The Journal of Engineering

7th International Symposium on Test Automation and Instrumentation (ISTAI 2018)

IET Journals
The Institution of Engineering and Technology

Parameter selection algorithm of DBSCAN based on K-means two classification algorithm

eISSN 2051-3305
Received on 12th October 2018
Accepted on 8th January 2019
E-First on 24th October 2019
doi: 10.1049/joe.2018.9082
www.ietdl.org

Shouhong Chen[1,2], Xinyu Liu[1], Jun Ma[1] ✉, Shuang Zhao[1], Xingna Hou[1]

[1]*Guilin University of Electronic Technology, Guilin 541000, People's Republic of China*
[2]*Jiangsu University, Zhenjiang, 212013, China, People's Republic of China*
✉ *E-mail: majun@guet.edu.cn*

Abstract: Clustering algorithm is one of the most important algorithms in unsupervised learning. For density-based spatial clustering of applications with noise (DBSCAN) density clustering algorithm, the selection of neighborhood radius and minimum number is the key to get the best clustering results. Aiming at the problems of traditional DBSCAN algorithm, such as the neighborhood radius and the minimum number of points, this article puts forward two classifications based on K-means algorithm, and gets two clustering centers. Where calculated between two data points and the cluster center-to -center distance, clustering, distance, statistics in a distance of data points within the scope of the search, the number of data points corresponding to the maximum distance value, and thus the parameters for the DBSCAN algorithm to estimate and selection of initial radius of neighborhood with the minimum number of clustering start critical value. When the parameters are iterated and optimized continuously, the data are divided into clusters, and the most suitable neighborhood radius and the minimum point number are obtained. The experimental data analysis show that the improved algorithm reduces the human factors in the traditional algorithm and improves the efficiency, so as to get the accurate clustering results.

1 Introduction

Machine learning is mainly divided into supervised algorithm and unsupervised learning. Compared with supervised learning, unsupervised learning does not need the labels of training samples. Without prior knowledge, the unlabelled samples are trained to learn the rules of the data, and the similar samples are classified as one class, and the dissimilar ones are classified as other classes. In unsupervised learning, clustering is the most widely applied method.

Clustering is a process of classifying data into different classes, so the objects in the same cluster have a great similarity, and the objects between different classes have a great difference. At present, clustering analysis is the pre-processing step of other algorithms, such as classification and qualitative induction algorithm. The goal of cluster analysis is to collect data on similar basis for classification.

Clustering analysis is an important research area in data mining. Aiming at clustering analysis, several methods are developed. It includes dynamic clustering, hierarchical clustering algorithm, density-based clustering algorithm and grid-based clustering algorithm [1]. In data classification, we will use an improved density clustering algorithm to classify data.

On the one hand, the traditional K-means algorithm [2] is the most classical algorithm in the dynamic clustering algorithm, but there are some problems in the traditional K-means algorithm: the algorithm random setting initialises the cluster centre, which makes the results of the clustering are not exactly the same; the algorithm usually ends with the local optimal, and the global optimal is difficult to obtain. When the data are too large, the computation efficiency is reduced, and it is difficult to get the clustering results quickly. Document [3] interconnects and merges sub-clusters generated by multiple sampling of data sets, so as to improve clustering results. In the literature [4], we find the best number of clusters by stratifying data and finding the similarity between classes based on hierarchical data.

On the other hand, DBSCAN algorithm can divide high density and connectivity data into clusters of arbitrary shape. Compared with K-means algorithm, it is easier to get global optimal. The same results will be generated by multiple operation. DBSCAN algorithm does not need to manually determine the number of classifications, but the neighbourhood radius and the minimum number of points need to be specified, but the selection of parameters is relatively difficult. When the amount of data is large, the memory consumption of DBSCAN algorithm to CPU is very large, resulting in low utilisation rate. Document [5] first uses K-means clustering algorithm to cluster the data, calculates the distance between samples after clustering, and selects the maximum distance value as the neighbourhood radius value of the corresponding category, and then calculates the minimum point number by the neighbourhood radius. The algorithm achieves desirable clustering results and improves accuracy, but it is difficult to select initial values.

Based on the traditional K-means algorithm, we will classify the data into two categories, and update two clustering centres until the end of iteration. The distance between the two cluster centres and the data points is calculated by the two clustering centres obtained, and the number of data points in a certain distance is counted, and the distance values corresponding to the number of data points are searched for the most. The parameters of the DBSCAN algorithm are estimated and selected.

2 Traditional K-means algorithm and DBSCAN clustering algorithm

2.1 Algorithm of K-means

The traditional K-means algorithm is the most classical algorithm in the dynamic clustering algorithm. It initialises the original data and selects some points randomly as the cluster centre. Through several iterations, the clustering centre is modified until the classification is reasonable. The advantages of the algorithm are simple logic, easy implementation, and good performance for some data. The choice of 'distance' has a direct impact on the results. The steps of algorithm of K-means:

Step 1: Initialise the cluster centre, set the cluster number K value and iteration number initial value.
Step 2: Load data and calculate Euclidean distance from data points to centre points one by one.

Step 3: Iteratively and continuously update the cluster centre until the cluster centre does not change.
Step 4: The data are finally divided into some classes.

K-means algorithm uses Euclidean distance to calculate the distance between data and centre points. The formula is as follows:

$$D = ||X_i - Z_i||^2 \tag{1}$$

That is:

$$D = \sqrt{(x_i - x_j)^2 + (z_i - z_j)^2} \tag{2}$$

Among them, X is every data value and Z is iterative centre point. The distances between the K centre points and the data points are calculated, respectively, and the minimum distance and the maximum distance can be determined. The least distance is classified as a class.

$$Z_k = \frac{1}{N}\sum X \tag{3}$$

Z is an iterative clustering centre, and it is determined whether the cluster centre is consistent with the centre point of the $N-1$ cluster after the N iteration until the end of the iteration, and the clustering results are determined.

A large number of data are found by multiple K-means clustering. When the number of cluster numbers is $K=2$, two cluster centres, X1 and X2, are obtained by experiments. The middle point X3 of two cluster centres is obtained by calculation. Through X3, the vertical bisector of X1 and X2 line segments is found, and two types of data are found on both sides of the vertical bisector, as shown in Fig. 1.

2.2 Density clustering algorithm

When clustering large data, clusters will appear as clusters of arbitrary shape, so density clustering algorithm will play a great role. The DBSCAN algorithm is typical [6]. The DBSCAN algorithm can divide the high-density and connected data into clusters of arbitrary shape. Compared with the K-means algorithm, it is easier to get the global optimal. The same results will be generated by multiple operation. The steps of algorithm of density clustering:

Step 1: Determine the value of the neighbourhood radius and the minimum point number of the parameter
Step 2: Loading and reading the data
Step 3: Get any points and from points to all data points connected to density.
Step 4: Determine whether each data point is expanded or not completed
Step 5: Find the object set, classify and output

DBSCAN algorithm divides data points into core points, boundary points and noise points. When the data point is within the neighbourhood radius (epsilon) and the number is greater than the minimum point number (MinPts), the data point is called the core point. When the number is less than the minimum point number, the data point is called the boundary point. When the two conditions are not consistent, the data point is called the noise point.

2.3 Improved density clustering algorithm based on K-means

In DBSCAN density clustering algorithm, the minimum number and neighbourhood radius need to be set manually. The minimum number of points will directly affect the number of data clustering, while the neighbourhood radius directly affects the number of noise points [7]. In the K-means algorithm, the clustering results directly cluster the noise points with the data points, which make the noise points directly affect the clustering results, and cannot

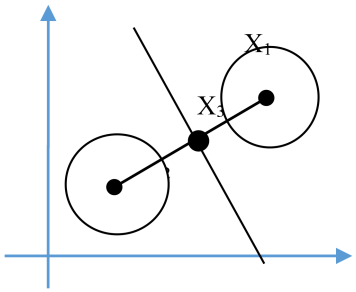

Fig. 1 *Distribution of data when K = 2*

form clusters of arbitrary shape through the connectivity of the data. So this paper proposes an improved density clustering algorithm based on K-means clustering to solve it.

After finding appropriate density clustering parameters, DBSCAN parameters are improved and optimised to achieve the best clustering results.

Step 1: The K-means algorithm is used to obtain two clustering centres, A and B, and the middle point C of AB, to calculate the Euclidean distance d1, d2 and d3 for each point to 2 cluster centres, A, B, and middle point C.

$$d = \sqrt{(x_i - x_A)^2 + (y_i - y_A)^2} \tag{4}$$

Step 2: According to the values of d1, d2, and d3, we calculate the number of points corresponding to m when d equals 1. D-M images are made, respectively, so as to find out three extreme points D1, D2, D3. We calculate the average value of D1 and D2, it is called 'D'.
Step 3: The value of M1 and M2 is used to determine whether the value is >50, so the value of initial neighbourhood radius is determined to beε_0. The minimum point number Minpts0 is determined according to the 1/2 of d3 value.
Step 4: Using ε_0 and Minpts0 to do DBSCAN training, we get the number of every kind of data points, and form X sets and statistics the number of elements in each set. When the number of points of a class decreases a little and the noise points appear gradually, after clustering the data in the whole data set, and we make density clustering.
Step 5: The relationship between the number of elements and the total number of each set is judged. Reducing m by 1 and iterating the ε.
Step 6: Until the point number of each class decreases a little after clustering, and the noise point appears gradually, the best clustering result is achieved.

3 Results

Here, we compare the accuracy of the improved algorithm through experiments. Data1 is a cluster data set provided by the literature [8]. Data2 is a cluster data set provided by the literature [9]. Data3 is verified and compared by the cluster data set provided by the document [7].

3.1 Experimental results of data1

When k equals 2, the two clustering centres of image are (19.709, 22.6331) and (19.4566, 7.6161). Making d-m images: $d=16.15$, m is about 70. $\varepsilon = D \div 4 = 4.0375$, Minpts = 6.85. Constant optimisation and operation of the epsilon value, and the final ε value is 1.3375. (Fig. 2)

3.2 Experimental results of data2

When k equals 2, the two clustering centres of image are (17.997, 36.0025) and (49.0037, 34.9763). Making d-m images: $d=16.15$, m is about 40. $\varepsilon = D \div 2 = 8.075$, Minpts = 11.5. (Fig. 3)

Constant optimisation and operation of the epsilon value, and the final ε value is 1.975.

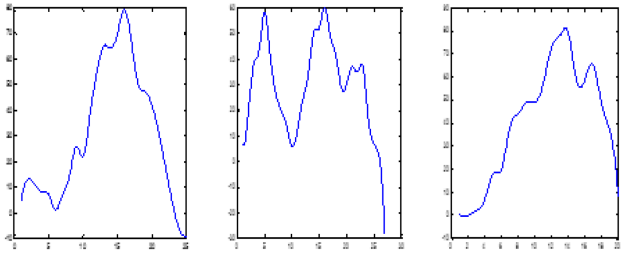

Fig. 2 *Data1 cluster centre d-m diagram*

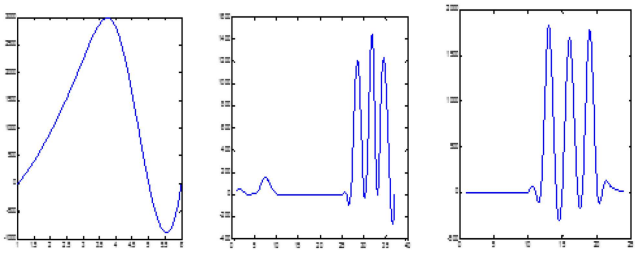

Fig. 3 *Data2 cluster centre d-m diagram*

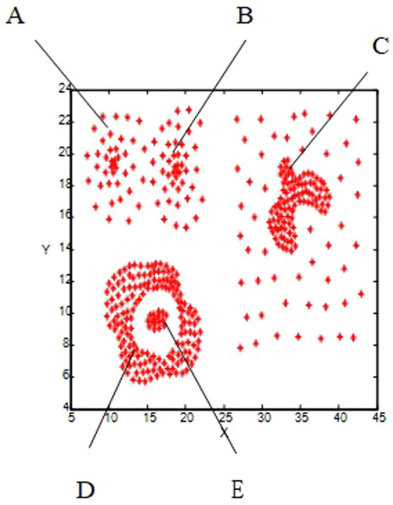

Fig. 4 *Original image of data3*

3.3 Experimental results of data3

The original image of Data3 is shown in Fig. 4:

When k equals 2, the two clustering centres of image are (34.3511, 16.3331) and (15.5095, 12.6654). Making d-m images: $d = 4.89$, m is about 45. $\varepsilon = D \div 2 = 2.445$, Minpts = 4.015. (Fig. 5)

Constant optimisation and operation of the epsilon value, and the final ε value is 1.352.

Using the improved algorithm, cluster analysis is carried out in block. The data points around the C class are processed according to the noise points, as shown in Fig. 6.

3.4 Data analysis

The clustering results are compared with the actual data labels, and the accuracy is shown in Table 1.

As can be seen from Table 1, the improved algorithm has appropriate clustering results. In data1 and data3, compared with the traditional K-means algorithm and DBSCAN algorithm, the improved algorithm improves the accuracy of the algorithm. For data2, DBSCAN density clustering algorithm and improved algorithm all have high accuracy. It is concluded that the improved algorithm reduces the human factors in the traditional algorithm, and improves the accuracy of the final clustering, and gets a better clustering result.

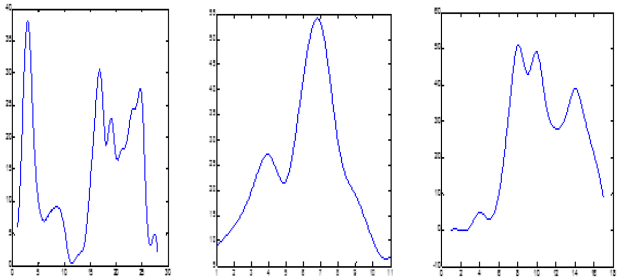

Fig. 5 *Data3 cluster centre d-m diagram*

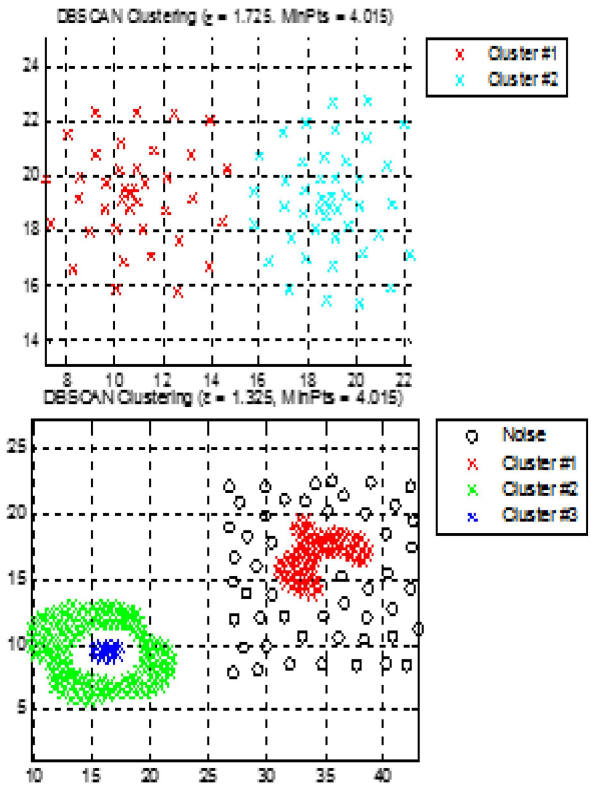

Fig. 6 *Data3 cluster*

Table 1 Comparison of data analysis of each algorithm

Clustering method	Data1	Data2	Data3
K-means	73.8%	75.8%	65.2%
DBSCAN	81.6%	100%	78.95%
the algorithm	99.1%	100%	87.47%

3.5 Application

3.5.1 Application to image clustering: Here, the improved algorithm is applied to image clustering. The picture used here comes from the handwritten open data set of MNIST.

MNIST is a classic demo for deep learning. These pictures are collected by different people from 0 to 9 handwritten digits. The Corinna Cortes of the Google laboratory and the Yann LeCun of the colon Institute at the New York University have a handwritten digital database. The training library has 60,000 handwritten digital images, and the test library has 10,000. The pixels of each picture are 28*28. Picture bit depth is 8 and the pictures are grayscale pictures. We have selected 24 pictures, and each number has eight pictures. The results of the classification are shown in Fig. 7:

The experimental results show that the first and second categories are correct, but the third class has four picture classification errors. The accuracy rate is 83.3%.

3.5.2 Application to UCI database of Iris: Here, the improved algorithm is applied to the clustering of natural data of Iris. The

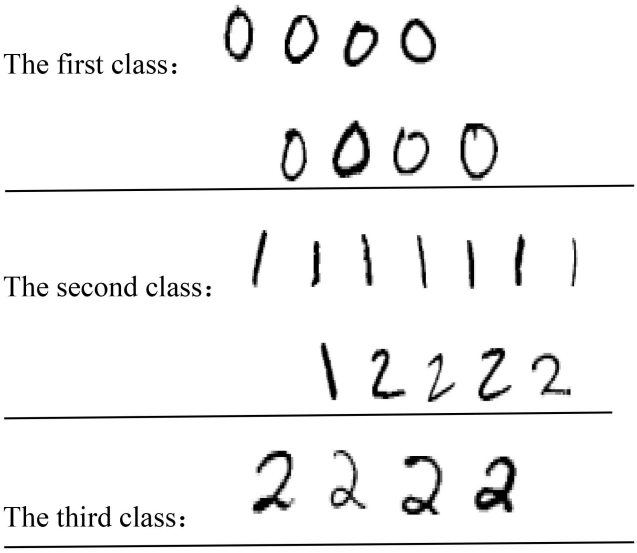

Fig. 7 *Clustering results of handwritten datasets*

UCI database is a database used by University of California at Irvine for machine learning.

There are 187 data sets in this database, and the number of them is increasing. The UCI data set is a common standard test data set. The iris data set is a two-dimensional table of 150 rows and five columns. The iris data set is a dataset used to classify flowers. Each sample contains four characteristics: calyx length, calyx width, petal length and petal width. We selected the dataset of IRIS as a test. IRIS database has four-dimensional data. By using the improved algorithm, the experimental results show that the accuracy of the method is 86.9%.

4 Conclusion

The neighbourhood radius and the minimum number of points have great influence on clustering. Choosing the appropriate neighbourhood radius and the minimum number of points is the key to get the ideal clustering.

Here, we adaptively select the parameters of the DBSCAN algorithm, which improves the computation speed and achieves the expected clustering results. The algorithm carries out two classification of the K-means algorithm, and then selects two important parameters of DBSCAN, Eps and Minpts, and adaptively selects the appropriate parameters by the improved algorithm. The improved algorithm overcomes the traditional algorithm of finding neighbourhood radius and minimum number of points. The experimental results show that the improved algorithm is applicable to the clustering of specific data and achieves the desired results. The improved algorithm solves the artificial interference.

However, the time complexity of the algorithm is relatively high, which will lead to the lower operation speed of the algorithm. In future research, we will focus on the optimisation and processing of the time complexity of the algorithm..

At the same time, the algorithm is applied to image clustering and natural data sets, and they achieve some good results. In the image clustering, the algorithm is effective. We can use this algorithm to filter out the bad pictures in the picture data set. It is also possible to categorisation natural data sets without labels, so as to exclude dissimilar objects.

5 Acknowledgments

The authors would like to acknowledge the support of Guangxi Natural Science Foundation (NO. 2015GXNSFDA13900, 2014GXNSFCA118017), the support of director fund projects of Guangxi Key Laboratory of Automatic Detecting Technology and Instruments (NO.YQ15101).

6 References

[1] Xiang, P.S.: 'Survey of clustering algorithm', *J. Southwest Univ. Nationalities*, 2011, **37**, (5), pp. 112–114

[2] Lei, X.F., Xie, K.Q., Lin, F.: 'An efficient clustering algorithm based on local optimality of K-means', *J. Softw.*, 2007, **7**, (7), pp. 1683–1692

[3] Wang, Y., Tang, J., Rao, L.F., *et al.*: 'High efficient K-means algorithm for determining optimal number of clusters', *J. Comput. Applic.*, 2014, **35**, (5), pp. 1331–1335

[4] Hu, R.F., Yin, G.F., Tan, Y.: 'A hybrid clustering algorithm and it is application', *J. Sichuan Univ. (Eng. Sci. Ed.)*, 2006, **38**, (5), pp. 156–161

[5] Wang, ZH.F., Shan, G.L.: 'K-means based method for dynamically selecting DBSCAN algorithm parameters', *Comput. Eng. Applic.*, 2017, **53**, (3), pp. 80–86

[6] Zhang, T., Liu, CH., Zhou, X.F.: 'Density clustering algorithm based on real core point', *Appl. Res. Comput.*, 2018, **35**, (12), pp. 3564–3568

[7] Zahn, C.T.: 'Graph-theoretical methods for detecting and describing gestalt clusters', *IEEE Trans. Comput.*, 1971, **100**, (1), pp. 68–86

[8] Gionis, A., Mannila, H.P.: 'Tsaparas: clustering aggregation', *ACM Trans. Knowl. Discovery from Data (TKDD)*, 2007, **1**, (1), pp. 1–30

[9] Rezaei, M., Fränti, P.: 'Set-matching methods for external cluster validity', *IEEE Trans. Knowl. Data Eng.*, 2016, **28**, (8), pp. 2173–2186