TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.5, May 2014, pp. 3786 ~ 3791
DOI: http://dx.doi.org/10.11591/telkomnika.v12i5.5110

A Hybrid Clustering Algorithm Based on Improved Artificial Fish Swarm

Lin Tian[1], Liwei Tian*[2]
[1]School of Electronic Engineering, Shenyang University,
21 South Wanghua Street, Shenyang, Liaoning, China, Ph: +8615904020165
[2]Liaoning Information Integration Technology Engineering Research Center of Internet of things
Lianhe Road, Shenyang Liaoning, China, Ph: +8613304035877
Corresponding author, e-mail: tianliwei@163.com

Abstract

K-medoids clustering algorithm is used to classify data, but the approach is sensitive to the initial selection of the centers and the divided cluster quality is not high. Basic Artificial Fish Swarm Algorithm is a new type of heuristic swarm intelligence algorithm, but optimization is difficult to get a very high precision due to the randomness of the artificial fish behavior. A novel clustering method based on improved global artificial fish swarm is proposed in this paper by analyzing the advantages and disadvantages of two algorithms, which has the ability to optimize the global clustering effect. The result of the experiment shows that quality of clustering is improved; the optimal central points and the clear division of data groups are obtained by the mathematical model combing improved fish swarm algorithm and K-medoids algorithm.

Keywords: *artificial fish swarm algorithm, K-medoids algorithm, clustering analysis*

1. Introduction

Cluster analysis is an important research direction of data mining; clustering is classifying data for different patterns based on the different characteristics of different objects [1]. The same objects have a high similarity degree, while objects of different groups vary greatly from each other, this form the law of distribution of the object and correlation between the data [2]. Since database collected lots of data, it requires scalability of algorithm and cluster quality. In this paper, K-medoids algorithm is used to divide clusters by calculating distance, dissimilarity, squared error and other parameters, this algorithm has strong robustness and flexibility, but it is susceptible to effects of the outliers and local extreme value, and randomly initialize parameters play a decisive role on the clustering results.

A novel bottomup optimization mode Artificial Fish Swarm Algorithm (AFSA) is used in this paper. AFSA use swarm intelligence of biosphere to solve optimization problems, as a generalized neighborhood search algorithm, by means of heuristic search strategy, its capacity of tracking changes rapidly gives algorithm the ability of global optimization, because of the characteristics of global convergence itself, the initial value can be set as fixed or random allowing parameters to be set in a wider scope [3]. AFSA has strong adaptability and parallelism, many behaviors combinations can be selected due to its good flexibility, and it can get better optimization performance which genetic algorithm and particle swarm optimization does not possessed. This artificial intelligence mode which is based on biological behavior is different from classical pattern, firstly is to design a single entity perception, behavioral mechanisms, then placed a group of entities in the environment so that they can solve the problem in environment interaction [4, 5]; however making the best reaction under the stimulation of the environment is the basic idea of AFSA. Literature [6] proposed reducing the search field to accelerate local search of artificial fish individual, but this optimization only took convergence speed into account making severe limitation of swarming and following behaviors of AFs, thus affecting the quality of the optimization. [7] introduced the K-means algorithm to speed up the iteration, but the performance was unstable because of many random processes in AFSA and it affected the practical application of the method. Using simulated annealing algorithm to improve AFSA, the approach in [8] modified preying behavior to avoid the degradation of artificial fish, although this hybrid algorithm overcame the shortcoming which

Received November 11, 2013; Revised December 28, 2013; Accepted January 10, 2014

easily fall into local minima, convergence time of the method was relatively long and it was not suitable to analysis huge data. Combining AFSA with clustering analysis algorithm based on grid and density, [9] obtained the number K of clusters automatically and it applied to arbitrary shape of data, better parallelism, but the quality of ultimately clustering quality was affected by the number and the size of grids which led to some limitations [10].

The traditional K-medoids has greater ability of local search, but is very sensitive to the initial cluster centers and easily falling into local optimum, if outliers are randomly selected as the initial centers, the whole quality of classification will decline. AFSA is less sensitive to initial values, even if its global optimization, has bad convergence and slower iteration rate in late period. Aiming at the advantages and disadvantages of both algorithms, this paper presents a global optimization idea to improve K-medoids clustering algorithm based on AFSA, the result of the test on a small data set shows that the improved algorithm obtains clear classifications and better performance.

2. Clustering Model

$X=(x_1, x_2, \ldots x_N)$ as the N data samples, x is the data representative point, C_i is an arbitrary cluster, O_i is the center of the cluster C_i, (j=1,2…,k). Algorithm is presented as follows.

Selected k objects in set X as the initial centers arbitrarily (O_1,O_2,…O_i…O_k), assigned the remaining data except for representative centers by the proximity principle to each cluster; in each cluster (C_i), chose a noncentral point O_j randomly, calculating total cost ΔE after using non-center instead of the original center point; If ΔE<0, then replace the original O_i with a non-center O_j, performing the above steps repeatedly until k centers is fixed [11, 12]. Cost function is used to evaluate the clustering quality improved. The function is defined as follows:

$$\Delta E = E_2 - E_1 \tag{1}$$

ΔE represents the change of absolute error standard, E_2 refers to the sum of dissimilarity degree between representative points and center points in the same cluster after replacing the centers, and E_1 represents the dissimilarity degree before replacing [13, 14]. Calculate ΔE, if ΔE<0, the effect of clustering has been improved, then use the new center.

3. Optimized AFSA

3.1. Description of the Basic Behaviors

Population of AFs is N, individual state of AF: $F = (f_1, f_2, \ldots f_n)$,[where f_i is optimization variables], the largest moving step is Step, vision is Visual, test time of preying behavior is Try_number, crowd factor is δ, food consistence $Y = f(F)$ (Y is the value of objective function).

3.1.1. Preying Behavior

As one of the basic habits of AF, the main principle is finding the area where there is a large food concentration by sense of sight and taste. Current state of AF is F_j, select a state F_j randomly around current location within its visual field, in the process of seeking optimal solution, if $Y_i < Y_j$, then F_j is a better state than the current one and move one step to this direction, default choose a new state and judge again, test Try_number times repeatedly, if still unable to get a better solution then move a random step [15, 16].

3.1.2. Swarming Behavior

To ensure the survival of fish populations, AF will gather to the center of adjacent partners. F_i still corresponds to the current state, perceive the AF number nf nearby and its central location F_c .if satisfied $Y_c / n_f > \delta \bullet Y_i$,it means the position was less congestion level, more food, then step forward to F_c, or implement preying behavior [17].

3.1.3. Following Behavior

In nature, when one or a few fishes have explored food, its neighbors will follow swarm to reach the food position [18]. Perception of the best state F_j within the vision, satisfied $Y_j / n_f > \delta \bullet Y_i$ which display the location was less crowing degree, more food, then make a step to F_j, or do preying behavior.

3.2. Improvement of AFSA

(a) In preying behavior, when a state of randomly selected F_j does not satisfy the moving condition it will choose random behavior, that is difficult to obtain high precision, AFs searching nearby the global extreme points circuitously at anaphase of convergence, which lead to an invalid calculation. In this paper, when preying failed, AFs choose to move a step to a better value comparing with the bulletin board records:

$$F_i(k+1) = F_i(k) + Step \bullet [F_{better}(k+1) - F_i(k)] \quad (2)$$

$F_i(k+1)$ and $F_i(k)$ denote respectively current position and next position after the movement, Fbetter is the better state recorded by bulletin board, comparing with random method it gives the possibility of a better forward and thus jump out of local optima, preventing AFs in the local concussion at a standstill.

(b) In AFSA, the parameter crowding factor δ is to avoid overcrowding of AF and δ is a fixed value in global algorithm, this approach that make δ a constant will lead to mutual exclusion between individuals which are adjacent to global optimization solution, so AFs cannot gather to extreme points accurately and contrast crowding condition after every iteration will increase the computational cost too. Improved method defines the initial congestion factor δ =0.75, when Try_number = 180, ignoring the congestion factor namely $\delta \bullet n_f = 1$ in initial stages, it needs to limit the size of artificial fish, but in the latter part fishes have already gathered in optimum, default δ can reduce calculation amount and execution time of the algorithm, in this way not only does it improves the operation efficiency of AFSA but also has no effect on convergence.

(c) In order to solve the problem of centers of K-medoids by AFSA, when swarming and following behavior failed, preying behavior is carried out, thus increasing the convergence time and computation. So we renew the behavior as follows: substitute random swim for preying behavior after failing in movement. And the step is adaptive step-size. The method overcomes the problem that AFs aggregated at local solution and missed the global ones and enhance the quality of solutions.

4. A Hybrid Clustering Algorithm Based on Improved ASFA

4.1. Definitions of Improved AFSA

Definition 1: (adaptive step-size of AF) Adaptive step-size represents the moving distance of AF changing with iterations. Adaptive step-size is defined as:

$$F_{i+1} = F_i + Step \bullet Rand() \quad (3)$$

Definition 2: (clustering evaluation criterion) Objective function measures dissimilarity between representative points and objects, which means the compact degree of data distribution between classes, the objective function is defined as:

$$E = \sum_{j=1}^{k} \sum_{X \in C_j} | X - O_j |^2 \quad (4)$$

4.2. The Procedure of the Mixed Clustering Based on Optimized AFSA

Step 1: Initialize the initial value of AF parameters, calculate food consistence at current position by objective function;

Step 2: Carry out the algorithm through behavior's condition, update the location of AFs by preying, swarming and following behaviors, data density refer to food concentration; contrast food consistence within vision distance to select solution, with its state recorded in the bulletin board, finally fishes gather in the areas of high data density;

Step 3: Each state of AF represents a decision variable, and the fitness value is computed by objective function, evaluate optimization degree and record; repeat 2) 3), update the location information of AFs until the termination condition is met;

Step 4: According to bulletin board information and the location of fishes, choose input parameters for K-medoids, namely the initial center and the number of clusters; using K-medoids for cluster analysis until meeting minimum within-class scatter of data. The minimum within-class M is presented as follows:

$$M = \min E \tag{5}$$

The flowchart shows procedure of approach in Figure 1:

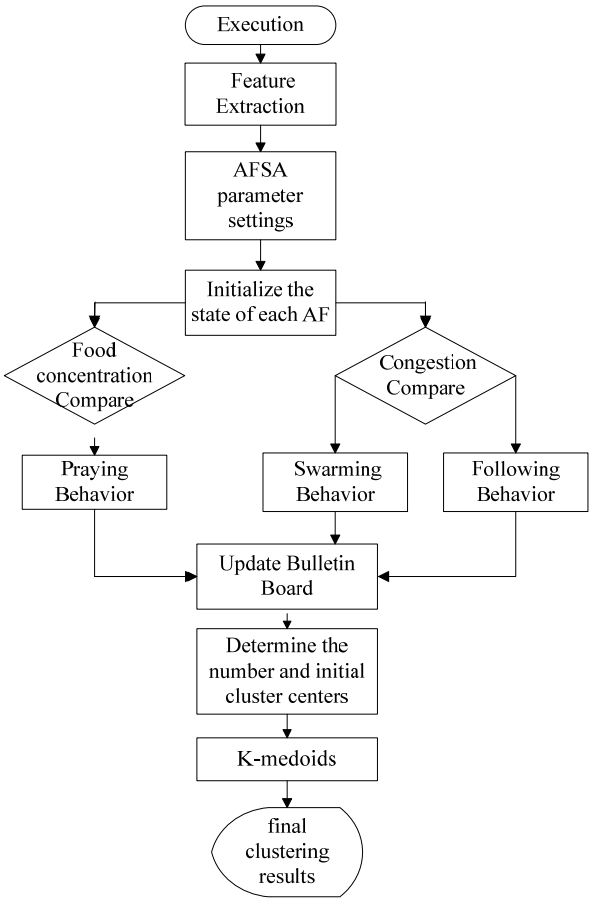

Figure 1. Flowchart of Clustering Algorithm Based on Optimized AFSA

5. Simulation

Simulation data include 300 3D data; running environment for experiment: Pentium(R), 3.00G; Programming environment: Matlab7.12.0 (R2011a); AFSA parameters are set as follows: Step is 0.2, Visual is 100, δ is 0.75, Try_number(iteration times) is 200,N (the total number of AF) is 50.

In the simulation, it classify the data by two hybrid clustering algorithms, comparison results of the approach this paper proposed and basic hybrid clustering algorithm. Operation

result of classic hybrid method shown in Figure 2, Figure 3 shows performance of improved approach.

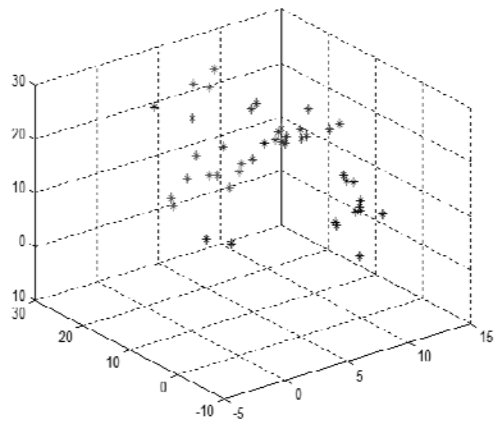

Figure 2. Optimization Graph of Basic Clustering Algorithm Based on AFSA

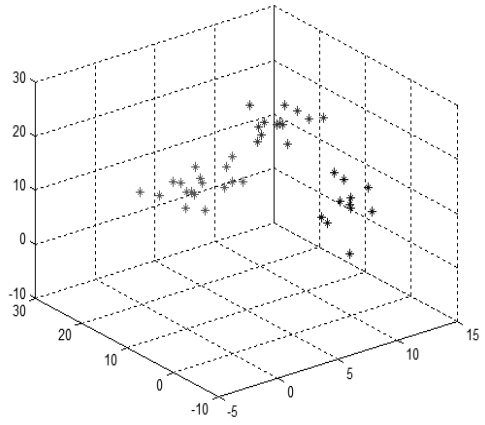

Figure 3. Optimization Graph of Improved Method

AFs find the centers in the 3D data, as shown in Figure 2 aggregation effect is not clear, a few individuals moves to local clusters; optimization result approximate to global data-intensive areas that can be seen from the iteration route in Figure 3; comparison of performance shows the edge of clusters is more obvious by improved method on the same condition, the aggregation of position is closer so that we can obtain a higher accuracy of the division to verify the advantages of this algorithm.

Table 1. The Results of Two Algorithms

	Total Number of AF	Iteration Times	Iteration Time /ms	Correct Rate
Method in [6]	50	200	762	89
Proposed Method	50	200	685	93

It is shown in Table 1 the proposed method reduced not only the iteration time but also calculation amount on the same condition, and the accuracy is also improved.

4. Conclusion

Hybrid clustering is widely applied in decision problem and early warning at current research. Comparison of experimental results shows improved AFSA hybrid clustering algorithm make similar data gather obvious, the model is more stable and accurate than the old one, distinguish samples precisely while also improving the cluster quality and obtaining better centers with clear division, reducing computation amount is also a breakthrough. The model of modern intelligence algorithm based on animal autonomous body combines K-medoids, this novel method avoids the weakness of dependency on Cluster initialization, and overcomes slow iteration speed in late period; its good parallelism can be effectively applied in various fields, it also plays a major role in knowledge discovery, information forecast and decision analysis. However, the convergence speed issue remains to be improved and researched.

Acknowledgements

This work was supported by the Liaoning National Natural Science Foundation (Grant No. 2013020011), the International S&T Cooperation Program of China (ISTCP) under Grant

2011DFA91810-5 and Program for New Century Excellent Talents in University of Ministry of Education of China under Grant NCET-12-1012.

References

[1] Jiawei Han, Kamber M. Data Mining: Concepts and Techniques. 2007; 252-272.
[2] Ningxia Xia, Yidan Su. An Efficient K-medoids Clustering Algorithm. *Journal of Application Research of Computers*. 2010; 27(12): 4517-4519.
[3] Xinmin Tao, Jing Xu, Libiao Yang. An Improved Hybrid Algorithm Based on Particle Swarm Optimization and K-means Algorithm. *Journal of Electronics &Information Technology.* 2010; 32(1): 92-94.
[4] Xiaohua Wang, Jie Shen, Rongbo Wang. A New Hybrid Algorithm Based on Ant Colony and Clustering. *Journal of Hangzhou Dianzi University*. 2010; 30(1): 26-27.
[5] Yichuan Shao, Xingjia Yao, Liwei Tian, Hanning Chen. *A Multi-swarm Optimizer for Distributed Decision Making in Virtual Enterprise Risk Management*. 2012.
[6] Yujun Fan, Dongdong Wang, Mingming Sun. Improved Artificial Fish Swarm Algorithm. *Journal of Chongqing Normal University (Natural Science Edition)*. 2007; 24(3): 24-26.
[7] Bai Liu,Yongquan Zhou. A Hybrid Clustering Algorithm Based on Artificial Fish Swarm Algorithm. *Journal of Computer Engineering and Applications.* 2008; 44(18): 136-138.
[8] Jia Liu. Improved Artificial Fish Swarm Algorithm and Its Applications in Function Optimization. *Journal of Shijiazhuang Institute of Railway Technology*. 2011; 10(3): 33-36
[9] Xudeng He,Liangdong Qu. *Artificial Fish Swarm Clustering Algorithm*. Application Research of Computers. 2009; 26(10): 3666-3668
[10] HongWei Zhao. *A Resource Discovery Mechanism on Cloud Computing System*. 2012.
[11] Juanying Xie, Shuai Jiang. *A Simple and Fast Algorithm for Global K-means Clustering*.Second International Workshop on Education Technology and Computer Science (ETCS). IEEE. Wuhan. 2010; (2); 36-40.
[12] Bingru Yang, Danyang Cao. *An Improved K-medoids Clustering Algorithm*.2nd International Conference on Computer and Automation Engineering (ICCAE). IEEE Singapore. 2010: 132-135.
[13] Lipo W, Xiuju F. Data Mining With Computational Intelligence. 2005.
[14] Yuzhen Zhao, Xiyu Liu, Hua Zhang. The K-Medoids Clustering Algorithm with Membrane Computing. *TELKOMNIKA Indonesian Journal of Electrical Engineerin*. 2013; 11(4): 2050-2057.
[15] Doaa M Atia, Faten H Fahmy, Ninet M Ahmed, Hassen T Dorrah. A New Control and Design of PEM Fuel Cell System Powered Diffused Air Aeration System. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012; 10(2): 291-302
[16] Chuxiao Li, Zhu Ying, Shi JunTao, Song JiQing. *Method of Image Segmentation Based on Fuzzy C-Means Clustering Algorithm and Artificial Fish Swarm Algorithm*. International Conference on Intelligent Computing and Integrated Systems (ICISS). IEEE Press. 2010: 254-257..
[17] Yongming Cheng, Mingyan Jiang, Dongfeng Yuan. *Novel Clustering Algorithms based on Improved Artificial Fish Swarm Algorithm*. Sixth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE Press. 2009; (3): 141-145
[18] Dongfeng Yuan, Mingyan Jiang. Artificial Fish Swarm Algorithm and Its Applications. 2012; 43-66: 110-130