Advances in Social Science, Education and Humanities Research, volume 105
International Conference on Sports, Arts, Education and Management Engineering (SAEME 2017)

Parallel CSA-FCM Clustering Algorithm Based on MapReduce

Chunchun Cui [a], Runtong Zhang [b,*]
School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China;
[a]2425738357@qq.com, [b]15120621@bjtu.edu.cn

Keywords: FCM, CSA, Noise Immunity, MapReduce.

Abstract. Fuzzy C-Means (FCM) algorithm is a kind of widely used clustering algorithm, which is widely used in pattern recognition, image processing, medical research and other fields. But FCM doesn't have better performance suppressing noise. A parallel clustering algorithm based on MapReduce is proposed in this paper, which combines Clonal Selection Algorithm and the algorithm uses intelligent optimization method to optimize the initial clustering center, and makes use of the global search ability of CSA to make the algorithm more robust. The algorithm process is designed to conform to the MapReduce programming model and it has the ability of dealing with large-scale dataset. The experiments prove that parallel Clonal Selection Algorithm-Fuzzy C-Means (CSA-FCM) can improves the searching performance and the noise immunity and has high speed up and scalability.

1. Introduction

In fuzzy cluster analysis, the fuzzy C mean (FCM) clustering is a typical algorithm. The algorithm firstly from randomly selected sample of a set of data as the initial clustering center, and then plug in membership and cluster center formula calculation, found that the objective function to minimize sample points so far, the whole process is a simple iteration. The fuzzy C mean (FCM) clustering algorithm and its derived algorithm are successfully used in the field of pattern recognition, classification, data mining and graphics processing. But as the scale of the data increases dramatically, the traditional serial computing model has failed to meet the demand for large data volumes.

At present, the researchers at home and abroad have proposed many parallel schemes for FCM algorithms. However, the FCM clustering algorithm also has its own shortcomings and disadvantages, such as the problem of the isolation point and the sensitivity of noise data to the algorithm robustness.

Clonal selection algorithm is used in parallel global random search strategy, so the algorithm can obtain the global optimal solution of the probability is higher also, at the same time the algorithm has the advantages of simple, general and strong robustness. Moreover, the algorithm has the characteristics of self-learning, self-organizing and adaptive, suitable for parallel processing. Computing model based on graphs, this paper puts forward a hybrid clonal selection algorithm and FCM algorithm of parallel CSA-FCM clustering algorithm, the problem of FCM algorithm to deal with noise data, which can improve the robustness of the fuzzy c-means clustering algorithm, and algorithm on Hadoop platform performance experiment.

Fuzzy C mean clustering algorithm has many disadvantages, such as choicing of isolated point, noise data, and fuzzy indicator m and be sensitive to the initial solutions and the iteration is easy to fall into local extremum, resulting in the accuracy of clustering results is not very high and the computational overhead of clustering large data sets is relatively large. At present, the typical work of fuzzy C means clustering algorithm mainly includes: [2]using the information entropy improves the method of Fuzzy C mean algorithm to get the formula of cluster center solving the problem of uniform shrinkage of the fuzzy C mean algorithm.[3] The initial clustering center is selected according to the maximum density principle of the data region, overcoming the defects of random selection of the initial cluster center is easy to make the algorithm converge to local minimum.[4] The concept of attribute weight interval supervision is applied to FCM clustering algorithm, But the decision maker's experience has a great influence on the clustering algorithm.[5] A new method for selecting cluster centers is proposed. The method takes the midpoint of the nearest two sample points

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

as the clustering center. However, it is possible to select the midpoint of the two noise points to make the algorithm sensitive to noise data.

2. Parallel CSA-FCM Algorithm

First, we should solve these key problems by using clonal selection algorithm for cluster analysis: The selection of each clonal operator is discussed, and the parameters of each operation parameter are determined according to the actual conditions. The second is how to encode the problem well into the antibody in the process of clustering. Finally, the antibody antigen affinity function is constructed. Only by solving these problems well can we make better use of the CSA algorithm, and make the improved fuzzy C mean clustering algorithm achieve the best results.

(1)Encoding

According to the real number encoding the initial cluster centers for the corresponding antibody in the clonal selection algorithm, the initial cluster center of each dimension quantization are encoding. The form of specific antibodies is: $P_{11}P_{12}```P_{1d}P_{21}```P_{2d}```P_{c1}P_{c2}```P_{cd}$. Among them, c is the number of cluster centers, and d is the dimension of sample data.

(2)Construction of Affinity Function

From the iterative process of fuzzy C mean clustering algorithm we know, in order to obtain better clustering results, the objective function value must be smaller, each class in the clustering center and each sample point between the more natural affinity. Therefore, we can construct the affinity function of the algorithm by using the fuzzy C mean objective function F(U,V), Formula is as follows:

$$f = \frac{1}{F(U,V)+1} = \frac{1}{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}{}^{m} d_{ij}{}^{2}(x_j, v_i)+1} \tag{1}$$

(3)Mutation

Suppose that the j dimension of the first i antibody in the sample is $P_i(j),\ i=1,2,```,n, j=1,2,```,d$,The antibody is mutated according to the following formula:

$$P_i(j) = P_i(j) + \lambda_i \times N(0,1) \tag{2}$$

Among them, N (0,1) is a random number between 0 and 1, which is the probability of antibody mutation. After the increase of the number of iterations, the clustering is stable, and the value decreases dynamically.

Specific steps are described below:

Step 1 set the initial population of the algorithm, the number of antibodies is K, and randomly generate the initial antibody population P (k), set the initial clustering number of C algorithm is equal to 2, and the maximum number of clustering is Cmax.

Step 2. first objective function using membership matrix U calculate formula in fuzzy clustering algorithm F (U, V), and then by the formula (1) to calculate the initial antibody population with each antibody samples in all individuals (i.e. the affinity of F antigen);

Step 3 K antibodies with the highest affinity were selected to enter the memory cell population. The clonal operator of clonal selection algorithm was used for cloning and mutation of antibody population concentrated on memory cells. Finally, the affinity between the new antibody population and all antigens was calculated, and the antibodies with lower affinity in the new antibody population were deleted. The new antibodies P'(k) were formed by inhibiting the antibodies with similar distances. If the termination condition of the clustering algorithm is satisfied, then the next step is returned, or else the step 2 is returned.

Step 4 Using the formula (1), the antibody with the highest affinity in the new antibody population was calculated, and the corresponding clustering center was used as the optimal clustering result when the number of clusters was c;

Step 5 If the number of clusters, then the formula is used (3) to calculate the intra class distance of each cluster, and then the fuzzy C mean algorithm is used to split the class with the largest distance within the class. The initial cluster centers are randomly generated, together with the original

clustering centers to form new antibody populations, and the number of clusters is set. Go back to step two and proceed to the next iteration. If the number of clusters is C=C+1, then turn to the next step;

$$S_i = \sum_{j=1}^{n} d_{ij}^{2}(x_{ij}, v_i) \tag{3}$$

Among them, the meaning of vi and dij is the same as that of the FCM algorithm.

Step 6 The fuzzy clustering algorithm of target function value of minimum output as the best antibody clustering results, and the maximum membership degree of each cluster antigen and antibody (data) is part of the cluster.

The execution of the algorithm is shown in figure 1:

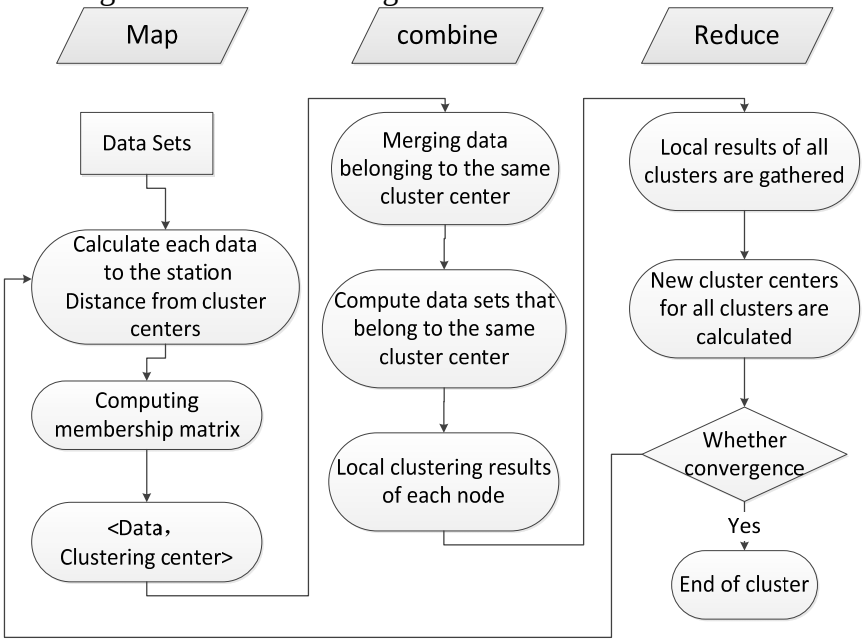

Fig. 1 flow chart of algorithm

3. Experiment Results

3.1 Environment

The experiment is based on the Hadoop framework, and it uses 8 nodes, one as master and the other as slaves. The hardware configuration of each node is as follows: CPU model is Intel-Core, main frequency 2.0GHz, memory 2GB.

The data used in this experiment are data from the UCI machine learning database selected from the US 1990 Census. The dimension of the data set is 68 dimensions, with 2458285 records. By processing the data set, the data set is divided into six different size datasets, each of which grows at a rate nearly 2 times the size. As shown in Table 1 below:

Table 1. Data Set settings

Data set name	Record number	The size of data
Dataset1	614571	81.1MB
Dataset2	1229142	162.3MB
Dataset3	2458285	324.5MB
Dataset4	4916570	649.1MB
Dataset5	9833140	1.26GB
Dataset6	19666280	2.53GB

3.2 Algorithm Performance Analysis

In the experiment, the running time of six data sets in 1 clusters, 2 nodes, 4 nodes and 8 nodes on Hadoop cluster was measured respectively. The experimental results are shown in Table 2 below:

Table 2. CSA-FCM algorithm running time (seconds)

Data Sets	1 nodes	2 nodes	4 nodes	8 nodes
Dataset1	1885.8	1824.9	1761.3	1876.5
Dataset2	2824.4	2061.3	2007.4	1843.2
Dataset3	5480.6	3686.5	2480.0	1977.8
Dataset4	8761.0	5387.4	3462.8	2132.4
Dataset5	17467.9	9437.0	6194.3	3048.5
Dataset6	34543.7	18611.5	11071.7	5363.9

In the analysis of the experimental results, the acceleration ratio and expansion rate are used as evaluation indexes. Speed-up ratio is an important evaluation index to evaluate the performance of serial processing and parallel processing algorithms, and it describes a performance improvement represented by running time reduction. It is equal to the serial processing time of the algorithm divided by the parallel processing time.

In experiments, a single node Hadoop cluster, that is, an Master and a Slave, is used as an approximation of the serial execution time for speedup calculations. The experimental results of the algorithm are shown in figure 2:

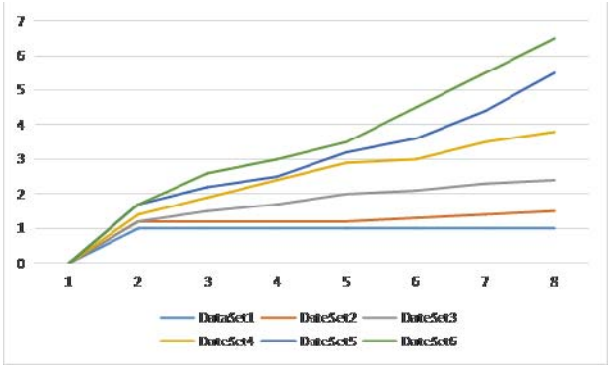

Figure 2. speedup of each data set

As we can see from Figure 2, the speedup ratio of the algorithm is nearly linear. When data size is small, because of the Hadoop cluster needs to take up time to start and communicate, this time occupies a larger proportion of the whole processing time, and the speed-up ratio performance of the algorithm is not very good. However, as the data scale increases, the speed-up ratio of the algorithm is better than the performance. Therefore, the speedup performance of the algorithm is better for large data processing.

3.3 Noise Immunity Analysis of the Algorithm

The scaling rate is used to test the change in operating efficiency as the size of the data increases with the number of nodes, the efficiency of the algorithm is the ratio of the speedup of the algorithm to the number of nodes. In the experiment, three groups of data were tested, and The first group tests the efficiency of different size data sets DataSet1, DataSet2, DataSet3 and DataSet4 on 1 nodes,2 nodes,4 nodes and 8 nodes' operating efficiency. The second groups tested the efficiency of different size data sets DataSet2, DataSet3, DataSet4 and DataSet5 on 1 nodes, 2 nodes, 4 nodes and 8 nodes' operating efficiency. The third groups tested the efficiency of different size data sets DataSet3, DataSet4, DataSet5 and DataSet6 on 1 nodes, 2 nodes, 4 nodes and 8 nodes' operating efficiency.

The experimental results of the algorithm are shown in Figure 3. As we can see from the graph, the running rate of the algorithm tends to be steady as the size of the data set and the number of nodes increase at the same rate. As we can see from the results, as the size of the data set increases, the performance of the algorithm is getting better and better.

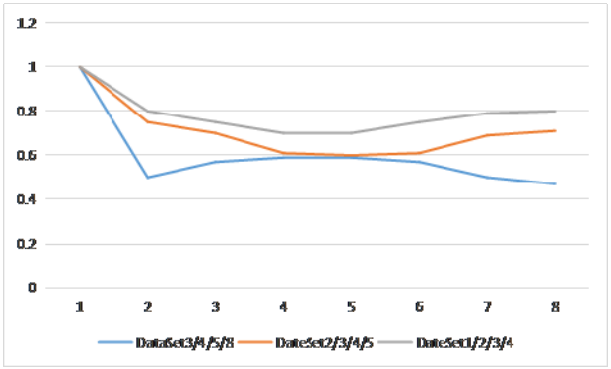

Figure 3. the efficiency of the data set on each node

4. Summary

In this paper, the parallel fuzzy clustering algorithm is sensitive to noise data, so the algorithm is not robust, and a parallel CSA-FCM clustering algorithm based on clonal selection algorithm and FCM algorithm is proposed. Parallel selection algorithm is a global random search using cloning strategies and self-learning, self-organizing and adaptive characteristics suitable for parallel processing, improve the anti-noise and large data execution efficiency of FCM algorithm. Experimental results show that the parallel CSA-FCM algorithm has good noise data processing ability and maintains good speedup and scalability in parallel environment.

Acknowledgments

This work was partially supported by a Key Project of National Natural Science Foundation of China (NSFC) with grant number 71532002.

References

[1]. Yuanzhuo Wang, Xiaolong Jin, et al.Network big data: present situation and Prospect. Journal of Computer Science. Vol. 36 (2013) No. 11, p. 1126-1136.
[2]. V. Cherkassky, F. Mulier. et al. Learning from Data Concepts, Theory, and Methods. New York: John Wiley and Sons, 2008, 413 – 417.
[3]. R. Krishnapuram, J. M. Keller. et al. An Approach to Clustering. IEEE Transactions on Fuzzy Systems, 2010 ,1(2): 98 - 110.
[4]. Tamalika Chaira. et al. A novel intuitionistic fuzzy C means clustering algorithm and its application tomedical images[J].Applied Soft Computing,2011,11(2):1711-1717.
[5]. Grira N, Crucianu M, Boujemaa N. et al.Active semi-supervised fuzzy clustering[J]. Pattern Recognition, 2008, 41(5): 1834-1844.
[6]. Hung M C , Yang D L. et al. An efficient fuzzy C-means clustering algorithm[C].IEEE International Conference on Data Mining (ICDM2001), California, USA, 2001:225-232.
[7]. Eschrich S,Ke J W,Hall L O. et al. Fast fuzzy clustering of infrared images[C].IFSA World Congress and 20th AFIPS International Conference,Joint 9th,Vancouver, BC, Canada, 2001, 2:1145-1150.
[8]. Wu Kuo lung,Yu Jian,Yang Miin shen. et al. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests[J].Pattern Recognition Letters,2005, 26(5):639-652.