Improvement of DBSCAN Algorithm Based on Adaptive Eps Parameter Estimation

Lu Zhu
School of Information
Yunnan University
Kunming, China
zhulu001@foxmail.com

Jiang Zhu
School of Information
Yunnan University
Kunming, China
zhujiang616@foxmail.com

Chongming Bao
School of Information
Yunnan University
Kunming, China
chmbao@ynu.edu.cn

Lihua Zhou
School of Information
Yunnan University
Kunming, China
lhzhou@ynu.edu.cn

Chongyun Wang
School of Information
Yunnan University
Kunming, China
cywang@ynu.edu.cn

Bing Kong
School of Information
Yunnan University
Kunming, China
kongbing@ynu.edu.cn

ABSTRACT

DBSCAN algorithm is a density-based clustering algorithm, which is widely used in various fields and can identify clusters of arbitrary shape. Aiming at the existing problems that the selection of neighborhood radius parameter Eps depends on manual experiment, judgment intervention or calculation estimation, the accuracy of clustering results is sensitive to Eps value, using uniform global parameters in non-uniform density data sets makes the clustering effect unsatisfactory. An adaptive Eps parameter estimation method based on Gauss kernel density is proposed. The kernel probability density of the point is calculated by the Gauss kernel density estimation, and then the positive correlation between the estimated kernel probability density and Eps value is found. It adaptively matches the appropriate Eps value for the current point, rather than setting a globally unified Eps parameter. Theoretical analysis and experimental results show that the method has good clustering effect on non-uniform data sets, and the algorithm is also effective for uniform data sets.

CCS CONCEPTS

• Theory of computation~ Parameterized complexity and exact algorithms • Computing methodologies~ Artificial intelligence • Computing methodologies~ Machine learning

KEYWORDS

DBSCAN, Eps, Gauss Kernel Density Estimation, Parameter Selection

ACAI '18, December 21–23, 2018, Sanya, China
© 2018 Association for Computing Machinery.ACM ISBN 978-1-4503-6625-0/18/12…$15.00
https://doi.org/10.1145/3302425.3302493

ACM Reference format:

Lu Zhu, Jiang Zhu, Chongming Bao, Lihua Zhou, Chongyun Wang and Bing Kong. 2018. Improvement of DBSCAN Algorithm Based on Adaptive Eps Parameter Estimation. In *Proceedings of 2018 International Conference on Algorithms, Computing and Artificial Intelligence (ACAI'18).* *Sanya, China, 7 pages.* https://doi.org/10.1145/3302425.3302493

1 Introduction

Cluster analysis [1] is widely used in mathematics, statistics, network, medicine and commerce. It can reorganize data without knowing the exact distribution structure and domain information, and classify data based on similarity. Compared with other clustering methods, density-based clustering algorithm is a very important method. It can find isolated data points or clusters of various shapes and sizes in noisy data. The typical representative of this kind of algorithm is Density-based spatial clustering of applications with noise (DBSCAN) [2]. As one of the most popular clustering algorithms in scientific literature, machine learning and data mining, it has attracted wide attention. In 2014, the algorithm was awarded the test of time award (an award given to algorithms which have received substantial attention in theory and practice) at the leading data mining conference, KDD [3].

DBSCAN algorithm needs to input two important parameters, Eps and minPts, the former is the neighborhood radius when defining density, and the latter is the threshold when defining core points (minimum number of points required to form a cluster). The clustering effect of DBSCAN algorithm is closely related to the choice of Eps and minPts. Usually, the setting of these two parameters depends on experiments, observations and judgments [4]. In the case of fixed minPts, if the selected Eps value is too small, a large number of objects will be marked as noise, and the natural cluster may be split into several clusters. If the value of Eps value is too large, a lot of noise will be added to the cluster, and some natural clusters that should be separated will also be merged into one cluster. At present, clustering algorithm is applied in various fields, and the density distribution of all kinds of data is different. Using globally unified Eps parameters can`t produce better clustering effect when the density of data isn't uniform [5]. Many literatures have proposed optimization and improvement methods for parameter selection of DBSCAN algorithm in non-uniform density datasets. Literature [2] is the traditional DBSCAN algorithm. It sets minPts=4 according to experience, and parameters Eps get more appropriate values by taking multiple values. This method relies too much on human experience observations, and the clustering results depend on human subjective judgment, but this observations is not very good in non-uniform density data sets clustering. Literature [6] still sets minPts = 4, while sorting the data objects in the data set, Eps is the corresponding value of the $\delta^2/(4c)C_n^2$ object. Although this method can obtain a more appropriate Eps value, it also introduces another parameter, the number of clusters c, which is generally difficult to predict. Literature [7], the record of cluster connection information is proposed, which can correct the problem of poor clustering results caused by improper selection of input parameters and reduce the sensitivity of DBSCAN to input parameters. Literature [8] analyses the eigenvalues of sample data by the kernel density estimation method in nonparametric theory. This method can automatically determine the clustering parameters Eps. Although it can determine a more appropriate value of Eps, the clustering results are not good enough when the density of data set is not uniform. Literature [9] aims at the problems of DBSCAN algorithm in Eps parameter and processing efficiency of massive data sets, a new algorithm OPDBSCAN is proposed. The algorithm divides data sets into multiple overlapping partitions, and obtains local Eps as partition Eps through k-dist graph of each partition, and then uses MapReduce for parallel clustering. Although this method reduces the impact of global Eps parameters, it still needs to specify the number of partitions.

In view of the unique advantages of DBSCAN and its wide application requirements, in order to avoid the problem caused by using global Eps parameters when the density of data sets is not uniform, and when Eps is too large, clusters with smaller density will be misclassified into clusters with larger density; when Eps is too small, one cluster will be misclassified into multiple clusters. In this paper, an improved strategy is proposed to estimate the Eps value of each data point by calculating the estimated value of the nuclear probability density of the data sample, and according to the positive correlation between the estimated probability value of the nuclear density and the Eps value. The algorithm automatically matches the probabilistic Eps radius of the current point in the process of searching the neighborhood of the point. When the density of a sample is large, the point may be the center point of a cluster or the point near the center point, then matching the larger Eps, the points of the same cluster can be included in the same cluster as much as possible; When the density of a sample point is small, the point may be the edge point or discrete point of a cluster, and the smaller matching Eps can reduce its impact on other clusters. Experiments show that the proposed algorithm can be better applied to non-uniform density data sets, and the clustering results have better performance.

2 Related works

In order to introduce the improved algorithm proposed in this paper, the related knowledge of the kernel density estimation and DBSCAN algorithm used in the improved algorithm is briefly introduced.

2.1 Definition and Theory of Kernel Density Estimation

One of the basic problems of probability statistics is to solve the distribution density function of random variables from a given set of samples. The methods to solve this problem include parameter estimation and non-parameter estimation. Rosenblatt (1955) and Parzen (1962) proposed a non-parametric estimation method, namely the kernel density estimation method [10]. The kernel density estimation method does not use the prior knowledge of data distribution and does not add any assumptions to the data distribution. It is a method to study the characteristics of data distribution from the data sample itself.

Definition 1. (Nuclear Density Estimation) Assuming data (x_1, $x_2 \ldots\ldots x_n$) is n sample points and its probability density function is f, The formula of probability density function is as follows:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad (1)$$

K is a kernel function (non-negative, integral is 1, accord with probability density property, and mean value is 0), h > 0 is a smooth parameter, called bandwidth.

Definition 2. (Kernel function) kernel function is the inner product of mapping relationship. The mapping function itself is only a mapping relationship and does not increase the dimension characteristics. However, we can use the characteristics of Kernel function to construct a kernel function that can increase the dimension. In theory, all smooth peak functions can be used as the kernel functions of KDE, as long as the area under the curve of normalized KDE is equal to 1 (the probability value of data points on the graph).

Definition 3. (Gauss Kernel) Gauss Kernel, also known as Radial Basis Function (RBF), is a commonly used kernel function. When choosing the kernel function, if there is no prior knowledge of the given data, the RBF kernel function is a better choice. Literature [11] establishes the relationship between the kernel function K and the regularization operator P to examine the generalization ability of some kernels, and illustrates that the RBF kernels can obtain very smooth estimates, often with good performance. Another advantage of RBF core is that its range of core values is (0,1), which makes the calculation process simple. Because of the convenient mathematical properties of the Gauss kernel, this paper uses the Gauss kernel function formula as follows:

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right) \qquad (2)$$

Definition 4. (Multidimensional Gauss Kernel Density Function) [12] Assuming data $x_1, x_2 \ldots\ldots x_n$ is n sample points, where x_i is d-dimensional data $(y_1, y_2, \ldots\ldots y_d)$, then its nuclear density estimation is defined as:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_{d,h}(x - x_i) \tag{3}$$

Then:

$$K_{d,h}(X) = \prod_{i=1}^{d} K\left(\frac{x - x_i}{h}\right) \tag{4}$$

2.2 Basic Definition of DBSCAN Algorithms

DBSCAN is based on a set of neighborhoods to describe the compactness of the sample set. The parameters (Eps, minPts) are used to describe the compactness of the sample distribution in the neighborhood. Eps describes the neighborhood distance threshold of a sample, minPts describes the threshold of the number of samples in the neighborhood where the distance of a sample is Eps. DBSCAN is mainly defined as follows [13]:

Definition 5. (Eps Neighborhood) The region of a given object point P with Eps as its radius is called the Eps Neighborhood of the object.

Definition 6. (Core Object) If the number of samples in the Eps neighborhood is greater than or equal to minPts, the object is called Core Object.

3 Research on Adaptive Eps Parameter Estimation Method Based on Kernel Density

3.1 Basic Thought

DBSCAN algorithm searches in the domain by selecting the appropriate Eps as the neighborhood radius. If there are more points than minPts, it can form clusters. The kernel probability density of each sample point in the data set can be obtained by estimating the kernel density. Inspired by literature [4] [5], we can conclude that the larger the sample density, the more likely it is to be the center of a cluster, and the larger the matched Eps is the neighborhood search radius, the more points of the same cluster can be included in the same cluster centered on that point. When the density of a sample point is small, the point may be the edge point or discrete point of a cluster, and the smaller matching Eps as the neighborhood radius can reduce its impact on other clusters. Because kernel density estimation does not make use of prior knowledge of the distribution of data sets, it can be used as a reference for selecting Eps for each point, and then clustering.

Because the accuracy of the algorithm is related to the setting of kernel density function and minPts parameters, in order to make the estimation probability of kernel density more accurate and reduce the error, the bandwidth of the kernel function and minPts are processed to get a more accurate result in this paper.

3.2 Bandwidth Selection of Kernel Functions

There are many factors affecting the estimation of kernel density, which are not only related to the attributes of data sets, but also to the kernel functions used in their selection. he choice of the bandwidth of the kernel function is directly related to the accuracy of the kernel density estimation[14]. Bandwidth is larger, probability distribution curve is flat, on the contrary, the probability distribution curve is steeper if the bandwidth is smaller. In order to reduce the estimation accuracy caused by h. Because the Epanechnikov[15] kernel is optimal in the sense of mean square error, the efficiency loss is also small. Therefore, this paper uses the mean integrated square error to measure the quality of h [16]. The formula is:

$$\text{MISE}(h) = E\int\left(\hat{f}(x) - f(x)\right)^2 dx \tag{5}$$

$$\text{AMISE}(h) = \frac{R(K)}{nh} + \frac{1}{4} m_2(K)^2 h^4 R(f'') \tag{6}$$

Then:

$$R(K) = \int K(x)^2\, dx \text{ ,} \ m_2(K) = \int x^2 K(x) dx \tag{7}$$

In order to minimize MISE (h), it is transformed into a problem of finding poles. Minimizing MISE (h) is equivalent to minimizing AMISE (h) [17]， For derivation, let the derivative be 0:

$$\frac{\partial}{\partial h}\text{AMISE}(h) = -\frac{R(k)}{nh^2} + m_2(K)^2 h^3 R(f'') = 0 \tag{8}$$

$$hAMISE = \frac{R(K)^{\frac{1}{5}}}{m_2(K)^{\frac{2}{5}} R(f'')^{\frac{1}{5}} n^{\frac{1}{5}}} \tag{9}$$

When the kernel function is determined, R、m、f'' in formula (6) can be determined, then:

$$h\text{AMISE} \sim n^{-\frac{1}{5}}, \ \text{AMISE}(h) = O\left(n^{-\frac{4}{5}}\right) \tag{10}$$

If the Gaussian kernel function is used for estimating the kernel density, the optimal selection of h (the bandwidth that minimizes the mean integral square error) is as follows:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}} \tag{11}$$

Here $\hat{\sigma}$ is the standard deviation of the data sample.

3.3 MinPts Selection

For the selection of minPts, minPts is the minimum expected cluster size, that is a cluster consists of at least minPts points. As a rule of thumb, the minimum value can be derived from the dimension D of the data set, such as $\text{minPts} \geq \text{D} + 1.\text{minPts} = 1$, which doesn't make sense because every point in itself is already a cluster. When $\text{minPts} \leq 2$， the clustering result is the same as that of hierarchical clustering using single chain metric. Therefore, at least three must be chosen. However, larger values are usually better for noisy data sets and produce more significant clusters. According to the rule of thumb [18]， this paper chooses $\text{minPts} = 2 \times \text{dim}$.

3.4 Eps Adaptive Selection

Through formulas(1),(2)and (3)we can get the probability density of each sample point $\hat{f}_h(x)$, through the definition of probability density, we can think that the larger the probability density, the larger the $\text{Eps}_{(xi)}$ can be given as the neighborhood radius; Points with smaller probability density can be endowed with smaller radius of $\text{Eps}_{(xi)}$. It can also be considered that there is a positive correlation between $\hat{f}_i(x)$ and $\text{Eps}_{(xi)}$.This paper sets a coefficient &, then:

$$\&*\hat{f}_i(x) = \text{Eps}_{(xi)} \quad (11)$$

In the process of multi-dimensional data clustering, DBSCAN algorithm uses $\text{Eps}_{(xi)}$ to form multi-dimensional hyperspheres as search neighborhood. For the relationship between probability density of Gauss distribution in different dimensions D and radius r, in high-dimensional space, most of the probability mass of Gauss distribution lies on a thin spherical shell with a certain radius. The volume of the multidimensional hypersphere is:

$$V_d = \begin{cases} \dfrac{\pi^{\frac{d}{2}}}{\left(\frac{d}{2}\right)!} R^d, \text{ d is even} \\ \dfrac{2^d \pi^{\left(\frac{d-1}{2}\right)}\left(\frac{d-1}{2}\right)!}{d!} R^d, d \text{ is } odd \end{cases} \quad (12)$$

The unit hypersphere volume formula is as follows:

$$V_d = \begin{cases} \dfrac{\pi^{\frac{d}{2}}}{\left(\frac{d}{2}\right)!}, \text{ d is even} \\ \dfrac{2^d \pi^{\left(\frac{d-1}{2}\right)}\left(\frac{d-1}{2}\right)!}{d!}, d \text{ is } odd \end{cases} \quad (13)$$

In the clustering process, we can approximate that the unit volume of a single point is approximately equal to the volume of a hypersphere in the Eps radius of the point divided by the number of points included:

$$V_d \approx \frac{V_{eps}}{2*dim} \quad (14)$$

Total volume can be considered:

$$n*V_d = \sum_{i=1}^{n} \frac{V_{eps}}{2*dim}, \ \mathbf{R} = \&*\hat{f}_i(x) \quad (15)$$

The coefficients & obtained are as follows:

$$\& = \left[\frac{2n*dim}{\sum_{i=1}^{n}\left(\hat{f}_i(x)^d\right)}\right]^{\frac{1}{d}} \quad (16)$$

Then:

$$Eps_{(x_i)} = \&*\hat{f}_i(x)^d \quad (17)$$

3.5 Improved Pseudo-code of DBSCAN Algorithm

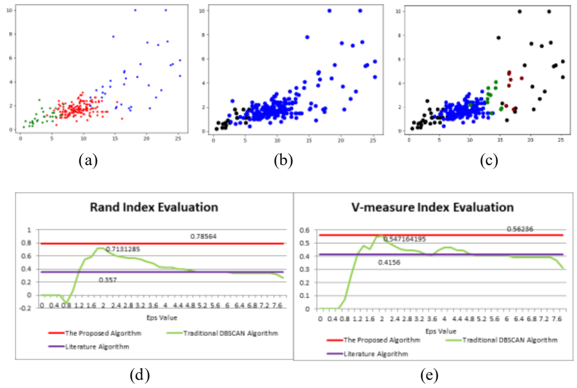
```
Algorithm: Improved Pseudo-code of DBSCAN Algorithm
Input:
    D: A Data Set Containing n Objects
    minPts: Neighborhood density threshold
Output: A set of density-based clusters
Method:
    Mark all objects as unvisited;
    Do
    Random selection of one unvisited object p;
    Marked as visited;
    If There are at least minPts objects in P domain:
        Create a new cluster C and add p to C.
        Let N be the set of objects in the $Eps_{(p)}$ neighborhood of p;
        For every point in the N:
            If p′ is unvisited then;
                Marked p′ is visited;
            If There are at least minPts objects in the $Eps_{(p′)}$domain of p' ,then add these points to N;
            If p′ is not yet a member of any cluster, then add p′ to C;
       End for
       Output C;
    Else marked p is noise;
    Accessing objects that are not marked as unvisited;
```

4 Experimental Results and Analysis

On the basis of several data sets, we compare the traditional DBSCAN algorithm (make Eps value float to get better results), DBSCAN algorithm in literature [8]. and improved DBSCAN algorithm in this paper. For the convenience of representation, the algorithm of literature [8] is uniformly expressed as the algorithm of literature. In order to verify the effectiveness and feasibility of the proposed algorithm, the non-uniform density data set DS1, special hypercube data set DS2, real data set Iris, and multi-dimensional Glass real data set are clustered. When judging the accuracy of the algorithm, the following two evaluation indices are used to evaluate the accuracy of the algorithm. Adjusted Rand index [19], the range of values is [−1, 1], the larger the value, the more consistent the clustering results are with the real situation. V-measure [20], Qualitative analysis can be made in terms of homogeneity and integrity, from 0 to 1 to reflect the worst to the best performance. Because it is not easy to plot in multi-dimensional space, this paper only makes two-dimensional scatter plots to assist in illustration.

4.1 DS1 Data Set

In order to verify the effectiveness of the proposed algorithm on non-uniform density datasets, the DS1 dataset selected in Experiment 1 is a non-uniform density dataset with 215 data samples, which contains four dimensions and is divided into three clusters. Figure 1 (a) shows the original cluster partition of DS1 dataset, (b) shows the two-dimensional rendering of clustering results of literature algorithms, and (c) shows the clustering results of the proposed algorithm. (d) and (e) represent the Rand and V-measure indices to evaluate the effectiveness of the three algorithms.



(a) (b) (c)

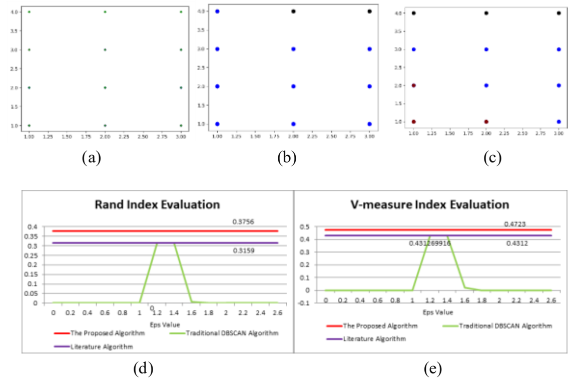
(d) (e)

Figure 1: DS1 Clustering Results

From Figure 1， the algorithm in this paper has better results than the traditional algorithm and the literature algorithm, and the accuracy is higher, and the algorithm in this paper is even better than the best effect of the traditional DBSCAN algorithm. From the graphs (a), (b) and (c) , it can be seen that although there are many noise points in this algorithm, three clusters are correctly divided, while only one cluster is partitioned in the literature algorithm. This is because the Eps value of the literature algorithm is large, and as a global unified parameter, it can`t divide the distance between clusters very well. Most of the other two clusters are also included. This shows that the global uniform parameters are sensitive to cluster partition in non-uniform data sets, and that the algorithm in this paper is more accurate in non-uniform density data sets. From the graph (d) and (e), the clustering effect of the proposed algorithm is better than that of the traditional algorithm and the literature algorithm, and is higher than the best effect of the traditional DBSCAN algorithm.

4.2 DS2 Data Set

In order to verify the effectiveness of the proposed algorithm on uniform density data sets, the DS2 data set selected in Experiment 2 is a special structured data set consisting of vertices of a 4-dimensional hypercube with 132 sample points. It is a uniform density data set, and the original data is divided into three clusters.



(a) (b) (c)

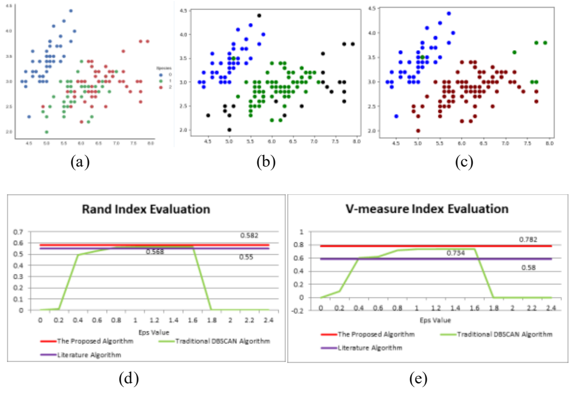
(d) (e)

Figure 2: DS2 Clustering Results

Figure 2 (a) shows the original cluster partition of DS2 dataset, (b) shows the two-dimensional rendering of clustering results of literature algorithms, and (c) shows the clustering results of the proposed algorithm.(d) and (e) represent the Rand and V-measure indices to evaluate the effectiveness of the three algorithms. From the results, the literature algorithm divides the data set into two clusters, and the algorithm in this paper divides into three clusters. The clustering of uniform data sets is better than that of literature. From the point of view of graph (d) and (e), the accuracy of this algorithm and literature algorithm is better than that of traditional DBSCAN algorithm, and this algorithm is better than the optimal value of traditional DBSCAN algorithm. It shows that the proposed algorithm is also effective on uniform data sets.

4.3 Iris Real Data Set

In order to verify the validity on real data sets, Iris data sets are selected as clustering objects. Iris data set is a common experimental data set of classification and clustering. It is a data set of multiple variable analysis. The data set contains 150 data sets, which are divided into three categories, 50 data in each category, and 4 attributes in each data set.



(a) (b) (c)

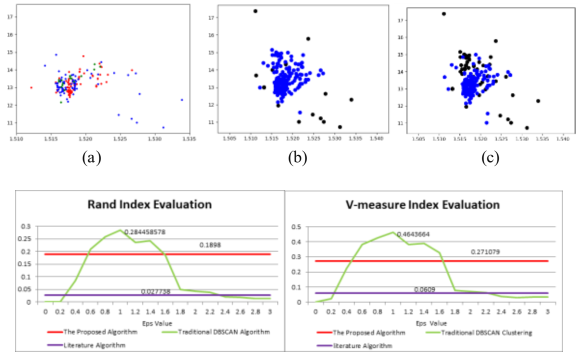
(d) (e)

Figure 3: Iris Clustering Results

From the Iirs clustering results in Figure 3 (a), (b) and (c), the proposed algorithm is accurately divided into three clusters, and the literature algorithm is divided into two clusters. From the graph (d) and (e), the accuracy of the proposed algorithm is also higher than the traditional DBSCAN algorithm and literature algorithm, and better than the traditional DBSCAN optimal value. Because some data of Iris dataset also have the characteristics of non-uniform density, this experiment not only shows the effectiveness of the proposed algorithm for non-uniform density dataset, but also shows that the proposed algorithm can achieve better results in real dataset clustering.

4.4 Glass Data Set

In order to verify the effectiveness of the proposed algorithm on multi-dimensional datasets, the Glass dataset selected in Experiment 4 is a real dataset with 9 features, which is divided into 6 clusters.



(a) (b) (c)

(d) (e)

Figure 4: Glass **Clustering Results**

From the experimental results of Figure 4（a）,(b) and (c), the proposed algorithm and the literature algorithm divide the data set into a cluster. The rest points are marked as noise points , which differs greatly from the original partition of the data set. From (d) and (e), although the accuracy of the algorithm is not good, it is close to the optimal value of the traditional DBSCAN algorithm, and better than the literature algorithm. Generally speaking, the results of three kinds of clustering are not ideal, and the clustering accuracy is relatively low. This is caused by the data characteristics of Glass dataset, because there is dimensionality disaster in high-dimensional data [21]. (Because in high-dimensional space, all data are very sparse, and the volume index increases, the distance difference between objects becomes more insignificant, which leads to great deviation in similarity measurement and distance calculation.).This has a great impact on DBSCAN algorithm and its improved algorithm based on density clustering defined by spatial distance of points.

By comparing the four experiments mentioned above, because the non-parametric density estimation theory is adopted in the proposed algorithm, and the global uniform parameters are not used, the neighborhood radius Eps of the current point is given in the process of neighborhood search. Compared with the traditional DBSCAN clustering method which requires prior knowledge to set parameters of unknown data sets and the literature algorithm which adopts global uniform parameters, the proposed algorithm can often get better clustering results. Moreover, the proposed algorithm completely relies on the characteristics of the data set itself in the clustering process, does not need human intervention, can adapt to a variety of forms of data sets, and can achieve adaptive clustering. Although the proposed algorithm is not better than the optimal value of traditional DBSCAN in multi-dimensional data sets, it is also close to the optimal value of traditional algorithm. In practical clustering, the traditional algorithm often can't select the appropriate parameters and can't get the optimal value, so the results of the algorithm in this paper are completely acceptable to approach the optimal value.

5 Conclusions and Future Work

The main purpose of this paper is to improve the DBSCAN algorithm on the self-adaptive selection of parameters Eps in clustering process. The improved method is to find the positive correlation between the estimated kernel probability density and the domain search radius Eps, and to match the neighborhood radius of the current search point adaptively. Thus, the DBSCAN method for setting global uniform parameters is improved. In this process, self-adapting clustering is carried out by the characteristics and properties of the data set itself, and better clustering results can be obtained without artificial determination of input parameters. Of course, this paper trades for clustering validity by increasing the computational complexity of the algorithm. For processing large data, the cost is high, and this paper does not study the selection of another parameter minPts. Therefore, how to determine minPts parameters adaptively, increase the accuracy of high-dimensional data clustering and reduce the computational complexity will be the next research direction.

ACKNOWLEDGMENTS

The authors thank the reviewers for their helpful comments and suggestions. This research was supported by the National Natural Science Foundation of China (61762090, 31760152), the Natural Science Foundation of Yunnan Province (2016FA026).

REFERENCES

[1] Zaki, M. (2010). Introduction to Data Mining. CHINA MACHINE PRESS.

[2] Ester, M., Kriegel, H. P., & Xu, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. International Conference on Knowledge Discovery and Data Mining (pp.226-231). AAAI Press.

[3] Martin Ester,(2014)."2014 SIGKDD Test of Time Award". ACM SIGKDD. 2014-08-18. https://www.kdd.org/News/view/2014-sigkdd-test-of-time-award

[4] Wang, W. T., Wu, Y. L. , Tang, C. Y. , & Hor, M. K. . (2015). Adaptive density-based spatial clustering of applications with noise (DBSCAN)

according to data. International Conference on Machine Learning & Cybernetics. IEEE.

[5] Yang, L. I., Li, M. A. , & Suohai, F. . (2016). Improved dbscan clustering algorithm based on dynamic neighbor. Computer Engineering & Applications.

[6] Shi-Hong, Y., Ping, L. I. , Ji-Dong, G. , & Shui-Geng, Z. . (2005). A statistical information-based clustering approach in distance space. Journal of Zhejiang University Science A(Science in Engineering), 6(1), 71-78.

[7] Yingkun Cai, Kunqing Xie, Xiujun Ma. (2004)An improved DBSCAN algorithm that shields the sensitivity of input parameters [J]. Journal of Peking University (Natural Science Edition) ,480-486.

[8] Zonglin Li, Ke Luo. (2016). Adaptive determination of parameters in Dbscan algorithm. Computer Engineering and Application, 52 (3), 70-73.

[9] He, Y., Tan, H., Luo, W., Mao, H., Ma, D., & Feng, S., et al. (2012). MR-DBSCAN: An Efficient Parallel Density-based Clustering Algorithm using MapReduce. IEEE, International Conference on Parallel and Distributed Systems (Vol.42, pp.473-480). IEEE.

[10] Parzen, E. . (1962). On estimation of a probability density function and mode. Ann. Math. Statis., 33(3), 1065-1076.

[11] Wang, S., Wang, J., & Chung, F. L. (2013). Kernel density estimation, kernel methods, and fast learning in large data sets. IEEE Transactions on Cybernetics, 44(1), 1-20.

[12] Kristan, M., Leonardis, A., & Skočaj, D. (2011). Multivariate online kernel density estimation with gaussian kernels. Pattern Recognition, 44(10), 2630-2642.

[13] Hinneburg, A. , & Keim, D. A. . (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. International Conference on Knowledge Discovery & Data Mining. AAAI Press.

[14] Xiaowei Xu. (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. Acm Transactions on Database Systems, 42(3), 19.

[15] Epanechnikov, V.A. (1969). "Non-parametric estimation of a multivariate probability density". Theory of Probability and its Applications. 14: 153–158

[16] Rosenblatt, M. . (1956). Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics, 27(3), 832-837.

[17] Wang, Y., Feng, C., Liu, Y., Zhao, Y., Li, H., & Zhao, T., et al. (2017). Comparative study of species sensitivity distributions based on non-parametric kernel density estimation for some transition metals . Environmental Pollution, 221, 343-350.

[18] Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: the algorithm gdbscan and its applications. Data Mining & Knowledge Discovery, 2(2), 169-194.

[19] Robert, V., & Vasseur, Y. (2017). Comparing high dimensional partitions, with the coclustering adjusted rand index.

[20] Xu, Z. . (2010). Fuzzy harmonic mean operators. International Journal of Intelligent Systems, 24(2), 152-172.

[21] Zhang, J., Zheng, X., & Li, L. (2017). Pdp: parallel dynamic programming. IEEE/CAA Journal of Automatica Sinica, 4(1), 1-5.