AN IMPROVED CLUSTERING METHOD BASED ON DENSITY AND DIVISION METHOD

Zhang Qiu-Ju
School of Management and Economics
Beijing Institute of Technology, Beijing
100081, China
zhangqj1983@hotmail.com

Abstract. Combining partitioning and density - based clustering method, an improved clustering method is proposed in this paper on the basis of objective clustering algorithm. Firstly, the points with greater density which are distant from each other were selected as the initial centers of K-means clustering. Then, Use K-means was used to roughly determine the elements contained in each class. Afterwards, the points with the largest density in each class were searched and taken as the centers to re-conduct K-means clustering. If a class has more than one maximum density points, then the class will have multiple clustering centers, which make the shape of the class not round any longer and facilitate the classification of irregular shapes. Finally, by the dipole idea of the objective clustering algorithm, the optimal number of clusters was determined. The improved algorithm proposed in this paper achieved very good results in the clustering tests on random data set and UCI data set

Keywords: clustering analysis, K-means, density-based clustering method, objective clustering algorithm.

1. Introduction

In the analysis and description of the world, class represents a group of objects with public characteristics. People divide objects into different classes to explore the commonality between the objects of the same class and the gap between the objects of different classes, which are the information that can not be obtained by other methods. In the case of data analysis, clusters are potential classes and clustering analysis is a technique for discovering these classes automatically.

Clustering analysis refers to the analysis process which divides the sets of physical or abstract objects into classes made up of similar objects, with a goal to collect data on a similar basis to classify. Clustering has been applied and developed in many fields, including mathematics, computer science, statistics, biology and economics, which is mainly used for describing data, measuring the similarity between different data sources and classifying data sources into different clusters.

Traditional clustering analysis methods are mainly partitioning method, hierarchical method, density-based method, grid-based method and model-based method, with algorithms that are widely used in each method, such as K-means

clustering algorithm in partitioning method, cohesive hierarchical clustering algorithm in hierarchical method, neural network clustering algorithm in model-based method, etc.

In general, a partitioning-based clustering method firstly requires a given number of clusters to create an initial division and then applies an iterative relocation technique to move the objects between partitions to achieve the final division. Each class contains at least one sample point, each of which belongs to and can only belong to one class. The typical division methods are K-means algorithm, CLARANS algorithm and FREM algorithm. Partitioning-based clustering requires the size of the number of clusters to be specified in advance. If the values are not appropriate, the clustering rationality will be affected. Besides, partitioning-based clustering basically belongs to spherical clustering and its division effect on classes with irregular shapes is not obvious.

Hierarchical clustering methods combine data objects into a clustered tree, which can be further divided into agglomerative type (with bottom-up hierarchical decomposition) and divisive type (with top-down hierarchical decomposition) hierarchical clustering. Aggregated hierarchical clustering initially treats each sample point as a class and then merges the different classes in each iteration process until the preset clustering effect is reached. Split hierarchical clustering initially treats all sample points as a class and then splits them in each iteration process until the termination condition is met. The representative hierarchical clustering algorithms are BIRCH algorithm and CURE algorithm. However, the two methods are time-consuming and their clustering complexity is too high.

The density - based clustering methods consider a cluster as a high-density region in the data space that is separated by a low-density region. Its main idea is to continue clustering as long as the density of the neighboring area (the number of objects or data points) exceeds a certain threshold. That is, for each data point in a given class, there must be at least a certain number of points in a region with a given range. The methods can be used to filter "noise", isolate point data and find clusters of arbitrary shapes. Typical density-based clustering methods are DBSCAN and OPTICS. However, density-based clustering requires two parameters to be preset: the radius of adjacent threshold and the number of data points. As the clustering results are highly sensitive to the two parameters, improper parameter setting can lead to poor classification results.

The grid-based clustering methods quantify the object space into a finite number of units to form a grid structure, where all the clustering operations are performed. Represented by STING algorithm and CLIQUE algorithm, the grid-based clustering methods have fast processing speed and their processing time is independent of the number of data objects and is only related to the number of units in each dimension in the quantization space. However, the clustering effect of the methods is sensitive to the division of the grid.

The model - based clustering methods assume a model for each cluster, looking for data to best fit the given models. Model-based clustering algorithms may locate clustering by constructing a spatial density function that reflects

the spatial distribution of data points. It also automatically determines the number of clusters based on standard statistics, taking into account "noise" data or isolated points, resulting in robust clustering methods. Neural networks and decision trees are typical model-based clustering methods, which often have strict assumptions on data distribution. Therefore, model - based clustering methods have their own limitations.

In this paper, an improved clustering algorithm is proposed based on an analysis of the advantages and disadvantages of various clustering methods. On the one hand, the objective clustering analysis algorithm is applied to determine the number of clusters. On the other hand, the idea of density - based clustering methods is introduced to realize the clustering of classes of various shapes, breaking the limitation that division-based clustering is only suitable for spherical shape.

2. The basic principle of objective clustering algorithm (OCA)

One of the major problems to be solved in clustering is the number of classes that should be divided. One of the basic principles of clustering is to ensure that the similarity of the samples within the class is high and the gap between classes is obvious. At present, the algorithms that can automatically determine the number of clusters are grid-based clustering algorithm, density-based clustering algorithm and model-based clustering algorithm. However, as mentioned above, the clustering results of grid-based clustering methods and density-based clustering methods rely heavily on input parameters while the model-based clustering method is relatively strict on the assumption of data distribution. Partitioning-based clustering and hierarchical clustering often determine the number of clusters by establishing an objective function, for example, taking the average distance of the elements in each class to the various centroids as the objective function, and the number of clusters which makes the objective function reaches the minimum value is determined as the optimal number of clusters. However, in the case that the clustering point is not clear, the situation that the objective function gets smaller and smaller with the increase in the number of clusters tends to appear, resulting in a best clustering result that each sample becomes a class.

The objective clustering algorithm is a nonparametric clustering method proposed by Academician A. G. Ivakhnenko [11, 12] of the Ukrainian Academy of Sciences that can automatically and objectively determine the number of classification categories. The method uses two criteria for clustering: inner criteria are used to generate classes, and outer criteria (consistency criteria) are used to determine the optimal number of classes.

The basic principle firstly uses the "dipole" idea to classify the sample data into two corresponding subsets A and B, where clustering is carried out based on the distance between the sample points, respectively, so that each class has similarity. Then, by applying the idea that the nearest two points should belong

to the same category, whether the corresponding dipole samples in the two sets are assigned to the same class is taken as a criterion for testing the rationality of clustering in order to determine the optimal number of clusters.

The specific method is as follows:

Step 1: Calculation of the distance between all the data samples in the training set. The distance between the samples is sorted from small to large, with each distance corresponding to a dipole.

Step 2: Assuming that the sample size is n, take a number of dipoles that do not have a common sample with each other and put them into set A and B. Similarly, sets C and D are obtained from the remaining dipoles and taken as detection sets.

Step 3: The samples in sets A and B are numbered and the two samples from the same dipole use the same number in sets A and B to correspond to each other.

Step 4: for $i = 1 : r - 2$.

Step 5: Find the classification scheme with the highest value. If there are multiple schemes with the same vale, then Step 4 is repeated on detection sets C and D. Afterwards, which scheme has the highest value on C and D among the schemes that have the highest value on A and B is considered and the clustering of this scheme is the optimal clustering and the corresponding number of clusters is the optimal number of clusters.

Obviously, objective clustering analysis is essentially a hierarchical clustering algorithm, so there exists the ubiquitous defects that hierarchical clustering algorithm have. When the sample size is very large, the number of times needed for clustering is large, resulting in very low calculation efficiency. Moreover, the algorithm is only suitable for the clustering of spherical clusters and can not find clusters of arbitrary shapes. Therefore, this paper combines OCA with K-means and density-based clustering method to propose a new clustering algorithm.

3. Improved objective clustering algorithm

The improved objective clustering algorithm takes into account the fact that dividing the data into too many classes has actually lost the meaning of clustering, so it sets one parameter to be the maximum number of classes. The data is clustered using the K-means method. The criteria for determining the best number of classes refer to the OCA algorithm. The specific algorithm is as follows:

Step 1: The corresponding two sets A and B, as well as C and D, are generated using the methods mentioned in the objective clustering algorithm, and each set contains one sample.

Step 2: Let the number of clusters be an integer which is large enough.

Step 3: Select the initial clustering center.

The Euclidean distance between the samples in set A is calculated to represent the Euclidean distance between points and points. The average distance

between all points is calculated and taken as the neighborhood threshold. Assuming that the two most distant sample points in the set are E and F, put them into the initial cluster center set; the distance between the remaining points in the set and point E and F is tested and corresponding points are selected and denoted by G and put into the set; the distance between the remaining points in the set and point E, F and G is tested and corresponding points are selected and denoted by H and put into the set. This process goes on until the point is finally found out.

Step 4: Select the previous point in the set as the initial cluster center. The data in set A is clustered as a class using the K-means algorithm. The average distance of the samples in this class is calculated as the neighborhood threshold, and the points with the highest density in each class are found as the new clustering centers. If there is more than one point with the maximum density, it suggests that the class is of irregular shape and the points should be taken as the new clustering centers of the class. K-means clustering is re-conducted with the new clustering centers to re-determine which class the sample points belong to, and to gradually obtain the final clustering center through iteration. The final cluster centers of each class are used as the initial centers of set B to conduct K-means clustering and set B is also clustered as a class.

Step 5: command.

Step 6: command, if, execute Step 4.

Step 7: The corresponding classification scheme is the optimal clustering scheme. If there are multiple schemes with the same value, clustering is carried out on the sets C and D based on these schemes. The clustering of the scheme with the highest value on C and D is the optimal clustering and the corresponding number of clusters is the optimal number of clusters.

Step 8: The clustering center point of the optimal clustering scheme in Step7 is used as the initial clustering center, and the whole data set is clustered to obtain the final clustering result.

4. Numerical experiment

The improved algorithm proposed in this paper is applied to the clustering of two random data sets, which is compared to the OCA algorithm and K-mean algorithm. The number of clusters of the K-mean algorithm is set to three. The clustering results are shown in Fig. 1 and Fig. 2.

It can be seen from Figure 1 and Figure 2 that the improved-OCA algorithm inherits the advantages of density-based clustering algorithm for class contour recognition, and has obvious advantages on non-spherical data clustering compared with other partition-based algorithms. Meanwhile, it is not as much sensitive as the density-based algorithm to input parameters and its clustering accuracy and number of clusters are independent from parameter setting. For example, the clustering process of the above random data sets verifies that the obtained clustering results are exactly the same, whether the K value is set to be

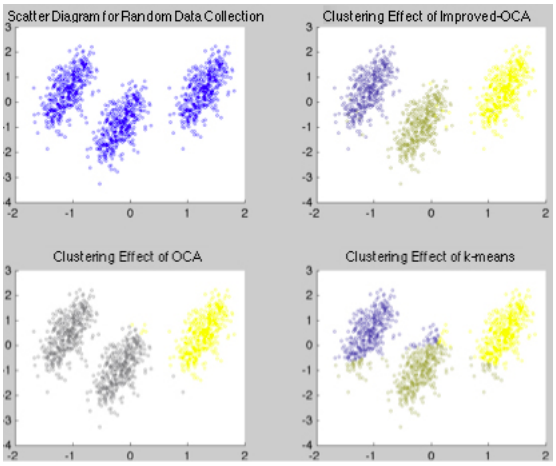

Figure 1. Comparison of clustering algorithms on random data set 1

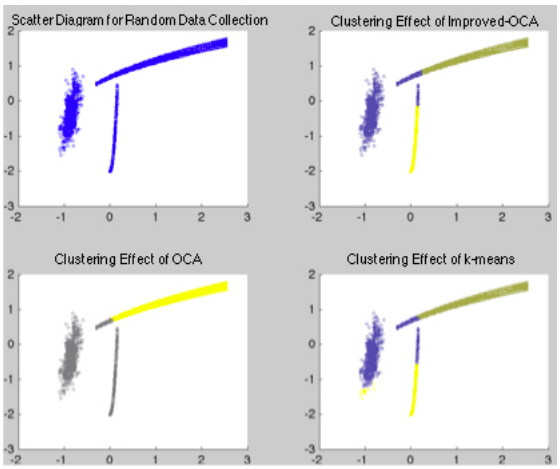

Figure 2. Comparison of clustering algorithms on random data set 2

20, 15 or 10. The improved-OCA algorithm automatically determines the number of categories and identifies the category outline. In addition, when the K value is set to be small, the number of operations is small, and the classification efficiency can be improved.

Among the judgments on the classes of the 1200 samples on data set 1 by the improved-OCA algorithm, there are only 16 wrong judgments, with an accuracy rate of 98.67%. While the accuracy rate of the judgment on the 1510 samples on data set 2 is 90.73%.

Then, the Iris dataset and the Breastcancerw dataset provided on the UCI dataset are clustered. The Iris dataset contains 150 sample points and is divided into three classes, each containing 50 sample points. Class 1 is completely separated from class 2 and class 3 while class 2 and class 3 are crossed. There-

fore, it is a reasonable division to divide the dataset into 2 or 3 classes. The improved-OCA algorithm classifies the first 50 sample points into one class and the remaining 100 sample points into another class, with a correct rate of 100diagnosis, contains 444 class 1 data and 239 class 2 data, which are correctly classified into 435 class 1 data and 168 class 2 data by the improved-OCA algorithm, with an accuracy rate of 88.29

The above random data set and UCI dataset clustering test proved the advantage of the improved-OCA algorithm in discovering clusters of arbitrary shapes relative to partitioning and hierarchical methods. Besides, the improved-OCA algorithm is not as sensitive as the density-based clustering method to input parameters. The only input parameter of the improved algorithm is the maximum number of classes permitted and the final clustering result does not depend on this parameter.

5. Conclusion

This study improved the OCA algorithm based on hierarchical clustering based on the analysis of the advantages and disadvantages of various algorithms by preserving its consistency criteria for determining the number of clustered categories, introduced the idea of density - based clustering algorithm, found out the point with the largest density in each class and took them as the centers of the classes to reconduct clustering so as to ensure that the classes are of arbitrary shapes, with no need of inputting parameters which may influence the clustering results. Finally, a test was carried out on the random data set and UCI data set, which proves the effectiveness of the improved-OCA algorithm proposed in this paper.

However, there are some shortcomings in the improved-OCA algorithm. It is essentially exhaustive K-means clustering in a small range to determine the value of the number of clusters, which is bound to affect the efficiency of clustering when the sample size is very large. How to judge the optimal number of clusters more quickly and efficiently is the problem which remains to be further studied.

References

[1] I.S. Dhillon, D.S. Modha, *Concept decompositions for large sparse text data using clustering*, Machine learning, 2001, 42(1-2), 143-175.

[2] A.K. Jain, M.N. Murty, P.J. Flynn, *Data clustering: a review*, ACM computing surveys (CSUR), 1999, 31(3), 264-323.

[3] C. Elkan, *Using the triangle inequality to accelerate k-means*, Fawcett T., Mishra N. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). Washington DC,USA: The AAAI Press, 2003, 3, 147-153.

[4] D. Arthur, S. Vassilvitskii, *k-means++: The advantages of careful seeding*, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Philadelphia, PA, USA: Association for Computing Machinery, 2007, 1027-1035.

[5] P.S. Bradley, U.M. Fayyad, *Refining Initial Points for K-Means Clustering*, Shavlik J. Proceedings of the 15th International Conference on Machine Learning (ICML98), San Francisco, USA: Morgan Kaufmann, 1998,98, 91-99.

[6] Zhao Yanchang, Song Mei, Xie Fan, et al., *Clustering Datasets Containing Clusters of Various Densities*, Journal of Beijing University of Posts and Telecommunications, 2003, 26(2), 42-47.

[7] L. Ertoz, M. Steinbach, V. Kumar, *Finding Clusters of Different Sizes, Sharps, Densities in Noisy, High Dimensional Data*, Proc of International Conference on Data Mining San Francisco, USA SIAM Press, 2003, 1-12.

[8] E. Martin, H.P. Kriegel, *A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, Proc of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 1996, 226-231.

[9] M. Patwary, D. Palsetia, A. Agrawal et al., *A new scalable parallel DBSCAN algorithm using the disjointset data structure*, High Performance Computing, Networking, Storage and Analysis Salt Lake City, Utah, USA, IEEE, 2012, 10-16

[10] T.N. Thanh, K. Drab, M. Daszykowski, *Revised DBSCAN algorithm to cluster data with dense adjacent clusters*, Chemo metrics and Intelligent Laboratory Systems, 2013, 120, 92-96.

[11] A.G. Ivakhnenko, *Heuristic self-organizing in problems of engineering cybernetics*, Automatic, 1967, 6, 207-219.

[12] A.G. Ivakhnenko, *The Review of Problems Solvable by Algorithms of the Group Method of Data Handling (GMDH)*, Pattern Recognition and Image Analysis, 1995, 5 (4), 527 - 535.

Accepted: 31.03.2017