The Research of Electrical Behavior Hybrid Computing Technology Based on FCM Clustering Algorithm

Jiangpeng Dai[1,*], Aihua Zhou[1], Wei Rao[1], and Lipeng Zhu[1]

[1]*Institute of Computing Technology and Applications, State Grid Smart Grid Research Institute, Nanrui Rd. #8, Nanjing, China*

Abstract. With the development of smart grid construction work and the innovation of the business model of the China State Grid Corp, power generation, power transmission, power transformation, power distribution, consumption and scheduling, and other business areas generate a large number of data every day. At the same time, power system is one of the important application areas of big data. In this paper, we study the model of large data transaction between power system and other industries based on the background of the Sino-Singapore Tianjin Eco-City smart grid constructions. First, the sources and characteristics of the multi energy sources in the Sino-Singapore Tianjin Eco-City are analyzed. Then, the problem of large data transaction is modeled as a multi attribute negotiation problem, and the utility function of multi attributes negotiation is proposed to reflect the degree of association among different attributes. Finally, a multi-objective optimization algorithm is proposed for the large data trading of smart grid by using the vector evaluation genetic algorithm, which can achieve a win-win situation for both the buyer and the seller.

1 Introduction

With the continuous deepening of the construction of the smart grid in the China State Grid Corp, Advanced information technology and digital communication technology are applied in the power generation, transmission, distribution, scheduling, consumption and customer service and other aspects [1]. At the same time, with the rapid development of economy and the improvement of people's living standards, on the one hand, the residents use electricity in constant growth, on the other hand, the individual user's personalized requirements for the use of electricity services are gradually improving. Electric power enterprises provide customers with electric energy products as well as undertake professional guidance on the use of electricity to improve power utilization efficiency and utilization level of the task [2-5]. The satisfaction of these requirements depends on the data acquisition and data analysis techniques.

Based on smart meter data, statistics and tap the power customer's power mode, a power enterprise not only can master the customer’s composing and understand the characteristics of the electric behaviour, but also provide personalized, meticulous service to achieve customer intelligence, lean management of the prerequisites. However, with the development of electric power communication technology, the use of electric power information acquisition system to generate electricity every day is the high frequency mass of data, which acquires user behaviour characteristic analysis technique to be capable of deal with high speed, high precision processing of a large number of data and data types to explore the requirements of high value information. This is in accordance with the typical characteristics of large data applications, but also means that the use of traditional computing structures and data mining methods cannot meet the above requirements.

In this paper, we analyse the different requirements of the parallel computing and real time computation in the field of the intelligent power consumption. A hybrid computing architecture for batch processing and stream processing is proposed. Besides, using fuzzy C mean clustering (Fuzzy c-Means clustering, FCM) to analyze the data of the users is proposed. Based on hybrid computing structure and FCM clustering algorithm, we can make a quick and accurate judgment of the user's electricity pattern.

2 Hybrid computing architecture for data analysis

Traditional large data computing architectures such as Hadoop and Fourinone are suitable for dealing with large scale and high concurrency. But the time needed to calculate the time may reach tens of minutes or even hours, which results in the high delay time limit the applicability of these computing architectures in high real time applications. However, the high real-time streaming computing framework such as Spark, Storm are suitable for the distributed real-time computation of high speed and large data streams, whose data structure design and object relationship are not suitable for large-scale parallel computing. In this paper, a new type of scalable real-time hybrid data processing framework is proposed to solve the contradiction between parallel computing and high delay. As shown in Figure 1.

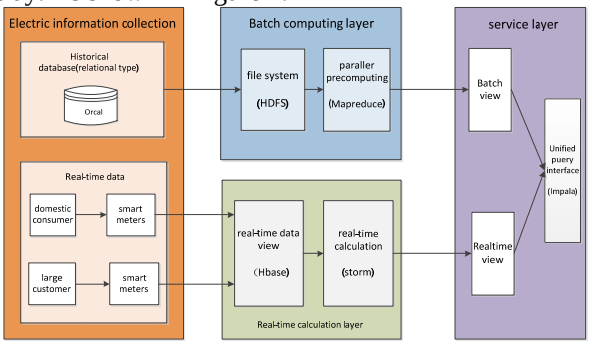

Figure 1. Analysis of hybrid computing architecture with data.

The data obtained by the electric information acquisition system are divided into two categories of

historical data and real time data. Historical data sample size is large, which is more accurate after some calculation such as clustering analysis, data mining and so on. And directly from the smart meter to the home users and large users with real-time data shows the characteristics of the flow data which is needed to be processed in real time. According to the different requirements of data processing, the hybrid computing architecture is divided into three layers:

(1) Batch computing layer: There are two main functions based on Hadoop Apache framework. The first function is to store the same batch data set, which is the same as that of the data set which is based on the distributed file storage system HDFS. The second is to use MapReduce to calculate the duration of the data view. The calculated object is the complete data set, so the calculated frequency will not be too high. For some big data applications, taking into account the possible size of the data set and the node computing power, the time required for a MapReduce cycle may be several hours.

(2) Stream processing layer: Based on Storm architecture and HBase Apache database, the data source of the stream processing layer is the real-time data acquisition. Its main function is to compensate the high delay time of the batch layer by using the Storm framework. The real-time view thus obtained can be used to complement the delay of a batch view, but because the data sample size is small, the accuracy is not as low as the batch view, so the life cycle will be replaced by a lot of view.

(3) Service layer: Based on the Cloudera Impala engine, the query interface is responsible for the external application of open data view, so that the external application can be based on real-time requirements of the batch and real-time view of the data to index and query.

2.1 FCM Based power consumption characteristics analysis with hybrid computing

2.1.1 Analysis of the characteristics of the power consumption based on FCM algorithm

The fuzzy clustering (Clustering Fuzzy) algorithm takes into account the uncertainty of the real data, and compared with the Hard Clustering. Fuzzy clustering algorithm allows a data object to be multiple different clustering, data objects and the degree of closeness of each cluster center can be measured by the degree of membership, so its application is more flexible. In this paper, we use the fuzzy C-means (C-means Fuzzy, FCM) clustering algorithm to analyze the power behavior. FCM clustering algorithm based on objective function is suitable for processing large amounts of data, and the algorithm is simple, so it is easy to be realized on the computer. It is suitable for dividing the complex data set based on time series, which is in agreement with the characteristics of the data. The core idea of FCM algorithm is solved by $\min\{J_m(U,P)\}$, thus obtaining the optimal partition matrix and clustering center matrix. For the classification of object model space contains n members set $X=\{x_1,x_2,\cdots,x_n\}$. U can be expressed as:

$$U=[\mu_{ik}]=\begin{bmatrix}\mu_{11} & \cdots & \mu_{1n}\\ \vdots & \mu_{ik} & \vdots\\ \mu_{c1} & \cdots & \mu_{cn}\end{bmatrix}, \text{ among them, } 1\le i\le c, 1\le k\le n \quad (1)$$

Among them, $\mu_{ik}=\mu_{x_i}(x_k)$ represents the subordinate relationship between the sample x_k and the subset $X_i(1\le i\le c)$. For FCM the range of μ_{ik} is [0, 1]. Namely, the subordinate relationship between each sample and the subset X_i can be represented by a real number of 0~1, and $P=\{p_i,1\le i\le c\}$ said the cluster center matrix class I subset of X_i.Optimization objectives can be expressed as:

$$J_m(U,P)=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2 \quad 1\le m<\infty \quad (2)$$

Among them, m is the smoothing factor, and the m control mode is the degree of sharing among the classes. The bigger the m is, the more fuzzy clustering results are obtained. In general, in order to control the clustering results not to be too vague, the m value is set to 2. d_{ik} indicates the distance between the sample k to the i cluster center p_i can be represented by different types of model. Using Euclidean distance representation in this paper:

$$d_{ik}=\left[\sum_{j=1}^{s}\left|x_{ij}-x_{kj}\right|^2\right]^{\frac{1}{2}} \quad (3)$$

FCM algorithm updates the membership degree μ_{ik} and clustering center p_i. When the iterative convergence is obtained, the membership degree and the clustering center can be used to classify the data sets and to determine the relationship between the data objects and the classification. Through the iterative process in the control stop valve ε and the number of iterations of the b, we can solve the following formula:

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

$$p_i=\frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \quad (5)$$

Residential users, large users of smart meter installed with PLC and wireless communication technology which is used to transmit the data of user with a certain frequency information acquisition system. We select four kinds of data objects as clustering analysis:

(1) x_{i1} power consumption: Daily total power;
(2) x_{i2} load rate: Average load / maximum load;
(3) x_{i3} peak electric coefficient: Peak time consumption/daily total power;
(4) x_{i4} valley electric coefficient: Valley power consumption/daily total power.

The acquisition frequency of the smart meter is a point every 15 minutes. So the daily collection is 96, and total daily electricity is the sum of the 96 data. The average load is /96. Total electricity consumption in peak time and valley power consumption are total electricity consumption in peak and valley time. So the x_k of each sample is a four-dimensional vector. FCM algorithm

based on the use of electrical behavior analysis process as shown in Figure2:

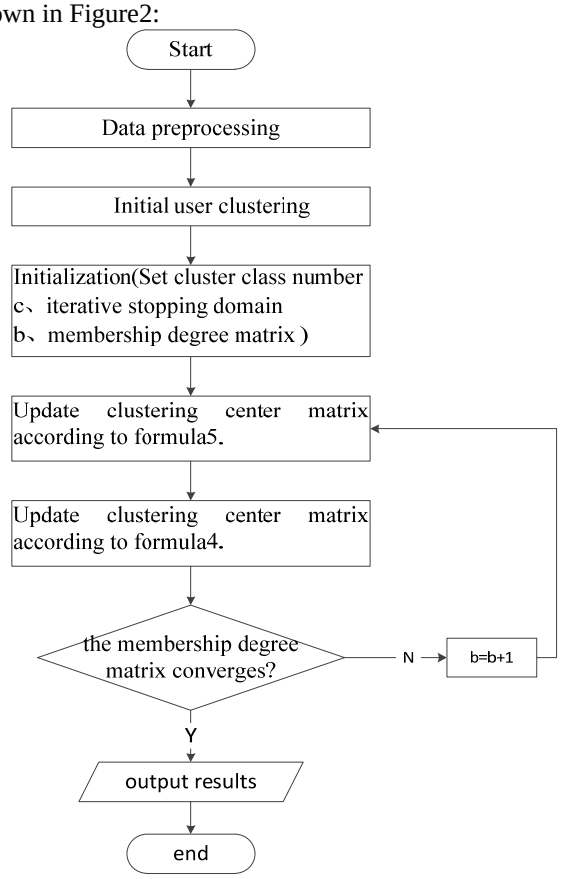

Figure 2. The analysis algorithm of electric behavior based on FCM

2.1.2 Hybrid computing method of FCM

The batch computing layer of hybrid computing architectures proposed in this paper can batch parallel computing for large-scale data by using Hadoop platform. With the use of the distributed file system HDFS (Hadoop distributed file system) and calculated model of MapReduce, batch computing layer can be directly calculated the massive historical electricity data based on files and can get the electricity behavior characteristics.

In order to adapt to the MapReduce calculation model, we need to in parallel reform the electrical behavior analysis algorithm which were based on FCM algorithm. The FCM iteration is decomposed into two stages of Map and Reduce, on Map stage, a function can effects different data sets in the different date node. The Map output is a set of records in the form of <key, value> pairs stored on that node. After the end of map stage, the calculated model will be transmitted to the node which will undertake the work of Reduce, and dispose the key(like merge etc.) from the Map stage ,than output the final results in the form of<key, value>. Due to the Map and Reduce steps can run distributedly on multiple computers and highly abstract process of distributed computing, so the calculated model of MapReduce can analyze large-scale data (1TB) conveniently and effectively.

After studied the calculated model of MapReduce and combined with process of FCM algorithm, we find the similar computing that using formula(3) to calculate the distance of the sample to the current clustering center is the most frequent count. To clustering process of FCM which have n samples of the object in the classification of k, each iterations needs to do distance calculation about n*k time and each calculations needs to do variance calculation of s dimension's characteristics. According to the idea which can greatly improve the working efficiency of FCM, we put forward the FCM clustering algorithm based on MapReduce, the process is shown in Figure3.

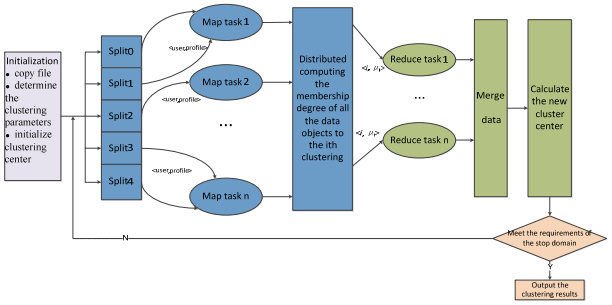

Figure 3. Parallel computation of FCM based on Mapreduce

(1) Copy the electric consumption data from relational databases (such as Oracle) to the distributed file system (HDFS), determine the clustering number C and stop domain ε according to the need of clustering;

(2) Based on the last clustering result, the initial clustering center can be determined and then transmit those data to the data nodes that participating in the distributed computing;

(3) Do some pretreatment to electric data and produce the key-value pairs of < user, profile >, in which the user is the user's unique identifier and profile contains the characteristics of the data object xi1 ~ xi4;

(4) All the key-value pairs of < user profile > is divided into several data subset, and transfer to the Map function. The Map function do membership degree calculation according to the formula (4) and store the result in the intermediate key-value pair of <i, μi> in which i is clustering number, and μi is membership degree about all objects in data subsets to the ith clustering;

(5) Transmit the calculation result of the Map function to Reduce nodes, Reduce task will merge the intermediate key-value according to the clustering number, then obtain a new clustering center by calculating formula (5);

(6) Repeat step (2) ~ (6), until the matrix of membership degree satisfies the stopping domain condition, the distributed FCM algorithm ends, export clustering results including the clustering number, the clustering center and the final membership degree of each clustering.

The algorithm is suitable for clustering analysis about the hybrid architecture batch calculation layer aim at massive historical electric data, and can get initial user

under the laws of the electricity. And real-time on-line analysis application of the electrical behavior require the real-time extraction of online power user's data from the electric energy data acquisition system and the real-time clustering of streaming data in the obtained initial user's clustering, but the batch computing layer's high delay characteristics is too difficult to meet the above requirements. Therefore, the real-time computing layer is needed to complete fast analysis and calculation of the flow data.

In batch calculation layer, we use the Storm framework to complete real-time clustering of electrical information. Different from Hadoop and other batch processing system, Storm is a distributed real-time computing system and it focused on the processing of streaming data and is mainly used in real-time analysis and continuous calculation. In the Storm framework, the logic of computing tasks is encapsulated into the Topology object, Topology is composed of Spout and Bolt. The processing logic of data is encapsulated in Bolt, and Bolt subscribes and receives streaming data from Spout, it also manages the data. The process of real-time computing layer in allusion to electricity behavior clustering is shown in Figure 4.

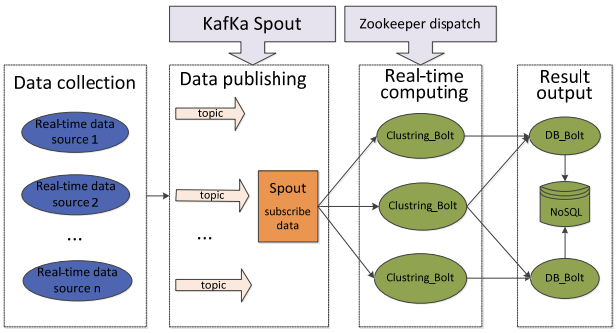

Figure 4. The online real-time clustering about electrical behavior based on Storm

The data interface of real-time electrical data obtained from the electrical information acquisition system can be pushed to the Kafka Topic through the producer servicing interface of Kafka, then be stored and wait for real-time access to the data. Spout subscribe Topic data and do distributed clustering computing on component of Clustering Bolt which has multiple computing nodes, under the dispatch of Zookeeper. The calculating results can be sustained by DB_Bolt or be released directly to clients by Topic Kafka. As has obtained the initial user's group, the real-time data can be calculated quickly in Clustering_Bolt.

Table 1. Clustering data object containing the initial user group

No.	Initial user group	x_{i1} power consumption	x_{i2} load rate	x_{i3}peak electric coefficient	x_{i4}valley electric coefficient
1	Vacant housing	0.5kWh	0.2 %	17%	16%
2	Office workers	150kWh	0.87 %	30%	10%
3	Residence for elderly	50kWh	0.64 %	15%	10%
4	Big family	200kWh	0.91 %	40%	15%
5	Commercial housing	400kWh	0.9 %	50%	8%
6	**Pending analysis data 1**	**230kWh**	**0.94 %**	**43%**	**18%**
7	**Pending analysis data 2**	**420kWh**	**0.82 %**	**57%**	**7%**

As shown in Table 1, when clustering the electrical behavior, it does not need to cluster all of historical data but need the characteristics of the initial users as the initial sample data of clustering. The number of rows is equal to the number of clusters c, incorporate the k pieces of electric data from the data to be analyzed into the data matrix, and the membership degree matrix $U_{c*(c+k)}$ is obtained after b iterations and finally converge. Among them, the data of the c+1 column to the c+k column is the membership degree of data to be analyzed to the initial users. Through the optimization of the distributed FCM algorithm, the size of sample data can be greatly reduced, so that the real-time requirements of the on-line clustering analysis can be met.

3 Experiment and analysis

In order to verify the FCM clustering algorithm proposed in the paper, the method of electrical behavior clustering analysis is implemented on the distributed computing architecture. In the laboratory environment we build up a distributed computing environment made up by five nodes, one is NameNode, others is DataNode. We collected 457 residents' electrical consumption data of a district and surrounding commercial tenant from Tianjin Eco-city, the data coverage from July 3rd, 2014 to October 28th, 2014. The sampling interval are 15 minutes, it means that each household take sample of 96 points data every day, we do research on the type of residential customers based on those.

After using the algorithm flow showed in Figure 3 to parallel clustering analyze the collected data of residential electrical data, eliminated the obvious unreasonable bad data which contains small amount of sample then we obtained four types of typical users, as shown in Figure 5.

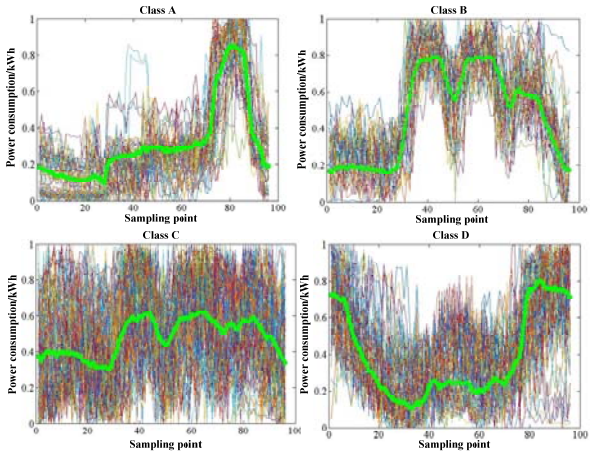

Figure 5. Clustering results of resident power consumption.

(1) Class A users' electrical consumption is big in the peak period of morning and night, the electrical consumption reached its peak in the evening and the power consumption is small at the rest of the time, so it's a typical office worker's. There are 274 users belong to class A.

(2) Class B users' electrical consumption is high during the day, the consumption began to fall after 19:00, it conform to the company characteristics which rent the office in the residential area. There are 84 users belong to class B.

(3) Class C users' electrical consumption is in a relatively average level throughout the day, the consumption in noon and evening is slightly higher than average level, it belongs to the business feature which operate all the days. There are 84 users belong to class B.

(4) Class D users' valley electric coefficient is higher than the peak power coefficient, and the valley power consumption is used in the high level, it should be the small production processing enterprises which produce in the valley time to avoid the high price in the peak time.

Acknowledgement

Fund Project. China State Grid Corp Fundamental Frontier Technology Project financing (Research on Power Big Data Hybird Processing and Complex Relation Analysis, Issue number SGRIJSKJ2015309), Starting and ending time:2015.1~2016.12.

References

[1] Cao junwei, Wan yuxi, Tu guoyu, etc. Information system Architecture for Smart Grids[J]. Chinese Journal of Computers, 2013(1): 144-167. (in Chinese)

[2] Hu xuehao. Smart Grid—A Development Trend of Future Power Grid[J]. Power System Technology, 2009(14): 1-5. (in Chinese)

[3] Yin qiang. Research on Sino-Singapore Tianjin Eco-City mode of operation[J]. Tianjin: Tianjin University of Technology, 2009. (in Chinese)

[4] Xie kai, Liu mingzhi, Yu jiancheng. Summary on smart grid integrated demonstration project of Sino-Singapore Tianjin Eco-City[J]. Journal of Electric Power Science and Technology, 2011(1): 43-47. (in Chinese)

[5] Li xiaoquan. Applications of Intelligent Power Equipment in Sino-Singapore Tianjin Eco-City Grid[D]. Baoding: North China Electric Power University, 2013. (in Chinese)

[6] Wang dewen, Song yaqi, Zhu yongli. Smart Grid information platform based on cloud computing[J]. Automation of Electric Power System, 2010(22): 7-12. (in Chinese)

[7] Song yaqi, Zhou guoliang, Zhu yongli. Present Status and Challenges of Big Data Processing in Smart Grid[J]. Power System Technology, 2013(4): 927-935. (in Chinese)

[8] Rusitschka S, Eger K, Gerdes C. Smart grid data cloud: a model for utilizing cloud computing in the smart grid domain[C]. Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference. Gaithersburg, MD: IEEE, 2010: 483-488.

[9] Liu keyan, Sheng wanxing, Zhang dongxia, etc. Big Data Application Requirements and Scenario Analysis in Smart Distribution Network[J]. Proceedings of the CSEE, 2015(2): 287-293. (in Chinese)

[10] JiaWei Han, Micheline K, Jian P. Data mining: Concepts and technique[M]. 3rd ed. New York: Elsevier, 2011.