Journal of Theoretical and Applied Information Technology
20th February 2013. Vol. 48 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE K-MEANS CLUSTERING ALGORITHM BASED ON CHAOS PARTICLE SWARM

[1] LI YI RAN, [2]ZHU YONG YONG, [3] ZHANG CHUN NA

[1.3] University of Science and Technology Liaoning, Anshan 114054, China

[2] Department of Economics and Business Administration, Chongqing University of Education, Chongqing 400067, China

E-mail: [1]lyr7879@yahoo.com.cn, [2]zhuyongyong@126.com, [3]zcn1979@yahoo.com.cn

ABSTRACT

Proposed the Algorithm of K-means (CPSOKM) based on Chaos Particle Swarm in order to solve the problem that K-means algorithm sensitive to initial conditions and is easy to influence the clustering effect. On the selection of the initial value problem, algorithm using particle swarm algorithm to balance the random value uncertainty, and then by introducing a chaotic sequence, the particles move speed and position in a redefined, thus solving the initial value sensitivity, while the algorithm with overall search capability, but also to avoid the local optimum. The algorithm add acceleration factor and escape factor in order to improve the time efficiency. Experiment result proved that the CPSOKM algorithm has a fast convergence speed, high stability, and good clustering effect.

Keywords: *K-means algorithm; Particle swarm; Chaotic sequence; Cluster analysis; CPSOKM algorithm*

1. INTRODUCTION

Clustering in data mining and knowledge discovery is an important analysis method, has been widely used in various fields of scientific research, in which the K-means algorithm is a classical algorithm in clustering analysis. The algorithm has the advantage of being able to handle large data sets, its shortcoming is that the initial clustering center and the selection of clustering, if it is not properly, the cluster effect will be difficult to ensure [1].

Data clustering divided the content similarity in grammatical or semantic in an unsupervised learning environment into one class, without manual marking pretreatment, the solution of the problem belongs to optimization category, the expected extremum of target function can be got in iterative way. In order to ensure the stability of the experimental results and obtain the global optimal solution, the search process cannot be excessively dependent on the initial value, should adopt the random strategy. The conventional algorithm of the above problems should include neural networks, ant colony algorithm, leapfrog algorithm, particle swarm algorithm etc.[2].

In recent years, related studies based on particle swarm hybrid algorithm is really more. Reference[3] introduced rough set theory based on the combination of particle swarm optimization algorithm and K-means algorithm aiming at image classification, and improved image recognition of the algorithm. Reference[4] is also a mixed strategy, it proposed quantum particle swarm algorithm which aiming at gene clustering, K-means algorithm uses a single strategy, not only improve the clustering speed but also the effect is good. Reference[5] is based on the improvement of simulated annealing particle swarm optimization algorithm, which adjusts the constant acceleration to particles in the global and local position. The experimental results show that the cluster effect was improved greatly. Reference[6] is based on K-mean hybrid particle swarm algorithm, the algorithm can effectively avoid depend on initial value, the value of the mutation process is better, the anti-interference ability is strong, the global search ability needs to be improved. Reference[7] combined C-means and quantum particle swarm algorithm with fewer parameters, and the global search capability is strong.

The above research focus on the initial value sensitivity, the ability of global search, local escape ability, convergence rate, periodicity, the point of tangency should not include, therefore this paper also need special consideration.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Particle swarm optimization algorithm (Particle Swarm Optimization), referred to as PSO, is a newly developed evolutionary algorithm, and its thought originated from birds, fish and other groups foraging behavior. Algorithms assume that the space of each particle can be used as a solution, optimal solution generated in the iterative process. Position changes of the particle determines by its own speed each iteration, the particle swarm is keeping the current optimal particle, and do its search, eventually find the individual and global extremum. The algorithm is simple and clear, without too much manual intervention, convergence speed is rapid[8,9]. Based on this, this article introduced K-means in clustering particle swarm optimization algorithm to implement data clustering.

2. CHAOS PARTICLE SWARM OPTIMIZATION ALGORITHM

2.1 Chaotic Sequences

Spatial scattered points belongs to the category of nonlinear, so chaos phenomena is, it has two major characteristics: ergodicity and regularity, which can have a regular traversal to all the given region of the state, and do not repeat. Particle search by adopting the chaotic strategies are clearly better than the simple random search[10]. Search procedure is as follows:

(1)Define initial area, and set N dimensional initial vector $R_0 = (R_{01}, R_{02}, \cdots, R_{0N})$, the various values in R_0 are adjacent, and the difference is very small.

(2)Use the logistics equation to calculate the initial vector R_0 , and generate chaotic sequence $m_1, m_2, \cdots, m_n$. after several iterations, the system will complete in a chaotic state. Vector layer can be expressed as:

$$m_{i+1} = m_i(1 - m_{i-1})\lambda \quad (1)$$

In the formula, λ is iterative control parameter.

(3)Suppose space particle X_i , and use the forum (1) to get better position of X_i , such as $X_i^{'}$.

$$X_i^{'} = r \cdot rnd \cdot m_j + X_i \quad (2)$$

In the formula, r is the activity radius of particle X_i , $rnd \in [-1,1]$, $j \in [0,n]$.

2.2 Algorithm Description

Particles in algorithm of PSO have some randomness in search of individual and global extreme sports, i.e. when the particles in the search process, if it arrive local optimal solution, other particles followed and fell into the region; and the particle's motion may not be able to keep it away from the local optimum bound, so it is easy to fall into local optimal. The particle motion has a regular pattern because of the introduction of Chaos , and it has traversal properties. At the same time, chaotic mechanism can also be reaction in the PSO to allow the particle quickly escape from local optimal, achieve better extremum search results, algorithm process is as follows:

(1)Particle swarm parameter initialization: Including particle swarm space, namely population size; particle initial velocity and position; search number of iterations, the length of the chaotic sequence and other necessary parameters.

(2) Fitness value calculation: Calculating the fitness value of the current particle, if the individual current value better than the previous value, and updating the particle positions; orderly statistics the current value of all particle, if the optimal solutions of the individual extremum better than the previous global optimal solution, then replaced.

(3)Chaos optimization: set the current optimal solution as $P_b = (P_{b1}, P_{b2}, \cdots, P_{bk})$, sign as P_{bi} , $i \in [1,k]$, mapped to forum (1), get $m_i = (P_{bi} - a_i)/(b_i - a_i)$, then use the logistics equation to iterate chaotic variable sequence, and inverse map to the solution space to obtain extremal solutions $P_b^{'} = (P_{b1}^{'}, P_{b2}^{'}, \cdots, P_{bk}^{'})$. Finally, traverse the solution space of each solution, calculate its fitness value, and obtain alternative solutions $P_b^{''}$.

(4)Search each particle in groups, if the optimized solution is better than that of the current solution, then replace current position with $P_b^{''}$.

(5)To view the current status whether satisfy a search condition, if it is satisfied, then the current solution as the optimal solution; Otherwise, return to calculate right value.

3. PARTICLE SWARM OPTIMIZATION IN K-MEANS ALGORITHM

3.1 K-Means Algorithm

K-means algorithm is a kind of local objective function algorithm based on the distance[11], by setting the initial number of clusters, and randomly

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

selected cluster center, after each iteration, the data points of cluster according to the distance from the clustering center is divided into a new cluster, at the same time will produce a new clustering center, algorithm assumes that the closer the data the greater the points of similarity.

Set finite set of Q dimensional space S^Q as $X=\{x_1,x_2,\cdots,x_n\}$, the initialization can be divided into k class randomly, denoted as $C_1,C_2,\cdots,C_k$, if a class has n objects, the i clustering center can be defined as $Z_1,Z_2,\cdots,Z_k$, $Z_i=\frac{1}{n}\sum_{j=1}^{n}x_j, j\in[1,k]$, the definition of the objective function as follows:

$$J=\sum_{i=1}^{k}\sum_{j=1}^{n_j}D^2_{x_j,z_i} \quad (3)$$

In the formula, $D^2_{x_j,z_i}$ represents the distance that the j text to a class i clustering center, i.e. Euclidean distance.

The core idea of algorithm is through continuous iteration to find the k optimal cluster centers of the sample data set, other data mobile to the cluster center, until the objective function value is minimum. Algorithm steps are as follows:

(1)Initialize the training sample set, randomly generated k cluster centers, credited as $Z^1=\{Z_1,Z_2,\cdots,Z_k\}$;

(2)Other remaining data points in clusters are divided into the class belongs according to the corresponding value of the objective function;

(3)Reference $Z_i=\frac{1}{n}\sum_{j=1}^{n}x_j, j\in[1,k]$ and recalculate cluster centre Z^k;

(4)Set the iteration limit condition, if it is $J<\delta$, then the process is terminated; Otherwise, continue to iterate, and return (2), until a termination condition is satisfied.

Optimizated algorithm pseudo code is as follows:

```
Public bool fn_CKM(int i_KC, double d_Limit,
double d_CZ[],int i_Count, ObjText obj_XText)
{
ObjNewC obj_Kc=new
ObjNewC(fn_InitZ(i_KC,d_CZ));
// initialize the cluster center, and redraw the intra
cluster data point
fn_CluZ(obj_Kc, obj_XText);
//recalculate cluster centre
return (d_CZ[i_Count+1]- d_CZ[i_Count])<
d_Limit;
// judgment on the algorithm terminates condition
}
```

3.2 Particle Velocity Optimization

The effect of clustering is mainly to inspect the satisfaction degree of convergence, each particle of groups has a certain velocity, and moments in the adjustment, the track from their flying experience, and affected by the other particle. Therefore, algorithm analysis should not be the individual as isolated points, but these points should be treated as contacted with each other and rely on each other.

Definition 1: the acceleration factor.

Usually, along with the population movement, if speed is reduced too fast, groups are fall into local optimum easily. Experimental measured that this phenomenon is particularly obvious in FCM, and it is often convergent fast when the algorithm is end. The key reason is that the speed will gradually decline, and ultimately will be reduced to 0 in the iterative process; when the speed drops a certain external value, search ability will decline, resulting is not jump out of the local optimum. Set threshold ε, if the speed is less than this value, then adjust the current speed, namely $v_i<\varepsilon$, then $v_i=\lambda v_0$. v_0 is the initial velocity, i.e. the maximum velocity of particles; λ is the acceleration factor and $0<\lambda<1$.

Definition 2: escape factor.

The species propagate have a features called clustering, that is to say, when a population gathered a large number of individuals, if the space is relatively narrow, part of the group will be separated out, and then construct the new community. The position of the particle is also based on this consideration, turn to acceleration factor, if speed is less than the threshold value ε, this paper set up an escape factor γ, the selection of a location as the current position to replace the best position of particles, and to the current rate of iteration, $v_i<\varepsilon$, then $p_g=p_i$. Here, the selection of alternative position need to be put forward specially should adopt the random strategy, which can guarantee the stability of the original groups, yet the diversity of population.

Now exists Q dimensional space, set up the i particle, can be expressed as $q_i=(q_{1Q},q_{2Q},\cdots,q_{iQ})$, its fitness value can be expressed as

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$p_i = (p_{1Q}, p_{2Q}, \cdots, p_{iQ})$, with a velocity $v_i = (v_{1Q}, v_{2Q}, \cdots, v_{iQ})$, for each iteration of the k ($1 \le k \le Q$) dimensional equation as shown below:

$$v_{ik}^n = \lambda v_{ik}^{n-1} + \alpha_1 \beta_1^{n-1}(p_{ik}^{n-1} - q_{ik}^{n-1}) + \alpha_2 \beta_2^{n-1}(p_{gk}^{n-1} - q_{ik}^{n-1}) \quad (4)$$

$$q_{ik}^n = q_{ik}^{n-1} + v_{ik}^n \quad (5)$$

In the formula, λ is acceleration factor, its value should be adaptive, not large or small, and be adjusted timely in the iterative process. $\alpha_1, \alpha_2; \beta_1, \beta_2$ are correction factors, α_1, α_2 are set the same value, $\in [0,3]$; β_1, β_2 are random numbers, $\in (0,1)$.

Through the correction of velocity, freedom degree of particles in the Q dimension space search process is greater, and not easy to fall into local optimum, moderate convergence rate, number of iterations is defined by conditions.

4. K-MEANS CLUSTERING ALGORITHM WITH CHAOS

4.1 Algorithm Description

Data information has one important characteristic - category diversity. The given page content is belong to multiple categories, the clustering process is divided into different groups based on the attribution of the class, similar individual distance in group is relatively near, but not the same for individual. The habitual use of data clustering are not good at study of relations between classes, the K-means algorithm is a data point distance to describe the groups category clustering methods, which assume that the initial clusters has been divided into k finite set, using the iterative redirection will cluster object in accordance with certain rules so as to change the shape of mobile, collection. Here needs to pay attention to two issues: random initialization strategy and the optimization strategy of algorithm, due to inherent characteristics, the former can lead to unstable clustering effect, the latter result into local optimum. Therefore, the text put forward in combination of K-means algorithm and particle swarm algorithm, not only improves the convergent speed of the algorithm, the clustering effect is ensured.

Hybrid algorithm also needs to consider two questions: first, how to change the text vector into particle swarm individual; second, what establish kind of fitness model to describe particle swarm clustering.

The algorithm map the text vector to particle swarm, and reflect as cluster coding format, a particle corresponds to possible solutions of a text set in groups, and its existence form is discrete. Turn to sample space 2.1, C common characters of text are abstracted, so any text can be represented as $x_i = \{x_{i1}, x_{i2}, \cdots, x_{iC}\}$; it has k cluster centers, defined as Z_i , construct the position of particle; the length of particle can be defined as $Q \times k$, velocity can be represented as $v_i = \{v_{i1}, v_{i2}, \cdots, v_{iC}\}$. Fitness model as shown below:

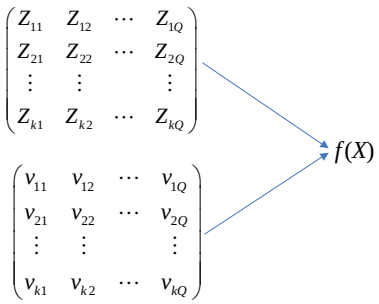


Figure 1: Cluster Coding Format

Based on the previous knowledge, the ultimate goal of K-means algorithm is through continuous adjustment of the data points, and reset the clustering center in order to obtain the stable objective function. Extreme value determined from forum (1) whether it obtain the optimal solution.

4.2 Algorithm Step

The optimized algorithm based on Chaos Particle Swarm is shown as follows:

Input: Characters vector extracted from data;

Output: The data has been divided the class;

Step 1: Finish feature vector into the training set;

Step 2: Initialize particle swarm size N , Set the number of iterations, start $m = 0$, The maximum number of iterations is $m_{\max}$, set the position ($x_i^k \mid x_i^1, x_i^2, \cdots x_i^N$) and velocity （ $v_i^k \mid v_i^1, v_i^2, \cdots v_i^N$ ） of the particle group, particle radius r of action in Chaotic sequence, correction factors $\alpha_1, \alpha_2; \beta_1, \beta_2$, set k cluster center Z^1 randomly;

Step 3: Calculate the initial fitness, and select the best position of each particle (the best fitness value), move particles and the best position as the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

iterative initial position to each particle, i.e. $P_{bi}^{m} = X_{bi}^{m}$. At the same time, the fitness function change accordingly, namely $f(P_{bi}^{m}) = f(X_{bi}^{m})$;

Step 4: The chaotic sequence is introduced in the current iteration depending on the forum (1), (2), calculate the position of the particle, and the current position is compared, if it is excellent, then replaced;

Step 5: In the process of position replacement, considering the acceleration factor and escape factor under the premise, speed should be updated according to forum (4), (5);

Step 6: Compare global fitness constantly in the iteration of particle swarm, and select historical records of the optimal global fitness value, use this position to replace the global optimum of the current particle location;

Step 7: Termination conditions:

① Whether the objective function touches the limit value δ;

② Whether the iteration number reaches the maximum iteration limit number $m_{\max}$;

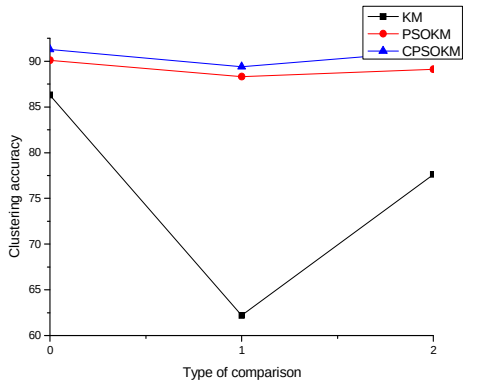
If meet either of these conditions, it can obtain the global optimal solution P_{gb}^{m}; otherwise, increase iterative step length factor m.

5. EXPERIMENT

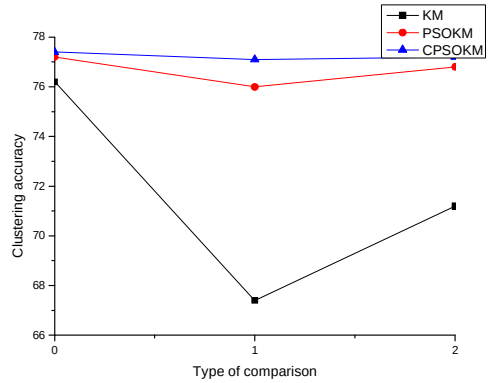
5.1 Data Source

Experimental data are taken from three data sets in the universal database UCI: Balance, Similar and Simple. Among them, balance is a low dimensional data set, the dimension is 4, 585; Similar is a high dimensional data set, the dimension is 16090, sample number is 280; Similar is a low dimensional data set, the dimension is 4, sample number is 302;

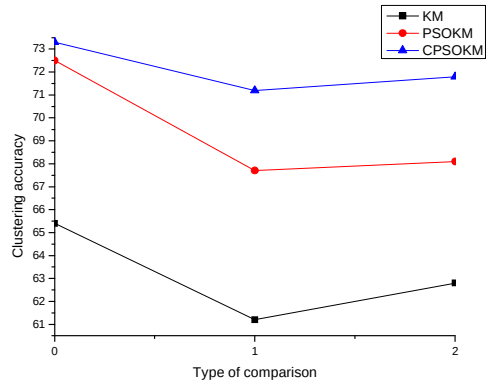
The complexity of above data set are not identical, subject word is more in the individual data, involving the three algorithms: the traditional K-means algorithm, denoted as KM; based on particle swarm algorithm K-means, mark PSOKM, as well as the text presented with chaotic particle swarm algorithm, denoted by CPSOKM algorithm, experimental comparison algorithm clustering accuracy in three aspects: the maximum value, minimum value and average value, a set of relevant data comparison is shown in figure 2.



(a) Balance



(b) Similar



(c) Simple

Figure 2: The Comparison Of Clustering Accuracy

5.2 Result Analysis

Figure 2 shows K-means algorithm clustering accuracy of minimum, maximum value and mean value of maximum difference in three kinds of algorithm, the reason is that the algorithm is sensitive to the initial value, leading to clustering result value jumps, not stable enough. PSOKM, CPSOKM are more stable than K-means, but the two had little difference; this is due to the particle

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

swarm algorithm combined with the K-means algorithm, K-means escape from local optimal limit, stability are improved in some degree, clustering effect is improved obviously. While the chaos mechanism is introduced into the CPSOKM algorithm, so the clustering results are more close to real value, stability is the strongest.

6. CONCLUSION

This paper combines PSO algorithm and K-means algorithm to search optimally, which not only solves the problem of K-means algorithm is sensitive to initial value and easily falls into local solutions, but also optimizes the clustering results. While introduced chaos sequence to improve the global search ability, so that the particles moving ability can be enhanced, and avoid randomness. Introduced the acceleration factor and escape factor, due to the decreased particle velocity, and readjusted the speed, so that the particles are more easily escape from local to manacle, and accelerated the convergence rate. The experiment result proved that in the comparison among KM, PSOKM and CPSOKM, the improved optimized particle swarm algorithm CPSOKM that this paper proposed whether it is the global search ability and stability are more superior to other algorithms.

REFRENCES:

[1] Dong-Xia Chang, Xian-Da Zhang, and Chang-Wen Zheng,"Agenetic algorithm with gene rearrangement for K-means clustering ", *Pattern Recognition* ,42(7):1210-1222.

[2] Sung-Kwun Oh, Wook-Dong Kim, Witold Pedrycz, and Su-Chong Joo, "Design of K-means clustering-based polynomial radial basis function neural networks (pRBF NNs) realized with the aid of particle swarm optimization and differential evolution " , *Neurocomputing*, 78(1):121-132.

[3] Chih-Cheng Hung and Hendri Purnawan, " A Hybrid Rough K-Means Algorithm and Particle Swarm Optimization for Image Classification" , *MICAI '08 Proceedings of the 7th Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence*, 2008,585-593.

[4] Jun Sun, Wei Chen, Wei Fang, Xiaojun Wun, and Wenbo Xu, " Gene expression data analysis with the clustering method based on an improved quantum-behaved Particle Swarm Optimization", *Engineering Applications of Artificial Intelligence*, 2012,25(2):376-391.

[5] Dae Sung Lee, Young Wook Seo, and Kun Chang Lee, "An adjusted simulated annealing approach to particle swarm optimization: empirical performance in decision making", *ACIIDS'11 Proceedings of the Third international conference on Intelligent information and database systems*, 2011,566-575.

[6] Cui X and Potok T E, "Document clustering analysis based on hybrid PSO+K-means algorithm", *Journal of Computer Sciences*, 2005(Special Issue):27-33.

[7] Hao Wang, Danyun Li, and Yayun Chu, "A New Scalability of Hybrid Fuzzy C-Means Algorithm", *Artificial Intelligence and Computational Intelligence (AICI)*, 2010(3):55-58.

[8] I.L.Schoeman and A.P.Engelbrecht, "A novel particle swarm niching technique based on extensive vector operations", *Natural Computing*, 2010,9(3):683-701.

[9] M. G. Epitropakis, V. P. Plagianakos, and M. N. Vrahatis, "Evolving cognitive and social experience in Particle Swarm Optimization through Differential Evolution: A hybrid approach", *Information Sciences*, 2012, 216:50-92.

[10] Leandro dos Santos Coelho and Viviana Cocco Mariani, "Use of chaotic sequences in a biologically inspired algorithm for engineering design optimization", *Expert Systems with Applications*,2008,34(3):1905-1913.

[11] Rupali Vij and Suresh Kumar, "Improved k-means clustering algorithm for two dimensional data", *CCSEIT '12 Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*, 2012, 665-670.