A Clustering Method Based on PSO-GA Optimization Algorithm

Yanrong Yang

Network & Education Technology Center, Northwest A&F University, Yangling, Shaanxi, 712100, China

Abstract

The conventional clustering methods have many drawbacks, and with the advent of information era, it brings the data with complicated structures and large scales. These large-scale and complicated data have greatly challenged clustering analysis techniques. Actually, it is very difficult for the traditional clustering analysis techniques to process data of different types. Nor do they have the capacities to find the clusters of other shapes than globular cluster. Besides, they have difficulty in overcoming such problems as uncertain number of clusters, random selection of initial cluster center and easiness to get trapped into prematurity. Therefore, they have greatly affected the uniformity and dispersion of clustering distribution as well as the clustering effect. In recent years, swarm intelligence technology has been increasingly combined with clustering analysis and new models and ideas have kept coming into being. Under this circumstance, it is not difficult to find that swarm intelligence technology has a bright development prospect and great potential in the field of clustering. In order to conquer various problems which can't be solved by traditional clustering algorithms, this paper proposes a PSO-GA K-means clustering method based on the characteristics of fast convergence, global optimization ability and complementation of PSO and GA and sets the corresponding fitness function. The experiment result proves that the method of this paper not only inherits the strong global optimization capacity of PSO and GA, but it also leads to better clustering uniformity, higher solution accuracy, stronger stability and higher efficiency.

Key words: Particle swarm optimization (PSO), Genetic Algorithm (GA), K-Means Clustering.

1. INTRODUCTION

Clustering is to classify the spatial physical data with similar features so as to minimize the difference of the data objects of the same class and maximize that of the data objects of different classes. Through clustering, people can better utilize and analyze the data and effectively acquire useful information from plenty of data. K-means clustering algorithm is one of the most frequently used method in clustering analysis, but its classification result relies on the selection of initial cluster center. Many researchers have attempted to overcome the said defect of K-means clustering algorithm with the excellent self-organization, adaptivity and self-learning ability of intelligence algorithms( Huang and Lai et al.,2015). Particle swarm optimization (PSO) and genetic algorithm (GA) have brought significant benefits to the improvements of clustering field and techniques as well as themselves. PSO is a newly-emerging bionic optimization algorithm. Based on swarm intelligence, this algorithm has the global optimization capability and every particle of the swarm can search the global optimal solution with their own experience or by learning from others(Örkcü and Özsoy et al., 2015). GA simulates the natural selection of Darwin’s Theory of Biological Evolution and the computation model in the biological evolution of genetics and it is a method which searches the optimal solution by simulating the natural evolution process. As it is easy to understand, easy to achieve and strong in global search ability, it has become an emerging heuristic global search algorithm with rapid development. Because a large number of basic clustering algorithms require the users to provide certain prior knowledge on clustering, such as the expected number of classes, it makes the clustering result very sensitive to the input parameters, which has greatly limited the adaptability of the algorithms, especially for the data sets which include high-dimensional data. On the other hand, the clustering algorithm based on PSO or GA does not demand prior knowledge so that it reduces the users' burden and improves the clustering result(Tlili and Krichen, 2015; Jadoun and Gupta et al., 2015).

Conventional clustering analysis methods can be divided into two kinds: hierarchical method and partitioning method. The operations of hierarchical method are irreversible. So, if a certain combination or split point is not well selected, the quality of clustering result may be quite low. As for the partitioning method, it requires identification of the number of clusters at first and it is very sensitive to noises and initial value. As partitioning method performs clustering according to the distance between the objects and the optimization of certain standards, it is easy to get trapped into local optimum and it can only find globular clusters. Nowadays, K-means clustering algorithm is the most common method due to its
concision and efficiency. In 1995, inspired by the research result of artificial life, James Kennedy and Russell Eberhart had proposed the design model of PSO with its design idea coming from the research on the preying behaviors of bird flock. They had found that at the beginning, all birds began to fly in a random manner, that is to say, they don’t have specific directions. However, as time passed, all birds flew along the same direction at the same speed and gathered to a certain point finally. James Kennedy and Russell Eberhart realized that in the foraging behavior of bird flock, the birds had accelerated the speed to find foot greatly by sharing information among the members (İnkaya, 2015; Netjinda and Achalakul et al., 2015). Through cooperation, the overall group interest they obtain has greatly increased. GA is a random search method evolved by reference to the evolutionary law of biological world and the genetic mechanism of “survival of the fitness”. Professor J. Holland in USA was the first person to come up with this algorithm in 1975. It directly operates on structure objects without limitations of derivation and function continuity. It has internal implicit parallelism and better metaglobal optimization capacity. It uses probability optimization method, it can automatically obtain and instruct the search space of optimization, adjust the search direction in an adaptive manner and it doesn’t need specific rules. The clustering analysis based on swarm intelligence algorithm is one of the important contents of data mining and pattern recognition (Alam and Dobbie et al., 2015; Ferraro and Giordani, 2015).

This paper firstly analyzes the principles of particle swarm optimization and K-means clustering algorithm. Then, it applies PSO and GA to the clustering analysis of data mining and proposes a PSO-GA clustering method. It builds the model on clustering problems with their characteristics and explains its principles in details. Finally, it is the experiment simulation and analysis. It conducts the simulation experiment with the widely-used, typical multi-modal standard test functions and compares its operation result with those of other algorithms in order to verify the global search ability and comprehensive capacities of the algorithm. This paper compares the method of this paper with other clustering algorithms and the result shows that the method of this paper is better in terms of convergence accuracy and classification performance.

2. PRINCIPLES OF K-MEANS CLUSTERING

At first, give \( n \) samples and the number of clusters \( k \) to be generated. Randomly select \( K \) objects as the initial cluster center. Then, calculate the distance between every object to each cluster center and put every object under the cluster where the nearest cluster center is located. After distribution completed, the center of every cluster will change. In this case, re-calculate the cluster center through means. Repeat the above procedures continuously until any of the following condition is met (Beauchemin, 2015).

(1) No change to any (or the fewest) cluster centers. In other words, adjustments have been completed and the average error criteria function has been converged.

(2) No (or the fewest) samples have been put under different clusters, namely that no classification of samples has been adjusted.

(3) Local minimum sum of squared error.

K-means clustering algorithm searches a classification scheme of \( k \) clusters through iteration and use the mean value of \( k \) clusters to represent the minimum overall error obtained from various samples. The foundation of K-means algorithm is the minimum-sum-of-squared-error criterion and its cost function is as follows.

\[
E = \sum_{j=1}^{k} \sum_{x \in C_j} \| x - m_j \| ^2
\]

In this above formula, \( m_j \) represents the mean value of the \( j \)th cluster. We hope to minimize the cost function. To be visual, the more similar the samples of the same class are, the smaller the square error between the sample and the mean value of that class is. Find the sum of the square-error obtained from all classes and it can be proven that when dividing into \( k \) classes, whether various clusters are the optimal. Partitioning cluster method needs to select a certain distance as the similarity measurement among the samples when performing clustering. The shorter the distance between two objects is, the more similar they are and the longer the distance is, the smaller the similarity is. Select the mean value of objects in the cluster to perform similarity computation (İnkaya, 2015; Peng and Wang et al., 2015).
3. PRINCIPLES OF PARTICLE SWARM OPTIMIZATION

In PSO, the position \( x_i = (x_{i1}, x_{i2}, \cdots, x_{in}) \) of every particle represents potential solution to the optimization problem. During the flight, the flying speed of particle is represented as \( V_i = (v_{i1}, v_{i2}, \cdots, v_{in}), i = 1, 2, \cdots, m \). In every iteration, the particle updates the position and speed by tracking two extremums. The individual extremum \( P^*_i = (p_{i1}, p_{i2}, \cdots, p_{in}) \) and the global extremum \( P^*_g = (p_{g1}, p_{g2}, \cdots, p_{gn}) \) represent the optimal solution generated by the particle from various iterations and the existing global solution. The particles of every generation will update its speed and position according to the following formulas.

\[
v_{id} = w v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \quad (2)
\]
\[
x_{id} = x_{id} + v_{id} \quad (3)
\]

In the above formulas, \( r_1 \) and \( r_2 \) are two uniformly distributed random figures in the scope of \([0,1]\) respectively; \( c_1 \) and \( c_2 \) are two positive constants, known as accelerating factor and used to adjust the maximum step length of the global optimal particle and the optimal individual particle respectively; the proper \( c_1 \) and \( c_2 \) can make it easy to converge and not get trapped into local optimal; \( \omega \) is the inertia weight factor and \( d \) is the number of dimensions in \( D \) dimensions (Montain and Blanco et al., 2015).

In the iteration and updating, the speed and position of every dimension of the particle have strict limited scope. If the speed \( V_i \) of a certain dimension of a particle exceeds the maximum speed \( V_{d\text{max}} \) set by the algorithm by accelerating the particle, the speed in that dimension will be limited to the maximum speed \( V_{d\text{max}} \) of dimension. Generally speaking, the selection of \( V_{d\text{max}} \) shall not go beyond the width scope of the particle. Set a bigger \( V_{d\text{max}} \) and ensure the global search capacity of the particle swarm. If \( V_{\text{max}} \) is small, the local search ability of PSO will be enhanced. But if \( V_{d\text{max}} \) is too big, it may cause the particle fly pass the position of the optimal solution and if it is too small, it might reduce the global convergence ability of the particle. PSO continuously tracks the search through many iterations until the stipulated error is met or the stipulated number of iterations are achieved (Yuan and Ji et al., 2015).

4. PRINCIPLES OF PSO-GA CLUSTERING ALGORITHM

In the clustering analysis based on PSO-GA, every particle represents the center of \( K \) classes. Therefore, for every particle \( z_i = (c_{i1}, c_{i2}, \cdots, c_{ik}) \), \( C_{ij} \) represents the coordinate vector of the center of the \( j \)-th class of the \( i \)-th particle. Particle swarm is made up of many candidate classification schemes. To evaluate the classification schemes is the key to apply optimization algorithms to perform clustering. Generally, the following fitness function is used.

\[
F(z_i) = 1/\sum_{i=1}^{k} \sum_{v_{j} \in C_{ij}} d(x_i, c_{ij}) \quad (4)
\]

This fitness function represents the sum of the inner-class distance of all classes. In this formula, \( C_{ij} \) is the corresponding cluster of \( c_{ij} \). Though there is certain relationship between the inner-class distance and the between-class distance in the clustering result, such relationship is not certain and it will cause insufficient evaluation strategies for such clustering result. The evaluation made by this paper on every particle is determined by the following fitness function \( f \).

\[
f(z_i) = w_1 d_{\text{min}}(z_i) - w_2 \overline{d}_{\text{max}}(z_i) \quad (5)
\]

In this formula, \( w_1 \) and \( w_2 \) are the given positive constants, \( \overline{d}_{\text{max}}(z_i) = \max_{j=1,2,\cdots,K} \{ \sum_{v_{j} \in C_{ij}} d(x_i, c_{ij}) / |C_{ij}| \} \), \( |C_{ij}| \) is the number of elements of \( C_{ij} \); \( \overline{d}_{\text{max}}(z_i) \) represents the maximum average inner-class distance in the
corresponding classification to $z_i$ and $d_{\text{min}}(z_i) = \min_{v_i \neq p} \{d(c_d, c_p)\}$ is the minimum between-class distance for the corresponding class to $z_i$. In this case, it only needs to search the minimum value of $f$ to make the classification scheme meet the conditions of maximum inner-class distance and maximum between-class distance. To adjust $w_1$ and $w_2$ can facilitate different priority search strategies.

The analytical steps of clustering algorithm based on PSO-GA are classified as follows.

1. Initialize the population parameters, including the population size $n$, the total number of generations $\text{max}_\text{gen}$ evolved in the mixed algorithm, the two learning factors $c_1, c_2$ in PSO, the speed vector and position of every particle, which refer to the vector constituted by $k$ vectors in $R^v$, the maximum speed $v_{\text{max}}$ and the number of generations $t$ evolved of particle swarm as well as the crossover probability $c_p$ and mutation probability $m_p$ in GA.

2. For every particle $z_i(t)$, calculate the distance between the set $\{x_1, x_2, ..., x_n\}$ to be classified and the corresponding $k$ centers to the particles and perform classification on $\{x_1, x_2, ..., x_n\}$ according to the distance.

3. Calculate the fitness $f(z_i(t))$ of particle from classification. Calculate the value of fitness function according to the given fitness function, sort $n$ individuals according to the value of fitness function, calculate the mean value of the fitness function value of every particle and directly pick up $m_k$ individuals with better fitness value than the mean value.

4. Evolve the rest $n-m_k$ individuals with GA.

5. Calculate the individual optimal solution and the population optimal solution and update the state of the particle through the speed $v_i(t)$ and position $x_i(t)$ of PSO.

6. Put the $m_k$ individuals directly picked out from PSO and the $n-m_k$ individuals obtained from GA together and form a new particle swarm.

7. $k = k + 1$, turn to (3) until the present conditions or the maximum number of iterations have been met. Output the optimal solution, namely the position of the particle and the value of the optimal fitness function.

5. EXPERIMENTAL TEST AND ANALYSIS

5.1. Test Functions

1. For the multi-modal Griewank function, $f(x) = \sum_{i=1}^{n} \frac{x_i^2}{4000} - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$. Here, $x_i \in (-600,600)$ and $\min f(x) = f(0,0,\cdots, 0) = 0$. Such function is the multi-modal function which can’t be separated in different dimensions. The local optimal scope becomes narrower with the increase of number of dimensions. Besides, the number of local optimal solutions grows exponentially and it is increasingly probably to find the global optimal solution.

2. As for the multi-modal symmetrical Rastrigin function, $f(x) = \sum_{i=1}^{n} (x_i - 10 \cos(2\pi x_i) + 10)$. Here, $x_i \in (-6,6)$ and $\min f(x) = f(0,0,\cdots, 0) = 0$. This function is a symmetrical multi-modal function. It uses cosine function to generate many local optimal points. This algorithm is easy to get trapped into local optimal during the search process and it is not easy to find global optimal point. The two-dimensional figures of this function are shown as follows.
5.2. Comparison of Algorithm Performance

As shown in the Tab.1, this paper has tested two complex functions with variables with dimensions of 10 and 20, which can show the optimization ability of PSO-GA. In order to avoid or reduce random impact, run every test functions 30 times. The parameters are set as follows, population size $N = 40, P_r = 0.7, P_m = 0.005, c_1 = 1.49445, c_2 = 1.49445, w\in[0.5,0.9]$, the number of iterations is 300.

<table>
<thead>
<tr>
<th>Demension</th>
<th>Iteration</th>
<th>GA</th>
<th>PSO</th>
<th>PSO-GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Griewank function</td>
<td>10</td>
<td>300</td>
<td>0.2447</td>
<td>0.4177</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>500</td>
<td>25.6932</td>
<td>32.6316</td>
</tr>
<tr>
<td>Rastrigin function</td>
<td>10</td>
<td>300</td>
<td>2.1568</td>
<td>3.4472</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>500</td>
<td>53.3755</td>
<td>74.6508</td>
</tr>
</tbody>
</table>

It can be seen that the algorithm of this paper can achieve better convergence in the test functions and effectively solve the function optimization problem and its actual output value is quite approximate to the expected output value, thus giving strong methodological support to K-means clustering classification.

5.3. Clustering Analysis and Test

First, a random two-dimensional Gauss distribution (normal distribution) of dataset is created, which including three centers, and then the clustering analysis is applied by using this method, distribution contour lines, clustering effects and thermal energy diagram are shown in the following Fig.2.
In order to verify the accuracy and effectiveness of the algorithm, simulation experiment has been conducted on the algorithm by selecting Forest Fires, Las Vegas Strip and Car Evaluation from UCI standard data sets. The characteristics of the data sets are shown in Tab.1. 517 samples in Forest Fires with 13 attributes for every sample, 504 samples in Las Vegas Strip with 20 attributes for every sample and 1728 samples in Car Evaluation with 6 attributes in every sample. See details in Tab.2.

Repeat the experiment for 30 times and see the statistics of clustering results by four different algorithms on four datasets in Tab.3, which has listed the error rate of different algorithms on different datasets.

6. CONCLUSIONS

As a research focus, clustering analysis has drawn sufficient attention in the field of data mining. With the predominant intelligence, diversity, excellent self-organization and adaptivity shown in the theory.
of swarm intelligence, the clustering methods based on intelligence computation have witnessed rapid development. This paper has given the model and steps for PSO-GA algorithm to be used in clustering analysis and simulation experiment has been made. The results show that it can lead to satisfactory result and it is more suitable for the clustering problems with high-dimensional space and non-uniform data distribution and it is not sensitive to noisy data.

REFERENCES

Alam Shafiq, Dobbie Gillian, Rehman Saeed Ur,(2015) “Analysis of Particle Swarm Optimization Based Hierarchical Data Clustering Approaches”, Swarm And Evolutionary Computation, 25, pp.36-51.


