Parallel clustering method of power marketing Big data based on DBIK means algorithm and coarse granularity

Wei Xing*, Xiujie Shi, Botao Wu

State Grid Customer Service Center, Tianjin, 300300, China

*Corresponding author's e-mail: 13584048844@163.com

Abstract. By analyzing and classifying a large amount of electricity marketing data, we aim to better understand user groups and market segmentation, and provide decision-making support for power companies or suppliers. Therefore, a parallel clustering method for Big data of power marketing based on DBIK means algorithm and coarse granularity is proposed. Combining DBSCAN algorithm and K-means algorithm, DBIK means algorithm is designed to complete parallel clustering of Big data of power marketing to be clustered. Ensure that the clustering effect is coarse-grained, optimize the solution results, and realize parallel clustering of Big data of power marketing. Experimental results show that the proposed method has better parallel clustering effect of Big data of power marketing, higher clustering accuracy and shorter clustering delay.

Keywords: DBSCAN clustering; K-means algorithm; Coarse grained calculation; Electricity marketing; Marketing Big data of electric power marketing; Parallel clustering

1 Introduction

With the rapid growth of power marketing Big data in the power industry, the traditional serial clustering algorithm has been unable to meet the demand for efficient processing and accurate classification of massive power marketing Big data [1-3]. Therefore, researchers have turned to parallel clustering methods to improve the analysis efficiency and accuracy of power marketing Big data. Awad et al. [4] proposed k-means clustering technology based on neural processor to cluster Big data. Wang et al. [5] proposed a marginal extension of FDP, called K-means find density peaks (KFDP), to improve clustering performance.

However, the current parallel clustering method still has some challenges when dealing with Big data of power marketing. First of all, the existing algorithms lack in-depth research on the specific background of the power industry, and cannot fully tap the characteristics and laws of the Big data of power marketing in the power industry. Secondly, existing algorithms rely too much on fine-grained features when dealing with Big data of power marketing in the power industry, resulting in high computational complexity and low efficiency. Therefore, the purpose of this study is to pro-

A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_117

pose an efficient and accurate clustering method based on DBIK means algorithm and coarse-grained parallel clustering method for Big data of electric power marketing, through in-depth study of specific background and characteristics of the electric power industry, to solve the problems in the analysis of Big data of electric power marketing.

2 Design of parallel clustering algorithm

2.1 Flowing DBSCAN

Since the historical Big data of electric power marketing will have an impact on the new Big data of electric power marketing, the delay factor α is introduced to judge its impact. When the value of α is 1, the impact of the historical Big data of electric power marketing is the same as that of the new Big data of electric power marketing. When the value of α is 0, the impact of the historical Big data of electric power marketing can be ignored.

Suppose x_t represents the newly arrived Big data of power marketing, c_t represents the clustering situation of the last time or historical time, c_{t+1} represents the new clustering, n_{t+1} represents the number of Big data points of power marketing in the new clustering, and the new clustering is calculated to obtain:

$$c_{t+1} = c_t n_t \alpha + \frac{x_t m_t}{n_t \alpha} \tag{1}$$

Among them, n_t represents the number of historical power marketing big data, and m_t represents the number of newly arrived power marketing big data.

2.2 K-means algorithm

In order to control the missing judgment of the power marketing Big data generated in the judgment process, use the aggregation parameter obtained in the aggregation process[6], take the absolute value of the parameter as the fixed distance of the node, integrate the power marketing Big data into a cluster, and divide the cluster into multiple components according to the number of attributes of the power marketing related information. After repeating the above processing process, Delineate a variety of distance categories to form a variety of cluster head nodes, and control the processing of power marketing Big data. Without considering external factors, the judgment state formed by aggregation may form multiple cluster heads, which can affect the energy efficiency of the actual operation of the power marketing related information system. In order to eliminate the additional energy consumption generated by external factors on the power marketing related information system, Gaussian elimination method is

used to re process the determined cluster heads. After determining the energy and position parameters of each cluster head, Simulate the energy generated during the actual operation of power marketing related information systems, build a linear matrix to determine the energy of cluster head nodes, and determine the number of rounds of actual power marketing Big data. And integrate the remaining energy parameters to determine the security status of the obtained information.

2.3 DBIK-means algorithm

The power marketing Big data DBIK means algorithm is designed through the K-means algorithm, which divides the power marketing Big data flow calculation into the following steps:

① Selection and utilization of historical power marketing Big data

Select representative historical Big data of electric power marketing to establish a training Big data set of electric power marketing to obtain different types of electricity use characteristics. Cluster the Big data of electric power marketing through DBSCAN algorithm [7] and label the Big data points of electric power marketing. The marked Big data points of electric power marketing are represented by p, and the attributes and types of Big data points of electric power marketing are represented by y and z, respectively. The following results are obtained:

$$p=\left(y,c_{t+1},z\right) \tag{2}$$

During the operation of the power system, the selected historical power marketing Big data will be updated to real power marketing Big data covering multiple time intervals. For the K-means algorithm [8], these power marketing Big data will become broadcast variables and be distributed to all nodes.

② Vector calculation of power marketing Big data in unit time

Using electricity consumption values to describe electricity consumption characteristics[9-10], assuming that the effective electricity consumption value is represented by I_e, the sampling period and sampling time are represented by T and $i\left(t\right)$ respectively, and the number of big data on electricity marketing within a sampling period is represented by N, we can obtain:

$$N=p\frac{I_e}{T}\int_{t=0}^{T} i\left(t\right)^2 dt \tag{3}$$

The K-means algorithm is used to calculate the Big data points of electric power marketing collected in the time interval and output the vector. The calculation is divided into two parts, namely Map and Reduce. The Map part is used to analyze the Big data of electric power marketing and obtain the Big data of demand electric power marketing; Reduce is used for power marketing Big data stipulation, consolidation and classification of power marketing Big data, and output vector.

③ User information join operation

The join operation is used to merge user information with household information, outputting the final electricity information vector. The K-means algorithm is supplied to the join operator, and the join completes the RDD merge through household numbers.

④ DBIK-means algorithm

When the quantity of Big data of sample centralized power marketing meets the conditions, if the power consumption information of a user is inconsistent with its historical information or information of similar users, the power consumption information may have problems and needs to be observed. The Big data of power marketing to be clustered is mainly judged by whether it shows clustering characteristics horizontally and vertically. If it shows clustering characteristics, it is normal; otherwise, it is abnormal.

⑤ Obtain Big data of power marketing to be clustered

The delay factor obtained from the historical Big data of electric power marketing is used to establish the Big data of electric power marketing in the latest time interval as a training set. Repeat the above steps to complete the parallel clustering of Big data of electric power marketing to be clustered.

2.4 Coarse grain division

After completing the parallel clustering of Big data of electric power marketing to be clustered, in order to ensure the clustering effect, coarse grain computing is used to optimize the solution results. The specific steps are as follows:

The research goal of rough sets is to create an information table that can be represented by quads, as shown below:

$$S = N\langle U, A, V, f\rangle \tag{4}$$

Where, S represents the Knowledge representation and reasoning system, U represents the universe of non empty finite object set, A represents non empty finite attribute set, V represents attribute value range set, and f represents mapping function. Thus, the following expression can be obtained:

$$S' = S\left(U, At, L = \left\{V_a \middle| a \in At\right\}, f = \left\{I_a \middle| a \in At\right\}\right) \tag{5}$$

The optimal solution is obtained through the above process to realize the parallel clustering optimization of Big data of power marketing.

3 Experiments and Analysis

3.1 Experimental setup

The experimental environment adopts Intel Core i5-4210 1.8GHz processor, with a running memory of 8GB, and all nodes are connected through a 100Mbps Ethernet converter. The software is Hadoop version 2.6. 14321 electric power marketing Big data are selected as experimental data, and some electric power marketing Big data are shown in Table 1.

Table 1. Power marketing Big data set at different times of a day

Index	10:45:00	11:44:00	12:43:00	13:42:00
Revenue data/GB	10.1	11.5	11.2	12.9
Electricity consumption data/GB	23.3	24.6	25.4	26.3
User Information Data/GB	16.1	17.4	18.4	19.3
Electricity pricing data/GB	20.3	19.4	20.8	20.1

To verify the effectiveness and feasibility of the method proposed in this paper, reference [4] and [5] methods were used as comparative methods for experiments.

3.2 Result Analysis

Keep the above experimental environment unchanged, control the aggregation parameters of the three clustering methods to be the same, count the power marketing Big data set prepared by the three clustering methods clustering experiment, define the clustering delay generated by the three clustering methods as the fluctuation time generated by clustering power marketing Big data, and count the clustering delay generated by the three clustering methods. The results are shown in Table 2:

Table 2. Clustering delay results generated by three clustering methods

Number of power marketing Big data sets	Cluster latency/s		
	Reference [4] Method	Reference [5] Method	Method in the paper
2	0.51	0.31	0.16
4	0.52	0.30	0.18
6	0.50	0.37	0.17
8	0.54	0.38	0.17
10	0.57	0.31	0.17
12	0.55	0.36	0.13
14	0.54	0.34	0.17
16	0.54	0.34	0.18
18	0.52	0.38	0.12
20	0.51	0.35	0.14

According to the delay results shown in Table 2, it can be seen that under the control of the three clustering methods, the delay generated by the proposed method is

about 0.1 seconds, which is compared to the clustering methods in the two literature, The clustering method designed in the article generates a shorter delay time, and the actual clustering process requires the shortest clustering time.

To verify the effectiveness and feasibility of the algorithm, compare the clustering accuracy of the three methods for power marketing Big data. Figure 1 shows the accuracy comparison of the three methods.

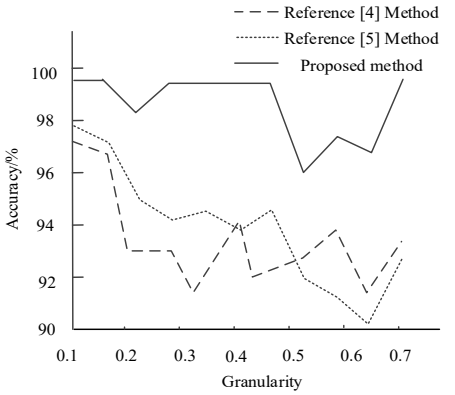

Fig. 1. Accuracy Comparison

According to the trend of each method in Figure 1, it can be seen that the accuracy of the two comparison methods continuously decreases with the increase of probability threshold, and the downward trend is very obvious; And the accuracy of the method in this article is higher, and the amplitude of change is smaller, indicating that the method has a certain degree of stability.

4 Conclusion

Based on DBIK means algorithm, this research designs and implements a parallel clustering method suitable for the power industry to improve the efficiency and accuracy of power marketing Big data analysis. The design experiments have verified the good advantages of this study. The results of this study are expected to provide an efficient and accurate clustering method for Big data analysis of big power marketing in the power industry. This will help power enterprises better understand and use the massive power marketing Big data, optimize marketing strategies, and improve competitiveness and market share. At the same time, this study can also provide reference and inspiration for big data analysis in other fields of big power marketing, promoting the development and application of parallel clustering methods in a wider range of application fields

Acknowledgments

This article is the result of the State Grid Customer Service Center-2023 SGCC business operation early warning analysis and risk control ability improvement project.

References

1. Zhang, Y., Ai, Q., Li, Z. (2021) Grouping of dynamic electricity consumption behaviour: An f-divergence based hierarchical clustering model. IET Generation, Transmission & Distribution, 15(22): 3163-3175.
2. Li, Y., Zhou, R., Liang, F. (2022) Application research of cloud computing in electricity marketing field measurement remote acquisition system. Mechatronic Systems and Control, 50(1): 55-60.
3. Chen, W, Fan, J., Du, H. (2023) Investment strategy for renewable energy and electricity service quality under different power structures. Journal of Industrial and Management Optimization, 19(2): 1550-1572.
4. Awad, F. H., Hamad, M. M. (2022) Improved k-Means Clustering Algorithm for Big Data Based on Distributed SmartphoneNeural Engine Processor. Electronics, 11(6): 883.
5. Wang, G., Fu, T., Ren, H. (2022) K-means Find Density Peaks in Molecular Conformation Clustering. Chinese Journal of Chemical Physics, 35(2): 353-368.
6. Wang, Q. M., Hu, D. C. (2022) Optimization of K-Means Fast Clustering Algorithm Based on Spark. Computer Simulation, 39(3): 344-349.
7. Kawtar, S., Damien, J., Roger, G. (2021) A data mining approach for improved interpretation of ERT inverted sections using the DBSCAN clustering algorithm. Geophysical Journal International, 225(2): 1304-1318.
8. Pan, H., Lei, Y., Yin, S. (2021) K-means clustering algorithm for data distribution in cloud computing environment. International journal of grid and utility computing, 12(3): 322-331.
9. Fang H, Wang Y W, Xiao J W ,et al. A new mining framework with piecewise symbolic spatial clustering[J].Applied Energy, 2021, 298(6):117226.
10. Kalmi P, Trotta G, Kazukauskas A. Energy ln elated Financial Literacy and Electricity Consumption: Survey–Based Evidence from Finland[J]. Journal of Consumer Affairs, 2021,55(3):1062-1089.