A New Student Achievement Evaluation Method Based on k-means Clustering Algorithm

Sumei Xi
School of Science
Qilu University of Technology
Jinan, China
Email: 37024953@ qq.com

Abstract—**This article puts forward an improved k-means clustering algorithm for the student achievement evaluation. Through the empirical analysis, we can explore the importance of evaluating the student achievement of the inner information hidden in the student score data, which can make up the drawbacks of the traditional methods.**

Keywords-clustering; achievement evaluation; teaching management; k-means

I. INTRODUCTION

Examination and teaching are inseparable in school education and exam that plays an important role in inspection of students' learning situation and state. Therefore, grade evaluation is very necessary for detecting and monitoring the quality of education, guiding the teachers' teaching behavior, and urging the students to study hard. Now in the school, where has a variety of systems and all kinds of databases, and already has accumulated a lot of student grade data, but due to the lack of related mining knowledge and technology, the staff can only get a small amount of information by simple statistics of Excel, and the most hidden information among these amount of data couldn't be used. Therefore, it is vital to improve the students' knowledge level that how to make statistical analysis of the student's early stage scores. Facing this challenge, data mining technology arises at the historic moment, and gradually shows strong vitality. As an important data mining algorithm, k-means algorithm is a kind of hard clustering method, namely dividing n sample data into k classes in an n-dimensional Euclidean space. Because the k-means clustering algorithm is sensitive to noise and outliers, and is very effective to deal with large data sets, k-means algorithm was applied to grade analysis in this article, so as to comprehensively analyze students' exam results.

Literature [1] expounded the role and the present situation of grade management, the shortage of the existing grade management, and showed the role of the decision tree algorithm and rough set theory in the grade management. Literature [2] introduced the k-means algorithm, and proposed an improved genetic k means clustering algorithm based on the classical k-means algorithm. Literature [3] improved it based on the analysis of the strengths and weaknesses of the k-means clustering algorithm, and compared the merits of the original algorithm and the improved algorithm through the experiments. Literature [4] introduced the classical data preprocess technology and implemented a heuristic session identification algorithm based on the reference documents of log request.

In this article, we will evaluate, count and analyze the students' grades by k-means clustering algorithm, so as to determine the relative location of students' grades among a group, which can prepare for improving the students' scores, providing the feedback information for the teaching work, and taking the corresponding remedial measures, and then to further improve the teaching quality of school.

II. THE GRADE EVALUATION SCHEME BASED ON CLUSTERING ALGORITHM

A. Overall Design Scheme

This article will design the overall scheme according to the figure 1. At the same time this article will choose the college grades database, which including the students' test scores of all courses.

Step 1, data acquisition. In order to ensure the data integrity and accuracy, the selection and finishing work of the raw data must be well done firstly. This article selects the course scores of some grade students in one semester of the college.

Step 2, data preprocessing. Data preprocessing is a gradually in-depth and step by step process. Preprocessing the data through data interview, data cleaning, data conversion and data verification four steps, we can solve the data collision and data inconsistency problems, eventually form a table of student achievements.

Step 3, clustering algorithm implementing. After determining mining tasks, we can realize the k-means algorithm on the process of students' score analysis via programming k-means clustering algorithm in the MATLAB.

Step 4, Clustering results evaluation. The information found in the clustering results will be explained and evaluated. After using k-means clustering algorithm, during the students' score evaluation, every class is a score group, different class corresponding to different score group, also the central grade of the different grades is given accordingly. These central grades are one of the reference standards of student grades division.

Step 5, putting forward the strategy. The dug out information will be provided to the teaching policymakers, adjusting the teaching strategy, further guiding the teaching work, improving the students' scores.

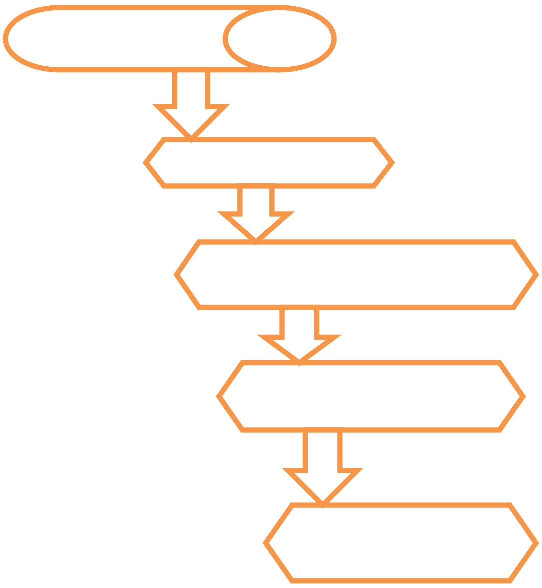

FIGURE I. DESIGN SCHEME IMPLEMENTATION

B. Design Principle Based on the K-means Algorithm

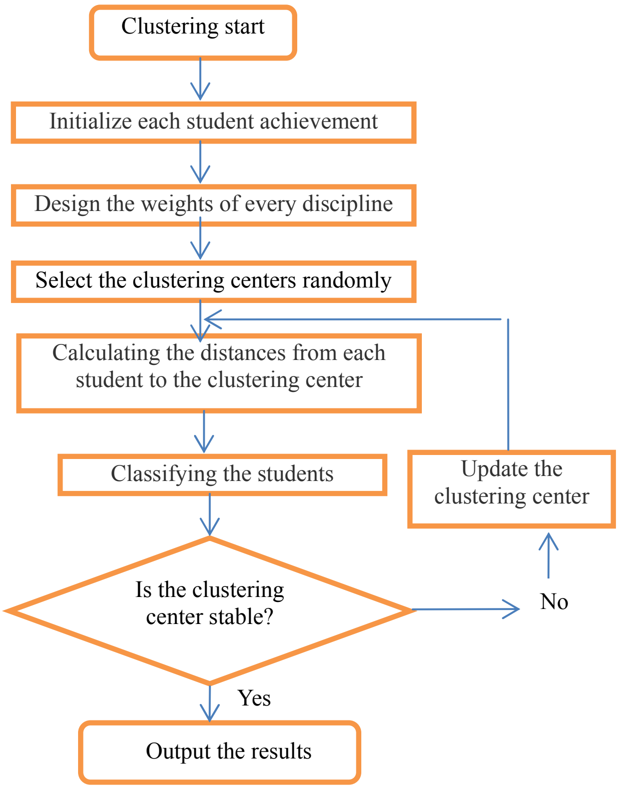
Clustering start
Initialize each student achievement
Design the weights of every discipline
Select the clustering centers randomly
Calculating the distances from each student to the clustering center
Classifying the students
Update the clustering center
Is the clustering center stable?
No
Yes
Output the results

FIGURE II. K-MEANS ALGORITHM OF STUDENT ACHIEVEMENT RESEARCH

Figure 2 shows the procedure of the k-means algorithm used in students' score processing. There are two key problems during the whole design process, and they are grade representation and the distance calculation of the score separately. For the first problem, this article treats each student's score in one subject as a q-dimensional vector, denoted by $x_i=(x_{1i}, x_{2i},\dots, x_{qi})$,(i=1,2,…,n), where x_{ki} denotes the score of k-th subject of which student number is i. The score uses a percentage grading system, and gives different weights according to different subjects. For the second problem, this article adopts Euclidean weighted distance to define the distance between the students' scores. The clustering number is set to P, and c_j（j=1,2,…, p）is the clustering center, then the distance from the score to the clustering center can be denoted by the formula (1).

$$\|x_i - c_j\| = \sqrt{\sum_{k=1}^{q} \omega_k (x_{ki} - c_{kj})^2},(1 \le j \le p) \qquad (1)$$

Where, q is the dimension of particle properties, ω_k for each property weight.

For all students grouping clustering results in all the subjects, the steps of our k-means clustering algorithm are as follows.

Step 1: suppose the students' score set is $Q=(x_1, x_2, \dots, x_{n-1}, x_n)$, where $x_i=(x_{1i}, x_{2i}, \dots, x_{qi})$;

Step 2: Randomly select a particle as the initial clustering center $c_1,c_2,\dots, c_P$ from each class;

Step 3: According to the formula (2), the objects x_i (i = 1, 2,..., n) in the student scores set Q, in turn, according to the Euclidean average distance, are assigned to the nearest center c_j (j = 1, 2,..., p).

$$\|x_i - c_j\| = \min(\sqrt{\sum_{k=1}^{q} \omega_k (x_{ki} - c_{kj})^2}),(1 \le j \le p) \qquad (2)$$

Where, q is the dimension of particle properties, ω_k for each property weight.

Step 4: calculating the new centers c_j(j=1,2,…, P) of P clusters according to the formula (3).

$$c_j = \frac{1}{N_j} \sum_{j \in S_j} x_i, j = 1,2,\dots,P \qquad (3)$$

Where, N_j is the particle number of the j-th cluster S_j.
Step 5: if each clustering center c_j(j=1,2,.., p) doesn't change any more, then finish, otherwise, return to step 3.

C. Student Management Strategy Based on the Grade Evaluation

In our strategy, we divide the students into four classes, which are excellent, good, moderate and underachiever, and propose some strategies to different categories of students from two aspects of self -development and teaching management, shown in table 1.

III. EMPIRICAL ANALYSIS

A. Instance Description and Evaluation Process

Step 1: data acquisition

According to the college database, we select some grade students' grades in some semester. The students all have eight courses in the semester, they are advanced math, college English, liner algebra, Chinese history, college computer, c programming, enterprise management and law fundamental, the corresponding weights is in turn are 0.2,0.2,0.2,0.1,0.1,0.1,0.05.0.05. Student achievement all are centesimal system, randomly selecting 200 students form a piece of original transcript.

TABLE I. STUDENT MANAGEMENT STRATEGY

class	Student self-development strategy	Teaching management strategy
excellent	While maintaining the excellent position, committed to the exercise and cultivate the ability of other aspects, so as to improve their comprehensive strength	Arrange competition, social practice, let the students can enhance their competitiveness
Good	Aligning with outstanding students, to find suitable learning methods for their own, improve the learning efficiency	Finding the potential of this part of students, to set up the incentive system, motivate their motivation beyond the outstanding students
Moderate	Looking for suitable learning methods for their own, improve the learning efficiency	Full attention and help the students, they would become the second or even the first category, or maybe back to the fourth class, this part of the student's effective management has a crucial role to improve overall student scores
underachiever	Listen carefully in class, finish the homework independently after class, strengthening basic courses learning, improve learning initiative	Arrange outstanding students to share their learning experience, improve this part of the students' study enthusiasm

Step 2: data preprocessing

This article integrates 200 students' original transcripts into a transcript. Through data processing, making each data in the table is the unique and no doubt, at the same time we can fill the empty data or delete them. We don't take the 0 scores into account just because the 0 scores in the database may affect the k-means algorithm. At the same time, we select the scores that less than 60 through Excel, the corresponding scores also not be processed by the k-means algorithm. Because once the score is below 60 points, the student must carry on the make-up exam, and the corresponding score will also be changed. the data collection about 200 students in this article totally have 10 people fail, then for the rest of the 190 students test scores we can do the k-means algorithm processing.

TABLE II. STUDENT ACHIEVEMENT TABLE AFTER DATA PREPROCESSING

82	58	75	72	71	83	85
78	74	55	76	79	88	83
80	71	78	74	55	79	88
78	57	77	73	72	80	81
76	77	53	74	80	82	84
74	55	76	75	71	71	88
79	75	74	69	52	79	84
88	68	59	77	72	80	84
76	56	79	69	74	85	88
77	76	75	66	55	77	83

Step 3: student achievement analysis and processing by k-means algorithm

Determining the clustering number value k, the clustering number is used to be close to the clustering variables number. Here we select k=4. By analyzing the initial data center, randomly selecting a few students' scores as the initial clustering center, we use the MATLAB algorithm to implement it.

The experimental results are as shown from figure 3 to figure 7.

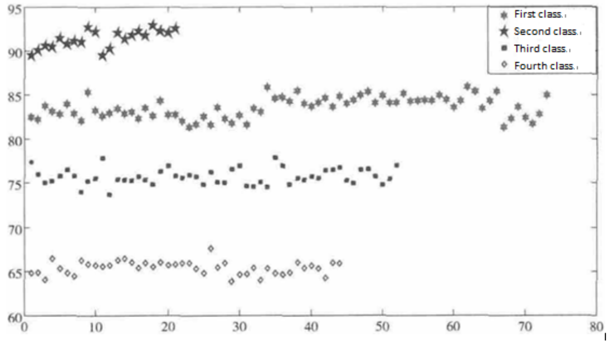

FIGURE III. CLUSTERING RESULTS OF STUDENT SCORES

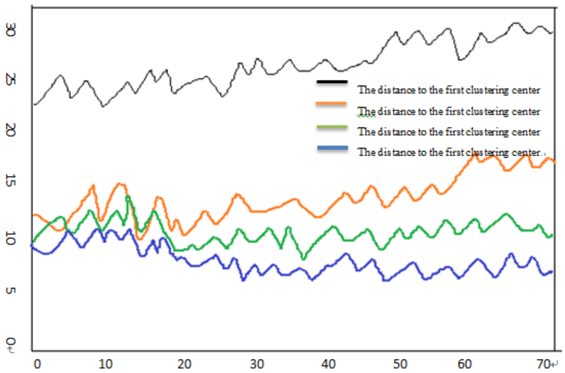

FIGURE IV. THE DISTANCES OF FIRST KIND OF STUDENT ACHIEVEMENT TO THE VARIOUS CLUSTERING CENTERS

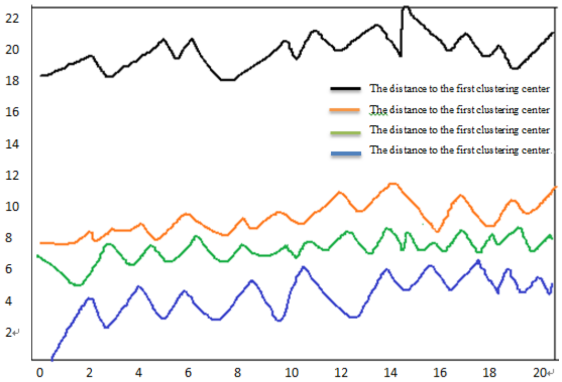

FIGURE V. THE DISTANCES OF SECOND KIND OF STUDENT ACHIEVEMENT TO THE VARIOUS CLUSTERING CENTERS

B. Empirical Results Analysis

1) The figure 3 shows that the second class of student

achievement is outstanding, the first class of student achievement is good, the third class student achievement is the medium, and the fourth class student achievement is the underachiever. Through calculation, the number of outstanding scores and the good scores account for 47% of the total, the number of medium and the underachiever account for 48% of the total, the rest is the proportion of students who have fail. It shows that the whole learning statement randomly selected from the major is needed to be improved and the relevant staff and teachers should take the necessary measures to improve students' learning motivation. By the research we can also find that, at the same time, each student’s achievement with the change of the center will affect the overall results distribution, especially like advanced math, college English, liner algebra, and other high weight subjects.

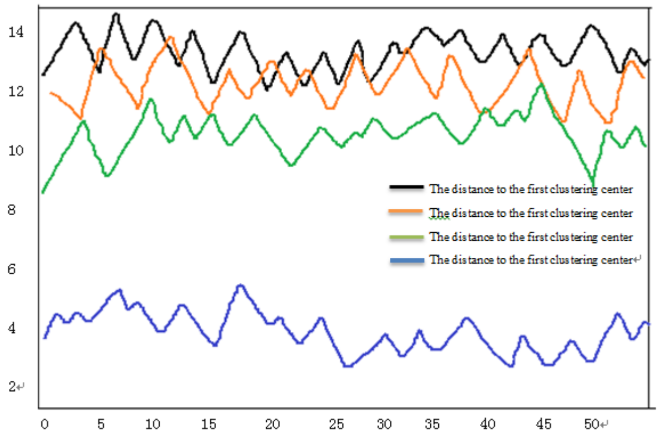

FIGURE VI. THE DISTANCES OF THIRD KIND OF STUDENT ACHIEVEMENT TO THE VARIOUS CLUSTERING CENTERS

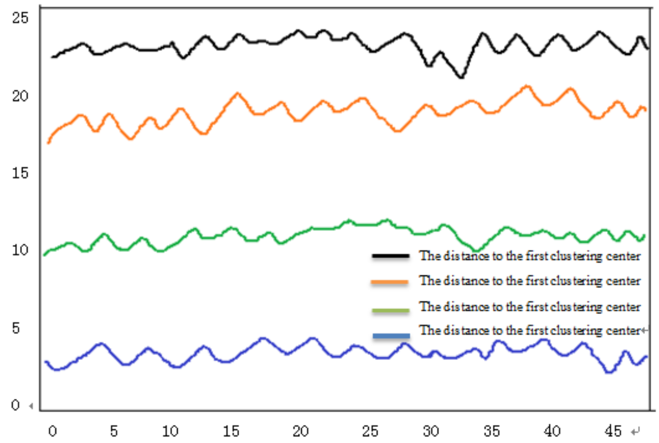

FIGURE VII. THE DISTANCES OF FOURTH KIND OF STUDENT ACHIEVEMENT TO THE VARIOUS CLUSTERING CENTERS

2) As shown in figure 4, 5, 6, 7, 21 students who are closer to the second clustering center gather into a class, 52 students who are closer to the third clustering center gather into a class, 44 students who are closer to the fourth clustering center gather into a class and 73 students who are closer to the first clustering center gather into a class. We can find that similar results were divided into the same class, which makes up the drawback of the traditional classification method, which is in the case of less difference of student achievement, after dividing the results may vary widely.

3) The application of clustering analysis technology can not only make 190 students knowing his own position overall all achievements, but also reflect the lack of certain types of the students in some disciplines, so as to remind the teaching staff taking relevant measures. Experimental results can provide a basis to the teaching personnel to develop a targeted solution, so as to improve students’ academic achievements in the late.

IV. CONCLUSION

This article research the application of k-means clustering algorithm in students’ achievement evaluation analysis. By the data preprocessing, using the k-means algorithm, analyzing the data by MATLAB, we make up the defect of the traditional statistical methods. We show some students' self-development strategy and teaching management strategy according to different types of students, so as to prepare and improve the teaching quality of student achievement for the late.

ACKNOWLEDGMENT

This work is supported by three projects of Shandong Province Higher Educational Science and Technology Program (J12LN09), China, Jinan Youth Science and Technology Star Project (No.20120104), China, and by natural science fundation Project of Shandong Province, China (No. zr2011fm028).

REFERENCES

[1] Qing Tan, “Analysis and Research of Grades of Examination Paper Ba sed on K-means Clustering Algorithm,” Journal of Henan University (Natural Science), vol. 39, No.4, pp. 412-415, Jul. 2009.

[2] Linhua Lu, Bo Wang, “Improved Genetic Algorithm-based Clustering approach,” Computer Engineering and Applications, vol.43, No. 21, pp.170-172, 2007.

[3] Aiwu Zhou, Yafei Yu, “The Research about Clustering Algorithm of K-means,” Computer Technology and Development, vol. 21, No.2, pp. 61-65, Feb. 2011.

[4] Liwei Zhang, Li Li, “The Research of Data Preprocessing Technology in Web Mining,” Computer Knowledge and Technology, vol. 6, No. 15, pp.4324-4325, May. 2010.