2008 International Conference on Intelligent Computation Technology and Automation

The Clustering Algorithm Based on Particle Swarm Optimization Algorithm

Pei Zhenkui[1,2] , Hua Xia[1] , and Han Jinfeng[1]
[1]*College of Computer and Communication Engineering, China University of Petroleum,*
Dongying 257061, China
[2]*School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044*
peizhk@hdpu.edu.cn

Abstract

After analyzing the disadvantages of the classical K-means clustering algorithm, this paper combines the core idea of K-means clustering method with PSO algorithm and proposes a new clustering method which is called clustering algorithm based on Particle Swarm Optimization algorithm. It used the global optimization of PSO algorithm to make up the shortage of the clustering method. The algorithm is evaluated on Iris plants database, Results show that the algorithm is more effective and promising.

1. Introduction

Clustering analysis is an important part of the Data Mining research; it is an important method of the unsupervised learning. It divides the data into certain polymerization classes according to the attribute of the data ,enable the element of every class to have the same characteristic as far as possible, the characteristic difference among the different polymerization class is as far as possible big[1].

The traditional clustering algorithm seeks the optimal solution of the researched question through the iterative Hill-climbing method, because it is a local search algorithm, it has some shortages. We take the K-means clustering algorithm[2] as the example; people have found two inherent drawbacks in the practical application, (1)the random selection of Starting value may lead to different clustering results, even has no solution ; (2) The algorithm is an algorithm based on the objective function; it usually solves the extreme value problem by the gradient method. Because the gradient method searches along the energy decrement direction, the algorithm falls into the local extreme value easily. These flaws have limited its application scope greatly.

PSO is an effectively global optimization algorithm, it guides optimization search by the Swarm Intelligence, which comes from cooperation and competition between particles, Compares with the evolutionary algorithm, PSO retains the global search strategy based on population, its operation is simple, and solution of each generation population has Dual advantages of Self-learning and learning from others. So it can find the optimal solution by lesser iterative times. This paper combines the core idea of k-means clustering method with PSO algorithm and proposes a clustering algorithm based on Particle swarm optimization algorithm [6,7].

2.Particle swarm optimization algorithm

As an optimization algorithm, Initially, Particle swarm optimization algorithm is used for optimization of the Continuous space, in the continuous space coordinate system; Mathematical description of PSO[3,4,5] is as follows:

We suppose population size is *N*, Each particle is treated as a point in a *D* dimensional space. The ith particle is represented as $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots, x_{id} \cdots, x_{iD})$, $\boldsymbol{x_i}$ is a latent solution of the optimized question. The rate of the particle *i* is represented as $\boldsymbol{v_i} = (v_{i1}, v_{i2}, \cdots, v_{id} \cdots, v_{iD})$, it is a position change quantity of particle in an iteration. The particles are manipulated according to the following equation:

$$v_{id} = \omega v_{id} + c_1 rand_1()(p_{id} - x_{id}) + \ c_2 rand_2()(p_{gd} - x_{id}) \qquad (a)$$

$$\begin{cases} v_{id} = v_{\max} & if \quad v_{id} > v_{\max} \\ v_{id} = -v_{\max} & if \quad v_{id} < -v_{\max} \end{cases} \qquad (b)$$

$$x_{id} = x_{id} + v_{id} \qquad (c)$$

In the equation (a), the historical best position of all the particles in the population is represented by p_{gd}, the historical best position of the current particle is represented by p_{id}, the particle's new velocity is calculated according to its previous velocity and the distances of its current position from its own historical

978-0-7695-3357-5/08 $25.00 © 2008 IEEE
DOI 10.1109/ICICTA.2008.421

best position and the group's historical best position. Variable ω is the Inertia weight, c_1 and c_2 are positive constants, $rand_1()$ and $rand_2()$are two random functions in the range [0,1]**.** In the equation (b), particles' velocities in each dimension are limited to a maximum velocity v_{max}, v_{max} decided the search precision of particles in solution space. If it is too big, the particles possibly fly the optimal solution, if it is too small; the particles easily fall into the local search space and have no method to carry on the global search. In the equation (c), the particle's new position is calculated according to its current position and the new velocity, finally, the performance of each particle is measured according to a predefined fitness function, then finding the optimal solution of the research problem.

3. The K-means clustering algorithm

In the R^n space, the clustering problem may describe as follows: a given point set including N points x_1， $x_2,\ldots,x_N$,we divide these points into K (known constant) sets G_1，$G_2,\ldots,G_K$ according to the similarity of them. They satisfy the following conditions:

1）$G_i \neq \varnothing, \quad i = 1,2,\ldots K$ ；

2）$G_i \cap G_j = \varnothing,$

$i, j = 1,2,\ldots,K$ ； $i \neq j$ ；

3）$\bigcup_{i=1}^{K} G_i = \{x_1, x_2, \ldots, x_N\}$

The basic K-means algorithm consists of the following steps:

(1) assigns the clustering number K.

(2) Random select K points C_1，C_2, …, C_K as the initial clustering centers from the given point set {x_1，x_2, …, x_N}

(3) Select C_1，C_2, …,C_K as the clustering centers and divide the set {x_1，x_2,…, x_N} according to the following regulation:

If $d\left(x_i, c_p\right) < d\left(x_i, c_q\right) \ \ p,q = 1,2,\ldots,K$ and $p \neq q$, then $x_i \in G_p$ (G_p is a class and its center is C_p)

(4) Recalculate the new clustering centers C_1^1, C_2^1, …, C_K^1 according to the equation

$$C_i^1 = \frac{1}{|G_i|}\sum_{x_j \in G_i} x_j ,$$

$i = 1,2,\ldots,K$,

$|G_i|$ is the point number in G_i

(5) If C_i^1=C_i i=1，2，…，K (or the algorithm has achieved the hypothesis biggest iterative times) then terminate the algorithm, else make C_i=C_i^1 and return (3)

4. The clustering algorithm based on PSO

In the clustering algorithm based on Particle Swarm Optimization algorithm, each particle Y_i= (y_1，y_2, …, y_K) represents centers of the K classes, y_j (j=1，2，…，K) represents the central point's coordinates vector of the j^th^ class in the i^th^ particle (the dimension of y_j is decided according to the actual situation). The particle swarm constitutes by many candidate classified plan. We know it is a key of clustering which use optimization algorithm to evaluate the quality of classification plan, so we propose an adaptability function f as follows:

$$f(Y_i) = \frac{\max(\bar{d}_1(Y_i))}{\min(d_2(Y_i))} \qquad \text{(d)}$$

$\max(\bar{d}_1(Y_i))$ is the maximum value of mean values of distances within same classes in the classification plan which is expressed by particle Yi , $\min(d_2(Y_i))$ is the minimum value of distances between classes in the classification plan which is expressed by particle Yi。

$$\max(\bar{d}_1(Y_i)) = \max_{j=1,2,\ldots,K}\Big(\sum_{\forall x_i \in y_j} \frac{d(x_i, y_j)}{|y_j|}\Big)$$

$|y_j|$ is the element number in the jth class.

$$\min(\ d_2(Y_i)\ ) = \min_{\forall i, j, i \neq j}(d(y_i, y_j))\ ；$$

i,j =1，2,…,K.

The minimum value of the adaptability function f simultaneously satisfies small distance within same class and big distance between classes, the classification plan is better.

The clustering algorithm based on Particle Swarm Optimization algorithm consists of the following steps:

(1) In the n dimension space, we set population size m, acceleration coefficient c_1 and c_2, the hypothesis biggest iterative times num, clustering number K, a given point set which has N points etc. Initialize a population of particles with random positions and velocities (the position and velocity vectors are constituted by K vectors of n dimension space), then set the historical best position of each particle p_{best} equal to the initial position and set global best position of particle swarm g_{best} equal to the best of all p_{best}

(2) For each particle Yi, recalculate distances between the set {x_1，x_2, …, x_N} and K centers and

divide the set {x_1, x_2, …, x_N} according to the distance regulation of K-means algorithm.

(3) For each particle Y_i, Calculate the fitness evaluation according to the expression $f(Y_i)$

(4) Compare and reset the historical best position p_{best} and the best fitness evaluation of each particle, Compare and reset the global best position g_{best} and the best fitness evaluation of particle swarm.

(5) Change the velocity and position of particles according to equations (a) and (c) and limit them according to equations (b) and (e).

$$\begin{cases} x_{id} = x_{\max} & if \;\; x_{id} > x_{\max} \\ x_{id} = -x_{\max} & if \;\; x_{id} < -x_{\max} \end{cases} \qquad \text{(e)}$$

In the expression (e), we select the maximum value of each dimension in all points as x_{max}

(6) Inspect termination condition (the algorithm has achieved the hypothesis biggest iterative times) if it is satisfied then terminate the algorithm, else return (2)

(7) Output classification result

5. Algorithm test and comparative analysis

We carried on the Clustering test to the new clustering algorithm and contrasted the test results with k-means clustering algorithm. The experimental data divides into two groups, the first group is city coordinates of Hopfield-10 TSP, and the second group is Iris (150, 4, 3), which is a benchmark example known as classification, in the clustering algorithm based on Particle Swarm Optimization algorithm, we set acceleration coefficient c_1 = c_2=1.3, ω linearly reduces from 1.0 to 0.3, two algorithms were carried on 10 times, we took the average optimal solution.

Distribution map of the first group data as follows:

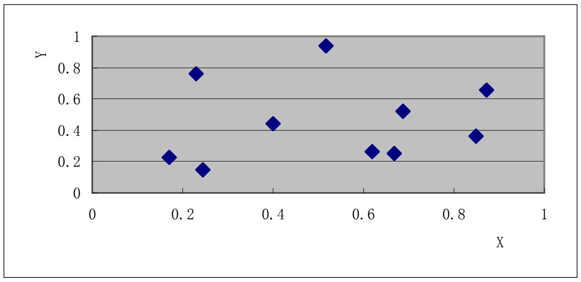

Figure 1 the scatter diagram of 10 city

In the test, we set population size is 10 and calculate the fitness evaluation of two algorithms according to expression (a).In the same classified result situation, the contrast result is as follows:

Table 1 contrast result1

method	clustering number	average best fitness value
clustering algorithm based on PSO	3	0.26224
k-means clustering algorithm	3	0.361840

the second group is Iris (150, 4, 3), This data set contains 150 sample records, it is from the flowers sample of setosa,versicolor and virginica, there are 50 records in every class, Each record has 4 attributes: sepal length, sepal width, petal length and petal width, the unit is the centimeter, considering visibility of graph, we only carry on the experiment to two attributes of petal, Figure 2 is the scatter diagram of the Iris data set petal attribute.

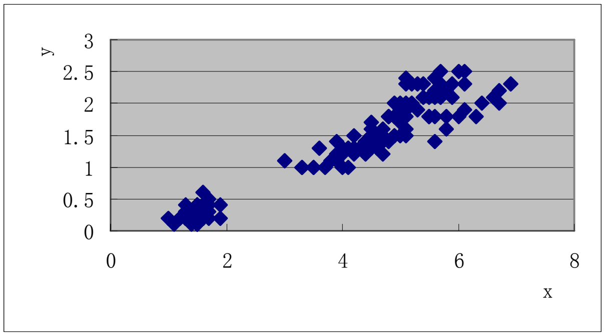

Figure 2 the scatter diagram of the Iris data set petal attribute

In the Figure 2, the setosa data separate from other data, but the separation property is very bad about versicolor and virginica, therefore, we select 100 data of versicolor and virginica and carry on the test. In the experiment we set population size is 130; the test result is as follows:

Table 2 contrast result2

method	clustering number	average best fitness value	error rate
clustering algorithm based on PSO	2	0.273040	5%
k-means clustering algorithm	2	0.340202	8%

From above two experiments, we can see both the k-means clustering algorithm and the clustering algorithm based on PSO obtain good classified result for dealing with simpler problems, but the latter' fitness evaluation is better. When we increase the data quantity, in the test result of the clustering algorithm based on PSO, the 28th and the 34th of versicolor are divided into virginica, the 7th, the 27th and the 39th of virginica are divided into versicolor. Total error rate is 5%, in the test result of the k-means clustering algorithm; the 28th and the 34th of versicolor are divided into virginica, the 7th, the 20th, the 24th, the 27th, the 28th and the 39th of virginica are divided into versicolor. Total error rate is 8%. The former error rate is lower than the latter. This has demonstrated the advantage of the clustering algorithm based on PSO.

6. Conclusion

The k-means clustering algorithm is one of the most widely used clustering method for its Simple idea, But the random selection of Starting centers may lead to different clustering results, even has no solution, considering the global search ability of PSO, this paper proposes the clustering algorithm based on PSO and designs corresponding adaptability function. The algorithm is evaluated on Iris plants database, Results show that the algorithm is more effective and promising. The experimental results indicate that the clustering algorithm based on PSO is better than the k-means clustering algorithm.

References

[1] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: *A Review*", *ACM Computing Survey*, 1999, pp.1-60.

[2] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations", *In: proceedings of the 5th Berkeley Symposium on mathematics Statistic Problem*, 1967, pp: 281-297.

[3] Y. Shi, R.C. Eberhart, "Fuzzy adaptive particle swarm optimization", *In: Proc. of the IEEE CEC,* IEEE Press, 2001, pp.101-106.

[4] Y.H. Shi,R.C. Eberhart, "A modified particle swarm optimizer", *In: Proc. of the IEEE CEC* ,IEEE Press, 1998, pp.69-73.

[5] Y. Z. Yao, Y. R. Xu, "Parameter analysis of particle swarm optimization algorithm", *Journal of Harbin Engineering University,*2007, pp.1242-1246

[6] S. Gao, J.Y. Yang, "New Clustering Method Based on Particle Swarm Algorithm", *Journal of Nanjing University of Aeronautics &Astronautics*, 2006, pp.62-65.

[7] L.H. Lu,B. Wang," Improved Genetic Algorithm-based clustering approach", *Computer Engineering and Applications,* 2007, pp.170-172.

[8] J. Kennedy, R.C. Eberhart, and Y. Shi, *Swarm Intelligence*, Morgan Kaufman Publishers, San Francisco, 2001.

[9]M. Clerc, J. Kennedy, "The particle swarm: explosion, stability, and convergence in a multi-dimensional complex space", IEEE Trans. Evolut. Comput. , 2002, pp. 58-73.

[10]S. Katare, A. Kalos, and D. West, "A hybrid swarm optimizer for efficient parameter estimation", *In: Proc. of the IEEE CEC*, IEEE Press, 2004, pp: 309-315.