# Application of XGBoost in Identification of Power Quality Disturbance Sources of Steady-state Disturbance Events

Liu Yang
*State Grid Corporation of China*
*Beijing, 100031, P.R. China*
yang-liu@sgcc.com.com

Yi Li
*Global Energy Internet Research Institute Ltd.*
Beijing, China
80726576@qq.com

Cheng Di
*North China University of Technology*
*Beijing, China*
10237528@qq.com

*Abstract*--**In the era of IoT, time-series pattern recognition is not a trivial task considering their fast dynamically changing characteristic. The distribution and trend changes of these time-series data such as current or voltage could reflect emerging environment event. In response to the increasing power quality disturbances in the power grid, the XGBoost algorithm are used to identify multiple sources of power quality. Firstly, statistical methods are used to extract features from power quality disturbance sources. These features are computed through different statistic method with aims to reflect the time-series distribution. Secondly, a training data set is constructed and a XGBoost classifier is trained based on the generated training data sets. Furthermore, the prior knowledge of some interference sources is added on this basis, and then it is applied to power quality interference source identification. Experimental results show that this method can effectively identify power quality disturbance sources, and the proposed method has good robustness and noise immunity.**

***Key words: power quality, steady-state disturbance event, XGBoost, time-series classification***

## I. Introduction

In recent years, large-scale new energy and disturbance loads such as wind power, photovoltaic power generation, electrified railways and distributed power sources are build. At present, most provinces and cities in China have established power quality monitoring systems and collected a large number of power quality monitoring data. The analysis of power quality data in power grid is the premise of discovering power quality problems. How to obtain features and identify them effectively from massive power quality monitoring data is the primary problem.

At present, the research on power quality of researchers domestic and foreign mainly focus on single disturbance signal and composite disturbance signal, but only a few researches on generating power quality disturbance sources, which are still in the exploration stage [1].

Most existing power quality analysis consist of feature extraction and pattern recognition. Some Commonly used feature extraction methods include short-time Fourier transform [2, 3], wavelet transform [4], S transform [5-7], Hilbert-Huang transform [8-10], etc. In addition, some commonly used pattern recognition methods are decision trees [10, 11], neural networks [12], support vector machines [13] and so on. In literature [3], the power quality signal is extracted by short-time Fourier transform, and the two-class support vector machine is used to realize multi-label classification.

In literature [4], a power quality disturbance classification method based on wavelet transform and neural network is proposed. This method performs multi-resolution analysis on the power quality disturbance signal to obtain the signal feature quantity, and then recognizes the disturbance signal through the 3-layer BP network. Although the method has high recognition rate, it needs to choose the appropriate

978-1-7281-1190-2/19/$31.00©2019 IEEE

wavelet base for transformation, and it needs huge computation resources. Moreover, the compute result is easy to fall into the local optimal solution. In literature [5-7] , they use S transform and support vector machine for classification. S-transformation can extract feature frequencies effectively, but its time-frequency resolution is not the best for extracting features. Literature [8-10] used Hilbert-Huang to extract the disturbance signal, and the classification feature tree is constructed by using the obtained features to realize the classification and recognition of the disturbance signal. However, such methods require a large amount of computation, and the obtained features are usually only used for single or compound disturbance. The classification of the signal is not suitable for the identification of power quality disturbance sources.

The XGBoost algorithm is a classifier composed of multiple decision trees. Compared with other classification algorithms, XGBoost algorithm tolerates noise well and has good generalization performance. It has become an important data analysis tool. In the propose method, we apply this fast and accuracy classifier as our base model to recognize different disturbance source.

## II. Gradient Boosting

### *A. Ensemble Learning*

The ensemble learning method [14] uses multiple weak classifiers to solve the same problem, which can effectively improve the generalization performance of the learning system. The weak classifiers of the serial integrated learning algorithm are generated in order, which utilizes the dependency relationship between the weak classifiers. During the training process, different weight value will be assigned to the samples which have been wrongly recognized. Furthermore, each weak learner will be trained based on the weighted samples with aims to improve overall classification performance.

### *B. Decision Tree*

A classification decision tree is a tree structure that describes the classification of instances. The decision tree consists of nodes and directed edges. There are two types of nodes in a decision model: internal nodes and leaf nodes. For details, the internal nodes represent a feature or attribute and nodes represent a class.

The decision tree algorithms mainly include ID3 algorithm [15], C4.5 algorithm [16] and CART algorithm [17]. This paper uses CART decision tree algorithm. Classification and regression tree (CART) is a widely used decision tree learning algorithm. It uses the Gini index as a feature selection criterion to divide the sample set.

If data set D have a class, the probability that the sample belongs to the i-th class is pi, then the Gini value of data set D can be expressed as

$$\mathrm{Gini(D)} = 1 - \sum_{i=1}^{n} p_i^2 \qquad (1)$$

Under feature A, the Gini index of set D is defined as

$$\mathrm{Gini(D,A)} = \sum_{i=1}^{n} \frac{|D_i|}{|D|} Gini(D_i) \qquad (2)$$

The Gini value represents the uncertainty of the set D, and the Gini index represents the uncertainty of the set D when segmented by feature A. The larger the Gini index, the greater the uncertainty of the sample set.

### C. *XGBoost Algorithm*

XGBoost [19] belongs to ensemble learning algorithm, which mainly inherits the idea of GBDT and improves it. GBDT uses the first-order information when calculating the negative gradient value, while XGBoost uses the second-order expansion of Taylor to get the first-order and the second-order information. At the same time, XGBoost borrows the idea of feature column sampling and data sampling, which not only increase the training speed, but also effectively prevents over-fitting.

## III. Power Quality Disturbance Source Identification Model

### *A. Steady-state Disturbance Event*

#### *1) Disturbance event definition*

According to the power grid operation requirements, the power quality disturbance event is defined as all abnormal operation events caused by power quality disturbances in the power grid. It mainly includes the monitoring point power quality index exceeding the standard value and the power quality index exceeding the contracted level. The end user of the system cannot operate normally due to the power quality disturbance.

*2) Disturbance event and disturbance event* *characteristics*

According to the definition of the disturbance event, the disturbance event classification library and the disturbance feature classification library are respectively established. Among them, the cause of the disturbance event is divided into the cause of steady-state disturbance and the cause of transient disturbance, which respectively represent the event disturbance source that produces the steady-state disturbance characteristics and the transient disturbance characteristics. The causes of steady-state disturbance events mainly include ordinary railways, high-speed railways, wind power, and photovoltaics.

The monitoring data collected from the existing power quality monitoring system is a multi-dimensional massive data. The main indicators in the monitoring data include frequency deviation, active power, positive sequence voltage, negative sequence voltage, negative sequence unbalance current, voltage variation frequency, interharmonic current phase angle, and fundamental apparent power. Therefore, this paper uses common iron, high-speed rail, wind power and photovoltaic as the identification target, and uses three-phase active power, negative-sequence unbalanced current, three-phase third harmonic and three-phase fifth harmonic as the classification characteristic to predict the unknown monitoring point data.

The time domain characteristics of high-speed rail, general iron, wind power and photovoltaic are shown in Fig.1. By comparing the time-domain diagrams of phase of different power quality disturbance event sources, the power values of high-speed rail and ordinary iron are mostly positive.

From the above analysis of the power quality disturbance characteristics, it can be known that statistical methods can be used to analyze the disturbance event features in the time domain to extract features. In this case, the classifier can better predict the type of disturbance events at the monitoring points.

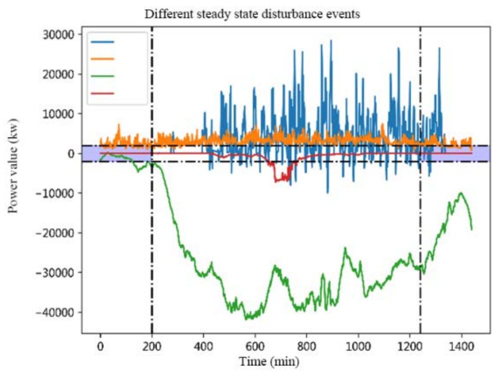


**Fig.1.** Time-domain variation diagram of phase A power for different steady-state disturbance events.

In this paper, the three-phase power, negative sequence unbalance current, three-phase third harmonic and three-phase fifth harmonic of the disturbance event are extracted according to the formula 3. In formula 3, x represents the feature used in this paper, and N represents the number of data points in the time domain.

$$\bar{x} = \frac{1}{N}\sum_{k=0}^{N-1} x_k \qquad (3)$$

The data distribution of power, negative sequence unbalanced current, third harmonic, and fifth harmonic pass formula are shown in Fig.2.

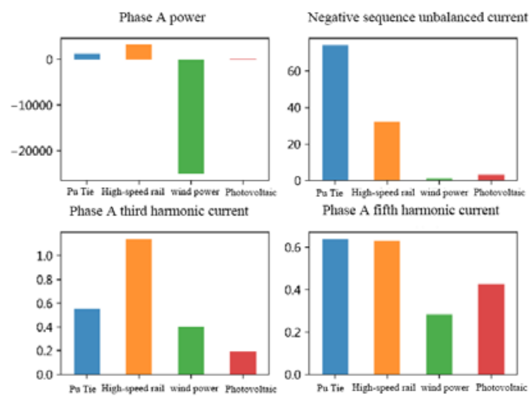


**Fig.2.** Statistical characteristics of different characteristics of steady-state disturbance events.

Through analysis and comparison, high-speed rail, general iron, wind power, and photovoltaic have obvious differences in three-phase power indicators. Additionally, there is a significant difference between the high-speed rail and the general iron in the negative sequence unbalanced current, the third harmonic and the power. There are also significant differences in the indicators of wind power and photovoltaic power and negative sequence unbalanced current. It can be seen that the characteristics transformed by the statistical method have significant differences between the steady-state disturbance sources.

The time domain waveform of the photovoltaic disturbance source is shown in Fig. 3. As can be seen from the figure, the photovoltaic is stable at both ends of the data, and the power value is small. In the middle section, the power value drops sharply which corresponding to the actual scene. And comparing with the graph 1, it shows that several other steady-state disturbance events do not have this characteristic of photovoltaics. Therefore, this feature of the photovoltaic disturbance source is taken as a priori knowledge, aims to improve the prediction accuracy of the recognition model.

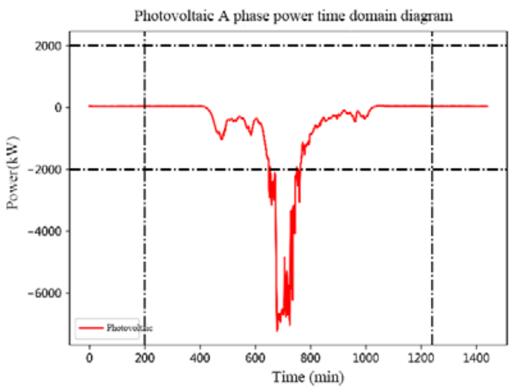


**Fig.3.** Photovoltaic disturbance event phase A power time domain diagram.

## B. *XGBoost Based Classifier Design*

This paper designs a power quality disturbance event classifier based on XGBoost. The classifier structure is shown in Figure 4. The classifier takes the key features of the steady-state disturbance event as the input vector of XGBoost, and uses the CART decision tree as the basic unit to build the model. The identification of the cause of the disturbance event is viewed as multi-objective classification problem.

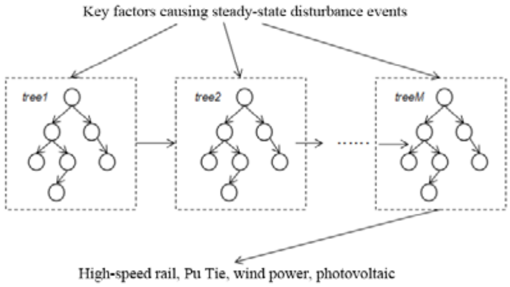


**Fig.4.** XGBoost classifier.

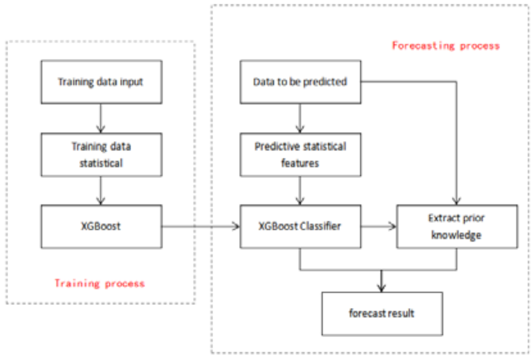


**Fig.5.** Model flow chart.

The workflow of the XGBoost-based power quality steady-state disturbance event recognition is shown in Figure 5. Specifically, it can be divided into a training phase and a prediction phase. The training process is as follows:

1) Read the monitoring point raw data;

2) The raw data is processed using statistical methods to obtain the characteristics of the model;

3) Using the XGBoost model shown in Figure 4, using the feature obtained in step 2, set XGBoost to consist of 30 CART trees;

4) Obtain a power quality steady-state disturbance event recognition model.

Use the trained XGBoost classifier and combine the prior knowledge to predict the test data. The specific process is as follows:

1) Perform statistical feature extraction on the predicted data and extract prior knowledge;

2) Input the feature extracted by the statistical method into the XGBoost classifier;

3) The classification result obtained in step 2 is merged with the prior knowledge extracted in step 1 to obtain the final prediction result.

## IV. EXPERIMENT AND ANALYSIS

In this paper, the actual power quality monitoring data of high-speed rail, general iron, wind power and photovoltaic in Henan Province is taken as the research object. Based on the XGBoost classification model, prior knowledge is added to identify the steady-state disturbance events of power quality.

### *A. Experimental data*

The 19-day monitoring data of several high-speed rail, general iron, wind power and photovoltaic monitoring points were used as the original experimental data. The statistical characteristics of the monitoring data of each monitoring point are extracted every day. Each data includes 10-dimensional features, which are A phase power, B phase power, C phase power, negative sequence unbalanced current, phase A third harmonic, phase B third harmonic, phase C third harmonic, phase A fifth harmonic, phase B fifth harmonic and phase C fifth harmonic. Based on the statistical characteristics of the test data, the prior knowledge is extracted. At the same time, set the sample label, set the label of the high-speed rail to 0, set the general iron to 1, the wind power to 2, and the PV to 3.

### *B. Experimental results and discussion*

The training set and the test set are based on the random selection. 70% of the original data is as the training set, and the other 30% is as the test set. After each model training and testing, the model accuracy is computed. The above process was repeated 100 times, and 100 precisions were averaged to obtain the accuracy of the method. The experimental results are shown in Table 1.

**Table1.** Model prediction result.

| Label | Prediction | Label | Prediction | Label | Prediction | Label | Prediction |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 0 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |

According to the classification results, the high-speed rail classifier can completely separate the high-speed rail from the general iron. There is a classification error between wind power and high-speed rail, and the overall classification accuracy rate is 98.7%. Moreover, the experimental results obtained after many experiments are stable and have good generalization performance.

In order to prove the effect of the model, this paper selects the Neural Network and the Decision Tree model as the base model for verification. After 100 times of ten-fold validation of the experiment, the accuracy of Neural Network model can reach 0.952 and the accuracy of Decision Tree can reach 0.9478. It is shown again that XGBoost has achieved satisfactory results in the identification of disturbance source categories in this paper, which exceeds other base models.

## V. CONCLUSION

In this paper, we studied the XGBoost algorithm and XGBoost-based power quality disturbance events, and the XGBoost-based classifier is applied to the new application field of power quality disturbance event analysis. The experimental results show that the designed features are useful in recognizing power quality disturbance event. More inclusive experiments illustrate the proposed method can identify the steady-state disturbance event well, and the algorithm computational complexity is low.

## ACKNOWLEDGEMENT

This work was supported by SGCC Scientific and Technological Project “Research and Application on Technologies of Disturbance Sources Analysis Based on Power Quality Monitoring Data” (SGRIDLKJ [2017] 463).

## REFERENCES

[1]Chen Hongkun, Huang Juan. Data Mining and Its Application in Power Quality Analysis[J]. Electric Power System and Automation, 2009, (5): 51-5.

[2]Xu Yonghai, Zhao Yan. Identification of Power Quality Disturbance Based on Short-Term Fourier Transform and Disturbance Time Orientation by Singular Value Decomposition[J]. Power System Technology, 2011, (8): 174-80.

[3]HUANG J, HEZUO Q U, XIAOMING L I. Classification for Hybrid Power Quality Disturbance Based on STFT and Its Spectral Kurtosis[J]. Power System Technology, 2016.

[4]Qin Yinglin, Tian Lijun, Chang Xuefei. Classification of Power Quality Disturbance Based on Wavelet Energy Distribution and Neural Network[J]. Electric Power Automation Equipment, 2009, 29(7): 64-7.

[5]Zhan Yong, Cheng Haozhong, Ding Yifeng, et al. S-Transform-Based Classification of Power Quality Disturbance Signals by Support Vector Machines[J]. Proceedings of the CSEE, 2005, 04): 53-8.

[6]ZHAO F Z, YANG R G. Power-quality disturbance recognition using S-transform [J]. Ieee T Power Deliver, 2007, 22(2): 944-50.

[7]BHUIYAN M J U, BEGUM T A, ALAM M R. S-Transform and Mahalanobis Distance Based Approach for Classifying Power Quality Disturbances[J]. 2017 International Conference on Electrical, Computer and Communication Engineering (Ecce), 2017, 681-5.

[8]Li Tianyun, Zhao Wei, Li Nan, et al. A New Method for Power Quality Detection Based on HHT[J]. Proceedings of the CSEE, 2005, 17): 52-6.

[9]TAO W Q, YIN S G, DING M, et al. Classification of Power Quality Disturbance Signals Based on S-Transform and HHT[J]. Chin Contr Conf, 2013, 3639-44.

[10]Li Xiaona, Shen Xinglai, Xue Xue, et al. Power Quality Disturbance Based on Improved HHT and Decision Tree[J]. Electric Power Construction, 2017, 38(2): 114-21.

[11]Huang Nantian, Peng Hua, Cai Guowei, et al. Feature Selection and Optimal Decision Tree Construction of Complex Power Quality Disturbance [J]. Proceedings of the CSEE, 2017, 03): 776-86.

[12]Guan Chun, Zhou Weiwei, Lu Weiguo. Recognition of Multiple Power Quality Disturbances Using Multi-Label RBF Neural Networks[J]. Transactions of China Electrotechnical Society, 2011, 08): 198-204.

[13]Qin Ye, Yuan Haiwen, Yuan Haibin, et al. Classification of Power Quality Disturbances Based on Optimized Least Squares Support Vector Machine[J]. Transactions of China Electrotechnical Society, 2012, 08): 209-14.

[14]KITTLER J, HATEF M, DUIN R P W, sss et al. On combining classifiers [J]. Ieee Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 226-39.

[15]QUINLAN J R. Induction of Decision Trees [J]. Mach Learn, 1986, 1(1): 81-106.

[16]QUINLAN J R. C4.5: programs for machine learning [M]. Morgan Kaufmann Publishers Inc., 1993.

[17]DE'ATH G, FABRICIUS K E. Classification and regression trees: A powerful yet simple technique for ecological data analysis [J]. Ecology, 2000, 81(11): 3178-92.

[18]Friedman J H . Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, 29(5):1189-1232.

[19]Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. 2016.