

PAPER • OPEN ACCESS

Identification of Power Quality Disturbances Based on Sample Entropy and Weighted Optimization Random Forest

To cite this article: Xiaohui Liao *et al* 2020 *J. Phys.: Conf. Ser.* **1486** 062018

View the [article online](#) for updates and enhancements.

You may also like

- [The Study of Power Quality Analysis by S-transform](#)
Jiajin Qi and Lin Lins
- [Analysis of Electrical Power Quality Disturbances Based on Empirical Mode Decomposition and Statistical Parameters](#)
Faqih Rofii, Agus Naba, Hari Arief Dharmawan et al.
- [Simulation and FPGA Implementation of Transient Power Quality Disturbance Detection](#)
Xionghui Lai



ECS
The
Electrochemical
Society
Advancing solid state &
electrochemical science & technology

DISCOVER
how sustainability
intersects with
electrochemistry & solid
state science research

Identification of Power Quality Disturbances Based on Sample Entropy and Weighted Optimization Random Forest

Xiaohui Liao, Chuanchuan Chen^{*}, Yuhan Wang, Mingjian Cui and Chuan Liu

School of Electrical Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China

^{*}Corresponding author's e-mail: 2301769908@qq.com

Abstract. With the increasing number of non-linear loads and impact loads in the actual power grid, power quality problems are becoming more and more serious. Accurate and rapid identification of power quality disturbance signals is of great significance for improving the power quality and ensuring stable operation of the power grid. For this reason, this paper proposes a multi-resolution analysis of wavelet transform and a combination of sample entropy to extract the characteristics of power quality disturbance signals. The classification effect is better than the initial samples. The weighted optimization random forest classifier is used to classify the features, which reduces the error caused by the mode voting method of the random forest. The simulation and actual test results show that the power quality disturbances detection algorithm based on sample entropy and weighted optimization random forest has better classification accuracy than the decision tree, extreme random tree and traditional random forest.

1. Introduction

In recent years, the development of smart grids has led to a large increase in non-linear loads and impact loads. For example, the use of power electronic equipment, electro-chemical industrial equipment and electric power-driven equipment inevitably produces non-sinusoidal waveform currents, which leads that stable operation of the power grid is challenged. The waveforms of the public connection points are severely distorted and the precision and sensitive electrical equipment has stricter requirements on power quality (PQ). Due to the unqualified PQ, the domestic and international power accidents are not uncommon. Therefore, the issues of PQ have become an actual problem that the power sectors and end users urgently need to consider and solve. The most important key point to improve PQ is the ability to accurately and quickly detect and identify PQ disturbances (PQDs)[1].

In view of the identification of PQDs, many scholars at home and abroad have conducted research. The identification of PQDs mainly includes two steps of feature extraction and classification. The commonly used PQ feature extraction methods can be roughly divided into: short time Fourier transform (STFT), wavelet transform (WT), S transform (ST), mathematics morphology (MM) and other methods[2]. In this paper, a multi-resolution analysis (MRA) method of WT combined with sample entropy (SE) is proposed to extract the characteristics of the PQD signals. The classification effect is better than that of the initial samples classification, which improves common methods for poor versatility and the low recognition accuracy problem. Several common PQD signals such as voltage rise, voltage sag, harmonics, and so on are decomposed into detail factors and scale factors by MRA method, and extracting its eigenvalues to form a feature vector. The samples value of PQD signals is obtained by SE calculation. Finally, the characteristics of the disturbance signals extracted



by the SE are classified and trained by the weighted optimization random forest (WORF). It is verified by experiments that this method is effective and feasible.

2. Feature extraction

2.1 Multi-resolution analysis method

In 1988, S.Mallat proposed the method of MRA when constructing orthogonal wavelets. The main idea of this technology is to filter the given signals by low-pass filter and high-pass filter when the signals are processed and decompose the signals to obtain approximate signals and detail signals (or low-frequency signals and high-frequency signals) 2 versions. The scaling scales and wavelet coefficients of the signal $f(t)$ on different decomposition layers can be obtained by the formula:

$$\begin{aligned} c_{j,k} &= \sum_{m \in \mathbb{Z}} h(n-2k)c_{j-1,m} \\ d_{j,k} &= \sum_{m \in \mathbb{Z}} g(n-2k)c_{j-1,m} \end{aligned} \quad (1)$$

Where $c_{j,k}$ is the approximate signal of the original signal reconstructed after the output of the low-pass filter; h is the coefficient of the low-pass filter; $d_{j,k}$ is the detailed signal of the original signal reconstructed after the output of the high-pass filter; g is the coefficient of the high-pass filter.

The decomposition principle is shown in figure 1. A3 shows that the approximate signal by decomposing the original signal through factor 2^3 is obtained after performing the 3-level decomposition; D3 is the detail signal of the 3rd stage[3-4]. As shown in figure 1.

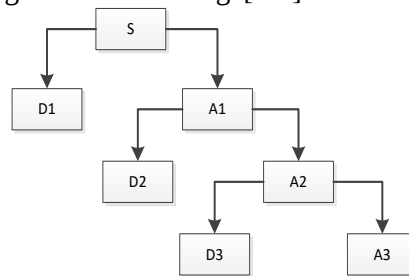


Figure 1. Signal decomposition tree structure diagram

2.2 Sample entropy

Moorman and Richman proposed the concept of SE on the basis of Pincus's approximate entropy (AE) theory. AE is mainly used to measure the probability of generating a new type of signal, the starting point is the complexity of the time series and the higher the complexity, the larger the approximate entropy. SE is also a new method for measuring time series and it is used to calculate the logarithm of sum. SE has two advantages over AE: the value of the SE is not affected by the data being too long or too short and the SE has good consistency. For example, if one time series has higher value than the other, then other m and n have higher values[5-6].

2.2.1 Calculation of sample entropy. In general, for the time series $\{x(n)\} = x_1, x_2, \dots, x_n$, the SE is calculated as:

Step1: First, a set of vector sequences of m dimensional numbers is grouped by sequence number, which is $x_m(1), \dots, x_m(n-m+1)$, in this sequence:

$$x_m(i) = \{x_m(1), \dots, x_m(n-m+1)\} \quad (1 \leq i \leq n-m+1) \quad (2)$$

Step2: Define $d[x(i), x(j)]$ to represent the one with the largest distance between $x(i)$ and $x(j)$, which is

$$d[x(i), x(j)] = \max_{k=0 \rightarrow m-1} \{|x(i+k) - x(j+k)|\} \quad (3)$$

Step3: Counting the distance between $x(i)$ and $x(j)$ is all less than the number n_i^m of r ,

($1 \leq i, j \leq N-M+1$), and then the ratio to the total number $N-M$ is calculated:

$$C_i^m(r) = \frac{n_i^m}{N-M} \quad (4)$$

Then find the average

$$C^m(r) = \frac{1}{N-M+1} \sum_{i=1}^{N-M+1} C_i^m(r) \quad (5)$$

Step4: Let $m=m+1$, repeat steps Step1, Step2, Step3, and calculate $C_{m+1}(r)$.

Step5: Calculate SE.

$$SampEn = -\ln \frac{C^{m+1}}{C^m} \quad (6)$$

The value of SE is related to the value of m and r . Generally, m is between 2 and 10, and r takes 0.1 to 0.25 times of the standard deviation (SD) of the original time series. In this paper, the SE is calculated after the sequence is standardized. When calculating the SE, m takes 2, and r takes 0.15SD. A flow chart of multi-resolution decomposition and SE calculation is performed as shown in Figure 2 for the PQD signals[7-8].

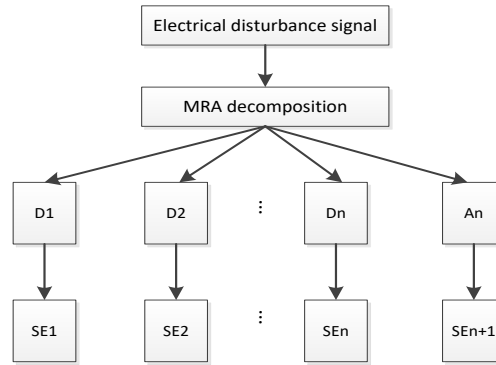


Figure 2. Sample Entropy Calculation Structure diagram

3.Introduction to Random Forest Algorithm

Random forest (RF) is a combined classifier composed of multiple independent decision trees (DTs). The basic principle is: the self-sampling method is adopted to extract data from the training sample set to form a sub-training set, and then the sub-training set is generated into a DT according to the method of randomly selecting a certain number of features to form a RF. After each sample data to be classified enters the generated RF, it produces a variety of classification results. Usually, the final output category is determined by the mode of the output category of this certain number of DTs. In this paper, the output category of the DT is determined by the weighted optimization model.

3.1Weighted optimization random forest model

The WOLF algorithm is an improvement on the basis of the traditional model voting by mode, and weighting of each DT, so that the DT with better classification performance has higher decision-making power than the poor performance DT. Assume that the K -decision tree weights of the random forest are $w_1, w_2, \dots, w_k$, and each DT must multiply this weight when voting[9]. The WOLF model output is:

$$f_{WRF}(a) = \arg \max_{i=1,2,\dots,k} \left\{ \sum_{k=1}^k [(f_k(a) = i)w_k] \right\} \quad (7)$$

3.2Determination method of random forest weight

In the process of constructing and training RF model, extract some samples of the training samples as

the training sample set of a single decision tree by randomly with replacement sampling. The without sampled data set was called out-bag data (OOB), and the evaluation of prediction ability of the decision tree with out-bag data was called OOB estimation. The construction of the decision tree and the OOB estimation of the decision tree can be completed serially. When the RF model is constructed, each decision tree can obtain a corresponding OOB evaluation value and assign the weight of the corresponding DT. For the DT $f(x)$, its voting weighted value is defined as, expressed as:

$$P_{OOB} = \alpha \times \frac{S_+}{S} \quad (8)$$

Where S_+ is the evaluation using OOB data to predict the number of samples in the correct decision tree; S is the total sample size participating in the decision tree, and α is the adjustment factor.

In the experimental process of training RF, we obtain k decision tree models $\{f_1(x), f_2(x), \dots, f_k(x)\}$ and their corresponding OOB prediction accuracy for each decision tree $\{P_1, P_2, \dots, P_k\}$, then the prediction result of the final model can be expressed as:

$$\max\{c / c_i = \sum_{j=1}^k P_{OOB_j} I(f_j(x) = l_i), l_i \in C, j = 1, 2, 3, \dots, k\} \quad (9)$$

Where c is the set of all category labels, l_i is the i -th classification label in set c , and $I(f(x) = l_i)$ is an indicative function. When the prediction result of the DT $f(x)$ is the category label l_i , the indicative function is equal to 1, otherwise equal to 0, p_j is the weighted value of the j -th DT in the voting process, c_i is the voting weighted result obtained by the i -th category label, and the final prediction result of the model is the largest one of the total number of weighted votes obtained by each classification label[10-11].

3.3 Disturbance classification method based on random forest

Step1: Construct a mathematical model based on the voltage rise, sag, harmonics, interruption, and voltage flicker fault of the PQ of the grid, calculate the eigenvalue of the PQDs through the WRA and SE and the sample set is established;

Step2: According to the construction theory of RF, the training set x' is generated from the original sample set x by randomly with replacement sampling;

Step3: According to the data characteristics in the training set x' and the simulation of the identification accuracy of PQ fault signals, the optimal number of decision trees in the random forest model is determined as n_{tree} , i.e:

$$n_{tree} \leftarrow \{Accuracy\}_{max} \quad (10)$$

Step4: Select the optimal feature from the feature according to the principle of minimum Gini coefficient for branch growth, establish an n_{tree} decision tree, and do not pruning during the growth process;

Step5: After the RF model is established, the input PQ fault signals are tested by the feature data set extracted by WRA and SE;

Step6: The RF model uses the weighted optimization voting method for all DT results to determine the final power fault disturbance type.

4. Experimental results and analysis

In this paper, the voltage rise, sag, harmonics, interruption, flicker, and several typical PQD signals are built by matlab simulation software. By using the MRA method, the approximate signals and the detailed signals in each frequency band are obtained. According to the mathematical models of five kinds of fault signals, 150 groups of fault signals are simulated, 30 sets of data for each fault signal, and the fault signals are decomposed into D1, D2, D3, D4, D5, D6, A6 seven components by MRA. The SE extraction of 150 sets of fault signals is realized by the method of SE calculation. Table 1 is a mathematical model of PQD signals[12]. Figure 3 to Figure 7 are PQD signals decomposition simulation graph through MRA. Table 2 shows part of the eigenvalues of the SE calculation.

Table 1. Mathematical model of PQD

fault type	mathematical model
voltage rise	$u(t) = u_m(1 + a \times (\varepsilon(t - t_1) - \varepsilon(t - t_2))) \times \sin(2\pi ft)$
voltage sag	$u(t) = u_m(1 - a \times (\varepsilon(t - t_1) - \varepsilon(t - t_2))) \times \sin(2\pi ft)$
harmonic	$u(t) = \sin(2\pi ft) + \sum_{k=2}^{k=13} a_k \sin(2\pi kft)$
voltage interruption	$u(t) = u_m(1 - (\varepsilon(t - t_1) - \varepsilon(t - t_2))) \times \sin(2\pi ft)$
voltage flicker	$u(t) = u_m \sin(2\pi ft) \times (1 \pm \frac{1}{2} \times \frac{\nabla u}{u} \sin(2\pi ft))$

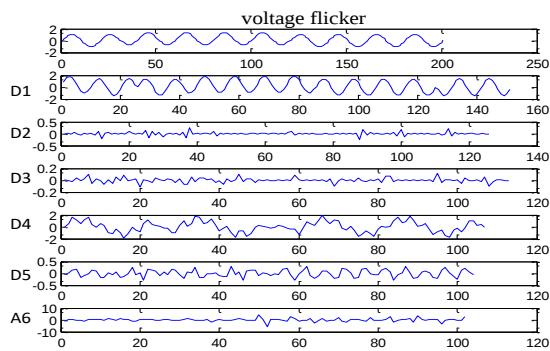


Figure 3. Voltage rise

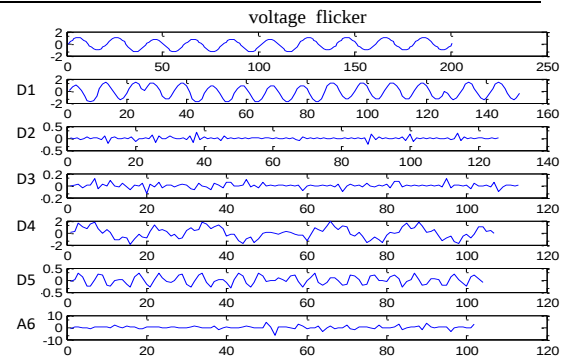


Figure 4. Voltage sag

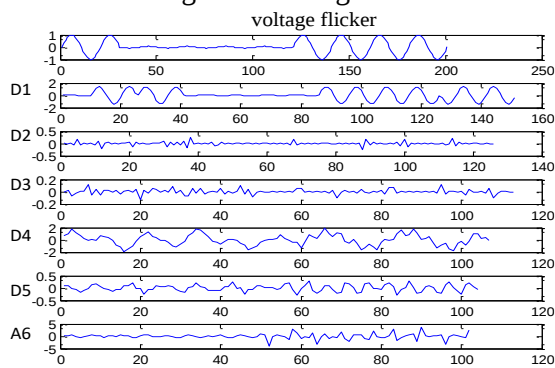


Figure 5. Voltage interruption

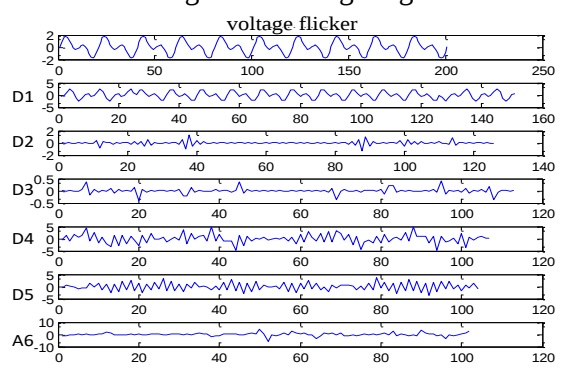


Figure 6. Harmonic

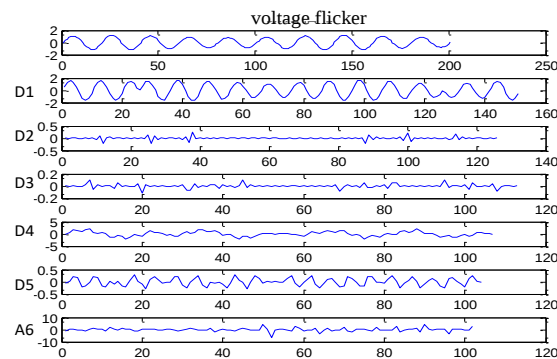


Figure 7. Voltage flicker

Table 2. Part of the eigenvalue by SE calculation

SamEn1	SamEn2	SamEn3	SamEn4	SamEn5	SamEn6	SamEn7	fault type
0.0753	0.0945	0.0065	2.2156	0.5051	1.9315	1.4586	voltage rise
0.0753	0.0914	0.0065	1.9095	0.4912	1.8056	1.5163	voltage rise
0.0553	0.1089	0.0065	1.6341	0.5546	2.1145	1.3863	voltage sag
0.0553	0.1086	0.0065	1.5404	0.5406	2.2513	1.4137	voltage sag
0.2234	0.1235	0.0134	2.5123	0.5351	1.4996	1.2993	Voltage interruption
0.1881	0.127	0.0271	1.754	0.371	1.4773	1.0761	Voltage interruption
0.1662	0.0731	0.0251	2.1145	0.9615	1.8589	1.4351	harmonic
0.0282	0.0762	0.1177	2.9704	1.1042	1.5626	1.9694	harmonic
0.2932	0.0939	0.0123	2.0149	0.5866	2.1832	1.4553	Voltage interruption
0.1996	0.1065	0.0043	2.0149	0.5519	2.3418	2.5257	Voltage interruption

SamEn1 represents the SE value calculated for the d1 component, and similarly, SamEn2 to SamEn6 corresponds to d2, d3, d4, d5 and d6. SamEn7 represents SE value calculated for the a6 component. Through Table 2, the feature vector V [SamEn1, SamEn2, SamEn3,..., SamEn7] can also be generated with each SE value as an element.

Use the jupyter notebook platform in anconda software to build the environment and install model libraries such as tensorflow and sklearn. The WOF model is trained with the calculated characteristics of 150 sets of SE. Through the simulation experiment, the model parameters of the RF are determined as shown in figure 8. When the $n_{tree} > 100$, the accuracy of the WOF gradually becomes stable. When the final $n_{tree} = 125$, the WOF identification accuracy is the highest. As shown in Table 3, at the same time, the DT, RF, and extreme random tree (ERT) are compared with the WOF and it is found that the WOF achieves better recognition accuracy for the PQ fault signals.

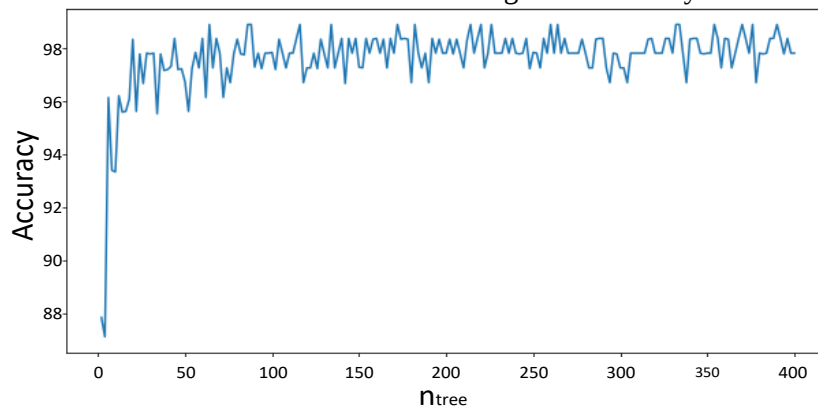


Figure 8. DT number and accuracy simulation graph

Table 3. Classification accuracy rate comparison results

Classifier	Decision tree	Random forest	Weighted optimization random forest	Extremely random tree
Accuracy	98.38%	98.89%	99.41%	98.38%

5. Conclusion

In this paper, a new method based on MRA and SE extraction is proposed to improve the problem of poor permeability and low identification accuracy. On this basis, the WOF algorithm is used to classify and identify five kinds of PQD signals, which reduces the generalization error caused by the traditional model voting method in random forest and effectively improves the recognition accuracy of PQD signals. Through a large number of simulation verification, the following conclusions are

obtained: 1) The MRA of WT can accurately analyse the frequency of the disturbance signals and the time of occurrence, which is a good reflection of the relationship between frequency and time. 2) The classification effect based on SE is better than the initial samples, but it is found through experiments that the calculation process of SE is complicated and the cycle is more, which leads to a longer running time of the SE algorithm. 3) Optimize the voting process of the model and improve the accuracy of feature classification and recognition based on the WOF to identify the PQDs.

References

- [1] Li, X., Shen, X., Xue, X., (2017) Identification of power quality disturbance based on improved HHT and decision tree. *J. Electric Power Construction*, 38: 114-121.
- [2] Yang, J., Jiang, S., Shi, G., (2019) Composite power quality disturbance identification based on segmented improved S Transform. *J. Power System Protection and Control*, 47:64-71.
- [3] Yang, F., (1999) *Engineering Analysis and Application of Wavelet Transform*, Beijing Science Press, Beijing.
- [4] Qu, G., Wang, C., Qi C., Zhao, L., Liu, Y., Chen, W., (2019) Study on power quality disturbance identification and diagnosis of SVM based on sample entropy and dragonfly algorithm. *J. Power Capacitor & Reactive Power Compensation*, 40:115-122.
- [5] Chen, y., Chen, y., Jian, j., (2017) Analysis and fault diagnosis of water pump unit based on wavelet packet sample entropy and SVM. *J. China Rural Water and Hydropower*, 3:165-175.
- [6] Li, G., Wang, H., Zhou, M., (2009) Identification of power quality disturbance based on improved wavelet entropy and support vector machine. *J. Transactions of China Electro technical Society*, 24: 161-167.
- [7] Guo, X., Tang, L., (2011) Research on heart sound recognition based on wavelet transform and sample entropy. *J. Journal of Computer Applications*, 27: 4555-4557.
- [8] ARIE, S., NIZAN, S., (2008) Magnetic anomaly detection using entropy filter. *J. Measurement Science and Technology*, 19: 52—55.
- [9] Li, G., Tie, Y., Qi, L., (2019) Multi-feature speech emotion recognition based on random forest classification optimization. *J. Microelectronics & Computer*, 36:70-73.
- [10] Ma, X., (2017). Optimization of stochastic forest model based on weighted decision tree. <http://www.doc88.com/p-2022558271473.html>.
- [11] Rodriguez-Galliano, V., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J., (2012) An assessment of the effectiveness of a random forest classifier for land cover classification. *J. ISPRS Journal Photogrammetry Remote Sensing*. 67:93-104.
- [12] Ning, K., (2018). Power quality analysis of distribution network with distributed power supply. <http://www.doc88.com/p-9793829859025.html>.