Advanced Science and Technology Letters
Vol.74 (ASEA 2014), pp.23-28
http://dx.doi.org/10.14257/astl.2014.74.05

G-DBSCAN: An Improved DBSCAN Clustering Method Based On Grid

Li Ma[1, 2, 3], Lei Gu[1, 2], Bo Li [1, 2], Sou yi Qiao[1, 2], Jin Wang[1, 2]

[1] Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China
[2] School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China
[3] Key Laboratory of Meteorological Disaster of Ministry of Education Nanjing University of Information Science & Technology, Nanjing 210044, China

Abstract. Clustering is one of the most active research fields in data mining. Clustering in statistics, pattern recognition, image processing, machine learning, biology, marketing and many other fields have a wide range of applications. DBSCAN is a density-based clustering algorithm. this algorithm clusters data of high density. The traditional DBSCAN clustering algorithm in finding the core object, will use this object as the center core, extends outwards continuously. At this point, the core objects growing, unprocessed objects are retained in memory, which will occupy a lot of memory and I/O overhead, algorithm efficiency is not high. In order to ensure the high efficiency of DBSCAN clustering algorithm, and reduce its memory footprint. In this paper, the original DBSCAN algorithm was improved, and the G-DBSCAN algorithm is proposed. G-DBSCAN algorithm to reduce the number of query object as a starting point, Put the data into the grid, with the center point of the data in the grid to replace all the grid points as the algorithm input. The query object will be drastically reduced, thus improving the efficiency of the algorithm, reduces the memory footprint. The results prove that G-DBSCAN algorithm is feasible and effective.

Key words: cluster analysis, DBSCAN, Grid, G-DBSCAN

1 Introduction

Data mining [1], also known as knowledge discovery, it was from a large number of incomplete, noisy, fuzzy, random data, extract potentially useful information and knowledge processes. The excavated knowledge can be applied to information management, decision support, process control, and many other applications [2, 3]. Now, data mining is one of the hot research fields in Computer Science.

Clustering analysis is an important part of data mining [4-6]. Clustering is the dataset into multiple classes, have a high degree of similarity between objects in the same class, the larger the gap between objects of different classes. By clustering, people can identify the dense and sparse regions, therefore, we can find the relationship between the data global distribution and data attribute [7, 8].

ISSN: 2287-1233 ASTL

Density-based method clustering objects according to the density. It generates clusters based on the density of the neighborhood object, or some kind of density function DBSCAN (Density - Based Spatial Clustering of Application with Noise) is a simple, effective density-based clustering algorithm [9-11]. Unlike the partitioning and hierarchical clustering method, DBSCAN defines clusters as the maximum density-reachable points. The high-density region can be divided into clusters. DBSCAN can find clusters of arbitrary shape, while fast clustering.

However, the DBSCAN algorithm is also has limitations. For example, when the data size increases, algorithm requires large memory support and larger I/O consumption. In the case of limited computing resources and a very large amount of data, the efficiency of DBSCAN will be greatly affected. The original DBSCAN algorithm will occupy a lot of memory problems, this paper proposes a DBSCAN clustering algorithm based on grid, namely G-DBSCAN. It attempts to use the grid to reduce the amount of data processed DBSCAN algorithm, while removing noise points, thereby reducing the memory footprint, improve efficiency of the algorithm. First, the data put into the grid, with the center point of all objects in the mesh to replace those objects in the grid as input data of DBSCAN, thus the number of data to be processed will be drastically reduced, but also can remove the noise point, this will further reduce the amount of initial input data, thereby improving the efficiency of the algorithm, reducing the memory consumption.

2 Methods

2.1 DBSCAN clustering algorithm

DBSCAN algorithm is a simple, effective density-based clustering algorithm, this algorithm uses the class of density connectivity quickly discover arbitrary shape clusters. The central idea of the algorithm: For a class of each object, the object within its given radius contains not less than a specified minimum number [12, 13].

Algorithm definitions:
Definition1. Core object: For a point P, take it as the center, taking Eps as the radius of a circle, the circle contains at least MinPts points, then the given object is the core object.

Definition2. Boundary object: In the Eps neighborhood of core objects, but do not meet the conditions of the core object, then these objects called boundary objects.

Definition3. Directly density-reachable: If P belongs to the Eps neighborhood of Q, and P is the core point, then object P is directly density-reachable from object Q

Definition4. Density-reachable: Object P is directly density-reachable from object Q, When there is a series of point P1, P2, P3……Pn=P started by Q, meet object Pi+1 is directly density-reachable from object Q, and Q is the core point, then object P is density-reachable from object Q.

Definition5. Density-connected: There exists a point O, if object P is density-reachable from object O, at the same time, object Q is density-reachable from object O, and then object P is density-connected to object Q.

Definition6. Noise points: P is neither core object nor boundary object, then it is a noise points.

The basic concept of DBSCAN:

Each point in the density-connected set is density-reachable. Choose any point P, if P is not classified, check that whether the P is the core point. If the point is the core point, find all points, they are directly density-reachable from object P. Form a new cluster with these points, assign an ID to each cluster. If P is a boundary object, then continue to access the next data point. Continue this process until all points have been processed. Finally, no ID points as noise points [14, 15].

2.2 The advantages and disadvantages of DBSCAN

Advantages:

Algorithm can find clusters of arbitrary shapes and sizes, automatically determine the number of clusters, isolated noise points, high efficiency and one scan can complete the clustering.

Disadvantages：

In the process of clustering, DBSCAN once found the core object, then this core object as the center outward expansion, this process will continue to increase core objects, unprocessed objects are retained in memory. If a large cluster exists in the database, it will require a lot of memory to store the core object information.

When the data density is not uneven, the quality of clustering is very poor. Input parameters sensitive. Parameter Eps, MinPts difficult to determine.

2.3 G-DBSCAN clustering algorithm

When understanding the G-DBSCAN algorithm, we need to know some relevant definitions.

Definition1. Grid size: grid side length defined according to the actual situation

Definition2. Noise-point threshold: the data grid is below the threshold, it will be treated as noise points. Threshold is generally designated by the artificial.

Definition3. Data center of the grid X:

$$X= (\sum_{i=1}^{n} x_i)/n \qquad (1)$$

Here, n is the number of data points in each grid.

Definition4. Distance calculation formula D:

$$D=\sqrt{\sum_{i=0}^{n} (x_i - x)^2} \qquad (2)$$

Here, n is the number of data attribute values.

Improved basic idea:

The average time complexity of DBSCAN algorithm is O (nlogn) (n is the number of data contained in the database). Most of the clustering process time is used in data query. In fact, DBSCAN clustering algorithm is a continuous process of data query. Therefore, if reduce the number of search data, we can reduce the memory footprint, improve the speed of clustering. Here, from the view of reducing the number of initial input data, we give a fast density-based clustering algorithm.

Although DBSCAN algorithm itself can remove noise points, it will also occupy memory space when judging the noise points. This also led to the processing speed of DBSCAN algorithm is slow. In order to improve the processing speed, in view of the above problem, we improved DBSCAN clustering algorithm.

1) Remove the noise point

We can first use the grid method to remove part of the noise points. The data points according to their attribute values assigned to the corresponding grid. Count the number of data points in each grid and calculate its data center. In the grid data number is less than a threshold point as noise points were removed from the dataset, thereby reducing the noise points in the data set.

2) Reduce the memory footprint

In the first step, the data has been assigned to each grid, and counts the number of data in each grid and its data center point. After removing noise points, the rest of the grid data center will be used as input data of DBSCAN. When get the clustering results, assign each grid point to the class that contains the center of the grid. Because the data is reduced, so the DBSCAN algorithm for data processing time is reduced.

3 Experiment Analysis

In this paper, the traditional DBSCAN algorithm and G-DBSCAN algorithm are compared. Test data sets taken from UCI database. UCI is a specialized database for testing machine learning, data mining algorithms. The data in the library have a certain classification, so it can be used to test the quality of clustering. In the UCI database, we selected Iris, Wine, Glass and Indian data sets to test, data set information as shown in Table 1. Where Number is the number of data, Attributes is the number of data attribute. In order to verify the improved algorithm, the distribution of test data remains unchanged.

In this paper, the original DBSCAN algorithm and the G-DBSCAN algorithm experiments were carried out four times. The noise threshold of G-DBSCAN algorithm is 1, that is, when the number of data in the grid is 1, we think it is a noise

point, and removing it from the dataset.

Table 1. dataset information

Datasets	Number	Attributes
Iris	150	4
Wine	178	13
Glass	214	10
Indian	768	8

Table 2 is a comparison of the experimental results, where Number is the number of data processing, Memory is memory footprint and Time is the time taken to achieve clustering.

Table 2. Experimental results Comparison

Datasets	DBSCAN			Grid-based DBSCAN		
	Number	Memory	Time	Number	Memory	Time
Iris	150	1.625696MB	46ms	27	0.975408MB	10ms
Wine	178	2.276008MB	114ms	51	1.300592MB	27ms
Glass	214	2.279256MB	133ms	16	0.975432MB	11ms
Indian	768	10.170304MB	368ms	58	1.300616MB	44ms

As can be seen from Table 2, the original DBSCAN algorithm has a lot of input data, takes up more memory and long running time. The G-DBSCAN algorithm is to divide some similar points to the same grid, after removing noise points, the rest of the grid data center will be used as input data of DBSCAN. When get the clustering results, assign each grid point to the class that contains the center of the grid and achieved the purpose of reducing the amount of input data. Data show that G-DBSCAN algorithm greatly reduces the memory footprint and program running time.

4 Conclusion

Now, spatial database is large in scale, and contains a large amount of information. The main task of data mining is to find useful information from the complex spatial database. Clustering analysis has been a hot research topic in data mining. Clustering is found similar data sets from large amounts of data. In this paper, through the analysis of the DBSCAN clustering algorithm, in view of its weaknesses, using the grid method to extend its performance so that it can effectively deal with large-scale spatial database. Experimental results show that the algorithm is feasible and effective. With the increasing scale of spatial database, data information becomes more and more complex. So in many applications it is difficult to select the appropriate clustering parameters. Therefore, the development of adaptive clustering algorithm will become an important part of our future research.

References

1. Chidanand Apte, Data mining: an industrial research perspective, Computational Science & Engineering, 1997, vol.4, no.2, pp. 6 – 9.
2. Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang, Combined Mining: Discovering Informative Knowledge in Complex Data, 2011,vol.41, no.3, pp.699-712.
3. Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamitra Bandyopadhyay, Carlos Artemio Coello Coello, A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I, Evolutionary Computation, 2014,vol.18, no.1, pp.4-19.
4. Xuqing Tang, Ping Zhu, Hierarchical Clustering Problems and Analysis of Fuzzy Proximity Relation on Granular Space, Fuzzy Systems, 2013,vol.21, no.5, pp. 814 - 824.
5. Ou Wu, Weiming Hu, Maybank,S.J, Mingliang Zhu, Bing Li, Efficient Clustering Aggregation Based on Data Fragments, Systems, Man, and Cybernetics, Part B: Cybernetics, 2012,vol.42, no.3, pp. 913 – 926.
6. Jianbin Huang, Heli Sun, Qinbao Song, Hongbo Deng, Jiawei Han, Revealing Density-Based Clustering Structure from the Core-Connected Tree of a Network, Knowledge and Data Engineering, 2013, vol.25, no.8, pp. 1876 - 1889.
7. Yihong Chu, Jenwei Huang, Kunta Chuang, Denian Yang, Mingsyan Chen, Density Conscious Subspace Clustering for High-Dimensional Data, Knowledge and Data Engineering, 2010,vol.22, no.1, pp.16-30.
8. Jian Hou, Xu E, Weixue Liu, Qi Xia, Naiming Qi, A density-based enhancement to dominant setsclustering, Computer Vision, 2013, vol.7, no.5, pp. 354-361.
9. Bin Jiang, Jian Pei, Yufei Tao, Xuemin Lin, Clustering Uncertain Data Based on Probability Distribution Similarity, Knowledge and Data Engineering, 2013, vol.25, no.4, pp.751-763.
10. Marzena Kryszkiewicz, Piotr Lasek, TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality, Rough Sets and Current Trends in Computing, 2010, Vol.6086, pp.60-69.
11. Smiti A, Elouedi Z, DBSCAN-GM: An improved clustering method based on Gaussian Means and DBSCAN techniques, Intelligent Engineering Systems (INES), 2012 , pp.573 – 578.
12. Andreas Thom, Oliver Kramer, Acceleration of DBSCAN-Based Clustering with Reduced Neighborhood Evaluations, KI 2010: Advances in Artificial Intelligence, 2010, Vol.6359, pp.195-202.
13. Abir Smiti, Zied Elouedi. DBSCAN-GM: An improved clustering method based on Gaussian Means and DBSCAN techniques, Intelligent Engineering Systems, 2012, pp.573-578.
14. Kellner D, Klappstein J, Dietmayer K, Grid-based DBSCAN for clustering extended objectsin radar data, Intelligent Vehicles Symposium (IV), 2012 , pp. 365 – 370.
15. Ming Huang, Fuling Bian, A Grid and Density Based Fast Spatial ClusteringAlgorithm, Artificial Intelligence and Computational Intelligence, 2009 , pp. 260 – 263.

