CHEMICAL ENGINEERING TRANSACTIONS
VOL. 46, 2015
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Peiyu Ren, Yancang Li, Huiping Song

ISBN 978-88-95608-37-2; **ISSN** 2283-9216
DOI: 10.3303/CET1546090

Research and Improvement of The Partition Clustering Algorithm Based on Distance Sum

Juan Li

School of Computer Engineering, Jinling Institute of Technology, Nanjing, 211169, China.
iamlj6@jit.edu.cn

This paper mainly improves the choice of the initial clustering center and isolated point problem of K-means algorithm. Firstly, we calculate the distances between all the data objects, and eliminate the influence of isolated points according to the ideas of the distance sum. Then we put forward a new method of the initial clustering center selection. In experiment part, we make the comparison between the improved algorithm and the original algorithm through the experiment. Experiment results show that our improved algorithm obviously decreases the influence of the isolated point, and the clustering result is closer to the actual data distribution.

1. Introduction

K-means algorithm leads to sensitivity to the isolated points by using the centroid as a reference point. The k-medoids algorithm selects the point closest to the center position in the clustering as the reference point, so it is not affected by the isolated points. PAM, the partition algorithm around the center, is one of the first proposed k-medoids algorithms. The PAM method is very effective for small data sets, but not ideal for large data set processing. CLARA (Clustering Large Application) is specially used for processing large amounts of data, which is an enhanced version of the PAM method [Z. Guofu et al (2006) reported]. The CLARA method improves the calculation speed in processing large amounts of data, but the result of the CLARA method is not the optimal solution of all the data. Generally speaking, the quality of the clustering results of CLARA method is determined by sampling method. The CLARANS method is also known as random searching algorithm for clustering [T. He et al (2012) reported]. It combines sampling technology and PAM, which is an enhanced version of CLARA method.

If the data clusterings are intensive, and the difference between each cluster is very big, K-means algorithm is more applicable. In addition, the complexity of the algorithm is O (nkt), wherein n represents the number of all data objects; k denotes the number of clustering; and t represents the iteration times of algorithm. So K-means algorithm can handle large datasets very well. However, the treatment of K-means algorithm for symbol attribute data is not good. The selection of initial clustering centers determines the quality of the algorithm. Meanwhile, the K-means algorithm is sensitive to the inputting sequence of data sample, and the isolated points have great influence on the clustering algorithm.

Since K-means algorithm has many shortcomings, lots of clustering algorithms have improved. For example, in order to reduce the influence of the isolated points on clustering effect, the authors adopt the thought, the clustering mean points separating with the clustering seeds [Yeli Li et al (2007) reported]. Some researchers adopt the pretreatment way for data to improve significantly the quality of the clustering algorithm [B. Yuanyuan et al (2009) reported]. In order to improve effectively the influence of the isolated points on the algorithm, the authors raise two aspects of improvement: data preprocessing and the selection of the initial clustering centers [L. Fengna et al (2008) reported].

There are two aspects problems in K-means algorithm: the selection of the initial clustering centers and the isolated points. This paper has mainly improved these two aspects. The second part of the paper introduces the K-means algorithm and the improvement ideas of the proposed algorithm. The third part is the concrete implementation of the improved algorithm. The fourth part conducts experiments and analysis on random data and standard data. It is proved that the advantages of the improved algorithm. Finally, the fifth part makes a conclusion for the paper.

Please cite this article as: Li J., 2015, Research and improvement of the partition clustering algorithm based on distance sum, Chemical Engineering Transactions, 46, 535-540 DOI:10.3303/CET1546090

2. The ideals of the improved algorithm

K-means algorithm is a partitioning clustering algorithm that takes mean values as the clustering centers. It is simple and fast. K-means algorithm is often applied in image analysis and pattern recognition etc. The first step of K-means algorithm for clustering is to select the initial cluster centers. The choice of the initial cluster centers directly determines the quality of the clustering algorithm. The second step is to sort all data objects according to the initial clustering centers. The last step needs to calculate the average value of each clustering, so as to continuously adjust the cluster centers. Through iterative loop optimize the clustering results continuously. The algorithm is designed to make internal data objects of each clustering very similar, but data objects in different clustering vary greatly.

The K-means algorithm is greatly affected by the initial clustering centers. There are two kinds of methods to determine the initial cluster centers. The first way is to put the entire data sample directly divided into K classes. Then take the average values of each clustering calculated as the initial cluster centers. The second way is to select multiple sets of initial cluster centers, and then clustering respectively. Finally, through the comparison find the best results. In real life, there are actual isolated points existing in the data sample. If the random selection method is adopted, the isolated points are likely to be the initial cluster centers. Because the isolated points are away from the data intensive areas, once the isolated point as the initial cluster center, it will have a great influence on the clustering result. In addition to this, the algorithm takes the mean point of each clustering as the new clustering centers for iterative processing while the mean points of the clustering are possible distant from the region with dense data, which will cause great impact on the clustering results. So in order to avoid the influence, the first step of the improved algorithm is to exclude the isolated points in the data sample by running the searching algorithm of the isolated points. And then cluster the remaining data objects in the database. The isolated points are in general processed after the end of the clustering algorithm. This improved algorithm is based on the ideas of Euclidean distance [Z. Zhonglin et al (2010) reported]. The data samples mainly use the two-dimensional data for clustering processing. Based on the ideas of distance sum mentioned in the literature [L. Shenglian et al (2004) reported], the isolated points are found out in the data sample. Firstly, calculate the distances of all the data objects between each other. Take the point whose distance from other data objects is the maximum as the first isolated point and remove it. Next, delete the second maximum isolated point. The number of the isolated points is mainly in accordance with the accuracy requirements. Generally speaking, in order to find the existing isolated points in the data samples, the distances of all data objects between each other need N2 calculation. If the data sample set N is too large, the time complexity of the algorithm will be high. So in order to solve this problem, it can be to carry on the random sampling of the data samples. Then by approximate calculation obtain the distance sum. It assumes that the uniform random sampling of all data objects is very efficient. If the distance sums between each data object and other data objects are expected to get, it is only necessary to calculate the distances between the data objects and the data objects in the random sample. Since the random sample is relatively small, the time complexity is reduced effectively.

3. Implement of the improved algorithm

The formula denoting the Euclidean distance between any two data objects in two dimensional space is as follows:

$$d = sqrt((x2 - x1)^2 + (y2 - y1)^2)$$

In the formula, x expresses the abscissa; y represents the ordinate; $(x1, y1)$ and $(x2, y2)$ show the two data sample points.

For the choice of the initial clustering center, the specific steps are as follows:

(1) First calculate the distances between two objects among Q data objects, so that output n order distance matrix D.

(2) Calculate the sum of the distances between each data object and all the other data objects, that is, find the isolated points based on the ideas of distance sum. The concrete implementation method is executing $A = sum(D, 2)$ through MATLAB for the sum of data elements on each row in the distance matrix D.

(3) Determine the first isolated point in the data samples. The concrete realization is for the matrix A generated in the second step executing $[q,1] = \max(A)$ to calculate the maximum point q in the matrix A, i.e. the point with the maximum distance sum, whose position is showed by 1.

(4) The third step is to identify the first isolated point, then after deleting the first isolated point, the algorithm needs to return to the first step for cycle processing. The remaining Q-1 data obtain the distance matrix, and then determine the second isolated point, i.e., find out the point with the second big distance sum. In order to

meet the requirement of the accuracy of the algorithm, exclude all isolated points meeting accuracy requirement through the continuous cycle processing.

(5) After deleting these isolated points, the remaining data objects get the distance matrix D.

(6) Firstly, the two points with the largest distance between each other are found out as the initial clustering centers. Then make these two data points into a straight line as a diameter to describe circle. The circle described in this way can include all data sample points basically.

(7) Then determine the third initial clustering center by dividing the circle into four equal parts. The third initial clustering center is the data object that is closer to the new diameter endpoints. If the cluster number is much more and K is more than 4. It is necessary to divide the circle into eight equal parts. The way is to draw another symmetric diameter. At last choose K data objects as the initial clustering centers.

After eliminating the isolated points, the remaining data objects run the K-means clustering algorithm for clustering results. The algorithm gets the final clustering centers. At the moment, calculate the distances between the isolated points and the final clustering centers to decide these isolated points belong to any clustering according to the principle of the nearest distance algorithm.

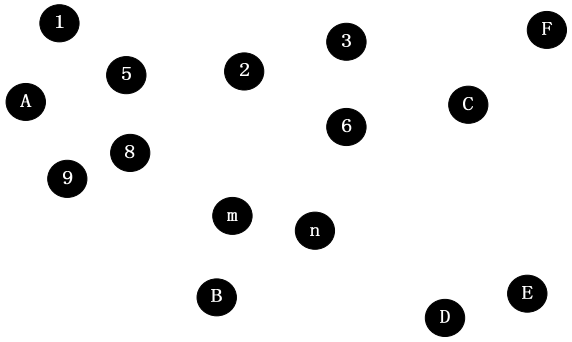

Figure 1: Distribution of 15 data objects

It assumes that there is a data sample set including 15 data objects, whose distribution is shown as Figure 1. Now carry out clustering based on the improved algorithm. Assume that K is 3, namely 3 classes, and it needs to eliminate three isolated points. Firstly, by calculating distances between each data sample, find out D, E, F, whose distance sums from other samples are the maximums. Filtrate them so as to eliminate the influence of the isolated points on the clustering centers. Then select the two points A, C with the largest distance in the remaining 12 points as the two beginning clustering centers. A and C are connected into a straight line which is selected as a diameter to describe a circle. Draw another diameter in AC vertical direction. Since B point is closer to the diameter endpoint, choose it as another cluster center. As K is 3, the initial clustering center satisfies the requirement. Based on the distances of the remaining objects to the three initial clustering centers (A, B, C), conduct clustering. Four data points, 1,5,8,9, are assigned to class A. m and n are assigned to class B. 2, 3 and 6 are assigned to class C. Then calculate the new clustering center. The clustering center of class A is the average value of the five points. The clustering center of class B is the average value of the three points. The clustering center of class C is the average value of the four points. Then take the new clustering centers as the reference points to conduct iteration cycle until the clustering object function (square error criterion) is the smallest, causing the clustering to achieve optimum. The improved algorithm only needs a single iteration. If select the initial clustering center randomly, the iteration number will markedly increase. After the completion of clustering, clustering the three isolated points D, E and F. According to the principle of the nearest to the center, calculate the distances between the isolated points and the final clustering centers. The results show that D and E belong to class B. F point belongs to class C.

4. Experimental analysis with random data

The hardware environment of the experiment mainly is the Intel core i3 dual-core processor, 380M, 2G memory. The software environment is win7, 64 bit operating system. The experiment uses two dimensional data of real type, processed in MATLAB environment.

By running $x = rand(80, 2)$ generate 80 data objects randomly to be the data sample for processing. The experimental data are showed as the following table 1.

The coordinates of random data range from 0 to 1. Since the data sample is small, K is 4. i.e., divide the data objects into four classes. The first clustering in the result diagram is solid points. The second clustering is pluses. The third clustering is circles. The fourth clustering is five-pointed stars. First, iterate the algorithm for

100 times to get the accurate clustering results, then compare and analyze how many times it at least needs to obtain the same accurate results.

Table 1: 80 random experiment points

(0.9501, 0.7948)	(0.2311, 0.9568)	(0.6068, 0.5226)	(0.4860, 0.8801)	(0.8913, 0.1730)	(0.7621, 0.9797)
(0.0185, 0.2523)	(0.8214, 0.8757)	(0.4447, 0.7373)	(0.6154, 0.1365)	(0.7919, 0.0118)	(0.9218, 0.8939)
(0.1763, 0.2987)	(0.4057, 0.6614)	(0.9355, 0.2844)	(0.9169, 0.4692)	(0.4103, 0.0648)	(0.8936, 0.9883)
(0.3529, 0.4235)	(0.8132, 0.5155)	(0.0099, 0.3340)	(0.1389, 0.4329)	(0.2028, 0.2259)	(0.1987, 0.5798)
(0.2722, 0.5298)	(0.1988, 0.6405)	(0.0153, 0.2091)	(0.7468, 0.3798)	(0.4451, 0.7833)	(0.9318, 0.6808)
(0.4186, 0.5678)	(0.8462, 0.7942)	(0.5252, 0.0592)	(0.2026, 0.6029)	(0.6721, 0.0503)	(0.8381, 0.4154)
(0.6813, 0.8744)	(0.3795, 0.0150)	(0.8318, 0.7680)	(0.5028, 0.9708)	(0.7095, 0.9901)	(0.4289, 0.7889)
(0.1897, 0.4983)	(0.1934, 0.2140)	(0.6822, 0.6435)	(0.3028, 0.3200)	(0.5417, 0.9601)	(0.1509, 0.7266)
(0.6979, 0.4120)	(0.3784, 0.7446)	(0.8600, 0.2679)	(0.8537, 0.4399)	(0.5936, 0.9334)	(0.4966, 0.6833)
(0.8216, 0.8392)	(0.6449, 0.6288)	0.8180, 0.1338)	(0.6602, 0.2071)	(0.3420, 0.6072)	(0.2897, 0.6299)
(0.5341, 0.5751)	(0.7271, 0.4514)	(0.3093, 0.0439)	(0.8385, 0.0272)	(0.5681, 0.3127)	(0.3704, 0.0129)
(0.5466, 0.6831)	(0.4449, 0.0928)	(0.6946, 0.0353)	(0.4565, 0.2714)	(0.7382, 0.1991)	(0.0579, 0.5828)
(0.6038, 0.7604)	(0.4660, 0.4611)	(0.0196, 0.3050)	(0.3046, 0.4387)	(0.6213, 0.6124)	(0.8998, 0.2126)
(0.3412, 0.375)	(0.7027, 0.3840)				

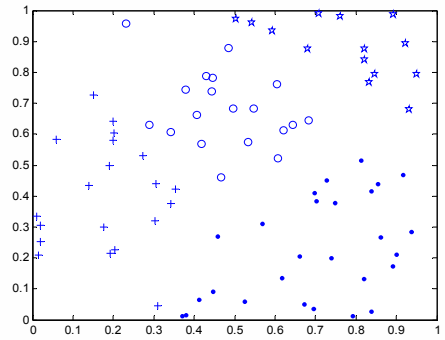

Figure 2. (Left) The first clustering diagram of 80 data points with the original algorithm

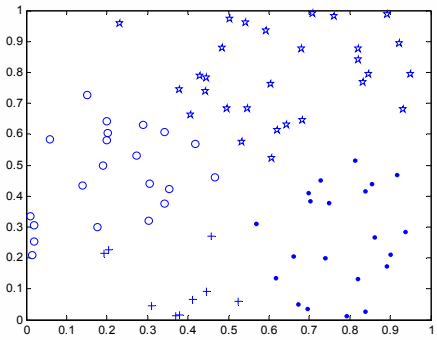

Figure 3. (Right) The second clustering diagram of 80 data points with the original algorithm

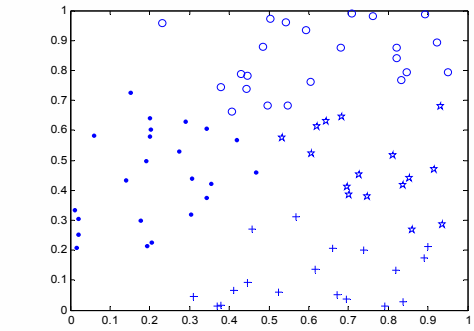

Figure 4. (Left) The third clustering diagram of 80 data points with the original algorithm

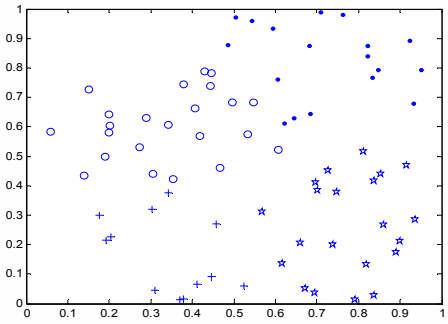

Figure 5. (Right) The clustering diagram of the improved algorithm after eliminating six isolated point

Improve the algorithm for clustering analysis. Firstly, eliminate six isolated point: (0.8936, 0.9883), (0.0153, 0.2091), (0.0185, 0.2523), (0.2311, 0.9568), (0.0099, 0.3340), (0.0196, 0.3050). After the step, only 74 data objects remain. Next, make sure two initial clustering centers, whose data coordinates are (0.9218, 0.8939) and (0.3093, 0.0439). Then, confirm the other two initial clustering centers, whose data coordinates are (0.4259, 0.7889) and (0.8913, 0.1730).

Input the initial clustering center and the clustering number K of the improved algorithm. The improve algorithm conducts clustering to get the final clustering centers: (0.7197, 0.8267), (0.3427, 0.1658), (0.3422, 0.6044), (0.7752, 0.2628). According to the principle of the nearest to the center, ensure the ownership of the isolated points. (0.8936, 0.9883) belongs to the solid point kind; (0.2311, 0.9568) belongs to the circle kind. (0.0153, 0.2091), (0.0185, 0.2523), (0.0099, 0.3340) and (0.0196, 0.3050) belong to the plus kind.

Table 2: Clustering table of random data

Algorithm	The initial clustering centers	The clustering centers after 100 iterations	The clustering results after 100 iterations	CPU time (second)
The original algorithm (Fig.2)	(0.7382, 0.1991), (0.3420, 0.6072) (0.6038, 0.7604), (0.5936, 0.9334)	(0.6988, 0.2235), (0.1828, 0.4015) (0.4775, 0.6804), (0.7721, 0.8817)	27 solid points, 20 plus points,19 circle points,14 five-pointed stars	At least 9 iterations 0.0470
The original algorithm (Fig.3)	(0.8385, 0.0272), (0.5681, 0.3127) (0.0185, 0.2523), (0.8214, 0.8757)	(0.7752, 0.2628), (0.3658, 0.1111) (0.2127, 0.4672), (0.6333, 0.7933)	21 solid points,9 plus points,21 circle points,29 five-pointed stars	At least 8 iterations 0.0460
The original algorithm (Fig.4)	(0.0153, 0.2091), (0.0185, 0.2523) (0.0099, 0.3340), (0.0196, 0.3050)	(0.2114, 0.4457), (0.6158, 0.1144) (0.6237, 0.8409), (0.7571, 0.4802)	23 solid points,18 plus points,23 circle points,16 five-pointed stars	At least 11 iterations 0.0490
The improved algorithm (Fig.5)	(0.9218, 0.8939), (0.3093, 0.0439) (0.4289, 0.7889), (0.8913, 0.1730)	(0.7197, 0.8267), (0.3427, 0.1658) (0.3422, 0.6044), (0.7752, 0.2628)	18 solid points,12 plus points,23 circle points,21 five-pointed stars	At least 5 iterations 0.0310

Experiments show that it is easy for the original K-means algorithm to make the clustering results generate errors due to the random selection of the initial cluster centers. In addition, the original iterations are too much. The difference between the clustering results and the real data distribution is obvious. The improved algorithm

eliminates all the isolated points with accuracy requirements, and chooses the proper initial cluster centers by the circle theory, so the clustering result of the improved algorithm has higher quality and better effect.

5. Conclusions

Partition-based K-means algorithm is a classical clustering algorithm, but the algorithm itself has some shortcomings. For example, the initial clustering number K must be specified in advance. In the initial clustering, the choice exists randomness, which the algorithm is easy to generate the local optimal solution, which is much affected by isolated points. This paper has proposed a new method of choosing the initial clustering centers, and before the clustering, exclude the impact of the isolated points on the algorithm. The experiments show that the proposed algorithm improves the clustering results, reduces the time complexity of the algorithm, and improve the quality of the clustering algorithm.

Acknowledgments

The work in this paper is supported by the Natural science fund for colleges and universities in Jiangsu Province, China (Grant No. 15KJD520008), and University Philosophy Social Science Fund Program of Jiangsu, China (2012SJD630079).

References

Bu Y.Y., Guan Z.R., 2009, the research based on the K-means clustering algorithm. Journal of Southwest University for Nationalities (Natural Science Edition), 35(1), 198-200, DOI: 10.3969/j.issn.1003-2843.2009.01.049.

He T., 2012, CLARANS of Uncertain Objects Based on Bayesian Probability. Proceedings of the 2012 Second International Conference on Electric Information and Control Engineering. IEEE Computer Society, 3, 203-206, DOI: 10.1109/ICEICE.2012.943.

Lian F.N., Wu J.L., Tang Q., 2008, An improved K-means clustering algorithm, Computer and information technology, 16(1), 38-40, DOI: 10.3969/j.issn.1005-1228.2008.01.014.

Li Y.L., Qin Z., 2007, An improved K-means algorithm, Journal of Beijing Institute of Graphic Communication, 15(2), 63-65, DOI: 10.3969/j.issn.1004-8626.2007.02.018.

Lu S.L., Lin S.M., 2004, Distance-based isolated point's detection research, Computer Engineering and Applications, (33), 73-75, DOI: 10.3321/j.issn:1002-8331.2004.33.022.

Zhao G.F., Qu G.Q., 2006, Analysis and implementation of CLARA algorithm on clustering. Journal of Shandong University of Technology (Science and Technology), 20(2), 45-48, DOI: 10.3969/j.issn.1672-6197.2006.02.014.

Zhong Z., Zhi C., Yuan L., 2010, The research of K-means algorithm based on the weighted Euclidean distance, Jo