Research on Density-Based K-means Clustering Algorithm

Shuxin Liu*, Xiangdong Liu

Institute 706, Second Academy of China Aerospace Science and Industry Corporation, Beijing China

*Corresponding author’s e-mail: 501660420@qq.com

Abstract: Cluster analysis is an unsupervised learning process, and its most classic algorithm K-means has the advantages of simple principle and easy implementation. In view of the K-means algorithm's shortcoming, where is arbitrary processing of clusters k value, initial cluster center and outlier points. This paper discusses the improvement of traditional K-means algorithm and puts forward an improved algorithm with density clustering algorithm. First, it describes the basic principles and process of the K-means algorithm and the DBSCAN algorithm. Then summarizes improvement methods with the three aspects and their advantages and disadvantages, at the same time proposes a new density-based K-means improved algorithm. Finally, it prospects the development direction and trend of the density-based K-means clustering algorithm.

1. Introduction

Clustering analysis is an unsupervised learning process[1], and is one of the main research directions of data mining. It explores and mines the differences and connections between data objects under the premise of unknown data relationships. Nowadays, cluster analysis is widely used in data analysis[2,3], image processing[4,5], pattern intelligent recognition[6], market research and other fields. Research on cluster analysis has been very popular in recent years, and there are many methods of cluster analysis (see Table 1), such as K-means, DBSCAN, BIRCH, CURE, CLIQUE, EM, etc.[7].

Table 1 Comparison of classical clustering algorithms

| Algorithm | Clustering Type   | Time Complexity |
|-----------|-------------------|-----------------|
| K-means   | Partition clustering | $O(nkd)$        |
| DBSCAN    | Density clustering | $O(n \log n)$   |
| BIRCH     | Hierarchical clustering | $O(n^2 \log n)$ |
| CURE      | Hierarchical clustering | $O(n^2 \log n)$ |
| CLIQUE    | Grid clustering    | $O(c^d + nd)$   |

Note: d is the dimension of the data sample, n is the total number of sample spaces, and k is the number of clustering divisions.

This paper mainly discusses the improvement of traditional K-means algorithm from the perspective of combining with density-based clustering algorithm. Firstly, there are described the basic idea and algorithm flow of two classical algorithms--K-means and DBSCAN. Then it summarizes the improved K-means algorithm from the perspective of density clustering aiming at the defects of K-means algorithm, and proposes a new improved K-means algorithm based on density clustering. Finally, it looks forward to the future research direction of the algorithm.
2. Classical algorithm

2.1 Traditional K-means algorithm
The traditional K-means algorithm is a classical algorithm based on partition and is also the most widely used algorithm in cluster analysis. It was first proposed by MacQueen and describes a process of dividing d-dimensional sample population into k sets based on samples [8].

2.1.1 Basic ideas
K-means algorithm is a classification clustering algorithm based on only one input. Similar data will be in the same cluster, and dissimilar data will be in different clusters. The basic idea is to select k data samples in the d-dimension dataset $X = \{x_i|i=1,2,...,n, x_i \in \mathbb{R}^d\}$ containing n data samples as the initial cluster center. Then divide k clusters according to the similarity of data samples to the initial cluster center. By iterating and updating the cluster center set until the cluster center set no longer changes, finally obtain the clustering result.

Definition 1 Distance: Euclidean distance is used to measure the similarity between two data samples. Euclidean distance formula is:

$$Dist(x_i, x_j) = \sqrt{\sum_{i=1}^{d} (x_{i} - x_{j})^2}$$  \hspace{1cm} (1)

Where $x_i$, $x_j$ are two different data samples.

Definition 2 Mean Value of Data Samples in Cluster: It is the mean distance of data samples in the same cluster. The formula is denoted as:

$$MeanCen_j = \frac{1}{|C_j|} \sum_{x_i, x_p \in C_j} Dist(x_i, x_p)$$  \hspace{1cm} (2)

$C_j$ is the j-th clustering, $|C_j|$ is the number of data samples in the j-th cluster, $\sum_{x_i, x_p \in C_j} Dist(x_i, x_p)$ is the sum of the distances of two data samples in the same cluster.

Definition 3 Error Sum of Squares Criterion (SSE): It is a standard to judge the quality of clustering results. The formula is denoted as:

$$SSE = \sum_{j=1}^{k} \left| x - c_j \right|^2$$  \hspace{1cm} (3)

2.1.2 Algorithm flow
Input: A dataset $X = \{x_i|i=1,2,...,n, x_i \in \mathbb{R}^d\}$ containing n data samples with d-dimension, Number of clusters k;

Output: k stable clusters.

(1) Randomly select k data samples as the initial center of clustering to obtain the initial center set $c = \{c_1, c_2, ..., c_k\}$;

(2) Calculate the distance between the remaining data samples and each cluster center according to Formula (1);

(3) Find the nearest clustering center $c_j$ according to the distance, and put the data samples into the clustering $C_j$;

(4) Get k clusters $C = \{C_1, C_2, ..., C_k\}$, then calculate the mean values of data samples within cluster according to Formula (2), and update the cluster centers $c' = \{c'_1, c'_2, ..., c'_k\}$;

(5) Calculate the sum of squares of total errors according to Formula (3);

(6) Repeat (2)-(5) until SSE converges, then the clustering ends.
2.1.3 Advantages and disadvantages of K-means algorithm
The principle of K-means algorithm is simple and easy to implement. It can process large data sets quickly and the algorithm is very efficient. It can be seen from table 1 that its time complexity is relatively low. However, the K-means algorithm also has certain limitations: (1) k value is artificially determined; (2) there is no basis for the selection of the initial cluster center; (3) very sensitive to outliers.

2.2 Density-based clustering
The essence of density-based clustering is to find high-density areas in a set of data sets mixed with low-density areas, among the data samples located in the low-density areas are usually treated as outliers[9]. Its biggest bright spot is that it can identify and analyze clusters of any shape by judging the density value of the data sample. According to different definitions of density, density-based clustering algorithms include DBSCAN, OPTICS, CLIQUE, etc.[10].

2.2.1 Basic ideas
The DBSCAN algorithm is a classic algorithm based on density clustering, which was proposed by Martin Ester et al. in 1996[1]. The DBSCAN algorithm uses two input parameters of neighborhood radius and boundary threshold to determine the density value of each data sample. It can effectively cluster and identify outliers.

For a dataset $X=\{x_i|i=1,2,...,n, x_i \in \mathbb{R}^d\}$ containing n data samples with d-dimension, the DBSCAN algorithm mainly has the following definitions:

Definition 4 Neighborhood Radius(Eps):There is an object $p$ in dataset $X$, then its neighborhood $Eps(p)$ is a set with object $p$ as the center and Eps as the radius, namely:

$$Eps(p) = \{q \in X | Dist(p,q) \leq Eps\}$$  (4)

Definition 5 Point Density: Refers to the number of objects contained in neighborhood $Eps(p)$ of the set with the object $p$ as the center and Eps as the radius, namely:

$$density(p) = |Eps(p)|$$  (5)

Definition 6 Core Points and Boundary Points: Given a boundary threshold $minPts$, density($p$) is compared with the $minPts$, and the core points and boundary points should satisfy the following formula:

$$\begin{cases} 
density(p) \geq minPts, \text{ Core Points} \\
\density(p) < minPts, \text{ Boundary Points} 
\end{cases}$$  (6)

The core idea of DBSCAN algorithm is to calculate point density through Eps and $minPts$, mark core points and boundary points, then combine the associated points into a cluster, realize effective clustering. Thus, the input parameters Eps and $minPts$ determine the result of DBSCAN algorithm.

2.2.2 Algorithm flow
Input: A dataset $X=\{x_i|i=1,2,...,n, x_i \in \mathbb{R}^d\}$ containing n data samples with d-dimension, two parameter Eps and $minPts$;
Output: k stable clusters.
(1)Initialize the core point set E, boundary point set B, and the number of cluster divisions k;
(2)Calculate the euclidean distance between two data samples, and calculate the Eps($p$) and density($p$) of each data sample according to the Eps;
(3)If the point density of the data sample is greater than or equal to the $minPts$, put the data sample into the set E, otherwise put it into the set B, and delete the data sample from the set X;
(4)In the set E, the data samples that meet the density can be divided into a cluster ;
(5)If the set E is empty, use the distance formula to cluster the data samples in the set B according to the closest principle;
(6)Scan dataset X again until the boundary point set is empty

3
If the dataset X is not empty, put the remaining data samples into the nearest cluster according to the minimum distance; on the contrary, the clustering ends

3. Density-based k-means clustering algorithm improvement

Among the advantages and disadvantages of K-means algorithm described above, it is very sensitive to the processing of values, clustering centers and outliers. Density-based clustering does not need to determine the number of clusters before the algorithm starts and is insensitive to data sets containing outliers. At present, many researchers put forward corresponding improved algorithms based on the idea of density clustering to solve the limitations of K-means algorithm. The improved algorithm is discussed from three different directions as follows.

3.1 Optimization of k value

In the classical algorithm K-means, k value which is determined in advance as an input item before the algorithm starts. On the premise of no prior knowledge, different choices of k value will produce different results. When the selected k value is less than the real value, the data samples in the cluster will differ greatly. On the contrary, the clustering results are meaningless. The upper limit of the value of k in the K-means algorithm is \( n \leq \sqrt{n} \) \(^{(11)}\), and Zhang et al.\(^{(12)}\) uses the default initial k of clustering as the upper limit. They determined the optimal k value through the distance formula ratio within and between clusters.

3.2 Optimization of the initial clustering center

In the K-means algorithm, the initial cluster centers are randomly determined so that if an outlier is selected, the clustering results may fall into local optimization. Therefore, it is very important to choose a suitable cluster center. When selecting the initial cluster center, if only the density of the data sample is considered, it is easy to lead to local optimal results; if only the maximum distance of the data sample is considered, it is easy to choose the outlier as the initial cluster center. For optimize the initial clustering center, Xie et al.\(^{(13)}\) calculated the average density of each data samples and put higher than minPts into E. The first data sample in E was used as the initial clustering center, and then dissimilar (k-1) data samples were selected from E as the initial clustering center. And He et al.\(^{(14)}\) selected the initial cluster centers based on the principles of maximum density and longest distance, and also considered the case where the density is not unique.

3.3 Processing of outliers

Outliers are relatively far from the cluster center and do not gather. If outliers are not properly processed in the clustering process, they will have a great impact on the clustering results. Luo et al.\(^{(15)}\) calculated the density of the data sample object by setting a specific weight in advance and eliminated the data sample object larger than the specified threshold. Xing et al.\(^{(16)}\) found outliers and put them in an independent empty set at the beginning of the algorithm, and then clustered the remaining data sample according to the algorithm rules. At this time, outliers would not affect the selection of clustering center value and clustering results.

3.4 Comparison of advantages and disadvantages of optimization algorithms

From the perspective of density clustering, and aiming at the limitations of K-means algorithm, improved algorithms have been proposed from the above three aspects. The advantages and disadvantages of each improved method are compared in Table 2.
Table 2 Comparison of advantages and disadvantages of the improved algorithm

| Improved algorithm | To improve the direction | advantages | disadvantages                  |
|--------------------|--------------------------|------------|---------------------------------|
| Literature on 12   | k value                  | Through k--, it can determine the best k value and improve the clustering accuracy. | Clustering centers and outliers will still affect the clustering results, and the algorithm will be slower. |
| Literature on 13/14| initial clustering center | Determine the cluster center in the high-density area to avoid the local optimization problem of clustering. | The algorithm does not consider the value of k, and relies on specific parameter values, which affects the clustering results. |
| Literature on 15/16| outliers                 | Reduce the influence of outliers on the algorithm by deleting or separating outliers. | It does not consider the value of k, and other conditions are derived from experience. |

4. A new density-based k-means improved algorithm

4.1 Improved algorithm idea

According to the principle of clustering algorithm, the similarity between clusters should be different as far as possible, and the similarity of each data sample in the cluster is similar. Therefore, a new density-based K-means improved algorithm is proposed which is superior to the above algorithm.

The improved algorithm is based on density clustering, and its central idea is to select the two most distant points in the high-density area as the initial cluster centers, and then evaluate whether it is the best k value by calculating the SSE. If not, split the existing largest cluster and re-cluster. Finally, return to the K-means algorithm to get the result.

4.2 Basic process of improving the algorithm

For a dataset $X=\{x_i|i=1,2,\ldots,n, x_i \in \mathbb{R}^d\}$ containing n data samples with d-dimension, define the $Eps=\max\{Dist(x_i,x_j)\}$, $\min Pts = \frac{1}{n} \sum_{i=1}^{n} \text{density}(x_i)$ and $\text{density}(x_i)<\alpha \times \min Pts$, the steps of dividing k clusters from n data sample are as follows:

1) Calculate euclidean distances, determine $Eps(p)$, $\text{density}(p)$ and $\min Pts$, initialize set $c$, set $E$ and set $N$, and $k=2$;
2) Put the outliers into set $N$, and delete the outliers from set $X$ to obtain $X'=X-N$, then compare the $\text{density}(p)$ in $X'$ with $\min Pts$. If $\text{density}(p)>\min Pts$, put the $p$ into set $E$;
3) Select the two data samples with the largest distance from set $E$ as the center $\{c_1,c_2\}$ and put them into set $c^{(1)}$, compare the distance between all objects in dataset $X'$ and the two cluster centers and divide them into two clusters $\{C_1,C_2\}$ according to the principle of nearest, meanwhile, calculate $\text{SSE}(1)$ when $k=2$;
4) Compare the size of the two clusters, select the larger one and then choose two new clusters $\{c_2',c_3\}$ based on the maximum distance condition, and use $\{c_1,c_2',c_3\}$ as the cluster center to cluster all objects in dataset $X'$ to obtain three clusters $\{C_1',C_2',C_3\}$, and calculate $\text{SSE}(2)$ when $k=3$;
5) Repeat the process of (4) continuously to obtain the relationship between the k value and the criterion of SSE;
6) When $\text{SSE}(k-1)$ does not change significantly, determine $k$ value and corresponding center set $c^{(k-1)}$, and then continue to clustering according to k-means algorithm;
7) Finally, divide the outliers into the nearest cluster, and the clustering ends.

The improved algorithm flow is shown in Figure 1.
Enter and initialize dataset X
Calculate the distance, density and average distance of each data sample, divide outlier points, and update outlier set N and dataset X'
Put the data with greater than the average density into the set E, select the two points with the largest distance from E as the cluster center, cluster the data in X', update the cluster center, and calculate SSE (1), at which time k=2.
Compare the size of k clusters, divide the larger cluster into two clusters according to the largest cluster and k++, calculate the new cluster center and SSE (2)

\[
\text{\{SSE(k-1)\} drop sharply?}
\]

Yes
Output the optimal k value, clustering center and optimal clustering

No
By clustering outliers in set N according to the minimum distance, k clusters are obtained
End

FIG 1. Flow chart of improved algorithm

4.3 Effect analysis of the improved algorithm
The improved algorithm initially defined k=2, and the optimal k value was determined by combining SSE in the process of constantly updating k value. In the clustering algorithm, k value is not particularly large, so if it is decreased, the process will be longer. Increasing traversal from the lower limit of k will reduce the number of traversals. With the increase of the cluster number k, the aggregation degree of each cluster will gradually improve, while the SSE will gradually decrease. When k is less than the optimal, SSE decreases greatly, while when k reaches the optimal, SSE decreases sharply, and then flattens with the continuous increase of k. Therefore, SSE can be used to determine the optimal k value and achieve the optimal clustering effect.

In this improved algorithm, the maximum distance value is used as Eps to calculate the density of each sample, and the average density is used to make minPts to determine the high density area E, which solves the randomness caused by the parameters introduced by the density idea and also reduces the range of selecting clustering centers. The algorithm also introduces a judgment parameter α to define an outlier, and deletes it at the same time to get a new set X'. This process shields outliers and avoids the influence of outliers on clustering results. In set E, if the principle of maximum density is used to determine the clustering center again, the clustering result will be locally optimized, so the clustering center is determined by the maximum distance. The combination of distance and density avoids local optimal clustering results or large deviation of cluster center selection caused by single use.

In the improved algorithm, each time k value is determined, the corresponding center set and cluster need to be calculated and stored, so the spatial complexity of the algorithm is high. The value of parameter α is (0,1), which is an important factor in determining outliers. When the parameter α is
small, the unscreened outliers will affect the updating of the cluster center, on the contrary, data objects with higher density will be excluded, which will also affect the center selection. Therefore, the value of \( \alpha \) needs to be considered.

5. Conclusion

From the perspective of density clustering algorithm, the idea of improving the traditional K-means algorithm mainly comes from the limitations of K-means algorithm. In the improved algorithm, the influence of uncertain factors and outliers on the algorithm result is avoided effectively. However, in the clustering algorithm, different measurement formulas will have different impacts on the clustering. To sum up the experience, it is necessary to optimize \( k \) value and initial clustering center in the future without increasing the complexity of the algorithm, or choose a more appropriate and suitable clustering measurement formula from the perspective of similarity measurement.

References

[1] Gan G, Ma C, Wu J. Data Clustering: Theory, Algorithms, and Applications[M]. 2nd. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2020.
[2] Lia Hananto April et al. Analysis Of Drug Data Mining With Clustering Technique Using K-means Algorithm[J]. Journal of Physics: Conference Series, 2021, 1908(1): 012024
[3] Liang Y, Wang H, Chen Z, Zhang Z, Wang Q, She G, Xie W,Liu Z, Sun Y. Power Consumption Data Recognition Based on Clustering Analysis[J]. Journal of Wuhan University (Engineering Edition), 2017,50(04):542-547+555.
[4] Xu G, Zhao T, Jiang S,Gao D. Finite Element Method for Surface Cracks [J]. Journal of Huazhong University of Science and Technology (Natural Science),2019,47(10):52-55+68.
[5] Wang X, Liu X,An Improved Clustering Algorithm Based on P Systems and its Application in Face Image Processing[J]. Laser Journal, 2019,40(03):74-78.
[6] Kahn A, Gonski J, Ochoa I, et al. Anomalous Jet Identification via Sequence Modeling[J]. ArXiv Preprint arXiv:2105.09274, 2021.
[7] Kurasova O, Marcinkevicius V, Medvedev V, et al. Strategies for Big Data Clustering[C]//2014 IEEE 26th international conference on tools with artificial intelligence. IEEE, 2014: 740-747.
[8] MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967, 1(14): 281-297.
[9] Campello R J G B, Kröger P, Sander J, et al. Density - Based Clustering[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, 10(2): e1343.
[10]Wang J, Xia L, Jing J. Maximum Density Clustering Algorithm[J]. Journal of University of Chinese Academy of Sciences, 2009, 26(4): 539-548
[11]Pakhira M K, Bandyopadhyay S, Maulik U. Validity Index for Crisp and Fuzzy Clusters[J]. Pattern recognition, 2004, 37(3): 487-501
[12]Zhang L, Chen Y, Ji Y, Zhang J. Research on K-means Algorithm Based on Density[J]. Application Research of Computers,2011,28(11):4071-4073+4085.
[13]Xie X, Li X, Mo L. Microblogging Opinion Analysis Based on An Improved K-means Algorithm[J].Computer Engineering and Science,2018,40(01):155-158.
[14]He Y B, Liu X J, Wang Z Q. Improved K-means Algorithm Based on Global Center and Nonuniqueness High-Density Points[J]. J. Comput. Eng. Appl, 2016, 52(1): 48-54.
[15]Luo J, Hong D. A K-means Clustering Algorithm Based on Density and Distance[J].Software Engineering,2020,23(10):23-25+4.
[16]Xing C, Gu H. K-means Algorithm Based on Average Density Optimizing Initial Cluster Centre[J]. Computer Engineering and Applications, 2014, 50(20): 135-138.