Feeder Fault Warning of Distribution Network Based on XGBoost

Jing Cai*, Yangyang Cai†, Hui Cai§, Shuilan Shi ‡, Yanting Lin$, Miaohong Xie$

1 State Grid Fujian Electric Power Co.,Ltd, Fuzhou, China
2 Fujian Yongfu Power Engineering Co.,Ltd, Fuzhou, China
3 China Construction Strait Construction Development Co., Ltd, Fuzhou, China
4 China Construction First Bureau Group Second Construction Co.,Ltd
5 Data Research Institute (Fujian) Information Industry Development Co., Ltd
*caijing@sgitg.sgcc.com.cn
*Corresponding author’s e-mail: ivyenglish131@126.com

Abstract. In this paper, the historical observation data of the power grid are used to build a predictive model of power outages for distribution network by using machine learning methods. By judging whether the distribution transformer network is about to fail, the maintenance and troubleshooting of the distribution network can be achieved in advance, thereby fundamentally reducing the occurrence of power fault of the distribution transformer. The data covers several dimensions such as distribution network loads, equipment ledgers, historical faults, weather and so on. The experiments show that the proposed method based on XGBoost is valid and efficiency for feeder fault early warning.

1. Introduction

More than 80% of feeder faults occur on the distribution side in the power system. For a long time, sorting out dangerous factors of distribution network blackouts and identifying hazard sources are mainly based on the experience of professional technical managers and operators, which are not comprehensive and objective, and lack effective methods and scientific tools for identifying the subtle signs of hidden blackout risks. It is difficult to identify risks and control them in time before the occurrence of a power outage. Part of the knowledge of hidden power hazard characteristics is not mastered in advance, and it cannot be eradicated after a power outage, resulting in frequent power outages.

In order to effectively solve the problem of repeated power outages, improve the reliability of power supply in distribution networks, and reduce customer complaints, it is urgent to dig out hidden factors that cause frequent power outages and study their mechanism of power outages. The active early warning application function provides a scientific basis for active early warning prevention and control of frequent power outages.

In recent years, many methods such as rough set [1,2], Bayesian network[3], have been proposed to solve the problem of power distribution network fault diagnosis, which are mainly grounded on the actions of protection devices or a particular electrical feature. With the development of massive power grid information systems, big data technics are available to mine and forecast such as support vector machine (SVM) [4] and neural networks [5,6]. However, it is hard for the methods mentioned above to
work out satisfactorily in dealing with power system faults. Another reason is that the decision rules obtained by these black boxes models cannot be understood from the perspective of human.

XGBoost[7] is one of scalable variety of the Gradient Boosting Machine (GBM), which has been proved an excellent tool among artificial intelligence methods as its easy parallelism and high prediction accuracy. In[8], a XGBoost-Based Algorithm is used to post-fault transient stability status prediction of power system, which indicates deep learning methods can be introduced to combine with power system to locate the fault of power system.

2. Model for feeder fault warning

2.1. Features of frequent power outages
The key to the technology is to build a set features that can comprehensively and accurately represents the status of the distribution network. The power outage of the actual distribution network is not only affected by the power supply facilities such as distribution equipment, overhead lines, poles and towers, but also multiple factors such as external force destruction and improper installation and maintenance, which affect the stability of the entire distribution network power supply. Therefore, the lack of any link in the process of early warning of frequent power outages in the distribution network will lead to deviations in the results.

Before the model is built, the technical reasons for historical faults, the causes of responsibility, the text description of historical faults in the distribution network, and the existing system data, the following major features were extracted: feeder characteristics, equipment account characteristics, equipment failure information, historical failure information, grid load data, and weather data. While the comprehensive features are considered, the GBDT (Gradient Boosting Decision Tree) method is used to re-select the features based on the characteristics of the features, which aims to extract the relevant features and the features with low correlation and negative impact on the accuracy of the model.

2.2. Construction of feeder fault prediction model
Based on a priori business understanding and data exploration, the factors that may be related to power outages are extracted to form the characteristics of the machine learning model. This model takes the feeder faults prediction as the research goal. By analyzing the causes of the feeder failure outage from January to June 2017 in the distribution network, the characteristics of weather, feeder operation period, feeder load, and historical number of feeder failures are extracted. Among them, the weather is the main feature that affects power outages, and this feature will be taken into account when data compression is performed. Internal data is extracted from various business systems of the power grid, and external weather data is obtained from the Meteorological Bureau.

Data processing is required, including outlier and missing value processing, normalization, and over/under sampling, then the model can be tuned. The algorithm adopted in this project is XGboost, and then perform a brute force search on the space of multiple hyperparameter combinations. Each set of hyperparameter combinations is substituted into the learning function as a new model and evaluated by the K-fold cross-validation method. The commonly used evaluation criteria are the AUC (Area Under Curve) in ROC (Receiver Operating Characteristic) curve. The vertical axis of the ROC curve is the TPR (True Positive Rate), which represents the proportion of the minority sample that is correctly predicted to all the minority samples, and the horizontal axis is the FPR (False Positive Rate), which represents the proportion of the majority of samples that are mispredicted to all the majority of samples. In the ROC curve, when TPR is higher than FPR, the area of the ROC curve will also be higher, indicating that the model works well. The whole flow for the proposed module is shown as Fig.1.
2.3. XGBoost
GBDT is a kind of boosting algorithm in integrated learning, which working mechanism is to train several classification decisions trees, each classification tree is a weak classifier, and the linear weighted combination of these weak classifiers constitutes a strong classifier.

Considering choosing an algorithm with regular terms, representative logistic regression was selected from the single classifier. While in XGBoost, regularization term is added based on the GBDT(Gradient Boosting Decision Tree).

The principle of the XGBoost algorithm is to split the original data set into multiple sub-data sets, randomly assign each sub-data set to the base classifier for prediction, and then sum the results of the weak classification according to a certain weight to predict the final result. By adding a regular term, XGBoost can effectively prevent overfitting, so it is suitable for this study. The target loss function of GBDT’s algorithm is

$$l(\phi) = \sum_{i=1}^{j} l(y_i^1, f_{m-1}(x_i)) + \beta T(x_i; \theta_m)$$

In the formula, $y_i^1$ represents the true value of the sample, $i$ is the sample number, $j$ is the total number of samples, $f_{m-1}$ represents the current classifier, $\beta$ is the weight, and $T(x_i; \theta_m)$ represents the current iteration. In the trained decision tree, $\theta_m$ represents its parameters, and $x_i$ represents samples. On this basis, XGBoost defines the structural complexity function $\Omega(f_k)$ by introducing the number of leaf nodes $T$ and the leaf score function $\omega$ of the decision tree, so that the target loss function is transformed into:

$$L(\phi) = \sum_i l(y_i^1, y_i^p) + \sum_k \Omega(f_k), \Omega(f_k) = \gamma T + \lambda ||\omega^2||$$

In the formula, $y_i^p$ represents the sample prediction value, $\gamma$ is the penalty term of L1 regularization, and $\lambda$ is the penalty term of L2 regularization.

2.4. Oversampling
An unbalanced dataset is one in which the target variable has more observations in one specific class than the others. The difference in the number of samples leads to the asymmetry of the information provided by the training algorithms of the different types of samples, resulting in the performance degradation of the machine learning model. The sample number of the minority data is relatively lacking, resulting in the influence of the minority data on the classifier is too low. In this project, the feeder that has a power fault during a specific day is a positive sample, and the rest are negative samples. The sample time window used is 16-17 years. As of 17 years, there are more than 16,000 feeders of Fujian
Electric Power. The cumulative feeder faults in 16-17 years are about 62,000, with an average of about 85 faults per day. The proportion of positive and negative samples ratio is about 1:188. Such a disparate positive and negative ratio is the difficulty of researching the blackout warning model. Normal positive sample random oversampling or negative sample random downsampling cannot achieve good results. The projects ultimately use positive sample data characteristics to simulate the distribution of real positive sample data, and use SMOTE (Synthetic Minority Oversampling Technique) algorithm for positive sample oversampling to effectively solve the problem of imbalance of classified samples.

3. Data Preparation

According to the indicator system for the impact assessment of power distribution outages, the quantifiable indicators corresponding to each criterion level were selected, including the amount of power lost by enterprises during power outages, the number of users affected, the importance of users, and the tendency of users to complain. First of all, we use historical data such as power outage duration and load to estimate the power loss of enterprises during power outages; on the other hand, the user complaint tendency has established an evaluation model, because the outage sensitive theme model of the marketing data foundation platform has an existing user complaint tendency Evaluation model, so extract the data of sensitive users of power outage as the data source of user complaint tendency, and construct a wide table of index data of power outage impact assessment model of distribution transformer.

Main characteristic data:
- Feeder characteristics, such as the number of historical faults, the number of faults in the same period, the number of historical alarms, the commissioning time, etc.
- Equipment selection, such as extracting the top-ranked equipment based on the statistical results of fault-related equipment and exploring the correlation between its characteristics and power outages. These equipment include towers, wires, cables, fittings, insulators, transformers, fuses, and lightning arresters.
- Equipment account characteristics, such as years of operation, manufacturer, model, rated current, rated resistance, length, diameter, etc.
- Weather data, temperature, humidity, wind, thunder, etc.
- Equipment defect information, such as the current undetected defects, the history has eliminated defects, etc.
- Feeder load information, such as operating power, load characteristics, etc.

4. Results and Analysis

4.1. Correlation analysis of consecutive fault-free days

The feature importance score is the highest in consecutive days without failure. Discretize it based on the fastest information gain method and analyze the difference in the failure rate between different values.

| Value interval | Fault rate |
|----------------|------------|
| 0–25           | 4.70%      |
| 25–50          | 1.96%      |
| 50–100         | 1.24%      |
| 100–600        | 0.67%      |
| 600–           | 0.12%      |

It can be seen from the table that the longer the continuous trouble-free days of the feeder, the better the overall stability of the feeder and the lower the probability of power outage. There are about 16,000 feeders in Fujian province, with an average of about 78 faults per day (based on data from year 2016 to 2017). The average daily fault rate is 0.5%. For feeders that have not been powered off for more than 600 days, the fault rate is about 0.12%, which is lower than the average failure rate.
4.2. Analysis of The Influence of Feeder Commissioning Period and Duration
Several features in the system are related to the commissioning period of the equipment, namely: transformer average commissioning period, cable average commissioning period, lightning arrester average commissioning period, pole tower average commissioning period and conductor average operation period. Taking feeder average commissioning years as an example, study the relationship between feeder commissioning years and power outages, and also discretize the value of transformer average commissioning years based on the maximum information gain.

Table 2. Commissioning period fault rate.

| Value interval | Fault rate |
|----------------|------------|
| 0~2.1          | 0.13%      |
| 2.1~3.7        | 0.39%      |
| 3.7~4.4        | 0.74%      |
| 4.4~10.1       | 1.11%      |
| 10.1~13.8      | 0.74%      |
| 13.8~          | 0.40%      |

It can be seen from the table that at the beginning, as the equipment age increases, the failure rate rises. After reaching a certain age, the failure rate begins to decline as the age increases. The reason may be that the quality of the equipment will gradually expose the problem when it is first put into use, and the equipment may be replaced accordingly. The longer the service life, the better the quality and the lower probability of failure.

4.3. Receiver Operating Characteristic (ROC)
The samples are sorted according to the prediction results of the learner, and the samples are predicted as positive examples one by one in this order, and each time the TPR and the FPR are calculated.
The area under the ROC curve becomes AUC. Generally, when AUC is greater than 0.7, it means that the model is valid. The value of AUC varies between 0 and 1, the larger the better. The AUC of the module we built is 0.8899, indicating that the model is valid. The relationship between features and whether there is a power outage is explored to a certain extent.

![AUC curve of the module](image)

The AUC score is 0.8899

5. Conclusion
The historical data of the Fujian province power grid between years 2016 and 2017 are used to build a predictive model of power outages for distribution network based on XGBoost. The data covers several
dimensions such as distribution network loads, equipment ledgers, historical faults, weather and so on. Correlation of consecutive fault-free days and feeder commissioning period rate are analyzed. The results of AUC which is 0.8899, indicates that the model is valid and efficient for feeder fault early warning.

References
[1] Yaotian, Z., Zhengyou, H., Jing, Z., Peng, Z., Ming, L., Jiantin G. (2007) A power network fault diagnosis method based on rough set theory and naive bayesian networks. Power System Technology, 31(1),37-43.
[2] Ran, L., Lieyong, Z., Xueping, Z. (2010) Distributed fault diagnosis of power networks applying the united rules mining algorithm based on rough set theory. Proceedings of the CSEE, vol. 30, no.4, pp. 28-34.
[3] Cai, B., Liu Y., and M. Xie. (2017) A Dynamic-Bayesian-Network-Based Fault Diagnosis Methodology Considering Transient and Intermittent Faults. IEEE Transactions on Automation Science and Engineering14(1), 276-285.
[4] Zhang, S., Wang, Y., Liu M., and Bao, Z. (2018) Data-Based Line Trip Fault Prediction in Power Systems Using LSTM Networks and SVM. IEEE Access 4, 7675-7686.
[5] Li, W., Monti, A. and Ponci F. (2014) Fault detection and classification in medium voltage DC shipboard power systems with wavelets and artificial neural networks. IEEE Trans. Instrum. Meas., 63(11),2651–2665.
[6] Bhattacharyya, S. (2015) Fault detection on a ring-main type power system network using artificial neural network and wavelet entropy method. Int. Conf. Comput., Commun. Autom., pp. 1032–1037 .India.
[7] Chen, T. and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 785–794.
[8] Chen, M., Liu, Q., Chen, S., Liu, Y., Zhang C., and Liu, R. (2019) XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. IEEE Access 7(1),13149-13158.