New Data Clustering Algorithm Combined of Ant Colony Algorithm and Improved Fuzzy C-Means Algorithm

Zhiming Zhang, Guobin Wu and Jie Luo*
Information engineering department Engineering College of CAPF Xi'an, China
*Corresponding author

***Abstract*—A new clustering algorithm combined of ant colony and improved Fuzzy C-Means (AC-LFCM) was proposed to resolve the shortage of Fuzzy C-Means (FCM) clustering algorithm on the presence of sensitive to initialization, easy to fall into local optimum and neglected the influence of local information of data. Firstly, aimed at the existing defects of FCM, a new algorithm named local-Fuzzy C-Means (LFCM) was formed through considering influence of data's neighborhood to target function; then introduced ant colony algorithm with great ability for disposing local extremum and parallel computing to fix on the initial numbers of clustering as well as the centers of clustering, combined with LFCM algorithm to find the whole distributing optimization clustering and achieve clustering analysis. And In the data clustering experiments on synthetic datasets and three datasets of UCI datasets by the LFCM and AC-LFCM algorithm, the results show that, compared with FCM, the algorithm has obvious advantage on the clustering performance.**

Keywords-data mining; ant colony algorithm; FCM algorithm; clustering

I. INTRODUCTION

Clustering is a kind of important data analysis technology in data mining, which is widely used in data mining [1]. There are multiple clustering methods in the existence of data mining with different work principle and basic algorithm characteristics, and fuzzy c-means clustering is a kind of common fuzzy clustering algorithm [2]. The basic idea is to a database containing n data object, to build data clustering by the data partition method according to membership size of data points to all kinds of data clustering centers, namely divided n data object into C group ($c \le n$). But the FCM clustering algorithm belongs to iterative optimization method with local search ability essentially, which is sensitive to the initial cluster centers, and easy to fall into local optimum, and difficult to guarantee the global optimal; beside, data set itself may contain a variety of attributes, while the objective function of FCM clustering algorithm only considers the distance between data points, errors may occur in the classification of large data sets classification.

Reference [3] ant colony algorithm, another swarm intelligence optimization algorithm following the simulated annealing algorithm, genetic algorithm, neural network algorithm of heuristic search algorithm [4,5]. The algorithm has ability of fast random search and global optimization, and adopts mainstream parallel and distributed calculation method. Applying the ant colony algorithm to data clustering to generated initial cluster centers of data object, can compensate for the problem of FCM clustering algorithm that is sensitive to initial clustering center. But based on the principle of positive feedback, ant colony algorithm is easy to appear the precocity and stagnation when iteration to a certain number of times, then use FCM algorithm for data clustering will achieved good effect in the optimization and time performance. This paper introduces the ant colony algorithm, combined with improved neighborhood Fuzzy C-means method, complementary advantages, to further optimize the ability of ant colony and fuzzy clustering algorithm.

II. IMPROVED FUZZY C MEANS CLUSTERING ALGORITHM

A. Fuzzy C Means Clustering Algorithm

The basic idea of the FCM clustering algorithm is as follows [6]: the object in the database is divided into C classes, and each class corresponds to a cluster center V and u_{ij} represents membership degree of arbitrary data points x_i belonging to the j-th category. FCM using error's square as the objective function, which describes square sum of weighted distances of various objects point to the cluster center, that is

$$J_m(U,V) = \sum_{\mathrm{k}=1}^{c}\sum_{i=1}^{n} u_{ik}^{m} d_{ik}^{2} \tag{1}$$

Based on the above analysis, the general steps of the FCM clustering algorithm can be described as:

$$u_{ij} = \frac{\left[\dfrac{1}{\left\|\mathrm{x}_i - v_j\right\|^2}\right]^{\frac{1}{\mathrm{m}-1}}}{\sum_{j=1}^{c}\left[\dfrac{1}{\left\|\mathrm{x}_i - v_j\right\|^2}\right]^{\frac{1}{\mathrm{m}-1}}} \quad \begin{matrix} i = 1,2,\cdots,n \\ j = 1,2,\cdots,c \end{matrix} \tag{2}$$

Step 1: parameter initialization. Random initialize cluster center V^0 and according to the formula (2) to calculate the initial membership matrix U^0 ; specify weighted power exponent constant M, let the number of iterations $l=1$.

Step 2: computing cluster center. According to the formula (3), the corresponding cluster centers V^l are calculated for the membership degree matrix U^l.

$$v_j = \frac{1}{\sum_{i=1}^{n} u_{ij}^m} \sum_{i=1}^{n} \left[u_{ij}\right]^m x_i \qquad j = 1,2,\cdots,c \quad (3)$$

Step 3: update membership matrix. According to the formula (2), calculate the membership matrix U^l with the cluster center V^l.

Step 4: for a given threshold $\varepsilon(\varepsilon > 0)$, if $\left\|v^{(l+1)} - v^l\right\| < \varepsilon$ the iteration stops; otherwise let $l = l+1$, then return to Step 2,and continue to iterative optimization.

Analysis the principle of FCM shows that, algorithm only considers the distance information of the data points as the study object, clustering and classification completely counts on the Euclidean distance between data objects. although can solve many problems in data clustering. However, when applied on a large data set of data classification has great limitations. As for classification of large data sets, data set itself may contain a number of interrelated attributes, in accordance with the FCM algorithm just calculates the Euclidean distance between the data objects as the basis for classification would ignores the neighborhood information between data objects, which leads to a error that different clustering of data points were assigned to the same class. The neighborhood of the data points contains a lot of information that can be used for data classification. Therefore, this paper introduces the neighborhood into the fuzzy C- means clustering, and gets the local - fuzzy -C mean clustering algorithm (LFCM).

B. Local - Fuzzy -C Mean Clustering Algorithm (LFCM)

In order to overcome the limitations of the FCM clustering algorithm, and LFCM algorithm increases effect of neighborhood information through changing the objective function of the FCM algorithm [7], the central idea is to increase a function key which expresses Neighborhood Information of data point to the objective function of the FCM, to exploit and make full use of the information of the data objects in the classification process, to ensure that the neighborhood of data objects in the data quantity and structure does not change. While changing the objective function, LFCM algorithm also changes the membership matrix and data clustering center, which directly improve the accuracy of the classification results. In addition, due to the impact of the consideration of the field information, part of data points will be directly assigned to the cluster centers of its territory, dramatically improve the efficiency of calculation and convergence. The expression of objective function of improved LFCM algorithm is as follows:

$$\begin{aligned} J_m(U,V) &= \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^m d_{ik}^2 = \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^m \left\|x_i - v_k\right\|^2 \\ &+ \gamma \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^m \left\|\overline{x_i} - v_k\right\|^2 \end{aligned} \quad (4)$$

In the formula (4), the two terms respectively represents the minimum distance of each data point from the cluster center and the minimum distance of the average value of the data from the cluster center. Comparing with the formula (1), we use the minimum distance of the average value of the data from the cluster center to describe the information of the data object.

The formula (5) can be got when solving the minimum value of the formula (4) using the Lagrange multiplier method.

$$\begin{aligned} J_m(U,V) &= \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^m \left\|x_i - v_k\right\|^2 + \gamma \\ &\sum_{k=1}^{c}\sum_{i=1}^{n} u_{ki}^m \left\|\overline{x_i} - v_k\right\|^2 + \lambda(1 - \sum_{k=1}^{c} \mu_{ki}) \end{aligned} \quad (5)$$

Then we obtain cluster center v_k and membership matrix μ_{ki} for dataset

$$u_{ki} = \frac{1}{\sum_{l=1}^{c} \left[\frac{\left\|x_i - v_k\right\|^2 + \gamma \left\|\overline{x_i} - v_k\right\|^2}{\left\|x_i - v_l\right\|^2 + \gamma \left\|\overline{x_i} - v_l\right\|^2} \right]^{\frac{1}{m-1}}} \quad (6)$$

$$v_k = \frac{\sum_{i=1}^{n} u_{ki}^m + \gamma \overline{x_i}}{\sum_{i=1}^{n} u_{ki}^m} \quad (7)$$

III. FUSION OF LFCM ALGORITHM AND ANT COLONY ALGORITHM (AC-LFCM)

Ant colony algorithm has stable global search ability and fast parallel computing ability, but when solving evolve to one certain algebra, due to increasing of information of the optimal path so that a large number of ants will gather in the current several pieces of the optimal paths, which result in precocity and stagnation of optimization. LFCM algorithm is sensitive to the initialization, in the previous clustering, due to random data clustering center, algorithm's convergence speed is slow, and easy to fall into local optimum. Therefore, the basic idea of combination of LFCM algorithm and ant colony algorithm is: Based on ant colony algorithm ability of fast global searching

in the early stages of the clustering to rapidly generate the initial clustering center of the data object, and later use the LFCM convergence mechanism to optimize clustering structure, which achieve advantage complementary.

A. Setting of Ant Colony Algorithm in AC-LFCM

A lot of experiments and analysis show that ants will release a substance called pheromone when pass a path and transfer information through the material with the other individual, then the follow ants can identify the presence and degree of the pheromone to determine their marches direction [8, 9]. Ant colony algorithm is based on the transmission of information by ants to find the optimal path information of positive feedback phenomenon to find the optimal solution.

Based on the above analysis and combined with the actual demand to use ant colony algorithm to generate initial data clustering center, the ant colony algorithm in the AC-LFCM is set as follows:

1) Question abstract: the ant colony individual is abstracted as the data sample point, the food source is abstracted as the cluster center, and the process of the ant colony foraging is the process of data clustering.

2) Parameter initialization: for a given data point set $X=\{x_1,x_2,\cdots,x_n\}$, set the iteration algebra for N, iteration number $l=0$, time $t=0$, pheromone of each path $\tau_{ki}^{l}(t)=0$, cluster radius is R, initial cluster center $V^{(l)}$; calculation $d_{ki}^{(l)}=\left\|x_k-v_i^l\right\|$.

3) Optimization process:

Selection mechanism: each ant individual represents a data point, for the pheromone of each path, if $d_{ki}^{(l)}\le r$, so $\tau_{ki}^{l}(t)=1$; otherwise $\tau_{ki}^{l}(t)=0$, then on the time t the probability $p_{ki}^{l}(t)$ of ants K moving to v_i^l from x_k defines:

$$p_{ki}^{l}(t)=\frac{\left[\tau_{ki}^{l}(t)\right]^{\alpha}\left[\eta_{ki}^{l}(t)\right]^{\beta}}{\sum_{j\in K}\left[\tau_{ki}^{l}(t)\right]^{\alpha}\left[\eta_{ki}^{l}(t)\right]^{\beta}} \tag{8}$$

Update mechanism: when $l=l+1$ a path search is completed, the pheromone of path (k,i) update can be defined:

$$\tau_{ki}^{l}(t)=\rho\cdot\tau_{ki}^{(l-1)}(t)(t-\Delta t^{(l-1)})+\Delta\tau_{ki}^{(l-1)} \tag{9}$$

4) Termination condition: In experiment, two ways were usually used to judge whether the iteration is terminated or not: $\left\|V^1-V^0\right\|\le\varepsilon$ or the number of iterations is greater than the preset maximum iteration number N.

B. AC-LFCM Clustering Algorithm

In the fusion of ant colony clustering algorithm and LFCM algorithm, firstly generate initial clustering center and the clustering number through the ant colony algorithm, then use the clustering center and number as the initial information of LFCM parameter initialization, then transfer the LFCM clustering process for data clustering and labeling. The algorithm of ant colony clustering and LFCM data clustering algorithm is as follows:

Step 1: Transfer the process of the ant colony algorithm to generate data clustering center $V=\{v_1,v_2,\cdots,v_c\}$ and the number of cluster *C*;

Step 2: Initialize the LFCM parameters according to the results of ant colony algorithm, including the number of cluster *C*, cluster center $V=\{v_1,v_2,\cdots,v_c\}$ fuzzy factor m, the average value of the data set $\bar{x}$ and the iterative threshold $\varepsilon(\varepsilon>0)$;

Step 3: Calculate the degree of membership matrix iteration according to the formula (9) $u_{ki}(k=1,2,\cdots,c,\quad i=1,2,\cdots,n)$;

Step 4: Compute cluster center $v_k^1(k=1,2,\cdots,c)$ according to the formula (10);

Step 5: If $\left\|V^1-V^0\right\|\le\varepsilon$, the iteration terminates, output the clustering result (U,V); otherwise, order $v_k^0=v_k^1(k=1,2,\cdots,c)$ and return step 3.

IV. Simulation Results

In order to verify the effectiveness of the algorithm, this paper uses the manual data point set and the experimental data set to test the clustering effect of the improved LFCM and AC-LFCM fused clustering algorithm. n the experiment of artificial data points, the data set used is 40 numerical small data points of artificial structure, the data set contains two numerical attributes, both can be used for data classification. Figure I is a sketch map of two-dimensional artificial data set.

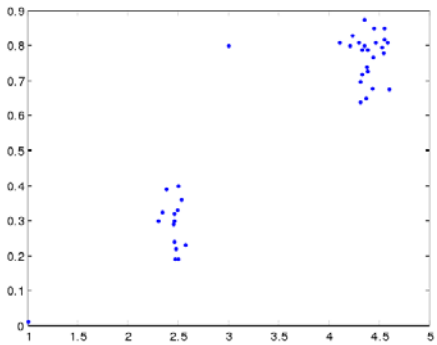

FIGURE I. SKETCH MAP OF 2D ARTIFICIAL DATA POINTS

From Figure I the data set includes 40 data points, according to the distribution of data points the dataset can be

divided into four categories, and respectively apply FCM algorithm and LFCM algorithm for clustering data, then use comparison of the data classification results to demonstrate the effectiveness and superiority of the LFCM clustering algorithm. Figure 2 is diagram of clustering data classification result using FCM algorithm, and the objective function iterates 15 times in experiment; Figure III is the classification results using improved LFCM clustering algorithm, which divide data set into four types, the objective function iterates 10 times.

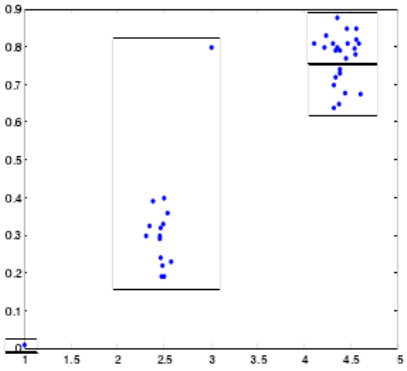

FIGURE II. DATA CLUSTERING RESULTS OF FCM ALGORITHM

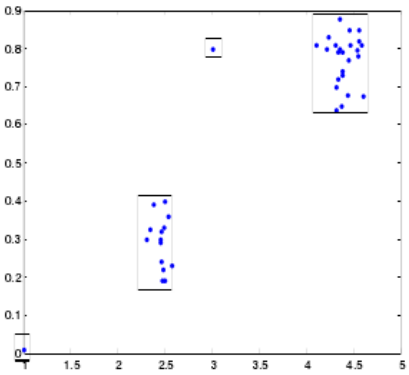

FIGURE III. DATA CLUSTERING RESULTS OF LFCM ALGORITHM

FCM algorithm in Figure II divided the data set to four categories, but some single data point was falsely assigned to other data types, and this should be divided into a class of data points was wrongly divided into two types, classification effect is not very good; Compared with Figure II, Figure III applied LFCM algorithm on the data set classification, each data point was accurately clustering to best class, classification effect is very great. Analysis data classification results of two kinds of clustering algorithm, it is concluded that accuracy of LFCM classification was significantly higher than that of FCM algorithm for small data set classification, and the iteration times of former target function less than that of the latter, so LFCM clustering has higher efficiency

Experimental data sets using three popular machine learning data sets, including Iris Flower data set, Wine data set, Identification of Glass data set, Each data set was suitable for the experimental correlations among the study variables with the number of attributes and classification, The specific data sets and their characteristics are shown in Table I.

Experiments select F-measure to evaluate the clustering performance of the algorithm, the F-measure is an external evaluation method combined with of information.

In experiments, we respectively use the fuzzy c-means algorithm (FCM), fuzzy ant colony clustering algorithm (AC-FCM) and fusion clustering algorithm based on ant colony and neighborhood fuzzy C-means Clustering (AC-LFCM) of three data sets in the table 1 to conduct 30 times of experiment, the average results of dataset according to the F-measure evaluation method is shown in Table II:

TABLE I. DESCRIPTION OF EXPERIMENTAL DATA SET

Dataset	Record number	Attribute Number	Classification number
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6

TABLE II. AVERAGE F-MEASURE OF DATASET CLUSTERING

Algorithm	Iris	Wine	Glass
FCM	0.905	0.891	0.903
AC-FCM	0.911	0.909	0.912
AC-LFCM	0.918	0.913	0.917

Experimental results shows, the performance of the fusion algorithm based on ant colony and neighborhood fuzzy C means clustering is better than that of FCM, fuzzy ant colony clustering algorithm and algorithm combined with FCM algorithm and ant colony algorithm.

V. CONCLUSIONS

This paper adopts neighborhood fuzzy c-means algorithm instead of fuzzy c-means algorithm and combines it with ant colony algorithm to form a new method of data clustering. using fast global search and local search optimization ability of ant colony algorithm, the algorithm overcomes the defects of fuzzy c-means algorithm that sensitive to initial value and easy to fall into local optimum, in addition, algorithm introduce the neighborhood information of the data point to improve FCM algorithm, which solves the problem that is easy to generate the of classification error of FCM algorithm. Simulation results show that the proposed algorithm is superior to the FCM algorithm, the fusion clustering algorithm based on ant colony and FCM, which has a certain advantage in optimizing performance and stability.

REFERENCE

[1] WU Y H. General Overview on Clustering Algorithms[J]. Computer Science, 2015, 42(z1):491-499.

[2] ZENG A P. A fuzzy Rough Set Approach for Incremental Feature Selection on Information Systems[J]. Fuzzy Sets Systems,2015:39-60.

[3] Dorigo M, Maniezzo V, Cnlnrni A. Ant system: ant system: optimization by a colony of cooperating agents [J].IEEE Trans On System ,M an, and Cybernetics,1996,26(1):29-41.

[4] XU X H, FAN Y F. Improved ants-clustering algorithm and its application in multi-attribute large group decision making[J]. Systems Engineering and Electronics, 2011, 33(2):346-349.

[5] SHA L ,BAO P M, LI N G. The research of a clustering algorithm based on the ant colony system[J]. Journal of Shandong University (Engineering Science), 2010, 40(3):13-18.

[6] ZHAO F, JIAO L C, LIU H Q. Kernel Generalized Fuzzy C-means Clustering With Spatial Information for Image Segmentation[J].Digital Signal Processing. A Review Journal,2013,23(1):184-199.

[7] LI J X. Anomaly detection method for datasets based on fuzzy clustering[D].Harbin: Harbin Institute of Technology,2015.

[8] BAI Y N, SI Y S. A Fuzzy Clustering Algorithm Based on the Self-adaptive Ant Colony Algorithm[J]. Journal of North China Institute of Water Conservancy and Hydroelectric Power, 2011,32(6):134-137.

[9] Gao W. Improved Ant Colony Clustering Algorithm and Its Performance Study.[J]. Computational Intelligence & Neuroscience, 2016, 2016:1-14.