K-means clustering algorithm based on bee colony strategy

Zhenrong Zhang¹, Jiayi Lan² and Zhenrong Zhang*¹
¹ College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong, 266590, China
² College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong, 266590, China
*Corresponding author’s e-mail: wutt61@chinaunicom.cn

Abstract. The traditional swarm intelligence optimization K-means algorithm has some problems, such as poor global search ability and blind area selection of initial center points, which leads to the reduction of clustering availability. In order to avoid the above limitations, this paper proposed IABC K-means algorithm. Firstly, the employed bees stage in the traditional artificial bee colony ABC algorithm uses the current colony optimal solution information to guide its optimization search. Secondly, this paper wanted to solve the search ability of the employed bees and expand the information sharing range between various individuals, a random guidance mechanism is proposed in the onlookers bees stage. Finally, chaotic sequence is introduced in the scout bee stage to accelerate the convergence speed of the algorithm. IABC algorithm is proposed and applied to K-means clustering algorithm to improve the poor global search ability of K-means algorithm and the random selection of initial center points. Experiments show that the IABC and IABCK-means proposed in this paper effectively improves the clustering availability.

1. Introduction

Artificial bee colony algorithm (ABC)¹ is a relatively novel algorithm inspired bee colony intelligent algorithm, which imitates the foraging (honey Collection) process of bees. Although many swarm intelligence optimization algorithms have emerged in recent years. For example genetic algorithm², differential evolution algorithm³, ant population algorithm⁴, particle swarm optimization algorithm⁵, whale optimization algorithm, etc., compared with them, ABC algorithm has few parameters, simple implementation, high precision and limited, and has been applied to the fields of data mining, information processing, clustering and so on.

In the artificial bee colony algorithm (ABC), we can divide bees into three categories, namely employed bees, onlookers bees and scout bees. Many scholars improve it in order to solve the problem of convergence speed of the algorithm and increase the population diversity. Artificial bee colony algorithm was proposed by Gao⁶ et al. based on differential evolution operator to aim at this problem. Zhu⁷ et al. used some optimization methods to enhance the local search of hired bees. Tsai⁸ et al. made use of the universal gravitation formula in the stages of employed bees and onlookers bees, which improved the interaction ability of artificial bee colony algorithm.

According to the defects of artificial bee colony algorithm and the characteristics of K-means algorithm. Firstly, the employed bees stage in the traditional artificial bee colony ABC algorithm uses the population optimal solution information to guide its optimization search. Secondly, this paper wanted to solve the search ability of the employed bees and expand the range of sharing information between
various individuals, a random guidance mechanism is proposed in the onlookers bees stage. Finally, chaotic sequence is introduced in the scout bees stage to increase the convergence speed of the ABC algorithm. IABC algorithm is proposed and applied to K-means clustering algorithm to improve the poor global search ability of K-means algorithm and the random selection of initial center points.

2. IABC algorithm

2.1. Employed bees stage based on current colony optimal solution

The search ability of the position update formula in the employed bees stage is insufficient, the iteration in the neighborhood search is random and the global optimal solution is easy to occur. Inspired by the differential evolution algorithm and the existing improved artificial bee colony algorithm, this paper makes the most of the optimal solution of the current colony to guide the optimization search in the employed bees stage, and puts forward the following search equation.

\[
V_{ij} = X_{best,j} + R_y \times (X_{best,j} - X_{kj})
\]  

Where \( X_{best} \) is the optimal feasible solution of the current colony; \( R_y (-1 \leq R_y \leq 1) \) is a random number.

2.2. Onlookers bees stage based on random guidance mechanism

The search ability of the employed bees stage is insufficient and the scope of information sharing among various individuals needs to be expanded, a random guidance mechanism is proposed in the onlookers bees stage, which can be described by equation.

\[
V_{ij} = X_{kj} + R_y \times (X_{ij} - X_{kj})
\]  

Where \( k \in (1,2,\cdots,NP) \) and \( k \neq i, i = 1,2,\cdots,NP \), \( NP \) is the number of honey sources.

2.3. Scout bees based on chaos theory

In the later stage of the algorithm, the similarity of food source location becomes higher and the location update speed is slow, resulting in the decline of search ability. In this paper, chaotic sequence is introduced in the scout bees stage to expand bees colony diversity and increase the convergence speed of the algorithm. In this paper, Kent mapping is used, which is defined by:

\[
C_i^{\beta+1} = \begin{cases} 
  \frac{c_i}{\beta} & \text{if } c_i < \beta \\
  \frac{1-c_i}{1-\beta} & \text{if } \beta \leq c_i \leq 1 
\end{cases}
\]  

Where \( C_i \) represents the chaotic mapping variable of the ith time. It is a random number and it ranges from 0 to 1. \( \beta \) is a parameter in the range of (0,1). For the convenience of calculation, we take \( \beta = 0.7 \).

The search formula of chaotic sequence based on Kent map is as follows:

\[
V_{ij} = X_{i,\min} + C_j \times (X_{i,\max} - X_{i,\min})
\]  

3. IABCK-means algorithm

In this section, we combined with the ABC algorithm search process and K-means algorithm idea, this paper proposes a new density based fitness function.

\[
\text{fitness}_i = \frac{\text{count}(D(C_i))}{\sum_{s_i \in C_j} d(x_i, C_j)}
\]
\[ \text{dist}(x_i, x_j) = \sqrt{|x_i - x_j|^2} \]  

(6)

Where numerator indicates the number of class ith points, denominator represents the sum of distances from the objects in the class to the center point.

The IABC K-means specific steps of clustering algorithm are as follows:

1. Initializing the bee colony, the amount of employed bees and onlookers bees are same as the amount of honey sources, the maximum amount of iterations is \( \text{MAXITER} \), the maximum amount of searches is \( \text{Limit} \), and the amount of cluster clusters is \( k \) and the resulting colony is \( \{Z_1, Z_2, \cdots, Z_M\} \).

2. Firstly, we cluster the initial bee colony and the fitness value of each bee was calculated according to equation (5). Then we sort according to the fitness value, and take the first half as the employed bees and the second half as the onlookers bees.

3. According to equation (1), each employed bees will find new honey sources by one near the honey source at the beginning, and then we calculate the fitness value of the honey resource. Employed bees judge whether to update the honey source according to the "greedy principle."

4. Probability selection of onlookers bees based on roulette principle. A employed bees is selected by the onlookers bees according to the probability. If the probability is higher, the possibility of choice is bigger. Then, the onlookers bee carries out neighborhood search according to formula (2), compares the advantages and disadvantages of the two honey sources according to the "greedy principle", and carries out selective reservation.

5. After the onlookers bees search, this algorithm takes the obtained position as the cluster center, cluster the data set, and update the bee colony with each new cluster center according to the cluster division.

6. If the result of an onlookers bees does not change after the maximum search times, the onlookers bee becomes a scout bee and carries out neighborhood search according to formula (4), it generates a new position to replace the original position.

7. If the current iteration is greater than the maximum number of iterations, the iteration ends and the algorithm ends; Otherwise, it will turn to step 2 and increase the number of iterations by one.

8. This algorithm output the clustering results and the algorithm ends.

4. Experimental results and discussion

In the IABC algorithm performance test, the amount of employed bees and onlookers bees is set to 10, the maximum amount of iterations is 2000, and the maximum amount of searches is 100. When the maximum amount of searches exceeds 100, the employed bee becomes a scout bee. This paper compares the algorithm with the ABC algorithm and EABC algorithm[12] on four standard test functions[13] and analyzes the effect. The IABCK-means algorithm parameters are set as follows: the maximum amount of iterations is 100, the maximum amount of searches is 10, and the amount of employed bees and onlookers bees is 10. The algorithm clusters on three data sets to see the change trend of observation results. Table 1 is the information of data set in our experiment and Table 2 is the algorithm test function information.

| Date set | Number of samples | Attribute dimension | Number of categories |
|----------|-------------------|---------------------|---------------------|
| Iris     | 150               | 4                   | 3                   |
| Glass    | 214               | 10                  | 6                   |
| Gamma    | 19020             | 11                  | 2                   |
Table 2. Expression of test function.

| Function  | Expression                                                                 |
|-----------|-----------------------------------------------------------------------------|
| Sphere    | $f_1(x) = \sum_{i=1}^{n} x_i^2$                                              |
| Rastrigin | $f_2(x) = \sum_{i=1}^{n} (x_i^2 - 10\cos(2\pi x_i) + 10)$                   |
| Rosenbrock| $f_3(x) = \sum_{i=1}^{n} 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2$              |
| Griewank  | $f_4(x) = \frac{1}{4000} \sum_{i=0}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$ |

Figure 1. Convergence diagram of three algorithms in Sphere function fitness.

Figure 2. Convergence diagram of three algorithms in Rastrigin function fitness.
Figure 3. Convergence diagram of three algorithms in Rosenbrock function fitness.

Figure 4. Convergence diagram of three algorithms in Griewank function fitness.

Figure 5. Convergence trend of IABCKmeans fitness on different data sets.

From Figure 1 to Figure 4, it can be seen that the traditional algorithm will have different degrees of slow convergence speed on the four test functions and it is easy to fall into the current population optimal solution. The number of iterations of the EABC algorithm is significantly reduced, but it still has disadvantages in the global optimization ability. This IABC algorithm improves the position update formula to further reduce the number of iterations and the slow iteration problem in the later stage. As can be seen from Figure 5, the new fitness function makes the IABC K-means algorithm change little on the three data sets, can quickly and efficiently search the location of the optimal solution, and find out the local optimal solution in the data set optimization. Therefore, the IABCK-means algorithm in this paper performs well in global optimization, and also has the ability to jump out of local optimal solution, and the algorithm is more effective.

5. Conclusions and Future Work
With the rapid development of artificial intelligence industry, our life is full of all kinds of data. As a branch of artificial intelligence, data mining has been widely used. Cluster analysis, as an important algorithm of data mining, plays an indispensable role. K-means clustering algorithm has become a common algorithm for cluster analysis because of its simplicity and fast convergence speed. Aiming at the problems of poor global search ability and random selection of initial center points of K-means algorithm, IABC algorithm is proposed and applied to K-means clustering algorithm. Experiments show that different strategies are adopted for the three bee colonies of ABC algorithm, which effectively enhances its global optimization and local optimization ability. The next work will focus on the combination of artificial bee colony algorithm and hybrid attribute clustering algorithm, and how to ensure the privacy disclosure of cluster center and cluster in the clustering process.

References
[1] Karaboga D . An idea based on honey bee swarm for numerical optimization[J]. 2005.
[2] Tang K S , Man K F . Genetic algorithms and their applications[J]. Signal Processing Magazine IEEE, 1996, 13(6):22-37.
[3] Storn R , Price K . Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces[J]. Journal of Global Optimization, 1997, 11(4):341-359.
[4] A M D , B C B . Ant colony optimization theory: A survey[J]. Theoretical Computer Science, 2005, 344(2–3):243-278.
[5] Tregle I C . The particle swarm optimization algorithm: Convergence analysis and parameter selection[J]. Information Processing Letters, 2003, 85( 6):317-325.
[6] Gao W , Liu S . Improved artificial bee colony algorithm for global optimization[J]. Information Processing Letters, 2011, 111(17):871-882.
[7] Zhu G , Kwong S . Gbest-guided artificial bee colony algorithm for numerical function optimization[J]. Applied Mathematics & Computation, 2010, 217(7):3166-3173.
[8] Tsai P W , Pan J S , Liao B Y , et al. Enhanced artificial bee colony optimization[J]. International journal of innovative computing, information & control: IJICIC, 2009, 5(12):5081-5092.
[9] Yang D , Liu Z , Zhou J . Chaos optimization algorithms based on chaotic maps with different probability distribution and search speed for global optimization[J]. Communications in Nonlinear Science & Numerical Simulation, 2014, 19(4):1229-1246.
[10] Alatas B . Chaotic bee colony algorithms for global numerical optimization[J]. Expert Systems with Applications, 2010, 37(8):5682-5687.
[11] A, Marco Dorigo , and C. B. B . "Ant colony optimization theory: A survey." Theoretical Computer Science 344. 2–3(2005):243-278.
[12] Raghav R S , Pothula S , Ponnurangam D . An enriched artificial bee colony (EABC) algorithm for detection of sinkhole attacks in Wireless Sensor Network. 2017.
[13] Abraham A , Jatoth R K , Rajasekhar A . Jatoth and A, Rajasekhar.: Hybrid Differential Artificial Bee Colony Algorithm.