A Parameter Adaptive Clustering Algorithm Based on Reference Points and Density

Cheng Ouyang
College of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China
E-mail: 1149733987@qq.com

Jun Yu
Institute of Bioinformatics
Chongqing University of Posts and Telecommunications
Chongqing, China
E-mail: yujun@cqupt.com

Jun Tan
Institute of Bioinformatics
Chongqing University of Posts and Telecommunications
Chongqing, China
E-mail: tanjun@cqupt.com

ZhiFan Zeng
China Southern Power Grid
Administration of power supply
Guangxi, China
E-mail: 550350617@qq.com

Abstract—CURD is one type of clustering algorithm based on reference point and density. This algorithm is similar to DBSCAN in processing arbitrary shape clustering ability and has linear time complexity of K-MEANS algorithm. CURD algorithm needs to set Radius and t, so the whole process of clustering needs manual intervention which has the similarity with most of clustering algorithm. This paper proposes SA-CURD clustering algorithm based on CURD which can automatically set Radius and t by analyzing dataset statistics, in order avoid manual intervention in the process of clustering and achieve complete automation. Experiments show that SA-CURD can rationally choose Radius and t and get highly precise clustering results.

Keywords-Data Mining; Clustering; SA-CURD; CURD

I. INTRODUCTION

As a technique of finding interesting information from massive data, Data Mining[1] has met extensive appliance day by day. Clustering is one important data digging technique, and it aims to put data set into different cluster and data in same cluster has higher similarity. Now the clustering algorithm has four types basically[2]:

Partitioning method: Including K-means, CLARA, K-mode. The main advantages of this category lie in lower complexity and relatively high efficiency and flexible in processing data set; the disadvantage is that it needs the number k of cluster firstly therefore contradict with clustering algorithm's original intention. Also, the choose of k initial points will have big influence for cluster result and this algorithm can only find non-undercut globular cluster which is very sensible to noisy data[3].

Hierarchical method: Including RIRCH, CURE and ROCK. This algorithm can process noisy data and CURE and ROCK can find arbitrary-shape cluster (BIRCH can find non-undercut cluster). However, CURE and ROCK have higher complexity so they have to adopt methods like sample and partition. Besides, this algorithm also needs to get the number k of cluster.

Density-based method: Including DBSCAN and DENCLUE. This algorithm can find arbitrary-shape cluster. It is insensitive to data's input order and it is unnecessary to appoint the number of cluster in advance. However, this algorithm has higher complexity[4].

Grid-based on method: Including STING, CLIQUE. The main advantage of this algorithm is its high efficiency. But this algorithm barely considers data's distribution[5], and it does not have good clustering quality because it uses statistics information of only one grid to substitute all points of this grid.

As researchers demonstrated above, the main problem of density-based clustering algorithm is high complexity[6]. In order to settle this problem, Dr. Ma Shuai[7] from Peking University proposed a clustering algorithm CURD based on reference point and density. The algorithm maintains all the merits of DBSCAN and has linear time complexity. All researchers need is manual choosing Radius and t before clustering.

II. RELATED WORK

A. Curd algorithm

The CURD algorithm use certain numbers of reference point to express a clustering zone and shape. The reference point is virtual and unstable and CURD use density to shield abnormal noisy data. Besides, adopt distance can let the processing of high dimension data switch into one-dimensional space. Therefore, at this point, CURD can process high-dimensional data.

Definition 1 (Density of point). For arbitrary point P and distance Radius in space, use P as central point and R as central distance from P, then the number of other points in the space is called density, denoted by Density (p, radius).

Definition 2 (Reference point). For arbitrary point p, distance radius and threshold t, if Density (p, radius)>=t, then researchers call p as reference point and call t as density threshold.

Definition 3 (Representative region). Every reference point represents a circular region whose center of circle is the reference point and researchers call the region as representative region.

Definition 4 (Neighboring reference). Given the distance radius and the threshold T, If reference point p and q can satisfy Dist(p, q)<=2*radius, researchers call p and q are neighboring reference.

CURD introduces reference point and use certain numbers of reference point to represent the region and shape of a cluster efficiently. The frame of CURD is shown in Fig. 1. This algorithm finds geometry reference point which can accurately reflects the characteristic of the input data space firstly, then establish the mapping between reference point and representative region point and then classify the reference point, make sure that each reference point can constitute basic information of a cluster. At last, researchers gather the same representative region points from the same reference point into a cluster.

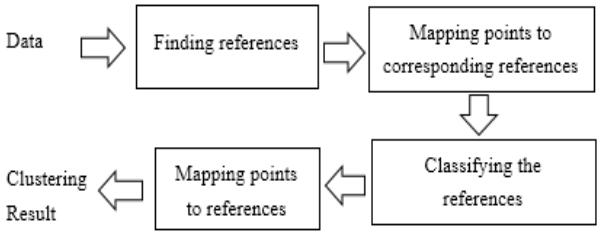

Figure 1. CURD algorithm

III. PARAMETERS DETERMINATION METHOD FOR CURD

The accuracy of CURD is related to the value of radius and t. When t is fixed, then density will increase with the decrease of Radius. If the radius was oversized, then more noisy data will enter the cluster and several decided natural cluster will be merged incorrectly in one cluster. When the radius is fixed, the density will increase with the increase of density. If the value of t is too small, it will result in massive points being marked as central point then it will bring noisy data into cluster. If the value of t is too big, it will cause the decrease of numbers of central point. Therefore, based on CURD, our research has proposed a distance-based self-adaption method to ascertain parameter radius and t, researchers call this method SA-CURD. The main idea of this method is using statistics characteristics[8] of dataset to choose the value of radius and t.

A. *Method for determining the parameter radius*

Definition 1 Given a clustering space K = {P, R}, where R representation dimension and P = { p_1,p_2, … , p_n}. Suppose d (i, j) is the distance between point p_i and p_j in sample and d_w is the average distance between all points in sample.

$$d_w = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d(i,j)}{n(n-1)/2} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\|p_i - p_j\|}{n(n-1)/2}$$

Definition 2 Given a clustering space K = {P, R}, where R representation dimension and P = { p_1,p_2, …, p_n}. Let d_k (i, j) is also the distance between point p_i and p_j in sample, which satisfies d (i, j) < d_w (d (i, j), and d_s represents the average value of all d_k.

$$d_s = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} dk(i,j)}{k(k-1)/2} = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\|p_i - p_j\|}{k(k-1)/2}$$

Definition 3 Given a clustering space K = {P, R}, where R representation dimension and P = { p1, p2, …, pn}.Mean difference value AvgSpt is the mean of absolute value of the difference between dk and ds.

$$\text{AvgSpt} = \frac{\sum_{i=1,j\neq i}^{k} |d_k(i,\ j) - d_s|}{k}$$

$$= \frac{\sum_{i=1,j\neq i}^{k} \left| \|p_i - p_j\| - \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\|p_i - p_j\|}{k(k-1)/2} \right|}{k}$$

The concrete steps of determining the radius parameter are as follows.

(1) Calculate the average value dk of Euclidean distance between all points;

(2) Calculate the mean absolute deviation AvgSpt of all points and mean distance values;

(3) Remove the point that greater than the average point d_k, and calculate the average Euclidean distance ds between the remaining points;

(4) Determine whether the value of (dk - AvgSpt) is greater than ds, if it is, turn to Step (6), if it isn't, then turn to Step (5);

(5) Regard the remaining points in Step (3) as all points, and turn to Step (1);

(6) Output radius;

$$\text{radius} = d_s - AvgSpt = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\|p_i - p_j\|}{\frac{k(k-1)}{2}}$$

$$- \frac{\sum_{i=1,j\neq i}^{k} \left| \|p_i - p_j\| - \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\|p_i - p_j\|}{k(k-1)/2} \right|}{k}$$

B. *Method for determining the parameter t*

Definition 1 Let C={C_1, C_2, …, Cn} are candidate reference points and Num(i)={N_1, N_2, …, Nn} is the number of sample point in candidate reference points. Suppose ANum is the average number of sample points for all candidate reference points.

$$\text{ANum} = \frac{\sum_{i=1}^{n} N_i}{n}$$

Definition 2 Let C ={C_1, C_2, …, C_n} are candidate reference points and Num(i)={N_1, N_2, …, N_n} is the number of sample point in candidate reference points. Suppose ASNum is the average value of the number of sample point which less than ANum. AvgT is the mean of absolute values of the difference between the items which less than ANum and ASNum.

$$\mathrm{AvgT} = \frac{\sum_{i=0}^{k}|N_i - ASNum|}{k} = \frac{\sum_{i=0}^{k}|N_i - \frac{\sum_{i=1}^{n}N_i}{n}|}{k}$$

The concrete steps of determining the t parameter are as follows.

(1) Calculate the average value ANum of the sample points for all candidate reference points;

(2) Calculate the mean absolute deviation AvgT;

(3) Output t ;

$$\mathrm{t} = \mathrm{ANum} - \mathrm{AvgT} = \frac{\sum_{i=1}^{n}N_i}{n} - \frac{\sum_{i=0}^{k}|N_i - \frac{\sum_{i=1}^{n}N_i}{n}|}{k}$$

IV. EXPERIMENT AND ANALYSIS

The computer configuration of this experiment is, CPU is 2.5GHz, memory is 4G, and the operating system is Windows7 flagship version. The algorithm uses Java programming language.

A. *Experimental data and results*

Experiments using two artificial synthetic data set DS1 and DS2. DS1 is a two dimensional data set of 180 objects, DS2 is the data set of 200 objects, the two and their clustering results are shown in Fig. 2, Fig. 3, Fig. 4, Fig. 5. By Fig. 3, Fig. 5 can be seen, SA-CURD can find the high density of the data focus area and make appropriate cluster division. This shows that the SA-CURD algorithm can effectively select the appropriate Radius and t parameters.

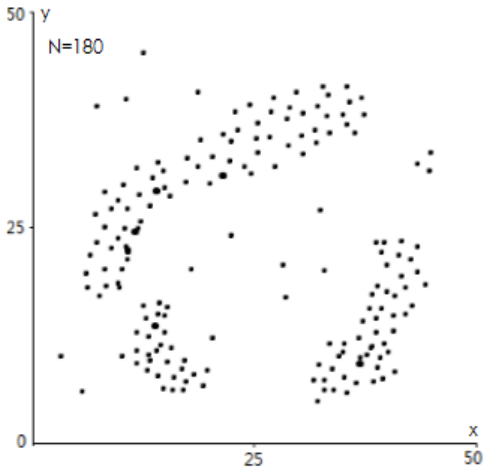

Figure 2. DS1

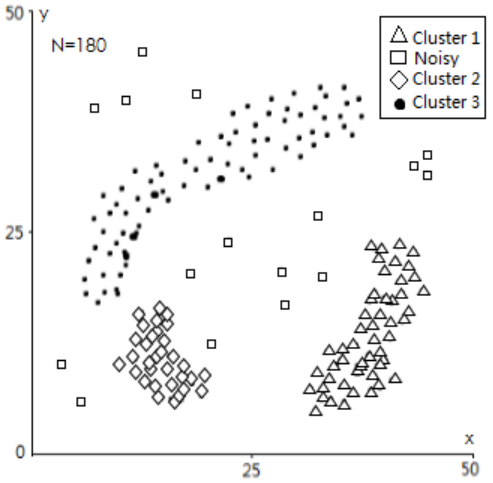

Figure 3. SA-CURD clustering results for DS1 (radius=3.58, t=4)

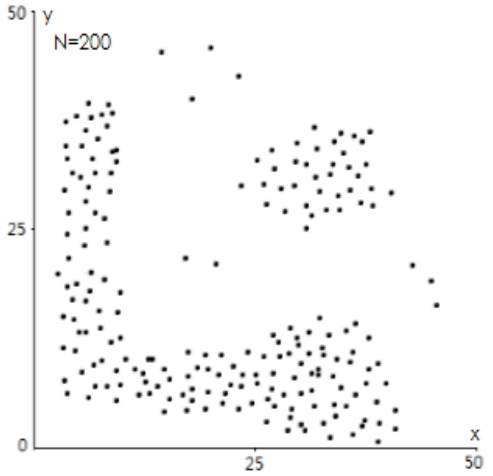

Figure 4. DS2

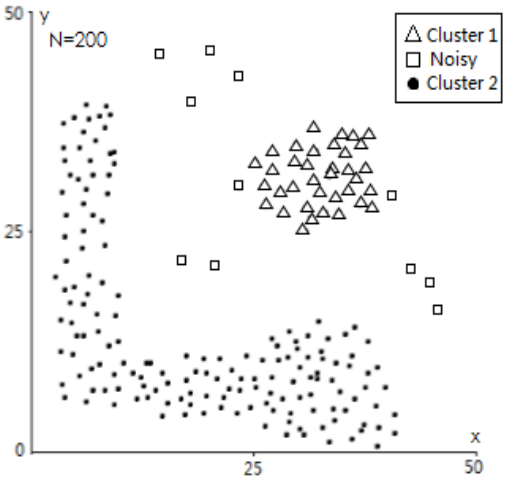

Figure 5. SA-CURD clustering results for DS2 (radius=4.82, t=5)

B. *Time characteristic analysis and clustering accuracy*

Researchers use supervised F metric[9] to detect the correctness of clustering. TABLE I gives the clustering results and accuracy index of DS2 and DS1 datasets, and compares it with the traditional CURD clustering algorithm.

TABLE I. THE TIME AND ACCURACY OF SA-CURD AND CURD OF COMPARISON OF TWO ALGORITHMS

Data set	Dimension	Number of objects	Clustering algorithm	radius	t	Running time(ms)	Accuracy(%)
DS1	*2*	*180*	*SA-CURD*	*2.58*	*4*	*231*	*99.44*
			CURD	*1*	*2*	*165*	*72.84*
DS2	*2*	*200*	*SA-CURD*	*3.82*	*6*	*248*	*98.5*
			CURD	*1*	*2*	*172*	*66.42*
DS3	*2*	*350*	*SA-CURD*	*6.9*	*8*	*322*	*96.28*
			CURD	*2*	*3*	*237*	*55.62*
DS4	*3*	*85*	*SA-CURD*	*1.74*	*5*	*188*	*95.54*
			CURD	*0.5*	*2*	*162*	*79.18*

From TABLE I, it can be seen that the improved SA-CURD is slower than the traditional CURD algorithm in time performance. This is because the SA-CURD takes an extra time for calculating the values of the radius and t parameters. However, its complexity is the same level of CURD algorithm. Are about data quantity N is linear complex.

Can be seen from TABLE I also, directly take 1/50 of the data space as the radius, take the average density of the candidate reference points as t for CURD clustering accuracy is not high. Therefore, it is necessary to determine the parameters of radius and t by analyzing the statistical features of data sets. The improved SA-CURD algorithm can determine the parameters according to the statistical characteristics of the data set, and the accuracy of the clustering results is relatively high.

V. FURTHER DISCUSSION

For the cluster density of very different data sets will greatly reduce the accuracy of the clustering of SA-CURD. This is the main problems existing in the CURD algorithm itself. The use of a global single parameter radius and t, causing that the clustering process is the only one measure of density. If the radius is selected to be large, it will lead to a high density of natural clusters are merged; and the choice was small, the result of low density of natural clusters are discarded[10].

VI. CONCLUSIONS

CURD algorithm is a clustering algorithm based on reference point and density, which holds all the advantages of DBSCAN and can find any shape of clusters, and has approximate linear time complexity. CURD algorithm needs two parameters of radius and T, resulting in the need of manual intervention for the clustering process. Based on the CURD, the SA-CURD clustering algorithm for radius and t is proposed. According to the statistical characteristics of the data set, the algorithm chooses the appropriate radius and T, and realizes the full automation of the clustering process, and improves the accuracy of the clustering results.

REFERENCES

[1] Han JiaWei, Kamber M. Data Mining: Concept and Techniques[M] . America: Morgan Kaufmann Publishers, 2001. 223-224.

[2] Sun Jigui, Liu Jie, Zhao Lianyu. Research on clustering algorithm [J]. software journal, 2008, 19 (1): 48-61.

[3] S.Guha, R.Rastogi, K. Shim, Cure: an efficient clustering algorithm for large databases, Information Systems 26 (1) (2001) 35-58.

[4] A.M. Fahim, G. Saake, A.M. Salem, F.A. Torkey, M.A. Ramadan, Dcbor: a density clustering based on outlier removal, International Journal of Computer Science 4 (3) (2009).

[5] Shao Rong Feng, Xiao Wenjun. An improved DBSCAN clustering algorithm quality new method [J]. Xi'an University of Electronic Science and Technology Journal (NATURAL SCIENCE EDITION), 2005,35 (3): 523-529

[6] B. Borah, D.k. Bhattacharyya, Ddsc: a density differentiated spatial clustering technique, Journal of Computers 3 (2) (2009) 72.

[7] Wang Tengjiao, Ma Shuai, Tang Shiwei, et al. A fast clustering algorithm based on reference point and density [J]. software journal, 2003,14 (6): 1089-1095.

[8] Edited by Wu MeiCun. The basic principles and methods of mathematical statistics [M] Chengdu: Southwestern University of Finance and Economics press, 2006.

[9] Steibach M, KarypisG,Kumar V. A comparison of document clustering techniques: technical report[R]. Minnesota: University of Minnesota-Computer Science and Engineering, 2000.

[10] A.Hinneburg, D.A.Keim, An efficient approach to clustering in large multimedia databases with noise, Knowledge Discovery and Data Mining 5865(1998).