International Core Journal of Engineering Volume 6 Issue 3, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

# Single-phase Ground Fult Location Method for Wind Farm Collector Lines Based on RF-XGBoost

Yujia Zhai[1, a], Yongli Zhu[1, b]

[1]State Key Laboratory of Alternate Electrical Power System with Renewable Energy Source, North China Electric Power University, Baoding 071003, China.

[a]1778543596@qq.com, [b]yonglipw@163.com.

## Abstract

**In order to solve the problem of abandonment of wind power caused by the single-phase grounding short-circuit fault of the collector line in the wind farm, this paper proposes a RF-XGBoost-based single-phase ground fault location method for the wind farm collector line. This method uses random forest algorithm and XGBoost algorithm to realize fault location. First, the historical fault data is used as the original data set, and the random forest algorithm is used to reduce the dimensionality of the original data set to remove the poorly correlated data; the retained strong correlation feature values and label values are input to the XGBoost algorithm for training and training, and finally fit The relationship between the label value and each characteristic value is used for fault location. This method has high positioning accuracy and good extensibility. It is not only suitable for wind farm current collection lines, but also for other complex radial topologies. PSCAD / EMTDC and Python simulations are used to verify the correctness of the proposed method.**

## Keywords

**Wind farm; Collector line; Single-phase ground fault; Fault location; Random forest; XGBoost.**

## 1. Introduction

With the increasing support of national policies for new energy generation, especially wind power generation, the rapid progress of wind turbine technology, more and more wind farms have been completed and put into operation [1]. However, wind farms are mostly built in mountainous regions with harsh climate conditions, and manual line inspection and maintenance are more difficult. Especially when the wind farm's current collection line fails, how to troubleshoot the fault point and perform maintenance in a short time is a problem. Focus issues. In a wind farm, the power generated by the wind turbine is usually transmitted to the overhead line through a short cable line. The overhead line sends the collected wind currents to the 35kV low-voltage bus line of the 110kV booster substation. There are grounding transformers, overhead lines and cables together to form the current collection system in the wind farm field. It can be seen that the wind farm's current collection system has the characteristics of multiple branches, mixed networks, and short lines. To sum up, when a fault occurs in the wind farm's current collecting line, it is difficult to find the fault point by manual method quickly. A large amount of power generated by wind turbines cannot be transmitted to the power grid, resulting in the phenomenon of abandoning wind power. Therefore, it is necessary to study the short-circuit fault location of the wind farm's current collector line.

At present, the research on the short-circuit fault location of wind farm collector lines is still in its infancy. The current research on fault location methods is mainly aimed at transmission lines and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

distribution networks. The distribution networks and wind farms are similar in terms of current collection lines. They have the characteristics of relatively short lines, complicated topology, more line branches, mixed cable-overhead lines, and more power sources. The research on the location method of wind farm current collecting lines has great reference significance. Classification based on the method principle, currently there are mainly fault analysis methods [2-5], traveling wave methods and artificial intelligence fault location methods.

The method of artificial intelligence is a research hotspot in recent years. It is usually based on various newly proposed algorithms as the basis for data processing and model building, and mainly obtains data by adding measurement points in the network. Therefore, the way of artificial intelligence can often better describe the network topology and other properties, so it has an advantage when dealing with more complex networks. In recent years, artificial intelligence methods have been increasingly applied to fault location of transmission lines, especially to fault location of distribution networks. Reference [6] reduced the dimensionality of the solution by constructing an active tree and a passive tree, by cutting off the passive branches without fault current, and quickly and accurately performing global optimization by using the harmony algorithm. Finally, DG-containing Fault location of the distribution network. Reference [7] adopted a voting method to form the relationship between the feeder status and the limit current value, and initially selected the possible faulty sections. Finally, the Bayesian probability model was used to evaluate the above sections, and the most selected ones were selected. The segment that may be faulty, this method can realize fault segmentation, and has better fault tolerance. Literature [8] proposed a method using Bayesian compressed sensing theory to solve the node negative sequence voltage equation and then use the reconstructed negative sequence current to locate the method. This method does not require high equipment and does not require strict synchronization. In [9], the bat algorithm was used to solve the fault location problem of the transmission line of the distribution network, and the characteristics of the cellular automata were used to improve the local optimization ability of the algorithm, which has good convergence. Reference [10] uses phase-mode transformation relationships and different sequence diagrams corresponding to different fault types, and selects different feature values for grey correlation calculation to identify different fault types and locate them. Literature [11] proposed a new method of fault segmentation based on a combination of matrix algorithm and optimization algorithm. His advantage is its higher fault tolerance rate. References [12-15] used relay algorithm, mathematical morphology, non-extracted wavelet analysis in morphology, and neural network for fault location respectively, all of which have advantages.

## 2. Selection of eigenvalues by random forest algorithm

Due to the special structure of wind farms, fault locating methods that perform well on power distribution networks and transmission lines are no longer applicable to wind farm collection lines. Based on the characteristics of wind farm collector lines, this paper proposes a method for reducing the dimensions, training, and learning of single-phase ground fault data of wind farms by using machine learning methods to obtain the location of fault points. That is, a single-phase ground fault location method for wind farm collector lines based on RF-XGBoost.

Take Figure 1 as an example to illustrate the importance of feature selection. Figure 1 is a schematic diagram of a wind farm structure. The wind farm's current collection system is composed of multiple overhead lines connected to the 35kV bus and is transmitted to the power system through a boost transformer. The boost becomes a "star-delta" connection, and the high-voltage side adopts a star connection. The low-voltage side is connected in a triangle, and no zero-sequence current flows. There is also a Z-type ground transformer on the 35kV bus. The measuring points in the figure are distributed at the connection point between each collector line and the 35kV bus, and each wind turbine is connected to the collector line and the grounding transformer connected to the 35kV bus (ie, points A-M 'and O in the figure ).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

Next, a single-phase ground fault point is artificially set on the current collector line B-G. Due to the short current collector line, the faults can be set relatively densely. A fault point is set every 50m here. Collect the measured data of all measuring points after a few cycles of the fault (that is, after the fault is stable).

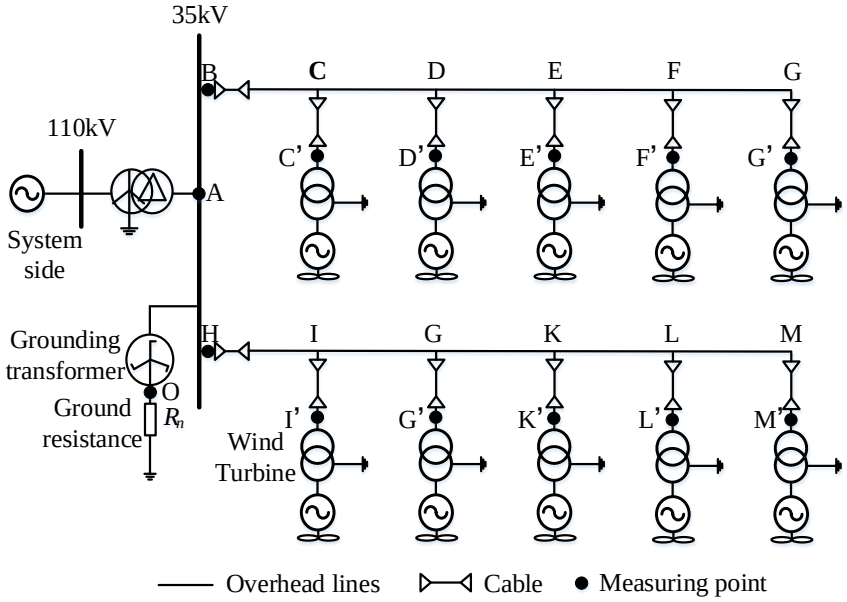


Fig.1 Schematic diagram of wind farm structure

It should be pointed out here that each measured data at this time has a large amount of data and many types of data. Taking Figure 2 as an example, assuming that a fault is set at 1500 meters on the BC section, there are 13 measurement points in the network shown in the figure, and each measurement point measures three-phase current and three-phase voltage respectively. For this topology, for At any fault point, a total of 78 data of 13 * 6 will be generated, and then the fault distance will be used as the label value, and the 78 data corresponding to the fault distance will be used as his attribute (ie, characteristic value) to participate in training.

At this point, it also means that this is a 78-dimensional regression problem. When the line topology is more complicated or the number of measurement points is greater, the data will have a higher dimension. When statistical analysis of data, too many variables corresponding to the data in theory will indeed carry more and richer information, but will greatly increase the complexity of the problem; in addition, among the many variables, especially the individual measurement of wind farms The data generated by a point is usually more or less electrically related. They are linearly related in them, and the information carried by these data overlaps with each other, that is, too many data dimensions carry redundant information. This redundant information not only complicates the problem, but also causes the classifier to be too scattered when performing regression or classification operations, weakening the accuracy of the results. It is therefore necessary to reduce the dimensionality of the input data.

This paper uses the random forest algorithm (RF) to reduce the dimensionality of the data for the following reasons:

1) The model establishment and data training of the random forest algorithm are faster and more efficient.

2) Random forest has better fault tolerance, and can better deal with imbalanced data of positive and negative examples.

3) Random forests can process higher-dimensional data, do not need to select feature quantities, and have a strong ability to resist overfitting.

4) Random forest does not rely on new test set data or cross-checking, but can perform unbiased estimation of data errors based on internal data during regression or classification. According to this

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

characteristic, the importance of each feature can be ranked. This is also the most important reason for using RF dimension reduction.

It is necessary to point out here that the selection of characteristic quantities can be determined artificially based on the experience of experts. For example, the single-phase ground fault of the wind farm collector line here, according to electrical experience, may be closely related to the single-phase ground fault. Electrical parameters and zero sequence parameters of faulty line fault phase. However, with the development of wind power scale and technology, especially the gradual popularization of electrical equipment condition monitoring, the number and types of sensors will increase, and more data volumes and data types will be monitored, such as equipment temperature, vibration, Factors such as partial discharge frequency, insulation aging, etc. will likely be factored into the analysis of transmission line failures. Therefore, it is necessary to use RF method for data dimensionality reduction. It can make the fault detection of collector lines in large wind farms efficient and intelligent.

## 3. Fitting of XGBoost Algorithm to Fault Features

After the input data is subjected to dimensionality reduction processing, several feature values that are most closely related to the label value will be obtained, and they will be used as inputs for learning and training to form the relationship between the fault distance and each feature quantity. However, in practice, the measured data (ie, each feature) at each measurement point of the wind farm is a very large data input. Even after dimensionality reduction, we still try our best to prevent overfitting, and at the same time let the fitted model reflect the internal law of single-phase ground fault location of the wind farm current collection line to the greatest extent. The XGBoost algorithm utilizes the idea of continuous iteration to continuously optimize the training model, so whether it is to combat overfitting problems, or to process complex data with multidimensional feature quantities, it is very outstanding. Therefore, this paper uses XGBoost algorithm to model the regression problem.

## 4. Fault location of RF-XGBoost algorithm

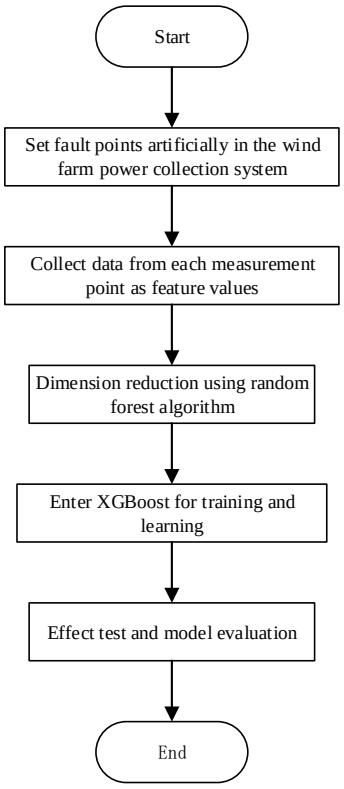


Fig.2 RF-XGBoost algorithm flowchart

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

In summary, this paper proposes a single-phase ground fault location method for wind farm collector lines based on RF-XGBoost. The specific algorithm steps are as follows:

Step 1: Set up a single-phase ground fault more densely in the wind farm model;

Step 2: The set fault distance is the label value of the training set, and the measured electrical parameters of each measurement point of the wind farm corresponding to the fault distance are used as the feature values of the training set;

Step 3: Use the random forest algorithm to reduce the dimensions of the original data set, and select several electrical parameters that are most closely related to the fault distance as the training feature values;

Step 4: The feature values and label values after the dimensionality reduction are input to the classifier XGBoost as the original data set for learning and training;

Step 5: Use the eigenvalues and label values after the dimensionality reduction as the test set to test and evaluate the regression calculation results.

The specific algorithm flow is shown in Figure 2.

## 5. Simulation verification

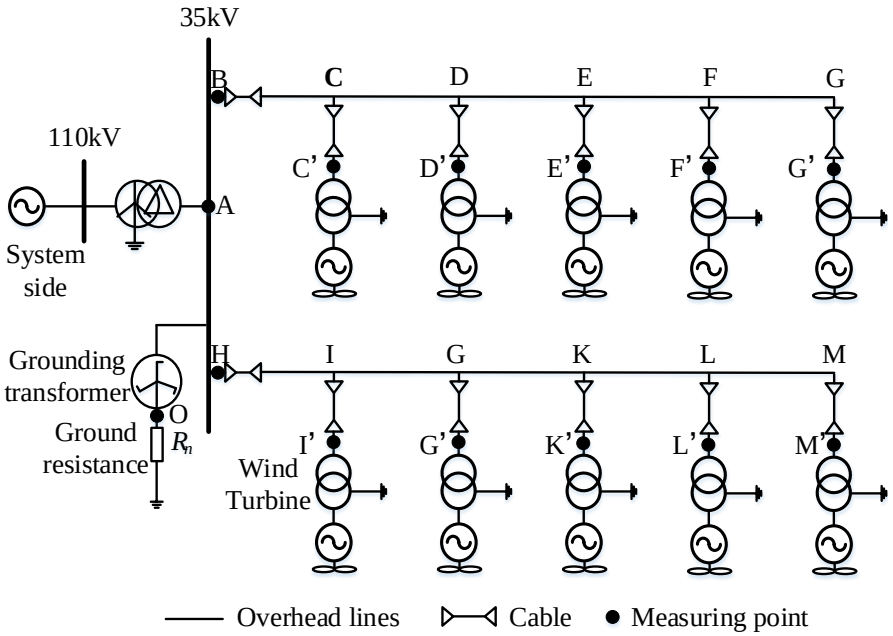


Fig. 3 Schematic diagram of simulation network

In this paper, PSCAD power system transient simulation software is used to verify the wind farms of two collector lines. As shown in Figure 3, two current collector lines are connected to the 35kV low-voltage side bus, and five wind turbines are connected to each of the current collector lines, and the unit models are the same. Each wind turbine access point is 2km adjacent, that is, the length of the overhead line section such as BC and CD is 2km; each wind turbine is connected to the overhead line by a 0.1km cable, that is, the overhead line sections C-C ', D -D 'equal length is 0.1km. The measuring points are distributed at the connection points between each collector line and the 35kV bus, each wind turbine connected to the collector line and the grounding transformer connected to the 35kV bus (ie, points A-M 'and O in the figure).

### 5.1 Location results of faults in different locations

In this paper, after setting the fault points as described above, the original data set is reduced by using a random forest, and the eigenvalues with larger weights are selected. Then, it is assumed that the phase A on the BG line is 0.33km, 1.42km, 2.505km, 3.707km, 4.882 A single-phase ground fault occurred at 10 points in total of 10 points, including km, 5.47km, 6.333km, 7.04km, 8.59km, and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

9.28km. A test set was formed by using the parameters of each measurement point and the corresponding fault distance during the fault to verify that the fault occurred on a line Positioning training effect at different positions. The results are shown in Table 1.

Table 1 Training results at different fault distances

| Actual fault distance (km) | Ranging results (km) | Error(km) | Relative error |
|---|---|---|---|
| 0.33 | 0.3335 | 0.0035 | 1.05% |
| 0.142 | 0.1421 | 0.0013 | 0.09% |
| 2.505 | 2.508 | 0.0033 | 0.13% |
| 3.707 | 3.705 | 0.0013 | 0.03% |
| 4.882 | 4.878 | 0.0038 | 0.08% |
| 5.470 | 5.466 | 0.0034 | 0.06% |
| 6.333 | 6.33 | 0.0026 | 0.04% |
| 7.04 | 7.038 | 0.0017 | 0.02% |
| 8.59 | 8.581 | 0.0086 | 0.10% |
| 9.28 | 9.271 | 0.0087 | 0.09% |

It can be seen from Table 1 that when the fault occurs at any position on the line BG, the algorithm proposed in this article can calculate the fault position more accurately, and the relative positioning errors are basically within 0.2%. In fact, the first fault The relative error of the dyeing is more than 1%, but due to the small base, in fact, the absolute error of the positioning of this sample is also about 3 meters, which is ideal. It can be seen that the RF-XGBoost algorithm can better solve single-phase ground faults that occur at different locations.

### 5.2 Positioning results under different transition resistance values

This example is based on the example in Section 5.1. The phase A ground fault is also set every 50m in the simulation diagram shown in Figure 3. The difference is that the group has different transition resistances for the same fault point. Control experiments were performed, so fault experiments were performed with transition resistances of 5, 10, 20, 50, 100, 150, and 200 ohms. Then, each simulation result data set corresponding to each transition resistance was reduced and trained separately. Finally, it is assumed that a phase A ground fault occurs at 5.47km from point B on BG, and this is the test set to verify the positioning effect at this time. The results are shown in Table 2.

Table 2 Positioning results under different transition resistance values

| Actual fault distance(km) | Transition resistance (Ω) | Ranging results (km) | Error (km) | Relative error |
|---|---|---|---|---|
| 5.47 | 5 | 5.475 | 0.0051 | 0.09% |
| 5.47 | 10 | 5.466 | 0.034 | 0.06% |
| 5.47 | 20 | 5.454 | 0.0156 | 0.29% |
| 5.47 | 50 | 5.473 | 0.0031 | 0.06% |
| 5.47 | 100 | 5.461 | 0.0087 | 0.16% |
| 5.47 | 150 | 5.486 | 0.016 | 0.29% |
| 5.47 | 200 | 5.452 | 0.0183 | 0.33% |

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

It can be known from Table 2 that when a phase A ground fault occurs at 5.47km from point B on BG, the relationship between the fault location and the input characteristic value can be accurately trained under different transient resistance values. The result is ideal.

### 5.3 Positioning results of different training algorithms

This paper proposes the RF-XGBoost algorithm trained by using random forest dimensionality reduction and then using the XGBoost algorithm, but there are currently several other widely recognized learning methods. This section will compare their performance with the RF-XGBoost algorithm. This article first uses the simulation method shown in Section 5.1 to obtain the original sample set, and then uses RF-XGBoost, XGboost, Nearest Neighbor Regression (KNN) and Support Vector Machine (SVR) four algorithms to learn and train, and uses Section 5.1 to show 10 The detection results of the fault points (numbered 1-10 from 0.33km-9.28km respectively) are shown in Figure 4.

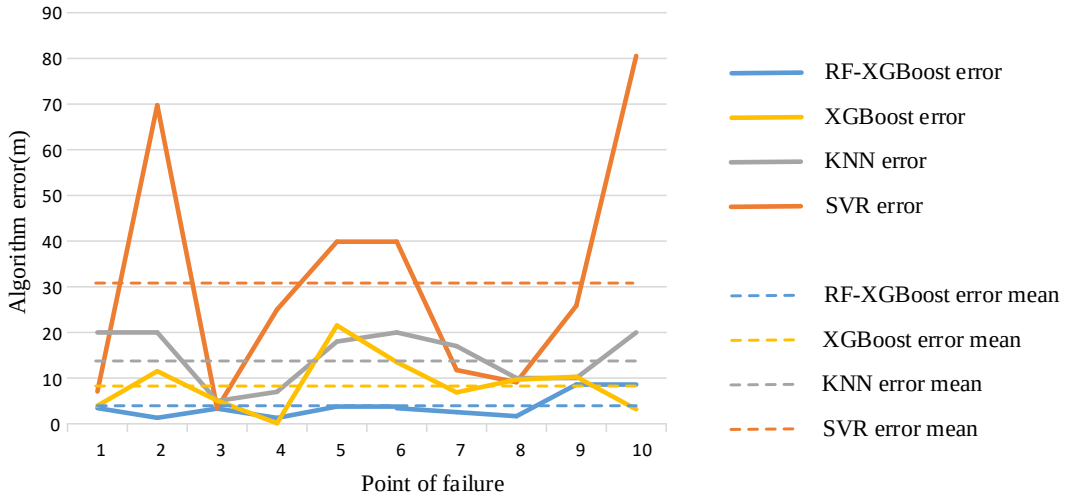


Fig.4 Positioning results of different training methods

As shown in the figure above, in the positioning results, except that the error of the RF-XGBoost algorithm is slightly higher than the XGBoost algorithm at the fourth fault point, it is superior to other algorithms in the other verification examples. In terms of the average positioning error of the four algorithms, the method proposed in this paper has the smallest error average, followed by the XGBoost method, and the KNN algorithm again, the SVR algorithm has the largest average error, and is much larger than the other methods. It is worth mentioning that the nearest neighbor regression (KNN) method is a lazy learning method. It uses a "remember" method to "remember" each piece of raw data that is involved in learning. When testing with a test set , It can only select the label value closest to the expected output result as the training result output by voting, but cannot get a value other than the label value in the training set. Therefore, in this paper, faults are set every 50 m. The training result of KNN can only be an integer multiple of 50. If the fault interval set in this paper is smaller, the training accuracy of KNN is higher.

In addition, in order to verify whether adding the random forest dimension reduction link can bring benefits to the single-phase ground fault location of the wind farm collector line, this section sets up the XGBoost algorithm for comparative experiments, that is, the original feature quantities are directly input without dimension reduction Trained in XGBoost. The experimental results show that the RF-XGBoost algorithm with random forest dimensionality reduction has improved accuracy.

### 5.4 Fault location under different operating conditions

Because the wind turbine's power generation is greatly affected by wind speed, it is necessary to perform simulation verification on fault location of wind farm current collection lines operating at different wind speeds. This article sets the fault to occur on the collector line B-G, and the fault type

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

is phase A short-to-ground. The fault is set in the overhead line section 3, 5, 7, and 9 km away from point B. Table 3 shows the fault location results.

Table 3 Location results of faults in different operating conditions

| Operating status | Actual fault distance (km) | Ranging results (km) | Error (km) | Relative error |
|---|---|---|---|---|
| Low wind speed | 3 | 2.998 | 0.002 | 0.07% |
| | 5 | 4.996 | 0.004 | 0.08% |
| | 7 | 6.994 | 0.006 | 0.09% |
| | 9 | 8.993 | 0.007 | 0.08% |
| Rated wind speed | 3 | 2.998 | 0.002 | 0.07% |
| | 5 | 4.997 | 0.003 | 0.06% |
| | 7 | 6.998 | 0.002 | 0.03% |
| | 9 | 8.995 | 0.005 | 0.06% |

As can be seen from Table 3, the method proposed in this paper can better solve the problem of single-phase ground fault location for wind farm collector lines operating under different operating conditions. The fault location can be ideally determined at both low wind speed and rated wind speed.

## 6. Conclusion

This paper proposes a RF-XGBoost-based fault location method for single-phase-to-ground faults in wind farm collector lines. A variety of simulation experiments have verified the accuracy of single-phase ground fault location under different fault distances, different transition resistance values, different training algorithms, and different operating conditions, with good results. The method described in this paper has high positioning accuracy, has the ability to be extended to other radial network topology structures, and has wider universality. However, simulation data has been used in the test and training of samples in this paper, and the method lacks field data to support the method, so further field verification is needed.

## References

[1] Zhang Liying, Ye Tinglu, Xin Yaozhong, Han Feng, Fan Gaofeng. Issues and measures related to large-scale wind power access to the power grid [J]. Chinese Journal of Electrical Engineering, 2010, 30 (25): 1-9.

[2] M.S.Sachdev, R.K.Agarwal. A technique for estimating transmission fault locations from digital impedance relay measurement[J]. IEEE Transactions on Power Delivery, 1998, 31(1):121-129

[3] L Eriksson, M M Saha, GD Rockefeller. An accurate fault locator with compensation for apparent reactance in the fault resistance resulting from remote end infeed [J]. IEEE Transactions on PAS, 1985, 104:424-436

[4] Wang Xuewen, Shi Fang, Zhang Hengxu, Xue Jingrun, Xi Zhihao, Xie Wei, Ling Ping, Liu Jinsong. Location Method of Single-Phase Ground Fault Section in Small Current Grounding System Based on Transient Energy [J]. Power System Technology, 2019 43 (03): 818-825.

[5] Zhang Ke, Sun Lizhi, Zhu Yongli, Liu Xuechun. Fault location method for wind farm collector lines based on vector deviation [J]. Automation of Electric Power Systems, 2019, 43 (10): 127-134.

[6] Liu Bei, Wang Yan, Chen Chun, Huang Haochuan, Dong Xuzhu. Application of Harmony Algorithm in Fault Location of Distribution Network Containing DG [J]. Transactions of China Electrotechnical Society, 2013, 28 (05): 280-284.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202003_6(3).0018

[7] Zheng Chenling, Zhu Gelan. Bayesian estimation of an intelligent distributed fault segment location algorithm for distribution networks [J / OL]. Power System Technology: 1-7 [2019-11-10]. Https: // doi. org / 10.13335 / j.1000-3673.pst.2019.0734.

[8] Jia Ke, Li Lun, Yang Zhe, Zhao Guankun, Bi Tianshu Research on Fault Location in Distribution Networks Based on Bayesian Compressed Sensing Theory [J]. Chinese Journal of Electrical Engineering, 2019, 39 (12): 3475-3486.

[9] Fujiacai, Lu Qingsong. Fault interval location of distribution network based on bat algorithm [J]. Power System Protection and Control, 2015, 43 (16): 100-105.

[10] Tong Xiaoyang, Zhang Shaoxun. Method of fault location and type identification for distribution network based on grey correlation [J]. Automation of Electric Power Systems, 2019, 43 (04): 113-118 + 145 + 119-124.

[11] Xu Biao, Yin Xiangen, Zhang Zhe, Pang Shuai, Li Xusheng. Fault Location in Distribution Networks Based on Matrix Algorithm and Optimization Algorithm [J]. Automation of Electric Power Systems, 2019, 43 (05): 152-161.

[12] Z. Lu,Ji T. Y.,Wu Q. H., et al. An adaptive distance relaying algorithm with a morphological fault detector embedded[C]. In Power & Energy Society General Meeting, 2009. PES '09.IEEE[C], 26-30 July 2009, 2009; 1-8.

[13] J. Zhen,Qiming Z.,Jianghai L., et al. A novel mathematical morphology filter for the accurate fault location in power transmission lines[C]. In TENCON 2009 - 2009 IEEE Region 10Conference[C], 23-26 Jan. 2009, 2009; 1-6.

[14] J. F. Zhang,Smith J. S.,Wu Q. H. Morphological undecimated wavelet decomposition for fault location on power transmission lines[J]. Circuits and Systems I: Regular Papers, IEEE Transactions on, 2006, 53 (6): 1395-1402.

[15] Tawfik M.M, Morcos M.M. ANN-based techniques for estimating fault location on transmission lines using Prony method[J]. IEEE Transactions on Power Delive ry, 2001, 16(2):219-224.