636	102
3	0.704075	0.999560	420

Table 8. Model summary.

Model	R	R square	Adjusted R square
1	0.974	0.949	0.949

Table 9. ANOVA.

Model	Sum of square	df	Mean Square	F	sig
1	Regression	467470.013	3	155823.338	10443.241
	Residual	25082.160	1681	14.921	
	Total	492552.173	1684		

Table 10. Coefficients

Model		Unstandardized coefficients		t	sig
		B	Std. Error		
1	Constant	− 2.545	0.188	− 13.517	< 0.001
	Video	0.141	0.006	23.989	< 0.001
	Unit	1.335	0.043	31.225	< 0.001
	Document	0.239	0.008	30.421	< 0.001

For the analysis of student attitudes, when the contour coefficients of the three algorithms are the largest, students' learning attitudes are divided into 4 categories. The three algorithms found the best for attitude 0 and the worst for attitude 2. For attitude 0, the Kmeans algorithm has the lowest accuracy, and for student learning attitude 1, attitude 2 and attitude 3, MiniBatchKmeans has the lowest accuracy. We draw the result of kmeans algorithm into a radar chart, as shown in Fig. 4. We can draw the following conclusions: the first type of students can be summarized as "negative and laziness" and their attitudes are: negative learning, laziness, and even giving up learning. The second class of students is "perfunctory and active" and their attitudes are: perfunctory, positive comments, and easy going. The third type of students is the " Medium-general " whose performance is

"medium grades", "average enthusiasm", and "sloppy". The fourth category is "proactive", which is manifested as "good scores", "high motivation", and "active learning". Among them, there are 993 people for "passive and lazy", 295 people for "perfunctory and active", 324 people for "Medium-general", and 73 people for "proactive".

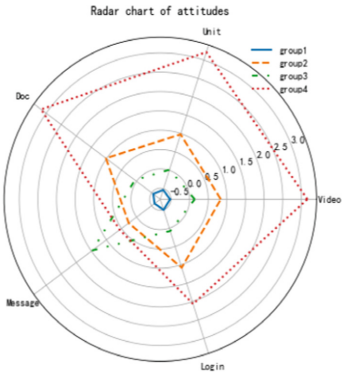

Fig. 4. Radar chart of student attitudes.

According to the function model obtained above, we only need to know that Video, Unit, and Discussion can predict students' performance. In order to evaluate the quality of the model, we use SPSS to test and evaluate the model. The evaluation results are as follows (Tables 8, 9 and 10) :

Through analysis, we can see that the coefficients corresponding to the number of final video views, detection times, and document reading obtained by using the multiple linear regression are 0.141, 0.1335, 0.239, respectively, the constant term is −2.545, and the significance of each independent variable is less than 0.001. The description shows that the influence of each independent variable on the dependent variable is significant. At the same time, the multi-correlation coefficient R of the model is 0.974 close to 1 and the significance is less than 0.001, indicating that the fit is good.

4 Summary

In his paper, we clear and preprocess students' behavior data, then use three clustering algorithms to classify students' learning personality and attitude and compare the results of the three algorithms. Finally, we select the K-Means algorithm to cluster the data and analyze the model. The personality of students is divided into three categories: "active", "ordinary", and "boring". Students' attitude is divided into four categories: "negative and lazy", "perfunctory and active", "medium-general" and "Proactive". Then the multiple linear regression algorithm is used to predict the student's performance and test the model. The study of student behavior data can provide targeted suggestions for future teaching practice, and can also provide a theoretical basis for continuous improvement of teachers' classroom teaching [10].

References

1. Cheng, P.: Learner personalized modeling and user portrait system for online education platform (2019)
2. Mostafa, S.M.: Clustering algorithms: taxonomy, comparison, and empirical analysis in 2d datasets. J. Artif. Intell. **2**(4), 511–524 (2020)
3. Cui, H.: A k-means++ based user classification method for social e-commerce. Intell. Autom. Soft Comput. **28**(1), 277–291 (2021)
4. Ruipérez-Valiente, J.A., Halawa, S., Slama, R., Reich, J.: Using multi-platform learning analytics to compare regional and global MOOC learning in the Arab world. Comput. Educ. **146**, 103776 (2020)
5. Ruipérez-Valiente, J.A., Halawa, S., Slama, R., Reich, J.: Using multi-platform learning analytics to compare regional and global MOOC learning in the Arab world. Comput. Educ. **146**, 103776 (2018)
6. Tseng, S.F., Tsao, Y.W., Yu, L.C., Chan, C.L., Lai, K.R.: Who will pass? analyzing learner behaviors in MOOCs. Res. Pract. Technol. Enhanc. Learn. **11**(1), 1–11 (2016)
7. Tang, M., et al.: Research on students' classroom behavior based on big data analysis and hidden Markov model. J. Phys: Conf. Ser. **1873**(1), 012084 (2021)
8. Zhang, X.: Research on the behavior model of college students based on big data analysis. Electron. Technol. Softw. Eng. **2**(2), 1–5 (2021)
9. Zhang, L.: Research on student online learning behavior based on data analysis. J. Yuzhang Norm. Coll. **2**(2), 87–91 (2021)
10. Deng, T.: Analysis of student learning behavior based on Mu classroom data. Journal **2**(2), 78–82 (2020)
11. Tian, C.: Analysis and research of student behavior based on comprehensive data of colleges and universities based on K-Means and DBSCAN clustering algorithm. Journal **2**(32), 86–88 (2021)
12. Zaffar, M.F.: A hybrid feature selection framework for predicting students performance. Comput. Mater. Continua **1**(70), 1893–1920 (2022)
13. Jin, X.: Prediction of college entrance examination results based on genetic algorithm and support vector machine model. Journal **2**(2), 62–65 (2020)
14. Zhao, Y.: Prediction of English scores of college students based on multi-source data fusion and social behavior analysis. Journal **2**(4) (2020)
15. Tian, Y.: College entrance examination score prediction based on multi-feature perception network. Journal **2** (2021)
16. Li, L.: Online learning performance prediction based on decision tree algorithm. Journal **2**(1), 130–133 (2021)
17. Ren, G.: Application of BP neural network in early warning of college student performance. Journal **2**(10), 53–55 (2020)
18. Yu, T.: Research on the application of SVR regression in performance prediction and early warning. Journal **2**(11), 76–80 (2020)
19. Ma, X.: Principal component analysis face recognition algorithm based on BP neural network. Journal **2**(1), 140–146 (2021)
20. Zhao, L.: Analysis and research of student behavior based on comprehensive data of colleges and universities based on K-Means and DBSCAN clustering algorithm. Journal **2**(36), 226–229 (2007)
21. Ma, J.: Research on big data visualization application based on pca dimensionality reduction. Journal **2**(2), 201–206 (2021)
22. Chen, K.: Analysis of college English band 4 scores based on multiple linear regression. Journal **2**(10), 37–39 (2018)
23. Zhang, X.: Application of multiple linear regression in analyzing student performance ranking prediction. Journal **2**(5), 154–160 (2018)