

Transient power quality assessment based on big data analysis

Huang Zhiwei¹ Gao Tian² Zhang Huaying¹ Han Xu² Cao Junwei² Hu Ziheng¹ Yao Senjing¹ Zhu Zhengguo¹
(1. Shenzhen Power Supply Co., Ltd 2. Tsinghua University)

ABSTRACT

A transient power quality assessment method is proposed in this paper, using Naive Bayes classification method which is based on big data processing architecture. In this architecture, data sources will be extended to the aspects of power grid monitoring data, the power customer data and the public data, and the assessment severity will be classified into the normal state, the abnormal state, the critical state, and the failed state, according to the Naive Bayes classification results. Based on the data type of transient power quality assessment, big data processing architecture used in this paper can be able to process distributed data and streaming data, so that it can ensure not only updates classifier rules regularly, but also the real-time condition assessment. In the classifier training phase, we use the massive historical data as the distributed learning object, and generate assessment rules periodically. In the state assessment phase, each assessment node will update the assessment rules generated by training phase, generate real-time evaluation of samples from stream processing framework, and evaluate the power quality state according to the current rule. On this basis, this paper designs a Naive Bayes classification method based on MapReduce processing, and realizes the map and reduce process method to compute the priori probability and the conditional probability in distributed way. Experiments show that the transient power quality evaluation method based on the big data analysis presented in this paper is feasible, and achieve good results both in classification accuracy and processing speed.

Keywords: Big data, MapReduce, Distributed data mining, Naive bayes classification.

INTRODUCTION

The power quality disturbances mainly refers to the deviation of frequency, voltage, current and other indicators during normal electrical operation. For a long time, people are more concerned about steady problems of power quality, such as voltage deviation, frequency deviation and three phase unbalance^[1]. However, with the extensive use of sensitive equipment, the transient disturbance of

power quality has become the main reason for the users' experience problems, especially voltage sag and short interruption. Therefore, the purpose of comprehensive assessment is obtained an accurate description of the situation, so that both sides of power system can accurately assess the severity of the transient disturbance, so as to confirm the responsibility belonging and provide an important basis for power quality management^[2].

Information and communication technology (ICT) is increasingly important for the power industry, because ICT is widely used in entire procedure of power production and business management, with the result of explosive growth in data. The data which generated from the power transmission, distribution and consumption can be used for transient power quality assessment. In the background of big data technology, the data can be divided into massive historical data and real-time data stream, which characterized by multiple sources, massive amount, heterogeneity, and space-time continuum^[3]. Therefore, the big data framework should possess both distributed computing capabilities and stream processing capabilities, and promote the breadth and depth of data analysis in transient power quality assessment, resulting in more real-time and reliable conclusion.

CURRENT RESEARCH ON TRANSIENT POWER QUALITY ASSESSMENT

Currently, the transient power quality disturbance has become an important factor affecting power system security, stability and reliable operation. The purpose of transient power quality assessment is real-time calculation of the disturbances severity, so that the sensitive power users can be warned by the potential equipment failure. The traditional method is to integrate multiple indicators into a single quantitative indicator scientifically and objectively, which is mainly divided into qualitative analysis and quantitative analysis. Both of their methods use coefficients weighting to reflect the relative importance of each indicator, and a one-dimensional multi-dimensional power quality disturbances to merge. The main problem with this approach is that the various power quality indicators are not

significantly relevant, leading to their corresponding weighting factor is difficult to determine.

From the perspective of machine learning, classification algorithm is a supervised learning method, whose goal is to classify unknown samples into different classes^[4]. The algorithm creates a model (usually called classifier) by studying the features of the various classes to classify samples which are not labeled. Transient power quality assessment of this article uses a distributed classification algorithm, by learning relevant data to train a classification model, and then according to the classification model to identify and predict power quality situation. The following three types of data is valuable to the power quality comprehensive assessment. (1) Power grid monitoring data, such as current, voltage, frequency, etc. (2) Public information data, refers to the external information which affect the power quality, such as weather, latitude and longitude, geological disasters. (3) Power user data, such as users' sensitivity, importance, load capacity, complaints and other information. According to the results of power quality assessment, the severity condition of the power quality are classified into four grades, including normal state, abnormal state, critical state, and fault state, which are defined as follows.

(1) normal state

Stable condition of the power grid, no risk to the power users even the sensitive ones.

(2) abnormal state

One or several key features of power quality arise slight transient disturbance, which might generate risk for the sensitive power user.

(3) critical state

One or several key features of power quality arise serious transient disturbance, which might generate risk for the most power users.

failed state

One or several key features of power quality are worse than standard situation badly, which lead to equipment failure for most power users.

TRANSIENT POWER QUALITY ASSESSMENT BASED ON BIG DATA ANALYSIS FRAMEWORK

The data that is used for transient power quality assessment includes the massive historical data and real-time streaming data, which means the big data processing platform in this article should be able to process both of them. This paper realizes data conversion, data integration, data processing based on Hadoop, and proposes a hierarchical framework for the power quality assessment, as shown in Figure 1. After obtaining relevant data, HDFS and

MapReduce are used to manage and process both of the distributed data and streaming data. Then, transient power quality condition are classified into normal state, abnormal state, serious state, and faults state based on the Naive Bayes classification algorithm.

(1) Data source layer: it realize data acquisition across different business systems and different security zone, which is generated from multi-sensors, public information and power users. The data is divided into massive historical data and real-time streaming data which are processed separately and used for the classifier training and classification.

(2) Distributed File Management layer: using a master-slave structure which is comprised of the name node and the data node, it can provide high throughput in data access, and realize dynamical duplication adjustment, data blocks management and data transfer control, with high fault tolerance and high portability.

(3) Big data processing layers: the MapRduce framework is used for distributed computing. Job tracker nodes are responsible for the job scheduling. As to a specific task, the Map processing and Reduce processing are critical phases to parallel processing of big data.

(4) Data Mining layers: based on MapReduce computing framework, this paper use the batch data retrieval and real-time data retrieval to get required data, and then run a distributed data mining algorithms to achieve transient power quality classification.

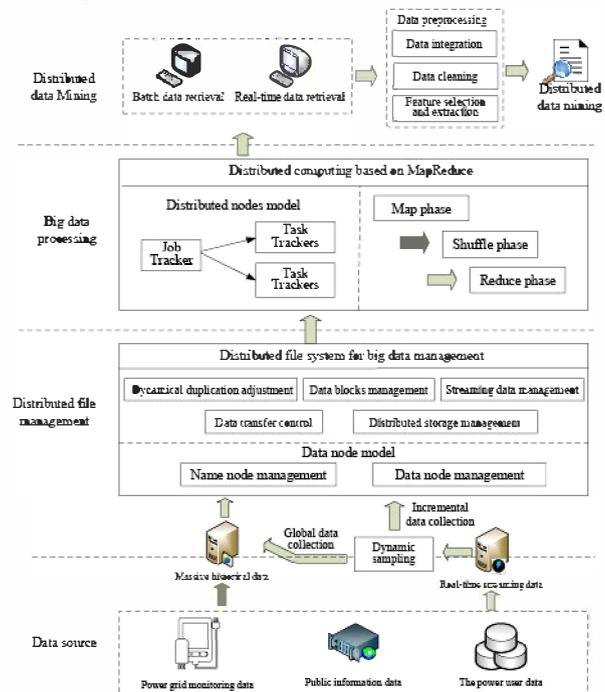


Fig 1. Big data processing framework for power quality assessment

POWER QUALITY ASSESSMENT BASED ON NAIVE BAYES CLASSIFICATION

Distributed computing based on MapReduce

In this paper, the Hadoop framework is used for comprehensive assessment, which means the MapReduce processing is used in data cleaning, feature extraction, Naive Bayes classifier training and power quality classification as a distributed computing model^[5]. The power quality condition will be divided into normal state, abnormal state, critical state and fault state. Using Hadoop distributed computing framework, the machine learning algorithms can be converted to MapReduce processing mode, which can realize parallel data processing to improve the performance of data mining algorithm. In general, with the growth of the training data and testing data, distributed classification algorithm can solve the classification problem of large-scale data sets by expanding the number of distributed processing nodes^[6]. When the scale of training example is relatively small, the stand-alone classification algorithm shows a better performance. However, with the increase of sample scale, the advantages of MapReduce on scalability and parallelism become apparent, and the distributed classification algorithm can efficiently be applied on big data. Transient power quality assessment based on big data analysis framework

The mathematic model of distributed Naive Bayes classification

Naive Bayes classification has a solid mathematical foundation with little sensitive to missing data, which can still get a relatively accurate classification result in the case of incomplete samples. The power grid monitoring data, public information and power user data are integrated as the training set for distributed classification. The big data is formalized as a vector with n dimensionality: $X = (x_1, x_2, \dots, x_n)$. The conditions of power quality such as normal state, abnormal state, critical state and fault state are formalized as C_i . Then, the Bayes theorem is shown as following formula.

$$(1) P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

The classification problem can convert into calculating the largest value of $P(C_i|X)$ and corresponding label C_i , as the following formula.

$$(2) f(X) = \max_{i=1,2,3,4} P(C_i|X)$$

Where, $P(C_i|X)$ is conditional probability of C_i when X occurs. $P(C_i)$ is the priori probability which can be calculated by training sample as the formula (3).

$$P(C_i) = \frac{S_{C_i}}{S} \quad (3)$$

Where, S_{C_i} represents the quantity of C_i , and S represents the total quantity of training set. The conditional probability $P(C_i|X)$ can be calculated by the formula (4), in where $P(x_j|C_i)$ represents the probability of x_j in class C_i .

$$P(X|C_i) = \prod_{j=1}^n P(x_j|C_i) \quad (4)$$

In the formula (3), the attributes of samples X should be discrete. As to the continuous attributes, they should be converted into discrete value to calculate $P(x_j|C_i)$. If the attributes are hard to be converted, such as the user complaints and user load capacity, this paper execute normalization processing firstly as formula (5).

$$f(x_k) = \frac{x_k - \min(x_i)}{\max(x_i) - \min(x_i)}, \quad i = 1, \dots, n \quad (5)$$

For the three types of data used in Classification model, power grid monitoring data is high sampling data, and the public information data and power user data show a lower sampling frequency. The data is divided into massive historical data and real-time streaming data, respectively used for classifier training and power quality classification. Massive historical data includes public information data and power user data, as well as the historical grid monitoring data, which is used as training and test sets in the distributed classification algorithm, in order to generate transient power quality classifier rules. The test set is used to assess the accuracy of classifiers. Then, the classifier will automatically deployed to all of the assessments nodes (usually is a power substation). In the assessment phase, the power grid monitoring data is sampled incrementally at regular time, combined with other data to generate the assessments samples. Based on the classifier at each assessment node, the classification algorithm is proceed to evaluate the severity of power quality, whose processing flow is shown in Figure 2.

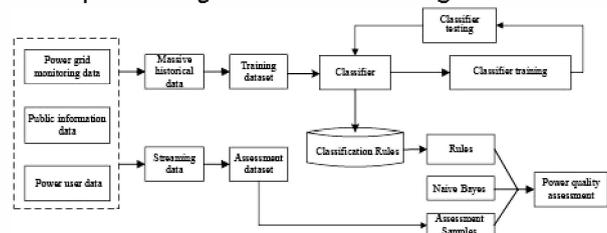


Fig 2. The processing flow of power quality assessment

Distributed classification algorithm

As the power quality comprehensive assessment is closely related to the monitoring device and regional users, the data storage and assessment requirement are great comfort to the parallel processing. The distributed data mining algorithm of this paper is based on the MapReduce computing models to

realize parallelization. The whole process of Naive Bayes algorithm, such as data cleansing, feature extraction, classifier training, are carried out based on the MapReduce model. The MapReduce model includes four entities: (1) Client, responsible for writing MapReduce program, tasks configuration and jobs submission; (2) JobTracker, responsible for initiating tasks, assign tasks to TaskTracker and coordinate the implement of the entire operation; (3) TaskTracker, responsible for the implement of Map or Reduce tasks on data fragment; (4) Hdfs: responsible for saving the task data, configuration information and the results of data. Based on the MapReduce computing framework, Naive Bayes classifier training algorithm is described as follows:

(1) The classifier training is divided into two MapReduce Job: Job1, Job2, corresponding to the calculation of prior probability and conditional probability, whose execution are controlled by two JobTracker. The input data is shown as $\langle Key, Value \rangle$ pattern, where Key is the sample number k , and $Value$ is the sample itself including relevant data features and class label, denoted by $[k, X]$.

(2) In the *Input Split* phase, MapReduce can calculate *Input Split* based on the input file. Input Split includes the position and the length of a data fragment, while the data physically stored as the data blocks in HDFS.

(3) Job1: calculating the prior probability. *Map* phase obtains data from *Input Split* phase, and calculates the sample number of classes. The output Key of *Map* phase is the class label. The *Shuffle* phase will sort the output Key for each Mapper, and guarantee the same Key of *Map* phase output is allocated to the same *Reducer* using the function $hash(key) \bmod 4$. In the *Reduce* phase, the algorithm will obtain all relevant data of each class, then calculate the total number of and.

(4) Job2: calculating the conditional probability. The processing phases of MapReduce is shown in Fig 3, which also presents the $\langle Key, Value \rangle$ of each phases. According to the formula (4), calculating should base on the result of each, which means the probability of attribute j equals on condition of the samples belonging to class. After *Map* phase obtains data from *Input Split* phase, each *Mapper* calculates the number of samples in class, whose attribute j equals. The output Key of *Map* phase is (j, C_j) . The *Shuffle* phase will guarantee the same Key of *Map* phase output is allocated to the same *Reducer* using the function $hash(Key) \bmod 4n$. In the *Reduce* phase, the algorithm will calculate the total number of samples in class, whose attribute j equals. Then, and can be calculated easily.

(5) Classification for the transient power quality. According to the formula (2), the algorithm calculates

for each sample to determine the classification result.

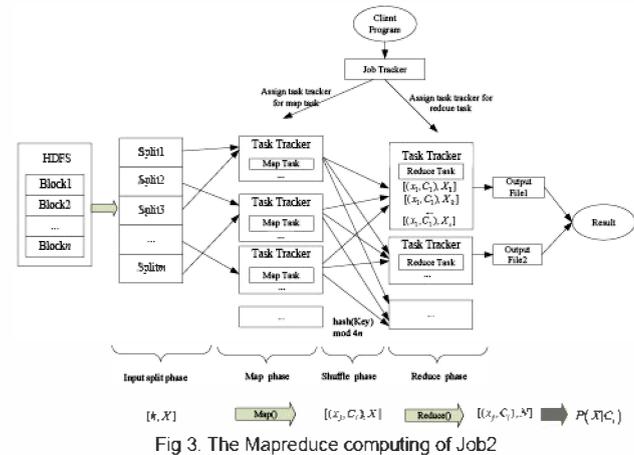


Fig 3. The Mapreduce computing of Job2

EXPERIMENT

This paper collects the monitoring data of three 10kV transmission lines during the years of 2010 to 2012, which are provide by a certain power grid of China. The data includes: the voltage sags, temporary rise, short interruptions and other monitoring data; the relevant user sensitivity, importance, load capacity; the humidity, temperature and weather information. In the simulation experiments, the above three base lines are simulated to distributed assessment nodes for classifier training and power quality assessment. Part of the training set and testing sets is shown in Table 1.

Table1. The structure of training/testing dataset

Distributed assessment nodes	Power grid monitoring data			User data			Public information data			Classification result of power quality
	Voltage sag depth (level 1-10)	Voltage sag last (level 1-10)	Current change (level 1-10)	Sensitivity (level 1-V)	Importance (level 1-V)	Load Capacity (level 1-V)	Temperatures (level 1-V)	Humidity (level 1-V)	Weather	
Transmission Line1	2	5	3	V	II	IV	III	IV	Hunderstorm	abnormal
Transmission Line2	3	9	7	II	IV	I	V	II	Snow	failed
Transmission Line3	8	2	7	I	IV	III	I	I	Rain	normal

According to the algorithm model and distributed data condition, this paper establishes a distributed computing environment upon five PCs, which includes a Master node and 4 Slave nodes. CentOS6.4 operating systems were installed on five nodes, and the Hadoop were setup and configured.

The master node configured NameNode and JobTracker role, which are responsible for distributed data management and job decomposition. Four Slave nodes configured DataNode and TaskTracker role, which are responsible for distributed data storage and algorithms proceeding. In the experiment, the Job1 and Job2 will be executed orderly, and will be calculated on each Slave nodes to obtain the classification result. The experiments scheme is as follows: the data of 2010, 2011 and 2012 are respectively treated as the training data sets, while the data of 2013 is treated as test set. The classification accuracy of the classification algorithm are shown as Fig 4.

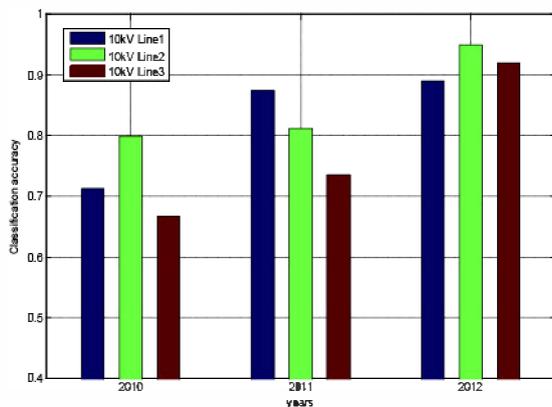


Fig 4. Power quality assessment result during the years of 2010-2012

According to the results shown in Figure 5, when the data of 2012 data is used as the classification training set, classification algorithms can achieve better accuracy, which prove that the distributed Naive Bayes algorithm can get the relatively precise result for power quality assessment. Meanwhile, the generation time of the training set greatly impact on classification accuracy, which indicates that periodic update of training set and classifier is necessary to improve the accuracy of the state assessment. This is because the power grid and user conditions are constantly changing, especially various types of power electronic devices (whether installed by power providers or users) are significantly affecting power quality.

CONCLUSION

With the rapid development of electricity industry information technology, the amount and source of data show explosive growth, which evolves to the big power data throughout all aspects of power transmission, power distribution and power management. In this paper, Bayes classification algorithm is based on the large data processing architecture. Heterogeneous big data is used to assess power quality, including the power grid monitoring data, the power user data, and public

data, to achieve the status of transient evaluation. As the result, the power quality is classified to normal state, abnormal state, critical state and fault state. Experimental results show that Naive Bayes classification algorithm based MapReduce processing can achieve better classification accuracy, thus helping to assess the power quality state effectively.

REFERENCES

- [1] Gnacinski, P., Mindykowski, J., Tarasiuk, T. A New Concept of the Power Quality Temperature Factor and Its Experimental Verification [J]. IEEE Transactions on Instrumentation and Measurement, 8(57), 2008: 1651-1660.
- [2] R. Dugan, M. McGrannaghan, S. Santoso, H. Beaty. Electrical. Power Systems Quality [M]. New York: McGraw-Hill, 2002.
- [3] S. Banerjee, N. Agarwal. Analyzing Collective Behavior from Blogs Using Swarm Intelligence [J]. Knowledge and Information Systems, 12(33), 2012: 523-547 .
- [4] R. Ahmed, G. Karypis. Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks [J]. Knowledge and Information Systems, 12(33), 2012: 603-630.
- [5] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering, 1(26), 2014: 97 - 107.
- [6] E.Y. Chang, H. Bai, K. Zhu. Parallel Algorithms for Mining Large-Scale Rich-Media Data [C]. Proc. 17th ACM Int'l Conf. Multimedia, 2009: 917-918.