IMPROVED FUZZY C-MEANS CLUSTERING ALGORITHM BASED ON SAMPLE DENSITY

[1]HUIJING YANG, [2]DANDAN HAN, [3]FAN YU
[1] School of Software, Harbin University of Science and Technology, Harbin 150040, Heilongjing, China
[2] School of Software, Harbin University of Science and Technology, Harbin 150040, Heilongjing, China
[3] Financial Department, Heilongjiang University, Harbin 150080, Heilongjing, China

ABSTRACT

Fuzzy clustering techniques, especially fuzzy c-means (FCM) clustering algorithm, have been widely used in automated image segmentation. The performance of the FCM algorithm depends on the selection of initial cluster center and/or the initial memberships value. if a good initial cluster center that is close to the actual final cluster center can be found. the FCM algorithm will converge very quickly and the processing time can be drastically reduced. In the paper for the problem that fuzzy c-means clustering algorithm is sensitive to the initial cluster centers, propose a method of selecting initial cluster centers based on sample density. At the end do experimental analysis and verification of the proposed key technologies. The results show that the proposed algorithm is superior to the FCM algorithms.

Keywords: *Hard C-Means Clustering; Fuzzy C-Means Clustering; Sample Density*

1. INTRODUCTION

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters [1]. It is a branch in statistical multivariate analysis and unsupervised machine learning, which has extensive applications in various domains, including financial fraud, image processing, medical diagnosis, and text categorization [2,3].

Many clustering strategies have been used, such as the hard clustering scheme and the fuzzy clustering scheme, each of which has its own special characteristics [4]. In real applications there is very often no sharp boundary between clusters so that fuzzy clustering is often better suited for the data. Among the fuzzy clustering methods, fuzzy c-means(FCM) algorithm is the most popular method because it has robust characteristics for ambiguity and can retain much more information than hard segmentation methods [5]. Although the conventional FCM algorithm works well but it has still many problems, such as the determination of the number of clusters or how to initialize the cluster centers, and so on.

In this paper, we study on the distribution of the data set, and introduce a definition of sample density as the representation of the inherent character of the data set. A regulatory factor based on sample density is proposed to correct the selection of initial cluster center in the conventional FCM. The data sets of experiments using Iris data are operated. Comparing with some existing methods, the proposed algorithm shows the better performance.

2. RELATED WORKS

Fuzzy c-means (FCM) are extensions of hard c-means (HCM). FCM has been shown to have better performance than HCM. FCM has become the most well-known and powerful method in cluster analysis [6].

2.1 Hard C-Means

Hard c-means algorithm is a typical dynamic clustering algorithm, in the algorithm, each cluster is represented the average of objects. First of all, Randomly choose K vectors as the initial cluster centers. Secondly, the rest data is classified into one of k class by principle of division of the minimum distance one by one, then generate a new cluster [7]. Finally, calculate the new various class center after re-classification.

Given a data set of input vectors x={x_1,x_2,..., x_n}. It has n data points. Where xi={x_{i1},x_{i2},..., x_{im}} is the data object with m dimension variables. the data set is divided into k classes: W_1,W_2,..., W_K. The guidelines function of the sum of squares error is defined as the objective function. the objective function and class center update formula are defined as following(1) and (2):

$$E = \sum_{j=1}^{k} \sum_{x_i \in W_j} \left\| x_i - c_j \right\| \qquad (1)$$

$$c_j = \frac{1}{n_j} \sum_{x_i \in W_j} x_i \qquad (2)$$

where, c_j(j=1,2,..., k) is the cluster center of the class W_j and n_j is the number of sample in class W_j. obviously, the objective function is a function of the samples and cluster centers, it is trying to figure out a class minimum objective function value, making the final clustering results compactly and independently. the value of the objective function E depends on the cluster center. If the higher the objective function value, the greater the error, the worse the clustering effect. Therefore, we should seek to make the minimum value of clustering results [8].

The algorithm can be written as follows.

Input: the data set of X, the number of cluster K, iteration counter t and the iteration termination condition .

Output: meeting iteration termination condition cluster centers and the number of iterations.

step1: Initialize the cluster centers $C_q(t)$(q=1,2,…,k). This is typically achieved by randomly selecting k points from among all of the data points. Where t is the number of iteration. in the first time t=1.

step2: Compute the distance $d(x_i, C_q(t))$ from the each rest data to the each cluster centers, then put x_i in the class that it has shortest distance from x_i to the cluster center.

step3: Update the cluster center with the equation as (3), then compute the objective function E1 using equation(1).

$$C_q(t+1) = \frac{\sum_{i=1}^{n_q} X_i}{n_q} \qquad (3)$$

where, n_q is the number of data in the q class, q=1,2,..., k.

step4: reallocate xi: if xi is the member of W_q and have the relation of (4), then allocate x_i to W_p class. to compute the objective function E_2 using equation(1).

$$C_q(t+1) = \frac{\sum_{i=1}^{n_q} X_i}{n_q} \qquad (4)$$

step5: if $|E_2-E_1|<$,then stop iteration, go to step6; otherwise go to step3.

step6: close algorithm, output the cluster centers and the number of iteration.

2.2 Fuzzy C-Means Clustering

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method was developed by Dunn in 1973 and improved by Bezdek in 1981 and it is frequently used in pattern recognition [9].

The basic idea of the fuzzy C-means algorithm is described as follows. Consider a set of vectors $x=\{x_1,x_2,..., x_n\}$, $x_i \in W_j$ where n is the number of sample points and j is the dimension of pattern vectors. The FCM algorithm focuses on minimizing the value of an objective function. the objective function measures the quality of the partitioning that divides a data set into C clusters [10].

The FCM algorithm measures the quality of the partitioning by comparing the distance from pattern x_i to the current candidate cluster center w_j with the distance from patter x_i to other candidate cluster centers. the objective function is an optimization function that calculates the weighted within-group sum of squared errors as equation (5).

$$J_m(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} d_{ij}^{2}(x_i, v_j) \qquad (5)$$

Where: n is the number of patterns in x; c is the number of clusters; U is the membership function matrix; the elements of U are u_{ij}; u_{ij} is the value of the membership function of the ith pattern belonging to the jth cluster; d_{ij} is the distance from x_i to w_j; V is the cluster center vector; m is the fuzzy factor which is a weighting exponent on each fuzzy membership, is any real number greater than 1 to control fuzziness or amount of clusters overlap.

The FCM algorithm focused on minimizing J_m, subject to the following constrains (6) on U :

$$u_{ij} \in [0,1], \quad \sum_{i=1}^{n} u_{ij} = 1, \quad \sum_{j=1}^{n} u_{ij} < n \qquad (6)$$

According the analysis the algorithm can be written as follows.

step1: Given a fixed number C, initial the cluster center matrix w0 by using a random generator from the original dataset. Record the cluster centers set t=0, m=2, and decide ε where ε is a small positive constant.

step2: Initialize the membership matrix U_0 by using formula (7)

$$u_{ij}(t)=\frac{1}{\sum_{k=1}^{c}(d_{ij}(t)/d_{kj}(t))^{2/(m-1)}} \qquad (7)$$

If $d_{ij}(t) = 0$， then $u_{ij} = 1$ and $u_{rj} = 0$ ($r \neq j$).

step3: Increase t by one. Compute the new cluster center matrix W_i using formula (8)

$$V_i(t+1)=\frac{\sum_{j=1}^{n}u_{ij}^{m}(t)x_j}{\sum_{j=1}^{n}u_{ij}^{m}(t)} \qquad (8)$$

step4: Compute the new membership matrix U_i by using formula (7)

step5: If $\max\{|u_{ij}(t)-u_{ij}(t-1)|\}\le\varepsilon$ then stop, otherwise go to step3.

2.3 Influential Factors of The FCM

FCM algorithm is very simple, but from the FCM clustering algorithm steps, you can see a few important factors that influence the FCM clustering effect.

(1) the selection of initial cluster center: the performance of the FCM algorithm depends on the selection of initial cluster center and/or the initial membership value. if a good initial cluster center that is close to the actual final cluster center can be found. the FCM algorithm will converge very quickly and the processing time can be drastically reduced.

(2) The cluster number: the cluster number c is determined in advance in classic FCM. Reasonable and best number of clusters is one of the key steps to decide that the result of FCM algorithm is good or bad . Typically, c is unknown, the range of values is in c (1, n]. Obviously, different values of c, clustering results must be different.

(3) Fuzzy factor m: it is a weighting exponent on each fuzzy membership, is any real number greater than 1 to control fuzziness or amount of clusters overlap. if m=1,then the FCM clustering algorithm degenerates into HCM; if m=∞, then FCM clustering algorithm is then lost clustering characteristics

3. IMPROVED FCM CLUSTERING ALGORITHM

In this paper, we study on the distribution of the data set, and introduce a definition of sample density as the representation of the inherent character of the data set. and propose a method of selecting initial cluster centers based on sample density. It can solved the problem of the center initialization of the FCM algorithm and the number of cluster, improved the clustering speed, accuracy and Stability.

3.1 The Principle of The Selection of Initial Cluster Center

Generally, the more intensive the sample points are in a region , the greater Sample distribution density it that has a sample point as a center has. Therefore, the choice of the initial class center should meet two conditions:

(1) The higher density of the cluster center sample point is, the better the degree of cluster is;

(2) The greater the distance of different cluster center is the better the degree of cluster is;

If you can find samples that meet the above conditions as the initial class center, you can avoid algorithm unsatisfactory by initialization generated.

1. Measure of the sample density

Set $D_{ij}=\|x_i-x_j\|$, it represents the distance between the two sample points. Calculate the Average distance from the sample point xi to other sample points using the formula (9).

$$\bar{D}_i=\frac{\sum_{k=1,i\neq k}^{N}D_{ik}}{N-1} \qquad (9)$$

Calculate the density of the sample xi using the formula (10). where, N is the number of all the sample points; t_i is the number of sample points that its distance is shorter than the Average distance; and $D_{i\text{-max}}$ is the biggest distance in the t_i.

$$P_i=\frac{t_i}{D_{i-\max}} \qquad (10)$$

we can find that The more points of the sample points x_i is around, the bigger value of the sample density is. And vice versa.

2. the selection of initial cluster center

Set the sample points that has the maximum sample density as the cluster center, delete the cluster center and the sample points that its distance to cluster center smaller than $D_{i\text{-max}}$ from the data set. Then, Iterate the above process until all the sample points are deleted from the data set or the number of the rest sample points is smaller than a default value. Now the cluster centers and its number has been generated. we can input them to the FCM algorithm to cluster. because the cluster centers and its number was generate according to the sample density, using the method in the cluster we can reduce noise and accelerate the cluster speed.

3.2 Algorithm Steps of The Selection of Initial Cluster Center in New FCM

Input: the data set of $x=\{x_1,x_2,..., x_n\}$, Where $x_i=\{x_{i1},x_{i2},..., x_{im}\}(i=1,2,...,n)$.

Output: the cluster centers of sample $Y=\{y_1,y_2,...,y_c\}$ and the number of the cluster.

Step1: Default initialization, set c=0,and Set sample end threshold value k=3.

Step2: Calculate the distance matrix D:

$$D=\begin{bmatrix} 0 & & & \\ D_{21} & 0 & & \\ ... & ... & ... & \\ D_{n1} & D_{n2} & ... & 0 \end{bmatrix}$$

Step3: Calculate the number t_i and radius $D_{i\text{-}max}$ using the formula (9) ,then calculate the density of each sample $P_i(i=1,2,...,n)$ using the formula (10).

Step4: Select the sample point as cluster center that it has the maximum sample density. $y_c=\max(P_i)(i=1,2,...,n)$.

Step5: If $D_{ij}\leq D_{i\text{-}max}$, then set $D_{ij}=0$.

Step6: Set $n=n-t_i$.

Step7: If $n\leq k$, then the algorithm ends and output Y and c, otherwise, set c=c+1, delete the sample, update the data set. then go to step 2.

4. EXPERIMENTS AND RESULTS ANALYSIS

Experimental platform for MATLAB2007, VC++ 2008. The experiments use classical IRIS as the test data set to verify the validity of the algorithm. It is internationally recognized the typical supervised clustering effect good or bad data. All experiments take fuzzy factor m=2 and take ε=10-4 .

Iris data set contains 3 classes of 50 instances each, and each data contains four kinds attributes, where each class refers to a type of iris plant [11]. One class is linearly separable from the other 2; the latter are not linearly separable from each other. The original location of IRIS data set's centers are shown in Table 1.

Table 1: Iris Data Set's Real Centers

Name of class	Clustering centers
Class 1	5.00,3.42,1.46,0.24
Class 2	5.93,2.77,4.26,1.32
Class 3	6.58,2.97,5.55,2.02

To verify the validity of the algorithm that the HCM ,the FCM and new FCM algorithms has been applied in a real case for IRIS. The results of experiments are shown in Table II.

Table 2: The Comparison Of New Algorithm And FCM Algorithm

algorithm	Selection of initial value of cluster center	Clustering centers	Number of iteration	Rate of correct

FCM	random	v1=(5.0036,3.4030,1.4850,0.2515)
		v2=(5.8892,2.7612,4.3643,1.3974)
		v3=(6.7751,3.0524,5.6469,2.0536)
			27	89.33

New FCM	Sample density	v1=(5.0028,3.4033,1.4850,,0.2514)
		v2=(5.8892,2.7607,4.3633,1.3979)
		v3=(6.7754,3.0530,5.6466,2.0539)
			13	89.33

From the above table we can find that the proposed algorithm is superior to the FCM algorithms. The user must be determine the number of clusters in the classic FCM before the algorithm was run, hence the experience of user will affect the cluster results. It will result in increasing the number of errors distributed over and the number of iterations.

The improved algorithm based on sample density can be determined cluster centers and the number of clusters by one calculation, , does not require multiple iterations optimal selection process.

The experiments show that the improved algorithm solved the problem of the center initialization of the FCM algorithm, improved the clustering speed.

5. CONCLUSION

In this paper, we proposed a novel method for efficient clustering that is better than FCM algorithm. An sample density conception based on distance is defined for identifying the center and the number of clusters. Iris data set were used to compare the performance of FCM and the new algorithms. Experimental results show that we reduce the computation cost and improve the performance by finding a good set of initial cluster centers.

REFERENCES:

[1] Graves Daniel, Pedrycz Witold, "Kernel-based fuzzy clustering and fuzzy clustering:A comparative experimental study", *Fuzzy Sets and Systems*, Vol. 161, No. 4, 2010, pp. 522-543.

[2] Pierpaolo D'Urso, Paolo Giordani, "A Robust Fuzzy k-means Clustering Model for Interval Valued Data", *Compurational Statistics*, Vol. 21, No. 2, 2006, pp. 251-269.

[3] Pierpaolo D'Urso, Paolo Giordani, "A Weighted Fuzzy C-means Clustering Model for Fuzzy Data", *Computational Statistics&Data Analysis*, Vol. 50, No. 6, 2006, pp. 1496-1523.

[4] Kuo-Lung Wua, Jian Yub, Miin-Shen Yanga, "A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests", *Pattern Recognition Letters*, Vol. 26, No. 5, 2005, pp. 639-652.

[5] Naresh S. Iyer1, Abraham Kandelb, Moti Schneiderb, "Feature-based fuzzy classification for interpretation of mammograms", *Fuzzy Sets and Systems*, Volume. 114, No. 2, 2000, pp. 271-280.

[6] S. M. Xiang, F. P. Nie, C. S. Zhang, "Learning a Mahalanobis distance metric for data clusteringand classi_cation", *Pattern Recognition*, Vol. 41, No. 12, 2008, pp. 3600-3612.

[7] D.A. Clausi, "K-means iterative Fisher(KIF) unsupervised clustering algorithm applied to image texture segmentation. Pattern Recognition", *Pattern Recognition*, Vol. 35, No.9, 2002, pp. 1959-1972.

[8] Evans A. N., Liu X. U., "A morphological gradient approach to color edge detection, IEEE Transactions on Image Processing", *IEEE Transactions on Image Processing*, Vol. 15, No.6, 2006, pp. 1454-1463.

[9] J. Wang, S. T. Wang, "Double indices FCM algorithm based on hybrid distance metric learning. Journal of Software", *Journal of Software*, Vol. 21, No. 8, 2010, pp.1878-1888.

[10] H. Timm, C. Borgelt, C. Doring, and R. Kruse, "An extension to possibilistic fuzzy cluster analysis", *Fuzzy Sets and Systems*, Vol. 147, No. 1, 2004, pp. 3-16.

[11] Daniel Graves, Witold Pedrycz, "Kernel-based fuzzy clustering and fuzzy clustering: A comparativeexperimental study", *Fuzzy Sets and Systems*, Vol.161, No.4, 2010, pp.522 -543.