Hierarchical Clustering Analysis Method Based on the Grid with Obstacle Space

Shan Donghong, Yang Zhaofeng
Software School of Ping Ding Shan University
He Nan, Ping Ding Shan 467000
China

Journal of Digital Information Management

ABSTRACT: *The advantage of grid-based clustering method is its fast processing speed. The speed of clustering algorithm and the number of data objects is unrelated. To discover any size and shape of the clusterÿit is by the number of units on each dimension in the data space. In this method, the amount of data and computation time does not matter, calculations and data entry of the order does not matter, does not require the number of k-means algorithm to pre-specified cluster and so on. Clustering problem with obstacle constraints has very strong practical value in the spatial clustering analysis, and has become a research hotspot in recent years. Under the condition of existing obstacles constraints, the vast majority of the spatial clustering algorithm can't effectively solve the problem of irregular obstructions. Thus it has a greater impact on the accuracy of the algorithm clustering results, and reduces the efficiency of the algorithm. To solve this problem, an obstacle constraint space grid-based hierarchical clustering algorithm, which is GSHCO algorithm, is proposed. The algorithm inherits the advantages of grid-based clustering algorithm, by defining the concept of barriers to grid to deal effectively with the obstacles of arbitrary shape, to achieve the purpose of found clusters of arbitrary shape; At the same time, the algorithm uses a hierarchical strategy which can effectively reduce the complexity of the algorithm with obstacles clustering and the algorithm is improved operating efficiency. The experimental results show that the GSHCO algorithm can deal with obstacles constrained clustering, and with higher performance and better clustering quality.*

Categories and Subject Descriptors: I.5.3 [Clustering] H.2.7 [Database Administration]: Data mining

General Terms: Data Mining Algorithm, K-means clustering, Grid algorithms

Keywords: Spatial Cluster Analysis, Cluster Ensemble, Grid Spatial Clustering, Multi-scale Clustering

Received: 16 August 2012, Revised 24 October 2012, Accepted 29 October 2012

1. Introduction

In the process of cluster analysis, spatial data mining refers to the unknown extracted from the data warehouse or a large database, the implied value of the mode or information. It is a great application value in the database research field, and combines many areas of statistics, databases, and machine learning theory and technology. Since the beginning of the 21st century, along with the rapid development of the information age, all walks of life have been widely used a lot of space database, resulting in spatial data mining areas of the rapid rise. Spatial clustering methods to solve and analyze the complex problem of the lack of background information reflects the tremendous superiority of the spatial clustering method has become a spatial data mining field of the most active and most widely used technology [1, 2] .

Cluster analysis is widely studied in data mining one of the topics to find the similarity between the data from the data, and so the data classification, to discover useful information or knowledge implicit in the data. Now in community, widely used clustering algorithm can be divided into the following types: level-based approach, based on the partitioning method, grid-based methods, model-based approach and density-based approach. However, for a particular clustering method, it is generally the results of a variety of clustering integration together. From point of view, it can't simply be attributed to a single category from the type.

A lot of data clustering algorithms have been proposed, one of the more well-known is BIRCH, CLARANS, CLIQUE and DBSCAN and so on. For the efficient clustering of large-scale and high-dimensional database analysis is still a open-ended questions to be examined. Spatial data processing in spatial data discretization commonly used method is the grid method. Clustering algorithm based on grid methods to achieve ease of high-dimensional data processing and incremental clustering algorithm in a wide range of applications. Researchers have proposed a variety of grid-based clustering algorithm, including STING, Wave Cluster and CLIQUE. STING methods use the statistical information stored in the grid cell; Wave Cluster is the wavelet transform method to the data object clustering; CLIQUE clustering method based on grid and density in high-dimensional data space.

The advantage of grid-based clustering method is its fast processing speed. The speed of clustering algorithm and the number of data objects is unrelated. To discover any size and shape of the cluster, it is by the number of units on each dimension in the data space. In this method, the amount of data and computation time does not matter, calculations and data entry of the order does not matter, does not require the number of k-means algorithm to pre-specified cluster and so on. However, the clustering of grid-based methods of clustering results by the influence of input parameters, and difficult to set these parameters. If the data has noise, the algorithm clustering quality will be poor in the case of without special handling. Moreover, the algorithm of the data dimension for the scalability is poor. Grid-based clustering technique is an important technical direction of the clustering technique. Firstly, the existing grid-based clustering algorithm is researched. From the grid meshing unit to the various steps of the information in the statistics grid is researched too. Finally, the clustering of grid-based method research is put forward.

2. Spatial hierarchical clustering algorithm

2.1 Introduction to Cluster Analysis

The so-called clustering is a data object to group making more than one cluster or class. After generated the cluster is a collection of a set of data objects. In the same cluster objects with each other have high similarity, but objects in different clusters are different.

1. The pros and cons of the measure of spatial clustering analysis

The purpose of the clustering algorithm is automatically classified the data object into meaningful clustering. The guiding principle of the clustering algorithm is pursuit of higher within-class similarity and the lower class of similarity between. The pros and cons of a clustering algorithm can be measured from the following aspects:

2.1.1 Scalability

Because of the spatial database contains extremely rich data, the structure is more complex, we must seek a more rapid and effective clustering algorithm. Its running time must be acceptable, predictable, so the time complexity of the algorithm of polynomial or exponential no practical value.

2.1.2 Ability to handle different data types

The spatial database contains data not only have complex spatial properties, but also contains the non-spatial attributes. Generally contain a variety of different data type, a good algorithm should be able to handle different data types.

2.1.3 Able to identify all type of spatial clustering of arbitrary shape

The specific features of spatial clustering is not known prior to analysis. Spatial clustering contains a variety of complex shapes, so a good algorithm should be able to identify clusters of arbitrary shape.

2.1.4 The ability to deal with noise

Often the database contains vacancies, outliers, bad data and unknown data in reality. If the clustering algorithm is sensitive to such data, it will lead to poor clustering results.

2.1.5 High dimensional spatial data

Spatial data will generally contain a higher dimension. How to clustering spatial data mining is special requirements on clustering in the high-dimensional space.

2.1.6 Usability and understandability

Results of clustering should be understandable and available for the user.

2.2 Spacial Data Mining

Spatial data mining is different from the general data mining and the conventional transaction database. Spatial cluster analysis is not only an important part of spatial data mining, but also widely used in spatial data mining and in-depth study of one of the elements.

Compared with traditional data mining, the main different point performance of spatial data mining is in the following areas:

(1) In traditional data mining processing, the processing content is the numbers and types. But in spatial data handling, processing is more complex data types, such as: points, lines and polygons and other objects.

(2) Traditional data mining is usually explicit input, but it Is often implicit in the spatial data mining data input.

(3) Assumptions in the traditional data mining: The data sample is generated independently. However, this assumption is not established in the spatial data mining. In fact, there is a high degree of correlation between the spatial data, for example: People with similar occupational characteristics and background are usually easy to gather in the same region.

3. Grid-based clustering method

3.1 The main idea based on the clustering of the grid (grid-based clustering)

The main idea is based on the clustering of the grid (grid-based clustering): In order to form a grid structure, the first object space should be quantified making a limited number of units. In each unit, the statistical parameters of the storage object, such as maximum, minimum, distribution type, variance and mean and so on. Then, you can all clustering operations on the space of this quantization. Grid-based algorithm execution time is not the number of data objects. In other words, it is come up with the number of modules, so this method to analyze the data speed is fast.

In STING methods, statistical information grid approach to spatial data mining algorithms. It is usually multi-resolution approach to clustering, and the lowest level of granularity determines the quality of the cluster. Wave Cluster is a multi-resolution clustering methods, this method is to transform the original feature space by the wavelet transform. Its main idea is: first to quantify the feature space, the data is mapped to a multidimensional grid. After wavelet transform, the grid unit cluster by searching for connected components. The complexity of wave Cluster in the calculation of the generated clustering is low, the advantage of this method is high efficiency, large-scale data processing can be a good isolated point, is not sensitive to the input data sequence, and the discovery of the complex structure of poly classes.

The clustering of high dimensional space is also a hot topic in the clustering algorithm, CLQUE comprehensive method based on grid and density. This is a very effective clustering algorithm for a clustering of high dimensional data in large databases.

3.2 The meshing of Spatial clustering method

Theoretically, grid clustering is using the cover technology, which is the merger of the adjacent high-density grid cell to form a cluster. However, this clustering method, the division of the fixed grid there are many defects: (1) fixed grid position will cause the loss of a small cluster. If a clustering is divided into a grid of multiple, adjacent, and there is no formation of high-density grid, so that you can't generate the clustering; (2) The accuracy of the clustering results by the size of the impact of the grid. If the grid cell is too large, then the clustering around a grid may contain noise; but if the grid unit is too small, will generate a large number of grid clustering efficiency have been seriously affected. So that the grid size have an impact not only on the precision of clustering, but also had an impact on the efficiency of the clustering; (3) Clustering accuracy is not high. If the boundary of the cluster falls into the low-density grid cell, the cluster will be lost boundary [3].

We can see that the fixed mesh technique is not applicable for high dimensional data clustering. For example, if there is a 100-dimensional data set, divided into 10 intervals in each dimension, so that you can generate 10^{100} mesh. In dealing with so large a number, the computer's performance will decline rapidly, but also the so-called "*dimension disaster*". Therefore, the fixed grid clustering technique is very suitable for high density of large data sets. As the dimension increases, the clustering efficiency of the fixed grid technology is rapidly declining. Therefore, the grid clustering technique in-depth study things is very necessary. An effective way to improve the clustering efficiency include the following: improve the efficiency of the search for neighbors and reduce the number of grid cells, leading to adaptive meshing technology; Split point in the calculation of optimized adaptive meshing technology-based clustering method requires a large amount of computation. In order to solve the problems of the above appear, here we propose a new coverage of the grid-based clustering algorithm CBOG(Clustering based on Overlapping Grid). This algorithm uses the generation of grid technology to reduce the number of grids dynamic; to eliminate noise and outlier data because the use of the concept of density; the use of border processing technology to improve the quality of the clustering, and eventually covering technology to connect the grid to the formation of cluster.

3.3 Meshing structure

The first step in the grid-based clustering algorithm is divided into a grid structure. Search sub-space strategy, there are mainly based by the end of the grid partitioning algorithm and based on the method of top-down meshing algorithm. The basic idea of subspace clustering algorithm is as follows: to find out clusters directly in high-dimensional space is more difficult, considering the original data space is divided into different sub-space from the corresponding sub-space to examine the presence of clustering. For subspace of the high-dimensional data space, it is difficult to understand the relatively abstract visual description high-dimensional space since the space.

Subspace clustering and feature selection is similar to the attempt to found clustering in the same data set to select a different subspace, expanding the feature selection task. Subspace clustering need to use appropriate search strategy and evaluation criteria to select the found subspace clustering search strategy and evaluation criteria, which has a significant impact on the clustering results.

At present, depending on the search direction the subspace clustering methods can be devided into two categories: from the bottom-up search and top-down search method.

3.3.1 Search method from the bottom

This approach narrows the search space by clustering the monotony of the dimension, or is the a priori nature of association rules: If a d-dimensional unit is dense, then its projection on the d-1 dimensional space should also be intensive, which is the space of closed; In contrast,

d-1 dimensional space given data unit is not intensive, it is in an arbitrary d-dimensional space is not dense. Such algorithms are that each dimension is divided into a number of grids by using subspace clustering algorithm based on density clustering and grid-based method of combining cluster analysis. In this strategy the most prominent problem is easy to produce overlapping clusters, that is, certain points are more than one cluster, or do not belong to any one cluster. What's more based on the density and the density of the grid method the problem threshold step size and grid parameters dependent still exists.

3.3.2 Top-down search method

First the entire data set is divided into k parts, and the same set of weights given to each cluster; Then, you want to repeat some strategy for continuous improvement of these initial clusters, and to update the cluster weights. This strategy have been established for each part of the data clusters, and so to avoid the cluster due to repeated, and then to ensure that a data point exists and exists only in a cluster. However, for large databases, repeat the process of computing the required computational cost is quite high, so most of these algorithms use some strategies to improve the performance of the algorithm on a small part of the inspection data as a data sample. Similarly, in this search strategy the same or similar cluster size, cluster number of parameters will produce heavily dependent. The number of samples is also an important parameter, using the algorithm of the sampling strategy, the values of these parameters on the final clustering results have a significant impact.

3.3.3 The comparison

For top-down division method, its main advantage is that users do not need to specify the division parameters. It can be divided according to the distribution of the data space, so it can be said that this division is more reasonable. In top-down grid method, the influence of the data space dimension is small. This method can quickly large-scale high-dimensional data set of cluster separated. This method is the computational complexity of the data set size and dimensions were tested linear relationship, so this method of handling high dimensional data is more appropriate. Because it is based on the data distribution of the mesh, but usually that noise in the entire space is uniformly distributed. As a result, from the top-down division method is not sensitive to noise. However, this method of grid unit volume is much larger than the Internet grid method in the grid cell size by the end of the description of the cluster accuracy, so the method of cluster description of accuracy than by the bottom-up grid method lower. Usually in the top-down meshing process, end the same area may contain different clusters. With a cluster may be divided into different areas, thus further reducing the accuracy of the algorithm. Another drawback of the method of such division is in the process of division, the need for the data sets, multiple scans.

In contrast, bottom-up divide just a linear scan data sets and can get higher precision. So, these two methods were applied on different issues. The former is suitable for handling high-dimensional data sets, which can effectively deal with the access cost of a large dynamic data and large data sets.

4. Hierarchical cluster analysis algorithm based on the grid with obstacle space

4.1 The ideology of the hierarchical cluster analysis algorithm based on the grid with obstacle space

Integrated with grid-based spatial hierarchical clustering algorithm from the obstacle constraints density and grid-based clustering algorithm, and introduced of the concept of barriers to the grid on the basis of the CLIQUE algorithm. Based on density and grid-based clustering process, the algorithm fully considers the impact of barriers to grid clustering results, and ultimately come to meet the actual needs of the clustering with obstacles constraints. That the proposed algorithm using a hierarchical clustering strategy to reduce obstacles constrained clustering algorithm time and in the cost of reduce space complexity. First of all barriers to the grid edge of the grid extension line of the data sample space is divided into several sub-region does not contain barriers to grid. Hierarchical clustering is in the various sub-regional accessibility constraints clustering. Then the center of a cluster is clustering two clustering with obstacles constraints in order to reduce the clustering with obstacles constraints on calculating the amount. Under the premise of spatial data with obstacles clustering the computational efficiency of the algorithm is to be improved. It can be found obstacles connectivity adjacent clusters formed by the merger of new clusters, to avoid spatial clustering algorithm with obstacles meaningless over-clustering problem [4].

GSHCO algorithm combines the advantages of density - grid algorithm, which can not only deal with the obstacles of arbitrary shape, but also be able to handle large-scale data sets and found that clusters of arbitrary shape. Also it is easy to carry out high-dimensional expansion. Similar to the CLIQUE algorithm, in order to effectively reduce the computation of the algorithm for high-dimensional processing, the processing of high-dimensional space first post-dimension data projected onto the k-1 dimension. Candidating dense units in the post-dimension space and then inferring k-1 dimensional space-intensive unit. GSHCO algorithm is for two-dimensional space, but the results above can be easily extended to high-dimensional data sample space. At the same time, adopting a classification strategy can effectively reduce the computational complexity, and reduce the point with obstacles clustering for dealing with computation. Eventually, this method can improve the efficiency of the algorithm, and incremental processing and effective treatment of data is the uneven distribution of the sample space.

Because of the presence of obstacles, the calculation of the distance between each data point is more complex. The complexity of distance calculation should be

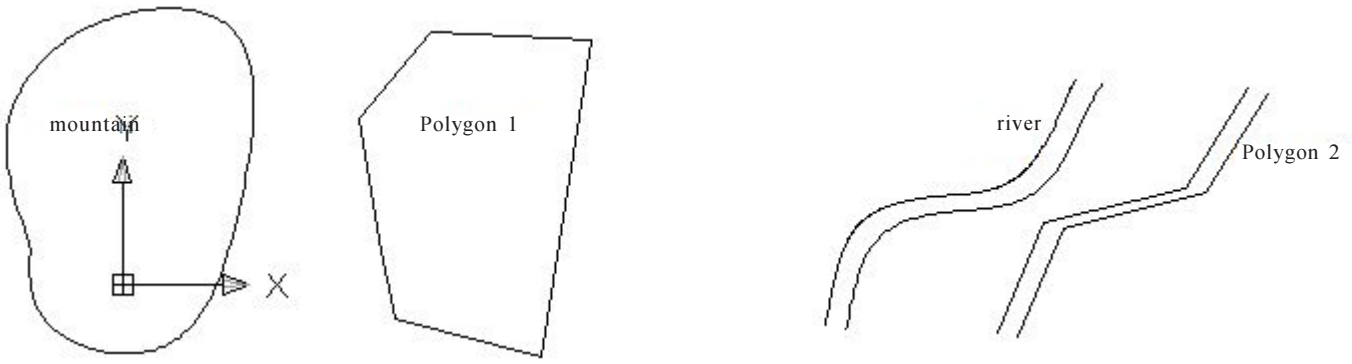

Figure 1. Obstacle Polygon Description

minimized in order to better the clustering process. The solution is to increase the pretreatment process to merge the obstacles in the space during the clustering. On the basis of the obstacle polygon description, in the pretreatment process merger is a merger of the polygon. In most cases, the merger of the polygon can reduce the number of its edge, thereby increasing the speed of data processing. Typical obstructions polygon merging process is shown in Figure 2.

4.3 The obstacle distance calculation

The presence of obstacle constraints directly affects the spatial distance between sample points. The distance of sample points to each cluster center is the basis for cluster operation, which is also the basis for clustering quality evaluation. Therefore, the distance calculation with obstacle constraints data space between the sample points with obstacles, naturally become the primary problem of spatial clustering with obstacles constraints.

If sample data space barriers between the two data points are directly up, we can use the Euclidean distance calculation two spacing; If two data points separated by obstacles blocking, we can use the visualization the calculated obstacle distance between two points, shown in Figure 3. For obstacle distance between two points, select the minimum visual distance. Figure 3, the obstacle distance of p, q between two points is visual distance of the obstacles $O1$ and $O2$ each vertex between the minimum V6V9. Thus, obstacle distance has the high computational complexity. So, effectively reduce the complexity of the clustering time and space with obstacles effective way to reduce the points of dyscalculia distance logarithmic. When obstacles exist connected, usually through the connectivity at the visual distance of the obstacle distance between two points.

4.4 Algorithm analysis

The first parameter putting u gridded data sample space is algorithm preprocessing GSHCO. Because of the presence of obstacles undermine the connectivity of the data sample space, deal with obstacles in the algorithm will merge polygon coverage and its borders through the grid to form a continuous low-density region of the obstacle grid. And through the obstacle grid boundary of the grid extension line of the sample space is divided into a number does not contain barriers to grid-connected sub region. Then, according to density threshold MinP , we can determine that the grid is the sparse grid or dense grid. When the pretreatment is completed, all of the sample space grid corresponding mark, marked as obstacles to the grid, dense grids or sparse grid. This algorithm uses the classification strategy. First of all, we

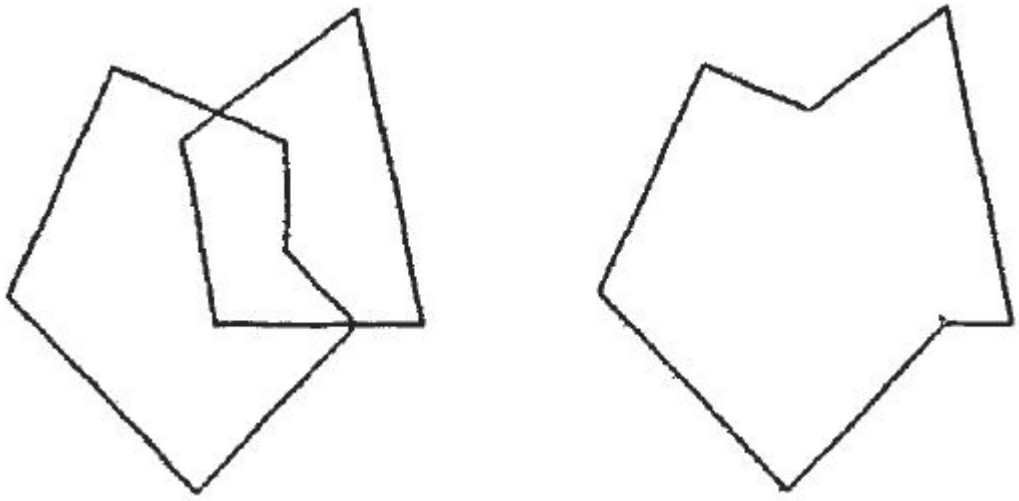

Figure 2. Obstacles Polygon

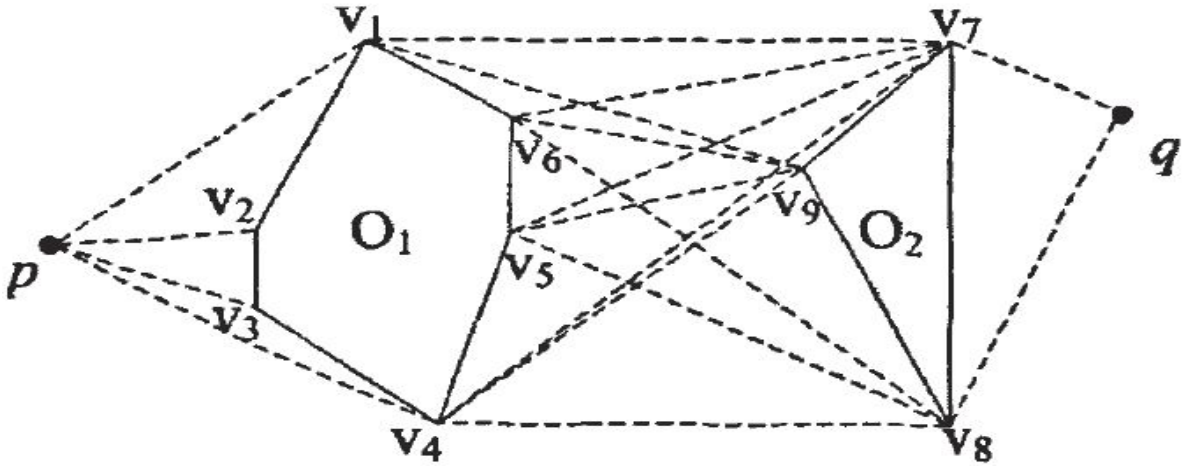

Figure 3. The obstacle distance

want to pair region accessible a cluster based on the grid - density method. After this two cluster for a cluster to the cluster center. The main purpose is the artificial division of the same cluster of data points caused by obstructions divided sub-cluster merger, and the discovery of the cluster of obstacles connected cross obstacles. Using a cluster represents points of the two clustering with obstacles. This will be able to reduce the computational complexity of clustering with obstacles constraints. Specific process GSHCO method, is described as follows:

Input: Obstacles collection D space of the data sample D, the density threshold MinPg and grid step size u;

Output: the sample space obstacle constraints clustering results;

Specific steps:

Step 1: Mesh

Processing the data sample space grid, first to determine the obstacles cover and border crossing grid barriers to the grid. To do the "*obstacle grid marks, and record barriers to grid density count* (Q) = 0". At the same time, calculate the number of data sample points count in the non-obstacle grid (C).Based on MinPg, we can make a judgment or dense grid for the sparse grid and do the marks;

Step 2: Sub-regional division

The division of the obstacle pair region does not affect the sub-regional connectivity. It should be divided in accordance with the extension line of each side of the polygonal obstacles in the combined grid for the entire sample space. If the divided sub-region still contains obstacles, should continue to be divided until each sub-region does not contain any obstacles:

Step 3: A sub-regional barrier-free clustering

For promoter region that does not contain any obstacles, these regions should be in accordance with the grid - the density clustering algorithm to a cluster. Traverse all the sub-region of the promoter region, and do the following sub-operations:

(1) All non-obstacle mesh "*unhandled*" tag;

(2) Determine whether there is ""*unhandled*" mark all grid in the promoter region, if not, deal with the next sub-region; perform (3);

(3) Grid marking to "*deal*", if the grid is sparse grid will have to return to re-run (2); Otherwise, the dense grid should be given a new cluster logo. Inspection of the grid adjacent grid, connected to the density and given the current cluster logo. Connected to the density of each dense grid density connected until the formation of density connected region. Ultimately each grid within the region are the current cluster logo;

(4) Modified cluster identification, and return to the implementation of (2) to the next cluster to find, until the promoter region of the grid is marked as "*processing*";

Step 4: Two clustering with obstacles based on the obstacle distance

A cluster generated by each cluster the cluster center as a representative point, according to the connectivity of the obstacle the obstacle distance. Carried out with the obstacles clustering. Because a small number of data points involved in the two clustering with obstacles, in this paper is COD. CLARANS algorithm;

Step 5: output clustering results

For the formation of clusters in the two clusters of data points, a clustering process represent a class of data objects in the end by belonging to a class, giving the logo with a cluster. Thus, the formation of the clustering results of the entire data sample space.

4.5 Algorithm evaluation

Theoretically, DBCLUC algorithms can effectively deal with the obstacle constraints clustering. For arbitrary shape obstacle is not ideal, which is easy to form scattered clusters. This article GSHCO algorithm inherited the advantages of grid-based clustering. It can handle the obstacles of any shape and form clusters of arbitrary shape, and because of obstructions caused by the scattered small clusters adjacent cluster integration. Avoid meaningless small cluster to form a new cluster. This makes the clustering result of the spatial data sets more

in line with the actual situation. Its degree of accuracy is higher than DBCLUC algorithm.

By using hierarchical clustering strategy GSHCO algorithm, the algorithm's computational complexity is reduced. And compared to its computational efficiency DBCLUC algorithm has improved significantly. Experimental the GSHCO operating efficiency of the algorithm were investigated under the conditions of the same barriers to change the scale of spatial sample collection. It can be seen through the experimental comparison, when changes in the number of sample points, GSHCO algorithm with running time and the number of sample points is basically a linear correlation. The algorithm DBCLUC with the purpose of the number of samples increases computation time increased even more sharply. GSHCO algorithm running time is always less than DBCLUC algorithm. If the data is large scale, the more apparent difference there is between the two. This verified the clustering process of this article GSHCO algorithm to adopt a classification strategy, and reduce the number of the points involved in the complex with the barriers calculated. Therefore we can say the GSHCO algorithm effectively reduces the time complexity of the algorithm and improve the efficiency of operations.

In summary, verified by experiments this article GSHCO algorithm for the validity of spatial data with obstacles clustering. And be able to find a clustering of obstacles exist, scattered small cluster merger, to avoid meaningless over-clustering; By classification strategy, the algorithm GSHCO can effectively reduce the computational complexity, and reduce the points for obstacle distance calculated on the number. This is an ideal method to cluster for large-scale sample space with obstacles.

5. Conclusion

Finally, the advantages and limitations of the grid clustering method grid-based clustering methods have been analyzed and summarized, including the definition of the grid, the grid cell density to determine the grid division method are summarized. In this paper, draw the following conclusions:

(1) Based on the grid - the density of spatial clustering algorithm does not consider the influence of data points within the adjacent grid currently examining the grid. This method the clustering results are not smooth, cluster boundary is not sufficiently clear. This need effectively ways find a close neighbor unit. When the huge data sets and data with geographic distribution feature, you need to develop effective algorithms to improve the processing speed.

(2) The GSHCO algorithm can effectively carry out the clustering of spatial data with obstacles. And this algorithm can find a cluster of obstacles exist, scattered small cluster merger to avoid meaningless over-clustering.

(3) Using classification strategy, the algorithm GSHCO can effectively reduce the computational complexity, and reduced point for obstacle distance calculated on the number of this approach. So the clustering of large-scale sample space with obstacles is very applicable.

(4) The optimization of existing grid algorithms is from different aspects to improve the effectiveness of the grid algorithm. On the basis of existing research and analysis, follow-up need to focus on resolving the problem. such as developing a compression algorithm and sparse grid density similar to the grid merging algorithm.

There is a broad space for grid clustering algorithm in the future, which will be used in more fields and will play a more powerful role.

References

[1] Kebiao, Mao., Li, Zhihao., Ruohong, Haitao, Zhou. (2002). Spatial data warehouse-based spatial data mining. Remote sensing information (theoretical research).19-26.

[2] Ruiju, Zhang., School, Tao. (2003). GIS and spatial data mining study of integration issues. *Investigation Science and Technology.* 2 (1) 21-24.

[3] Huiping, Chen., Yu Wang, Jiandong. (2007). New Progress in Research of the subspace clustering algorithm. *Computer Simulation*. 24 (3) 6-10.

[4] Wu, Xiekun, Ye Bin, Bi Xiaoling. (2007). Based on unit area of high dimensional data clustering algorithm. *Computer Research and Development.* 44 (9) 1618-1623.