Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **1**, January **2019**

Clustering Research Based on Feature Selection in The Behavior Analysis of MOOC Users

Lin Xiao[1,2]

[1]College of Information Science and Engineering,
Fujian Provincial Key Laboratory of Big Data Mining and Applications,
Fujian University of Technology, Fuzhou, 350118, China

[2]College of Information Management,
Jiangxi University of Finance and Economics Nanchang
Jiangxi,330013, China
*Corresponding author: xiaolin201@qq.com

Received July 2018; revised August 2018

Abstract. *This paper proposes a K-Means feature selection algorithm for clustering analysis of MOOC users. The algorithm is divided into three steps. First, the weight calculation method is designed to select the important characteristics according to the weight, and the two is to optimize the algorithm of the initial cluster center; and the three is to design the equilibrium discriminant function to determine the optimal number of clustering. Finally, by comparing the experimental results with the other two traditional algorithms, the efficiency of the K-Means feature selection algorithm is verified in three aspects, such as the algorithm running time, the average iteration number and the clustering accuracy rate. The algorithm in this paper is an important step in the analysis of user behavior. Its clustering results are the input of the next user performance prediction module.*
Keywords: MOOC; User Behavior; Cluster; Feature Selection.

1. **Introduction.** The popularity of MOOC began in 2012. The full name of "MOOC" is Massive Open Online Course. These four words fully reflect the connotation and characteristics of MOOC, including large-scale, open, online status and curriculum which are not limited by time and space [1]. To sum up, MOOCS are open courses that are distributed across the Internet in order to enhance knowledge dissemination by individual organizations with the spirit of sharing and collaboration.

However, although there are many advantages, but there are many cases of much criticism, such as they have a large number of registered users, but few of the final completion of the course users; and it has a high dropout rate [2]; and there is a lack of interaction between the users and the teachers, and the teachers can not effectively provide learning guidance for each user.

In view of the above problems, a solution is to analyze the learning behavior data of MOOCs users, and find ways to overcome the limitations of MOOC. In the case of user behavior data, comprehensive discipline knowledge is used to analyze and excavate the characteristics of the user and the law of learning behavior, and these rules are applied to the practice [3]. The main purpose of the study is to pursue the maximum teaching benefit, and to plan the corresponding learning content and learning strategy according

to the individual attributes of different users' knowledge level, character goal, interest and so on.

2. **Problem Statement and Preliminaries.** The analysis of the behavior of MOOC user is to find out the information that is ignored by the user and the hidden learning rules in the learning process. The research results can provide personalized guidance and learning strategy advice for each user and improve the efficiency of learning [4-5]. At the same time, it should warn the users who are less successful and have the risk of dropping out. And the teachers can improve their teaching methods and provide users with more reasonable counselling and suggestions based on the analysis and prediction results.

The content of this paper is clustering algorithm research, which is an important part of the whole research. Its main contents include three points. The first is the analysis of MOOC users' learning behavior data. The two is to select the feature by weight calculation. The three is to improve the clustering algorithm and verify the effectiveness of the algorithm. The next step is the performance prediction module, but the results of this study can serve as the input for the next user performance prediction. Only when these two parts are completed can we support the evaluation and learning strategy recommendations of the whole user behavior analysis.

2.1. **Data sources.** The data of MOOC user behavior will be generated through interaction between users and teachers, users and learning resources, users and users. Because the data contain many privacy information, in order to protect user's privacy, a series of data preprocessing is needed, including hiding identity information [6]. After anonymity processing, some open data sets have been released internationally for global research and sharing. In this paper, Canvas data set is selected as the research object. The Canvas open data-set is an aggregate of more than 320 thousand anonymous data from the Canvas Network Open Courses platform from January 2014 to September 2015 [7]. Table 1 enumerates some of the feature attributes contained in the Canvas data-set.

2.2. **preliminary analysis.** (1) the basic information of the user. User age ranges on Canvas Network platform range from 34 to 54 years old, and the major user groups have bachelor's degree and master's degree. This is in line with the characteristics of MOOC users who are mostly busy with all kinds of needs. The research on the reasons of user participation shows that the purpose of user learning will affect the tendency of users to choose courses.

Table 1. characteristic attributes of the Canvas data set

Serial	Characteristic ID	Meaning	Field type
1	*Course_id_DI*	Course ID	numerical
2	Discipline	Discipline	string
3	*Userid_DI*	User ID	numerical
4	Grade	User grade	numerical
5	Completed	Course completed	numerical
6	*LoE_DI*	Basic level of user education	string
7	*Age_DI*	User age	numerical
8	Nevents	Course interaction times	numerical
9	*Ndays_act*	Number of interactive days in the course	numerical
10	*Nforurn_posts*	Number of speeches in the Forum	numerical
11	*Course_length*	Course length	numerical
12	*Primary_reason*	Cause of participation	string
13	*Expected_hour_week*	Number of expected days per week	string

(2) user type. This article divides user types into active users, passive users and negative users. In the analysis process, we found that passive users account for only 26.36% of the total, while active users account for less than 3.99%.

(3) the factors that affect the performance. Many factors have a certain relevance to the performance, including the number of courses, the number of intercourse days, the number of course chapters, the number of forum speeches, and the length of the course. The number of speeches in the forum was positively correlated with the scores, and the length of courses was negatively correlated with the scores.

3. **The main process of the algorithm.** Based on the high sensitivity of K-means algorithm and the applicability to high-dimensional data, this paper chooses to improve K-means for data clustering analysis of MOOC users. First, feature selection method is used to select the first w features with the greatest weight (i.e. the greatest contribution to clustering). The two is to use the optimized algorithm to select the optimal initial cluster center. The three is to use the balanced discriminant function to select the optimal number of clusters. When the equilibrium discriminant function converges, the function number and the clustering number are recorded by the array. The number of clusters corresponding to the minimum function value is the optimal clustering number. The whole idea is to use the method of feature selection to reduce the characteristics of high dimensional data and then to cluster. It has a significant effect on solving the problems of low precision and high timeliness of high dimensional data clustering. The specific steps of the K-means feature selection algorithm are as follows.

The first step, initializes the weight value of each feature.

Second step, it uses feature selection algorithm to update and output feature weight vector W.

Third step, According to the descending order of weights, select the top m features with the greatest weight value.

Fourth step, using the optimal selection algorithm of initial clustering centers to get the optimal K initial clustering centers.

Fifth step, data objects are partitioned according to Euclidean distance, and each data object is partitioned into clusters belonging to the nearest distance center.

Sixth step, update the center of each class cluster.

Seventh step, judging whether the equilibrium discriminant function is convergent.

Eighth step, if the function converges, the algorithm ends; otherwise, go to the fifth step and continue the iteration.

Hypothesis data set $s = \{s_1, s_2, ..., s_m\}$, the number of characteristics of each data object is p. That is,$s_i = \{s_{i1}, s_{i2}, ..., s_{ip}\}$,$1 <= i <= m$. The category of s_i is c_i, and $c_i \in C$, $C = \{c_1, c_2, ..., c_k\}$, C is a collection of k categories. The method first selects a data object s_i from the data set, and then selects d data objects from each category, which are closest to s_i. Among them, the d data objects of the same category as S constitute the set H(c). In addition, different types of data objects with d form a collection M(c) according to their categories. According to set H(c) and M(c), the weight vector W ($W = \{w_1, w_2, ..., w_p\}$) is updated. The weight of the characteristic t ($1 <= t <= p$) is calculated as shown in formula (1).

$$w_t^{i+1} = w_t^i - \sum_{x \in H(c)} \frac{diff(t, s_i, x)}{(n \times d)} + \sum_{c \in class(s_i)} [\frac{p(c)}{1 - p(class(s_i))} \sum_{x \in M(c)} diff(t, s_i, x)]/(n \times d) \quad (1)$$

Among them, n is the number of sampling times. The function (t, s_i, s_j) is the difference function between the data object s_i and s_j $(1 <= i \neq j <= m)$on the characteristic t. As shown in formula (2) and (3).

If the features t is continuous, A and B are the minimum and maximum values of the characteristic t in the data set, then (t, s_i, s_j) is as follows in formula (2).

$$diff(t, s_i, s_j) = |\frac{s_{it} - s_{jt}}{max_t - min_t}| \tag{2}$$

If the feature t is discrete, then $diff(t, s_i, s_j)$ is as follows in formula (3).

$$diff(t, s_i, s_j) = \begin{cases} o & \text{if } s_{it} = s_{jt} \\ 1 & \text{if } s_{it} \neq s_{jt} \end{cases} \tag{3}$$

4. **Control design.**

4.1. **Feature selection based on weight.** According to the above method, the weight vector W is obtained, which is arranged in descending order according to the size of the weight value, and the first m eigenvectors of the maximum weight value constitute the final feature subset. The method selects the nearest neighbor of a feature based on the distance between the data objects. The selection of the nearest neighbor's choice is associated with the accuracy of the weight value of the characteristic attribute. The larger the weight of the feature, the stronger the ability of distinguishing the data object from the feature, and the greater the contribution to the clustering. On the contrary, it shows that the feature has less ability to distinguish data objects and less contribution to clustering.

4.2. **Optimization of initial cluster center.** Suppose that X is a set of data with N data objects. $X = \{x_1, x_2, ..., x_n\}$. Each data object contains p features, that is $X_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$. Now it plans to divide the data set X into k class clusters. The class cluster is described as $C_j(j = 1, 2, ..., k, k < n)$. Then, The $j(1 <= j <= p)$ feature attribute of the $i(1 <= i <= n)$ data object can be defined as X_{ij}.

Define: The Euclidean distance between arbitrary data objects X_i and X_j $(1 <= i \neq j <= n)$ is defined as follows in formula (4).

$$d(x_i, x_j) = \sqrt{\sum_{a=1}^{p}(x_{ia} - x_{ja})^2} \tag{4}$$

Define:The distance density function $density(x_i)$ corresponding to the data object $x_i(1 <= i <= n)$ in the data set is defined as follows in formula (5).

$$density(x_i) = \sum_{j=1}^{n} \frac{d(x_i, x_j)}{\sum_{l=1}^{n} d(x_l, x_j)} \tag{5}$$

Define: The neighborhood radius R_i of data object $x_i(1 <= i <= n)$ in data set is defined as follows in formula (6).

$$R_i = n^{cR} \times \frac{1}{n} \sum_{i=1}^{n} e^{-density(x_i)} \tag{6}$$

Among them, $cR(0 <= cR <= 1)$ is the neighborhood radius adjustment coefficient. According to experience, when cR was equal to 0.13, it had a good clustering effect.

Suppose that there is a data object $x_i(1 <= i <= n)$ in the dataset X, then take x_i as the center of a circle, within the neighborhood of x_i, the number of data objects in the spherical domain with a radius of R_i is the point density of x_i, which is written as $D(x_i)$. Its formula is shown in (7). The larger the $D(x_i)$ value is, the higher the density of the spherical region of the data object x_i is.

$$D(x_i) = |\{p|d(x_i, p) \leq R_i, p \in X\}| \tag{7}$$

Suppose the density average value of all data objects in the dataset is $MD(x)$, the formula $MD(x)$ is shown in (8).

$$MD(x) = \frac{1}{n} \sum_{x \in X} D(x) \tag{8}$$

The first step of optimizing the initial cluster center algorithm is to calculate the distance between every two data objects in the data set according to formula 4, so as to construct the distance matrix D, then calculate the point density $D(x_i)$ of each data object, and the average density value $MD(x)$ of the data set. The second step is to compare the point density $D(x_i)$ and $MD(x)$ of each data object and divide the data object greater than or equal to $MD(x)$ to the set M. The third step is to select the data object with the highest density in set M as the first initial clustering center C_1 and add it to C, that is, $C = C \cup \{C_1\}$. The fourth step is to select the data object whose distance from C_1 is larger than its neighborhood radius R_1 and its point density is only next to C_1 from the data set M. This data object is used as the second initial clustering center C_2, and it is added to C, that is, $C = C \cup \{C_2\}$. By this way, until the k initial cluster centers are selected.

4.3. **Design of equilibrium function.** When a data set is clustered, the criterion function is generally used to discriminate and determine whether there exists similar data objects in the data set. The criterion function is also called the objective function, and the clustering algorithm determines whether the similarity degree in the class cluster can be maximized by calculating the objective function, and whether the degree of dissimilarity between the clusters tends to maximization. Based on this idea, this paper selects a balanced discriminant function as a criterion to detect the differences within clusters and the differences among clusters in cluster C. It effectively balances the inconsistency between the differences within clusters and the differences among clusters, and improves the overall quality of clustering. When the value of the equilibrium discriminant function reaches the minimum, the clustering result and the optimal clustering number can be obtained under the optimal conditions.

Assuming that the data set X and collection C are expressed as such, $X = \{x_1, x_2, ..., x_n\}$, $C = \{c_1, c_2, ..., c_n\}$. The set C is a set of K classes, where $c_i(1 <= i <= n)$ is the center of the i class.

The intra cluster difference is a measure of the compactness of clustered clusters. It calculates the square sum of the distance from each data object in the cluster to the center of the cluster, as shown in formula (9).

$$w(c) = \sum_{i=1}^{k} w(c_i) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)^2 \tag{9}$$

The difference between clusters is obtained by calculating and judging the Euclidean distance between cluster centers. Assuming that c_i and c_j are the centers of the No.i cluster and the No.j cluster respectively, then the difference is named $b(c)$ between the two clusters as shown in formula (10).

$$b(c) = \sum_{1 \le j \le i \le k} d(c_j, c_i)^2 \tag{10}$$

This paper presents a balanced discriminant function, as shown in formula (11), in which the difference w(c) within the cluster and the difference b(c) between the clusters need to be normalized first, and the k is the number of clustering.

$$W(c,k) = \frac{1}{1 + e^{b(c)-w(c)}} \tag{11}$$

5. **Simulation experiment and result analysis.** In order to verify the effectiveness of the algorithm, the experimental environmental conditions selected include: Window7 experimental platform, Matlab language programming, 2GB memory. Based on a preliminary analysis of the Canvas dataset, the main features used for feature selection in the data set are selected, which is illustrated in table 2. Feature *last_start_time* is the difference between characteristic *start_time_DI* and *last_event_DI*, which means the time difference between the first course interaction and the last course interaction of the user .

5.1. **An experimental analysis of the method of feature selection.** The weight value of each feature is calculated by the method of feature selection above, and the order output is descended in descending order. In view of the actual situation of this article, the 5 characteristics of the maximum weight value are selected. Because the algorithm needs to select a random data object named R in the running process, the difference of R will lead to a certain discrepancy in the weight value. Therefore, this paper adopts the average value method, setting the number of operation of the algorithm is 10, and calculates the average value of each feature weight value. As shown in table 3, the column of the form is the characteristic number, and the row of the form is the average of each computation result.

As shown in Figure 1, the weight values of the features are arranged in descending order. The weight relationships of each feature are as follows: $feature2 > features8 > features6 > features4 > features1 > features9 > features7 > features5 > features3$. According to the foregoing setting, this paper selects 5 features with the greatest weight value, so excluding feature 9, feature 7, feature 5 and feature 3.

Table 2. Main Features for Feature Selection

Feature number	Feature name
1	*Grade_reqs*
2	nevents
3	*course_length*
4	*ndays_act*
5	*last_start_time*
6	*nforum_posts*
7	ncontent
8	*Completed_persent*
9	explored

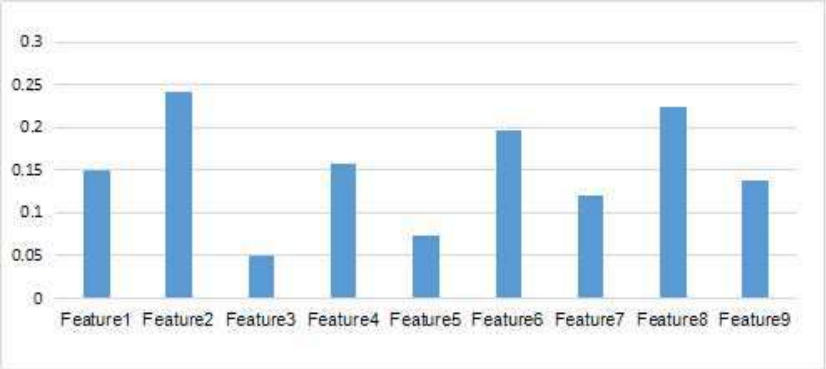

FIGURE 1. Average weight of each Feature

The analysis shows that the frequency of the course interaction (feature 2) is the characteristic attribute of the maximum weight value, which explains the clustering results of the maximum degree, followed by the degree of curriculum completion (feature 8) and the number of forum speakers (feature 6).

5.2. **Experimental analysis of K-Means feature selection algorithm.** In this paper, three algorithm experiments are carried out on Canvas open dataset to compare the results. The three algorithm is the K-Means feature selection algorithm, the traditional K-Means algorithm and the density based K-Means algorithm. The experimental results are analyzed and compared from three aspects, such as the algorithm time, the average number of iterations of the algorithm, and the accuracy of algorithm clustering. All experimental results show that the equilibrium functions are convergent.

(1)Comparison of algorithm time

The running time of the three algorithms is shown in table 4. From the graph, we can see that the K-Means feature selection algorithm has better stability, but it takes much more time in running time than the traditional algorithm. Because it takes a lot of time in feature selection and optimization to select the initial cluster center stage. The density based method runs longer because the algorithm is usually time-consuming when selecting the initial cluster center.

(2)comparison of the average iteration number of the algorithm

Running on the same data set, the average iteration number of the traditional K-Means algorithm is 7.9, the average iteration number of the density based K-Means algorithm is 6.4, and the average iteration number of the K-Means feature selection algorithm is 4.3. This is because the K-Means feature selection algorithm uses the method of selecting the initial cluster center on the basis of feature selection, which reduces the number of

TABLE 3. Average Features Weighted

Running times	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9
1	0.1406	0.2123	0.0668	0.1641	0.0821	0.1944	0.1083	0.2245	0.1120
2	0.1488	0.2491	0.0467	0.1240	0.0870	0.1724	0.1210	0.2083	0.1470
3	0.1535	0.2604	0.0359	0.1693	0.0679	0.2012	0.1035	0.2226	0.1285
4	0.1865	0.2337	0.0679	0.1434	0.0757	0.2219	0.1280	0.2302	0.1694
5	0.1554	0.2391	0.0564	0.1703	0.0604	0.2054	0.1335	0.2038	0.1424
6	0.1460	0.2366	0.0319	0.1642	0.0738	0.1995	0.1299	0.2075	0.1510
7	0.1439	0.2351	0.0530	0.1616	0.0812	0.2062	0.1035	0.2219	0.1366
8	0.1443	0.2399	0.0347	0.1490	0.0785	0.1913	0.1224	0.2507	0.1205
9	0.1483	0.2615	0.0470	0.1538	0.0672	0.1870	0.1314	0.2516	0.1318
10	0.1315	0.2517	0.0536	0.1707	0.0703	0.1946	0.1232	0.2120	0.1373

iterations. Experiments show that the K-Means feature selection algorithm can effectively reduce the number of iterations, and its execution efficiency is higher than the density based K-Means algorithm.

(3)Comparison of accuracy of clustering algorithm

The traditional K-Means algorithm has different results in each operation, the accuracy rate is low and unstable. On the contrary, the clustering accuracy of the K-Means feature selection algorithm is relatively high, and there is no fluctuation with the change of the number of running times. As shown in table 5, it shows that the algorithm has good clustering results and has a certain application value. This is because the feature is selected according to the feature weight value, and the unrelated features are eliminated, so the similarity between the data objects is closer to the actual situation, thus improving the accuracy of clustering.

6. **Conclusion.** In this chapter, a K-Means feature selection algorithm is proposed. Firstly, the feature selection method is used to select the first few features with the greatest weight. Secondly, the selection of the initial cluster center is optimized. The high density set is obtained by comparing the density of data objects and the average density of data sets. In the high-density set, the neighborhood radius is used to separate the data and determine the initial cluster center. Thirdly, the weight value is used as the basis to measure the contribution of feature attributes. In the iterative process of clustering, the criterion is that the similarity of objects within clusters is high and the difference among clusters is as large as possible. Finally, according to a large number of experiments, the traditional K-Means algorithm, the density based K-Means algorithm and the K-Means feature selection algorithm are compared and analyzed, and the high efficiency of the K-Means feature selection algorithm is verified from the three aspects of the algorithm running time, the average iteration number and the clustering accuracy rate. Although it is not until the next performance prediction module is completed, it can support the overall evaluation and learning strategy recommendations of the whole user behavior analysis. But the results of this study can be used as the input of the next

Table 4. Algorithm Comparison of Running Time

Running time	The traditional K Means	The density based K-Means	The K-Means feature selection algorithm
1	1.1032	1.8597	1.5107
2	1.1067	1.5904	1.5269
3	1.0801	1.5348	1.5649
4	1.1698	1.5251	1.6305
5	1.0231	1.5186	1.5119
average	1.0966	1.6057	1.5107

Table 5. Algorithm Comparison of Clustering Accuracy

Running time	The traditional K Means(%)	The density based K-Means(%)	The K-Means feature selection algorithm(%)
1	69.9	77.5	91.3
2	63.2	69.2	90.2
3	61.6	73.1	91.5
4	54.1	67.7	89.6
5	74.8	73.8	89.7
average	64.7	72.2	90.5

user performance prediction module, and it is also an important step in the analysis of MOOC users behavior.

Acknowledgment. Acknowledgment. This work is partially supported by The National Natural Science Foundation Project (71561010) and Fujian Province Education Planning projects (FJJKCG15-051) and Jiangxi Province Education Planning Project(15YB023). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] K. H. Xu, The Connotation and Characteristics of MOOC and Its Enlightenments to Lifelong Education in China, Vocational and Technical Education. (2014) 227-231.

[2] J. S. Pan, T. T. Nguyen, T. K. Dao, et al. Clustering Formation in Wireless Sensor Networks: A Survey, *Journal of Network Intelligence*, vol. 2, no. 4, pp. 287-309, 2017.

[3] B. Myroniv, C. W. Wu, Y. Ren, et al. Analyzing User Emotions via Physiology Signals, *Data Science and Pattern Recognition*, vol. 1, no. 2, pp. 11-25, 2017.

[4] C. Grainne , W. Sandra. Representing learning designs-making design explicit and shareable, *Educational Media International*, vol. 50, no. 1, pp. 24-38, 2013.

[5] G. Sun, T. Cui, W. Guo, et al. Micro Learning Adaptation in MOOC: A Software as a Service and a Personalized Learner Model, *Lecture Notes in Computer Science*, 9412, pp. 174-184, 2015.

[6] R. Wazirali, Z. Chaczko. Anticipatory Quality Assessment Metric for Measuring Data Hiding Imperceptibility, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, no. 2, pp. 404-412, 2017.

[7] G. Allione, R. M. Stein. Mass attrition: An analysis of drop out from principles of microeconomics MOOC, *Pier Working Paper Archive*, vol. 47, no. 2, pp. 174-186, 2016.