The research of K-medoids clustering algorithm based on density

Ping Liu[1], Hao Zhou[1], Junping Yang[2], Taorong Qiu[1]*

[1] *The school of Information Engineering, Nanchang University, Nanchang 330031, China*

[2] *Affiliated Hospital of Jiangxi College of Traditional Chinese Medicine, Nanchang 330006, China*

**Coresponding author's e-mail: qiutaorong@ncu.edu.cn*

Received 15 October 2013, www.cmnt.lv

Abstract

In view of that the clustering result of the traditional k-medoids clustering algorithm being sensitive to initial cluster centers. A new k-medoids clustering algorithm based on density was proposed in this paper. It conducted a rough clustering to generate several particles at first. Then select the centers of the k densest particles as the initial clustering centers. Tested by using UCI data sets, the validity of the proposed algorithm is demonstrated.

Keywords: k-medoids, density, clustering, cluster centers

1 Introduction

Clustering is the process of dividing a set of objects into several clusters. And then the objects are similar to each other in the same cluster. But the objects in different clusters are dissimilar. The traditional k-medoids clustering algorithm is not sensitive to noise, and it is simple and it has a fast convergence rate and strong local search ability, so it is widely used [1-8]. But the k-medoids clustering algorithm has the drawback that it is sensitive to the initial clustering centers. In order to solve this problem, many domestic and foreign researchers have done some efforts to improve the k-medoids clustering algorithm [5].

In article [6], a simple and fast k-medoids clustering algorithm was proposed. It solved the problem that the clustering result is sensitive to the initial clustering centers. In addition, it improves the convergence rate, but the initial clustering centers, which selected in this way may be in the same cluster. However, if there are some initial centers in the same cluster, then it will have a bad impact on the clustering accuracy. In article [7], a new algorithm was also proposed that the local search process is embedded in the iterative local search process, but it does not improve the clustering accuracy. Ma Qing (2012) developed a new k-meoids clustering algorithm based on granular computing. And it is necessary to improve its clustering accuracy [8].

Therefore, a new k-medoids clustering algorithm that can effectively improve the accuracy is proposed in this paper. It first conducted a density-based clustering to generate several particles. And then select the centers of the k densest particles as the initial centers. Experiment results show that the proposed algorithm has better performance than the k-mdoids algorithm based on granular computing.

2 Traditional k-medoids clustering

K-medoids clustering algorithm is a classical partitioning-based clustering algorithm. It is less sensitive to outliers than k-means clustering because it is based on the most centrally located object in a cluster. The basic idea of k-medoids clustering [9] can be described as follows: It randomly selects k objects in data set as the initial clustering centers, then it assigns each object to the nearest cluster. After each object is assigned to a cluster or marked as noise, the new clustering centers is decided.

Though other measures can be adopted in k-medoids, the Euclidean distance will be used as a dissimilarity measure in the algorithm. The Euclidean distance between object $x(x_1, x_2, ..., x_n)$ and $y(y_1, y_2, ..., y_n)$ is given by:

$$d(x, y) = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2 + ... + |x_n - y_n|^2} . \quad (1)$$

The objective function that evaluates the clustering effect can be given by:

$$E_{sum} = \sum_{t=1}^{k} \sum_{x_j \in s_i} d(x_j, c_i), \quad (2)$$

k is the number of clusters that is given by user, c_i is the clustering center, s_i is a cluster that its clustering center is c_i , x_j is the object that is assigned to s_i .

3 K-medoids algorithm based on density

3.1 BASIC CONCEPTS

Definition 1 (Particle Density) Given n objects, $o_1, o_2, ..., o_n$ and it is divided into $\{X_1, X_2, ..., X_m\}$, $o_i \in X_j (1 \le i \le n; 1 \le j \le m)$

$X_j (1 \le j \le m)$ is the cluster. m is the subset number of objects. Then the particle density can be defined as follows:

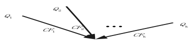
$$pd(X_j) = |X_j| / n . \quad (3)$$

$|X_j|$ is the cardinality of the set X_j

Definition 2 (Eps-neighborhood). Given a data set D and radius (Eps). The eps-neighborhood of a point p can be defined as follows [10-11]:

$$N_{Eps}(p) = \{q \mid q \in D, dist(p,q) \leq Eps\}$$

$$dist(p,q) = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n|. \quad (4)$$



$, q = (y_1, y_2, ..., y_n)$

Definition 3 (Core object) Given an object p, a minimum number of other objects (MinPts) [9]. Then p is a core object if $N_{Eps} > MinPts$

Definition 4 (The clustering Center). Given n objects, $o_1, o_2, ..., o_n$ and it is divided into $\{X_1, X_2, ..., X_m\}$, suppose that $X_j = \{x_{j1}, x_{j2}, ... x_{jt}\}$, X_j is called a cluster, then the clustering center of X_j is defined as:

$$m_j = \{x_{jp} \mid \min_{p=1}^{t} | x_{jp} - \frac{1}{t}\sum_{p=1}^{t} x_{jp} |\}$$

3.2 IMPROVED K-MEDOIDS ALGORITHM

In order to overcome the shortcomings of the traditional k-medoids algorithm and select the effective initial centers objects, an improved algorithm which is based on the density is proposed. The main idea of the algorithm is to divide the data set into several particles based on density and then select k densest particles. Here we calculate the distance between two objects based on formula(4). Finally, select the initial centers from these k particles according to Equation (3). The proposed k-medoids algorithm can be described as follows:

Algorithm 1: The selection of the initial centers.

Input: dataset D, $MinPts$, Eps

Output: k initial centers

Step1.1:For $p \in D$ do

{ if *p* is already included in a cluster

then continue

else

{if *p* is core object

then find the $N_{Eps}(p)$

else mark the object *p* is treated

}

}

end for

Step1.2: merge all clusters that have a common core object

Step 1.3: select *k* densest clusters and calculate their centers according to definition 4.

Algorithm 2: Assign object to centers

Input: *k* initial centers, data set *D*

Output: *k* clusters

Repeat

{Step2.1: for $p \in D$

{calculate the distance between *p* and the centers according to formula(1), then the object *P* is assigned to the nearest cluster.}

end for

step2.2: calculate the current cost of each cluster $E_i = \sum_{x_j \in s_i} d(x_j, c_i)$ and the total cost $E_{sum} = \sum_{t=1}^{k} \sum_{x_j \in s_i} d(x_j, c_i)$

Step 2.3: compute the new center (o_i) of every cluster according to definition 4, and calculate its cost $E_{temp} = \sum_{x_j \in w_i} d(x_j, o_i)$.

Step 2.4: if $E_{temp} < E_i$, then replace the old centers with the new centers

Step 2.5: calculate the total cost $E_{sum\_new} = \sum_{i=1}^{k} \sum_{x_j \in w_i} d(x_j, o_i)$, w_i is a cluster that its clustering center is o_i

}

Until $E_{sum} = E_{sum\_new}$

4 Experimental results analysis

4.1 TESTING ENVIRONMENT

Software environment: Windows xp, eclipse3.7.0, Jdk1.7.0_21.

Hardware environment: CPU: AMD A10-5800K (Quad core), Memory: 2G.

Programming language: Java.

4.2 TESTING DATA

In order to test the validity of the proposed algorithm, the method is applied to five UCI data sets. Their true classes are known. Then the data in the five data sets are reclassified with the improved k-medoids algorithm. The accuracy is the proportion of objects that are correctly grouped. The five data sets are shown in the table 1.

TABLE 1 Data sets

Data set	Number of distance	Number of attribute	Number of class
Iris	150	4	3
Wine	178	14	3
Soybean	47	35	4
Haberman	306	4	2
Ionosphere	351	33	2

4.3 TESTING RESULTS

In the step 1 of the proposed algorithm, it requires users to enter *Eps* and *MinPts*. However it is difficult to determine the precise values of *MinPts* and *Eps*. So we used a probable range for their values [13-15]. The test results are shown in the table2-table 6.

TABLE 2 Iris accuracy

Eps / MinPts	3	5	7	9	11
0.5	92.67%	92.67%	66.67%	66.67%	33.33%
0.7	89.33%	89.33%	66.67%	59.33%	59.33%
0.9	92.67%	66.67%	66.67%	66.67%	66.67%

TABLE 3 Wine accuracy

Eps / MinPts	3	6	9	12
40	70.79%	70.79%	70.79%	70.79%
45	61.8%	69.1%	70.79%	59.33%
50	61.8%	61.8%	70.79%	66.67%

TABLE 4 Soybean accuracy

Eps / MinPts	2	3	4	5
3	74.47%	46.81%	38.30%	36.17%
4	80.85%	80.85%	72.34%	38.30%
5	59.57%	59.57%	59.57%	59.57%

TABLE 5 Haberman accuracy

Eps / MinPts	2	4	6	8
2	51.63%	53.26%	51.96%	51.63%
4	75.16%	75.49%	75.49%	75.49%
6	73.86%	77.56%	75.49%	75.49%
8	77.12%	75.49%	75.49%	75.49%

TABLE 6 Ionosphere accuracy

Eps / MinPts	2	5	8	11
5	69.94%	69.94%	69.94%	69.94%
7	69.94%	63.20%	63.20%	63.20%
9	69.94%	63.20%	63.20%	63.20%
11	63.76%	63.20%	63.20%	63.20%

The changes of accuracy are shown in Figure 1, Figure 2, Figure 3, Figure 4 and Figure 5.

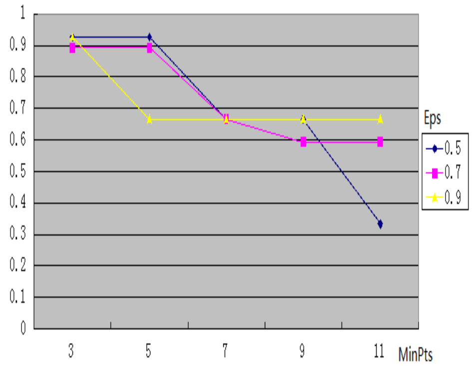


FIGURE 1 Iris

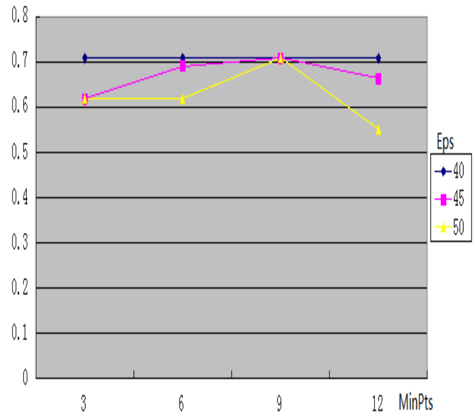


FIGURE 2 Wine

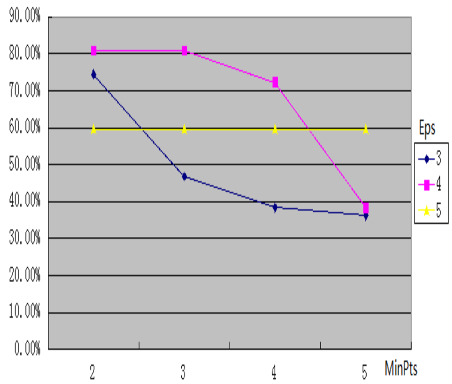


FIGURE 3 Soybean

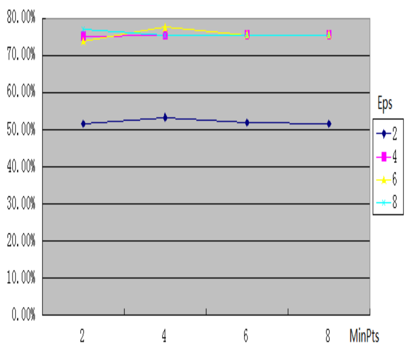


FIGURE 4 Haberman

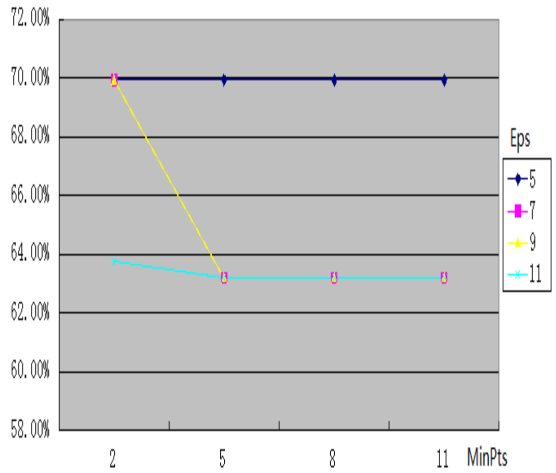


FIGURE 5 Ionosphere

From table 2 to table 6, it can be seen that the best accuracy are 92.67%, 70.79%, 80.85%, 77.56% and 69.94%. As figure 1, figure 2, figure 3, figure 4 and figure 5, we can find that accuracy decreases along with MinPts increasing. On these data sets clustering experiments, various clustering algorithms were compared with this algorithm. The results of the comparison are shown in table 7. Where SFK denotes simple and fast K-medoids clustering algorithm in article [6] and GCK denotes new k-medoids clustering algorithm based on granular computing in article [8] and PK denotes the improved K-medoids clustering algorithm in this article.

TABLE 7 Comparison of clustering algorithms

Data sets	SFK	GCK	IK
Iris	89.33%	90.00%	92.67%
Wine	70.79%	70.79%	70.79%
Soybean	72.34%	80.85%	80.85%
Haberman	73.20%	74.51%	77.56%
Ionosphere	60.11%	60.11%	69.94%

From table 7, it can be seen that the improved k-medoids clustering algorithm in this article has the highest clustering accuracy. It indicates that the initial cluster centers have a greater impact on the clustering accuracy.

5 Conclusion

In this paper, a new improved k-medoids clustering algorithm based on density is proposed. It first conducted a density-based clustering to generate several particles. And then select the centers of the k densest particles as the initial centers. The proposed algorithm is applied in several UCI data sets, and the experimental result shows that the proposed algorithm has better performance than the fast k-medoids clustering algorithm and the k-medoids clustering algorithm based on granular computing. However, although the proposed algorithm is capable of accurately grouping the data set, it is difficult to determine the value of Eps and $MinPts$. So, in future this can be the major area of research.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (No. 81260578 and No. 81460769 and No.61163023 and No.61070139)

References

[1] Pardeshi B, Toshniwal D 2010 Improved K-medoids clustering based on cluster validity index and object density *Advance Computing Conference (IACC), 2010 IEEE 2nd International. IEEE* 379-84

[2] Ren X, Li X, Liu X 2013 *Computer Modeling and New Technologies* **17**(4) 66-73

[3] Li Peizhe, Jian Lirong, Zhang Kun, Pei Shanshan 2014 *Computer Modelling and New Technologies* **18**(2) 274-80

[4] Shi Shaobo, Yue Qi, Wang Qin 2014 *Computer Modelling and New Technologies* **18**(2) 135-42

[5] Ren X, Li X, Liu X 2013 *Computer Modelling and New Technologies* **17**(4) 66-73

[6] Park H S, Jun C H 2009 A simple and fast algorithm for K-medoids clustering *Expert Systems with Applications* **36**(2) 3336-41

[7] Jinglan W, Wenxing Z 2004 An iterated local search algorithm for k-medoids clustering *Journal of computer research and development* **41** 246-52

[8] Qing M, Juanying X 2012 New k-medoids clustering algorithm based on granular computing *Journal of Computer Applications* **32**(7) 1973-7

[9] Yicheng J, Xia L, Qi Z 2011 *Principles and practice of data mining* Beijing: Publishing House of Electronics Industry 120-1

[10] He Y, Tan H, Luo W, et al. 2011 Mr-dbscan: An efficient parallel density-based clustering algorithm using MapReduce *Parallel and Distributed Systems (ICPADS), 2011 IEEE 17th International Conference on. IEEE* 473-80

[11] Zhou H, Wang P, Li H 2012 Research on Adaptive Parameters Determination in DBSCAN Algorithm *Journal of Information & Computational Science* **9**(7) 1967-73

[12] Li L, Xi Y 2011 Research on clustering algorithm and its parallelization strategy *Computational and Information Sciences (ICCIS), 2011 International Conference on. IEEE* 325-8

[13] Birant D, Kut A 2007 ST-DBSCAN: An algorithm for clustering spatial–temporal data *Data & Knowledge Engineering* **60**(1) 208-21

[14] Yang C, Wang F, Huang B 2009 Internet traffic classification using dbscan *Information Engineering ICIE'09. WASE International Conference on. IEEE, 2009* **2** 163-6

[15] Zhen J, Gui W Y 2010 Genetic Clustering Algorithm Based on Dynamic Granularity *Computing, Control and Industrial Engineering(CCIE), 2010 International Conference on. IEEE* **1** 431-4

Authors

Liu Ping, 1973.5, Jiangxi province, China

Current position, grades: The lecturer
University studies: Nanchang University.
Scientific interest: rough sets, GrC, and knowledge discovery
Publications: 2
Experience: She received her master's degree in Engineering from Nanchang University in 2005

Zhou Hao, 1990.5, Jiangxi province, China

Current position, grades: Graduate students
University studies: Nanchang University.
Scientific interest: rough sets, GrC, and data mining

Yang Junping, 1973.12, Jiangxi province, China

Current position, grades: Associate Professor, associate chief technician
University studies: Jiangxi University of Traditional Chinese Medicine
Scientific interest: Cellular immunity of integrated traditional Chinese and Western Medicine , The etiology and pathogenesis and knowledge discovery
Publications: 2
Experience: He graduated from Jiangxi University of Traditional Chinese Medicine in 2004, master, associate chief technician, is mainly engaged in the research of traditional Chinese medicine and Immunology.

Taorong Qiu, 1964.12, Jiangxi province, China

Current position, grades: professor, Doctor
University studies: Nanchang University.
Scientific interest: rough sets, granular computing, Intelligence information Processing and knowledge discovery.
Publications: 4
Experience: Taorong Qiu received the M.S. degree in Nanjing University of Technology, in 1991 and the Ph.D degree in computer application technology from Beijing Jiaotong University, Beijing, China, in 2009. Currently, he serves as a professor with the Department of Computer, Nanchang University.