Application and Research of Distance and Density on Improved K-means

Hongbo GU
College of Computer & Information Technology, Northeast Petroleum University, China
a Hongbo Gu: zjhgb@163.com.cn

Abstract K-means algorithm is introduced. An improved algorithm is proposed for the disadvantages of randomly selecting the initial center of clustering and the vulnerability to the effects of outliers. The Distance Mean method (DM) was used to remove the outliers, then high-density and max-distance (HDMD) was used to improve the selection of the initial center of clustering. The comparison experiment before and after the improvement was carried out. Experimental results show that the improved algorithm is stable and accurate. The improved algorithm was applied to computer language teaching and achieved good classification effects. The improved algorithm is used to research and analyse mobile customer records, the effect is in conformity with the actual situation.

1. Introduction
The most commonly used clustering algorithm is k-means. [1] This algorithm is more efficient; it has been applied widely in various fields. Its disadvantages are different initial cluster centers will affect the final clustering results [2]. The effect of the algorithm is affected by the outliers.

2. Improved algorithms
The data is not very specific, the number of attributes is not very large, the data set can be used the following improved method.

2.1 Remove the outlier
The outliers involved in the improvement are based on the distances [3], so you need to use definitions to remove the outliers. The distance between the two data objects is calculated, and the sum of the two data objects is calculated. Then calculate the mean of the sum of the data objects ŝ.

Definition 1 Outlier [4] Get the deviation of the data object \( x_i \) from other data objects \( T_i = x_i - \bar{x} \), and be stored as the data set \( T_i \). If \( T_i > 0 \), the data \( x_i \) is an outlier based on distance.

According to the above definition, algorithm 1 can be obtained.

Algorithm 1: In initial time , the number of data sets is \( n \), the number of dimensions is \( n \) dimensions, the data set \( U \) has \( n \) data objects, and \( t \) is used to record the number of outlier, \( t = 0 \), the set of outlier is \( A_t \).

1. The distance of all data objects in \( U \) is calculated and the data set \( S \) is obtained.
2. Scan the \( U \) and calculate the deviation data set for each data object in the \( U \). Scan \( T_i > 0 \) \( x_i \) is outlier, \( t++ \), and \( A_t = \{x_i\} \), \( U = \{U - x_i\} \);
3. Repeatedly scan the \( D_i \) until all of the data object \( T_i \leq 0 \); and \( A_c = \{x_1, x_2, x_i\} \), finally get out the outlier data set \( U' = U - A_i \).
2.2 Selection of initial cluster center

Another shortcoming of K-means is selected randomly k for the initial cluster center. Because of the random selection, the initial points are different, and the results of the clustering are different [5]. In literature [6] and [7], the selected K in high density region which is the farthest distance from each other, and that is proposed as the initial cluster center. A method based on HDMD (High-Density and Max-Distance) is proposed to determine the initial clustering center.

**Definition 2** The parameter of density of 2 object [8] For any data object \( x_i \), take the \( X_i \) for the center, the number of data objects in the area with the radius of \( \text{avgD} \) is called the density parameter of \( x_i \), and it is called \( \text{denP}_i(x_i, \text{avgD}) \).

The parameter \( \text{avgD} \) is usually calculated by the empirical values. It uses a calculation that depends on k.

\[
\text{avgD} = \frac{\text{MaxDis}}{2 \times k}.
\]

And \( \text{MaxDis} \) is the maximum distance between two data objects.

**Algorithm 2** [9]:
1. Calculate the distance between every two data objects \( d(X_i, X_j) \);
2. Calculate the density parameter of every data object \( \text{denP}_i \), and obtain the data object set M in high-density area;
3. Obtain \( \text{Max} \left( \text{denP}_i \right) \), make this object point to the first center \( c_1 \), \( C = \{c_1\} \);
4. Take the data object with the longest distance of \( c_1 \) as the second initial center, \( c_2 \), \( c_2 \in M \);
5. The \( c_3 \) to is the \( \text{Max}(\text{Min}(d(x_i, c_1), d(x_i, c_2))) \), \( i=1,2...,n \), \( j=1,2...,k-1 \). The data objects \( x_i, c_j \in M \), repeat step (5), until you find the number of samples in the M reached K.

Algorithm 2 is used to deal with initial clustering centers. There is no real clustering.

2.3 Improved algorithm

Suppose that the number of the data object set U is n and has p dimensional data objects, which are divided into k classes. Step: (1) Use the algorithm 1 to remove outlier and set \( A_t \). (2) Use the algorithm 2 to select the initial cluster center. (3) The classical K-means is used for cluster analysis.

2.4 Improved experiment and analysis

In this experiment, famous data sets, such as Wine, Iris and Glass were used to compare the classical k-means and HDMD k-means. The UCI database is a database that is used to test machine learning and data mining algorithms. The data objects in the library have a clear classification, so UCI data can be used to detect the accuracy of the clustering algorithm. The above 3 sets of test data sets are tested with the K-Means and the HDMD K-means. Table 1 to table 3 are the clustering accuracy of the 3 sets of data sets under different clustering algorithms.

The accuracy rate here is \( AR = m/M \). The m represents the number of data objects that is correctly clustered, and M represents the number of data objects that should be present.

**Table 1** The result of both improved and classic on Wine

| AR %         | Max | Min | Average | Data set |
|--------------|-----|-----|---------|----------|
| K-means      | 84.01 | 54.37 | 69.19   | Wine     |
| HDMD         | 84.57 | 60.21 | 72.39   |          |

Thirteen attributes are the 13 chemical components of a wine with 178 samples, divided into 3 categories. The best is 84.01 and the worst is 54.37. The best change of HDMD k-means is not very obvious, but the worst result has increased by nearly 6%.

**Table 2** The result of both improved and classic on Iris

| AR %         | Max | Min | Average | Data set |
|--------------|-----|-----|---------|----------|
| K-means      | 91.23 | 66.57 | 78.9    | Iris     |
| HDMD         | 92.39 | 76.28 | 84.34   |          |
The classic k-means algorithm can reach 91.23 at best and 66.57 at worst on the Iris data set. This indicates that this data set is suitable for this algorithm. The highest accuracy rate of using mdhd-k-means reached 92.39, and the lowest accuracy rate reached 76.28%.

Table 3 The test result of both improved and classic on Glass

| AR %  | Max   | Min   | Average | Data set |
|-------|-------|-------|---------|----------|
| K-means | 74.71 | 51.36 | 63.04   | Glass    |
| HDMD  | 80.12 | 51.38 | 65.75   | Glass    |

For Glass datasets, the highest accuracy rate based on the HDMD k-means was 80.12, and the lowest was 51.38. The highest accuracy rate has increased by 5%.

You can see the best results apply to the Iris data set. Because the data attribute and quantity of this dataset are less, this result is the best.

The Iris test data set was tested by the k-means and HDMD-k-means. Table 4 and 5 is the accuracy rate of two algorithms in Iris dataset.

Table 4 The accuracy rate of classic clustering on Iris

| order | cluster center | accuracy rate |
|-------|----------------|---------------|
| 1     | 38,53,109      | 80.67         |
| 2     | 52,45,115      | 80.17         |
| 3     | 35,67,130      | 91.23         |
| 4     | 118,48,87      | 81.37         |
| 5     | 48,105,22      | 77.34         |
| average |             | 82.16         |

As you can see, these data centers are close to the real dataset.

Table 5 The accuracy rate of the improved clustering on Iris

| order | HDMD-K-means |
|-------|--------------|
| cluster center | accuracy rate |
| 1     | 30,102,55    | 83.47         |
| 2     | 55,114,32    | 92.39         |
| 3     | 72,44,105    | 86.58         |
| 4     | 47,92,146    | 81.37         |
| 5     | 79,102,55    | 81.48         |
| average |             | 85.06         |

From the above experiments, it can be seen that the classical k-means can be accurately classified, and the random selection of Iris data set data will definitely influence the classification effect. The experiment shows that HDMD is used to find the initial clustering center, and the classification effect is better than the classical k-means. The randomness of the cluster center is removed, and the initial
clustering center is more suitable for the actual data distribution, and is also suitable for clustering the actual data.

3. Application and analysis

3.1 Application in computer teaching in colleges and universities

Students’ achievement plays a certain directive role in the whole teaching management activities. For example, the school’s identification of the students’ quality is almost linked to the results or has a very close relationship. On the one hand, analyse these test scores, so we can find out which knowledge points are well mastered and which need to be paid attention to in the future teaching. On the other hand, it is necessary to cluster the students’ performance data and analyse the useful information of students’ learning, so that the teachers can make corresponding teaching rules base on the experimental results and teach them according to their aptitude. The following test is a collection of data collected from the data collected by the class of 2,055 students in 76 classes. Select the data closely related to the teaching, and intercept some standardized data. The number after the first line of text is the score of each item. In order to find out the reasons that their grades are low and the achievement is unreasonable. 7 students missed the test, and 10 students took the test with 0. So 14 people took the exam, but the score was less than 20, which is the appearance of abnormal information, which can be regarded as outlier. Table 6 is the average of the 10 times accuracy of classical algorithm clustering analysis, clustering analysis after remove outliers and clustering analysis using HDMD.

The experiment was carried out according to the improved algorithm. Development environment: MATLAB R2015a, each algorithm runs 10 times, taking its average results.

| Conditions | Average correct rate of 10 times (%) |
|------------|------------------------------------|
|            | classic | Remove isolated | HDMD     |
| k          |         |                 |          |
| 4          | 51.29   | 51.26           | 57.15    |
| 5          | 59.19   | 60.07           | 61.49    |
| 6          | 61.72   | 65.43           | 69.29    |
| 7          | 61.27   | 63.78           | 65.07    |
| 8          | 58.17   | 59.75           | 62.37    |

3.2 Application analysis

After clustering the data, it is impossible to determine the state of the 2055 students in advance, so the number of clusters in this experiment varies from 4 to 7. In this application, there are 31 outliers, which can detect the effectiveness of the outliers based on the improved algorithm, and then select the initial clustering center using HDMD. From the algorithm accuracy rate, the classification results of classical algorithm are lower. After remove the outliers, the classification effect is better than it, which proves that the outliers have great influence on the results of cluster analysis. So, the first step in improved the algorithm will remove the outliers. It can be seen that, under either condition, the classification is best when k=6. The first kinds of students are proficient in learning programming language. And they have mastery of language and they are best. The learning habits of these students are almost amazing, and they are used to learn every course, preview carefully before has the class, perfect notes in class, and review in time outside of class, never miss the class. Language is based on complexity, after learning the knowledge points; we should pay more attention to the learning methods. Each program should be implemented in different ways and discover the problems and innovate. The second kinds of students can draw inferences from others. They listen carefully, and review in class.

Table 6 Comparison of the algorithms
But the program can also design 2 or 3 methods. Third class of students can do it according to the book. They learn from books carefully and remember them carefully. They don't study as hard as the second students, and there are always missing knowledge points, there is always a kind of problem is defective, such as sorting data, functions or sub. The fourth kind of students are “who know what they are”, and they are also dealing with teachers' demands and studying for grades. The fifth kind of students are an introduction, they think that learning is only a small part of life. In class, they copy their homework from here to there, and such as the students usually failed the exam on some subjects. The sixth kind of students is the layman of computer language.

From the above analysis, the improved algorithm removed the influence of results based on outlier, then calculated according to the initial clustering center after HDMD to find data, the resulting initial cluster center is more consistent with the actual distribution of data. But the running time of this experiment is longer than the classical algorithm, and the calculation quantity is relatively large, after all, that is not suitable for large data. We can find the application effect analysis in the UCI database that is better. It has something to do with the size of the data and the amount of attributes. It's something that we’ll work out later.

4. The improved algorithm is applied in practice

The final goal of user’ analysis will obtain the detailed classification results that can be explained technically and executed operationally. And we make a reasonable marketing plan and develop a reasonable marketing plan. First, the user analysis using clustering technology is determining the target customer base. The target of this application is all mobile phone customers in a certain area, dated from January 1 to 20, 2017. The data table includes 10 variables, including total cost, local phone fee, monthly fee, domestic traffic fee, provincial traffic fee, local traffic fee, provincial call, domestic call, foreign call, short message, and so on. There are 12,495 data objects which there are some abnormal data. According to the business requirements, these data are cluster analysis. The clustering results shown in table 7 are obtained.

| Conditions | Average correct rate of 20 times (%) |  |
|------------|----------------------------------|---|
|            | classic              | Remove isolated | HDMD |
| 4          | 50.79                | 52.27           | 58.15 |
| 5          | 63.31                | 65.78           | 70.17 |
| 6          | 59.31                | 62.71           | 65.39 |

From the experimental results, the data accuracy rate is highest when divided into 3 categories. It can be seen that when there is a large amount of data and there are certain isolated points, the classical algorithm accuracy rate is lower than the latter two algorithms. When the amount of data is large, the improved algorithm is more efficient than the original algorithm. From the business view, according to their consumption costs can be divided into high-class mobile customers, medium-sized customers, and ordinary customers. As the last step in the application of cluster analysis, customer group strategy is formulated.

5. Conclusions

In this paper, the classical k-means is presented, and the effect of the initial cluster center on the clustering results is proposed in the algorithm, and an improved algorithm is proposed. The experimental results show that the improved algorithm can obtain high and stable accuracy, and it is more suitable for clustering with less real data and less properties.
References

[1] Zhiwei Ye, Yu Jie Yin. A Clustering Approach Based on Cuckoo Search Algorithm. MiE.Comp., 32,104-110 (2015)

[2] Marques JP. Pattern Recognition Concepts, Methods and Applications. (2002)

[3] QiLin Yu. Optimization of Initial Clustering Centers Selection Method for K-Means Algorithm. Comp.Sys & App, 26,170-174,(2017)

[4] Hongbo Gu. An Improved k-means Algorithm Based on Outliers and Original Clustering Center. J. Yangtze U.(Nat Sci Edit),9,45-49,(2009)

[5] Jintao LI,Ping AI. Improvement of clustering algorithm based on K-means. F. Elec Meas T.,36,9-13,21,(2017)

[6] Fang Yuan, Zhi-yong, Zhou K-means Clustering Algorithm with Meliorated Initial Center. C. E.33,65-66,(2007)

[7] Yuxia Lai, Jianping Liu. Optimization study on initial center of K-means algorithm.CE.APP.44,147-149,(2008)

[8] Hongbo Gu. The optimization of original clustering center based on the clustering algorithm. J.Xi'an Polytechnic U.24,222-226,(2010)

[9] Jiazhi He,Yinghua Xie. Study on K-means algorithm of optimized initial clustering centers based on density. Mic.Com ITS App.34,17-19,(2015)