Research on Fault Diagnosis of Smart Meters Based on Machine Learning

Changbao Huang*, Youxiang Wang, Ying Liu, Chao Wang and Guanhe Wang
China Power Telecommunication (Beijing) Technology Co., Ltd
*Corresponding author’s e-mail: zhongneng@power-peak.com.cn

Abstract: To strengthen the real-time monitoring and management of on-site meters and to detect hidden problems of meters for timely correction, this paper proposed a fault diagnosis method of smart meters based on the decision trees of machine learning. By establishing a fault analysis model for smart meters, this method analyzed the laws of smart meter failures. It extracted typical characteristics from the massive data produced by the operating smart meters and studied the correlation among these characteristics as well as weight coefficients to obtain the probability of smart meter failures. Furthermore, the model was constantly optimized according to the feedback of on-site meter troubleshooting from operation and maintenance personnel, which effectively enhanced the timeliness and accuracy of fault detection and correction and reduced the difficulty of operation and maintenance.

1. Introduction
As the smart grid construction of State Grid Corporation of China progresses, the data application scope of the power utility information collection system is expanding, the function application is increasing, the business support capability is improving, and the importance of the system is rising. In this context, power marketers urgently need to rely on information technology to provide users with accurate services[1]. The new situation has made more challenging requirements on the reliability of the power utility information collection system and the efficiency of operation and maintenance. It has become a pressing need for the system to support intelligent collection, operation, and maintenance by detecting various faults in time, discovering meter failures in advance, and correcting relevant problems as early as possible based on the above-mentioned massive data[2].

It has become the focus and difficulty of current research to use big data analysis and machine learning technology to optimize the diagnosis model with better accuracy and timeliness in fault analysis for intelligent collection, operation, and maintenance[3].

2. Domestic and Foreign Research
Abroad, power meter fault diagnosis started relatively early. In the 1960s, many intelligent analysis methods for power grids based on Petri nets emerged in Germany. Since the 1980s, AI technologies have become increasingly mature and widely used in power grids, with the integrated applications of various methods, such as neural network, fuzzy theory, Bayes, and support vector machine, in power grid fault diagnosis[4,5].

Comparatively, the fault analysis technology of power meters appeared late in China. At present, the most commonly used method is the diagnosis and analysis model of power meter faults based on the master station of the power utility information collection system. The model is mainly the rules set by experts for determining different faults based on experience and business knowledge, and the basic
A way for determination is the threshold method. Although this method can detect power meter faults to some extent, the field verification has proved its low hit rate. As the threshold value in this model is a range, setting an optimal value necessitates a lot of human resources, which means that the efficiency is not well guaranteed. In recent years, with the promotion of "big data", "artificial intelligence", and "machine learning" in China, some methods of machine learning such as neural network and wavelet analysis have emerged in the power grid. However, there is no mature, efficient, and widely applicable method for power meter fault analysis in current research[6].

3. Introduction of Algorithms

Smart meter fault analysis is a matter of classification. That is, the output is either faulty or non-faulty. Since many fields in power meter fault analysis are nonlinear, the most suitable model here is GBDT. Compared with traditional GBDT, XGBoost eases its complexity later by pruning so that the final learning model is not prone to over-fitting. Besides, it can automatically execute in parallel using multi-threads with appropriate hardware resources and enhance the prediction accuracy through algorithm optimization.

XGBoost is an improved version of the boosting algorithm based on GBDT (an iterative decision tree algorithm) with better preciseness and faster computation speed as it can automatically utilize CPU multi-threads for parallelism.

The regular term in XGBoost is used inside the cost function to control the algorithm complexity. The regularization term contains a regularization term that controls the weight of model complexity, and the higher the weight, the less likely the model is to be over-fitted.

XGBoost supports parallelism and uses multi-threads to speed up tree construction. However, XGBoost parallelism does not mean multiple trees being constructed at the same time, and the iterations also have to finish one after another. The parallelism of XGBoost is reflected in the characteristic particle size. In decision tree learning, ranking feature values is one of the most time-consuming steps (since the best division point has to be determined). Before training the model, XGBoost sorts the data and saves it as a block structure that is used repeatedly in later work, thus greatly reducing the computational amount. The node splitting is done by opening multi-threads, calculating the gain of each feature, and selecting the one with the largest gain for practice – a concrete realization of XGBoost's support for parallelism.

4. A Model Case

This study uses the power utility data of a region and performs fault analysis and prediction for the corresponding smart meters. The specific processes are data preprocessing, feature engineering, algorithm modeling, model evaluation, and model optimization (see Figure 1).
4.1. Data Preprocessing
This is mainly to assess the quality of imported data and to repair or complete the abnormal/missing data, which is necessary for data analysis to be carried out accurately. Three specific phases are involved:

Data Quality Assessment: to analyze the quality of imported data, mainly over whether any data is abnormal (burst increment, burst decrement, wrong time scale, etc.) or missing (determinant attributes), or whether there is any inconsistency among the data, for overall knowledge of data quality.

Relevant Algorithm Design: to propose data rectification measures and design relevant data cleaning algorithms for each problem identified during the data quality assessment phase. There are two major types of conventional algorithms: first, data repair algorithms based on data trend prediction and power business rules; second, specific selection based on factors such as accuracy requirements of business application scenarios and data processing efficiency.

Data Cleaning Execution: to apply data cleaning algorithms and carry out specific operations to solve the problems of data anomalies, missing, and inconsistencies based on the results of data quality assessment.

4.2. Data Engineering
Feature terms refer to the feature data loaded by the model. Appropriate feature terms can reduce the feature dimension of the model, speed up the computation, lessen the influence of irrelevant features on the classification effect, and improve the accuracy of the analysis results. Various methods are used to select the features suitable for the model. The flow of feature engineering is shown in Figure 2.

4.2.1. Data analysis. This is to read the training data, take a small number of samples for observation, select the data involved according to the definitions of all anomalies, and conduct mathematical calculations on the data. When analyzing the distribution of features in each line, view the distribution of features in different categories for different data through a histogram or label column.

4.2.2. Data conversion. There are diverse data types in the source data, and in the XGBoost model, the input features can only be numeric. Therefore, non-numeric data requires initial conversion. Different data types correspond to different conversion methods, such as data dimensionality reduction by clustering, data conversion by category mapping, etc.

4.2.3. Weight giving. The feature terms are given corresponding weights through the entropy weight method so that each feature for a failure has a specific weight. The higher the weight, the greater the influence of that feature term on the failure.
4.3. Algorithm Modeling
The modeling process is shown in Figure 3.

![Diagram showing the modeling process]

The data processed by the model is mainly derived from the metering data of the power utility information collection system; a preliminary analysis of the operating condition of each power meter type is conducted to check whether the meter is properly working. According to the fault analysis results, the data samples are labeled: if a fault occurs in the meter, it is marked as “1”, and otherwise, “0”. The labeled data is loaded into the XGBoost algorithm model as training data, and the model performs supervised learning, finding the differences between faults and non-faults and the common characteristics among users of faulty meters through machine learning.

When setting parameters, select the optimal ones according to the actual application scenario. For some parameters, their values can be set first and then modified on the basis of the model training effect. According to a smart meter fault, set the relevant parameters involved in XGBoost and improve the model prediction accuracy. The specific parameters involved that are relatively important for XGBoost modeling are listed below for elaboration:

- 'booster': 'gbtree' – select the model for each iteration as: tree-based model;
- 'objective': 'binary: logistic' – this parameter is to define the loss function that needs to be minimized;
- 'eval_metric': 'auc' – this parameter refers to the measurement of valid data and takes the area under the AUC curve;
- 'lambda': 50 – this parameter refers to the L2 regularization term of the weights, used to control the regularization part of XGBoost with a greater role in reducing over-fitting;
- 'eta': 0.1 – the robustness of the model can be improved by reducing the weight of each step.

4.4. Model Evaluation Indexes
The smart meter fault model based on machine learning is a matter of classification whose evaluation indexes mainly include AUC.

4.4.1. AUC. This model is a dichotomous one with the output of probability values from 0 to 1. It needs AUC evaluation indexes to determine the optimal threshold to classify the samples into positive and negative categories.

In a dichotomous model, a threshold is determined to split the output results into two classes for the obtained continuous results. If the threshold is reduced, faultier meters can be identified, increasing the proportion of identified faults to all faults, i.e., the true positive rate (TPR). However, it also treats more non-faults as faults, raising the proportion of non-faults misidentified to all non-faults identified, i.e., false positive rate (FPR). Take FPR as the x axis and TPR as the ROC curve bounded by the y axis in the two-dimensional coordinate system. The area below the ROC curve, or AUC, can be used as an index to evaluate the average performance of the model, and if the model is ideally perfect, AUC = 1. AUC is usually a probability value; the model performance is considered better if the value is closer to 1.
4.5. Model Optimization

After the model is built, it is optimized according to the effect of data training. The optimization measure is mainly to modify the parameters of the model and the feature terms of the data.

The main ways of model optimization by adjusting the model parameters are as follows.

1. Choose a higher learning rate.
2. For a given learning rate and number of decision trees, tuning is performed on specific parameters of the decision trees (max_depth, min_child_weight, gamma, subsample, colsample_bytree).
3. Tune the regularization parameters for XGBoost (lambda and alpha).

It is also a main method of model optimization to adjust the sample feature terms through the training results. According to their influence degree on the output results, the sample features are appropriately deleted, added, and converted.

4.6. Experimental Results

A region was selected as a pilot site for this model. During the 4 months of trial operation, the model identified 64 failures of the smart meters. According to the field verification, 54 failures predicted were real, and 8 failures were missed by the model. The accuracy rate and the recall rate of the model were 84% and 87%, respectively. Although the accuracy is yet to be improved, the above algorithm can realize the analysis and diagnosis of smart meter faults and enhance the detection exactness to a large extent. The model is continuously upgraded through supervised learning to accurately predict faulty smart meters and abnormal power users for fast on-site detection.

5. Conclusion

When a fault diagnosis model for smart meters based on machine learning is established, it can monitor and analyze smart meters in real time and provide more accurate troubleshooting information for operation and maintenance personnel to quickly locate and correct the failures[7]. It not only greatly improves the efficiency and accuracy of fault detection, but also saves considerable labor and material costs, as it can discover hidden problems of power meters and potential defaulted power utility in advance for an earlier appropriate solution to be taken. Furthermore, it may help enterprises raise the quality of power marketing services and guarantee normal power utility for customers, promoting the stable development of the entire power industry.

References
[1] Hideichi Kikuchi, Kazuhisa Murata, Fumio Takahashi. Automatic Meter Reading Systems for Bulk Customers[J]. Fuji Electric Journal, 1998(17).
[2] YUWEN Xiao-di. Research and Application of Power User Electric Energy Data Acquire System [D]. Baoding: North China Electric Power University, 2011.
[3] WANG Xiao-feng, LI Geng-qing. Power Utility Information Collection System Development New Trends [J]. Power Demand Side Management, 2010, 12(5): 59-61.
[4] Fan Jie, Chen Xiao, Zhou Yu. An intelligent analytical method of abnormal metering device based on power consumption information collection system [J]. Electrical Measurement & Instrumentation, 2013, 50 (11): 5-9.
[5] Song Siyang. Research on electricity consumption information collection system building in the electric power marketing[D]. Beijing: North China Electric Power University, 2011.
[6] HE Chun. Research on Access Control of Power Wireless Network [D]. Beijing: North China Electric Power University, 2012.
[7] XU Wei, WANG Bin, JIANG Yuan-jian. Power Line Carrier Communication Technology and its Applications in Electric Energy Data Acquisition System[J]. Electrical Measurement & Instrumentation, 2010, 47 (7A) : 44-47.