Attribute Weighted Optimization of Fuzzy C-Means Clustering Algorithm

Ruijuan Li

College of Mathematics and Information Science, Langfang Teacher's College, Langfang 065000, Hebei, China

Chuiwei Lu

Computer Institute, Hubei Polytechnic University, Huangshi 435003, Hubei, China

Abstract

According to the standard fuzzy C-means clustering algorithm performed poor in the clustering effect during the clustering process. This paper presents an objective function optimization based on the attribute weighted and the objective function optimization. Firstly, use a little prior knowledge as the labeled sample. These calibrated samples information are used as the prior knowledge, and then use the self-creating method and the statistical characteristics based on the data to optimize the attribute weights in FCM algorithm, and then introduce the kernel function to improve the search ability of the fuzzy C-means clustering algorithm, simplify both the clustering center and membership matrix with lagrange multiplier approach. The simulation experiment shows that, contrast to the original algorithm and K-means clustering algorithm, the fuzzy C-means clustering algorithm optimized by the attribute weighted method presented in this paper has a better clustering effect.

Key words: FUZZY C-MEANS CLUSTERING ALGORITHM, CLUSTER ANALYSIS, ATTRIBUTE WEIGHTS, SELF-CREATING METHOD, LAGRANGE

1. Introduction

Cluster analysis is a powerful information process method which can discover the useful information. The traditional cluster analysis is a hard division method, it divides the object to be identification into different kinds [1]. The cluster analysis called fuzzy cluster analysis is based on the fuzzy mathematic, and the fuzzy cluster analysis introduced the concept of fuzzy mathematic is a multivariate statistical analysis method to research "like attract like", which is both a young branch of numerical taxonomy and an important branch of Unsupervised Pattern Recognizing Method [2]. By cluster analysis, we can extract the interesting information, regulation and others from the database, and observe or browse from different aspects. The discovered knowledge could be used in the decision, process control, information management and query processing.

Fuzzy clustering algorithm has always been the hot-spot of the research in the academia because of the widely application in real life. Zhang presented the fuzzy clustering algorithm based on the partition, which divided the dataset contained N data objects in K kinds or K clusters

and had a favorable robust performance and high accuracy of the clustering results. The requirements are as follows: (1) K should less than N and more than 1. (2) Each object in the cluster is not absolutely belonged to the only cluster, but possibly belong to the others in divided K clusters. The relation between each object and the cluster is fuzzy, not rigidity [3]. Chen presented an algorithm based on the level which had a good performance towards the cluster partition of the dataset. The algorithms are divided into two types: condensation and division. The search direction of condensation level algorithm is from down to up, each data object in the dataset will be considered as one kind, and then these kinds will combined slowly until to meet a pre-suppositional condition. The direction of the divisive analysis cluster-algorithm is from up to down, the whole dataset is considered as one kind, and then decomposed according to a certain condition until to decomposed to the needed kinds [4]. Yue presented an algorithm based on the Grid clustering algorithm which could solve the problems of the sensitivity of the algorithm towards the input parameters well. The algorithm transforms the space structure of the cluster dataset in the grid structure, and then carrys on various clustering operation, so that the cluster speed will increase [5]. Peng got a high accuracy cluster result with the fuzzy clustering method based on the knowledge foundation of the rough set. The rough set theory was introduced into the algorithm, considering the interaction of the attributes comprehensively, had a good performance in the dataset composed of the data objects with fuzzy boundaries [6]. Pal presented a fuzzy clustering grid algorithm which solved the problem that FCM has a poor performance in the sample dataset cluster effect of ball distribution [7]. Xu introduced the penalty factor into the FCM algorithm, presented the new algorithm of competitive learning, the new algorithm solved the problem that the cluster number need to be specified manually. [8]Tamika presented a W-K means clustering algorithm which take the whole dataset as one kind, then divided the dataset according to the attributes of the kinds until to meet the needed number of the kinds. This algorithm has a favorable performance in Sparse high-dimensional dataset[9]. Yao introduced the Mathematical Morphology into K-means clustering algorithm so that the cluster effect of algorithm was improved. The improved algorithm was applied successfully in the fish image identification [10].

According to the standard fuzzy C-means clustering algorithm performed poor in the clustering effect during the clustering process. This paper presents an objective function optimization based on the attribute weight and the objective function optimization, carries out the experiment simulation, verify and improve the effectiveness of strategies.

2. Defect analysis of Fuzzy C- means clustering algorithm

More popular and widely used method of fuzzy clustering is fuzzy C- means clustering algorithm, which is based on the basis of hard C-means algorithm.

Hard - Means (HCM) algorithm is the traditional method of partitioning the data set, which can set the cluster to the data of hyperellipsoidal configuration.

$$(\min)J_1 = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij} \left\| x_j - \upsilon_i \right\|^2 \quad (1)$$

Among them, $U = \left[u_{ij} \right]_{c\times n}$ is the cluster centers of HCM partition matrix representing the Euclidean distance.

Specific processes of HCM algorithm is as follows

(1)Initialization: category specified cluster c, $2 \le c \le n$, n is quantity, define iteration stops threshold ε,initialize the clustering centers V^0,define iteration counter $b = 0$;

(2)Calculate or update the partition matrix $U^b = \left[u_{ij} \right]$ according to the formula(2);

$$u_{ij} = \begin{cases} 1 & \left\| v_i - x_j \right\| = \min\limits_{1\le k\le C} \left\{ \left\| v_k - x_j \right\| \right\} \\ 0 & other \end{cases} \quad (2)$$

(3) update the cluster centers $U^b = \left[u_{ij} \right]$ according to the formula (3) ;

$$v_i = \frac{\sum_{j=1}^{n} u_{ij} x_j}{\sum_{j=1}^{n} u_{ij}}, 1 \le i \le c \quad (3)$$

(4) If $\left\| V^b - V^{b+1} \right\| \le \varepsilon$,interrupt the algorithm and output U and V,or define $b = b + 1$,then perform(2).

The FCM algorithm run the HCM algorithm to vague specifications by the fuzzy division rule. By squaring the distance between each sample and between each cluster center, add the class square J_1 to the weight and error objective function J_2 of square error of classes and functions.

$$(\min)J_2 = \sum_{i=1}^{C}\sum_{j=1}^{N} {u_{ik}}^2 \left\| x_k - \upsilon_i \right\|^2 \quad (4)$$

It is comprehensive described as follows:

$$(\min)J_m = \sum_{i=1}^{C}\sum_{k=1}^{N} {u_{ik}}^m \left\| x_k - \upsilon_i \right\|^2 \quad (5)$$

Among them , $m \in [1, \infty)$ is weighted index. Specific processes of FCM algorithm is as follows:

(1)Initialization: category specified cluster c, $2 \le c \le n$, n is quantity, define iteration stops threshold ε,initialize the clustering centers V^0,define iteration counter $b = 0$;

(2)update the cluster centers $U^b = \left[u_{ij}\right]$ according to the formula (6) ;

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{\left\|x_j - v_i\right\|^2}{\left\|x_j - v_k\right\|^2}\right)^{1/(m-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \le n \quad (6)$$

(3) update the cluster centers V^{b+1} according to the formula (7);

$$v_i = \frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m}, 1 \le i \le c \quad (7)$$

(4) If $\left\|V^b - V^{b+1}\right\| \le \varepsilon$,interrupt the algorithm and output U and V,or define $b = b+1$,then perform(2).

In order to analyze the FCM clustering results, use K-means clustering algorithm to analyze comparatively .

K-means algorithm first randomly k points as initial cluster centers, and then calculate the distance between each data object to the cluster centers, and return the data object to the class which is the nearest cluster center located; calculate the adjusted new cluster centers, if two adjacent cluster centers have no change, then indicate the end of the data object adjustment, an Clustering criterion function J_c is converged. The algorithm framework as follows:

(1)Given the data set with size n.Let be $I = 1$. Selecting k initial cluster centers.

$$Z_j(I), j = 1, 2, 3, ..., k \quad (8)$$

(2)Calculate the distance of each data object and cluster centers

$$D(x_i, Z_j(I)), i = 1, 2, 3, ..., n, j = 1, 2, 3, ..., k \quad (9)$$

If satisfy

$$D(x_i, Z_j(I)) = \min\left\{D(x_i, Z_j(I)), i = 1, 2, 3, ..., n\right\} \quad (10)$$

Then $x_i \in w_k$.

(3)Calculate k new cluster centers

$$Z_j(I+1) = \frac{1}{n}\sum_{i=1}^{n_j} x_i^{(j)}, j = 1, 2, 3, ..., k \quad (11)$$

(4)Judgment: if $Z_j(I+1) \ne Z_j(I)$, $j = 1, 2, 3, ..., k$,then $I = I+1$,return to(2); else, the algorithm ends.

Select a set of standard data IRIS data set as a test sample, then simulate the fuzzy C- means clustering algorithm K-means clustering algorithm , and the trend of objective function with the number of iterations is shown as the Figure 1.

Figure 1 shows two kinds of objective function value change curve with clustering algorithm iterations. With the increase of the objective function value of iterations can meet expectations, but the convergence speed fuzzy C-means algorithm is much better than K-means clustering algorithm.

The influence of the noise element inside the data collection inside to the entire cluster division process is too strong .Now a lot of noise in algorithms is not processed , so the noise affects the whole partition data collection. Or the noise processing results are not so satisfactory, and data processing of noise data is too complex or the influence of the noise element is not reduced substantively.

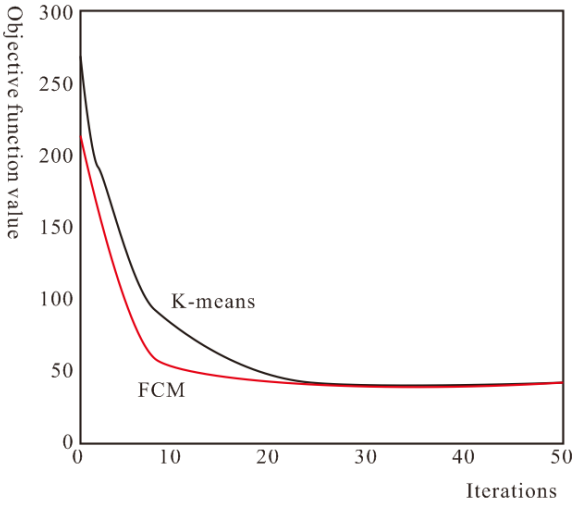

Figure 1. The simulation results of two kinds of Comparative clustering algorithm

3. Attribute weighted optimization of fuzzy C- means clustering algorithm

3.1 Optimization of the weight value based on the self-generation method

This paper select a small amount of a priori knowledge as the identification samples, which is a small part of sample calibrated from the whole sample in advance. Then use these sample information which has been calibrated as prior knowledge and self-generating method based on the statistical characteristics of the data, to optimize the FCM algorithm.

For a sample $X = \{x_1, x_2, ..., x_n\} \subset R^S$ set to be classified, and $x_j = \{x_{j1}, x_{j2}, ..., x_{jm}\}^T$,do B times sampling,generating $x^b = \{x_1^b, x_2^b, ..., x_n^b\}$ sand $x_j^b = \{x_{j1}^b, x_{j2}^b, ..., x_{jm}^b\}^T$. And then introducing the concept of precision, the standard deviation of the individual samples ,the relative standard deviation and the weight of each sample feature are calculated by the resulting sample for each sample

using equation (12) to (14) :

$$S_k^b = \sqrt{\frac{\sum_{j=1}^{n}(X_{jk}^b - \bar{X}_k^b)^2}{n-1}} \tag{12}$$

Among them, $\bar{X}_k^b = \frac{1}{n}\sum_{j=1}^{n} X_{jk}^b$, $k = 1,2,...,m$.

$$RSD_k^b = \frac{S_k^b}{X_k^b} \tag{13}$$

$$W_k^b = \frac{RSD_k^b}{\sum_{k=1}^{m} RSD_k^b} \tag{14}$$

Experiments show that when the number of samples self-generating method is sufficiently large ,the original sample set to be classified which concentrate the weights of the characteristics of each sample number can be represented by the following formula:

$$W_k = \frac{1}{B}\sum_{b=1}^{B} W_k^b \tag{15}$$

For a sample set to be classified, define the cluster center as $p_i = \{p_{i1}, p_{i2},..., p_{im}\}^T$,introducing the prior knowledge optimization strategy based on the self-generation method , the clustering objective function is rewritten as:

$$\begin{aligned} J_m(U,P) &= \sum_{i=1}^{C}\sum_{j=1}^{n}(u_{ij})^m (d_{ij}^w)^2 \\ &+ \alpha\sum_{i=1}^{C}\sum_{j=1}^{n}(u_{ij} - l_{ij}b_j)^m (d_{ij}^w)^2 \end{aligned} \tag{16}$$

Among them, $b = [b_j](j = 1,2,...,N)$ and $b_j = 1$ represent x_j is the sample calibrated in advance ; $L = [l_{ij}](i = 1,2,...,C;\ j = 1,2,...,n)$ represents the membership of calibration sample with respect to each category. Let be d_{ij}^w as the distance between sample $x_j(x_{jk}(k = 1,2,...,m))$ and Cluster centers $p_i(p_{ik}(k = 1,2,...,m))$,And satisfy the following conditions :

$$d_{ij}^w = \sqrt{\sum_{k=1}^{m} W_k (X_{jk} - p_{ik})^2} \tag{17}$$

In this case, the process to calculate the minimum of formula (16) is the process of clustering , as follows:

$$u_{ij} = \frac{1}{1+\alpha}\left\{ \frac{1+\alpha(1-b_j\sum_{i=1}^{c} l_{ij})}{\sum_{f=1}^{c}\left(\frac{d_{ij}}{d_{fj}}\right)^{\frac{2}{m-1}}} + \alpha l_{ij} b_j \right\} \tag{18}$$

$$p_i = \frac{\sum_{j=1}^{n}[u_{ij}^m + \alpha(u_{ij} - l_{ij}b_j)^m]X_j}{\sum_{j=1}^{n}[u_{ij}^m + \alpha(u_{ij} - l_{ij}b_j)^m]} \tag{19}$$

Specific steps of the algorithm is as follows:

(1)Set value of B ,then calculate W_k ;

(2)FCM cluster the sample data first, and the clustering center was taken as a sample marked; Or select mark based on a priori understanding of the data sample and the number is n_r ;

(3)Set C , m , according to the specific circumstances of the study . Initialize the membership matrix $U = [u_{ij}]_{c\times n}$. Set Iterative ε and iterations f ,then calculate the cluster center, and constantly optimize the division matrix ;

(4)Calculate the error. When the error is less than or greater than the number of iterations, end the algorithm.

3.2 Optimization of objective function based on kernel function

Introduced the kernel functions, in order to increase the search ability of optimization C- fuzzy clustering algorithm. Assume cluster centers v_k^Φ on high-dimensional space can find the original image v_k in the original space , then the target function becomes:

$$J = \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^m \left\|\Phi(x_i) - \Phi(v_k)\right\|^2 \tag{20}$$

It comes from Mercer nuclear definition:

$$\begin{aligned} d^2(x_i, v_k) &= \left\|\Phi(x_i) - \Phi(v_k)\right\|^2 \\ &= K(x_i, x_i) + K(v_k, v_k) - 2K(x_i, v_k) \end{aligned} \tag{21}$$

Thus, the objective function which improves the C-fuzzy clustering algorithm becomes

$$J = \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^m K(x_i, x_i) + K(v_k, v_k) - 2K(x_i, v_k) \tag{22}$$

seek the cluster centers of the method and the iterative formula of the membership matrix using lagrange multiplication :

$$u_{ki} = \frac{(K(x_i, x_i) + K(v_k, v_k) - 2K(x_i, v_k))^{\frac{1}{1-m}}}{\sum_{r=1}^{c}(K(x_i, x_i) + K(v_k, v_k) - 2K(x_i, v_k))^{\frac{1}{1-m}}} \tag{23}$$

$$\Phi(v_k) = \frac{\sum_{i=1}^{n} u_{ki}^m \Phi(x_i)}{\sum_{i=1}^{n} u_{ki}^m} \tag{24}$$

Clearly, formula (24) cannot be obtained directly. Using nuclear mapping definition, both sides of the equation is multiplied by $\Phi^T(x_j)$, then

$$K(x_j, v_k) = \frac{\sum_{i=1}^{n} u_{ki}^m K(x_j, x_i)}{\sum_{i=1}^{n} u_{ki}^m} \quad (25)$$

Therefore, improved C-fuzzy clustering algorithm C- process is as follows：

(1)Initialization: Given weighted index m,clustering class number $c(2 \le c \le n)$,set the parameter value of the selected kernel function ε. Initialize the membership matrix $U^{(0)}$,iteration counter $b = 0$.

(2)Calculate $K(x_i, v_k)$ 和 $K(v_k, v_k)$ by formula (16).

(3)Update the membership matrix $U^{(b)}$ by formula (14).

(4)If $\left\|U^{(b)} - U^{(b+1)}\right\| < \varepsilon$,Then stop updating the membership matrix U,else let be U,turn to(2). $\|\cdot\|$ is a kind of appropriate norm.

If the kernel chooses Radial Basis Function($K(x, x) = 1, \forall x \in X$),the objective function of improved C-fuzzy clustering algorithm becomes

$$J = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ki}^m (1 - K(x_i, v_k)) \quad (26)$$

At this point, the cluster center and membership matrix multiplication obtained by the Lagrange reduced to

$$u_{ki} = \frac{(1 - K(x_i, v_k))^{\frac{1}{1-m}}}{\sum_{r=1}^{c} (1 - K(x_i, v_k))^{\frac{1}{1-m}}} \quad (27)$$

$$v_k = \frac{\sum_{i=1}^{n} u_{ki}^m K(x_i, v_k) x_i}{\sum_{i=1}^{n} u_{ki}^m K(x_i, v_k)} \quad (28)$$

So, the improvement strategies of C-fuzzy clustering algorithm proposed in this paper can reduce the complexity of the algorithm is reduced. Analyze the time complexity and space complexity of the improved fuzzy C- means clustering algorithm in this paper, mainly by the selection of the centroid , calculation of the coefficient of variation, update of membership updates and centroid .the time complexity of the center of mass required is $O(n^2)$; the time complexity of computing the coefficient of variation is $O(n)$; the time complexity of update of the membership degree is $O(cmns)$; the time complexity of update of centroid is $O(cmns)$. In time complexity representation, c is the centroid number of the cluster, n is the total number of objects, m is the dimension of dataset, s is the iterations, and the overall time complexity of the algorithm is $O(n^2 + n + 2cmns)$.

The space complexity of algorithm contains dataset, calculation of the distance between the objects, centroid, the coefficient of variation, membership matrix to occupy the data space, among them the space complexity of the dataset is $O(nm)$,The space complexity of the calculation of the distance between the objects is $O(n^2)$,The space complexity of centroid is $O(cm)$,The space complexity of the coefficient of variation is $O(m)$,The space complexity of membership matrix is $O(cn)$.So the overall space complexity of the algorithm is $O(n^2 + nm + cn + cm + m)$.

4. Performance simulation of algorithm

To verify the performance of the improved algorithm presented by this paper, choose a set of standard data IRIS dataset as the test sample, simulate the fuzzy C-means clustering algorithm, K-means clustering algorithm and the improved fuzzy C-means clustering algorithm in this paper. The IRIS dataset is shown in Figureure2.The Figure 3-5 are the fuzzy C-means clustering algorithm, improved clustering fuzzy C-means clustering algorithm and K-means clustering algorithm proposed in this paper respectively.

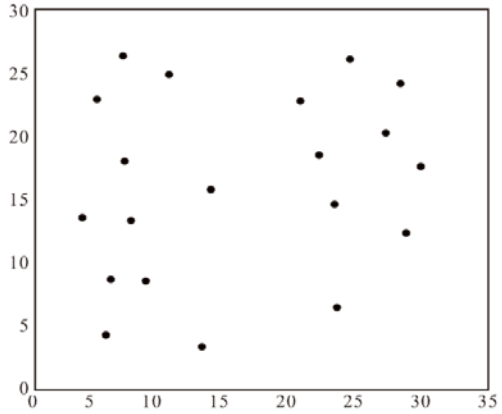

Figure2. IRIS data set

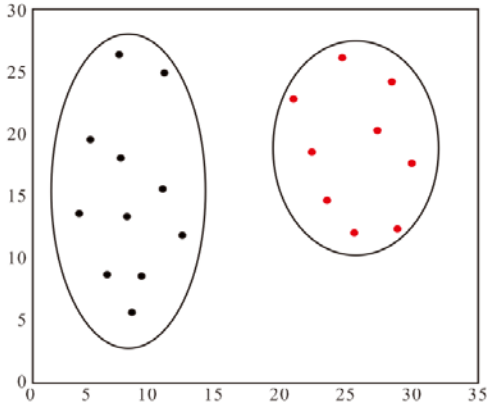

Figure 3. FCM clustering algorithm

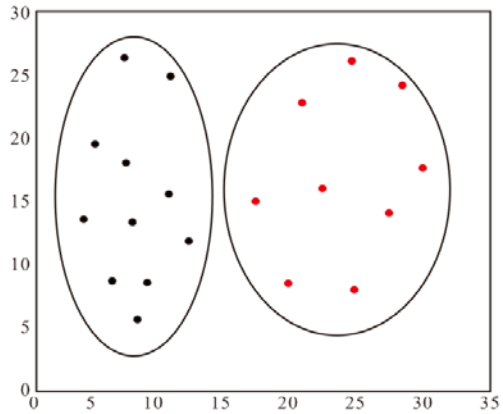

Figure4. K-means clustering algorithm

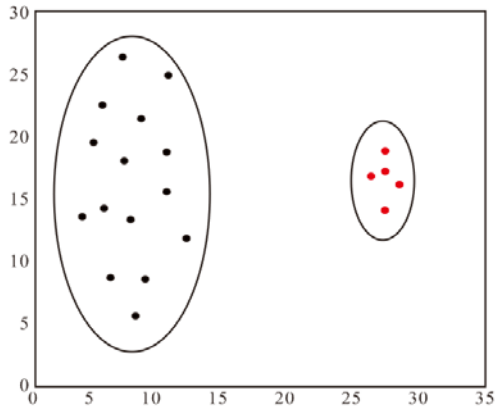

Figure 5. Improved FCM clustering algorithm

From the above results can be seen, contrast to the original algorithm and the K-means clustering algorithm, the attributes weighted optimization fuzzy C-means clustering algorithm proposed in this paper has a better clustering effect.

5. Conclusion

Recently, many algorithms appeared in the data mining direction. Aiming at the explosive growth of the data in society, these algorithms are with the good and the bad, and different kinds of information on the web are increasing. Although been proposed early and developed many related great algorithms by the scientists, FCM is also been regarded as a relatively good algorithm. Aiming at the defect of FCM in cluster, this paper presents a fuzzy C-means algorithm based on the attribute weights and the objective function optimization. The simulation results shows, contrast to the original algorithm and the K-means clustering algorithm, the attributes weighted optimization fuzzy C-means clustering algorithm proposed in this paper has a better clustering effect.

Acknowledgements

This work was supported by Langfang Teacher's College: LSZQ201305, the Key Project of Hubei Provincial Department of Education (D20144403), the Outstanding Youth Science and Technology Innovation Team Project of Hubei Polytechnic University (13xtz10).

References

1. Xi Y.(2015) A type-2 fuzzy C-means Algorithm and its Application. *Journal of Jinan University(Science & Technology)*, 29(5), p.p.372-376.
2. Hu FJ.(2013)A rapid eye-to-hand Coordination Method of Industrial Robots. *Journal of Information and Computational Science*, 10(5), p.p.1489-1496.
3. Xiao MS, Xiao Z.(2015) Interval Type Fuzzifier Parameter Model in Fuzzy C-means Clustering. *Systems Engineering and Electronics*, 37(4), p.p.868-873.
4. Zhu ZY, Wang LM.(2015) Initialization Approach for Fuzzy C-means Algorithm for Color Image Segmentation. *Application Research of Computers*, 32(4), p.p.1257-1260.
5. Liu XN, Lu YN.(2015) A Multi-Objective Image Segmentation Method based on FCM and Discrete Regularization. *Journal of Computer-Aided Design & Computer Graphics*, 56(1),P.P.142-146.
6. Cui ZH. (2014) Mean Shift based FCM Image Segmentation Algorithm. *Control and Decision*, 29(6), p.p.1130-1134.
7. Guo XC.(2014) Improved Fuzzy C-Means Clustering Algorithm. *Journal of Jilin University: Sci Ed*, 52(6), p.p.1293-1296.
8. Yu CJ, Zhang R.(2014) Research of FCM Algorithm based on Canopy Clustering Algorithm under Cloud Environment. *Computer Science*, 41(11), p.p.316-319.
9. Zhou SB.(2014)Data-weighted Fuzzy C-means Clustering Algorithm. *Systems Engineering and Electronics*, 36(11), p.p.2314-2319.
10. Wang Y, Pang YN.(2014)Improved FCM Algorithm based on Polar Coordinates Transformation for Iris Location. *Journal of Jilin University : Sci Ed*, 52(3), p.p.515-518.