Journal of Computers Vol. 28, No. 6, 2017, pp. 195-203
doi:10.3966/199115992017122806017

Clustering of College Students Based on Improved K-means Algorithm

Zhongxiang Fan[1], Yan Sun[2] and Hong Luo[3]

[1] School of Computer Science and Technology, Beijing University of Posts and Telecommunications,
Beijing, 100876, China
fzx_1223@163.com

[2] School of Computer Science and Technology, Beijing University of Posts and Telecommunications,
Beijing, 100876, China
sunyan@bupt.edu.cn

[3] School of Computer Science and Technology, Beijing University of Posts and Telecommunications
Beijing, 100876, China
luoh@bupt.edu.cn

Received 14 December 2016; Revised 14 January 2017; Accepted 14 February 2017

Abstract. Many colleges have accumulated a large amount of information, such as achievement data and consumption records. According to the above information, we attempt to identify the student group from various aspects. Based on this, we can acquire the characteristics of students in different groups, then get the relationship between students' different behaviors by association rules mining method. In this way, the college can have a better understanding of students to accomplish the reasonable management. In order to obtain more accurate cluster results, we proposed an improved K-means algorithm. Specially, we effectively detect outliers based on the grid density. In addition, we design a new method to produce initial cluster centers which replaces the traditional random way. Real experiments are conducted and the results show the iteration time is reduced and clustering precision is improved.

Keywords: college student, initial cluster centers, K-means, density, outlier

1 Introduction

With the development of information technology, in some university, the digital system has been set up to improve the management work. As a result, a large number of data with very important value have accumulated. These data can help us understand students more comprehensive. In this respect, we analyze these data based on the data mining methods to acquire the characteristics of different students and the relationship of students' different behaviors.

Aiming to identify different groups of students in various aspects, we focus on the achievement data and consumption records for analysis. Through analyzing the corresponding data based on clustering methods, we can obtain the characteristics of students in different groups. In addition, we can acquire the relationship among achievement, consumption and other attributes. Using the result of analysis, we can provide decision support for student management systems.

K-means method is one of the most popular clustering algorithms. Although K-means algorithm has the great advantage of being easy to implement, it still has some drawbacks. In view of the shortcomings of the traditional K-means clustering algorithm, we proposed an improved K-means algorithm which can improve these problems. In the proposed algorithm, a method based on grid density was used to remove the outliers firstly. Secondly, we use a new method to generate the initial cluster centers to replace the original random way. Finally, the improved K-means algorithm was used to analysis student data, which can help us to get better clustering result of students.

This paper is organized as follows. Section 2 presents the related works. Section 3 introduces the proposed improved algorithm. Section 4 experimentally demonstrates the performance of the proposed algorithm. And we use the proposed algorithm to analyze the student data, get the relationship between different behaviors. The Section 5 describes the conclusion.

2 Related Work

Data mining has been applied in the education. The International Educational Data Mining Society' is concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students. There are already a lot of scholars working on this.

Xue used K-means algorithm to analyze university students' online behavior. The algorithm is improved by considering the element of user information [1].

Based on the support vector machine classification algorithm, the non-linear multi-class classification of student's campus card consumption is realized, and the relationship between consuming behaviors is found by using the correlation analysis method [2].

Fan used association rules mining method to study of the relationship between students' achievement [3].

Shankar, Sarkar and Sabitha aimed to analyze the performance of the students based on different attributes with respect to their country. They analyzed the big data of Harvard University online course to find the performance metrics of registered students from different countries by K-means clustering method [4].

This paper discussed the data mining technique used to identify the significant variables that affects and influences the performance of undergraduate students. The CHAID algorithm was used to predict students' academic success [5].

In order to get better clustering results, we decided to use K-means algorithm for clustering. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest distance. K-means is very sensitive to the initial cluster centers so that the clustering results will be very different from different initial cluster centers. The outliers will also affect the algorithm. The result would be inaccurate if there is outliers exist in data.

In the proposed K-means algorithm, one sample will be removed when its grid density is less than a certain threshold. After removed the outliers, initial cluster centers can be produced. In tradition algorithm, the initial cluster centers are produced randomly, then, we use a new method to produce the initial cluster centers. In the method, each dimension data were divided into K segments, the average value of each segment would be the coordinate values of the corresponding initial cluster center in this dimension.

3 Improved K-means Algorithm

In this section, in order to solve the influence of outliers on the K-means algorithm, a method based on grid density is proposed to remove the outliers. And we propose a new method to generate the initial clustering centers to improve the efficiency of the algorithm.

3.1 Removal of Outliers

In order to reduce the effect of outliers on the K-means algorithm, we first need to determine which are outliers in all the data and remove the outliers. To detect the outliers, we're going to calculate the density of every points, when the density value of a point reaches a certain threshold we judge this point as an outlier.

In most cases, the density of a point represents the number of points in a circular range. In this way, to obtain the density of a point we have to calculate the distances of all other points to this point, so the time complexity is $O(n^2)$. Here we detect the outliers based on the grid density. For a point, calculate the density value in the grid space, that is, the number of objects in this grid space. The density of the grid space can also reflect the distribution of a point, but in this way the time complexity is greatly reduced. In the algorithm, we sort all points from a dimension, and calculate the number of points in a certain range.

Previously, the threshold needs to be defined, once the density of a point is smaller than this threshold, it would be recognized as an outlier and removed. For the data set D with dimension m, set the density threshold MinDs, the grid side length is $\{l_1, l_2, \ldots, l_m\}$, and the specific steps to remove the outlier are as follows:

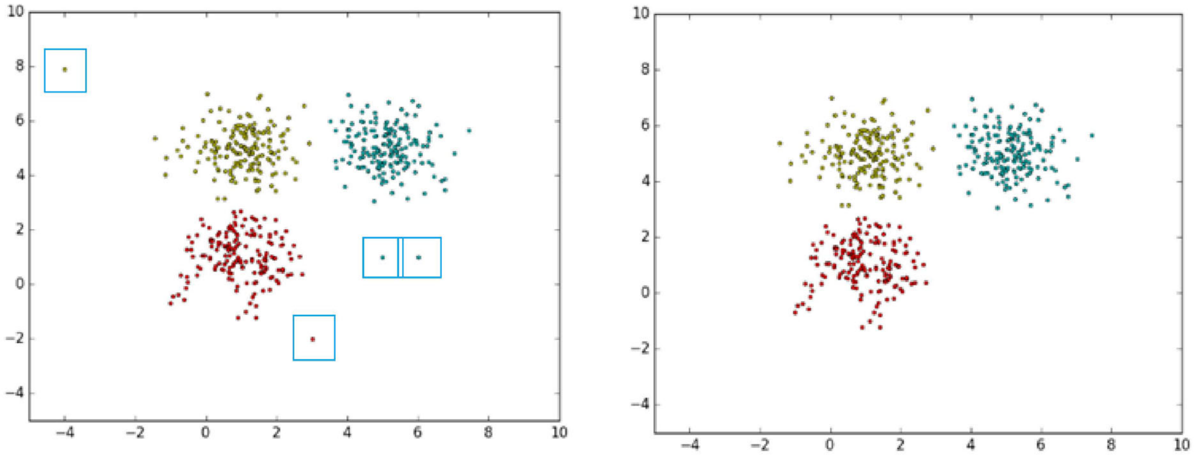
(1) sort all points in the data set according to the value of a dimension;

(2) traverse all the points $P_1, P_2, \ldots, P_n$ after sorting, calculate the density value of each point;

(3) for the current point $P(p_1, p_2, \ldots, p_m)$ traversing the other points forward and backward respectively. And determine whether the value of other points in the $(p_1 - l_1, p_1 + l_1)$ range, traverse the next point if the value in the range, or continue to judge the other dimensions, until all dimensions are judged;

(4) it is judged as an outlier if the grid density of point P is less than MinDs.

Data with outliers was used to verify the effect of the improved K-means algorithm. To be convenient to display, this data is two-dimensional. As shown in Fig. 1.



(a) The points with outliers

(b) The points after removing outliers

Fig. 1. Comparison before and after removing the outliers

Fig. 1 shows the comparison before and after removing the outliers. We can see four obvious outliers are really removed.

The value of the density threshold has a great effect on removal of outliers. In general, we are unable to determine a best threshold to detect the outliers. But we can get use some methods to get a good threshold. For example, we rank density of all points from small to large first. Then, we can observe the change of density value directly from the sequence of density. In general, we chose the point of maximum change as the threshold. In the example above, we give the first 25 density values after sorted: 0, 5, 16, 30, 94, 94, 109, 112, 116, 121, 125, 126, 127, 131, 131, 135, 136, 142, 143, 144, 144, 146, 146, 146, 148.

The change is very obvious when the density is 30. The density of points change more gently when these points belong to same cluster, because the distribution of the surrounding points is very close. While the outliers are significantly different, it has few points around, so the density change is very obvious. We take the point whose change is most intense as density threshold. However, sometimes the sorted density sequence changes smoothly between each point, it becomes difficult to determine threshold by the above approach. As shown in Fig. 2.

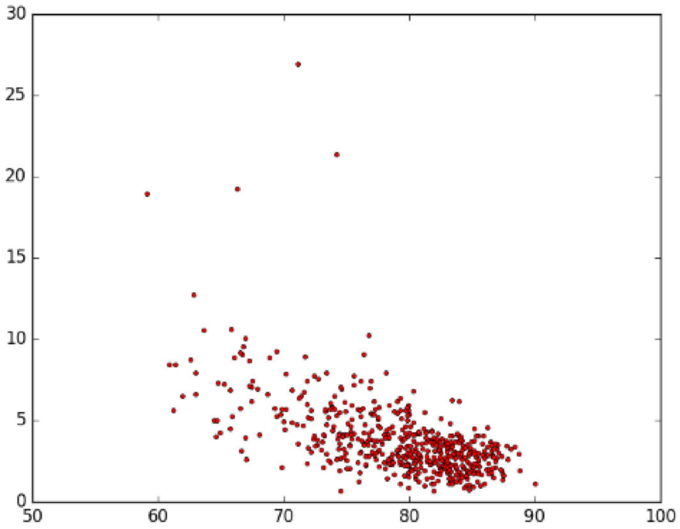


Fig. 2. Example 2

These points in Fig. 2 look more like a whole. The four points on the top of the figure are more likely to be outliers. Like the above, we also get its density ranking, and give the first 30:
0, 0, 0, 0, 1, 8, 11, 14, 22, 22, 24, 25, 26, 32, 33, 37, 38, 44, 48, 48, 51, 52, 57, 62, 68, 77, 84, 88, 88, 88.

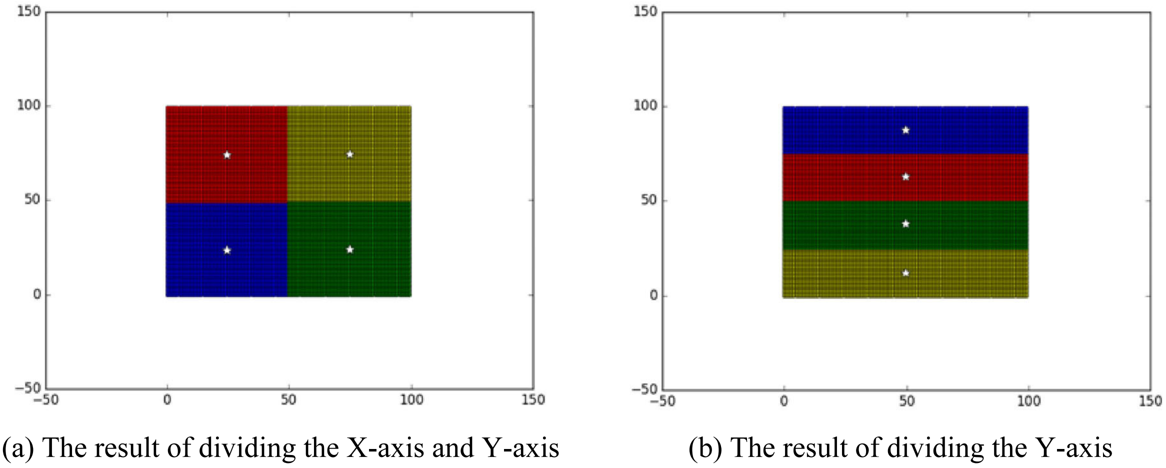
The change between these points is relatively smooth. In this case, we need to make decisions based on the data context. At this point, we detect the point whose density value is 0 as the outliers. In general, we set the smallest density value as the threshold.

3.2 Initial Cluster Centers

After the removal of outliers, the next step is choosing initial cluster centers. The final clustering result is very sensitive to the initial cluster centers, so in order to get a better result, more effective initial cluster centers should be produced. The traditional K-means algorithm generates initial cluster centers randomly. But this random method makes the result uncertain and has a low efficiency.

Cluster data can be divided into two types: one is the data distribution has certain segmentation and significant difference between each cluster; the other one is the distribution of the data has no distinct segmentation. In the first case, no matter which method is used to produce the initial centers, the final clustering result is almost same. However, in the second case, the method used is particularly important.

We take 100x100 evenly distributed points as an example, the results of different initial clustering centers are completely different, as shown in Fig. 3



(a) The result of dividing the X-axis and Y-axis

(b) The result of dividing the Y-axis

Fig. 3. Comparison of the results of different initial clustering centers

In Fig. 3 (a), the initial cluster center is selected by dividing the X-axis and Y-axis at the same time. This is also the result of the random clustering center. The initial cluster centers selected are the results of the division of the Y-axis only in Fig. 3. (b). The X-axis coordinates of the initial clustering centers are the same. Where the white star represents the final cluster center, but also the initial cluster center. This very regular data is rare in reality, most of the actual data will not be so regular. So in most cases the choice of the initial cluster center will also affect the clustering results, but in reality the impact of time even greater.

We use a new improved method to produce initial cluster centers. We divide data into K segments from each dimension. The average value of each segment will be the coordinate of corresponding initial cluster center in this dimension. In this way, the distance between each initial cluster center is large, which can make the differences between clusters more apparent. These initial centroids lead to a better clustering results.

Taking the previous data as an example, the results are shown in Fig. 4, the three triangles represent three initial clusters. We can see that the distance between the initial clusters is relatively far.

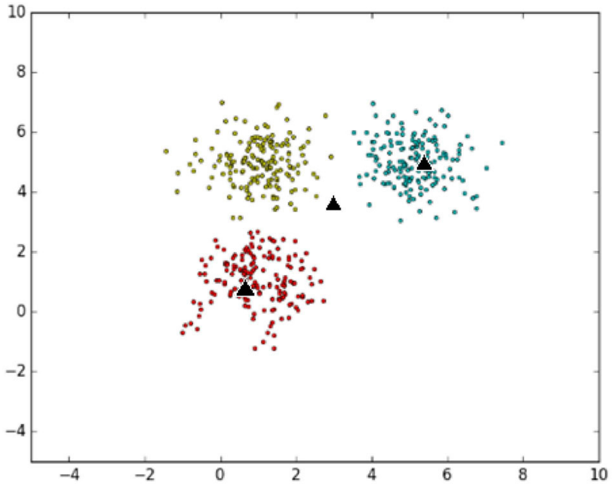


Fig. 4. Initial Cluster Centers

3.3 Improved K-means Algorithm

After the improvement of the shortcomings of the original K-means algorithm described above, the following is the specific process of the improved K-means:

Input: Dataset D (set of n samples, dimensions for m), number of clusters k, density threshold MinDs.

Output: A set of k clusters.

1: Calculate maximum and minimum values J_{min} and J_{min} in each dimension J ($0 \leq J \leq m$). Set the side length of grid in each dimension $GL_J = (J_{max}\text{-}J_{min}) / (k+1)$.

2: Calculate the grid density of each sample.

3: Remove the sample whose grid density is less than the threshold.

4: After removing the outliers, sort the value of each dimension J. Divide values into k segments, get average value of each segment $\{p_{j1,} p_{j2}, \cdots , p_{jk}\}$.

5: Set the value in step 4 as the coordinate of k initial cluster centers $\{C_1, C_2, \cdots , C_k \}$, $C_k = \{p_{1k}, p_{2k}, \cdots , p_{mk}\}$.

6: Calculate the distance between each sample to every cluster centers. Assign this sample to the nearest cluster.

7: For each cluster, recalculate the cluster center.

8: Repeat 6, 7 until the center of the cluster no longer change.

The determination of K value is a common problem in clustering process. In here, we use the "elbow method" to solve it. Give an appropriate cluster indicator, such as average radius or diameter of clusters. The indicator changes slowly when the number of cluster is larger than real K and quickly when it's smaller than real K, as show in the Fig. 5.

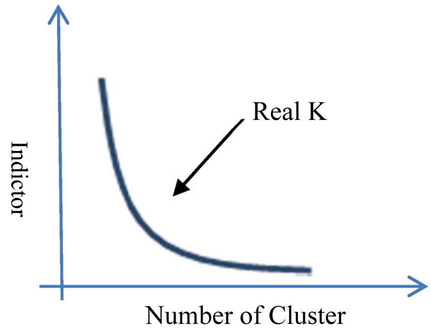


Fig. 5. K Value - Indicator

Firstly, we select an interval of K, and cluster the data under different K value. Then choose the K by the above method.

4 Experiment and Application

In order to validate the efficiency of improved algorithm, we test both algorithms for the dataset with known clustering. These data are from UCI. We conducted 100 experiments on each data, and take the average accuracy and time of all experiments as result. Dataset 1 - 6 are IRIS, Glass Identification, ILPD, Pima Indians Diabetes, Car Evaluation, Seeds. The results are shown in Fig. 6 and Fig. 7.

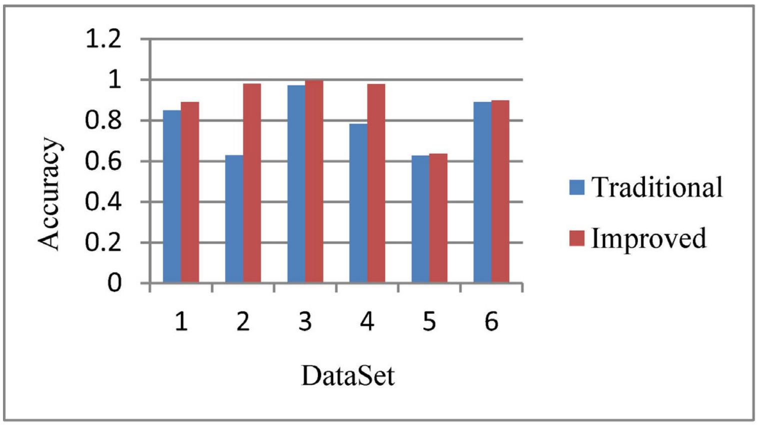


Fig. 6. Clustering accuracy

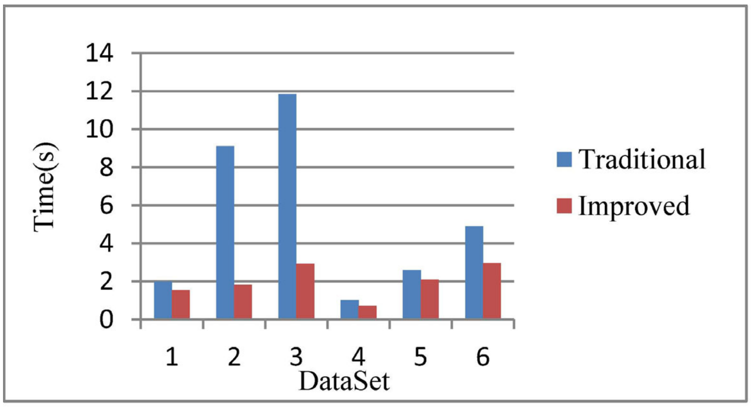


Fig. 7. Clustering time

The results show that the proposed algorithm produces better clustering results compared to the traditional algorithm in less computational time.

As mentioned above, an improved K-means algorithm was proposed. In this part, the proposed algorithm is applied to the data of 2014 students of computer institute. The achievement data and consumption data are clustered respectively. The achievement data is related to the average score of each semester, as shown in Table 1.

Table 1. Features of student achievement

ID	1	2	3	4	5	6	7
Xxx	91.4	88.5	89.2	90.3	90.1	86.2	89.5
…	…	…	…	…	…	…	…

According to the above table, ID means the student id. 1-7 columns respectively represent the average grade of each semester. The consumption data is shown as Table 2.

Table 2. Features of student consumption

ID	TM	AM	TC	AC	TI	AI
Xxx	3016	5.6	4028	8.3	370	24.7
…	…	…	…	…	…	…

ID also means the student id. TM and AM represent the total consumption and average consumption in market. TC and AC represent the total and average consumption in canteen. Similarly, TI and AI represent the total and average Internet charge.

According to the student's performance and consumption data, we are going to cluster student. The K value is 4 for achievement data and 5 for consumption data. The clustering results are shown in Fig. 8 and Fig. 9.

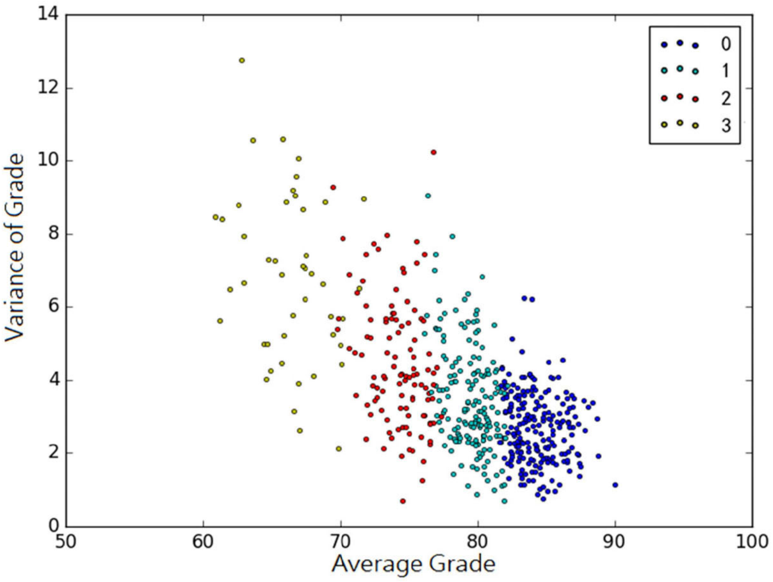


Fig. 8. Student achievement distribution

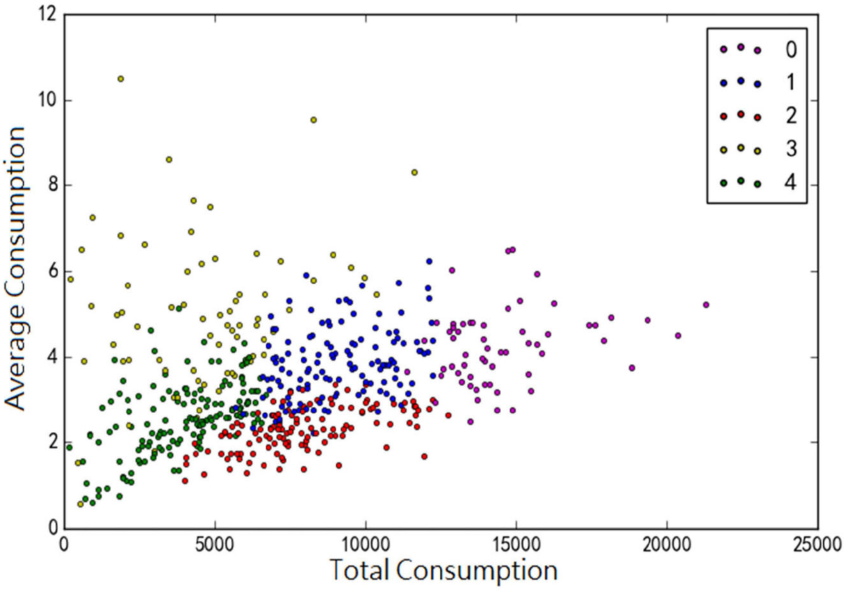


Fig. 9. Student consumption distribution

It can be seen that the average grade declines and the variance rises from cluster 0-3. We can consider that the performances of students from cluster 0 to 3 are getting worse. There is a linear relationship between the average grade and the variance, as the average grade increases, the variance decreases. It means that the higher the grade, the more stable the student achievement.

It also can be seen that the total consumption declines from cluster 0-4, and the average consumption changes without apparent rules. It means that there is no linear relationship between the total consumption and average consumption.

Next, we are going to mining the association between the clustering results and the student's other behaviors. In order to facilitate the association analysis of student behavior, we represent the student's different behavior categories with numbers, as shown in Table 3 below.

Table 3. Student behavior category

Consumption	Achievement	Library Fines	Network Charge	Sex
0	0	0	0	0(female)
1	1	1	1	1(male)
2	2	2	2	
3	3	3	3	
4		4		

The consumption and grade categories correspond to the categories in the above clustering results. Students are divided into five categories according to the library fines. From 0 to 4 the fine of students gradually increased, 0 class students are not fined. Students are divided into four categories according to the net recharge amount. From 0 to 3 all kinds of students the amount of network fees gradually increased, 0 class students did not recharge. The Table 4 shows the association rules between the different behaviors we get and the confidence of the association rules. We set the threshold of confidence as 0.7, so we only give the association rule with confidence greater than 0.7.

From the above association rules, according to the class achievement 1, 2, 3 are related to male, we can know that the majority of students with poor grades are boys. The confidence of the three rules gradually increased, indicating that students with lower grades are more likely to be boys. The class consumption 0, 1, 3 are related to male, consumption 2 related female, it means that the average consumption of boys is higher than that of girls. And according to the relationship between achievement 0, net charge 0 and female, we can see students with good grades generally charge less. Boys spend more energy on the Internet than girls, so girls' achievement generally better than boys.

Table 4. The association rules between behaviors

Antecedent	Consequent	Confidence
Consumption 3	Male	0.959
Achievement 3	Male	0.934
Consumption 0	Male	0.925
Consumption 1	Male	0.865
Net Charge 3	Male	0.851
Achievement 2	Male	0.844
Net Charge 1	Male	0.822
Library Fine 0	Male	0.812
Consumption 2	Male	0.800
Library Fine 4	Male	0.794
Consumption 2, Achievement 0	Net Charge 0, Female	0.784
Female	Net Charge 0	0.765
Consumption 2	Net Charge 0	0.763
Achievement 1	Male	0.726
Consumption 4	Net Charge 0	0.712

5 Conclusion

K-means is one of the most popular algorithms in clustering algorithm. However, the result depends extremely on initial cluster centers. Meanwhile, the outliers have a great impact on the result. To avoid these problems, an improved K-means algorithm based on the grid density was proposed. It can reduce the influence of outliers on the results apparently. In addition, this algorithm generates the initial cluster centers by dividing data from each dimension to produce more accurate result.

After that, the proposed K-means was used to get the cluster of students from different aspects, get the specific differences between students of different clusters. According to this, we obtained association between the sex, achievement, consumption and other behaviors of students, so as to improve the management work of students. In this paper, we mainly analyze the achievements and consumption of students. In the future, we can get a more comprehensive analysis of students from the perspective of more.

References

[1] L. Xue, W. Luan, Improved k-means in user behavior analysis, in: Proc. International Conference on Frontier of Computer Science and Technology, 2015.

[2] D. Wang, Research and application of data mining in campus card consumption behavior analysis, [dissertation] Harbin, China: Harbin Engineering University, 2010.

[3] T.K. Fan, J.Y. Sun, Analysis and application of college students' academic record based on data mining, Computer & Modernization 1(3)(2013) 82-84.

[4] S. Shankar, B.D. Sarkar, S. Sabitha, Performance analysis of student learning metric using k-means clustering approach, in: Proc. 6th International Conference-Cloud System and Big Data Engineering, 2016.

[5] A.C.K. Hoe, M.S. Ahmad, T.C. Hooi, M. Shanmugam, S.S. Gunasekaran, Z.C. Cob, A. Ramasamy, Analyzing students records to identify patterns of students' performance, in: Proc. International Conference on Research and Innovation in Information Systems, 2013.