ISSN : 0974 - 7435 Volume 10 Issue 14

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(14), 2014 [8219-8224]

Ddbscan: a density detection dbscan algorithm in e-commerce sites evaluation

Jianhua Jiang[1,2]*, Haiyan Bian[1], Yumian Yang[1]

[1]School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun, (CHINA)

[2]Libratory of Logistics Industry Economy and Intelligent Logistics, Jilin University of Finance and Economics, Changchun, (CHINA)

E-mail : jianhuajiang@foxmail.com

ABSTRACT

To solve the problem of the uneven density data of DBSCAN algorithm, this paper proposes a density detection DBSCAN algorithm, which is named as DDBSCAN. Firstly, the density detection functions are designed as the evaluation standard of data density; secondly, high-dimensional data are classified into several partitions based on different density values; thirdly, Eps and MinPts parameters are set up in these partitions automatically; finally, the DBSCAN algorithm is applied to each partition respectively. Experimental results show that the proposed DDBSCAN algorithm is superior to the original DBSCAN in uneven density data clustering perspective.

KEYWORDS

Cluster analysis; Density detection; Data partition; DBSCAN; High-dimensional data.

INTRODUCTION

DBSCAN algorithm[1] was proposed by German scholar M. Ester et al. in KDD96. Until now, the algorithm has already been applied in many fields, such as image analysis[2-3], DNA diagnosis[4-5], text clustering[6], ray detection[7], and so on. Recently, many foreign and domestic scholars have proposed various improved methods. For example, in 2012, Patwary M, an American scholar, presented a new parallel DBSCAN algorithm, named as PDSDBSCAN, integrating graph concepts with the DBSCAN algorithm and applied it to both balanced workload and shared memory[8]. In 2013, scholar Kellner D from Germany proposed the grid-based DBSCAN which is applied to the analysis of radar data[9]. Domestic scholar Jiang Hua, in 2011, came up with a new hybrid method, the PACA-DBSCA algorithm, integrating partitioning-based PDBSCAN with ant clustering that could deal with multi-dimensional data[10]. In 2013, Xu Haixiao brought up an improved algorithm, which took advantage of classification on the level of density and applied it to classification of high-performance computing center users[11]. All these improved algorithms above adopt the thought of data partition to improve the quality of clustering. However, data partitions are cutting data directly and lack of flexibility. Therefore, it's an essential demand for DBSCAN that how to divide the uneven density data into several partitions with arbitrary shape.

This paper proposes a density detection clustering algorithm by means of introducing coefficient of variation in statistics and integrates it with the approach of the binning method with equal-depth in data mining to solve the problem of the uneven density data with high-dimension of DBSCAN algorithm. Firstly, originally high-dimensional data are carried out to density detection analysis. Secondly, the data are divided into several partitions based on detection result and *Eps* and *MinPts* parameters are set up in these partitions automatically. Finally, each data partition adopts DBSCAN algorithm to cluster and then the clustering results are merged. The key contributions in this paper are as follows:

(1)The improved DBSCAN algorithm can detect different density partitions of any arbitrary shape and cope with the high-dimensional and uneven density data intelligently, thus it can be applied to wider areas ;

(2)Density detection methods are designed to evaluate the density distribution of data.

DBSCAN ALGORITHM

DBSCAN algorithm[1] can discover clusters of arbitrary shape, and handling the noise points effectively when it is dealing with spatial data. The algorithm needs two parameters: *Eps* (the radius of the cluster), *MinPts* (minimum points of a point in its *Eps*-neighborhood). The DBSCAN algorithm judges whether it is a core point though checking out the neighborhood of the node in the dataset D and decides how to expand clusters.

From the DBSCAN algorithm, the value of global variable *Eps* affects the clustering quality, especially the uneven distributed data. Therefore, it is a better solution that data can be divided into partitions according to the level of density.

DDBSCAN: AN IMPTOVED DBSCAN ALGORITHM WITH DENSITY DETECTION

To solve the problem of the uneven density data of DBSCAN algorithm, this paper proposes an improved density detection DBSCAN algorithm, named as DDBSCAN.

It contains two main steps:

(1)Detecting the density of the original data, and forming different density partitions of arbitrary shape;

(2)Setting parameters' value automatically in the different density partitions for DBSCAN algorithm clustering.

Density detection functions

(1) Density formula of node *i*:

$$\rho_i = \frac{\left|Pts(i)\right|}{\pi \cdot Eps^2} \quad \textbf{(1)}$$

where *Eps* can be set according to the experimental situations, which refers to the radius of a circle whose center is *i*; *Pts* (*i*) is a set of nodes which are in the circle whose center is *i* and radius is *Eps*, and thus, |*Pts* (*i*)| is the number of the set *Pts* (*i*).

$$\rho_k = get\_\rho(Pts_k), \quad Pts_k \in Pts(i) \quad \textbf{(2)}$$

where Pts_k is a point of the set *Pts* (*i*), and *k* is the density of the point Pts_k, which ρ_k presents the density of *Eps*-neighborhood of Pts_k.

(2) The mean formula of the density in the *Eps*-neighborhood of node *i*:

$$\overline{\rho}_i = \frac{\sum_{k=1}^{|Pts(i)|} \rho_k}{\left|Pts(i)\right|} \quad \textbf{(3)}$$

where $\overline{\rho}_i$ is the average density of *Eps*-neighborhood of node *i*.

(3) Formula of the density variance:

$$s^2 = \frac{1}{n-1}\left[\sum \rho_i^2 - n\cdot\left(\overline{\rho}_i\right)^2\right], \quad n = \left|Pts(i)\right| \tag{4}$$

where, s^2 is the density variance of the *Eps*-neighborhood of node *i*, which reflects the deviation degree between each point in the *Eps*-neighborhood and the mean, *n* is the number of points in the *Eps*-neighborhood.

(4) The density coefficient of variation:

$$cv_i = \frac{s}{\overline{\rho}_i}\ , s = \sqrt{s^2} \tag{5}$$

where, cv_i reflects the variation of density of node *i*, which happens in its *Eps*-neighborhood.

As can be seen from equation (1-4) that in the area whose radius is *Eps*, the more the points are, the relevant value of *cv* is smaller; conversely, the value of *cv* is greater.

Data partition based on binning method with equal-depth

Binning method with equal-depth, at first, needs to determine the depth of the box, and as can be seen from TABLE 1, a schematic case that we assume that the depth of each box is 5, with a total of 100 points. To measure the variation of each box, this paper introduces range change rate that is equal to the range value of one box divided by the value of its followed box. Setting λ as the threshold. If the range change rate is larger than λ, the division point is the critical point between the two boxes for data partitions; on the contrary, when the range change rate is smaller than λ, there is no data partition. Schematic is shown in TABLE 1:

TABLE 1 : Schematic of binning method with equal-depth

Box ID	Dataset	Range	Range change rate
1	2,3,4,5,6	4	5/4 = 1.25
2	6,8,9,11,11	5	24/5 = 4.8
3	18,21,26,40,42	24	...
...	...	...	...
20	58,59,62,69,78	20	——

The value of range and range change rate of each box can be calculated according to the data in TABLE 1. If the threshold value λ is set to 2, there is no data partition between *box 1* and *box 2*. And the range change rate of *box 2* is 4.8, larger than the threshold value λ, so data partition is needed between *box 2* and *box 3*, and the division point of the data partition can be the maximum 11 of *box 2* or the minimum 18 of *box 3*.

Automatically set of the parameter *Eps* in each partition of DBSCAN algorithm

Eps is a global variable of the original DBSCAN algorithm, and usually be set according to personal experience. Especially when facing the process of uneven density data, it is difficult for DBSCAN algorithm to set proper *Eps* values to get the optimal clustering result. The value of *Eps* in each partition is considered as the mean distance *Eps* (*i*) between every node and its first |*Pts* (*i*)| nearest neighbor. The specific steps are as follows:

(1) Determining the value of *MinPts*. The value of *MinPts* of the cluster which the *cv* is relatively small is 4; the value of *MinPts* of the cluster which the *cv* is relatively large is 2. And it can be adjusted automatically based on concrete data density;

(2) Calculating the number of nodes in the *Eps*-neighborhood of each node by the equation $\left|Pts(i)\right| = \rho_i \cdot \pi \cdot Eps^2$, which density value of each node ρ_i is got by Equation (1);

(3) Calculating the distance between every node and its first |*Pts* (*i*)| nearest node by the equation $Eps(i) = \frac{\left|Pts(i)\right|}{MinPts}\cdot Eps$;

(4) Considering the mean of *Eps*, i.e. *E* (*Eps*) as the *Eps* of each partition : $E(Eps)=\frac{\sum_{i=1}^{n} Eps(i)}{n}$, and *n* is the number of nodes in sample.

DDBSCAN algorithm

The improved DBSCAN algorithm introduces the basic idea of density detection to solve the problem that DBSCAN algorithm can generate a large number of outliers in the process of the uneven density data with high-dimension. The proposed density detection method is to query its *Eps*-neighborhood of *n* nodes at first, and thereby to obtain the value of *cv* of each node. The data partition is segmented by the value of *cv*. The algorithm is as follows:

Algorithm 1 : DDBSCAN algorithm

Input: the dataset D containing *n* objects and the radius *Eps*.

Output: the clustering result.

(1) Firstly, querying the Eps-neighborhood of each node, and calculating its density respectively;

(2) Calculating the mean density of each node according to the result of step (1), and furthermore achieving the variance s^2 of each node;

(3) Putting the result from step (2) in formula (5) and calculating the value of *cv* of each node;

(4) Partitioning data based on binning method with equal-depth;

(5) Setting the value of *Eps* and *MinPts* automatically;

(6) Running DBSCAN algorithm in every data partition;

(7) Getting the clustering result and ending the entire algorithm.

E-EOMMERCE SITES EVALUATION WITH DDBSCAN ALGORITHM

Sample selection

This paper selects the data from 100 E-Commerce demonstration enterprises in 2013-2014 selected by the Ministry of Commerce of the People's Republic of China. Access Per Million Users and Site Linking In[12] is used as the evaluation index of the E-commerce websites. Parts of the E-commerce websites are shown in TABLE 2:

TABLE 2 : The original data of E-commerce websites (parts)

Websites	Access Per Million Users	Site Linking In
1688.com	9684000	88426
jd.com	3828000	7913
6666688888.com	684	207
mmb.cn	300	35
....	....	....
lusen.com.cn	56	7
coffee58.smehi.cn	52	30

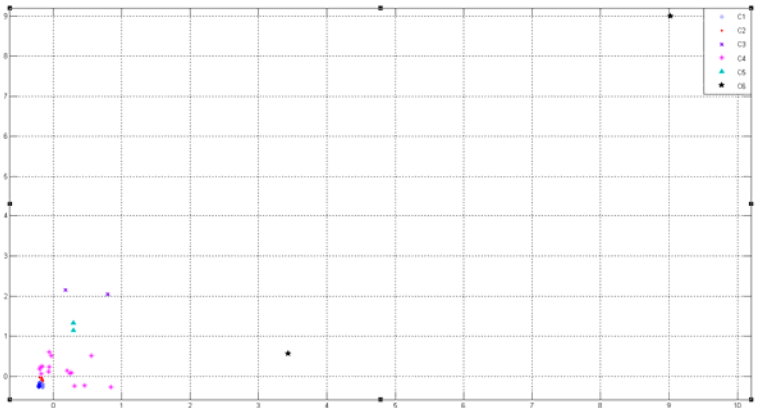

Figure 1 : The clustering result of DDBSCAN algorithm

The simulation and application of DDBSCAN algorithm in Weka

In this paper, DDBSCAN algorithm is executed in the Weka3[13] platform and implemented on the basis of the original DBSCAN package. The sample data can be divided into two partitions by means of binning method with equal-depth. One is partition 1(C1, C2), and the other is partition 2(C3, C4, C5, C6). Then the values of *Eps* and *MinPts* are calculated by the step 3.3 referred previously: *Eps1*=0.4, *MinPts1*=4; *Eps2*=0.02, *MinPts2*=2. The clustering result is shown in Figure 1. A color represents a cluster in the Figure. In addition, the black "□" indicates the outliers.

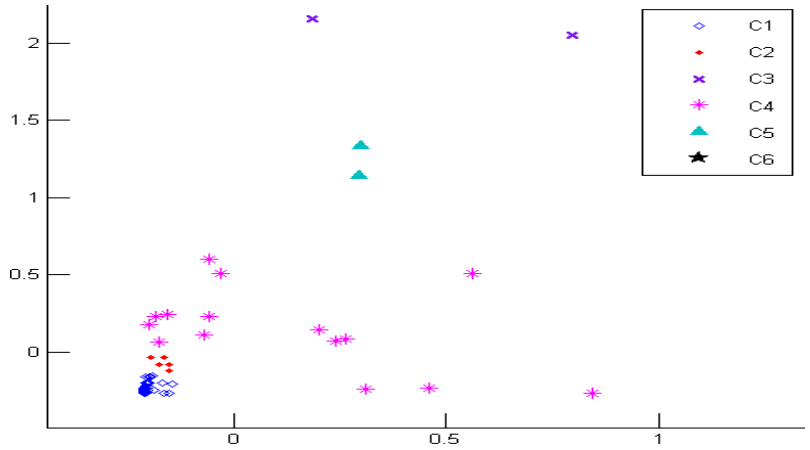

Figure 2 : The clustering result of DDBSCAN algorithm without outliers

The clustering result of the original DBSCAN algorithm applied to sample data is shown in Figure 3 to compare with that the DDBSCAN algorithm.

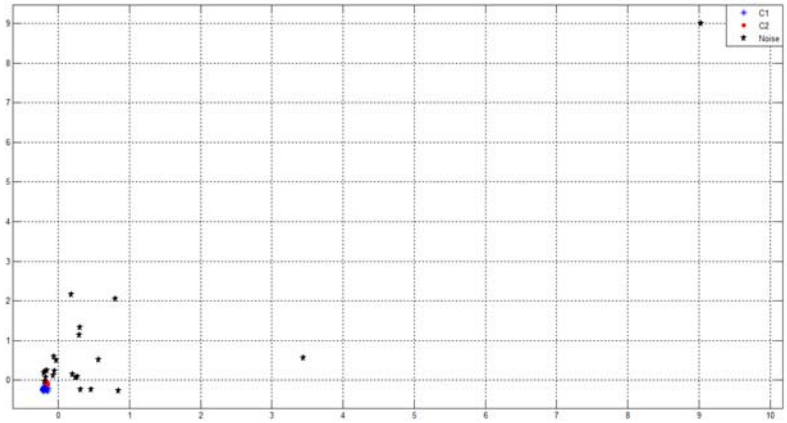

Figure 3 : The clustering result of the original DBSCAN algorithm

From the clustering results of the two algorithms we can see that the clustering effect of the original DBSCAN algorithm is not satisfactory in terms of the uneven density data. The relatively sparse data (C3, C4, C5) in Figure 2 are processed to outliers. Instead, DDBSCAN algorithm divides the sample data into partitions and automatically determines different parameters depending on the circumstances of each partition and the clustering result of the relatively sparse data on the upper right corner is, in general, in line with the data distribution. Besides, the number of data regarded as the outliers (C6) by DDBSCAN algorithm is significantly less than that by the original DBSCAN algorithm, which conforms to the characteristics of the data distribution. Experimental result shows that the DDBSCAN algorithm can get a better clustering result when data density is uneven.

CONCLUSIONS

DBSCAN algorithm can discover clusters of arbitrary shape with noise inside, but for the uneven density data, it is less effective in discovering clusters because of global settings of the parameters Eps and *MinPts*. This paper presents the DDBSCAN algorithm relying on density detection. Firstly, the original data with high-dimension are divided with the density detection functions, so that each partition can be of any shape and the density distribution is relatively uniform. Secondly, the value of *Eps* is set automatically for each partition in the process of clustering. Experimental result shows that the DDBSCAN algorithm is superior to the original DBSCAN in uneven density data clustering perspective.

ACKNOWLEDGEMENT

Jianhua Jiang is corresponding author. This research has received financial support by the National Natural Science Foundation of China (No. 61202306, 61170004, 61472049, 61402193), by the Foundation of Education Bureau of Jilin Province (No. 2012188), and by the Foundation of Jilin University of Finance and Economics (No. XJ2012007, 2013006).

REFERENCES

[1] M.Ester, H.P.Kriegel, J.Sander, X.Xu; A density-based algorithm for discovering clusters in large spatial databases with noise, In Proceedings of 2nd International Conference on Knowledge Discovering in Databases and Data Mining, Portland, Oregon **(1996)**.

[2] I.Lee, G.Cai, K.Lee; Mining points-of-interest association rules from geo-tagged photos, 46th International Conference on System Sciences (HICSS), IEEE, 1580-1588 **(2013)**.

[3] K.Rahul, R.Agrawal, A.K.Pal; Color Image Quantization Scheme Using DBSCAN with K-Means Algorithm/Intelligent Computing, Networking, and Informatics, Springer India, 1037-1045 **(2014)**.

[4] Z.Francis, C.Villagrasa, I.Clairand; Simulation of DNA damage clustering after proton irradiation using an adapted DBSCAN algorithm, Computer methods and programs in biomedicine, **101(3)**, 265-270 **(2011)**.

[5] M.Dos Santos, C.Villagrasa, I.Clairand, et al; Influence of the DNA density on the number of clustered damages created by protons of different energies, Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms, **298**, 47-54 **(2013)**.

[6] P.D.Turney, P.Pantel; From frequency to meaning: Vector space models of semantics, Journal of artificial intelligence research, **37(1)**, 141-188 **(2010)**.

[7] A.Tramacere, C.Vecchio; γ-ray DBSCAN: a clustering algorithm applied to Fermi-LAT γ-ray data I. Detection performances with real and simulated data, Astronomy & Astrophysics/Astronomie et Astrophysique, **549 (2013)**.

[8] M.Patwary, M.Ali, D.Palsetia, et al; A new scalable parallel dbscan algorithm using the disjoint-set data structure// International Conference for High Performance Computing, Networking, Storage and Analysis (SC),IEEE, 1-11 **(2012)**.

[9] D.Kellner, J.Klappstein, K.Dietmayer; Grid-based DBSCAN for clustering extended objects in radar data IEEE, Intelligent Vehicles Symposium (IV), IEEE, 365-370 **(2012)**.

[10] H.Jiang, J.Li, S.Yi, et al; A new hybrid method based on partitioning-based DBSCAN and ant clustering, Expert Systems with Applications, **38(8)**, 9373-9381 **(2011)**.

[11] H.Xu, J.MA, Q.WU; Application Research of DBSCAN Algorithm Based on High-Performance Computing Center Users Classification, Journal of Jilin University (Information Science Edition), **5**, 528-534 **(2013)**.

[12] Zhi-ping Hou; Analysis on application and research of factor and clustering in E-commerce website evaluation, Science and Technology Management Research, **18**, 144-147 **(2011)**.

[13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, H.Ian Witten; The WEKA Data Mining Software:An Update, SIGKDD Explorations, **11(1), (2009)**.