A clustering algorithm based on the combination of screening strategy and swarm  

yudifei  
Jitong college of liaoning university of petrochemical technology, Fushun 113000, China  
1113092789@qq.com  

Abstract. aiming at the problem that the k-means clustering algorithm is affected by the initial clustering center, a k-means clustering optimization algorithm combining the screening strategy and artificial colony (ABC) was proposed. This algorithm operates in an unsupervised way, separating data from outlier points through screening, combining the advantages of ABC algorithm, using the objective function of ABC algorithm as the measurement function of initial clustering, and improving the effectiveness and accuracy of k-means clustering by changing the initial clustering center.

1. Introduction  
With the development of science and technology, more and more data are produced in various industries. In the face of these massive network data, the way of manual processing is no longer realistic. The use of big data technology can store a variety of different data, through the data mining of these not directly related data, in order to extract useful data for visualization, to promote the development of the industry. Cluster analysis is an effective technique in big data, through which knowledge can be mined that cannot be found only through the data itself.

Cluster analysis is the analysis of unrelated and disorderly data set objects, the use of a variety of algorithms to classify the data set, which have a similar degree of acquaintance data together. The commonly used classification method is to classify the data by the degree of difference between the data, which is often determined by the multidimensional attributes of the data. The k-means clustering algorithm is a relatively efficient clustering method, which is a kind of partition clustering proposed by MacQueen. It is not only easy to operate, but also has high performance. However, this algorithm also has some disadvantages, such as its clustering results are susceptible to noise, when the initial clustering center is selected, the clustering results will be different each time. Therefore, many scholars are committed to the improvement of k-means and use the combination of various algorithms to improve the shortcomings of clustering algorithm. For example, Lu shaojuan et al. [6] combined with artificial fish swarm algorithm to improve the local convergence of k-means clustering algorithm. Jia Ruiyu, Li Zhen et al. [7] proposed hierarchical k-mean clustering algorithm, using hierarchical ideas to divide and merge.

Artificial colony algorithm (ABC algorithm) is a kind of extensive swarm intelligence algorithm, which is an algorithm developed by Karaboga after observing the foraging behavior of bees and discovering its rules. ABC algorithm only needs to consider the problem itself, without knowing the special properties of the problem. The flow of the algorithm is similar to the flow of bees collecting
honey. Among them, honey represents a solution, and the swarm gets an optimal solution, that is, the optimal solution. This algorithm has high precision and is an effective classical algorithm.

2. Introduction to relevant algorithms

2.1. **K-means clustering algorithm**

K-means clustering algorithm is one of the clustering algorithms, which belongs to partition clustering. In this kind of clustering problem, the clustering effect is measured by distance. So we can set up a data set \( X = \{x_1, x_2, ..., x_n\} \), where the data is d-dimensional data, expressed as \( x_i = \{x_{i1}, x_{i2}, ..., x_{id}\} \), set the cluster of data clustering as \( C = \{C_1, C_2, ..., C_k\} \). The initial clustering center is \( c_i(j=1,2,...,k) \), and the number of data in the ith cluster is \( N_i \).

K- algorithm flow of mean clustering algorithm is as follows:

Step(1) select k initial clustering center points in data set X;

Step(2) calculates the similarity of data in the whole data set according to formula (1), and distributes the objects with high similarity into a cluster to make the data objects in each cluster as close as possible in the initial stage. This gap is usually expressed by Euclidean distance.

\[
d(c_i - c_j) = \sqrt{\sum_{k=1}^{d}(c_{ik} - c_{jk})^2}
\]  

(1)

Where \( c_i, c_j \) is the data in the KTH cluster.

Step(3) continuously updates the average value of \( c_1, c_2, ..., c_k \) objects in each cluster according to formula (2)

\[
c_i = \frac{1}{N_i} \sum_{j=1}^{N} x_j \quad (x_j \in C_i)
\]  

(2)

Where \( x_j \) is the data sample in cluster \( C_i \) and the mean of the data objects in cluster \( C_i \).

Step(4) calculates the value of criterion function E according to formula (3). If convergence occurs, the algorithm is terminated; otherwise, step 2 is returned.

\[
E = \sum_{i=1}^{k} \sum_{j=1}^{N} dis(x_j, c_i)^2
\]  

(3)

Step(5) the final output clustering results.

2.2. **Artificial bee colony algorithm**

The principle of artificial colony algorithm is relatively complex, but the implementation is relatively simple. In this algorithm, there are three kinds of bees: foraging bees, observing bees and detecting bees. Among them, honey bees will share the location of the source by rocking dance, recruiting observation bees to mine together. The observed bee, which makes up half the population, waits in the swing zone, selected from known nectar sources. The individual in the swarm will always ensure the optimal location of its choice and conduct a domain search based on the information. When a honey source reaches the maximum mining times, the current position of the bee will become a scout bee. Scout bees search the neighborhood for new sources of honey, and then become bee hunters.

ABC algorithm flow as follows:

Step (1) initialization: by default, the number of random honey source generated is SN; the number of bee colony individuals is NP; SN=NP/2; the maximum number of cycles is MCN; the maximum number of collection is limit; the data dimension D; the SN honey source solution is \( X_0(i=1,2,...,SN; j=1,2,...,D) \); the fitness of the solution is \( fit \).
\( X_j = X_{j\max} + \text{rand}(0,1)(X_{j\max} - X_{j\min}) \) \hspace{1cm} (4)

\[ \text{fit} = \frac{1}{1 + \text{fit} + \text{abs}(...)} \] \hspace{1cm} (5)

The \( \text{rand}(0,1) \) in formula (4) is the random number between (0, 1), which controls the search range of the algorithm, and the \( X_{j\max} \) and \( X_{j\min} \) represent the j dimension data.

Step (2) bee picking stage: Bees were collected from formula (6) to obtain new honey sources, and their fitness \( \text{fit} \) was calculated. After comparison, honey sources were selected according to the greedy criterion.

\[ V_j = X_j + \Phi_j(X_j - X_w) \] \hspace{1cm} (6)

In formula (6), \( \Phi_j \) is a random number between [-1, 1], \( k \neq j \).

Step(3) observation bee scale: After completing the honey source location information sharing, calculate the probability \( P_i \), the observation bee will choose according to the roulette principle. If the roulette is successful, follow the bees, collect new sources of honey, and update the location of the source. If the new honey source fitness is better than before, observing the bee will save the new honey source solution, otherwise it will save the original honey source solution, at the same time the number of honey source collection plus 1.

\[ P_i = \frac{f_{is}}{\sum_{i=1}^{SN} f_{is}} \] \hspace{1cm} (7)

Step (4) bee detection stage: The honey source collection times are counted. If the honey source collection times reach the limit, the colony still fails to find the solution with the highest fitness, then the bees will discard the honey source, and their identity will change to the detection bee, and the search will be conducted again through formula (6) until a new honey source is found.

Step (5) determine the current optimal solution. If it is the global optimal, the solution will be output to complete the algorithm. Otherwise, the algorithm will go to step (2) and loop again. In this process, the three species of bees are constantly changing identity, division of labor, circular search.

3. **K- Optimization of initial center of mean clustering algorithm**

3.1. **Initial cluster screening strategy**

k mean clustering algorithm can achieve better clustering effect when analyzing large-scale data sets, especially for spherical or spherical-like structure data sets. k- mean clustering algorithm can quickly discover the structure of data and cluster data with high similarity. However, when the data set is a non-convex structure, the k clustering algorithm often has poor recognition ability. Meanwhile, k mean clustering algorithm is easy to be affected by noise points and edge points, and reduces the clustering effect, which makes the k mean clustering algorithm suitable for processing data limited. However, when the data set is non-convex, k clustering algorithm often has poor recognition ability. At the same time, the k-means clustering algorithm is susceptible to the influence of noise points and edge points, and the clustering effect is reduced, which makes the k-means clustering algorithm has limited data suitable for processing.

In the k-means clustering problem, the better the cluster center is selected, the higher the similarity of data in the class, and the larger the data difference between the classes, the higher the clustering
efficiency will be. In the algorithm, Euclidean distance is generally used to represent the degree of similarity between the data.

In order to solve the problem of abnormal point influence in the center of the initial cluster and make the data of the initial cluster center more complete, the initial data set can be processed. In the initial stage of data, the filtering strategy is used to select the data sets to obtain a more concentrated and complete cluster. In order to ensure that the data after a filter is sufficiently concentrated, the density \( Q \) needs to be calculated for each data point \( x_i \) in the initial dataset \( X = \{x_1, x_2, \ldots, x_n\} \).

\[
Q_x = \frac{1}{\sum_{x \in G} D(x_i, x)}
\]  

(8)

\( G \) is the set of \( m \) nearest neighbor data points of \( x_i \); \( D(x_i, x) \) is the distance between the data point \( x_i \) and \( x \). The smaller \( D(x_i, x) \) is, the tighter the data elements are.

According to the calculation of density, a new dataset \( X \) can be obtained by deleting the data whose density is less than \( \sum_{n \in X} n \) from the dataset \( X^* \). By this method, the effect of outliers is reduced to obtain a relatively tight initial data set.

3.2. Improved bee colony algorithm

The essence of k-means algorithm is the minimum optimization process of criterion function \( E \). The algorithm’s random selection of the initial clustering center easily leads to the algorithm falling into local optimization. In order to have better clustering effect, the similarity degree and difference degree of the members in the set are the key. Therefore, the clustering criterion function can be used as the objective function of the artificial bee colony by combining with the artificial bee colony algorithm.

Searching for the best honey is the process of selecting the best initial center in the cluster. Its correspondence is shown in the following table:

| Honey gathering process | Honey location | Honey Source Size | scouting speed | Optimal Honey Source | Optimal cluster centers |
|-------------------------|----------------|-------------------|---------------|----------------------|------------------------|
| Cluster iterations      | Cluster Center | Location          | convergence rate |                      |                        |

In combination with the artificial colony algorithm, the new data set \( X^* \) after screening is initialized, and sample \( Y = \{y_i \in X^*, i = 1, 2, \ldots, n\} \), \( d \) is data dimensions. The initial population number is \( SN \) bees, each bee represents the division of a cluster, \( Z = \{z_1, z_2, \ldots, z_k\} \) is a \( d \)-dimensional vector, and \( k \) is the number of clusters. Design new fitness function as \( \text{fit}_{Y} = \frac{1}{1 + E} \), with artificial swarm algorithm, where \( E \) is the criterion function defined by formula (3).

The k-means clustering algorithm based on artificial bee colony [11] is as follows:

Step(1) initializes the parameter and sets the number of colonies; the maximum iteration number is \( MCN \); the maximum collection number is limit; the initial number of clusters is \( k \).

Step(2) initializes the colony and generates SN bees randomly. The bees are composed of \( k \times j \) dimension vectors. The data are divided and \( \text{fit}_{Y} \) is calculated according to the new fitness function.

Step(3) carry out field search according to formula (6). When honey sources are found by bee picking, a comparison is made. If the new honey source fitness is higher than the original honey source, keep the new honey source position, otherwise keep the original honey source position, the number of iterations plus one.
Step(4) observe the bee will roulette principle to select, according to formula (8) to search the domain, the new solution will be as the clustering center.
Step(5) if there is no change after the limit collection, the scout bee will randomly generate a new position to replace the bee.
Step(6) determines the number of iterations, and when the maximum is reached, Step (8) is carried out; otherwise, step (3) is returned.
Step(7) outputs the initial clustering center and relative fitness value, and the algorithm ends.

4. Experimental data
To test the effectiveness of the improved algorithm in this article, the Iris, Wine, and Red Wine data sets from the machine learning database were used. Experimental environment: the host CPU is i7, the main frequency is 4.6ghz, and the running memory is 8.0g. The software runs the original k-means algorithm, the artificial colony k-means algorithm and the algorithm in this paper 20 times respectively, and calculates the mean value and standard deviation obtained by the three algorithms. Since the number of samples and the number of data attributes in the Iris, Wine and Red Wine data sets are different, the clustering Numbers are set respectively. See table 1 for the data. The maximum search iteration number MCN of the colony algorithm is set to 500, and the maximum collection number of the same solution is limited to 10. The calculated experimental data are shown in table 2 to 4.

| Data set name | Number of samples included | Number of samples | Number of clusters |
|---------------|---------------------------|-------------------|------------------|
| Iris          | 140                       | 4                 | 3                |
| Wine          | 174                       | 13                | 3                |
| Red Wine      | 1453                      | 12                | 3                |

Table 3 effect of Iris data clustering class

| Algorithm      | Convergence time | E Min     | E Max     | E average | E standard deviation |
|----------------|------------------|-----------|-----------|-----------|----------------------|
| K- mean        | 0.0061s          | 100.9403  | 122.2654  | 115.1545  | 18.2189               |
| ABC-k mean     | 2.35s            | 95.5842   | 105.2742  | 103.0478  | 0.0310                |
| Algorithm of this paper | 2.24s          | 93.4354   | 97.5437   | 96.4756   | 0.0135                |

Table 4 effects of Wine data clustering

| Algorithm      | Convergence time | E Min     | E Max     | E average | E standard deviation |
|----------------|------------------|-----------|-----------|-----------|----------------------|
| K- mean        | 0.0085s          | 1.6425e+4 | 1.8425e+4 | 1.7458e+4 | 482.1538              |
| ABC-k mean     | 5.632s           | 1.6348e+4 | 1.6789e+4 | 1.6414e+4 | 25.5498               |
| Algorithm of this paper | 5.324s          | 1.59478e+4| 1.6042e+4 | 1.5984e+4 | 16.4564               |
Table 5 Red Wine data clustering effect

| Algorithm          | Convergence time | E Min      | E Max      | E average | E standard deviation |
|--------------------|------------------|------------|------------|-----------|----------------------|
| K-mean             | 0.0258s          | 2.2705e+6  | 2.6451e+6  | 2.4215e+6 | 658.1584             |
| ABC-k mean         | 15.354s          | 2.2541e+6  | 2.4385e+6  | 2.3025e+6 | 35.5489              |
| Algorithm of this paper | 13.152s          | 2.1354e+6  | 2.1756e+6  | 2.1551e+6 | 24.3524              |

From the above data analysis, it can be seen that the original k-means algorithm converges quickly and completes fast clustering, but the test function value changes greatly. When the initial clustering center is selected differently, the result deviation is large. Therefore, the original k-means algorithm is unstable and greatly affected by the center, and local optimization can be obtained. In the experiment of ABC-k mean clustering algorithm, it can be seen that although the combination with ABC algorithm increases the time complexity and makes the convergence time longer, the fluctuation of E value is reduced to some extent and the clustering effect is improved. From the experiment, it can be seen that compared with the other two algorithms, the convergence time is improved, the change of E value is more stable, and the clustering effect is improved. With the addition of filtering strategy, the central data is more centralized, and the application scope of clustering algorithm in data types is more extensive.

5. Conclusion

In this paper, considering the advantages and disadvantages of k-means clustering algorithm and ABC algorithm, aiming at the problem that the clustering effect is affected by the initial clustering center, this k-means clustering algorithm is proposed to optimize the initial clustering center to improve the clustering effect of the algorithm. This paper starts with the initial data set and gets a data set more conducive to clustering by filtering the initial data set. Combined with the advantages of ABC algorithm, the objective function of ABC algorithm is taken as the measurement function of initial clustering. By changing the initial clustering center, the effectiveness and accuracy of k-means clustering are improved.

Acknowledgments

I would like to thank all the teachers and classmates, especially my roommates, for their encouragement and support in helping me complete the research of the thesis.

References

[1] Navy Ding, Qingxian Feng. Manual swarm algorithm based on boltzmann selection strategy [J]. Computer Engineering and Applications, 2009(31):53-55(in Chinese).

[2] Zhang Liangjun, Yang Tan, Xiao Gang, et al. MATLAB data analysis and mining actual combat. Beijing: Machinery Industry Press, 2016(in Chinese).

[3] Uribe C, segura B, baggio H C, et al. Hierarchical cluster analysis for Parkinson's disease with early cortical atrophy [J]. Parkinson's disease and related diseases, 2008, 50:3-9(in Chinese).

[4] Pengfei Liu, Yong Wang, Li Haiyan. Using philosophical thought to explore the mystery of bees' endless life. Chinese Bee Industry, 2010, 61(3):14-16(in Chinese).

[5] Xiong Zhongyang, Chen Ruotian, Zhang Yufang et al. an effective K-means clustering center initialization method. Computer Applied Research, 2011, 28(11):4188-4190(in Chinese).

[6] Lu Shaojuan, Zhang Guizhu. a clustering method of fusion K-means algorithm and artificial fish swarm algorithm. Computer Applications and Software, 2015, 32(9):240-243(in Chinese).

[7] Jia Ruiyu, Li Zhen. hierarchical k-means clustering algorithm based on the minimum spanning tree. Microelectronics and Computers, 2016, 33(3):86-88(in Chinese).

[8] Cesar L K, Kvalheim O M, cewith N b. hierarchical clustering analysis to identify interfering
matter spectrum metabolomics in untargeted mass by technical replication [J]. Journal of analytical chemistry, 2008, 1021: 69-77 (in Chinese).

[9] Yu X, Jiang J, Zhao B, et al. A self-adaptive artificial beecolony algorithm based on global best for global optimization [J]. Soft Computing, 2018, 22(9): 2935-2951 (in Chinese).

[10] Nan Jieqiong, Wang Xiaodong. particle swarm optimization algorithm for improving inertial weights [J]. Journal of Xi'an University of Engineering, 2017, 31(6): 835, 840 (in Chinese).

[11] Huang Weikai, Ma Xing. analysis of tactical missile test data based on improved k mean clustering. Air Defense, 2019, 10(4): 2094-4641 (in Chinese).

[12] Cao Yongchun, Cai Zhengqi, Shao Yabin. improved artificial colony clustering algorithm based on K-means. Computer Applications, 2014, 17(01): 204-207 (in Chinese).