Analysis method of big data based on K-means algorithm

Yichi Zhang¹, Tao Shu¹, Pengxiang Zhang¹, Yan Xu²

¹Air and Missile Defense College, Air Force Engineering University, Xi’an Shaanxi 710051, China
²Mathematics and Statistics Science College, Shaanxi Normal University, Xi’an Shaanxi 710119, China

*Corresponding author’s e-mail: zyc412181588@ahlctl.com

Abstract. Efficient and accurate clustering of massive data is the basic requirement of deep excavation for internal knowledge of data in the information era. Firstly, in the distributed parallel processing framework of Hadoop, the MapReduce model is applied to operate the K-means clustering algorithm, then the genetic algorithm is utilized to transform the clustering problem of K-means algorithm into the solution of extremum to solve the randomness confirmed in cluster center. After testing, this method has achieved the ideal effect in terms of clustering efficiency and clustering precision, and plays certain role in deeply excavating the internal knowledge rule of data.

1. Introduction

The development and application of internet technology and sensor technology has accumulated massive data information for the social progress and the computer is deemed as the transfer station of handling massive data. The explosive data generation speed facilitates people's deep understanding and perception of things in every field, but it also brings a burden to the computer data processing performance. As a result, both data processing efficiency and quality reduce [1]. K-means clustering algorithm is deemed as the classic method in processing big data clustering analysis. But the clustering results directly influence the advantages and disadvantages of knowledge excavation, modern information field proposes very high requirements to data clustering quality, besides, traditional K-means algorithm contains the following disadvantages at large: low processing efficiency of calculated amount big data, unsatisfactory convergence effect, etc. [2]. Therefore, the K-means algorithm is operated in the large-scale data distribution parallel processing framework of Hadoop, and the genetic algorithm is adopted to transform the clustering problems into solution of extremum in order to improve the accuracy and efficiency of K-means algorithm in the analysis of massive data.

2. The related model

2.1 K-means clustering algorithm

Clustering is a method of clustering research objects in the same data set according to similarity, which belongs to unsupervised learning algorithm. K-means clustering is one of the most classical clustering algorithms. It is a distance-based clustering method, that is, the distance is used as an index to judge whether two objects are similar or not. The data is more similar when the distance is closer, and then the objects are divided into the same data set for further mining and analysis, in which Euclidean distance is taken as the standard measurement. K-means clustering algorithm is very fast,
only need to calculate the distance between the point and the clustering center.

K-means algorithm is proposed by Lloyd, which can divide the data into \( k \) clusters that minimize the sum of squares of errors through continuous iterative calculation. The algorithm is widely used in all walks of life because of its simple and efficient operation, strong scalability, nearly linear time complexity and suitability for processing large data sets.

### 2.2. Improved K-means clustering algorithm

Two kinds of strategies are adopted to improve the operating performance of traditional K-means algorithm. On the one hand, the MapReduce model is applied in the large-scale data distribution parallel processing framework of Hadoop to achieve the parallelization of algorithm, optimize the operating efficiency of algorithm and save the time [3]; On the other hand, the clustering computing of K-means algorithm is completed based on genetic algorithm, and the clustering is transformed into the solution of extremum to improve the self-adaptive computing ability of algorithm and make the clustering results more accurate. In general condition, processing of big data classification problem with K-means algorithm is difficult to judge the size of clustering number \( k \) [4]. Therefore, the value \( k \) needs to be confirmed in the computing process for the optimization of K-means clustering with genetic algorithm as expounded in article. The computing process for optimization of K-means algorithm based on genetic algorithm is as below:

**Step 1:** Coding genetic operation. Since the clustering number \( k \) of clustering algorithm is confirmed in the change process, the code of \( k \) is added to the code of the original chromosome to update the design method of chromosome. \( 'W = w_1, w_2, \ldots, w_k' \) is the coding form of chromosome after updating. The length of expression form of chromosome depends on the clustering number \( k \), and the coding method is relatively flexible.

**Step 2:** Setting of clustering population reduction. The clustering population generates at random. After \( k \) individuals are selected from it at will, the original cluster center is represented with these individuals; Later, the coding method of real number is utilized to complete coding this group of individuals to work out a chromosome.

**Step 3:** Confirmation of fitness function of clustering. After the generation of \( K \) cluster centers in chromosome, the input vector used in clustering algorithm is the vector of each sample. The \( X \) is represented in the following method, and the input vector dimension is \( m \);

\[
X = [x_{11}, x_{12}, \cdots, x_{1m}]' \tag{1}
\]

The Euclidean distance is calculated based on the formula.

\[
\| x_i - w_i \| = \min_j [x_i - w_j] \tag{2}
\]

The sample is divided into the category with cluster center as \( w_i \). Inter-class spacing and intra-class spacing are the significant variables which need to be considered in the objective function in order to constrain the K-means algorithm to present the optimal clustering results. Therefore, those two variable design objective functions are considered as below:

\[
H = \frac{\sum_{i=1}^{k} \sum_{x_i \in w_i} \| x_i - w_i \|}{1 + \sum_{m \neq n} \| w_m - w_n \|} \tag{3}
\]

In formula (3), inter-class spacing and intra-class spacing are represented as \( \sum_{m \neq n} \| w_m - w_n \| \), \( \sum_{i=1}^{k} \sum_{x_i \in w_i} \| x_i - w_i \| \); in which \( m \) is 1, 2, \ldots, \( a - 1 \). \( N \) is \( i+1, \ldots, a \). The smaller objective function value is corresponding to comparatively excellent clustering results. At this time, the inter-class spacing is as big as possible and the extra-class is as small as possible, based on which, the objective
function \( f = \frac{1}{1+H} \), after transformation is used as fitness function.

**Step 4**: Genetic operation. After confirmation of basic parameters and functions of genetic algorithm iteration, the genetic operation such as selection operator, cross operator, mutation operator, etc. can be selected.

**Step 5**: Judgment of algorithm termination condition. The clustering results can be output after genetic algorithm reaches the upper limit of iteration times or the error reaches the termination condition; Otherwise, repeat operation of Step 3 till the clustering output results are worked out. The improved K-means algorithm conducts the following optimization and iteration based on MapReduce model:

1. Input the computer network dataset to be conducted the clustering analysis, and output k cluster centers to complete data clustering.
2. The output results of reducer function are confirmed based on genetic algorithm and transferred to the mapper function to calculate the distance between each sample data and the closest cluster center of genetic operation. At this time, the cluster of data samples and samples can be worked out [5].
3. The mapper function transfers the output results to the combine function, and the same cluster objects are merged in the following method: The sum of dimensions of different cluster sample data is gained and the quantity of basic data objects is summarized.
4. The combine function inputs the operation results into the reducer function, and then accurate cluster center quantity k is got through optimization of genetic algorithm. At this time, the sum of dimensions of all data samples and the total number of sample objects are summarized, the center value obtained through genetic operation is used as the effective K-means clustering algorithm center, and the algorithm convergence condition is checked [6].
5. The cluster center is obtained based on the K-means clustering algorithm after optimization of genetic algorithm to achieve the clustering analysis of data, with steps as below:
   1) The input sample data set is input into the mapper function;
   2) The optimal center branch confirmed finally is utilized to calculate the distance between sample data and k clusters and obtain the cluster of data objects;
   3) The mapper function inputs the results after calculation into the reducer function to complete collecting data, further obtaining the optimal network data clustering results.

3. Experimental analysis

3.1. The time analysis
In order to test the effectiveness and advantages of K-means algorithm improvement in the processing of massive computer network, the distributed parallel computing environment is constructed, and 5 sets of computers are set in LAN to constitute Hadoop cluster, in which one set of computer is defined as the Master node. Besides, the computers are defined as Slave nodes; Computer adopts 2GB memory, with 3.20GHz. 2.0×10^4 data are selected in the authoritative network database as the sample dataset used in clustering experiment. The fuzzy K-means algorithm and traditional K-means algorithm are used as the comparison test method, and both of them operate in the distributed operation framework in order to verify the article method and solve the clustering performance of massive data.

The article method operates in the stand-alone environment in the experiment at first, and the clustering time of different data volume is recorded in that method; Later, the article method, fuzzy K-means algorithm and traditional K-means algorithm are adopted for the distributed parallel clustering in the Hadoop clustering MapReduce model, the corresponding time is recorded, and the clustering time in four kinds of condition is summarized, as shown in Fig. 1.
In Fig. 1, the operation time of article method in stand-alone environment is lower than the operation time in distributed environment, because the improved clustering algorithm in distributed environment needs to traverse all dataset after a period. The overall trend of figure displays that, with the increase of data scale, the clustering time of parallel computation algorithm data in distributed environment reduces, with the trend of traditional K-means algorithm $>$ fuzzy K-means algorithm $>$ article method, in which the article method costs the least time and obtain the highest efficiency, because it adopts genetic algorithm to transform the clustering problems into the solution of extremum, without manual setting of cluster center $k$, which improves the adaptive degree of clustering algorithm. So the processing time of massive data reduces naturally.

3.2. The error analysis
The mean square error condition of cluster testing in each method in section 2.2 is shown in Table 1.

| The amount of data/10^4 | Stand-alone method/10^7 | K-means algorithm/10^7 | Fuzzy K-means algorithm/10^7 | Method of this paper/10^7 |
|------------------------|------------------------|------------------------|-----------------------------|--------------------------|
| 0.5                    | 3.125                  | 2.312                  | 2.001                       | 1.985                    |
| 1.0                    | 3.688                  | 2.678                  | 2.156                       | 1.648                    |
| 1.5                    | 3.979                  | 2.901                  | 2.579                       | 1.607                    |
| 2.0                    | 4.560                  | 3.117                  | 2.069                       | 1.355                    |

The table 1 shows that the clustering method in article in stand-alone operation environment not only has low efficiency, its clustering accuracy is also unsatisfactory. When the data volume is $2.0 \times 10^4$, the corresponding mean square error reaches $4.125 \times 10^7$; In the environment with the same data volume, the clustering mean square error operating in the article method in the Hadoop cluster MapReduce model is about $1.713 \times 10^7$, with the maximum accuracy. The algorithm is improved because the article method applies genetic algorithm to solve the problems confirmed in cluster center, and the inter-class spacing and intra-class spacing of K-means algorithm are considered in confirmation of the fitness function of genetic operation to guarantee the optimal clustering results.

4. Conclusion
New outlet of clustering analysis for massive network data is provided in the Hadoop distributed parallel processing framework which distributes the data clustering task in many computer nodes to change the previous situation of data processing in a single computer, greatly save the clustering analysis time of data and improve the data analysis efficiency. The improved K-means algorithm in article can improve the clustering efficiency of network data based on that method, in the meanwhile, the clustering problem is deemed as the problem of solution of extremum based on the genetic
algorithm, and the randomness of cluster center selection reduces, which guarantees the precision of clustering results of massive data, and provides effective solution for dividing the category of massive network data.

Acknowledgments
The paper was supported by The National Natural Science Foundation of China named active regulation mechanism research about in the process of medium or high carbon steel sheet in the fine blanking strain injury under the action of pulse current. (No. 51605488).

References
[1] Zhang Chengchang, Zhang Huayu, Luo Jianchang, etc. Based on cloud computing and K-means algorithm of massive amounts of electricity data analysis method [J]. Journal of computer applications, 2018, 38 (1) : 159.
[2] Cheng Ping, Zhang Li. Under the big data K-means clustering algorithm based on improved K tax risk identification [J]. Accounting monthly (accounting), 2019, 40 (21) : 143.
[3] Yin Aiying, Wu Yunbing, Zhu Minchen, etc. Based on graphs within the framework of K-means algorithm [J]. Computer application research, 2018, 35 (8) : 2295.
[4] Wang Jiayi. For large data processing calculation method of the parallel optimization clustering K-means [J]. Information and communications, 2018, 32 (5) : 32.
[5] Song Yang, Shi Hongyan. Based on the graphs under the framework of K-means clustering algorithm [J]. Computer and modern, 2019, 35 (8) : 28.
[6] He Yingsi. Based on large data of the improved fuzzy K-means algorithm [J]. Journal of Chongqing university of science and technology (natural science edition), 2018, 32 (12) : 145.