AN EFFICIENT CLUSTERING METHOD FOR DBSCAN GEOGRAPHIC SPATIO-TEMPORAL LARGE DATA WITH IMPROVED PARAMETER OPTIMIZATION

Jingwen Li [1,2], Xiaoqiang Han [1,2], Jianwu Jiang [1,2,*], Yao Hu [1,2], Lei Liu [1,2]

[1] Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin University of Technology, Guilin 541004, China
[2] Guilin University of Technology, Guilin 541004, China

KEY WORDS: Data Mining,Clustering Analysis, DBSCAN Density Clustering

ABSTRACT:

How to establish an effective method of large data analysis of geographic space-time and quickly and accurately find the hidden value behind geographic information has become a current research focus. Researchers have found that clustering analysis methods in data mining field can well mine knowledge and information hidden in complex and massive spatio-temporal data, and density-based clustering is one of the most important clustering methods.However, the traditional DBSCAN clustering algorithm has some drawbacks which are difficult to overcome in parameter selection. For example, the two important parameters of Eps neighborhood and MinPts density need to be set artificially. If the clustering results are reasonable, the more suitable parameters can not be selected according to the guiding principles of parameter setting of traditional DBSCAN clustering algorithm. It can not produce accurate clustering results.To solve the problem of misclassification and density sparsity caused by unreasonable parameter selection in DBSCAN clustering algorithm. In this paper, a DBSCAN-based data efficient density clustering method with improved parameter optimization is proposed. Its evaluation index function (Optimal Distance) is obtained by cycling k-clustering in turn, and the optimal solution is selected. The optimal k-value in k-clustering is used to cluster samples. Through mathematical and physical analysis, we can determine the appropriate parameters of Eps and MinPts. Finally, we can get clustering results by DBSCAN clustering. Experiments show that this method can select parameters reasonably for DBSCAN clustering, which proves the superiority of the method described in this paper.

1. INTRODUCTION

In recent years, with the rapid development of Electronic Science and technology and the deployment of a large number of sensor information devices, the efficiency of space-time data acquisition continues to improve. In our life, there are huge space-time data including location, time and environment attributes. Spatio-temporal data is growing at an explosive speed, with traditional data. Compared with these complex spatio-temporal data, there are more and more profound mining value. More and more researchers pay attention to mining useful information from massive spatio-temporal data. Cluster analysis method in data mining field can well mine the knowledge and information hidden in complex and massive spatio-temporal data. It has always been an extremely important part of data mining, and its participation has greatly affected the efficiency and effect of data mining. Higher attention has been paid to the subject of geographic information science.

Clustering is to use certain criteria to divide data sets into multiple classes or clusters, which can maximize the difference of different classes of data objects and minimize the similarity of similar data objects. Generally speaking, it is a classification method. [1] That is to classify data into different classes or clusters. With the deep understanding of spatio-temporal data information and the practical needs of data mining, clustering analysis has shown a new trend of integration and complementarity in recent years. A single clustering method is no longer suitable for massive and complex spatio-temporal data. This trend has also guided the attention of scientific researchers.

Clustering methods can be generally divided into five categories: partition-based method, hierarchy-based method, density-based method, grid-based method and model-based method [2,3].

The classical clustering algorithms include K-means algorithm, DBSCAN algorithm and hierarchical clustering algorithm.

K-means algorithm is a typical partition-based method. It is based on a given clustering objective function. The algorithm adopts an iterative updating method. Firstly, K class centers are randomly calculated as the starting point. Then the data are distributed to the nearest class center by distance, and the new class center is calculated accordingly. Finally, the steps of the above distribution and the calculation of class center are repeated until the class center is not changed or the number of iterations is limited. Finally, the clustering results are obtained to achieve better classification results. Although the operation is relatively simple, it also has some unavoidable defects. For example, it needs to pre-set the number of classifications, and its clustering effect is not ideal for data with complex structure. At the same time, it can not eliminate the interference of noise points, so it is not suitable. Clustering analysis of massive and complex spatiotemporal data.

Hierarchical clustering principle is to divide data sets at different levels to form a tree-like clustering structure. It includes bottom-up clustering and top-down clustering methods[4]. It does not need to pre-set the number of clusters, but because of the high complexity of hierarchical clustering, the computational power required is strong. And the singular value can also have a great impact. Comparatively speaking, the

* Jiang Jianwu - E-mail: fengbuxi@glut.edu.cn

operation of DBSCAN clustering based on density is more reasonable.

DBSCAN(Density-based spatial Clustering of Applications with Noise) is a spatial data clustering method based on density proposed by MartinEster， Hans-PeterKriegel in 1996.[5,6] It is also the most commonly used clustering method. The principle of this algorithm is to take the region with enough density as the cluster center, then calculate the density connection between them according to the density distribution between samples, and finally generate clusters that need clustering according to the region growing continuously according to the connected samples in order to obtain the final clustering results.

The extension of region generation according to density connectivity determines that it is suitable for clustering of arbitrary shapes and can effectively detect noise points and outliers. Clustering of arbitrary shape and dense data can be realized without pre-setting the number of clusters. Clustering results are relatively less disturbed by other factors, which can avoid the situation that is greatly affected by initial values. For complex space-time vector data such as shape and structure, clustering is very good. Therefore, this paper chooses the classical DBSCAN spatio-temporal density-based data efficient clustering algorithm to realize the clustering of large-scale geo-spatial data. Experiments show that this algorithm has strong universality and can effectively realize the clustering of vector spatio-temporal data of various shapes, sizes and densities under the interference of voice.

However, the traditional DBSCAN clustering algorithm requires a higher combination of Eps neighborhood and MinPts density, and different parameter combinations have a greater impact on the final clustering effect. It will also greatly affect the results of data mining.

For example, although the Eps field can be roughly obtained by k-distance graph method according to the basic guiding principles, if the parameter is too small, it is easy to cause data can not be clustered; if the parameter is too large, many data clusters will be highly concentrated in one class, and different clusters will be merged, resulting in wrong clustering. At the same time, it will also cause the problem of density sparsity. Similarly, the selection of MinPts density is the same. These artificially set parameters will have a great impact on the clustering results.

To solve the problem of misclassification and density sparsity caused by unreasonable parameter selection in DBSCAN clustering algorithm. In this paper, a DBSCAN-based data efficient density clustering method for parameter optimization is proposed. Optimal Distance is obtained by cyclic k-clustering, the optimal solution is selected, K-means clustering is carried out, and the clustering sample cluster is used for mathematical analysis to obtain reasonable parameters. DBSCAN clustering was performed.

2. METHOD

2.1 DBSCAN Density Clustering

The core idea of DBSCAN density clustering is: firstly, each point in the data set is set as the core point. If the density of MinPts can be reached in the Eps domain of the point, then a cluster will be formed with that point as the core. The unsatisfactory points are regarded as noise points. The boundary points of all the core points satisfying the density threshold within each distance threshold are determined, and each group of connected core points is divided into a cluster. Finally, the boundary points are allocated to the cluster of related core points. Continuous growth to achieve clustering results.

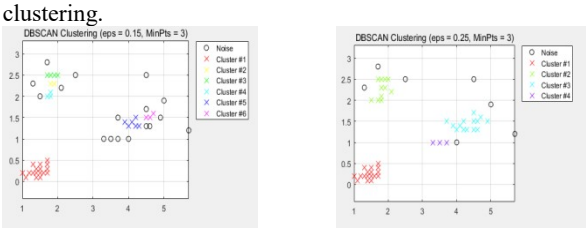
The specific steps are as follows:
1) Firstly, every sub-sample Ci in the sample set D= (C_1, C_2,..., C_m) is judged. If the sample set contained in the field of C_i Eps is greater than or equal to the density of MinPts, the sample is set as the core object, and then the core object is identified and added to the new core object set A.

2) Select a core object C_i randomly in the core object set A, and record the new core object sample set as A*={C_i}, the cluster object as k={C_i}, and update the unaccounted sample set &=&-C_i.

3) Take a core object from the cluster core object A* above. For the first time, only C_i can be obtained. The sample set contained in the C_i eps domain is marked as Ci*, Q=C_i*∩&. The intersection of the core object domain and the unavailable set is taken here to avoid duplicating the common parts between different core object sample subsets and to get the cluster. Initial cluster set k = k ∪ Q, and Q has been visited, so it needs to be removed from the sample set that has never been visited, so &=&-Q. Update cluster core object A* = A* ∪ (Q ∩A) - C_i.

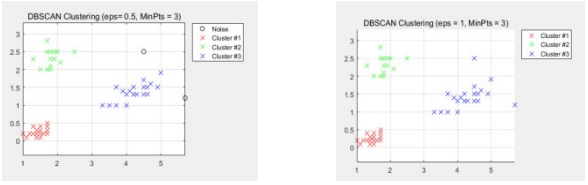
4) Because of the new core object sample set, non-accessed set and initial cluster object, we continue to take the core object for steps 2 and 3, and finally output: cluster partition k.

Above is the core step of DBSCAN density clustering. We find that the initial MinPs and Eps need to be set for clustering. As shown in the figure below, when the MinPs are the same, different Eps are selected. When the Eps are too small (Fig. 1), many samples are regarded as noise points (white points), and when the Eps are large, noise points cannot be eliminated. （ Fig. 2) Even if it is too large, most of the population will merge into one cluster, which can not achieve the purpose of clustering.



(a) Eps =0.15 (b) Eps =0.25

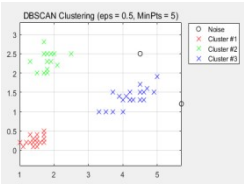
Figure 1.



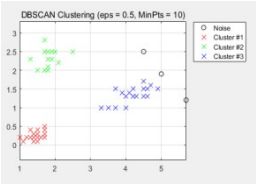
(c) Eps =0.5 (d) Eps =1

Figure 2.

As shown in the figure below, select different MinPts when Eps are the same.When MinPts is too small (Fig.3), most of the sample points will be clustered, and noise points can not be eliminated. When MinPts is too large, many sample points are misclassified as noise points (Fig. 4).

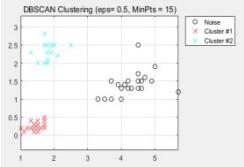


(a) MinPts =5

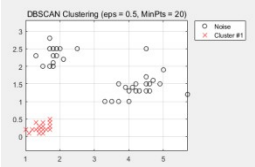


(b) MinPts =10

Figure 3.



(c) MinPts =15



(d) MinPts =20

Figure 4

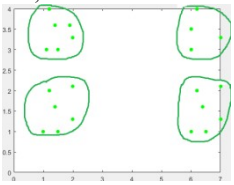
3. EFFICIENT CLUSTERING OF DBSCAN DATA WITH IMPROVED PARAMETER OPTIMIZATION

3.1 Basic thought

Improving parameter optimization is actually how to obtain appropriate parameters for DBSCAN clustering. Because K-means clustering algorithm has simple mathematical logic and principle and relatively easy operation, it can be used to obtain k-value, that is, classification number. Then, according to the relationship between the distance between the core object and the sample points in the clustering results, an evaluation index function is established to select the appropriate k value. Finally, the clustering result of K-means is obtained. According to the clustering result, the Eps neighborhood radius and MinPts density threshold for density clustering can be calculated by mathematical and physical analysis, and the final result can be obtained by DBSCAN clustering. It not only makes use of the characteristics of K-means clustering, but also makes full use of DBSCAN clustering to eliminate noise points. The following is a detailed deduction process:

1) Cyclic K-means clustering is used to get K-means clustering results with different K values.

2) The K-means clustering of a random set of data as shown in the following figure, in many cases, judging the validity of K-means clustering can not be separated from visual observation or empirical judgment, but the clustering effect can not be guaranteed when the data with complex features and large amount of data are used. As follows, when selecting k = 2 and k = 4, it is not intuitive to judge which kind of clustering is valid.



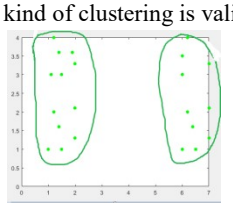


Figure 5.

In this paper, an evaluation index function is proposed to judge the reasonableness of k-clustering results. The characteristic of DBSCAN clustering is to satisfy the continual expansion of a regular clustering cluster according to the rule to achieve the clustering purpose. The K-means clustering results are generally round cluster shape, if the density is evolved according to the K-means clustering results. Clustering needs to decompose and merge the results of its clustering, so that the distance between the merged clusters is as small as possible, because it needs to meet the constraints of Eps and other constraints, while ensuring that the distance between the non-merged clusters and its sum is smaller, thus forming a variety of density-reachable classification clusters. K-means has a rule. For data set D, if the k value of clustering is x, that is, there are x clustering centers, for these clustering centers, the larger the k value is, the smaller the sum of the distance between the clustering center and the sample points with the center as the core point is. When the k value is larger, the sum of the distance between the clustering centers will be. The larger the distance, the more appropriate the k value is, the smaller the distance is. So the evaluation index function we set up can take the distance as the evaluation index. Because the original K-means clustering calculation takes the square sum of the distance between the sample and the center of mass to cluster, the distance in the evaluation index function here is replaced by the square of the distance. The formula is as follows:

$$S(k)=\frac{1}{N_K}\sum_{i=1}^{k}\sum_{x_i}(x_i-c_i)^2+\sum_{j=1}^{k}\sum_{l=1}^{k}(c_j-c_l)^2 \qquad (1)$$

Among them, k is the number of clusters, that is, the number of classifications, N_K is the number of samples with cluster i, x_i、c_i are the sample points and core points with cluster i, and c_j and c_l are the core points. The k value of the minimum index function is the required number of clusters.

3) Determine the values of each Eps and MinPts. The results of k clustering above are recorded as C_i, (0 < i < k). For the samples i n Ci, the distance S_K (0 < k< n (n-1)/2) between the two samples is calculated, where n is the number of sample points in the sample cluster C_i. The following is the distance interval and the number of distributions in the interval.

min(S)~min(S)+d	min(S)+d ~min(S)+ d·2	…….	min(S) +d·k~ max(S)
N_1	N_2	…….	N_K

Table 1

Because of the characteristics of clustering large data and the centroid distance as the criterion of k clustering, S_K and N_K in the distance set S of sample points within the cluster basically satisfy normal distribution, in which min (S) and Max (S) are the least, they are located on the left and right sides of the normal distribution graph. The S value with the largest probability in normal distribution is the Eps_i distance we want.which is clustering without noise points, when there are noise points, the number of S_K and it can only basically meet the characteristics of normal distribution. Here, according to the characteristics of left-right symmetry of normal distribution image, we can exclude some of the more obvious inconsistent sample point distance, which can be re-determined after excluding. The set of sense distances S, then there are

$$Eps_i = (\min(S) + \max(S)) / 2 \qquad (2)$$

Then the Epsi distances of each cluster can be obtained. The maximum number of samples in each cluster is MinPts, because if the number of samples in other clusters is taken, the subsequent clusters will be clustered into a large cluster, and the noise points can not be eliminated.

4) After obtaining the Epsi distances and the total MinPts of each cluster after K clustering, because the principle of

DBSCAN density clustering is the maximum density connected sample set derived from the density reachability relation, this joint sample set is a classification cluster we want. we can use the C_i cluster with distance of Min (Eps_i) as the starting point. Cluster DBSCAN density clustering, Judging whether the point b_i of all core points a density reachable in C_i is another core point, here judging whether b_i is the core point needs to use Eps and MinPts of cluster b_i belongs to. If b_i is another core point, this meets the idea of density-linked sample set in DBSCAN clustering. We can merge the samples with a and b_i as the core points into a new cluster. At the same time, we can update Eps to the larger values of the two samples, and then complete the density clustering of the whole data set by analogy. Third, if b_i is not a new core point, we can remove it as a noise point.

4. EXPERIMENT

Sklearns is an important machine learning library of Python. This paper chooses to use sklearns to generate data sets, totaling 4000 pieces of data, and draws scatter point graphs using matplotlib. According to the above improved method, when k = 4, the evaluation index function is 4.845573546695297, Eps = 0.1, MinPts = 10, as shown in the following figure. Fig. 2 is the result of kmeans clustering. It can be seen that the clustering results are very unsatisfactory. Fig. 3 is the result of DBSCAN density clustering after the improved parameter optimization. It can be seen that not only the clustering effect is reasonable, but also the noise points are eliminated, which proves the rationality of this experiment.

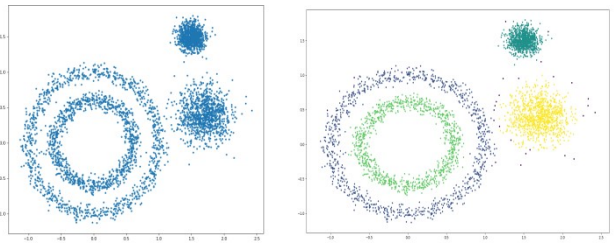


Figure 6.

5. CONCLUSION

Experiments show that this improved parameter optimization method can effectively avoid the problem of misclassification and density sparsity caused by unreasonable parameter selection of DBSCAN clustering algorithm. This method can reasonably select the parameters of DBSCAN clustering, and proves the superiority of this method.

REFERENCES

[1]Jain A K. Data Clustering: 50 Years Beyond K-means[M]// Machine Learning and Knowledge Discovery in Databases. 2008.
[2]Ester M, Kriegel H P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// International Conference on Knowledge Discovery & Data Mining. 1996.
[3] Shibing Zhou.Research and Application of Determining the Best Cluster Number in Cluster Analysis[D].wuxi.JiangNan University,2011
[4] Bianxia Shi.An Improved Hierarchical Clustering Algorithm[J] Microelectronics and Computer,2010,27(12):55-56.
[5] Pant J K, Lu W S, Antoniou A. A new algorithm for compressive sensing based on total-variation norm[C]// IEEE International Symposium on Circuits & Systems. 2013.
[6]Sander J , Ester M , Kriegel H P , et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications[J]. Data Mining and Knowledge Discovery, 1998, 2(2):169-194.