An Improved KFCM Algorithm Based on Artificial Bee Colony

Xiaoqiang Zhao* and Shouming Zhang

College of Electrical and Information Engineering, Lanzhou University of Technology,
Lanzhou, 730050, China
Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou,
730050, China
xqzhao@lut.cn, zhsm-2008@163.com

Abstract. Kernel fuzzy C-means（KFCM）clustering Algorithm is one of the most widely used methods in data mining, but this algorithm still exists some defects, such as the local optima and sensitivity to initialization and noise data. Artificial bee colony (ABC) is a very simple, robust, stochastic global optimization tool which is used in many optimization problems. In this paper, an improved KFCM algorithm based on ABC (ABC-KFCM) is proposed. It can integrate advantages of KFCM and ABC algorithm. According to the test, compared with the FCM and KFCM clustering algorithm, the proposed algorithm improves the optimization ability of the algorithm, the number of iterations is fewer, and the convergence speed is faster. In addition, there is also a large improved in the clustering result.

Keywords: Data mining, kernel fuzzy C-mean clustering, artificial bee colony, ABC-KFCM.

1 Introduction

With development of era, the number of data in various ways has increased rapidly. It is difficult to take full advantage of the useful knowledge stored in these data by the traditional approaches. Data mining technology emerges as the times require and is abroad applied on every field. Data Mining is defined as a process that gets information and knowledge which are connotative, unknown and useful from practical data which is substantial, incomplete, noise, ambiguous and stochastic[1].

During the past decade, clustering analysis as one of the main method of data mining causes the attention of people more and more. Clustering is the process of assigning data objects into a set of disjoint groups called clusters so that objects in each cluster are more similar to each other than objects from different clusters.

* Corresponding author: Xiaoqiang Zhao. This works is supported by Natural Science Foundation of Gansu Province (No. 0809RJZA005) and Foundation of Advanced Control Key Laboratory of Industrial Process of Gansu Province (No.XJK0907).

H. Deng et al. (Eds.): AICI 2011, CCIS 237, pp. 190–198, 2011.

K-means [2]is one of the most popular hard clustering algorithms which partitions data objects into k clusters where the number of clusters. This model is inappropriate for real data sets in which there are no definite boundaries between the clusters.

Since Zadeh proposed fuzzy sets that introduced the idea of partial memberships described by membership functions, it has been successfully applied in various areas. Especially, fuzzy sets could allow membership functions to all clusters in a data set so that it was very suitable for cluster analysis[3]. Ruspini first proposed fuzzy c-partitions as a fuzzy approach to clustering. Later, the fuzzy c-means (FCM) algorithms with a weighting exponent m=2 proposed by Dunn, and then generalized by Bezdek with m>1 became popular[4]. Fuzzy c-means clustering is an effective algorithm, but the random selection in center points makes iterative process falling into the local optimal solution easily.

Recently, tremendous works focus on using kernel method, which first maps the data into high dimension space to gain high discriminant capability, and then calculates the measure of the samples in their original data space with kernel function. Kernel fuzzy C-Means (KFCM) is proposed by substituting the Euclidean distance with kernel function.

KFCM not only to certain extent overcomes limitation of data intrinsic shape dependence and can correctly clustering, but also overcome sensitivity to initialization and noise data and improve the algorithm robustness. However, like FCM algorithm, KFCM algorithm still exists some drawbacks, such as the sensitivity to initialization and the tendency to get trapped in local minima. Therefore, an improved kernel fuzzy C-Means based on artificial bee colony (ABC-KFCM) is put forward.

2 KFCM Algorithm

FCM partitions a given dataset $X = \{x_1, x_2, \cdots, x_n\} \in R^p$, into c fuzzy subsets by minimizing the following objective function

$$J_m(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \left\| x_k - v_i \right\|^2 \tag{1}$$

where c is the number of clusters and selected as a specified value, n the number of data points, u_{ik} the membership of x_k in class i, m the quantity controlling clustering fuzziness, and V the set of cluster centers ($v_i \in R^p$). The matrix U satisfies

$$U \in \left\{ u_{ik} \in [0,1] \middle| \sum_{i=1}^{c} u_{ik} = 1, \forall k \quad and \quad 0 < \sum_{k=1}^{N} u_{ik} < N, \forall i \right\} \tag{2}$$

Define a nonlinear map as $\Phi : x \to \Phi(x) \in F$, where $x \in \mathrm{X}$. X denotes the data space, and F the transformed feature space with higher or even infinite dimension. KFCM minimizes the following objective function

$$J_m(U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \left\| \Phi(x_k) - \Phi(v_i) \right\|^2 \tag{3}$$

Where $\|\Phi(x_k)-\Phi(v_i)\|^2 = K(x_k,x_k)+K(v_i,v_i)+2K(x_k,v_i)$ (4)

Where $K(x,y)=\Phi(x)^T\Phi(y)$ is an inner product kernel function. If we adopt the Gaussian function as a kernel function, i.e., $K(x,\mathrm{y}) = \exp(-\|x-y\|^2/\sigma^2)$, then $K(x,x)=1$, according to Eq.(3) and Eq.(4), can be simplified to

$$J_m(U,V)=2\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m(1-K(x_k,v_i)) \tag{5}$$

Minimizing Eq.(5) under the constraint of U,we have

$$u_{ik}=\frac{\left(1/(1-K(x_k,v_i))\right)^{1/(m-1)}}{\sum_{j=1}^{c}\left(1/(1-K(x_k,v_j))\right)^{1/(m-1)}},\forall i=1,2,\cdots c;k=1,2,\cdots n \tag{6}$$

$$V_i=\frac{\sum_{k=1}^{n}u_{ik}^m K(x_k,v_i)x_k}{\sum_{k=1}^{n}u_{ik}^m K(x_k,v_i)x_k},\forall i=1,2,\cdots c \tag{7}$$

The proposed kernelized fuzzy C-means algorithm can be summarized in the following steps[5]:

```
Step 1: Set c, tmax, m and $\varepsilon > 0$ for some positive constant.
Step 2: Initialize the membership matrix $u_{ik}^0$.
Step 3: For t =1,2,……, tmax, do:
update the cluster centers $V_i^t$ with Eq. (7);
update membership matrix $u_{ik}^t$ with Eq. (6);
compute $E^t = \max_{i,k}\left|u_{ik}^t - u_{ik}^{t-1}\right|$, if $E^t \le \varepsilon$, stop;
end.
```

3 Artificial Bee Colony Algorithm

Artificial Bee Colony (ABC) algorithm was proposed by Karaboga for optimizing numerical problems in [6]. The algorithm simulates the intelligent foraging behavior of honey bee swarms. It is a very simple, robust and population based stochastic optimization algorithm.

In ABC[7] algorithm, the colony of artificial bees consists of three groups of bees: employed bees, onlookers and scouts. A food source represents a possible solution to the problem to be optimized and the nectar amount of a food source corresponds to the quality (fitness) of the associated solution, calculated by:

$$fit_i=\frac{1}{1+f_i} \tag{8}$$

In the algorithm[7][8], the number of the employed bees or the onlooker bees is equal to the number of solutions (the cluster centers) in the population. At the first step, the ABC generates a randomly distributed initial population of SN（the size of population）solutions (food source positions), Each solution x_i （i=1, 2, ..., SN）is a D(the number of optimization parameters)dimensional vector. After initialization, the population of the positions is subjected to repeated cycles, C =1,2, ...,MCN, of the search processes of the employed bees, the onlooker bees and scout bees. Each employed bee moves onto her food source area for determining a new food source within the neighborhood of the present one, and then evaluates the nectar amount. If the nectar amount of the new one is higher, the bee memorizes the new position and forgets the old one. After all employed bees complete the search process, they share the nectar information of the food sources and their position information with the onlooker bees on the dance area. An onlooker bee evaluates the nectar information taken from all employed bees and chooses a food source with a probability related to its nectar amount. If the nectar amount of the new one is higher, the bee memorizes the new position and forgets the old one.

An artificial onlooker bee chooses a food source depending on the probability value associated with that food source, p_i , calculated by:

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \tag{9}$$

In order to produce a candidate food position from the old one in memory, the ABC uses the following expression:

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \tag{10}$$

Where $j \in \{1,2,\cdots,D\}$ and $k \in \{1,2,\cdots,SN\}$ are randomly chosen indexes. Although k is determined randomly, it has to be different from i . ϕ_{ij} is a random number between [-1, 1]. As can be seen from (10), as the difference between the parameters of the xij and xkj decreases,the perturbation on the position xij decreases, too. Thus, as the search approaches to the optimum solution in the search space, the step length is adaptively reduced.

In ABC, providing that a position can not be improved further through a predetermined number of cycles (called limit), then that food source is assumed to be abandoned. Assume that the abandoned source is xi, then the scout discovers a new food source to be replaced with xi. This operation can be defined as in

$$x_i^j = x_{\min}^j + rand(0,1)(x_{\max}^j - x_{\min}^j) \tag{11}$$

ABC algorithm is a robust search process, exploration and exploitation processes are carried out together. The global search performance of the algorithm depends on random search process performed by scouts and neighbor solution production mechanism performed by employed and onlooker bees. Therefore, ABC algorithm is an efficient optimization tool since it combines exploitative local search and explorative global search processes efficiently.

4 ABC-KFCM Algorithm

Comparison of clustering and bee colony foraging is shown as Table1.

Table 1. Comparison of clustering and bee colony foraging

colony foraging	clustering
Position of a food source	A possible solution(clustering center)
The nectar amount of a food source	The quality of the associated solution
Speed of bee forage	Solution speed
The maximizing rewards	The best clustering result

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a set of n objects, where x_i is a D-dimensional vector. In ABC, a bee denotes a cluster center, i.e. $V = \{v_1, v_2, \cdots v_c\}$, where v_j is also a D-dimensional vector. In ABC-KFCM algorithm the same as other evolutionary algorithms, we need a function for evaluating the generalized solutions called fitness function. In this paper, Eq.(12) is used for evaluating the solutions.

$$fit_i = \frac{1}{1 + J_m(U, V)} \tag{12}$$

Where $J_m(U, V)$ is objective function of KFCM algorithm given in Eq.(5). The smaller is $J_m(U, V)$, the higher is the individual fitness fit_i and the better is the clustering result.

ABC-KFCM algorithm uses the capacity of global search in ABC algorithm to seek optimal solution as initial clustering-centers of KFCM algorithm, and then use KFCM algorithm to optimize initial clustering-centers, so as to get the global optimum. Detailed description of each step is given below:

Step 1: Initialize the parameters of ABC and KFCM including population size SN, maximum cycle number MCN, limit, clustering number c, m and ε;

Step 2: Compute kernel matrix $K(x_k, v_i)$ and initialize the membership matrix U^0 by Eq. (6)

Step 3: Generate the initial population (cluster center) cij by Eq. (7), and evaluate the fitness of the population by Eq. (12)

Step 4: ABC algorithm

4.1 Set cycle to 1

4.2 Set s to 1

```
 4.3 FOR each employed bee {
Produce new solution vij by using (6)
Calculate the value fiti
Apply greedy selection process}
 4.4 Calculate the probability values pi for the solutions (cij) by (9)
 4.5 FOR each onlooker bee {
Select a solution cij depending on pi
Produce new solution vij
Calculate the value fiti
Apply greedy selection process}
```

4.6 If the searching times surrounding an employed bee s exceeds a certain threshold limit, but still could not find better solutions, then the location vector can be reinitialized randomly according to Eq. (11), go to step 4.2

4.7 If the iteration value is larger than the maximum number of the iteration (that is, cycle> MCN, output the best cluster centers. If not, go to Step 4.1.

Step 5: KFCM algorithm

5.1 Update membership matrix u_{ik}^{t} with Eq. (6);

5.2 Update the cluster centers $v_i^{\,t}$ with Eq. (7);

5.3 Compute $E^t = \max_{i,k}\left|u_{ik}^{t} - u_{ik}^{t-1}\right|$, if $E^t \le \varepsilon$, stop; If not, go to Step 5.1.

5 Experiment Results

Three classification problems from UCI database which is a well-known database repository, are used to evaluate the ABC clustering algorithm. The three datasets are well-known iris, wine and glass datasets taken from Machine Learning Laboratory. The experimental data sets can be found in Table 2.

Table 2. Experimental data sets

Data set name	Number of sample	Class	Dimension
IRIS	150	3	4
wine	178	3	13
glass	214	6	9

Iris dataset is perhaps the best-known database to be found in the pattern recognition literature. The data set contains three categories of 50 objects each, where each category refers to a type of iris plant. One category is linearly separable from the other two; the latter are not linearly separable from each other. There are 150 instances with four numeric features in iris data set. There is no missing attribute value. The attributes of the iris data set are sepal length, sepal width, petal length and petal width.

Wine dataset contains chemical analysis of 178 wines, derived from three different cultivars. Wine type is based on 13 continuous features derived from chemical analysis: alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyaninsm, color intensity, hue, OD280/OD315 of diluted wines and praline. The quantities of objects in the three categories of the data set are 59, 71 and 48, respectively.

Glass dataset is another biggest number of classes (6 classes) in the problems that we tackle. It is used to classify glass types as float processed building windows, non-float processed building windows, vehicle windows, containers, tableware, or head lamps. Nine inputs are based on 9 chemical measurements with one of 6 types of glass which are continuous with 70, 76, 17, 13, 9, and 29 instances of each class, respectively.

There are three control parameters in ABC algorithm, the swarm size SN, the maximum cycle number MCN and the limit. They are set as follow: SN=20, MCN=2000, limit=100.The weighting exponent m is set to 2.

These data sets cover examples of data of low, medium and high dimensions. For every dataset, algorithms performed 20 times individually for their own effectiveness tests, each time with randomly generated initial solutions. The experimental results are shown in table 3, table 4 and table 5.

As shown in these tables, the hybrid ABC-KFCM obtained superior results than others in all of data sets. ABC-KFCM's search ability has being enhanced, and its optimize speed is faster. Also the experimental results show that when the size of data set (number of objects or clusters) is large, the new algorithm has the better clustering results.

Table 3. Clustering result of IRIS data

Algorithms	Average Error Number			Average Accuracy Rate (%)	Iterations
	Setosa	Versicolo	Virginica		
ABC-KFCM	0/50	2/50	2/50	97.33	10
KFCM	0/50	5/50	6/50	92.67	13
FCM	0/50	6/50	10/50	89.33	15

Table 4. Clustering result of wine data

Algorithms	Average Error Number			Average Accuracy Rate (%)	Iterations
	class1	class2	class3		
ABC-KFCM	3/59	8/71	2/48	92.70	12
KFCM	7/59	12/71	3/48	87.64	15
FCM	10/59	16/71	5/48	82.58	22

Table 5. Clustering result of glass data

Algorithms	Average Error Number						Average Accuracy Rate (%)	Iterations
	1	2	3	4	5	6		
ABC-KFCM	3/70	3/76	117	1/13	0/9	2/29	95.33	10
KFCM	6/70	8/76	3/17	1/13	1/9	3/29	89.72	15
FCM	9/70	10/76	4/17	2/13	2/9	5/29	85.05	17

6 Conclusions

Artificial bee colony algorithm is a new, simple and robust optimization technique. In this paper, an ABC algorithm is developed to solve clustering problems which is inspired by the bees' forage behavior. ABC-KFCM algorithm uses ABC algorithm to seek optimal solution as initial clustering-centers of KFCM algorithm, and then use KFCM algorithm to optimize initial clustering-centers, so as to get the global optimum. Above all, it solves the problems of KFCM. Experimental results show that the new algorithm is more accurate in clustering, higher efficiency and fewer the number of iterations.

References

1. Fayyad, U.M., Piatsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining, 1st edn. AAAI Press, Menlo Park (1996)
2. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
3. Izakian, H., Abraham, A.: Fuzzy clustering using hybrid c-means and fuzzy particle swarm optimization. In: Proceedings of the 2009 World Congress on Nature and Biologically Inspired Computing, Coimbatore, pp. 1690–1694 (2009)
4. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
5. Yang, M.S., Tsai, H.S.: A gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction. Pattern Recognition Lett. 29, 1713–1725 (2008)
6. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
7. Liu, J., Xu, M.: Kernelized fuzzy attribute C-means clustering algorithm. Fuzzy Sets Syst. 159, 2428–2445 (2008)
8. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
9. Zhang, D.Q., Chen, S.C.: A novel kernelized fuzzy C-means algorithm with application in medical image segmentation. Artificial Intelligence Med. 32, 37–50 (2004)

10. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
11. Karaboga, D.: An idea based on honey bee swarm for numerical optimization. Technical report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, Kayseri/Turkiye (2005)
12. Karaboga, D., Ozturk, C.: A novel clustering approach: Artificial bee colony (ABC) algorithm. Applied Soft Computing Journal (2008), doi:10.1016/j.asoc.2009.12.025
13. Zhang, C., Ouyang, D., Ning, J.: An artificial bee colony approach for clustering. Expert Systems with Applications 37, 4761–4767 (2010)