An Improved K-means Clustering Algorithm Based on Meliorated Initial Centre

Xiang Li[1,2,3,*], Zhenwei Wei[2,3,4] and Lingling Li[1,2,3]
[1]School of Computer Science, Zhengzhou University of Aeronautics, Zhengzhou, China,
[2]Collaborative Innovation Center for Aviation Economy Development,
[3]Institute of Aviation Industry Technology, Henan Aviation Economy Research Center,
[4]School of Aeronautical Engineering, Zhengzhou University of Aeronautics, Zhengzhou, China
*Corresponding Author

Abstract—**The initial clustering center of traditional K-means clustering algorithm is selected at random that different initial clustering center will get different clustering results, which have great randomicity and poor stability. To improve the K-means clustering algorithm optimized by adopting local outlier index, we adopt a positive approach by calculating the local outlier index of all data samples. Then k local dense points with furthest mutual distance were selected as the initial clustering center. At last, in this paper we eliminate the effects of local outlier using in the improved algorithm. The experimental results showed that this enhanced algorithm could reduce the susceptibility by using K-means clustering algorithm to select the initial clustering center, and also short number of iterations. In brief, the K-means Clustering Algorithm Based on Meliorated Initial Centre obtains a more accurate clustering result.**

Keywords-clustering; outlier index; initial clustering center; k-means clustering

I. Introduction

Data clustering (or clustering analysis) is to obtain the natural classification relation of chart pattern, point set or object set [1]. With reference to the idea of "Things of one kind come together", the clustering analysis classifies the data object into multiple categories (or cluster) by extracting the potential structure of sample data through research, so that makes the objects in the same category to have a higher similarity, but great difference between the objects in different category. Since the formation of category is completely driven by the data, without requiring any priori information and assumptions, therefore the clustering analysis is an unsupervised learning method, which is widely used in the data mining and data analysis [2].

K-means clustering algorithm is a partition-based clustering algorithm, this algorithm is simple, rapid and widely applied. [3]However, the initial clustering center of traditional K-means clustering algorithm is selected at random in the data set, and the quality of clustering result depends on the selection of clustering center. So this algorithm usually gets the locally optimal solution rather than the global optimal solution, and the clustering results are very unstable. Aiming at this disadvantage, the scholars have conducted extensive researches, the recent research achievements, for example, Literature [4] proposes a K-centroid combination clustering algorithm based on the competitive learning, Literature [5] proposes an improved algorithm of density-based algorithm optimizing initial clustering center, Literature [6] builds k data sets by using greedy algorithm, taking the means of data in the set as the initial clustering center. These studies had one thing in common that all of them started from the selection method of clustering center, they essentially made the initial selection for the data set according to a certain strategy, the preference was given to those having higher cohesion as the alternative of clustering center, from which choose and specify it as the clustering center of K-means clustering algorithm.

For the defect of initial clustering center selected at random, this paper proposed an improved K-means clustering algorithm based on the local outlier index, which firstly adopts LOF (local outlier factor)[7] algorithm to calculate the local outlier index of every point in the data sample, then exclude the outlier in the sample from the selection of initial clustering center, select k local dense points with furthest mutual distance as the initial clustering center, eliminating the fluctuation of clustering results. The experiment showed that compared with K- traditional means clustering algorithm, this initial clustering center optimal algorithm could effectively improve the clustering algorithm to obtain the stable clustering result with higher accuracy rate.

II. K- Traditional Means Clustering Algorithm

A. *K- traditional Means Clustering Algorithm*

Traditional K-means clustering algorithm is described as follows:

Input: data set X with n objects and clustering grouping number k.

Output: k clusterings to minimize the objective function E.

(1) Randomly select k objects from the sample data set as initial clustering center;

(2) Allocate each object of data set to the corresponding cluster according to the principle of nearest distance;

(3) Recalculate the clustering center for each cluster, the new clustering center is the mean of all objects in the cluster;

(4) Calculate the objective function E;

(5) If the objective function E converges, the algorithm ends, otherwise turn to (2).

B. Limitation of Traditional K-means Clustering Algorithm

The initial clustering center in K- traditional means clustering algorithm is selected at random, and the clustering result varies from different selection of initial clustering center. Using K-means clustering algorithm for Iris data in UCI database is divided into 3 categories, 8 times of experimental results are shown in the data in Table I.

It can be seen from the experimental results that K-traditional means clustering algorithm is quite heavily influenced by the initial clustering center. The accuracy rate of categories is different according to the different selection of initial clustering center, the difference between the best rate and the worst rate in 8 times of experiments is 40.7%. Concerning this problem, the common method is to calculate clustering results by randomly generating initial clustering center for many times, select one time of calculation result with minimum objective function as the final clustering result, like the implementation of K-means cluster in Matlab.

TABLE I. EXPERIMENTAL RESULTS FOR IRIS DATA BY K-TRADITIONAL MEANS CLUSTERING ALGORITHM

Initial center	Correct Category	Wrong Category	Accuracy
3, 7, 87	87	63	58%
9, 18, 101	86	64	57.3%
11, 56, 53	132	18	88%
35, 23, 88	72	78	48%
36, 25, 56	77	73	51.3%
14, 11, 97	86	64	57.3%
35, 6, 21	132	18	88%
100, 105, 144	133	17	88.7%

The disadvantage of this method is that of long time-consuming, unstable test results and it may not be able to get the optimal result. Therefore a reasonable choice of initial clustering center is the key to improve K-means clustering algorithm.

III. OPTIMIZATION OF INITIAL CLUSTERING CENTER BASED ON LOF

In order to overcome the problem of K-means clustering algorithm sensitive to initial clustering center, this paper introduced a method of calculating the local outlier index of data object in the improvement of clustering algorithm, avoiding initial clustering center to choose the local outlier, so as to enable the clustering center to reflect the distribution characteristics of data.

A. Local Outlier Index

D. Hawkins provides the definition of outlier [8]: the outlier refers to the object of observation deviating from other objects and generated by the exceptional mechanism principle, and the outlier is usually called exceptional point, isolated point, singular point. There are many detection methods of outlier[9,10], this paper adopts LOF algorithm, which calculates the local outlier index of each object in the data set, determining the local outlier by comparing the size of the index. The bigger the index, the smaller the distribution density of object in the area near this outlier, and this object is more isolated.

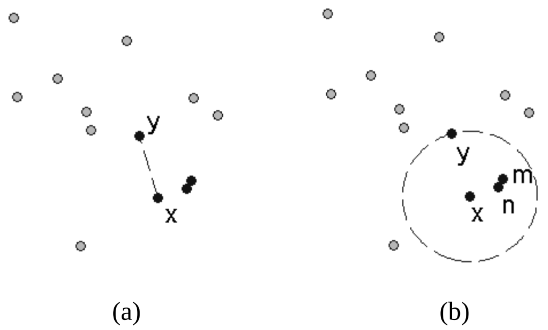
LOF algorithm is described as follows:

(1) Calculation of k-distance of object x. For a given positive integer k, k-distance of object x is denoted by $k\text{-dis}(x)$, in the sample space, the object y exists, the distance between it and object x is denoted by $\text{dis}(x,y)$.

If the following two conditions are met, we believe that $k\text{-dis}(x) = \text{dis}(x,y)$

has at least k objects z in the sample space, get $\text{dis}(x,z) \le \text{dis}(x,y)$;

- has at least k-1 objects z in the sample space, get $\text{dis}(x,z) < \text{dis}(x,y)$.



(a) (b)

FIGURE I. CALCULATION OF LOCAL OUTLIER INDEX

As shown in Figure I (a), when k = 3, calculate the distance of all points in the bi-dimensional sample to data object x, the third data object with the nearest distance is y, as shown in Figure I (b), it exists that the distance x of the two of m and n data objects is less than $\text{dis}(x,y)$.

(2) Finding out k-neighborhood of object x. After k-distance of object x is calculated, the collection of objects with the distance between object x of not more than k-distance is called k-neighborhood of object x, which is denoted by $N_{k-dis(x)}(x)$. As shown in Figure I (b), that is object m, n and y.

(3) Calculation of reachable distance of object x The reachable distance is denoted by $reachdis_k(x,y)$. The reachable distance of object x relative to object y may be calculated by using the following formula:

$$reachdis_k(x,y) = \max\{k - dis(y), dis(x,y)\}$$

(4) Calculation of local reachable density, as Equation (1). The local reachable density is reciprocal of mean value of reachable distance based on k-closest point of x.

$$L_{den}(x) = \frac{\sum\limits_{y \in N_{k-dis(x)}(x)} reachdis_k(x,y)}{\left|N_{k-dis(x)}(x)\right|} \qquad (1)$$

(5) Calculation of local outlier index, as Equation (2).

$$LOF(x)=\frac{\sum_{y\in N_{k-dis(x)}(x)}\frac{L_{den}(x)}{L_{den}(y)}}{\left|N_{k-dis(x)}(x)\right|} \tag{2}$$

The above are the steps of calculating the local outlier index, the bigger the local outlier index of object x, the higher the abnormality degree of x, otherwise it may be smaller. The objects are arranged in the descending order according to the size of local outlier index of data collection object, so that the further forward arranged, the bigger the outlier of object; the further backward arranged, the higher the density degree.

B. *Advanced K-means Clustering Algorithm Based on Local Outlier Index*

The advanced K-means clustering algorithm based on local outlier index is described as follows:

Input: data set X with n objects and clustering grouping number k.

Output: k clustering to minimize the objective function E.

(1) Calculate the local outlier index of each data object;

(2) Delete the data object points of local outlier index ranking at the first n, obtaining the data object collection D situated at the intensive area.

(3) Consider the minimum data object of local outlier index as the first clustering center o_1, and add o_1 into the initial clustering center collection; find the point o_1, which is farthest from the point o_2, from collection D as the second clustering center, and add it into the initial clustering center collection, and delete o_1 and o_2 from D.

(4) Find the maximum sum of the distance between o_3 and o_1, and between o_3 and o_3 from collection D, then add o_2 into the initial clustering center collection, and delete o_3 from collection D, and continue to find all the points of object collection with the furthest object distance from the initial clustering center from collection D as the clustering center till the No. k is found.

(5) Carry out clustering to the entire data set with k clustering centers obtained and the K-means clustering algorithm.

IV. Experimental Results and Analysis

Test the refinement algorithm of this paper with multiple data sets in the UCI database, and the results and analysis are as follows.

A. *Stability Analysis*

Iris, as one of the data sets that are widely used, contains 150 data, which can be divided into three classifications with 50 data for each one. Each data describes the properties of flowers with four real numbers, symbolizing the length and width of sepal, and the length and width of petal respectively.

Carry out the clustering analysis to Iris data with the method of this paper and take the three attribute data as coordinates. The test results obtained are as shown in Figure. II. It can be concluded from the figure that in the Iris data set, the first data, setosa, is further from other data, while the second data, versicolor, and the third data, virginica, are overlapped partially.

In the example shown in Figure II, LOF will be able to capture both anomalies due to the fact that it considers the density around the data instances. Several researchers have proposed variants of LOF technique. Some of these variants estimate the local density of an instance in a different way. Some variants have adapted the original technique to more complex data types. With the increase of value parameter K, local outlier factor (LOF) and values have no significant change trend of increase or decrease.

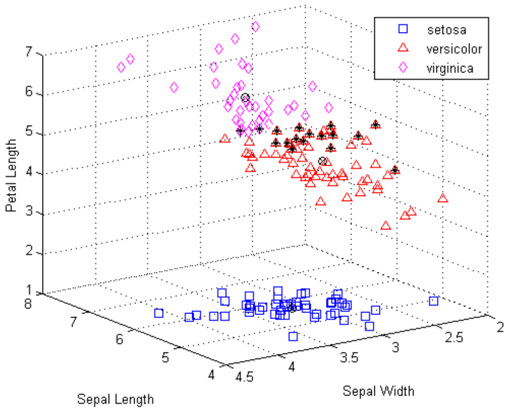


FIGURE II. IRIS DATA CLUSTERING RESULTS

As shown in Figure III, the value of the parameter K increased from 2 to 10, the maximum LOF change is up and down, but there is obviously no significant magnificent change.

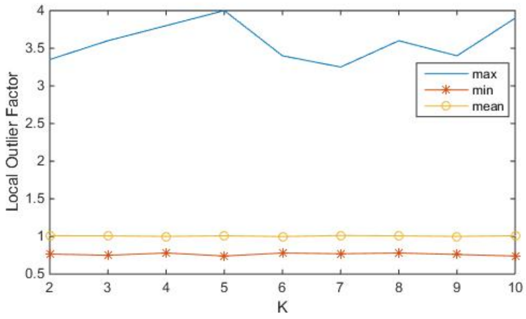


FIGURE III. THE RELATIONSHIP BETWEEN K VALUE AND LOCAL OUTLIER INDEX

But it should be noted that with the increase of K value, the number of adjacent objects need to compute are dramatically increasing, and the complexity of the algorithm is also increased accordingly. It involves the K adjacent object query efficiency. According to the experience of multiple experiments, the entire data object with the appointed local outlier index of over 1.5 shall be considered as the local outlier, and while calculating the local outlier index, the parameter k shall be set as one-tenth of the amount of the entire sample data set.

The experimental results of Iris data set show that the improved clustering algorithm has a high accuracy of 86.7%, and it can be known by comparing with the results of the traditional K-means clustering algorithm that the classification accuracy of the algorithm of this paper is superior to the average accuracy of the traditional algorithm. Besides, the improved algorithm can optimize the clustering center through the local outlier index calculation, so this algorithm has a better stability in clustering results and has favorable robustness to obtain a better clustering accuracy as a whole. In addition, the data that are classified incorrectly by the improved algorithm are mainly between versicolor and virginica, which is the same with the incorrectly classified data of the traditional K-means clustering algorithm in distribution. To a certain extent, it shows that the two classifications of data in Iris data set have poor distinguishability.

B. *Analysis of Clustering Accuracy and Time Complexity*

Test Iris, Haberman and Hayes-Roth in UCI database, and after the algorithm of this paper and the traditional K-means clustering algorithm operate for 20 times, the contrast of average clustering accuracy, average iterations and operating time is as shown in Table II.

TABLE II. RESULTS COMPARISON OF K-MEANS CLUSTERING ALGORITHM AND THE ALGORITHM OF THIS PAPER

Algorithm	Data Set	Clustering Accuracy	Iterations	Operating Time(ms)
K-means Clustering	Iris	81.6%	9	16
	Haberman	51.96%	7	12
	Hayes-Roth	69.3%	12	20
Algorithm of this paper	Iris	86.7%	4	125
	Haberman	73.82%	3	110
	Hayes-Roth	80.4%	5	164

It can be concluded from Table II that the clustering accuracy of the algorithm of this paper in Iris, Haberman and Hayes-Roth data sets is much higher than the average accuracy of traditional algorithm, and the iterations are less, but the operating time of improved algorithm is a bit longer. This shows that the improved algorithm clustering accuracy proposed by this paper has been improved significantly, but due to the introduction of local outlier index calculation into the improved algorithm, which complicates the algorithm and adds one order of magnitude to the operating time. It can be concluded after overall consideration that the algorithm of this paper can be applied if the data set has a small number of data and lower requirements to time.

V. Conclusion

This paper proposes a new kind of initial clustering center selection algorithm, which together with the introduction of local outlier index, effectively overcomes the sensibility of K-means clustering algorithm to the initial clustering center.

The experiment of multiple data sets shows that the improved K-means clustering algorithm based on the outlier index has stable clustering results and keeps higher accuracy. However, the introduction of local outlier index calculation increases the operand, so the next step should aim to research the method of filtrating data with pertinence and discuss the method of decreasing the operand of algorithm so as to improve the arithmetic speed.

Acknowledgment

This research is founded by the foundation item of National Natural Science Foundation NO. 41171341, and also the Colleges and universities in Henan province scientific research key project of 15B416003, 16A590003.

References

[1] Jain. A. K. Data clustering: 50 years beyond k-means[J]. Pattern Recognition Letters. 2010, 31(8), pp. 651-666.

[2] HAN J. W., Kamber Micheline, PEI J. Data mining concept and techniques[M] . 3rd ed. [S.l.]: Morgan Kaufmann –Elsevier, 2012 , pp. 19-26. Feng Chao. Research on the K-means Clustering Algorithm [D]. Dalian: Dalian University of Technology, 2007, pp. 25-29.

[3] Zhang Yu, Shao Liangshan. K-centroid Combination Clustering Algorithm Based on the Competitive Learning [J]. Computer Engineering, 2011, 37 (15), pp. 40-42.

[4] Fu Desheng, Zhou Chen. Improved K-means Algorithm Based on Density and Implementation [J]. Computer Application, 2011, 31(2), pp. 432-434.

[5] Tong Xuejiao, Meng Fanrong, Wang Zhixiao. Optimization of K-means Initial Clustering Center [J]. Computer Engineering,and Design 2011, 32 (8), pp. 2721-2723.

[6] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, et al. LOF: Identifying Density-Based Local Outliers [C]. Proceedings of ACM SIGMOD International Conference on Management of Data, Dalles, TX, 2000, pp. 93-104.

[7] Xu Xiang, Liu Jianwei, Luo Xionglin. Research on Outlier Mining [J]. Application Research of Computers. 2009, 26(1), pp. 34-39.

[8] Xue Anrong, Ju Shiguang, He Weihua, et al. Research on Local Outlier Mining Algorithm [J]. Chinese Journal of Computers. 2007, 30(8), pp. 1455-1463.

[9] Yu, Q., Luo, Y., Chen, C. et al. Outlier-eliminated k-means clustering algorithm based on differential privacy preservation [J]. Applied Intelligence. 2016 , pp. 1-13.