DBSCAN Clustering Algorithm Based on Density

Dingsheng Deng *
Sichuan University Nationalities, Kangding, Sichuan, China
*Corresponding author's e-mail: dds0904@scun.edu.cn

2020 7th International Forum on Electrical Engineering and Automation (IFEEA) | 978-1-7281-9627-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/IFEEA51475.2020.00199

***Abstract*—Clustering technology has important applications in data mining, pattern recognition, machine learning and other fields. However, with the explosive growth of data, traditional clustering algorithm is more and more difficult to meet the needs of big data analysis. How to improve the traditional clustering algorithm and ensure the quality and efficiency of clustering under the background of big data has become an important research topic of artificial intelligence and big data processing. The density-based clustering algorithm can cluster arbitrarily shaped data sets in the case of unknown data distribution. DBSCAN is a classical density-based clustering algorithm, which is widely used for data clustering analysis due to its simple and efficient characteristics. The purpose of this paper is to study DBSCAN clustering algorithm based on density. This paper first introduces the concept of DBSCAN algorithm, and then carries out performance tests on DBSCAN algorithm in three different data sets. By analyzing the experimental results, it can be concluded that DBSCAN algorithm has higher homogeneity and diversity when it performs personalized clustering on data sets of non-uniform density with broad values and gradually sparse forwards. When the DBSCAN algorithm's neighborhood distance eps is 1000, 26 classes are generated after clustering.**

Keywords-DBSCAN Algorithm; Density Clustering; Machine Learning;Algorithm Research

I. INTRODUCTION

Since the middle of the 20th century, machine learning has been developing rapidly in theoretical research. However, due to the lack of hardware computing performance and data, it has not been widely applied [1-2]. With the rapid development and upgrading of economic and technology, electronic products gradually into everyone's life, only this kind of products has become a necessity of People's Daily life, in the process of the interaction of the people with all sorts of machine has been produced a large amount of data, paving the way for the widespread use of machine learning [3-4]. There are huge social and commercial values buried in the huge data generated all the time, and it has become the common goal of both academic and industrial circles to dig out more values from these data. Machine learning has gradually emerged in the commercial application of data mining, produced remarkable excellent results and has important commercial value, and has gradually become an important solution for data mining [5]. Clustering algorithm is an important technology in the field of machine learning. It is deeply used in a wide range of data mining scenarios, such as commodity recommendation, numerical prediction, pattern recognition and so on [6-7].

In our daily life, we often encounter this kind of situation: the marketer of the shopping mall will put the goods with the highest sales volume in the same place, in order to increase the possibility of the goods being bought at the same time; Housing sales need to understand the characteristics of people who need to buy houses; According to the characteristics of patients with the same disease, medical scholars will take measures to deal with the disease or propose a cure. Cluster analysis is needed in all of these cases. The explosion of stored and transient data has spurred a thirst for intelligent data-processing tools and cutting-edge technologies. New technologies and tools that keep pace with The Times can turn massive amounts of data into information and knowledge that supports decision making in an intelligent way. Therefore, with the rise of the concept of big data in recent years, data mining has also been applied in a wider range of fields.Clustering is one of the most commonly used algorithms in data mining. Clustering has been widely used in image processing, market research, pattern recognition and data analysis [8-9]. The purpose of using clustering algorithm is to discover clusters in data, which is a collection of data objects. The property of class is that objects in the same set are most similar to each other; Among them, density-based clustering algorithm is widely used for data clustering analysis due to its simple and efficient characteristics. It can be used to cluster arbitrarily shaped data sets.DBSCAN algorithm uses the given clustering radius and density threshold to randomly select core points to conduct clustering in the manner of neighborhood expansion.

This paper first introduces the concept of DBSCAN algorithm, and then carries out performance tests on DBSCAN algorithm in three different data sets. By analyzing the experimental results, it can be concluded that DBSCAN algorithm has higher homogeneity and diversity when it performs personalized clustering on data sets of non-uniform density with broad values and gradually sparse forwards. When the DBSCAN algorithm's neighborhood distance eps is 1000, 26 classes are generated after clustering.

II. DBSCAN CLUSTERING ALGORITHM

A. DBSCAN Clustering Algorithm

The behavior of grouping similar data objects in a collection into the same class is usually called clustering, and the collection of data objects is called Cluster E31. A collection of data objects is divided into a group, which can be considered a form of data compression. One of the characteristics of clustering is that it is unsupervised, which is different from classification, which requires high overhead in object modeling. Another characteristic of clustering is its adaptability to changes in data. Clustering can automatically discover and identify the sparse and dense areas of the object data set, and discover the potential correlation between the global distribution and data attributes. In the business field, cluster analysis helps market analysts to analyze and describe the

978-1-7281-9627-5/20/$31.00 ©2020 IEEE
DOI 10.1109/IFEEA51475.2020.00199

characteristics of customer groups, so as to gain an understanding of the consumption direction, consumption ability, consumption willingness and so on of different customer groups. Cluster analysis is also helpful for the analysis of geographical data. The analysis of rainfall, geology and other data is closely related to the timely prediction of natural disasters.

DBSCAN algorithm has good clustering results in the application, which is a typical representative of density algorithm. The algorithm has the characteristic of finding any shape class in the data set. The DBSCAN algorithm requires the user to set the global constant parameter neighborhood distance and threshold before running. The neighborhood distance sets the radius of the neighborhood range of the sample point. The threshold is set by changing the minimum number of sample points within the radius of the neighborhood range to be marked as a core point. It is worth noting that the neighborhood distance and threshold are both constants, which have been set before the program started and do not change after setting. DBSCAN algorithm is insensitive to noise points and can be applied to data sets containing noise points. It can identify noise points and exclude them from clustering results. For each class in the clustering result, the density within the class is higher than the density at the edge of the class. The density of the noise point is lower than that of the edge. According to the data distribution characteristics, the algorithm uses the density difference to identify different density regions, and marks the clustering results.

The algorithm steps are as follows:

Enter the minimum radius e and the minimum density threshold minp.

Sequential reading of data into a text file. The data is sequentially read into a text file that holds the original two-dimensional X and Y coordinates of the points, and stored into a pointList that holds the Point structure (which records information about the input points).

Determine if the point is a core point. Read a point from pointList (read in order), if the point is not marked (not part of a cluster), then calculate the distance between the point and all other points, if the distance between the two points is less than or equal to the minimum radius e, put the two points (eliminate the same points) into the tmpLst array, and count; If the distance between two points is greater than the minimum radius e, then skip this point and proceed to the next point. In the end, when the total number is greater than or equal to the minimum density threshold, the elements in tmpLst are marked as grouped, and the elements with excessive groups are put into the resultList array as a cluster (stored as an element). If the point is marked, the point is skipped and the judgment of the next point continues. Until all the points are traversed once.

Merge the clusters and merge the elements in the resultList. The clustering of the core points in the resultList is judged and compared. If you have the same element, combine the two clusters (the cluster where the core point is located) to form a new cluster, and do the same until no new clusters are produced.

Output the result of clustering and noise point.

B. Density Peak Point Discovery

The neighborhood of a data point. Given a minimum number of points (minPts), the neighborhood center of data point P is defined as a set of Np data points closest to P.

The local density of a data point. Given a minimum number of points (minPts), the local density ρ_p of data point P is defined as:

$$\rho_p = \min P_{ts} / \max_{x \in N_p} Dist(p,x) \tag{1}$$

Where, N_p is p's neighborhood contains $\min P_{ts}$ data points closest to P, and Dist() is the distance function. In order to introduce the algorithm in this paper, the distance between data points is Euclidean distance. According to formula (1), $\max_{x \in N_p} Dist(p,x)$ is the local radius of P (EPS). The larger the local density of P is, the smaller the local radius is, the more minimum points can be met. In order to measure the difference between peak density point and other core points, the difference metric δ_p of data point (P) is defined as:

$$\delta_p = \min_{\rho_x > \rho_p} Dist(p,x) \tag{2}$$

That is, δ_p is the minimum distance of all points Np with a local density greater than point P. When the local density of P is the largest, P is most likely to be a density core point, and then δ_p is defined as:

$$\delta_p = \max_{x \in D} Dist(p,x) \tag{3}$$

III. Experimental Design of DBSCAN Algorithm on Different Data Sets

A. Data Acquisition

In order to experiment the effect of DBSCAN algorithm, this paper simultaneously uses DBSCAN algorithm clustering on three data sets. These three data sets are D31 data set, R15 data set and credit card user data set respectively. D31 and R15 data sets are sample points obeying Gaussian distribution, and credit card user data sets are real consumption data sets of telecom customers. The characteristics of the data set are shown in Table 1.

TABLE 1. CHARACTERISTICS OF DATA SET

Data set name	Size	Number of classes	Dimension
D31	3100	31	2
R15	599	15	2
Credit card data set	300	-	2

B. *Experimental Environment*

Ubuntu 18.04 using Linux system has a CPU of 2.5ghz, a host memory of 16GB and a hard disk capacity of 500GB. The development language uses version 3.5 of Python, and the IDE development environment uses Pycharm.

C. *Performance Indicators*

The result class standard deviation and community difference are used in this paper: for a result class, if the standard deviation and community difference are small, it means that the result class is highly homogeneous. If the difference between standard deviation and community is small, it means that the data within the result class fluctuates greatly and the homogeneity is low.

IV. Experimental Results of DBSCAN Algorithm on Different Data Sets

A. *Experimental Results of D31 and R15 Data Sets*

All data sets were put into DBSCAN algorithm. After the clustering of the algorithm, the clustering results were counted. DBSCAN algorithm was used for experiments on D31 and R15 data sets. The generation of each class in D31 and R15 data sets is gaussian, and there are 31 categories annotated in the original data set of D31. The clustering results are shown in Table 2 and Figure 1.

TABLE 2. CLUSTERING RESULTS

	mean	std	min	median	max
Number of in-class sample points	24.5882	23.573	5	11	72
Standard deviation of in-class sample points	4.45443	2.8370	0.26323	3.64988	11.48042
Dimension 1 community difference	1.2477	2.8661	0.1173	0.6578	26.5855
Dimension 2 community difference	1.01498	0.8820	0.138	0.6325	6.0401

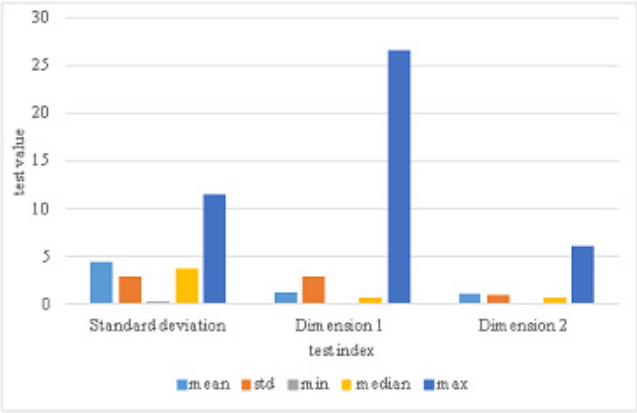

Figure 1. Clustering results

As can be seen from Figure 1, after the Eps parameter of DBSCAN algorithm is reduced, the result class after clustering can be increased, which increases the diversity to a certain extent. However, from the values of standard deviation and community difference within the result class, the difference within the result class is very large and unstable. It shows that the homogeneity within the class is not considered in the clustering process, and there is some randomness under the influence of initialization. DBSCAN algorithm increases the diversity of results by reducing Eps, resulting in the effect of large intra-class differences of result classes.

The range of eigenvalue range of R15 data set is smaller than that of D31 data set. It can be seen from the above table that the clustering results are 15 classes, which are the same as the big categories annotated with the original data, indicating that DBSCAN algorithm is excellent in the classification of big categories, but it cannot meet the controllable increase of diversity and the clustering process is uncontrollable.

B. *Experimental Results of Credit Card User Data Set*

After the clustering of the credit card user data set by DBSCAN algorithm and CEAV-DBSCAN algorithm, the result class statistics after the clustering of DBSCAN algorithm are shown in Table 3 and Figure 2.

TABLE 3. CLUSTERING RESULTS

	mean	std	min	median	max
Number of in-class sample points	36.5	68.54	0	7.5	176
Standard deviation of in-class sample points	65.68	158.27	0	1.15	388.71
Dimension 1 community difference	5.666	7.229	0	3.5	18
Dimension 2 community difference	5.333	8.477	0	0.6325	19

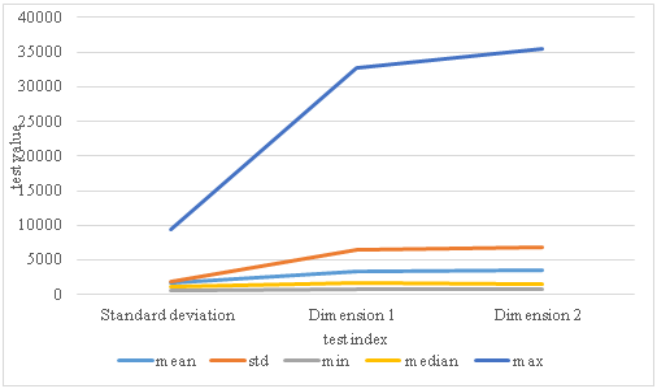

Figure 2. Result class statistics after DBSCAN algorithm clustering

When Eps take 10 neighborhood distance of DBSCAN algorithm, clustering algorithm to produce six classes, that is too little, and the results are at a neighborhood of 10 near axis zero gathered in data space, you can imagine, for characteristic value for more than 1000 samples, the length of 10 neighborhood distance is too short, hard clustering. Clustering can only be carried out on data below 1000. When the fixed

neighborhood is too small, effective clustering can only be carried out on the area close to the coordinate axis in the space, and the effective clustering cannot be carried out on the area of the principle coordinate axis.

Adjust neighborhood of DBSCAN algorithm distance Eps for 100 continue to experiment, after DBSCAN algorithm of clustering, our statistics class assessment data, when the neighborhood of DBSCAN algorithm from Eps in 100, 15 classes after clustering algorithm, the results of the number of classes is more than when Eps take 10, but the total is not much, the neighborhood of 100 results to data gathered in the middle of the space, you can imagine, for characteristic value for more than 10000 samples, the length of 100 neighborhood distance is too short, hard clustering, for eigenvalues of 1000 the following samples , the neighborhood distance with a length of 100 is a little large, especially for the samples with an eigenvalue less than 100, the neighborhood distance with a length of 100 completely covers the data space, and the data below and around 100 are classified into one category. When Eps is set at 100, the data can only be effectively clustered in the range of 1000 to 10000; when the neighborhood is fixed, the data can only be effectively clustered in a certain region of the space, and the data cannot be effectively clustered in other regions. After the clustering of DBSCAN algorithm, we counted the evaluation data of the result classes, as shown in Table 4.

TABLE 4. CLUSTERING RESULTS

	mean	std	min	median	max
Number of in-class sample points	25.8	70.48	5	6	280
Standard deviation of in-class sample points	301	354.03	0	160	1269
Dimension 1 community difference	132.666	185.94	0	92	731
Dimension 2 community difference	154.066	181.56	0	108	726

Eps do continue to neighborhood of DBSCAN algorithm distance adjustment, is set to 1000 continue to experiment, after DBSCAN algorithm of clustering, our statistics class assessment data, when the DBSCAN algorithm of Eps in 1000, the neighborhood distance clustering algorithm using the 26 classes, the results of the number of classes is more than when Eps from 10 to 100, but the total is not much, the neighborhood of 1000 results in the data gathered in the area of the space, you can imagine, for characteristic value for more than 100000 samples, the length of 1000 neighborhood distance is too short, hard clustering, right For samples with an eigenvalue less than 10000, the neighborhood distance with a length of 1000 is a little large, especially for samples with an eigenvalue less than 1000, the neighborhood distance with a length of 1000 completely covers the data space, and the data below 1000 and nearby are grouped into one class, which makes it impossible to conduct effective clustering for samples in this region. When Eps is 1000, effective clustering can only be carried out for data between 10000 and 100000.After the clustering of DBSCAN algorithm, we counted the evaluation data of the result classes, as shown in Table 5.

TABLE 5. CLUSTERING RESULTS

	mean	std	min	median	max
Number of in-class sample points	68	269	5	7.5	1378
Standard deviation of in-class sample points	1706	1816	513	1086	9382
Dimension 1 community difference	3369	6378	822	1757	32738
Dimension 2 community difference	33435	6884	773	1528	35538

V. CONCLUSION

In the era of big data, human electronic interaction is transformed into a series of data, which contains great value. Machine learning has shown excellent results in data mining, and has gradually become the main technology of data mining. However, the lack of labeling data in actual production makes unsupervised learning more adaptable. Clustering algorithm is an important technique in unsupervised learning. It is widely used in many scenarios, such as commodity recommendation and numerical prediction. However, in these scenes, the numerical range of data is very wide, and some of them have customized personalized services, which not only requires the clustering algorithm to be suitable for the non-uniform density data set with numerical vast density gradually sparse, but also needs to have diversified results and high homogeneity. Cluster analysis plays an extremely important role in data mining and can make a very important contribution in a large number of data analysis businesses. Nowadays, the data volume is increasing rapidly, so it is urgent to improve the efficiency and reliability of clustering algorithm in the stage of clustering analysis.

ACKNOWLEDGMENT

This paper was financially supported by the Key Project of Natural Science of Sichuan Provincial Education Department (No. 17ZA0295), 2017 Applied Demonstration Course Project of Sichuan University for Nationalities (No. sfkc201705) and Key Project of Natural Science of Sichuan University for Nationalities (No. XYZB19001ZA).

REFERENCES

[1] Li S S , An Improved DBSCAN Algorithm Based on the Neighbor Similarity and Fast Nearest Neighbor Query, IEEE Access, PP:99,2020

[2] Lee S , A Hybrid Framework using Fuzzy if-then rules for DBSCAN Algorithm, International journal of computational intelligence research, pp:403-412,2018

[3] Chen G , Cheng Y , Jing W , DBSCAN-PSM: an improvement method of DBSCAN algorithm on Spark, International Journal of High Performance Computing and Networking,pp:417, 2019

[4] Malik N , Supernova Type Ia Diversity: A Study using DBSCAN Algorithm, International Journal of Advanced Trends in Computer ence and Engineering,pp:3398-3402, 2020

[5] Kazemi-Beydokhti M , Ali Abbaspour R , Mojarab M, Spatio-Temporal Modeling of Seismic Provinces of Iran Using DBSCAN Algorithm,Pure and Applied Geophysics, pp:1937-1952, 2017

[6] Zhang, Wang, Liang,Short-Term Wind Power Prediction Using GA-BP Neural Network Based on DBSCAN Algorithm Outlier Identification,Processes, 8(2):157-, 2020

[7] Zhang H , Liu P , Guo Y, Blind modulation format identification using the DBSCAN algorithm for continuous-variable quantum key distribution, Journal of the Optical Society of America B, 36(3):B51,2019

[8] Zhang T , Ma F , Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function, International Journal of Computer Mathematics, pp:663-675,2017

[9] Memon K H , Lee D H , Generalised fuzzy c-means clustering algorithm with local information, Fuzzy Sets & Systems, pp:1-12,2018