Multi-source Partial Discharge Identification of Power Equipment Based on Random Forest

Ran Deng¹, Yongli Zhu¹, Xuechun Liu¹, Yujia Zhai¹

¹State Key Laboratory of Alternate Electrical Power System with Renewable Energy Source, North China Electric Power University, Baoding, Hebei, China
liferan@foxmail.com

Abstract. At present, research on partial discharge type identification of power equipment mainly focused on single source discharge, and a small amount of multi-source discharge research also focused on pulse separation of multi-source signals. Signal separation could lose a lot of valid signals and wasted information resources. In this paper, a multi-source partial discharge type identification method based on Random Forest (RF) is proposed. Firstly, in the feature extraction, the statistical characteristics of the multi-source data are extracted directly instead of signal separation, and the discharge information is fully utilized. Secondly, in terms of the choice of classifier, considering that the traditional SVM method can not handle the value of 0, this paper selects the random forest strong classifier with good anti-noise ability to identify the multi-source discharge type. The test results show that this method is effective and the recognition rate of PD is as high as 98%.

1. Introduction

Power equipment malfunction is a huge hidden danger of the safe operation of the power system, and partial discharge can effectively reflect the malfunction of the power equipment[1-2]. Identifying the type of partial discharge is of great significance in determining the approximate location and severity of the fault. Partial discharge can be divided into single source partial discharge and multi-source partial discharge[3-5]. The identification of single-source PD is relatively mature while there are few studies on multi-source discharge. At present, the multi-source discharge feature extraction method is almost always to separate a single source signal from a multi-source signal[6-7]. This requires higher digital signal processing techniques, increasing the time cost of signal processing. In this paper, the method of extracting features directly from multi-source signals is adopted, which saves processing costs and, in addition, preserves the integrity of information to the greatest extent.

In this paper, the traditional phase resolved partial discharge(PRPD) spectrum statistical features are used as the feature set of the data. It mainly has 11 original features of positive and negative half-period skew Sk, steepness Ku, local peak point Pe and phase median Mv, cross-correlation coefficient CC reflecting the difference of the profile, discharge amount factor QF and corrected cross-correlation coefficient Mcc[8].

With the rise of artificial intelligence, various intelligent algorithms have also been applied to the identification of partial discharge types. Many scholars use Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree algorithm to identify partial discharges. SVM is the most widely used and has achieved good recognition results. However, when the SVM algorithm trains the model, the symbol function classifies only the points greater than 0 and less than 0. Therefore, there is an error at the point where the actual value is 0, which greatly limits the prediction accuracy of the
SVM. The black box algorithm random forest algorithm has strong anti-noise ability, and can directly extract features from signals, and the classification effect is good. At present, the random forest algorithm is widely used in agriculture, economy, medicine, etc. However, it is rarely used in the fault diagnosis of PD. Therefore, this paper selects the random forest classifier to identify multi-source PD. In terms of example simulation, this paper selects three common partial discharge forms of transformers: plate-to-board discharge, corona discharge and suspension discharge. Multi-source discharge data were constructed based on these three types of discharge. The results show that the proposed method can effectively identify multi-source PD.

2. Basic principles

2.1. Bootstrap sampling

Bootstrapping is a tool for statistics that derives the sub-data set $D'$ from an existing data set $D$. Supposing that there are $N$ data in data set $D$, firstly select one datum from it, put this data back, and then select one datum from $N$ data. The $N'$ times are repeatedly selected to obtain $D'$. Some of the data in $D'$ may be the same, and there may be some data in $D$ but not $D'$. Assuming that the number $N'$ of bootstraps is equal to $N$, then the probability that a certain sample $(x_n, y_n)$ is not in $D'$ can be calculated as follows:

$$
(1 - \frac{1}{N})^N \approx \frac{1}{e}, N \rightarrow \infty
$$

(1)

It can be obtained that when $N$ approaches infinity, about 1/3 of the samples are not drawn.

2.2. CART algorithm

RF is made up of a number of decision trees. Generate $k$ decision trees using the CART algorithm, as shown in Figure 1.

![Figure 1. Build process of RF](image)

![Figure 2. Gini coefficient change chart](image)

After the decision tree has been branched, the overall category is more purer. The calculation formula of the branch condition is as shown in Equation 2, where $v$ represents the total attributes of a certain feature, $D^v$ represents the number of samples of a certain type of attribute, and $D$ represents the total number of samples of the feature.

$$
b(D, a) = \sum_v \frac{|D^v|}{|D|} Gini(D^v)
$$

(2)

The smaller the value of $b(D, a)$, the lower the purity. Therefore, the condition that minimizes $b(D, a)$ should be chosen to fork the tree. The purity classification algorithms of decision trees mainly include ID3, C4.5 and CART. ID3 adopts information entropy increase as the partition attribute. If a feature has no classification value and the characteristics of all group data are different, ID3 algorithm will preferentially select this feature to divide data, which loses the meaning of classification. The C4.5 algorithm uses the information entropy increase rate as the partition attribute, it solves the
shortcomings of ID3 with the CART algorithm which uses the gini coefficient as the partition attribute. The CART algorithm takes the form of a binary tree and is simpler in structure. The logarithmic calculation used by the C4.5 algorithm is more complicated than the CART calculation using the square of the gini coefficient, so the random forest defaults to the CART algorithm. The Gini coefficient can be calculated as follows:

$$Gini(D) = 1 - \sum_{k=1}^{c} p_k^2$$  \hspace{1cm} (3)

Where $c$ is the total number of sample categories and represents the probability of the $k$th sample. Figure 2 is a graph showing the change in the Gini coefficient when $c = 2$. It can be seen from the figure that when the value is 0.5, the Gini coefficient is the largest and the purity of the data is the lowest.

2.3. Out of bag data——OOB
The data that appears in D but not in D’ is called out-of-bag data, which is recorded as OOB. These data are not used to train random forests, so they can be used to test the quality of the model.

3. Algorithm steps
- Constructing dual source data.
- Extracting the statistical features directly from all samples.
- Training RF classifier.
- Analysing classification results.

4. Simulation example

4.1. Single source data

The laboratory simulates three common discharge models of plate-to-plate, corona discharge and suspension discharge. The physical diagram is shown in Figure 3.

```
 a. plate-to-plate discharge   b. corona discharge   c. suspension discharge
```

Figure 3. Three common discharge models

When 200 data are selected in each type of data obtained, a total of 600 single source data can be obtained.

4.2. Constructing dual source data
There are 200 groups in each of the three single-source data, and each set of data has two-dimensional $\varphi - n$ spectra, $\varphi - q_{min}$ spectra, $\varphi - q_{ave}$ spectra and $\varphi - q_{max}$ spectra. The $\varphi - n$ data is selected for each of the three discharge types. When generating a dual source data between different types, a total of 600
sets of dual source data can be obtained. Combined with 600 sets of single source data, there are a total of 1200 sets of data. Figure 4 is a plate-to-plate discharge spectrum; Figure 5 is a corona discharge spectrum, and Figure 6 is a spectrum of suspension discharge.

4.3. Simulation steps

- Extract 11 statistical features directly from the $\phi-n$, $\phi-q_{sum}$, $\phi-q_{ave}$ and $\phi-q_{max}$ spectra of each set of data, and obtain 44 statistical features for each set of data. Therefore, the size of the sample set is $1200 \times 44$.
- Train the RF classifier. The number of random features of each split of RF is chosen to be 6, 7, and 8, respectively. Adjusting the maximum number of leaf nodes and the number of decision trees in RF makes the training effect the best. Table 1 lists the degree of importance of each feature when the number of random features is taken as 7, and Table 2 lists the recognition rates for every random features number.

| Feature number | Feature importance |
|----------------|--------------------|
| 1—7            | 0.0073459 0.0031237 0.0025368 0.0073196 0.0067386 0.0065711 0.0654381 |
| 8—14           | 0.0711243 0.0185669 0.0019973 0.0214531 0.0378914 0.0300544 0.0761755 |
| 15—21          | 0.0034907 0.0289114 0.0170679 0.0390476 0.0318343 0.0384547 0.0070657 |
| 22—28          | 0.0289154 0.0013516 0.0025029 0.0037452 0.0079664 0.0037570 0.0046522 |
| 29—35          | 0.0108367 0.0164297 0.0153687 0.0745318 0.0298047 0.0439112 0.0157389 |
| 36—42          | 0.0158468 0.0356254 0.0268934 0.0456327 0.0263367 0.0146970 0.0146778 |
| 43—44          | 0.0167543 0.0357367 |

It can be seen from Table 1 that the 44 features are of similar importance, and the discharge type can be identified by the difference of the multi-source signal spectrum.
Table 2. Classification results

| Number of random features | 6 | 7 | 8 | Suspension discharge recognition rate /% | 100 | 10 | 93 |
|---------------------------|---|---|---|-----------------------------------------|-----|----|----|
| Number of decision trees in RF | 50 | 0 | 500 | Board to board + corona discharge recognition rate /% | 93 | 10 | 93 |
| Board to board recognition rate /% | 93 | 96 | 93 | Board to board + suspension discharge recognition rate /% | 93 | 10 | 96 |
| Corona discharge recognition rate /% | 93 | 96 | 100 | Corona + suspension discharge recognition rate /% | 93 | 96 | 96 |
| Average highest accuracy of all discharges /% | 94 | 98 | 95 |

It can be seen from Table 2 that when the number of random features selected by RF splitting is 7, the recognition rate is the highest, which is 98%. That is, directly extracting the spectral features of the multi-source signal can accurately identify the type of discharge.

5. Conclusion

- In this paper, a multi-source partial discharge type identification method based on Random Forest (RF) is proposed. The random forest classifier has good anti-interference ability and strong adaptability to multi-source features.
- Directly extracting the characteristics of multi-source data, without separating single source from the multi-source signals, fully retains the information of the multi-source data and saves the cost of separating signals. The classification accuracy is high.

References

[1] Shen Dingshen, Nie Dexin, Wu Zhirong, et al. Discrimination of PD and Interference Pulse Using Graphical Analysis[J]. 1997, 44(1): 43-44.
[2] Shang Haikun. Feature Extraction and Pattern Recognition Method for Partial Discharge Signals of Power Transformers[D]. North China Electric Power University, 2014.
[3] Schichler U, Koltunowicz W, Endo F, et al. Risk assessment on defects in GIS based on PD diagnostics[J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2013, 20(6): 2165-2172.
[4] Zhang Xiaoxing, Shu Na, Xu Xiaogang, Li Xin, Tang Ju. Partial Discharge Recognition of GIS Based on Chaotic Characteristics of 3D Spectrogram[J]. Transactions of China Electrotechnical Society, 2015, 30(01): 249-254.
[5] Wang Caixiong, Tang Zhiguo, Chang Wenzhi, Zheng Shusheng, Li Chengyu. A Multi-Source Partial Discharge Signal Separation Method[J]. Proceedings of the CSEE, 2013, 33(13): 212-219.
[6] Wang Caixiong, Tang Zhiguo, Chang Wenzhi, Zheng Shusheng, Li Chengbiao. Separation method of anti-interference and multi-source discharge signal for partial discharge UHF detection[J]. Power System Technology, 2012, 36(03): 46-50.
[7] Cheng Yangchun, Zhang Zhenliang. Diagnosis of multi-source partial discharge of transformer based on random forest[J]. Proceedings of the CSEE, DOI:10.13334/j.0258-8013.pessee.171469
[8] Wang Liuwang, Liu Jia, Yang Lingbo, Yang Fugang, Fu Changhong. Application of Random Forest Method in Fine Identification of Corn-Soybean[J]. Acta Agronomica Sinica, 2018, 44(04): 569-580.