Managing and Analyzing Big Error Data for Intelligent Electric Energy Meter Quality Monitoring

Guo Yu\textsuperscript{1,a}, Sike Wang\textsuperscript{2,b} and Shengqing Wei\textsuperscript{3,c}

\textsuperscript{1}State Grid HuBei Electric Power Research Institute Wuhan, China
\textsuperscript{2}State Grid HuBei Electric Power Company Measurement Center Wuhan, China
\textsuperscript{a}guoy_0701@sina.com, \textsuperscript{b}wongssko@sina.com, \textsuperscript{c}492980899@qq.com

Keywords: Intelligent Electric Energy Meter, Big data analysis, Hadoop Distributed File System, Luby Transform codes, HMM.

Abstract. With the installation and application of intelligent electric energy meters, the need for the veracity of meters from the users is increased thereupon. How to monitor and predict the quality of intelligent electric energy meters based on big error data acquired from intelligent electric energy meters is non-trivial. However, managing and analyzing large volumes of intelligent electric energy meter error data is a great challenge and is consequently hindering the effective utilization of the big dataset collected. In this paper, techniques for collecting, storing and analyzing large volumes of error data is presented to reduce the uncertainty of intelligent electric energy meters quality estimation. First, the error data of intelligent electric energy meters is sampled automatically and accurately by a designed device called Self-Service Error Calibrator Device. Then a storage system based on fountain codes is designed to store the error data on the HDFS platform. At last, an improved data mining technology named HMM (Hidden Markov Model) is leveraged to predict the variation performance tendency of intelligent electric energy meters varied with power load, manufacturers. Simulation results show that our proposed system can provide efficient, accurate, and cost-effective big error data management for intelligent electric energy meters quality monitoring.

Introduction

Intelligent electric energy meter is an advanced metering device based on modern computer technology, communication technology and measurement technology, which is the advanced measurement device for data acquisition, processing and management of electric energy information. With the rapid development of electricity information collection system, intelligent electric energy meters have been deployed widely.

As an emerging technology, the quality stability of intelligent electric energy meters have not been testified fully and the electricity consumers become more leery of the veracity of intelligent meter\textsuperscript{[1]}. The conventional way that monitoring the intelligent electric energy meters quality artificially and afterward remediably lacks of alarm correlation analysis\textsuperscript{[2]}. Thus it can not provide alarm and fault information required for fault location. In addition, the way that submitting the running broken meters in a bottom-up model lacks continuous tracking and monitoring of intelligent electric energy meters quality. It is unable for the users to locate and early warn the faults in time. Moreover, as an important mean to ensure and improve the quality of intelligent electric energy meters, the way of supervising and evaluating the intelligent electric energy meter suppliers is lack of objective and reliable evaluation basis. Therefore, it is extremely crucial and urgent to develop a system that can provide alarm correlation analysis, traceability, early warnings and comprehensive evaluation and decision aids to the quality monitoring of the intelligent electric energy meters.

As a new scientific trend, big data technology has penetrated into every industry and business functions fields currently\textsuperscript{[3-4]}. Driven by data analysis, big data technology works out data correlations to gain insight to the inherent mechanisms. Data in power systems have increased dramatically, leaving gaps and challenges; data processing is a major concern and its urgency increases with data growth. The 4Vs data (data with features of volume, variety, velocity, and veracity)\textsuperscript{[5]} in smart grids, which can hardly be handled within a tolerable elapsed time or hardware resources by traditional
model-based tools, have encouraged the development of an emerging paradigm—big data technology for power systems\cite{6-7}. Actually, big data technology has already been successfully applied as a powerful data-driven tool for numerous phenomena, such as quantum systems\cite{8}, financial systems\cite{9-10}, biological systems\cite{11}, as well as wireless communication networks\cite{12-14}. Major tasks of the architecture for these applications seem similar: 1) big data collecting; 2) big data storing, and 3) big data analyzing. It is believed that big data technology will also have a wide applied scope in power systems, and the results will be fruitful.

In this paper, a scheme of quality monitoring for intelligent electric energy meters is proposed. First, the error data of intelligent electric energy meters is sampled automatically and accurately by a designed device called Self-Service Error Calibrator Device. Then a storage system based on fountain codes is designed to store the error data on the HDFS (Hadoop Distributed File System) platform. The system mentioned above adopts a family of fountain codes known as Luby Transform (LT) codes, rather than the holistic approaches (i.e., whole-sector replication) to the simultaneous provision of adequate levels of data availability, integrity, and confidentiality. At last, an improved data mining technology named HMM (Hidden Markov Model) is leveraged to predict the variation performance tendency of intelligent electric energy meter varied with power load, manufacturers.

System Design

The block diagram of the big error data collecting, storing and analyzing system shown in Fig.1 includes the error calibration device of self-service electric energy meter, big data HDFS storage system of the error data based on the fountain codes and hidden Markov data mining.

![Block diagram of the big error data collecting, storing and analyzing system](image)

Intelligent Energy Meter Error Data Collecting

An error calibration device for self-service electric energy meter is designed in Fig.2 to sample the error data with the purpose of timely assessing the influence of the fault meter and tracing back the warning. The whole device includes the following modules: PC, program controlled single-phased power source, automatic stitch connection and removal unit of the standard single-phased meter, single-phase energy meter to be calibrated, error calculator, human-computer interaction touch screen and data processing/printing part. The client can place the intelligent energy meter into the gauge stand according to the prompts, and the transparent mechanical enclosure shut down automatically, then the stitch connection and removal unit works to connect the voltage/current circuit and the impulse terminal. Next, the client can select the load current to start the error calibration through the human-computer interaction touch screen, which controls the power source by the detection software in PC to output the voltage and current with adjustable amplitude and phase for the standard and calibrated meters. The impulses of the two meters under different loads are given to the error calculator and the error value of the meter to be calibrated can be calculated. Other potential experiments can also be selected to conduct. After all the selected trials finish, the result is shown on the touch screen and reminds whether to print. Finally, reminding the client to take away the calibrated energy meter.
Data Storage

HDFS which can run on the cheap commercial servers is the basis of the data storage and management under the cloud computing Hadoop platform. It meets the need of streaming data availability and super file dealing due to its features of high fault-tolerant capability, high reliability and high throughput rate\(^{[15]}\).

The way of data storage and collection in the Hadoop platform can directly influence the availability of the electric data, which is mainly accessed by the data durability in the network. The traditional solution of improving the data durability in the HDFS is to deploy a certain number of redundant nodes\(^{[16]}\). The perception data in the network is copied and stored at other nodes. If arbitrary nodes break down, the data can still be acquired. But there are problems with low data recovery rate and high complexity, in addition, when every file needs to be copied 3 times with only 33% space utilization. On the other hand, clients will lose the control of the data if the mass energy meter data are managed via the cloud server provided by the third-party resource platform\(^{[17]}\). Despite the cloud service providers may supply more reliable data backup and recovery, it can’t be ensured that the data kept in the cloud are leaked, hidden, lost, even more, suffer malicious attacks by the providers on account of their benefit or other interferences. Therefore, an effective and reliable electric data storage system is required to improve the storage space utilization, guarantee the benefit of the client and ensure the reality and completeness of the data.

On account of above reasons, an error data storage system for the intelligent energy meter based on the fountain codes is proposed, which includes HDFS client end, Hadoop big data storage platform. In the service of users, the error calibration device samples the error index of the intelligent energy meter; Then the samples are preprocessed by the HDFS client end and encoded by the fountain codes application server\(^{[18]}\); Finally, the encoded packages are uploaded to the Hadoop.

In fig.3, the error data sampled from the device are sent to the HDFS client end. With the sparse character of the data, the redundancy information is removed by the sparse processing. Then the sparse error data are encoded by the fountain codes. The encoded packages are managed by the HDFS server to ensure the safe and effective storage and read-write of the data.

The write workflow of the HDFS distributed file system based on the fountain codes in fig.4 is:

1. The TCP/IP connection is set up after the identification verification on the client end, which then connects to the NameNode and starts the RPC remote request via a configurable port.
2. The NameNode checks whether the file to be established exists and whether the creator has the permission to operate. If success, a record is created for the file, otherwise, throwing exception.
3. The error data of the intelligent energy meter is read into the client end. After the sparse representation of the read data, the files are equally divided into packets by the proxy server on the
HDFS end. Then the packets are encoded to be a little more than the source message in the form of data queue, and the encoding matrix information are stored in the server.

![Figure 3. The architecture of the error calibrator device.](image)

(4) The data begins to write into the nodes according to the data address provided by the NameNode. The data package firstly saves in the first DataNode in form of stream, after finishing, the second DataNode begins to write. This process continues until the last DataNode is written, then produces a feedback information, which indicates the accomplishment of the error data writing.

![Figure 4. The flow diagram of writing error data.](image)

As depicted in Fig.5, the read workflow of the system is as follows:

(1) The client server firstly sends a read data request to the NameNode, then the NameNode provides the address of the data package in the queue.
(2) According to the address information, the client end reads the data from relative DataNode. Combined with the received encoded package and the encoding matrix information, the source information package can be restored by the OFG decoding algorithm [19].

(3) After successful decoding, the client will be notified and cut off the connection with the DataNode.

(4) Compared with existing techniques, the proposed system has many advantages. Firstly, the redundancy information has been removed and the storage efficiency has also been raised via the sparse process on the error data. The codec process by the fountain codes instead of backup storage promotes the storage space utility. Simultaneously, the low complexity and high decoding success rate of fountain codes solves the data loss problem caused by the ineffective DataNode.

(5) Finally, with the encoding matrix stored at the HDFS end, unauthorized users can’t decode despite they can get the encoding package, which ensures the privacy of the data.

Data Mining

Based on the mass data from sampled energy meter error data and the information about the relative meter manufacturer and the users’ community, the statistic model can be set up. Moreover, with the analysis theory of big data, the time changing trends of the meters’ error property, the relativity of the electrical load and the error and the meter’s quality can all be obtained.

A. Error Data Organization

In the database, the error data is stored and connected via four datasheets: user information table, intelligent energy meter information table, error data table, other data table. The E-R diagram (Entity Relationship diagram) among the four tables is shown in fig.6.

The relationship between the user and the meters usually is 1:1 or 1:N after installation. User information mainly includes user ID, user name, installed meter ID, installed meter barcode, user address and telephone number, among which user ID is the key information. The intelligent meter information consists of meter ID, meter barcode, manufacturer, product type, nominal voltage, nominal current, switch type, production time, installation time, the key of which is meter ID. It can relate the user information and the meter information. The error data directly connects to the user in 1:1 or N:1 via self-service calibration, and also has indirect relationship with the meter information. It is made up of meter ID, user ID, calibration time, error proneness, error value, which can be used to find the relationship between the user information, meter information and error information. The error data can also relates to other data as 1:N in time domain, for example, electrical load, which can discover the time changing trends of the meters’ error property, the relativity of the electrical load and the error and the meter’s quality can all obtain.
B. HMM for Error Data

In order to find the relationship between the tables, HMM (Hidden Markov Model) is applied on the error data. It consists of two parts. One is Markov process of error data to describe the state transition. The other is random process of the observed value to show the correspondence between the data state and the observed value. The process exhibits in fig.7.

![Diagram of HMM for Error Data](image)

Figure 7. The flow diagram of reading error data.

The parameter set of HMM can be described as \( \lambda = (N, M, A, B, \pi) \). Here, \( N \) is the number of hidden states, \( M \) is the number of observed values, \( A = \{a_q\} \) is the probability distribution matrix of the state transition, \( B = \{b_q\} \) is the probability distribution matrix of observed values, \( \pi = \{\pi_1, \pi_2, \ldots, \pi_t, \ldots, \pi_T\} \) is the initial distribution of the error data. The parameter set is usually simplified as \( \lambda = (A, B, \pi) \).

The application of HMM mainly solves three basic problems, in this paper, we takes the two and relative solution. One is adjusting \( \lambda = (A, B, \pi) \) to make the probability of observed values \( P = (O \mid \lambda) \) be the biggest, which is a parsing problem to make the parameters represent the most ideal observed error data. The other is for given observed value sequence \( O = \{q_1, q_2, \ldots, q_t, \ldots, q_T\} \) and \( \lambda = (A, B, \pi) \), to select an optimal state sequence \( \{q_1, q_2, \ldots, q_t, \ldots, q_T\} \), which is the prediction problem to indicate the hidden information and calculate the possible path created by the model to generate the observed value sequence.

Simulations Results and Discussion

In this section, we perform discussion and simulations to verify the improvement the proposed scheme can obtain. The performance of the proposed quality monitoring scheme will be assessed from three aspects: Data confidentiality, Storage efficiency, Capacity of predicting the variation performance tendency of intelligent electric energy meters varied with power load, and manufacturers.

A. Data Confidentiality

Storing the mass energy meter data on third-party resources platform potentially exposes them to unauthorized accesses, while preventing the owner of those data to enforce specific access policies. Thus, it is necessary to ensure the confidentiality of data, which entails to make sure that they are not disclosed to unauthorized users. In the following, we discuss how we obtain confidentiality in face of a threat model in which an attacker is able to break only storage nodes while we assume that HDFS clients are trusted entities that cannot be compromised.

In this paper, meter data confidentiality is achieved by disclosing only coded meter information, whereas the coding key (i.e., the seed of the random generator used for random rateless encoding) is safely stored by the HDFS clients only. This implies that the coding matrix for each DataNode is private information of the owner of the data. Therefore, by keeping the coding key secret, and by limiting the number of fragments stored on each node, the proposed storage system is able to resist to different types of attacks to confidentiality.

On the other hand, if an attacker collects enough coded meter information to decode the DataNode, security is obtained by making information reconstruction computationally hard, since an attacker cannot access the coding key as long as it is kept secret. It follows that our scheme provides the so called computational security relying on the external and safe storage of the coding key. Higher security levels could be achieved by encrypting the coding key.

Thus, confidentiality is obtained by keeping the coding matrix secret or making meter data reconstruction computationally when HDFS clients are trusted and cannot be compromised.
B. Storage Efficiency

The conventional way to provide data availability is through whole-sector replication. It provides much lower availability levels than coding-based solutions at a much higher storage cost. In this paper, a storage system based on rateless codes is proposed to improve the storage efficiency.

To show the improved storage efficiency performance of the proposed storage scheme, we carry out a set of experiments in which we compare its performance against those attained by the whole-sector replication system. Fig. 8 shows the storage efficiency performance comparison over different values of packet length. The improved performance of the storage efficiency of our proposed storage scheme over the conventional storage system based on replication can be observed in Fig. 8. For instance, when every DataNode needs to be copied 3 times, the storage efficiency is only 33%, however the storage efficiency of our proposed scheme is almost 90% when the packet size is equal to 1024 bits. In the other word, only 1045 coded bits are needed to successfully decode the source meter data for our proposed scheme, while the conventional storage system need 3072 bits to ensure the availability of the source meter data.

![Figure 8. The storage efficiency performance comparison for different packet sizes.](image)

C. Capacity of Predicting the Variation Performance Tendency

In this subsection, we carry out a set of experiments to predict the variation performance tendency of the intelligent meter. The average self-checking variation data of a number of intelligent electric energy meters sampled from a community are organized through the E-R diagram first. Then the big data mining technology based on HMM is applied to predict the variations of the intelligent meter.

First, we evaluate the variation performance tendency of the intelligent meter over different inspection times and manufacturers. In each inspection time, the average variation data are calculated from 1000 intelligent electric energy meters. Note that when the variation $\sigma$ is less than 0.01, we say that the variation can be negligible, and $0.01 \leq \sigma \leq 0.2$ represents that this group of meters has little error. When $\sigma$ is larger than 0.2, it means that this group of meters should be calibrated.

![Figure 9. The variation tendency of the intelligent meter over different manufacturers and inspection times.](image)
As shown in Fig.9, there are 6 manufacturers contributing to the average variation data of meters. On account of the historic variation data, the variation classification modeling based on HMM is applied to model the tendency of the variation performance. The solid line and dashed line in Fig.9 represent the historic data and prediction data, respectively. It can be obtained that the forecast data is predicted in conjunction with the historic data and thus suitable levels of reliability of the prediction data can be achieved from the HMM big data mining technology.

Fig.10 shows the relationship between the variation classification and the average daily power consumption. The variation data are selected from one of the manufacturers and classified based on HMM model. As depicted in Fig.10 that the variation of the group meters can be negligible when average daily power consumption is less than 20 KWH, however the average daily power consumption which is larger than 50 KWH can lead to little error and thus should inform the power consumer that the intelligent electric energy meters should be calibrated. Moreover, it can also be observed that large average daily power consumption gives a positive error and small average daily power consumption gives a negative error. This observation from Fig.10 agrees with the practical situation as larger average daily power consumption rotates the intelligent electric energy meters faster and the variation becomes non-ignorable.

![Figure 10. The variation tendency of the intelligent meter over different average daily power consumption.](image)

**Summary**

In this paper, a scheme of quality monitoring and predicting for intelligent electric energy meters is proposed. To reduce the uncertainty of intelligent electric meters quality and provide the power consumer with the right to know the variation performance of the intelligent electric energy meters, techniques for collecting, storing and analyzing big error data are proposed. First, device called Self-Service Error Calibrator is designed to collect the error data of the intelligent electric energy meters automatically and accurately. Then HDFS storage platform based on fountain codes is proposed to improve the storage efficiency and provide adequate levels of data availability. Due to the low coding complexity of fountain codes, cost-effective big data storage and integrity can also be obtained. Moreover, suitable levels confidentiality can be achieved by keeping secret the coding matrix. At last, we propose an improved data mining technology named HMM to predict the variation performance tendency of intelligent electric energy meters varied with power load, manufacturers. We conduct a set of experiments to confirm the efficient and effective of our proposed method in big data management for intelligent electric energy meters quality monitoring.

**References**


