K-means Clustering Algorithm and Its Improvement Research

YanPing Zhao and XiaoLai Zhou
Information Center, Beijing JiaoTong University, Beijing 100044, China
Email: ypzh@bjtu.edu.cn

Abstract. Clustering is a typical unsupervised learning method, and it is also very important in natural language processing. K-means is one of the classical algorithms in clustering. In k-means algorithm, the processing mode of abnormal data and the similarity calculation method will affect the clustering division. Aiming at the defect of K-means, this paper proposes a new similarity calculation method, that is, a similarity calculation method based on weighted and Euclidean distance. Experiments show that the new algorithm is superior to k-means algorithm in efficiency, correctness and stability.

1. Introduction
Cluster analysis is a very important research topic in data mining. The goal of cluster analysis is to collect and classify data on a similar basis. Clustering comes from many fields, including mathematics, computer science, statistics, biology, and economics. Many clustering techniques have been developed in different application fields. At present, many clustering algorithms have been proposed, such as K-means [1], DBSCAN [2], CURE [3], FCM [4]. But on the whole, in the face of large-scale, high-dimensional data, have not achieved a better effect. K-means algorithm is a common partitioning based clustering method. It is theoretically reliable, simple, converges quickly, and can effectively deal with small and medium-sized data sets, so it is widely used. However, its results are sensitive to the initial clustering center, and the clustering results of different similarity measurement methods will be different. This paper improves the k-means algorithm.

1.1. Related work
K-means algorithm finds K clusters according to a specific standard (clustering is also known as cluster), and generally adopts the standard of square error. The time complexity of k-means algorithm is $O (NTK)$, where $N$ is the total number of data sets, $K$ is the total number of partitions, and $T$ is the number of iterations in the clustering process. The implementation efficiency of k-means algorithm is relatively high, but clustering should be carried out in the case that $K$ is known, and the initial center point is randomly selected, so that the clustering result has a high fluctuation.

In recent years, many improved methods have been proposed. For example, literature [5] defines the concepts of neighbourhood coupling and separation of objects by using the upper and lower approximation of object neighbourhood in the neighbourhood model, and proposes an initial clustering center selection algorithm. Literature [6] proposed the k-means clustering algorithm based on the optimization of artificial fish swarm, which can obtain the global optimal division.

DBSCAN is a dense-based clustering algorithm. The basic idea is: for each object in the clustering, the number of objects in the neighbourhood of a given radius should not be less than a given number,
and then the objects with the characteristic of density connection are clustered. This algorithm divides the regions with sufficient density into clusters, and finds clusters with arbitrary shapes in the spatial database with "noise". If a spatial index is used, the computational complexity of DBSCAN is O(nlogn), where n is the number of objects in the database. Otherwise, the computational complexity is O (n^2). But the algorithm is sensitive to user-defined parameters.

Maximum and minimum distance algorithm [7] is a tentative algorithm in the field of pattern recognition. Based on Euclidean distance, the idea is to take the object as far away as possible, avoid the situation that the initial center is too dense when k-means algorithm randomly selects, and improve the partition efficiency.

Coefficient of variation [8], also known as "dispersion coefficient", is a normalized measure of the dispersion degree of probability distribution. The formula is the ratio of standard deviation to the mean. Coefficient of variation is an objective weighting method, which can be used to calculate the weight of each attribute in the clustering algorithm. Then, the similarity between the data is calculated according to the weight of each attribute.

1.2. The paper work
This paper presents a clustering algorithm based on maximum and minimum distance and weighted similarity calculation. First, the maximum and minimum distance method is used to search out the optimal initial clustering center. While determining the initial clustering center, isolated points are removed to avoid being affected by noise. Then, K-means algorithm is adopted to carry out clustering. This algorithm uses density method to screen the influence of abnormal data on the algorithm, which is close to DBSCAN algorithm in handling arbitrary shape clustering, and has the same time complexity as K-means algorithm, so the efficiency is high.

This paper first introduces the improved algorithm, then analyses the performance of the improved algorithm from multiple angles, and finally discusses the improved algorithm.

2. Improved algorithm

2.1 The basic definition
Definition 1(density of point): For any point \( X \) distance \( \varepsilon \) in space, the number of points in the region with radius \( \varepsilon \) centering on point \( P \) is called point \( P \). Based on the density of distance \( \varepsilon \), it is denoted as \((X, \varepsilon)\), and \( \varepsilon \) is the density radius.

Definition 2(high-density data): for any point \( X \), distance and threshold \( \text{Minpts} \) in space, if the Density\((X, \varepsilon)\geq \text{Minpts}\) then \( P \) is high-density data; otherwise, it is low-density data. At the same time, we call \( \text{Minpts} \) a density threshold.

Definition 3(weight \( w \) ) : Suppose there are \( t \) attributes in the data set \( D \), and the variation coefficient of each attribute is \( v_1, v_2, \ldots, v_t \), then the weight \( w_i \) of the \( i \) attribute is:

\[
 w_i = \frac{v_i}{\sum_j v_j}  \quad (1)
\]

Definition 4(similarity): Let two \( t \)-dimensional vectors \( x_i = (x_{i1}, x_{i2}, \ldots, x_{it}) \) \( x_j = (x_{j1}, x_{j2}, \ldots, x_{jt}) \), whose similarity is:

\[
 d(x_i, x_j) = \sqrt{d_2(x_{i1} - x_{j1})^2 + \ldots + d_2(x_{it} - x_{jt})^2} \quad (2)
\]

Definition 5 (weighted similarity): The data set \( D \) has \( t \)-dimensional data, and the weight of each dimension is \( w_1, w_2, \ldots, w_t \). Let two \( t \) -dimensional vectors \( x_i = (x_{i1}, x_{i2}, \ldots, x_{it}) \) \( x_j = (x_{j1}, x_{j2}, \ldots, x_{jt}) \), and their weighted similarity is \( d_w(x_i, x_j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + \ldots + w_t(x_{it} - x_{jt})^2} \quad (3)\)

2.2 Improved algorithm
The main idea of the improved K-means algorithm is to first calculate the density of each sample in sample data set \( D \), so as to obtain a sample set \( E \) based on the density standard. Then, on the basis of \( E \),
the initial cluster center is selected according to the idea of the maximum and minimum distance, and then the weighted similarity is used to divide each data into corresponding clusters. Such an algorithm is called MW-K-means. The clustering algorithm for selecting the center point based on the maximum and minimum distance is called M-K-means algorithm, and the clustering algorithm based on the weighted Euclidian-distance is called W-K-means algorithm. Mw-k-means algorithm steps are as follows:

Input: (1) total number of clusters $K$, (2) set $D$ of $N$ data point (3) density radius $\varepsilon$ (4) density threshold $Minpts$.

Output: $K$ clusters and number of iterations

Step1: calculate the distance between each sample and other samples according to definition 4;

Step2: Determine the high-density data point set $E$ according to parameter value $\varepsilon$ and $Minpts$ and definition 1 and 2;

Step3: Take the point with the highest density in $E$ as the first point $C_1$ of the initial data center set $C$;

Step4: Take a high density point farthest away from $C_1$ as the second cluster center $C_2$;

Step5: Calculate the distance $d(X_i, C)$ from data object $X_i$ in $E$ to $C$ and the distance $d(X_i, C_j)$ from $C_j$ according to definition 4; $C_i$ can obtain $K$ initial clustering centers in order to satisfy the following equation: Data object $X$ of $\text{Max} (\text{Min}(d(X, C)), \text{Min}(d(X, C_2)),...,\text{Min}(d(X, C_{i-1})))$;

Step6: Calculate the weight of each attribute of set $E$ according to definition 3;

Step7: Calculate the similarity between each data point and the cluster center according to definition 5, and group it into the cluster with the greatest similarity;

Step8: Calculate the average value of the points of each cluster to get the new cluster center;

Step9: Iterate through Step7 and Step8 until the clustering center is no longer changed.

In this algorithm, the maximum and minimum distance algorithm is adopted to avoid the situation that clustering seeds may be too close to each other in the initial value selection of k-means algorithm. When calculating the weight, the abnormal points are removed first, which reduces the sensitivity of noise and outliers. The weighted similarity calculation shows the importance of each attribute in clustering, which is effective for the correct division of data.

In this method, density parameters $Minpts$ and $\varepsilon$ need to be input, and the setting method we adopted is:

$$MinPts = beta \frac{N}{[X]}$$  \hspace{1cm} (4)

$$\varepsilon = theta \frac{\sum \sum d(X_i, X_j)}{N^2}$$  \hspace{1cm} (5)

Where, $N$ is the number of objects in the data set, $beta$ and $theta$ are input by the user, and are the point density coefficient and the search radius coefficient respectively.

3. Algorithm performance analysis

In order to verify the effectiveness of MW-K-means algorithm, Iris, Wine and Haberman three sets of data in UCI machine learning database (an international common database specially provided for scientific researchers to test classification and clustering algorithm) were selected as experimental data sets. The detailed description of each data set is shown in Table 1.In this paper, the MW-K-means algorithm is compared with the traditional K-means algorithm, M-K-means algorithm and W-K-means algorithm on three data sets respectively. The experimental results are shown in Table 2, and the F-measure values of the four algorithms are shown in Figure 1.

| Data set | Number of objects | Number of attributes | Number of clustering | Number of objects of various |
|----------|-------------------|----------------------|----------------------|----------------------------|
| IRIS     | 150               | 4                    | 3                    | 50,50,50                   |
| Wine     | 178               | 13                   | 3                    | 59,71,48                   |
| Haberman | 306               | 3                    | 2                    | 225,81                     |

Table 1: Description of experimental data set
Table 2. Comparison of clustering results

| algorithm | Iris | Wine | Haberman |
|-----------|------|------|----------|
|           | T    | P    | R        | F        |
| k-means   | 8    | 83   | 83       | 83       |
| M-K-      | 42   | 89   | 89       | 89       |
| W-K-      | 6    | 87   | 87       | 87       |
| MW-K-     | 40   | 91   | 91       | 91       |

Note: T represents the total clustering time (MS); P represents accuracy rate (%); R stands for recall rate (%); F represents F_Measure (%) [10]

Figure 1. F-measure value comparison broken line

Each of the above algorithms is partitioned by creating an initial partition and then iteratively dividing the cluster to which each object belongs. It can be seen from Table 2 and Figure 1 that the traditional K-means algorithm and W-K-means show a faster speed due to the random selection of the initial center, but the quality of clustering is improved because W-K-means considers the contribution of each attribute to partition. On the premise of satisfying high density, m-k-means algorithm tries to sparse initial clustering center as far as possible, reflecting the distribution of data set, so as to avoid the large fluctuation influence brought by the random selection of initial center point by traditional K-means algorithm. Mw-k-means algorithm integrates the advantages of w-K-means and M-K-means algorithm, with faster convergence speed, more accurate clustering results, and higher stability, which can be used for practical clustering.

4. Conclusion

K-means algorithm is widely used, such as customer relationship management, student achievement analysis. Among them, the clustering of students according to their grades is helpful to understand the situation of students of various categories. Through the analysis and research of each class of students, relevant factors affecting grades are found out and different measures are taken to help students get on the normal learning path, so as to comprehensively improve the teaching quality of the school. In order to improve the quality of clustering, many improvements have been made. In this paper, the maximum and minimum distance and coefficient of variation are studied and applied to the improved K-means algorithm. Therefore, a clustering algorithm MW-K-Means based on density and weighted Euclidean distance is proposed. Mw-k-means algorithm maintains the advantage of fast clustering of K-means algorithm, obtains the initial clustering center with good dispersion and representation, avoids the influence of abnormal data on clustering results, and considers the contribution of each attribute to classify clusters when calculating similarity, thus improving the efficiency of classification. The comparison experiment shows that the method is reasonable for the selection of center point and the calculation of similarity, and it can get better clustering effect. Today, data in many areas is high dimensional, complex, dynamic, and large-scale. Therefore, how to use this algorithm to analyse high-dimensional data and massive data needs further research.
References
[1] HAN Jia-wei, MICHELIN K. Data mining concepts and techniques [M]. 2nd ed. Singapore: Elsevier, 2006: 383-419.
[2] Ester M, Kriegel HP, Sander J, et al. A density based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining Portland: AAAI Press, 1996: 226-231.
[3] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases[C]//Proc of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1998: 73-84.
[4] Bezdek J C. Pattern recognition with fuzzy objective function algorithms [M]. New York: Plenum Press, 1981
[5] Johannes Schneider, Michail Vlachos: Fast Parameterless Density-Based Clustering via Random Projections [C]//Association for Computing Machinery, New York: ACM Press, 2013: 861-866.
[6] Das S, Abraham A, Konar A. Automatic clustering using an improved differential evolution algorithm[J]. IEEE Trans. on System, Man cybernetics-part: System. Humans, 2008, 38(1): 218—237
[7] Waters B. Ciphertext-Policy Attribute-Based Encryption: An Expressive, Efficient, and Provably Secure Realization[C]// International Workshop on Public Key Cryptography. Springer, Berlin, Heidelberg, 2011: 53-70
[8] Han J, Susilo W, Mu Y, et al. Improving Privacy and Security in Decentralized Ciphertext-Policy Attribute-Based Encryption[J]. IEEE Transactions on Information Forensics & Security, 2015, 10(3): 665-678
[9] HAN Jia-wei, MICHELIN K. Data mining concepts and techniques [M]. 2nd ed. Singapore: Elsevier, 2006: 383-419.
[10] Feijiang Li, Yuhua Qian, et al, ACM Transaction on Knowledge Discovery from Data, 2018: 1-27.