Acta Technica *62* No. 2B/2017, 563–572 

# The large data particle clustering algorithm based on minimum variance response[1]

Zeng Jun[2]

**Abstract.** Large data clustering has good application value in the field of pattern recognition and fault diagnosis. A large data clustering algorithm based on minimum variance response is proposed. Firstly, the standard particle swarm algorithm is analyzed, the principle of minimum variance response and large data clustering is studied. Using the method of particle swarm optimization to reconstruct and extract the feature vector of large data information flow. Two approaches are proposed based on the limitation conditions of the minimum variance distortionless response beamforming [1]. The first one translates the uncertainty Constraints into the uncertainty for the whole extended steering vector. The second one uses the structure information of signal. Simulation results show that the algorithm can effectively improve the accuracy of data classification, reduce the error rate and improve the performance of data mining and feature extraction.

**Key words.** Large data clustering, particle swarm, minimum variance distortionless, response, beamforming.

## 1. Introduction

With the development of information computing science, the research of biology is applied in computational science, which realizes the design of intelligent bionic optimization algorithm and improves the processing and analysis ability of large data. The intelligent bionic algorithm mainly includes ant colony algorithm, Particle swarm optimization and quantum swarm optimization. Through the biomimetic principle of this kind of biological group or particle, the mathematical model of man and nature is simulated to realize the design of group intelligent optimization algorithm. The group intelligent optimization algorithm has better performance in

[1]This work was supported by the Project of Education Commission in Chongqing: Research on parallel mining of large data based on Hadoop architecture, KJ15012021 and "Chunhui" project of Ministry of education: Preliminary application of big data on intelligent agriculture platform of Internet of things, S2016038.

[2]College of Computer Engineering of Yangtze Normal University, Chongqing, China, 408100

artificial intelligence design, data clustering analysis and computer control [2].Large data clustering has a good application value in the fields of pattern recognition and fault diagnosis. In this paper, we introduce particle swarm optimization algorithm to improve the data clustering algorithm, and propose a large data particle clustering algorithm based on minimum variance response. Particle Swarm Optimization (PSO) is a population-based adaptive stochastic optimization algorithm, which is derived from Kennedy and Eberbar's research on foraging behavior of bird population. Now particle swarm optimization (PSO) is widely used in pattern recognition, data mining and intelligent control. In this paper, the standard particle swarm algorithm is analyzed, and the principle of large data clustering is studied by particle swarm optimization algorithm. Particle swarm space recombination method is used to reconstruct and extract the feature vector of large data flow, and optimize clustering is realized. Finally, performance tests were carried out by simulation experiments, showing better data clustering performance.

## 2. Standard particle swarm optimization and data clustering

### *2.1. Standard particle swarm optimization*

Particle swarm optimization (PSO) is based on the simulation of biological activities, and a new intelligent optimization algorithm is constructed by simulating the ability of the group to cooperate with each other. But the particle swarm algorithm itself comes from the phenomenon of biological groups, the theoretical basis is not complete. And because of its stochastic approximation optimization algorithm, it is mainly applied to continuous region, so the algorithm has the disadvantages of premature convergence and difficulty in applying discrete problem. Therefore, it is very important to study the theoretical analysis, algorithm improvement and discretization of particle swarm optimization [3].

The standard particle swarm algorithm and the discrete binary particle swarm algorithm are analyzed and improved. We obtain the following results: (1) Particle swarm algorithm is a heuristic stochastic optimization algorithm, each particle chase its own optimal particle and global optimal location search, and chase with random factors. Particle swarm algorithm in this random search process, the particles eventually converge to the group optimal particles. In this paper, it is theoretically proved that the trajectories of the particles converge to the optimal particle position of the population under the condition of increasing the randomness and the optimization of the particle optimal point. According to the theoretical results of the analysis, the principle of algorithm weight selection is further explained.

(2) Since the particle trajectory finally converges to the population optimal particle, this paper defines the concept of similarity between particles, and designs the concept of calculating the diversity of the population particles. By calculating the average similarity of the population particles and the optimal particles of the population, and measure the diversity of particle groups. Based on the population clustering degree and its similarity with the optimal particle size, each particle is randomly generated [4]. Thus, an improved algorithm for the standard algorithm is

constructed to improve the global search ability of the algorithm, to avoid premature convergence, and to improve the standard performance of the algorithm.

(3) The weight of the standard algorithm is the parameter of the global search and local search ability of the equilibrium algorithm, and its value affects the performance of the algorithm.The weights of the standard algorithm are linearly decreasing from the early to the later, but the weight of each particle is the same. In this paper, according to the similarity between particles and the optimal particles of the population, different weights are given to different particles, so that the weight of each particle is different, and it changes with the iteration of the algorithm. This constructs a particle swarm algorithm with dynamic change of weight.

(4) The theoretical analysis of the algorithm constructs a mathematical model, and the mathematical model clearly reflects the mathematical meaning of the algorithm itself from a mathematical point of view [5].By using this mathematical model instead of the updating formula of the original algorithm speed and position, a new evolutionary algorithm is obtained, and the selection of new evolutionary algorithm parameters is analyzed. The new algorithm can directly reflect the mathematical thinking of the algorithm. The simulation results show that the new algorithm is not worse than the standard algorithm.

### *2.2. The basic flow of particle swarm optimization*

The basic particle swarm algorithm is as follows:

Step 1: Initialization: The position and velocity of the particles are randomly generated in the dimension space of the problem space.

Step 2: Evaluation of particles: For each particle, evaluate the applicable value for the dimension optimization function.

Step 3: Update the best: 1) Compare the particle value and its individual optimal value $p_{\text{best}}$, if better than $p_{\text{best}}$, its $p_{\text{best}}$ position is set to the current particle position. 2) Compare the value of the particle to the population's optimal value $g_{\text{best}}$. If the current value is better than $g_{\text{best}}$, set the $g_{\text{best}}$ position to the current particle position.

Step 4: Update the particles: change the speed and position of the particle.

Step 5: Stop condition: the loop returns to Step 2 until the termination condition is met, usually satisfying the applicable value and the maximum iteration algebra.

Corresponding to the above algorithm flow, the basic framework of particle swarm algorithm is shown in Fig. 1.

### *2.3. Clustering algorithm overview*

Traditional clustering algorithms such as hierarchical, average-linkage clustering algorithm, segmented K-Means clustering algorithm, neural network-based pre-SOM clustering algorithm and density-based DBSSAN algorithm, etc. They have been proven effective in many applications, but they also have some problems: scholars have not yet proved that they can produce the global optimal cluster, and these clustering algorithms can not be efficient and accurate processing of high-dimensional

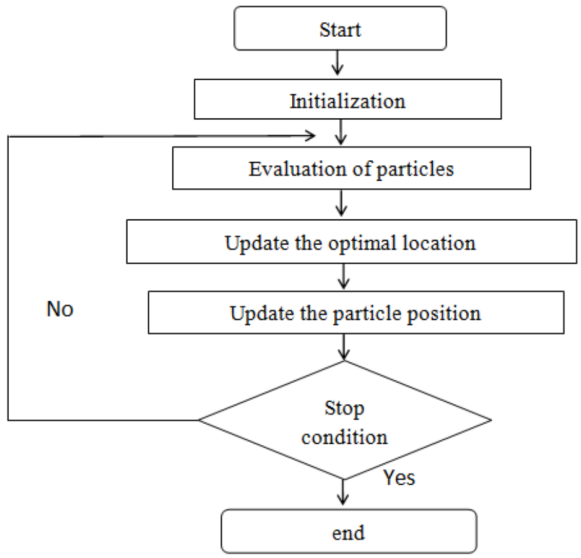


Fig. 1. The basic framework of particle swarm optimization

large number of data clustering. So we propose a large data clustering algorithm based on the nature and characteristics of the minimum spanning tree MST in graph theory. The algorithm tries to simplify the clustering problem into the edge of the MST. It is also the edge of the maximum weight of the MDT. The MST is divided into several sub-trees [6]. Each sub-tree is an optimal cluster. The process does not need to consider the size of the data and the distribution of the data shape, not only solve the above problems, but also efficient and accurate processing of large quantities of data. The algorithm can find abnormal points, and can be used in combination with other clustering algorithms, which has some expansibility.

Based on the theory of graph theory, the data set is preprocessed by quantifying the interrelatedness of each data object point. The data clustering analysis based on the network is based on the large data of $K$th high dimension. Constructing adjacency: $W$ data points are vertices, and the adjacency matrix between the data is the weight of the edge, and a full graph is constructed: and then the MST of the whole graph is generated. According to the actual problem and the distribution state of the data, $A$ sub-tree is the sub-tree of the smallest spanning tree, and a sub-tree is an optimal clustering of the data set.

An adjacency matrix is a data matrix that uses a two-dimensional data set to represent the relationship between data points. If graph $G$ is the weighted network graph, we can define it as:

$$A_{ij} = \begin{cases} w_{ij} & \text{if } (v_i, v_j) \text{ or } < v_i, v_j > \in E(G), \\ 0 & \text{if } (v_i, v_j) \text{ or } < v_i, v_j > \notin E(G). \end{cases} \tag{1}$$

Here, $w_{ij}$ represents the weight of the edge, and Figs. 2 and 3 show the process of generating the adjacency matrix from the graph

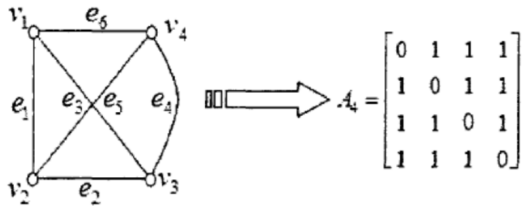


Fig. 2. The original distribution of large data

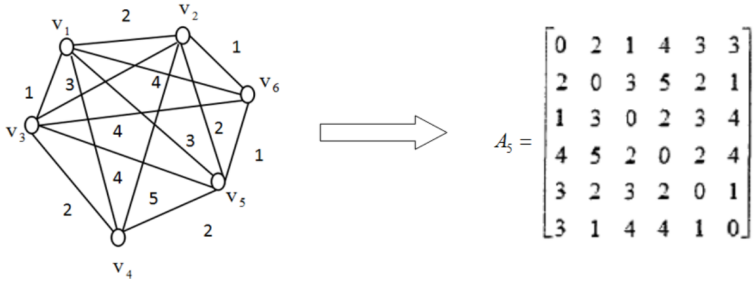


Fig. 3. Generate adjacency matrix by empowering the whole graph

From this we can see that the adjacency matrix of the network graph is a symmetric matrix, and the elements of the $i$th row in the matrix are the distances $w_{ij}$ between the vertex $V_i$ and the vertex $V_j$ in the weighted graph.

In this paper, we obtain the adjacency matrix between the data object points, so we get the full graph represented by the adjacency matrix, and let the weight of the data object point. The algorithm of the minimum tree gets the MST of the full graph, and then divides the edge of the smallest tree. And then in accordance with the size of the MST edge of the division of the smallest tree edge, get a number of the smallest tree subtree. Each subtree is an optimal Cluster.

The clustering algorithm is an unsupervised machine learning method, which is a method of dividing data between the degrees of data objects. It can classify data with high similarity or high relevance as a cluster, but the algorithm's own subjectivity is very strong; the algorithm uses different, the starting point assumes that the difference in the number of clusters specified or pre-defined will make the results different.

### *2.4. Realization of large data optimization clustering*

The method of particle swarm optimization is used to reconstruct and extract the feature vector of large data flow. The optimization of the algorithm is described as follows: The hybrid algorithm based on particle swarm optimization is a pair of chromosomal data with three segments. In the process of calculating the use of each

region of the data parallel operation to improve efficiency [7], as a basis for large data clustering to get clustering center vector distance:

$$(d_{ik})^2 = ||x_k - V_i||^2 . \tag{2}$$

The particle swarm optimization is performed to obtain the particle recombination state:

$$\sum_{i=1}^{c} \mu_{ik} = 1, \ k = 1, 2, ..., n . \tag{3}$$

Set the population size $M$, the total number of iterations of the algorithm $T$, the current number of iterations $T_N$, and obtain the phase shift velocity $V_i = (v_{i1}, v_{i2}, ..., v_{iD})^{\mathrm{T}}$ and position $x_i = (x_{i1}, x_{i2}, ..., x_{iD})^{\mathrm{T}}$ of the zero point trajectory of the multi-beam particles. Based on the constraint condition, the maximum value of the clustering objective function is obtained by using the particle swarm depth zero point trajectory optimization theorem [8]

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \frac{d_{ik}}{jk} \right)^{\frac{2}{m-1}}} , \tag{4}$$

$$V_i = \frac{\sum_{k=1}^{m} (\mu_{ik})^m x_k}{\sum_{k=1}^{n} (\mu_{ik})^m} . \tag{5}$$

The disturbance variables loaded into the population of individuals are

$$x_{n,G} = x_{n,G} + \Delta x_i . \tag{6}$$

For each particle, a random number between [0, 1] will be generated. If $r \leq m$, and the subscript is the $i \neq g_{\mathrm{best}}$, the large data stream is reconstructed and extracted according to the information of each particle in the population $P_i$ and $P_g$, we obtain the characteristic space of the multi-beam particle group depth zero locus:

$$x_i = (x_{i1}, x_{i2}, ..., x_{iD})^{\mathrm{T}} . \tag{7}$$

The multi-beam particles are added into the population to produce the recombination Schwefel 1.2 function $z$ of the particle swarm space, which is the C×D dimension. Through the above processing, the particle swarm optimization data clustering algorithm is improved [9].

## 3. Simulation experiment and result analysis

In order to test the performance of this algorithm in the realization of large data clustering, we conducted a simulation experiment, the experimental use of personal PC as a hardware environment, PC operating system: Windows 7, the processor Intel (R) Core (TM) 2 Duo CPU frequency 2.93 GHz. The number of particles is $N_{\mathrm{s}} =$ 200, 500, 700, and 1000. The sampling interval is 1024 dB, the interference signal

to noise ratio is -3 dB, the other parameters are $n = 30$ K, $m = \{20,\ 50,$ and 100, the data is sampled at a sampling time of 20 s, Each component of the cluster population is between (0, 1) [10]. According to the simulation environment and parameter design, the large data clustering simulation is carried out. First, the distribution of the large data samples is given as shown in Fig.4.

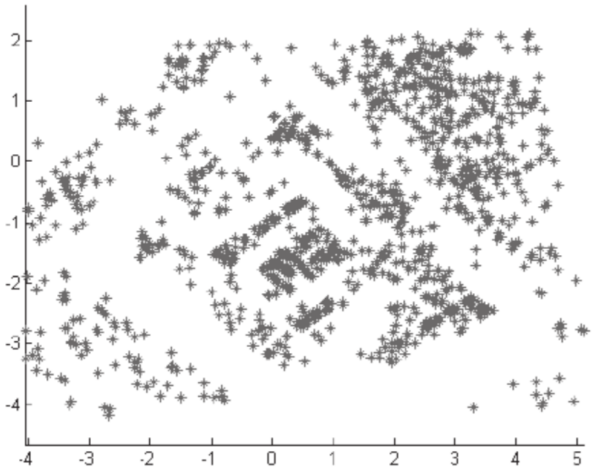


Fig. 4. The original distribution of large data

With the above data as test samples, large data feature extraction and classification processing, improve the data fusion and mining capabilities, we get clustering results shown in Fig. 5, we can see from the figure, this algorithm can effectively achieve large data clustering. The accuracy of clustering is 99.39 %, the precision is higher, the error rate is 12 % lower than that of the traditional method, and the application value is better.

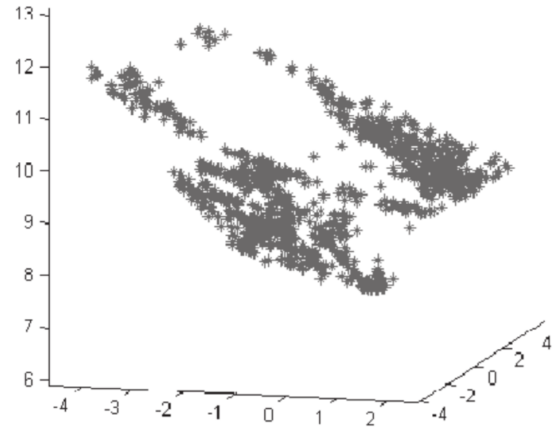


Fig. 5. The original distribution of large data

Simulation experiments are carried out to simulate the algorithm in this paper. Based on the minimum variance response, the large data particle clustering algorithm has obvious advantages both in the accuracy of the result, the complexity of the calculation or the efficiency of the calculation. This algorithm has the advantages of dealing with high - dimensional large - scale data, and it is an accurate, efficient and fast clustering algorithm, which can solve the shortcomings of traditional clustering algorithm. And in today's large data for the mainstream of the era of the algorithm has a very good practicality, can quickly and efficiently solve the problem.

## 4. Conclusion

This paper expounds the importance of data mining under the era of large data, and points out that the main content of this paper is based on the analysis of large data particle clustering based on minimum variance response. As an important means of data mining, clustering analysis is a kind of important data mining method, which divides the similarity data objects into the same data class cluster through certain criterion. Each subset in the sample cluster has some similar Attributes. However, the traditional clustering algorithm presents the shortcomings of low clustering efficiency, low precision and long calculation time when faced with high-dimensional large data set. Based on the nature of the minimum variance response, this paper proposes a clustering analysis of large data by using particle swarm optimization method. The adjacency matrix between each data object is constructed by correlation analysis. Through simulation experiment using the clustering method, we compare the algorithm with the traditional clustering algorithm: the large data particle clustering algorithm based on the minimum variance response, regardless of the accuracy of the result, the complexity of the calculation or the efficiency of the calculation. The algorithm has the advantages of dealing with high-dimensional large-scale data.

**References**

[1] I. S. Dhillon, D. S. Modha: *Concept decompositions for large sparse text data using clustering.* Machine Learning *42* (2001), Nos. 1–2, 143–175.

[2] H. Takizawa, H. Kobayashi: *Hierarchical parallel processing of large scale data clustering on a PC cluster with GPU co-processing.* Journal of Supercomputing *36* (2006), No. 3, 219–234.

[3] R. J. Hathaway, J. C. Bezdek: *Extending fuzzy and probabilistic clustering to very large data sets.* Computational Statistics & Data Analysis *51* (2006), No. 1, 215–234.

[4] S. Asharaf, M. N. Murty: *An adaptive rough fuzzy single pass algorithm for clustering large data sets.* Pattern Recognition *36* (2003), No. 12, 3015–3018.

[5] J. Cervantes, X. Li, W. Yu, K. Li: *Support vector machine classification for large data sets via minimum enclosing ball clustering.* Neurocomputing *71* (2008), Nos. 4–6, 611–619.

[6] R. Mendes, J. Kennedy, J. Neves: *The fully informed particle swarm: Simpler, maybe better.* Journal IEEE Transactions on Evolutionary Computation *8* (2004), No. 3, 204–210.

[7] Y. del Valle, G. K. Venayagamoorthy, S. Mohagheghi, J. C. Hernandez, R. G. Harley: *Particle swarm optimization: Basic concepts, variants and applications in power systems.* IEEE Transactions on Evolutionary Computation *12* (2008), No. 2, 171–195.

[8] I. C. Trelea: *The particle swarm optimization algorithm: Convergence analysis and parameter selection.* Information Processing Letters *85* (2003), No. 6, 317–325.

[9] C. A. C. Coello, G. T. Pulido, M. S. Lechuga: *Handling multiple objectives with particle swarm optimization.* IEEE Transactions on Evolutionary Computation *8* (2004), No. 3, 256–279.

[10] M. Wolfel, J. McDonough: *Minimum variance distortionless response spectral estimation.* IEEE Signal Processing Magazine *22*, (2005), No. 5, 117–126.

Received September 12, 2017