FCM Clustering Algorithm Based on Laplacian Coefficient Optimized Objective Function

Sheng Huang  
Information Technology & Engineering, Hunan International Economics University, Changsha 410001, Hunan, China

Qiao Li  
Information Technology & Engineering, Hunan International Economics University, Changsha 410001, Hunan, China

Jun Li  
Information Technology & Engineering, Hunan International Economics University, Changsha 410001, Hunan, China

Abstract

FCM (Fuzzy c-means) clustering algorithm can be used to build sample generic uncertainty description. In this paper, we propose a FCM clustering algorithm based on the Laplacian Coefficient optimization objective function. The Laplacian Coefficient is introduced in the objective function, the structure of information between the object into the weight, thus improve the quality and efficiency of algorithm, and then through the compactness and separability measure two parts to optimize clustering validity, and use the method of maximum effectiveness will function for standardization, in order to improve the anti-noise performance of improved algorithm. In the UCI standard data sets of Iris and Wine simulation experiments show that the proposed improved FCM algorithm compared with standard algorithm has more accurate clustering effect, and less affected by noise and higher robustness.

Key words: Laplacian Coefficient, Objective Function Optimization, Compactness Measurement, Separation Measurement, Validity Function.

1. INTRODUCTION

Clustering technology is a data mining technique which searches a limited kind collection or clustering collection, and identifies, and then describes the data. As an important function of data mining, clustering analysis can get the distribution of data, thus the characteristics of each class can be observed and further analysis of certain classes can be made centrally (Lu and Hao, 2010). In addition, clustering analysis can be used as a preprocessing step of other algorithms (such as correlation analysis and classification), as a result, the efficiency of these algorithms are greatly improved by dealing with the generated classes, so clustering analysis has become a very active research field in data mining area. Many effective clustering algorithms have been developed, and new algorithms are continuing to emerge (Hou et al., 2011).

In order to improve the performance and effect of the clustering algorithm greatly, some scholars proposed that combine the other field methods with clustering algorithm to make up some defects of clustering algorithm in the data mining field, and bring the optimal performance of the clustering algorithm into full play. The methods used usually are: genetic algorithm, immune algorithm, ant algorithm, etc. Dai Wenhua and others (Dai et al., 2012) used a new type of variable length chromosome coding scheme to form chromosomes by selecting sample points as the initial cluster centers randomly, the local optimal solution was effectively avoided, and the optimal number of clusters and clustering results were got by genetic variation within population, and parallel evolution, marriage between populations combined with the high efficiency of k-means algorithm and global optimization capability of parallel genetic algorithm. The research on artificial immune system rising in recent years is a new field of application. It has brought new vitality to the field of clustering analysis with the development of immune algorithm. Chen Xi and others (Chen et al., 2011) has successfully applied the immune algorithm to clustering analysis. In the studies of fuzzy clustering algorithm, Ruspini (Ruspini, 2009) first proposed the fuzzy clustering algorithm based on the objective function, but the real effective algorithm called FCM (Fuzzy C-Means) is given by Dunn (Dunn, 1974). FCM algorithm is simple, efficient, and widely used in the field of application. The application of image processing, Chinese text mining, radar signal recognition in literature (Huang, 2012; Geng and Wang, 2012; Chen and Luo, 2008) is a good example. While the defect of the FCM algorithm is obvious, Li Lili (Li et al., 2012) proposed a fuzzy clustering algorithm of optimization based on simulated annealing particle swarm for the defects of FCM algorithm such as sensitivity to the initialization, and easily getting into local extreme point. This algorithm overcomes the shortcomings of the fuzzy C-means
clustering algorithm by using the ability for strong global optimization of the particle swarm and the ability for jumping out of the local extremum of the simulated annealing algorithm. According to the problem that the FCM algorithm is easy to fall into the local optimal, people introduce the idea of evolutionary computation into FCM, in order to achieve the purpose of global optimization. There are some algorithms, mainly particle swarm optimization algorithm (Zhang and Zhou, 2014), algorithm based on simulated annealing (A sultan and Selin, 2013), genetic algorithm (Buckles et al, 2014) and evolutionary strategy algorithm (Babu and Murty, 2014), etc.

2. DEFECT ANALYSIS OF FCM CLUSTERING ALGORITHM

The basic idea of FCM algorithm (fuzzy C means clustering algorithm) is: the membership function matrix whose samples belong to different categories and the fuzzy coefficient \( m \) are introduced based on the HCM algorithm. Thus the classification of data, the calculation of clustering point, and the objective function are adjusted compared with the HCM algorithm.

In a given dataset \( X = \{x_1, x_2, \ldots, x_n\} \), each sample contains \( s \) properties. The FCM algorithm is to divide the data set \( X \) into \( c \) \( (2 \leq c \leq n) \) classes, and \( v = \{v_1, v_2, \ldots, v_c\} \) are \( c \) cluster centers. \( u_{ij} \) indicates the membership of sample \( j \) belonging to sample \( i \), \( u_{ij} \in [0,1] \), \( \sum_{i=1}^{c} u_{ij} = 1 \).

The objective function of FCM clustering algorithm is:

\[
J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^m d_{ij}^2
\]  

\( d_{ij} = \|x_i - v_j\| \) represents the Euclidean distance between sample \( x_i \) and center \( v_j \), \( m \geq 1 \) is a fuzzy weighted parameter. So the FCM algorithm is to find the minimum value of objective function \( J \) under the condition of meeting \( u_{ij} \in [0,1] \), \( \sum_{i=1}^{c} u_{ij} = 1 \).

In order to achieve the minimum for the objective function \( J(U,V) \) of formula (1), the cluster center \( v_i \) and the membership matrix \( U \) can be calculated by the following formula:

\[
v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, i = 1,2,\ldots,c
\]  

When \( d_{ij} = 0 \), so \( u_{ij} = 1 \), \( u_{ik} = 0 \), \( i = 1,2,\ldots,n \).

The specific process of the FCM algorithm is:
1. Enter the number of clustering \( c \), fuzzy factor \( m \) and iterative termination conditions \( \delta \);
2. Initialize the clustering center \( v_i^0 (i=1,2,\ldots,c) \);
3. Calculate \( u_{ij} (i=1,2,\ldots,c, j=1,2,\ldots,n) \) using formula (3);
4. Calculate \( v_i^t (i=1,2,\ldots,c) \) using formula (2);
5. If \( \|v_i^t - v_i^{t-1}\| \leq \delta \), the iteration is terminated, turn to step (6), or \( v_i^0 = v_i^t (i=1,2,\ldots,c) \), turn to step (3);
6. Export the clustering result \( (V,U) \).

Clustering analysis of FCM types is a kind of partition method, no matter whether there is a natural structure in the data and the feature space, if a classification number \( c \) is given, the \( c \) partition of data set must be exported. Therefore, the algorithm has an unreasonable assumption: the data to be analyzed is clustered. This unreasonable assumption leads to that the clustering possibility of data set is not analyzed by the present algorithms of FCM types, but a certain membership of the data is applied hard. This will result in that the fuzzy \( c \) partition is applied to data set unjustifiably by algorithms of FCM types, even if the data is distributed uniformly in the feature space, and there is no any clustering structure, thereby it is difficult to explain the results of clustering analysis. As shown in Figure 1, this is a data set distributed uniformly, and there is no natural structure, but when \( c = 3 \), the division in Figure 2 will be produced by the FCM algorithm, and different data partition is made by different initialization.
Therefore, it is difficult to make a reasonable explanation of the clustering results, and it is impossible to reveal the structural information contained in the data, and help users to generate new ideas or to form a new hypothesis.

3. FCM CLUSTERING ALGORITHM BASED ON LAPLACIAN COEFFICIENT OPTIMIZED OBJECTIVE FUNCTION

3.1. Optimized Objective Function Based on Laplacian Coefficient

In this paper, the Laplasse coefficient is used to optimize the objective function of the standard FCM algorithm. The Laplacian coefficient \( S = \{S_j\} \) is given to represent the weighted coefficient of the sample \( j \) belonging to cluster center \( i \). The definition of \( S_j \) is as follows:

\[
S_j = e^{V_{ij} - 1/n_i} \tag{4}
\]

The objective function weighted is:

\[
F(U, S, C) = \sum_{j=1}^{n} \sum_{i=1}^{c} (s_j u_{ij}^n \|v_i - v_j\|) \tag{5}
\]

Where \( U = \{u_j\} \) is a membership matrix, and, when \( 0 < u_{ij} < 1 \), the constraint is:

\[
\sum_{j=1}^{n} u_{ij} = 1, i = 1, 2, ..., c, j = 1, 2, ..., n \tag{6}
\]

The extreme value of formula (5) can be derived by using the Laplacian optimization theory. If the algorithm is convergent, we can get the values of the cluster centers \( v_j \) and membership matrix \( U \).

The iterative formula of the cluster center value is shown as formula (7).
The iteration formula of the membership matrix $U$ is shown as formula (8).

$$u_i = \frac{1}{\sum_{k=1}^{c} \left( \frac{s_{ij} \| x_j - v_i \|^2}{s_{ij} \| x_j - v_i \|^2} \right)^{1/(c-1)}}$$

(8)

3.2. Clustering Validity Optimization Based on Separation Degree Analysis

The clustering validity was optimized through compactness measurement and separation measurement in this paper after the optimization of the objective function.

1) Compactness measurement

The method of compactness measurement proposed in this paper is as follows:

$$\text{var}(U,V) = \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij} d^2(x_j, v_i) / n(i) \right] \cdot \left( \frac{c+1}{c-1} \right)^{1/2}$$

(9)

Where $n(i)$ represents the number of data $i$, $d(x,y)$ represents a measurement, its definition is as follows:

$$d(x,y) = \left[ 1 - \exp \left( -\beta \| x - y \|^2 \right) \right]^{1/2}$$

(10)

$\beta$ is defined as the reciprocal of the sample covariance:

$$\beta = \left( \frac{\sum_{j=1}^{n} \| x_j - \bar{x} \|^2}{n} \right)^{-1}$$

(11)

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$$

(12)

We can find from formula (9), when the number of clusters is close to the number of samples called $n$, $\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij} d^2(x_j, v_i)$ is decreasing monotonically, while $\left( \frac{c+1}{c-1} \right)^{1/2}$ and $\frac{1}{n(i)}$ increase with the increasing of $c$ since $\lim_{c \to \infty} \left\| x_j - v_i \right\|^2 = 0$. Therefore, the reduction of the measurement value is limited by $\frac{1}{n(i)}$. In this way, we can achieve the goal that the compactness measurement is as much as possible while the number of clusters is as small as possible.

When we take the noise point (as shown in Figure (3) as a separate type called $(t)$, $\frac{1}{n(t)} = 1$, its weight is greater than other classes, it will also affect the value of the entire effectiveness metric, so that var$(U_i, V_j)$ is more robust to the noise. $\left( \frac{c+1}{c-1} \right)^{1/2}$ is only used to make minor adjustments in order to achieve a better division of results.
Figure 3. Data set with noise

(2) Separation measurement

Supposing there is a data set \( X = \{x_1, x_2, \ldots, x_N\} \), \( F(X) \) represents a set of all fuzzy sets, and \( F = \{F_1, F_2, \ldots, F_c\} \) represents a fuzzy \( c \)-partition of \( X \). The separation measurement proposed in this paper is obtained by calculating the distance between fuzzy sets. In order to get the distance between fuzzy sets, we used the similarity measurement, that is, the similarity between fuzzy set \( F_i \) and fuzzy set \( F_j \), its formula is shown as follows:

\[
S(F_i, F_j) = \max_{x \in X} \min(\mu_{F_i}(x), \mu_{F_j}(x))
\]

For example, when the similarity of fuzzy set \( F_i \) and fuzzy set \( F_j \), \( S(F_i, F_j) = 0.4 \), it means that the degree of the similarity or relevance is at least 0.4. On the contrary, we can also say that the degree of the irrelevance or isolation between the two fuzzy sets is 0.6. The definition of separation measurement for fuzzy division based on the conclusions above is as follows:

\[
\text{Sep}(c, U_c) = 1 - \max_{i \neq j} S(F_i, F_j)
\]

Since the compactness measurement and the separation measurement usually are not in the same order of magnitude, so it is standardized by the method of maximum value.

\[
\text{Com}^c(V_c, U_c) = \frac{\text{Com}(V_c, U_c)}{\text{Com}_{\text{max}}}, c = 2, 3, \ldots, c_{\text{max}}
\]

\[
\text{Sep}^c(c, U_c) = \frac{\text{Sep}(c, U_c)}{\text{Sep}_{\text{max}}}, c = 2, 3, \ldots, c_{\text{max}}
\]

Where

\[
\text{Com}_{\text{max}} = \max_c \text{Com}(V_c, U_c)
\]

\[
\text{Sep}_{\text{max}} = \max_c \text{Sep}(c, U_c)
\]

So, the definition of clustering validity function \( V_c \) is as follows:

\[
V_c(V_c, U_c) = \frac{\text{Com}^c(V_c, U_c)}{\text{Sep}^c(c, U_c)}
\]

\( V_c(V_c, U_c) \) is defined as the ratio of compactness measurement and separation measurement, the smaller the compactness measurement, the smaller the difference of the data in the same class; the bigger the compactness measurement, the bigger the difference of the data in the same class. Thus, the minimum value of \( V_c \) corresponds to the optimal fuzzy \( c \)-partition or the optimal number of clusters \( c^* \).
4. ALGORITHM SIMULATION EXPERIMENT

In order to verify the performance of the improved algorithm proposed in this paper, the simulation test is carried out. First, the datasets Iris and Wine of the standard dataset UCI were selected. According to the research on FCM parameter, set \( m = 2 \), \( \delta = 0.01 \), and after 20 experiments, the results showed that the data set was divided into 3 categories consistently. While compared the improved FCM algorithm with standard FCM algorithm, if the number of clusters was different, the results obtained were also different. The results are shown in table 1.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Data set</th>
<th>Cluster number</th>
<th>Accuracy/%</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM</td>
<td>Iris</td>
<td>3</td>
<td>81.6</td>
</tr>
<tr>
<td>Improved FCM</td>
<td>Iris</td>
<td>3</td>
<td>92.4</td>
</tr>
<tr>
<td>FCM</td>
<td>Wine</td>
<td>3</td>
<td>85.2</td>
</tr>
<tr>
<td>Improved FCM</td>
<td>Wine</td>
<td>3</td>
<td>95.3</td>
</tr>
</tbody>
</table>

We can find from table 1, compared with the standard FCM algorithm, the improved FCM algorithm is more accurate.

Then the Bupa Liver Disorder dataset and Butterfly dataset were used to test the effectiveness of the de-noise ability of validity function. The data set is shown in figure 4-5.
The two data sets above were tested by the validity function, and several experiments were conducted with different initial values, since the results obtained were identical, so we just list one result of the experiments, the changes of validity function for different data sets are shown in figure 6-7.

![Figure 6. Change of Bupa Liver Disorder schematic diagram](image)

![Figure 7. Change of Butterfly schematic diagram](image)

In order to verify the robustness of function $V_c$, 100 noise points were added randomly to the Bupa Liver Disorder data set and Butterfly data set which are shown in figure 8-9, and the change schematic diagrams are shown in figure10-11.

![Figure 8. Bupa Liver Disorder data set added with noise](image)
From figure 10-11, it is easy to find that the validity function proposed can give the correct number of clusters for the two data sets, and the experimental results show that the function $V_w$ will not be affected by noise, and has a strong robustness.

5. CONCLUSIONS

Fuzzy clustering has become the mainstream of clustering analysis gradually since it can describe the intermediary of sample category, and can reflect the real world objectively. Among the numerous fuzzy
clustering algorithms, FCM algorithm can be said to be the most widely used and the most sensitive algorithm. According to the defects existed in FCM clustering algorithm, a FCM clustering optimization algorithm of objective function based on Laplacian coefficient is put forward. The simulation results show that the clustering effect of improved FCM algorithm is more accurate than standard algorithm, and worthy of spreading.

ACKNOWLEDGEMENTS

China hunan provincial science&technology department project: Design and implementation of smart home based on internet of things(No. 2013GK3113)

REFERENCES