Send Orders for Reprints to reprints@benthamscience.ae

Research on Improve DBSCAN Algorithm Based On Ant Clustering

Fang Yuankang[1,2,*], Huang Zhiqiu[1], Luo Yuping[2], Ye Zan[2] and Liu Ying[2]

[1]*Information Science and Technology School in Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu Province, 210016*

[2]*Computer Department in Chizhou College, Chizhou Anhui Province, 247000*

Abstract: DBSCAN algorithm is sensitive to the input parameter of Eps, especially when the data density is non-uniform. It gets poor result in clustering using the same global Eps. In addition, the algorithm has difficulty with high-dimension of data. In this paper, an improved DBSCAN algorithm LF-DBSCAN is proposed, which uses ant clustering algorithm in data preprocessing phase to classify the datasets and to get several values of parameter Eps, then call DBSCAN algorithm with different values of Eps to cluster the non-uniform datasets. Experimental results demonstrate the effectiveness of the improved algorithm.

Keywords: Ant clustering algorithm, A set of values of Eps, Ant Clustering, DBSCAN, LF-DBSCAN, similarity.

1. INTRODUCTION

Clustering is an important way to find or discover knowledge; it generally can be divided into five categories: Based on division, level, density, model and grid methods. DBSCAN Algorithm is a classical method based on density, it can find clusters of arbitrary shape, and does not need to lay down a pre-determined number of clustered dataset which will be clustered, and is also unaffected by noise. However, when the density of data set is uneven, the use of a unified global parameter Eps cause poor clustering. For the shortcomings of the DBSCAN, many scholars have performed a lot of research and study for further improvements. To simplify the input of parameters, Li Xia etc. [1] proposed a new definition method of dynamic density, and gave a dynamic density clustering algorithm DDBCA on the basis of this definition. The algorithm only needs a single input parameter (the number of nearest neighbors' k) and then will be able to dynamically identify clusters of uneven density clustering. Literature [2] proposes a new algorithm (PACA-DBSCAN) using the improved ant clustering algorithm and DBSCAN, this algorithm can reduce the impact of inputting parameters on the quality of clustering results, and can effectively deal with datasets of different densities. Yang Jing etc. [3] referred to the idea of data field, introduced the concept of the average potential difference and dynamically determines Eps, it avoids the disadvantages of Eps by manually determining, and effectively realizing the datasets with large density differences Eps. Literature [4] is based on the dataset so as to find the maximum effect of clustering index (CEI) to determine minPts according to the dataset characteristics, and to get the minPts according to the max CEI value, and to get Eps value by the Gaussian distribution, in order to achieve an adaptive threshold determined Eps. In order to free parameters and to discover clusters of arbitrary shapes and densities, literature [5] uses its one-dimensional projection analysis combined with Gaussian kernel density estimation methods to determine the density parameter (Eps and minPts). Other DBSCAN improved algorithm was also proposed by Zhou Dong's [6], the CURD algorithm proposed by Ma Shuai [7], Pan Ling ling's kernel DBSCAN algorithm [8] and so on.

Based on the above study, ant colony clustering algorithm combining high-dimensional data is projected onto a two-dimensional grid randomly to achieve effective handling of high-dimensional data, and has the advantages of the clustering properties of self-organization, and easy combination with other algorithms. This paper raises an improved DBSCAN algorithm LF-DBSCAN. Uneven density of high-dimensional data set is divided into several sub-datasets with the use of ant colony clustering algorithm and thus obtains different densities Eps value group. Using Eps parameter values for these different data sets to make algorithm DBSCAN cluster the uneven distribution of the density of high-dimensional data sets efficiently, to some extent. Experimental results demonstrate the effectiveness of LF-DBSCAN algorithm.

2. IMPROVE DBSCAN ALGORITHM BASED ON ANT CLUSTERING

As the classic DBSCAN algorithm selects only the data sets of global parameters (global parameter), making cluster algorithm cluster the different densities at the same time difficult. Real data sets tend to be high-dimensional, and are unevenly distributed and global parameters cannot characterize its intrinsic clustering structure. In order to achieve a non-uniform density dataset efficiently cluster, this paper

proposes an improved algorithm DBSCAN LF-DBSCAN algorithm.

2.1. DBSCAN Algorithm

DBSCAN algorithm [9] is a high-density-based clustering algorithm for the connection area, it defined cluster as the largest set of the density of dots connected, which divide cluster with sufficiently high density region. In order to identify a cluster, the algorithm needs to input two global parameters: a given radius of the neighborhood Eps and the minimum number MinPts. The basic idea of the DBSCAN algorithm is to optionally choose an object "p" from the data, the number of objects within a radius of Eps if, "p" region contains greater than or equal to MinPts months, then "p" is a core object and create a new cluster. Then find all density-reachable objects from "p", these objects will be marked as the same cluster. Otherwise, "p" is marked as noise or ignored. Then, DBSCAN handle all the other objects in the dataset similarly.

2.2. Ant Clustering Algorithm

According to the foraging behavior of real ants in nature, the Italian scholar Dorigo M *et al.* in 1991 first proposed the basic model of ant colony algorithm. Puts the ant colony algorithm into clustering analysis, the accumulation of ants from ant larvae corpse and classifies the larva behavior. Deneubourg JL [10] establishes a basic model based on ant colony clustering phenomena (Basic Model, BM). Lumer E and Faieta B extend the model to the areas of data analysis, and designed the LF algorithm used for data clustering [11]. The main idea is to make the high-dimensional data objects be randomly projected onto a two-dimensional plane, placing a number of artificial ants into the plane at the same time. Each ant randomly selects a data object, according to the similarity of the data object in the part of the area, and gets the probability of the ants picking up or leaving down the object. This probability guides the next action of the ants. After a finite number of iterations, the data object plane and the similarity of their neighboring area combine together to achieve self-organization clustering process.

Neighborhood similarity f (i) represents the average similarity of a data object and its neighborhood i ant objects found at the site between the r, f (i) by equation (1) to calculate

$$f(i)=\max\{0,\frac{1}{S^2}\sum_{j\in Neigh_{s\times s}(r)}1-\frac{d(i,j)}{\alpha}\} \quad (1)$$

In the formula, $\alpha\in[0,1]$ is the similarity regulatory factors; Neighs ´$Neigh_{s\times s}(r)$ represents area around r which is the side length of the square local area with s; d (i, j) means that the object i and j in the property distance in space, usually Euclidean distance or cosine distance function.

Probability of the ants constantly moving in a two-dimensional plane and repeatedly calculates the similarity according to the act of putting down and picking up the object, the higher the similarity of the neighborhood is, the lower the probability of picking up an object is; it also works on the contrary. If ants overload the data object, using the formula (2) to calculate the probability of ants putting down Pd; If ants have no load, press the formula (3) to calculate the probability of picking up ants Pp. Where, kp and kd are threshold constants.

$$P_d(i)=\begin{cases}2f(i) & iff(i)<k_d\\ 1 & others\end{cases} \quad (2)$$

$$P_p(i)=(\frac{kp}{kp+f(i)})^2 \quad (3)$$

LF algorithm can discover clusters of arbitrary shape, can effectively deal with high-dimensional data, and are easy to combine with other algorithms; but it takes more time and with low efficiency of the algorithm. In the LF algorithm, the movement of ants is random, which may take a lot of time for the ants find the sample object. Therefore, this paper uses the literature [12] LF proposed improved algorithm, artificial ants and data binding, eliminating the presence of unload ants to reduce the time cost.

2.3. Description LF-DBSCAN Algorithm

The core idea of LF-DBSCAN is: First, the use of LF algorithm to map high-dimensional data sets onto a two-dimensional grid, and bind artificial ants and data to move data objects freely. Achieve the initial division of data set and extract the center of each data subset (the distance between the point and the minimum cluster of all other points in the current), the average distance k-nearest neighbor (distance of the center of each neighborhood) serve as a threshold value Epsi; Then the value of all the neighborhood threshold Epsi should be in ascending order, followed by Epsi as a parameter called DBSCAN algorithm; After each call, the clustering of all data points have been marked, all marked points are no longer involved in the following until all are used up, the remaining data not processed is a noise point. By this way, ultimately realize the effective cluster of these uneven density dataset

Description LF-DBSCAN algorithm is as follows:

/ * Algorithm begins * /

For each data object in the data set Oi do

Randomly scatter Oi on the two-dimensional grid

End for

Initiate parameters Max,s,α,β,ε.etc

For iterations t<maximum number of iterations Max do

For all data objects Oi do

The average degree of similarity (1)calculating

a data object according to the formula

End for

Table 1. Characters of the datasets.

Dataset	Natural Number of Clusters	Data Dimensions	Data Set Size	Remark
Dataset1	3	2	346	Uniform Data Set
Dataset2	4	2	368	Non-uniform data sets
Iris	3	4	150	High-dimensional data sets
Wine	3	13	178	High-dimensional data sets

Based on the average similarity of all data objects in ascending order

For all objects sorted Oi do

If (f(Oi)<threshold p

Oi move, calculate f(Oi)

While ($f(o_i) < \varepsilon + \beta \exp(\frac{1-n}{\gamma})$) do

Oi move, calculate f(Oi)

End while

End if

End for

End for

For all object sorted Oi do

Recursively other objects Oi adjacent area-included within the same cluster

End for

For all sub-clusters i do

All objects in the distance matrix calculated Mi;

Calculate the distance matrix Mi digits of each column to find the distance and minimum points, this point in mind for the center cluster Midi;

Mi distance matrix of each column in ascending order and get transposed matrix KMi, calculated before KMi point in Midi row vector corresponding to k+1 components and the average value that is defined as a cluster of neighborhood threshold Epsi;

End for

All the neighborhood threshold Epsi arranged in ascending order;

While all neighborhood threshold Epsi untreated do

In order to select the neighborhood threshold Epsi, DBSCAN data clustering for cluster success point mark;

A call to all points of DBSCAN unmarked attend the next;

End while

/ * Algorithm terminates * /

3. EXPERIMENTAL RESULTS AND ANALYSIS

To verify the effectiveness of LF-DBSCAN four data sets are simultaneously used, and for each data set both DBSCAN algorithm and LF-DBSCAN clustering algorithm are separately used. The algorithms are implemented in MATLAB, carried out in CPU 2.1GHZ +2 G RAM + on Windows XP platform.

All experimental details of the data set are as shown in Table **1** and the dataset's Dataset1 and Dataset2 are artificial datasets. The distribution of the data objects are shown in Fig. (**1**) and Fig. (**2**); Iris and Wine from UCI Machine Learning Repository represents the datasets of data mining

To avoid completely dominated by the clustering results of a large range of variation of the attribute, and reducing the impact of the data size of the different properties of the absolute value to the calculated experiment, use the equation (4) for normalizing all data set; and use the F-Measure metric contrast DBSCAN algorithm and LF-DBSCAN effectiveness [13]. F-Measure metric is a comprehensive consideration of the precision and recall performance analysis method, which is defined by equation (5):

$$rg(x_{if}) = \frac{x_{if} - \min(f)}{\max(f) - \min(f)} \tag{4}$$

$$F(i,j) = \frac{2 \times p(i,j) \times r(i,j)}{p(i,j) + r(i,j)} \tag{5}$$

For samples of m data set, F-Measure defined in Equation (6):

$$F = \sum_j \frac{m_j}{m} \max_i F(i,j) \tag{6}$$

In the formula (3) above, x_{if} attribute represents the f value of the i-th data point, rg (x_{if}) represents the new value x_{if},Min (f) and max (f) denote the minimum and maximum desirable Properties f; p (i, j) for the accuracy, r (i, j) is a recall;maxf (i, j) takes on all clusters i, m_j is the number of objects in the class j.

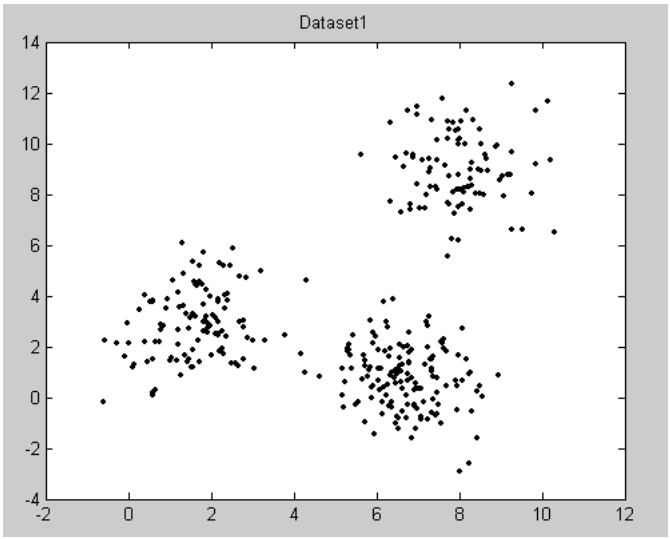


Fig. (1). Dataset 1.

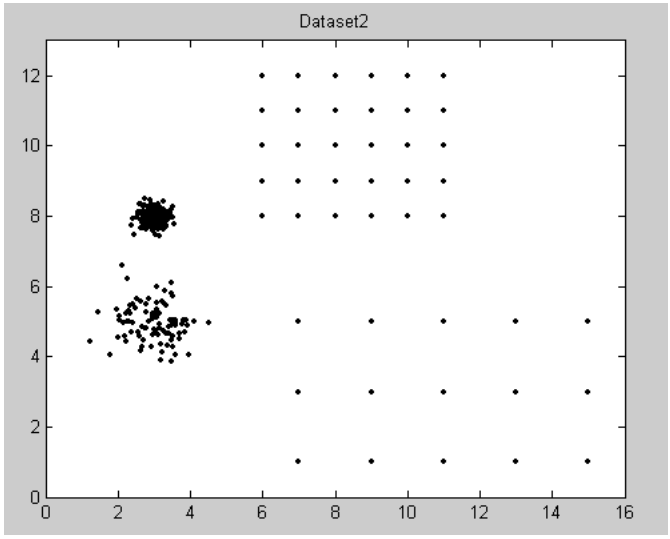


Fig. (2). Dataset 2.

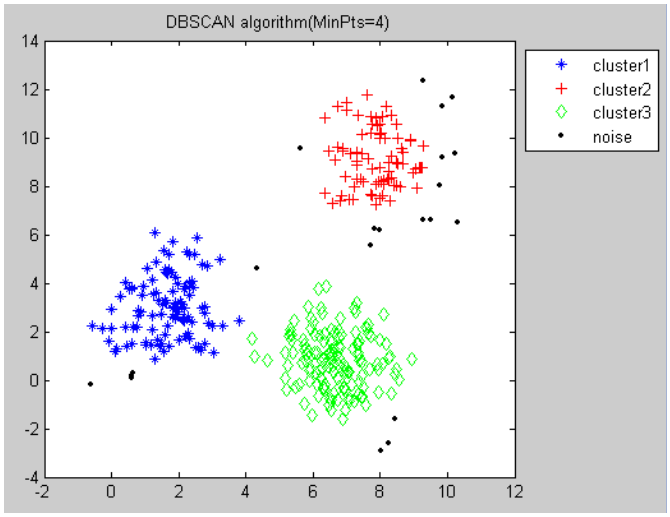


Fig. (**3**). Contd…

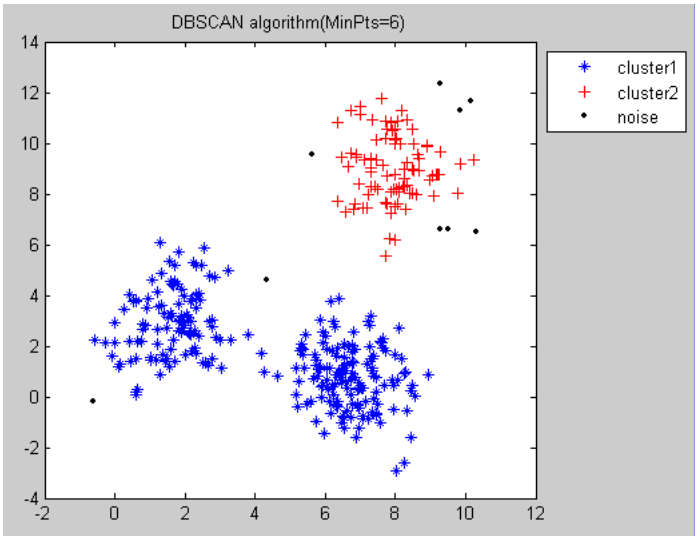
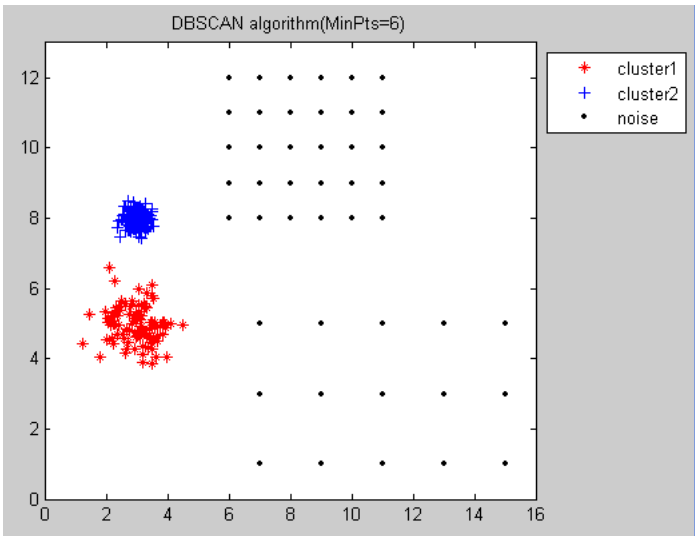


Fig. (3). Clustering results of DBSCAN on Dataset 1.



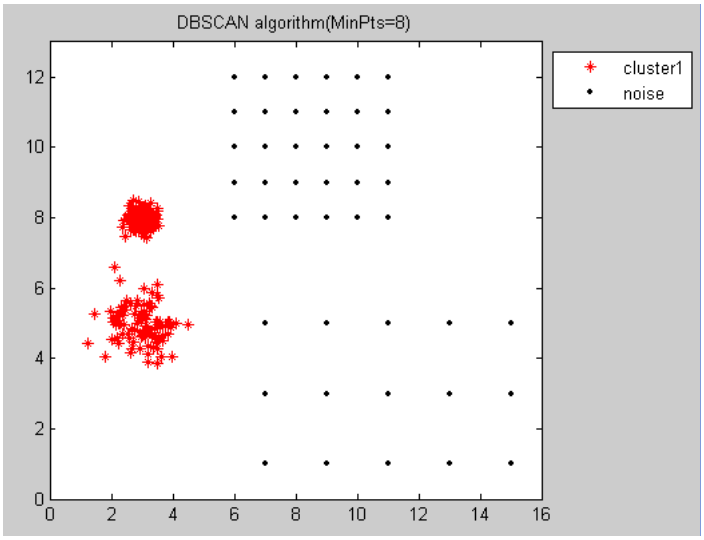
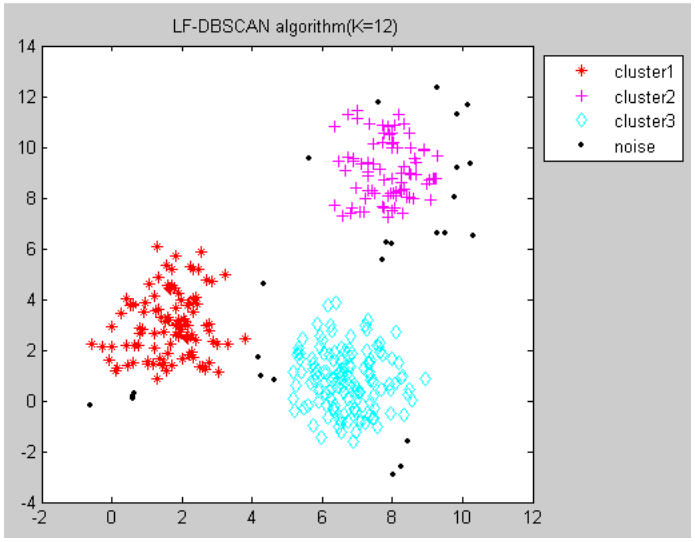


Fig. (4). Clustering results of DBSCAN on Dataset 2



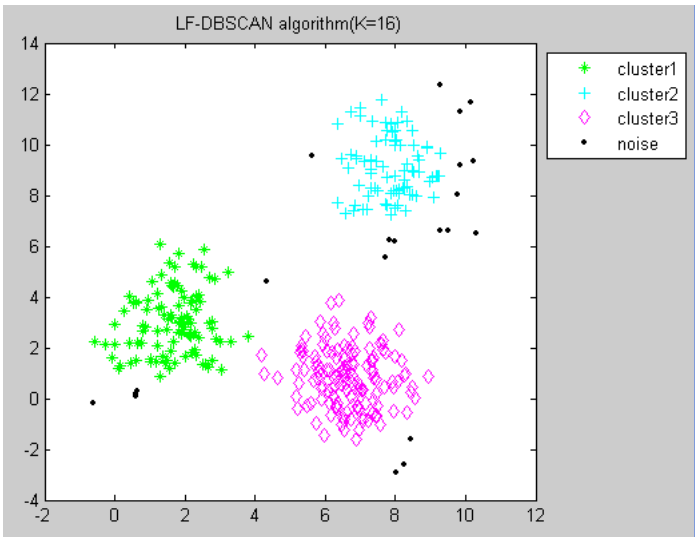


Fig. (5). Clustering results of LF-DBSCAN on Dataset 1.

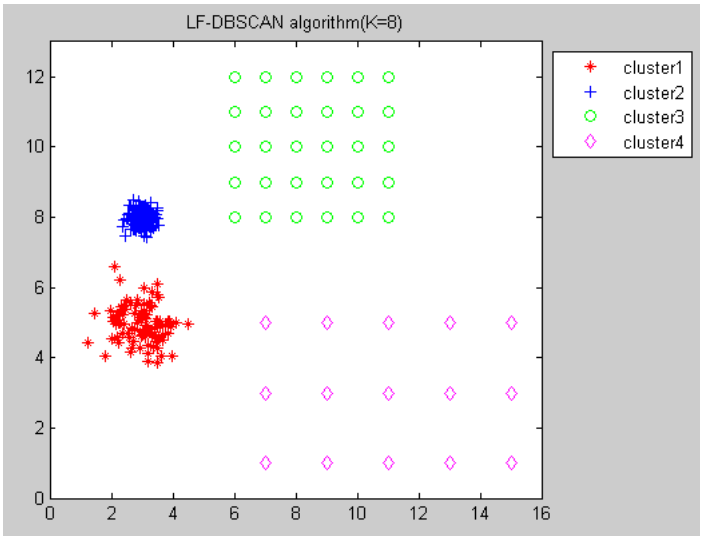


Fig. (**6**). Contd…

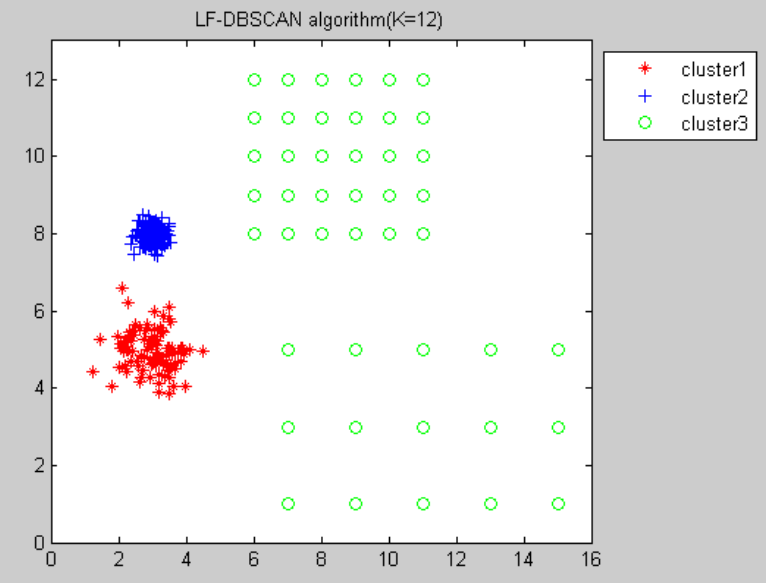


Fig. (6). Clustering results of LF-DBSCAN on Dataset 2.

APPENDIX:

Table 2. Result of DBSCAN and LF-DBSCAN algorithms on datasets.

Dataset	Algorithm Name	Neighborhood Threshold			Number of Clusters	F-Measure
Dataset1	DBSCAN	MinPts=4		Eps=0.0607	3	0.9681
		MinPts=6		Eps=0.0743	2	0.7494
	LF-DBSCAN	MinPts=6	K=12	Eps1=0.0523,Eps2=0.0546 Eps3=0.0788	3	0.9619
			K=16	Eps1=0.0629,Eps2=0.0574 Eps3=0.0864	3	0.9681
Dataset2	DBSCAN	MinPts=6		Eps=0.0720	2	0.8777
		MinPts=8		Eps=0.0832	1	0.6329
	LF-DBSCAN	MinPts=6	K=8	Eps1=0.0534,Eps2=0.2726 Eps3=0.4523	4	1.0000
			K=12	Eps1=0.0671,Eps2=0.3270 Eps3=0.5125	3	0.9633
Iris	DBSCAN	MinPts=6		Eps=0.3001	2	0.7778
		MinPts=8		Eps=0.3224	2	0.7778
	LF-DBSCAN	K=3 MinPts=8		Eps1=0.1085, Eps2=0.1115 Eps3=0.1501	3	0.8981

Table 2. contd…

Dataset	Algorithm Name	Neighborhood Threshold			Number of Clusters	F-Measure
Wine	DBSCAN	MinPts=4		Eps=0.7522	1	0.5060
		MinPts=6		Eps=0.7760	1	0.5060
	LF-DBSCAN	MinPts=6	K=8	Eps1=0.4758, Eps2=0.5823 Eps3=0.6313	2	0.6883
			K=12	Eps1=0.4956, Eps2=0.6281 Eps3=0.6604	2	0.6962

Experimental use DBSCAN clustering [14] proposed a method of calculating the Eps. According to the Literature [15] in the recommended formula, for the LF algorithm N data points scattered in the $\sqrt{10\times N}\times\sqrt{10\times N}$ grid, the number of iterations is 2000× N times and other parameters are initialized to : α=0.15、β=0.1 s=3、ε=0.5 and so on. To validate the algorithm LF-DBSCAN, apply DBSCAN algorithm and LF-DBSCAN algorithm of clustering a set of data, the clustering results are shown in Table **2** (Appendix). Fig. (**3-6**) is Dataset1 and Dataset2, DBSCAN clustering algorithms are invoked and the LF-DBSCAN consolidated. Clustering results can be seen Table **2** and Fig. (**3-6**), for uniform density datasets Dataset1, DBSCAN and LF-DBSCAN can effectively identify the natural shape of the data, but you can see DBSCAN sensitive to parameters when MinPts of the value is 6, it can only identify two clusters. For uneven density datasets, Dataset2, DBSCAN algorithm can effectively identify the number of clusters and their natural parameters are sensitive to relative LF-DBSCAN algorithm and can identify the natural clusters; For high-dimensional datasets Iris and Wine, DBSCAN is relatively poor in treatment effect, and therefore become less sensitive to the parameters MinPts, and the effect of LF-DBSCAN clustering better than algorithm DBSCAN. From Table **2** Comparison called on four datasets DBSCAN algorithm and F-Measure value after LF-DBSCAN, we can see the effect of LF-DBSCAN clustering algorithm is relatively better.

CONCLUSION

Using global parameters for traditional DBSCAN algorithm leads to poor cluster multi-density dataset, as well as the high-dimensional data processing result is not satisfactory enough. This paper presents a LF-DBSCAN algorithm to achieve effective non-uniform data set clustering. By comparing the experimental nature of the clustering effect of four different sets of data, the results demonstrate the effectiveness of LF-DBSCAN algorithm. But LF-DBSCAN algorithm still has room for improvements, such as the need to subjectively determine the parameters MinPts and k. It was found that the value k impacts the results of clustering effects, how to further reduce the influence of k is the next issue to be studied.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (61100034 and 61170043), China Postdoctoral Science Foundation (Grant 20110491411), Jiangsu Planned Projects for Postdoctoral Research Funds (Grant 1101092C), Anhui university provincial science research project (Grant KJ2011B108), and National Training Programs of Innovation and Entrepreneurship for Undergraduates (Grant No.201311306015).

REFERENCES

[1] X. Li, S. Jiang, Q. Zhang, and J. Zhu, "A dynamic density-based clustering algorithm appropriate to large-scale text processing", *Acta Sci. Nat. Univ. Pekin.*, vol. 49, no. 1, pp. 133-139, 2013.

[2] J. Hua, J. Li, S. Yi, X. Wang, and X. Hu, "A new hybrid method based on partitioning-based DBSCAN and ant clustering", *Expert Syst. With Appl.*, vol. 38, no. 8, pp. 9373-9381, 2011.

[3] J. Yang, J. Gao, J. Liang, and Y. Liu, "An improved DBSCAN clustering algorithm based on data field", *J. Front. Comput. Sci. Technol.*, vol.6, no. 10, pp. 903-911, 2012,

[4] G. Chen, B.-Q. Liu, and Y. Wu, "An Adaptive DBSCAN Algorithm Based on Gauss Distribution", *Microelectron. Comput.*, vol. 30, no. 3, pp. 27-30. 2013.

[5] M. Qian, and D. Ye, "Parameter free multi-density clustering using one-dimensional projection analysis", *J. Chin. Comput. Syst.,* vol. 34, no. 8, pp. 1866-1871, 2013.

[6] D. Zhou, and P. Liu, "VDBSCAN: varied density based clustering algorithm", *Comput. Eng. Appl.*, vol. 45, no. 11, pp. 137-141, 2009.

[7] S. Ma, T. Wang, S. Tang, D. Yang, and J. Gao "A fast clustering algorithm based on reference and density", *J. Softw.*, vol. 14, no. 6, pp. 1089-1095, 2003.

[8] L. Pan, Y. Zhang, and T. Xu, "Application of kernel DBSCAN algorithm in civil aviation customer segmentation", *Comput. Eng.*, vol. 38, no. 10, pp. 70-73, 2012.

[9] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", In: *Proc. 2nd Int. Conf. Knowl. Dis. Data Min. (KDD' 96)*, pp. 226-231, 1996.

[10] J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chretien, "The dynamics of collective sorting: robot-like ants and ant-like robots", In: *Proceedings of 1st International Conference on Simulation of Adaptive Behavior: From Animal to Animals*, pp. 356-365 1991.

[11] E. Lumer, and B. Faieta, "Diversity and adaptation in populations of clustering ants", *Proceedings 3rd International Conference on Simulation of Adaptive Behavior: From Animals to Animals.* Cambridge, MIT Press, pp. 501-508, 1994.

[12] Z. Ye, Z. Huang, "Clustering algorithm based on fusion of ant colony algorithm and K-medoids", *J. Electron. Measure. Instrum.*, vol. 26, no. 9, pp. 800-804, 2012.

[13] M. Steinbach, G. Karypis and V. A. Kumar, "Comparison of Document Clustering Techniques: Technical Report", *Minnesota: University of Minnesota Computational Science Engineering*, 2000.

[14] Q. Liu, L. Deng, Z. Jia, X. Qin, "Carrier frequency optimization of several cells based on enhanced DBSCAN algorithm", *Comput. Eng. Appl.*, vol. 50, no. 8, pp. 85-89, 2014.

[15] J. Handl, J. Knowles, M. Dorigo, "*Ant-based Clustering: a Comparative Study of its Relative Performance with Respect to K-means, Average Link and 1d-som"*, Technical Report TR/IRIDIA/2003-24, IRIDIA, Universite' Libre de Bruxelles, 2003. [http:// www. Handl. julia.de].

Received: November 06, 2014 Revised: December 05, 2014 Accepted: December 13, 2014