Improved K-means Algorithm Based on Threshold Value Radius

Jiaqi Song, Fei Li and Ruxiang Li
School of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China
Email: 571714030@qq.com

Abstract. K-means algorithm is a classical clustering algorithm, which needs to specify K value artificially and chooses the initial clustering center randomly. However, it is easy to fall into local optimal solution. In order to overcome the shortcomings of K-means algorithm, an improved algorithm based on threshold value radius is proposed in this paper. The initial clustering center and threshold radius are determined automatically by the average value of Euclidean distance between data. The experimental results show that the method proposed can effectively and accurately overcome the shortcomings of the traditional K-means algorithm.

1. Introduction
With the arrival of the era of big data and the rapid increase in the amount of information, how to extract valuable information from huge data has become a research hotspot, and data mining technology has emerged [1-2]. The main data mining techniques include cluster analysis, estimation, prediction, etc. Among them, cluster analysis is the most widely used technology. It divides data into multiple clusters or classes according to the similarity between different data sources, so that the similarity of data between the same cluster and class is high, while the similarity between different clusters or classes is low [3-5].

The K-means algorithm is a typical partition-based clustering algorithm, which is widely used because of its reliable theory, simple algorithm, fast convergence and efficient processing of large data sets [6-7]. However, the k-means algorithm itself still has some defects. Firstly, the effect of the k-means algorithm is greatly affected by the isolated points. Secondly, the initial clustering center number k of the algorithm needs to be manually selected and the k cluster centers are randomly selected, so that the k-means algorithm has unstable clustering results and is easy to fall into local optimal solutions [8-9].

Aiming at the shortcomings of the algorithm, a large number of scholars have proposed improved methods. For example, Weiqing Cheng [10] proposed an adaptive clustering algorithm, which selects the initial clustering center according to the maximum and minimum distance between samples, and the number of clusters is automatically determined according to the change trend of the sum of squared errors (SSE) and the value of k. Since SSE acts on a specific data set, an inconspicuous "elbow point" may occur, resulting in some errors. Gang Chang [11] proposed to evaluate the k value of the clustering result by establishing a k-value validity index. Firstly, clustering each data sample, getting the distance matrix of sample aggregation and constructing the minimum spanning tree by this matrix. Then, k clusters after initial partition was obtained by pruning. This method can improve the clustering effect to some extent, but the time complexity of the algorithm is large. Ruiyu Jia [12] proposed a self-determined cluster number and initial center K-means algorithm to denoise and determine the cluster center by calculating local density and Euclidean distance. This method solves...
the blindness of the initial cluster center selection. The clustering effect on data with sparse
distribution is not very well. Bo Gao [13] et al proposed to replace the number of clustering k with the
threshold radius R, and divided the relationship between the distance from the real-time data to the
clustering center and R. Although this method overcomes the local optimal solution and improves the
clustering efficiency, the value of R needs to be set artificially just like the value of k and it does not
solve the problem of k value selection in essence.

In order to solve the problem that the k - means algorithm in k values on the selection and
determination of the initial clustering center of the problem need to be manually specified We
improve the defect of K-means algorithm by introducing the concept of threshold radius. The first
initial clustering center is determined by means of the distance between sample data, and the cluster
number is automatically determined by threshold radius.

2. K-means Algorithm

2.1. Introduction of K-means Algorithm
K-means algorithm is a typical partition-based clustering algorithm [14], which was proposed by
James Macqueen in 1967. The algorithm randomly selects k initial cluster centers. In the process of
each iteration, the distance between the calculated data and the clustering center was divided into the
nearest cluster, and the clustering center was recalculated within the cluster. The algorithm was
determined to end by judging whether the value of the criterion function reached the threshold.

2.2. Main Steps of K-means Algorithm
Input: data set \( D = \{x_1, x_2, x_3, ..., x_n\} \), cluster number k.

Output: set of k clusters meeting convergence conditions: \( C = \{C_1, C_2, ..., C_k\} \)

The main steps of K-means algorithm are as follows:
Step1. K data objects are randomly selected from the data set D as the initial cluster center.
Step2. Calculate the distance from each data object in the dataset to the k cluster centers
\[
dist(x_i, c_j) = \left\| x_i - c_j \right\|^2, i = 1, 2, ..., n, j = 1, 2, ..., k
\]
(1)
Step3. According to the \( dist(x_i, c_j) \), each data object is divided into the nearest cluster, that is, satisfying:
\[
dist(x_i, c_j) = \min\{dist(x_i, c_j), i = 1, 2, ..., n, j = 1, 2, ..., k\}
\]
(2)
Step4. Updating cluster centers:
\[
c_j = \frac{1}{n} \sum_{x_i \in C_j} x_i, j = 1, 2, ..., k
\]
(3)
Step5. Computational criterion function:
\[
L = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i - c_j \right\|^2
\]
(4)
Step6. if \( L \leq \xi \), the algorithm ends, otherwise return step2 to continue the calculation.

3. Improved K-means Algorithm Based on Threshold Radius

3.1. Determination of Initial Cluster Center
In the traditional K-means algorithm, the method of randomly determining the initial clustering center
makes the clustering result uncertain, so the selection of the initial clustering center is very important.
In this paper, the average value of all samples in the data set is calculated as the initial clustering center. The main purpose is to find the points that can reflect the average level of samples and avoid creating local optimal solutions.

3.2. Determination of Threshold Radius

In order to solve the problem that K value needs to be specified artificially in traditional K-means algorithm, threshold radius (denoted as R) is introduced to replace K value in this paper. After selecting the first initial clustering center, the distance (denoted as D) from the sample in the data set to the initial clustering center is calculated. If D is not less than or equal to R, the sample data is divided into the current cluster. If D is larger than R, the current object is regarded as a new cluster center and a new cluster is established. The maximum clustering constraint (Cmax) and the minimum clustering constraint (Cmin) are set simultaneously. If the number of cluster samples has reached Cmax, no new samples will be added into the cluster. If the number of samples in the cluster is less than Cmin, the cluster will be deleted, and cluster samples will be divided into the nearest cluster, so as to achieve the effect of no need to determine the number of clusters k artificially.

The samples in the data set all have multiple attributes, so a sample data will be mapped in an n-dimensional space. The Euclidean metric can be used to calculate the real distance between two points in the n-dimensional space, and the calculation formula is as follows:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

(5)

In order to select the appropriate distance as the threshold radius, the mean value of Euclidean distance from all samples in the data set to the initial clustering center was calculated as the threshold radius. The calculation formula is as follows:

$$R = \text{avg}\left[\sum_{i=1}^{n} d(x_i, C)\right]$$

(6)

3.3. The Main Steps of Improve The Algorithm

Input: data set $D = \{x_1, x_2, x_3, \ldots, x_n\}$, maximum clustering constraint $C_{\text{max}}$, minimum clustering constraint $C_{\text{min}}$.

Output: The Set of Clusters Satisfying Convergence Conditions: $C = \{C_1, C_2, \ldots, C_n\}$

The main steps of the improved algorithm are as follows:

Step1. Standardize the data set and calculate the initial clustering center.

Step2. Calculate the distance between the data and each cluster center, record the nearest cluster and distance value. If the distance is greater than the threshold radius, create a new cluster centered on the current data. Otherwise, divide the data into the current cluster and update the cluster center. When dividing data, if the number of samples in the current cluster is greater than or equal to $C_{\text{max}}$, create a new cluster with the current data. Update the cluster center formula as follows:

$$c_j = \frac{1}{n} \sum_{x_i \in C_j} x_i, \ j = 1, 2, \ldots, k$$

(7)

Step3. After all sample partitions are completed, if the number of samples in the cluster is less than $C_{\text{min}}$, the cluster is deleted and the samples are divided into the nearest cluster. And the number of samples in all clusters does not exceed $C_{\text{max}}$.

Step4. Computational criterion function:

$$L = \sum_{i=1}^{n} \sum_{j=1}^{k} \|x_i - c_j\|^2$$

(8)

Step5. If $L \leq \xi$, the algorithm ends, otherwise return step2 to continue the calculation.
4. Experimental Process and Results Analysis

4.1. Experimental Description
In order to verify the effectiveness of the improved algorithm, Iris, Wine and Immunotherapy data sets in UCI data set specially used for testing the algorithm are selected as the test data. UCI data sets have clear classification, and the quality of clustering can be directly obtained [15-16]. The test data information is shown in Table 1.

| Data set       | Sample size | Number of attributes | Number of categories | Category distribution |
|----------------|-------------|----------------------|----------------------|-----------------------|
| Iris           | 150         | 4                    | 3                    | 50,50,50              |
| Wine           | 178         | 13                   | 3                    | 59,71,48              |
| Immunotherapy  | 90          | 7                    | 2                    | 19,71                 |

The Iris data set has 150 samples, each sample contains 4 attributes, which are divided into 3 categories, each category has 50 samples; the Wine data set has 178 samples, each sample contains 13 attributes, which are divided into 3 categories, corresponding to 59, 71, 48 samples respectively; the Immunotherapy data set has 90 samples, each sample contains 7 attributes, which are divided into 2 categories, corresponding to 19, 71 samples. Three data sets were tested separately using the traditional K-means algorithm and the improved K-means algorithm, and the test results were compared to verify the effectiveness of the algorithm.

4.2. Experimental Environmental
The experimental hardware environment is: a personal desktop computer with 16G RAM memory; Inter (R) Core (TM) i5-8500 CPU @ 3.00GHz processor; 256G solid-state hard disk and 1T hard disk; NVIDIA GeForce GTX 1060 graphics card; software environment is: 64 bits of Windows 10 professional edition; Pycharm, pandas, anaconda and other third-party Python libraries. All the algorithms in this paper are implemented by python.

4.3. Experimental Results and Analysis
By comparing the clustering effect of the improved algorithm and the original algorithm on three data sets, the experiment results are counted by choosing different iteration times. Since the selection of initial clustering center by the traditional k-means algorithm is random, the iterative accuracy of the traditional k-means algorithm is the average value of the results obtained after multiple executions, thus reducing the randomness of the obtained results. In order to better display the experimental results, some clustering effect diagrams are drawn, as shown in Figure 1, Figure 2, Figure 3, and Figure 4. The experimental statistics are shown in Table 2.

| Iteration number | Iris K-means accuracy | Improver algorithm accuracy | K-means accuracy | Improve algorithm accuracy | K-means accuracy | Improve algorithm accuracy |
|------------------|-----------------------|-----------------------------|------------------|---------------------------|------------------|---------------------------|
| 1                | 52.67%                | 90.67%                      | 54.35%           | 67.42%                    | 41.11%           | 66.67%                    |
| 2                | 66.67%                | 92%                         | 58.71%           | 71.35%                    | 43.33%           | 72.22%                    |
| 3                | 67.33%                | 93.33%                      | 64.17%           | 74.16%                    | 45.56%           | 76.67%                    |
| 4                | 73.57%                | 94.67%                      | 66.33%           | 76.40%                    | 51.11%           | 78.89%                    |
It can be seen from Table 2 that the improved k-means algorithm has significantly improved the accuracy of the original k-means algorithm, and the cluster tends to be stable after several iterations.
The traditional k-means algorithm Because different initial cluster centers and iterations make the clustering results have great volatility, it takes a lot of time to iteratively calculate to get a better clustering result. It can be seen that the improved k-means. The algorithm is practically feasible and improves the efficiency and accuracy of the algorithm.

5. Conclusion
The k-means algorithm is simple and widely used, but the existing k values need to be artificially selected. The shortcomings of the initial clustering center have a serious impact on the clustering results. In this paper, by introducing the threshold radius instead of the k value, the threshold radius is calculated to perform cluster analysis. Experiments show that the method can effectively improve the accuracy of clustering and solve the problem of k-value artificial selection and initial cluster center randomization. But the improved algorithm still has shortcomings and further research is needed: how to improve the clustering effect of the algorithm on non-spherical clusters and whether there is a more suitable method to determine the threshold radius.

6. Acknowledgments
This work is supported in part by the Natural Science Foundation of Sichuan Province of China under Grant 2019YFG0201 and 2018GZDZX0011.

7. References
[1] Chen Jie. Improvement of data mining classification algorithm. Nanjing University of Posts and Telecommunications, 2018.
[2] Li Fei, Zhang H, Wang J, et al 2018 Toward Intelligent Vehicle Intrusion Detection Using the Neural Knowledge DNA Cybernetics & Systems pp 412-419
[3] Sun Jigui, Liu Jie, Zhao, Lianyu, Clustering algorithms research. Journal of Software, 2008, 9(1): pp48-61.
[4] Tao LJ, Hong LY, YanH. The improvement and application of a K-means clustering algorithm //2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE, 2016.
[5] Jain, Anil K. Data Clustering: 50 Years beyond K-means// European Conference on Machine Learning and Knowledge Discovery in Databases. Springer-Verlag, 2008, pp 651-666.
[6] Khan S S, Ahmad A. Cluster center initialization algorithm for K-means clustering[J].Pattern Recognition Letters, 2014, 40(18): pp 7444-7456.
[7] Du Hongbo, Bai Azhen, Zhu Lijun. K-means Algorithm Based on Improved Density Peak Algorithm [J].Statistics & Decision Making, 2018, 34(18): pp 20-24.
[8] Tanir, Deniz & Nuriyeva, Fidan. (2017). On selecting the Initial Cluster Centers in the K-means Algorithm. 1-5. 10. 1109/ICAICT. 2017. 8687081.
[9] Tang Dongkai, Wang Hongmei, Hu Ming, Liu Gang. Improved K-means algorithm for optimizing initial clustering center. Mini-micro Systems, 2018, 39(08): pp 1819-1823.
[10] Cheng Weiqing, Lu Yanhong. An Adaptive Clustering Algorithm Based on Maximum and Minimum Distance and SSE [J].Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2015, 35(02): pp 102-107.
[11] Chang Gang, Luo Zuomin. Design of Intrusion Detection System Based on Clustering Algorithm [J].Automation & Instrumentation, 2018(09); pp 110-113.
[12] Jia Ruiyu, Li Yugong. The K-means algorithm for the number of clusters and the initial center point self-determination. Computer Engineering and Applications, 2018, 54(07): pp 152-158.
[13] Gao Bo, Huang Min. Research on Intrusion Detection System Based on Improved k-means Algorithm. Computer Knowledge and Technology, 2014, 10(31): pp 7425-7427and7431.
[14] Wang Jianren, Ma Xin, Duan Ganglong. Improved K-means clustering k-value selection algorithm. Computer Engineering and Applications, 2019, 55(08): pp 27-33.
[15] Xing Changzheng, Gu Hao. K-means algorithm for optimizing initial clustering center based on average density. Computer Engineering and Applications, 2014, 50(20):pp 135-138.

[16] Khan M M R, Arif R B, Siddique M A B, et al. Study and Observation of the Variation of Accuracies of KNN, SVM, LMNN, EnN Algorithms on Eleven Different Datasets from UCI Machine Learning Repository. 2018.