2015 Fifth International Conference on Communication Systems and Network Technologies

A Clustering K-means Algorithm Based on Improved PSO Algorithm

Long Tan
Computer science and technology in heilongjiang
Heilongjiang Heilongjiang China
e-mail: tanlong01@163.com

Abstract—Because of the shortcomings of the traditional K-means algorithm which is sensitive to select the initial clustering centers and easy converges to local optimization, the paper proposes An Clustering K-means Algorithm Based on Improved the Particle Swarm Optimization Algorithm. The algorithm uses a powerful global search capability of the Particle Swarm Optimization algorithm to optimize the selection of the initial clustering centers: dynamically adjusting the inertia weight and other parameters to enhance the performance of the Particle Swarm Optimization; taking advantage of the fitness variance of the group to decide the conversion timing between the front part of the Particle Swarm Optimization algorithm and the rear part of K-means algorithm; setting the variables to monitor the changes of the optimal values of each particle and particle population, timely take the premature convergence particle to the mutation operation, thus we can find the global optimum initial clustering centers for K-means algorithm, then the clustering results are not affected by the initial clustering centers, it is easy to get global optimal solution. The experimental results show that the clustering accuracy rate, clustering quality and the global search capabilities of the improved algorithm is higher than the traditional clustering algorithm proposed in this paper.

Keywords-component; initial clustering centers, K-means Particle Swarm Optimization algorithm, Global Search

I. INTRODUCTION

The cluster analysis is an unsupervised classification techniques, according to a certain similarity criterion to classify the data set, so that objects of the class are the same as possible, but it is the diversity as possible between dissimilar objects, the K-means algorithm algorithm is a classical clustering algorithm based on partition, K-means algorithm has many advantages such as easy to understand, simple, fast convergence and so on, but it has the deficiencies: sensitive to select the initial clustering centers and easy converges to local optimization[1].

Selection of the initial clustering centers determines the division of the result of the K-means algorithm and the main processes of the data of the K-means algorithm, different initial clustering centers may get different results. The selection of clustering centers chosen may cause the algorithm into local optima. The introduction of particle swarm optimization algorithm by many researchers can solve this problem at the time of clustering. The literature [4] proposed the adaptive nonlinear and inertia weight algorithm base on the Particle Swarm Optimization algorithm and K-means algorithm, The literature [5] proposed the hybrid clustering Algorithm base on Particle Swarm Optimization algorithm and K-means algorithm, classical particle swarm optimization algorithm improve the initial clustering centers of the K-means algorithm, and improve the accuracy of clustering results. In the literature [7] integration of K-means clustering algorithm and the new clustering of PSO algorithm combines the advantages of K-means and PSO, then it mprove the quality of clustering algorithms. The above algorithms enhance effect of the clustering algorithms in some extent, they can alleviate K-means algorithm dependent on initial value, but Particle Swarm Optimization may appear premature convergence phenomenon in the classical particle swarm algorithm, then algorithm is still likely to fall into local minima. The literature [7] study the clustering algorithm of the K-means based on improved particle swarm optimization, and it can treat particles of local extremum to jump out local optima. Although the algorithm inherits the global search ability of the Particle Swarm Optimization, but they can not fully and effectively use the local search capabilities of K-means algorithm. Based on the analysis of the above algorithm. In this paper we presents an clustering optimization of K-means algorithm base on the improved Particle Swarm Optimization algorithm. We can get the global optimum k clustering by the evolution of the iterative search base on the Particle Swarm Optimization algorithm, then the initial clustering centers perform the K-means algorithm and achieve the desired clustering division, thus it can make full use of global optimization ability of the Particle Swarm Optimization algorithm and the local search capability of the K-means algorithm, and dynamically adjust the coefficients of the inertia weight of the particle and flight time of the particle, and real-time monitor the status of each particle and particle swarm, timely take the premature convergence particle to the mutation operation, then we can avoid the phenomenon of the premature convergence in the Particle Swarm Optimization algorithm, and eliminate the dependence of the selection on the initial cluster centers about K-means algorithm, then the clustering results are not affected by the initial clustering centers, it is easy to get global optimal solution[2].

II. THE INTRODUCTION OF RELATED ALGORITHMS

This section describes the K-means algorithm and Particle Swarm Optimization algorithm

A. K-means algorithm

K-means algorithm based on the similarity of the criteria divide the data set into k categories, the K-means algorithm is described as follows:

978-1-4799-1797-6/15 $31.00 © 2015 IEEE
DOI 10.1109/CSNT.2015.223

1) Randomly selected k initial clustering centers from the data set;

2) Each data object in the data set: calculate the distance from it to all the cluster center, and in accordance with the principles of the nearest neighbor we will divide it into the nearest class;

3) Recalculated each newly formed cluster centers clustering;

4) Repeat the procedure (2), (3), the algorithm ends until the clustering center do not occur the changes.

Wherein the measure of the similarity we will use the distance calculation method of Euclidean, cluster centers is the average of all data objects within classes[3].

B. Particle Swarm Optimization algorithm

Particle Swarm Optimization algorithm (Kennedy and Eberhart proposed) is a simulation of the process group behaviorof the birds foraging, it is a new type of swarm intelligence algorithm, and it is the the most widely method that used to solve NP-hard problem of clustering. Currently there are many studies apply the particle swarm optimization algorithm to the cluster analysis,for example, the literature [5] use the optimization algorithm of particle swarm to improve the data stream clustering algorithm based on sliding window model, then we can solve the problem of the traditional data stream clustering algorithm. PSO algorithm particles (personal best position) according to their own experience, and social experience of the shared groups (group optimum position) change the speed and position, then they can fly to the global optimum, and through a predefined fitness function on the current position we can evaluate the performance of the particles[4].

Supposing the population size of the particle swarm is m, the search space is n latitude, the optimal location of the individual particles is: $Pbest_i$, $Pbest_i = (pbest_{i1}, pbest_{i2}, \dots, pbest_{in})$, the optimal location of groups is : $Gbest_i$, $Gbest = (gbest_1, gbest_2, \dots, gbest_n)$, the speed of each particle is V_i, $V_i = (v_{i1}, v_{i2}, \dots, v_{in})$, and the position is $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$, i =1, 2, … ,m, thus we can adjust the algorithm according to the following formula:

$$V_i(t+1) = \omega V_i(t) + c_1 \times r_1 \times (Pbest_i - X_i) + c_2 \times r_2 \times (Gbest - X_i) \quad (1)$$

$$X_i(t+1) = X_i(t) + V_i() \quad (2)$$

In the formula t is a variable number of iterations; ω is called the inertia weight coefficient; constants c1, c2 is the learning factor; r1, r2 is the random number of evenly distributed within the range [0,1].

III. K-MEANS CLUSTERING ALGORITHM - BASED ON IMPROVED PARTICLE SWARM OPTIMIZATION

The clustering converges of K-means algorithm is quicker than the Particle Swarm Optimization algorithm, and K-means has the relatively strong ability of local search. However, the initial cluster centers affect the clustering results of the K-means, different initial centers may result in different clustering results, clustering results are highly volatile, if the inappropriate choice of the initial clustering ,the K-means algorithm is very easy to fall into local optima, and the Particle Swarm Optimization algorithm has strong capability of global search, but its convergence rate becomes slow in the latter, thus convergence rate of the Particle Swarm Optimization algorithm is not ideal, and there may be premature phenomenon in the particle swarm of the classical PSO algorithm. To solve these problems, the paper presents the K-means clustering algorithm based on improved Particle Swarm Optimization algorithm, it can dynamically adjust inertia weight coefficient of the particle and flight time of the particle, and enhance the global search capability of the Particle Swarm Optimization algorithm. Through group fitness variance,we can determine the timing of the two combined, then we can take advantage of outstanding ability of the global optimization of the Particle Swarm Optimization algorithm and fast search capabilities of K-means algorithm, and accelerate the convergence speed of the particle swarm, and real-time monitor the status changes of the particle swarm, timely take the premature convergence particle to the mutation operation, thus we can find the global optimum initial clustering centers for K-means algorithm, and avoid the particle swarm to fall into local extremum, thus we can use the particle swarm to search the optimal initial cluster centers of the global, the K-means algorithm suppress the sensitivity of the selection of the initial clustering center, then we can get a good result of clustering division.

A. Dynamically adjust inertia weight and flight time

In the paper, we can adjust parameters the Particle Swarm Optimization algorithm, this can help to enhance the global search ability of the algorithm. higher the inertia weight w, stronger global search capability of the algorithm, and then smaller the inertia weight w, the stronger local search ability of the algorithm the value, the ideal PSO algorithm has the strong ability of global search at the early stage and the capability of local search at the later stage, therefore, the w's value of the article in the Particle Swarm Optimization algorithm can use the following linear adjustment strategies:

$$w(t) = w_{max} - {(w_{max} - w_{min})t}/{t_{max}} \quad (3)$$

In the formula, w_{max} is the maximum of the inertia weight, its general value is 4; w_{min} is the minimum of the inertia weight, its general value is 0, t is the current iteration number; t_{max} is the maximum number of iterations of the particle swarm[5].

After the study found that the flight time of the location migration of bird flock in the actual situation are constantly changing every time. Therefore, the particles can be increased the flying time factor to avoid the oscillation around the optimal solution, then we can speed up the convergence time of the particle swarm. Particle's formula of position adjustment can be changed:

$$X_i(t+1) = X_i(t) + H_0(1 - {t}/{t_{max}})V_i(t+1) \quad (4)$$

The constant of the flight is H0, its general value is 1.5, t is the current iteration number, t_{max} is the maximum number of iterations of the particle swarm.

If the particle flight out of the search space, its position can be set the half of the difference between the upper

boundary and the lower boundary of the particle. The position of the boundary of the particle is the $[X_{max}, X_{min}]$, $[X_{max}, X_{min}]$ presents the the maximum value and the minimum value of the data of the each dimension in the the clustering data, the range of speeds is $[-p * X_{max}, p * X_{max}]$, if the velocity of the particle exceeds the boundary, it is set to the half of the maximum velocity[6].

B. encoding and fitness function of the particle swarm

Each particle of the particle swarm represents potentially viable solution in the problem space of the solution. The clustering algorithm should solve the problem that determine the cluster centers which meet the conditions. Therefore it can be mapped to the particle of the particle swarm, the position x_i of the particle can be expressed the vector Z_j ($1 \leqslant j \leqslant k$) composed by the k clustering centers. If clustering data set is q-dimensional vector, the position and velocity of the particles are q*k-dimensional vector. Therefore, the particle can adopt the following form of real-code:

$Z_{11}\, Z_{12} \ldots Z_{1q}\, Z_{21}\, Z_{22} \ldots Z_{2q} \ldots Z_{k1}\, Z_{k2} \ldots Z_{kq} \,||$

$V_{11}\, V_{12} \ldots V_{1q}\, V_{21}\, V_{22} \ldots V_{2q} \ldots V_{k1}\, V_{k2} \ldots V_{kq} \,||\, f(x)$

We can lead the Particle Swarm Optimization algorithm into K - means algorithm, the criterion function which evaluated the quality of the clustering can be used the fitness function of the particle swarm.

Fitness function is defined as:

$$f(x) = \sum_{j=1}^{k} \sum_{S_i \in C_j} ||S_i - Z_j|| \quad (5)$$

Fitness value of the particle represents the similarity of each data objects. The smaller fitness value, more closely the degree of integration of the data objects within the class, the better clustering. So the goal of the improve the Particle Swarm Optimization algorithm is to search to particle positions which make the fitness value become the smallest particles, then the corresponding cluster of the particle position center is the optimization value of the initial cluster centers.

C. Particle mutation

There may be the problems of premature convergence in an iterative process, this can conduct the Particle Swarm fall into local optimal solution. Therefore, part of the text in the Particle Swarm Optimization algorithm, then we can set two variables to real-time monitor the state of each particle and particle swarm. when it detects phenomenon of the premature convergence in the particles or particle swarm, thus we can timely make mutation operation, and increase the diversity of particles, and make it out of the local extreme bondage, thus we can start the new search in the multidimensional solution space.

Specific practices are:

1) We set the control variable N_{pi} ($1 \leqslant i \leqslant m$, m is the number of particles) to real-time monitor the changes of the optimal value of each particle (i). If the fitness value f (x_i) of the particle (i) is no better than the corresponding individual extreme f ($Pbest_i$) of the personal best position, this iteration of particle (i) can not improve their personal best value, the variable N_{pi} which present the cumulative number of the particles's state that can not be improved increase again. If N_{pi} exceeds a predetermined threshold value: $thre_p$, it indicates that the particle (i) appeared prematurity, we should deal particle (i) with mutation operation, then it can depart from local minima as soon as possible.

2) We set a control variable N_g to monitor real-time changes of optimal value of the particle swarm. If the individual extreme value of all the particles are not less than the global maximum in this iteration, the optimal value of the particle swarm can not be improved, the variable N_g which present the cumulative number of the particles's state that can not be improved increase again, If N_g exceeds a predetermined threshold value: $thre_p$, it indicates that the the particle swarm is likely to fall into the local optimal solution. Then we can sort the accommodate value of all particles, it select the smaller fitness value of its particle mutation, thus it can increase the diversity of the particle swarm, then it overcome premature convergence problem of the particle swarm.

Specific approach of the mutation is to perform a K-means on the variation of particles, the position, the velocity and the personal best position of the particles can be re-initialized.

IV. CONVERSION TIMING BETWEEN PSO ALGORITHM AND K-MEANS ALGORITHM

Because the main goal of the algorithm in front stage is to look for the initial value of the optimizing clustering center,then the paper only execute the Particle Swarm Optimization algorithm, thus we can quickly search the initial clustering centers of the K-means algorithm. The degree of the convergence of the particle swarm can be reflected by the fitness variance, the fitness variance of the group σ^2 can be defined as:

$$\sigma^2 = {}^1\!/_m \sum_{i=1}^{m} \left[f(x_i) - f_{avg}\right]^2 \quad (6)$$

m is the particle population size; f (x_i) is the fitness value of particle i; f_{avg} is the fitness mean of all particles. When the fitness value of the variance σ^2 is small, it indicates that volatility of the fitness value of the particle swarm is very small, and the state of the particle swarm tends to converge. When σ^2 is less than the predetermined threshold value $thre_\sigma$, then it will terminate the execution of the improved the Particle Swarm Optimization algorithm and instead began to perform the K-means algorithm, thus it can accelerate the convergence speed of the Particle Swarm Optimization algorithm in the later stage.

V. THE DESCRIBES OF THE ALGORITHM IN THIS PAPER

This paper presents an clustering K-means algorithm based on improved the Particle Swarm Optimization algorithm , it is described as follows:

Input: Wait for clustering data set is S, the number of clusters is k, the population size of theparticle swarm is m, the maximum number of iterations is t_{max}.

Output: the number of the clustering division is k.

Algorithm is as follows:

Step1: the particle swarm is initialized, and randomly selected the number k of central point from the data set S, then we can see this value as the initial value of the particle position X_i. At the same time, the velocity of the particle V_i is initialized, personal best optimal position is $Pbest_i$ and the corresponding individual extreme f ($Pbest_i$), optimal location of populations is the Gbest and their corresponding global extreme f(Gbest). This process cycle m times, it can complete the initialization structure of the particle swarm[7].

Step2: Use the PSO algorithm perform iterative search of the particle swarm.

Perform the following operations on each particle of the the particle swarm:

a) According to the formula (3) it can dynamically adjust the inertia weight, then the formula (1) and the formula (4) are used to update the speed and the position of the particl.

b) In accordance with the principle of the nearest neighbor, it can divide the datasets, and calculate the value of the fitness of the particle.

c) If the fitness of the particle is smaller than individual extreme, it can update individual extreme and personal best optimal locations of the particle ($Pbest_i$).

d) If N_{pi} is greater than the predetermined threshold value $thre_p$, then the mutation operation should perform on the particle i.

e) If σ^2 (fitness variance of the particle swarm) has judged the converge which the particle swarm will tended to or the number of the loop iterations that just achieve maximum t_{max},it will terminate the iteration of the particle swarm and use the Gbest as corresponding initial value of the Step3, otherwise, turn the (b) and continue execute the iteration.

Step3: Execute the K - means algorithm, and output the final results.

VI. EXPERIMENT AND ANALYSIS

In the experiment every value of parameters of the algorithm: the scale size of the particle swarm m = 20, maximum number of iterations of the particle swarm t_{max} = 100, $thre_p$ = 4, $thre_g$ = 5, p = 0. 4, the fitness variance threshold of the particle swarm threσ = 0.1. The test result takes the average. It contrasts the accuracy of the clustering and fitness value of the algorithm, the Table 1 and Table 2 show the result of the comparison.

TABLE I. COMPARISON OF THE ACCURACY OF THE K - MEANS, PSO - KM AND PROPOSED ALGORITHM IN THE PAPER

actual data collection	accuracy	PSO-km	K- means	proposed algorithm
	Highest	86.25	85.14	86.95
BreastCancer	Lowest	85.14	85.14	85.60
	Average	86.10	85.14	86.57
	Highest	88.91	88.82	89.30
Iris	Lowest	88.56	87.90	88.42
	Average	88.69	87.99	90.13
	Highest	78.56	76.10	76.95
Wine	Lowest	76.95	78.56	76.06
	Average	77.90	77.84	76.10

TABLE II. COMPARISON OF THE FITNESS OF THE K - MEANS, PSO - KM AND PROPOSED ALGORITHM IN THE PAPER

actual data collection	fitness	PSO-km	K- means	proposed algorithm
	Highest	116.25	92.72	126.95
BreastCancer	Lowest	95.14	98.38	125.60
	Average	106.10	93.15	126.57
	Highest	48.92	48.82	59.30
Iris	Lowest	48.54	47.91	58.42
	Average	48.65	47.92	50.13
	Highest	138.56	136.15	126.35
Wine	Lowest	136.97	138.58	126.51
	Average	137.98	137.84	126.23

In order to verify the performance of the convergence of the algorithm we test the Iris data set, then we can get three kinds of the convergence graph of the algorithm (Figure 1). As it shown on the Figure 1, the converges of the K-means algorithm is fastest. Because of premature of the particle swarm of pso-km algorithm, its convergence is slower than the K-means algorithm.

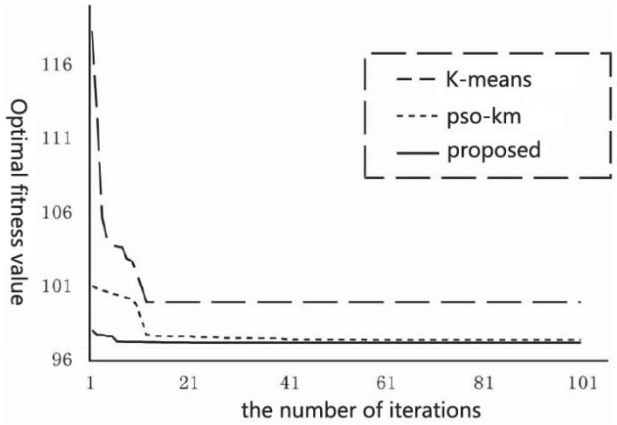

Figure 1. Test the Iris data set

VII. CONCLUSION

In this paper, because of the shortcomings of the traditional K-means algorithm which is sensitive to select the initial clustering centers, the paper presents an clustering K-means algorithm based on improved the PSO Algorithm, the algorithm dynamically adjusts the inertia weight coefficient of the particle and the flight time to enhance the global search ability of the particle swarm. The paper set the variables to monitor the changes of the optimal values of each particle and particle population, timely take the premature convergence particle to the mutation operation, thus we can find the global optimum initial clustering centers for K-means algorithm, then the clustering results are not affected by the initial clustering centers, it is easy to get global optimal solution. The experimental results show that the clustering accuracy rate, clustering quality and the global

search capabilities of the improved algorithm is higher than the traditional clustering algorithm proposed in this paper. However, compared with the traditional K-means algorithm, the computation time of the proposed algorithm has increased, this problem is the further research of the author needs.

ACKNOWLEDGEMENT

The research of author is partially supported by National Natural Science Foundation of China under Grant No.81273649 , and Provincial Natural Science Foundation of Heilongjiang , China under Grant No. F201434.

REFERENCES

[1] Han Jiawei, Kamber M.Data mining: Concepts and techniques.2nd ed.Beijing: China Machine Press,2006

[2] Kennedy J, Eberhart R .Particle swarm optimizationProceedings of IEEE international conference on neural net-works.Perth, 1995:

[3] Kim T S,May G S.Time series modeling of photosensitive pol-ymer development rate for via formation applications.IEEE transactions on electronics packaging manufacturing,2002

[4] Liuyue Ting, Li Lan. The hybrid of particle swarm and K-means clustering algorithm based on adaptive weights 2012

[5] Lv Yi Qing, Lin Jinxian. MPI-based parallel hybrid PSO K-means clustering algorithm 2011

[6] Fu Tao, Sun Ya Min. Based on PSO K - means algorithm and its application in network intrusion detection 2011

[7] Berens J,Finlayson G D.Image indexing using compressed color histogram,2000