International Journal of Multidisciplinary Engineering in Current Research
ISSN: 2456-4265, Volume 6, Issue 11, November 2021, http://ijmec.com/

METHODOLOGICAL IMPLEMENTATION OF ACADEMIC ORGANIZATIONAL OPERATED DATA WITH CLUSTERING MECHANISM USING AN IMPROVED K-MEANS ALGORITHM

Mr. Pathan Ahmed Khan
Assistant Professor,
Department of Computer Science and Engineering,
ISL Engineering College.

Mr. Mohammed Rahmat Ali
Assistant Professor,
Department of Computer Science and Engineering,
ISL Engineering College.

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. Given this, we can acquire the characteristics of students in different groups. In this way, the college can have a better understanding of students to accomplish the reasonable management. 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- *K-means; density; outlier; initial cluster centers; college student;*

I. 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 amount 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.

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 Kmeans 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. In section 5, we use the proposed algorithm to analyze the student data, and show the result. And the Section 6 describes the conclusion.

II. IMPROVED K-MEANS ALGORITHM

For enhancing the performance of k-means clustering algorithm, we proposed an improved algorithm based on the grid density. To detect the outliers, we're going to calculate the density of every point, 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 distance of this point with all other points 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(n2). Here we detect the outliers based on the grid density, so that the time complexity decreases. 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.

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

International Journal of Multidisciplinary Engineering in Current Research
ISSN: 2456-4265, Volume 6, Issue 11, November 2021, http://ijmec.com/

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. So 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.

Improved K - means algorithm:

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

1. Calculate maximum and minimum values Jmin and Jmin in each dimension J (0<=J<=m). Set the side length of grid in each dimension GLJ = (Jmax-Jmin) / (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 {pj1, pj2, O, pjk}.
5. Set the value in step 4 as the coordinate of k initial cluster centers {C1, C2, …., Ck }, Ck = {p1k, p2k,…, pmk}.
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. 8: Repeat 6, 7 until the center of the cluster no longer change.

III. EXPERIMENT AND ANALYSIS

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 follows:

.

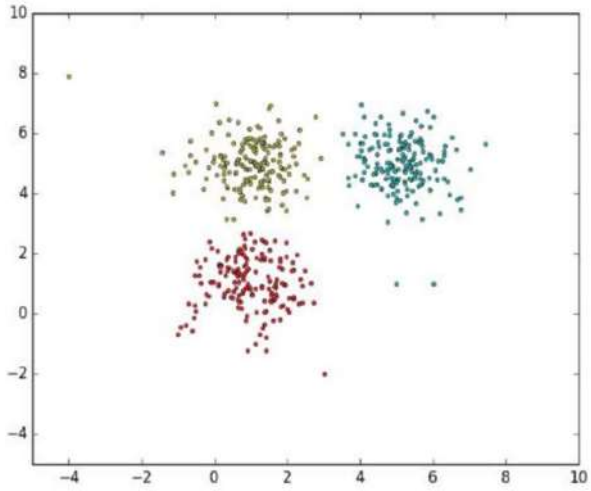

Figure 1

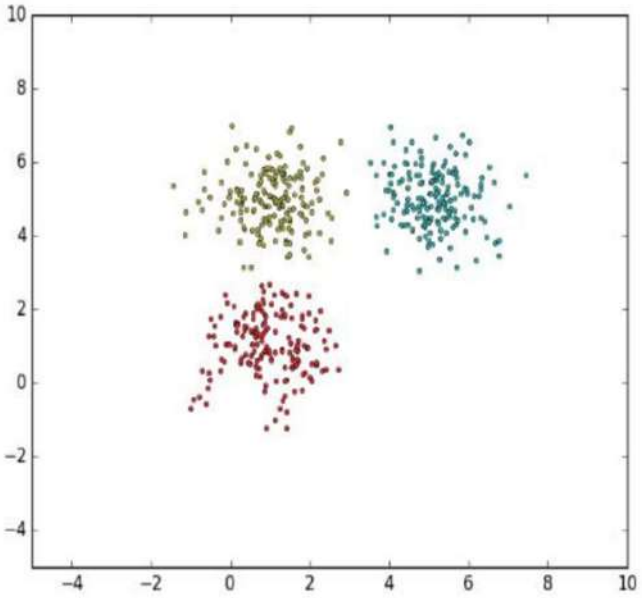

Figure 2

Fig. 1 shows the data with noise points. Fig. 2 shows the data removed the noise points. We can see 4 obvious outliers are really removed in Fig. 2.

In the improved K-means algorithm, 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

International Journal of Multidisciplinary Engineering in Current Research
ISSN: 2456-4265, Volume 6, Issue 11, November 2021, http://ijmec.com/

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 follows:

These points in Fig. 3 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.

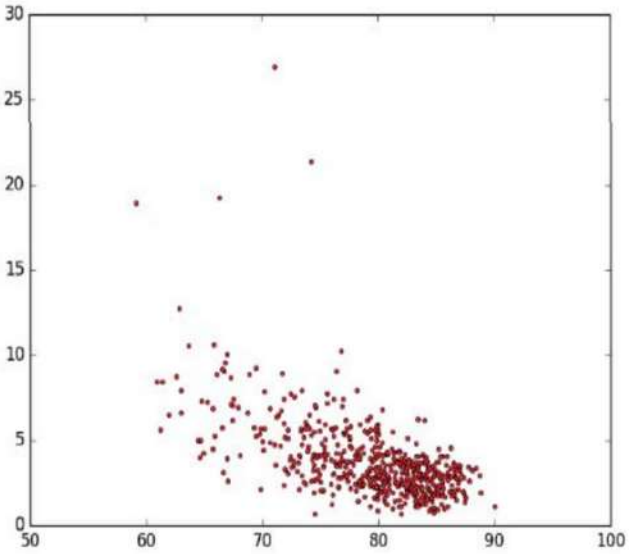

Figure 3

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.

In the above, we know the cluster result is sensitive to initial cluster centers. Next, we will discuss how to choose a suitable method to produce initial clustering centers. In the algorithm we proposed, we divided the data from every dimension. But if we want to reflect the difference of a certain dimension, we can only divide the data of this dimension, As follows.

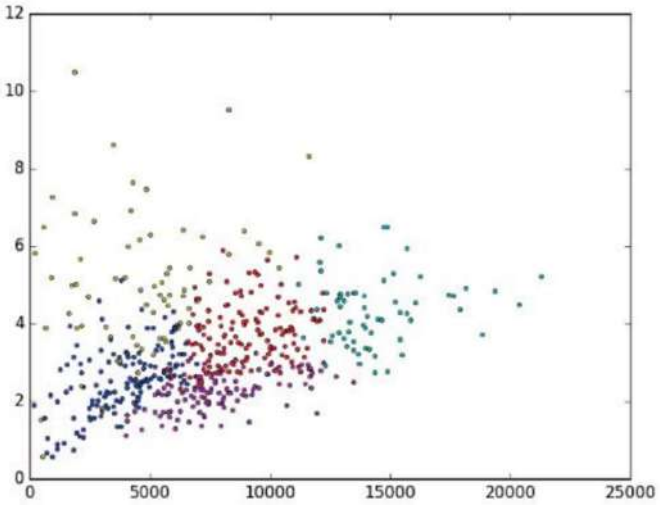

Figure 4

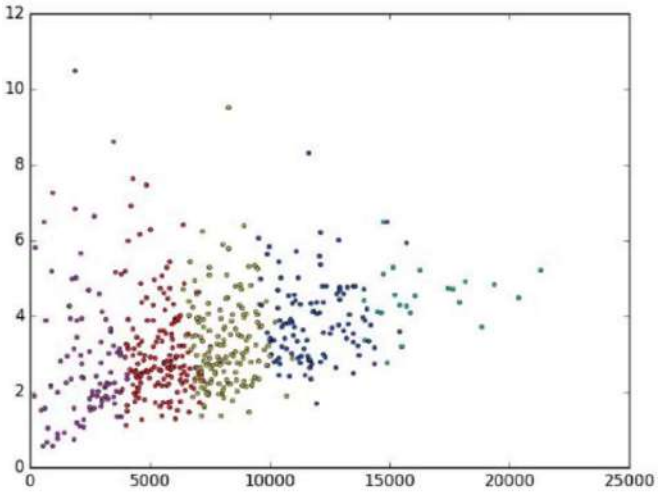

Figure 5

Fig. 4 is the clustering results divided from every dimension; Fig. 5 is the results divided only from X axis. As we can see, if we pay more attention to the differenceof the X axis, we can only divide the data of the X axis to determine the initial cluster center.

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

International Journal of Multidisciplinary Engineering in Current Research
ISSN: 2456-4265, Volume 6, Issue 11, November 2021, http://ijmec.com/

Identification, ILPD, Pima Indians Diabetes, Car Evaluation, Seeds. The results are as follows:

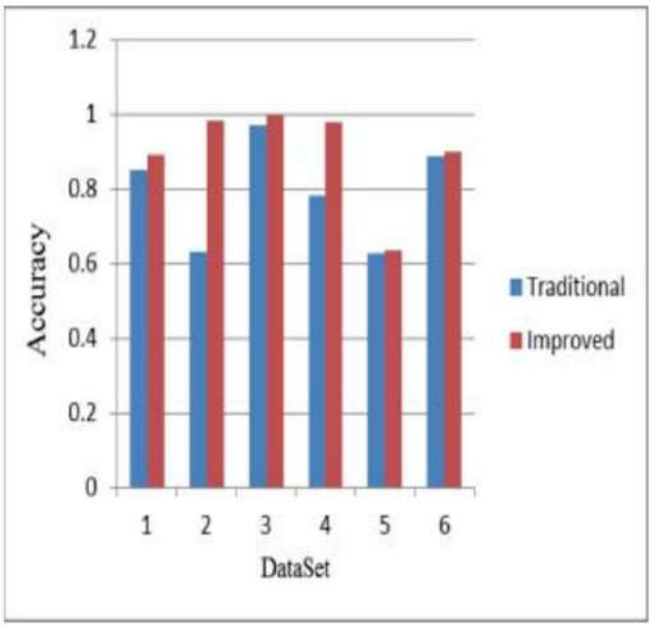

Table 1-Clustering accuracy

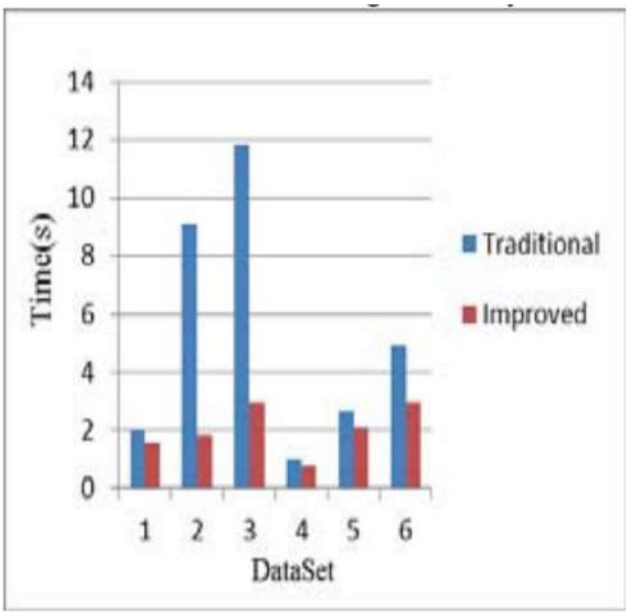

Table 2-Clustering time

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

IV. 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, and get the specific differences between students of different clusters. According to this, these results can be used to obtain the relationship between different 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.

V. REFERENCES

[1] Syed Rizvi, Nathan Showan, John Mitchell, “Analyzing the Integration of Cognitive Radio and Cloud Computing for Secure Networking,” Procedia Computer Science, vol. 61, pp.206-212, 2015

[2] Claudio Mazzariello, Roberto Bifulco, Roberto Canonico, “Integrating a Network IDS into an Open Source Cloud Computing Environment,” International Conference on Information Assurance and Security (IAS), Atlanta,USA, 2010, pp.265-270

[3] Vieira, K., Schulter, A., Westphall, C., &Westphall, “Intrusion Detection for Grid and Cloud Computing,” IT Professional, vol. 12, no.8, pp.38-43, 2010

[4] Kumar, P., Nitin, N., Sehgal, V., Shah, K., Shukla, S. S. P., & Chauhan, “A novel approach for security in Cloud Computing using Hidden Markov Model and clustering,” IEEE, World Congress on Information & Communication Technologies, 2011, pp.810-815

[5] Modi, C. N., Patel, D. R., Patel, A., &Muttukrishnan, “Bayesian Classifier and Snort based network intrusion detection system in cloud computing,” International Conference on Computing Communication and Networking Technologies, 2012, vol. 90, pp.1-7

[6] Anand Kannan, “Performance evaluation of security mechanisms in Cloud Networks,” Ph.D. dissertation, HEC, Stockholm , Sweden, 2012