The feeder outage prediction model based on Optimization and Improvement of Random Forests Algorithm

Jiajia Zhu, Lanbo Zhou, Liming Lu*
State Grid Shanghai Shibei Electric Power Company, Shanghai, China

*Corresponding author e-mail: zhao_rb@shphqs.com

Abstract. The accurate prediction of feeder monthly fault level is the key to distribution network operation and maintenance. Aiming at the problems of low prediction accuracy, a method based on improved random forest to predict feeder monthly fault level is presented. Compared with the similarity calculation method of original random forest sample, this one increases the measurement of leaf node path distance, and applies the improved sample similarity to classification problem. Through the experimental comparison in the fault database of distribution network, the improved method can achieve a better classification effect than the original one, thus proving the effectiveness of the improved method.

1. Introduction
At present, various data mining algorithms have been applied to fault prediction of distribution network. The application scope and advantages of the algorithm are different. How to get a good use to the advantages of the algorithm while avoiding the disadvantages of the algorithm is critical [1]. In recent years, the predictive modelling algorithm of data mining based on machine learning theory has emerged in this field. There are three typical representatives, respectively; Artificial neural network (ANN), support vector machine (SVM) and decision tree algorithm. ANN is a learning algorithm based on the principle of empirical risk minimization, but it has some problems such as over fitting, large generalization error, and difficulty in determining the number of hidden elements. Compared with ANN algorithm, SVM can obtain better generalization performance, but there are also some problems that kernel function can only be selected by experience, and kernel parameters and penalty parameters are difficult to determine[2,3]. Decision tree algorithm has higher precision and antinomies ability, and by integrating multiple decision tree method to gather the random forest algorithm (RFA) relative to the traditional decision tree algorithm has better generalization ability and classification effect, and can avert over fitting which has generalization error limit and few parameters adjustment. Therefore, many scholars choose random forest algorithm to construct feeder fault prediction model in order to forecast feeder monthly fault level. The prediction results of random forest algorithm and three classification algorithms are compared and analyzed[4,5].

The higher the classification accuracy of any algorithm is, the better it will be. Therefore, the improvement of classification accuracy is an eternal topic in the research of classification algorithm optimization. Although many scholars have conducted extensive research on random forest and obtained many significant research results, random forest still has some limitations and shortcomings and has some room for improvement. In this paper, the method of calculating the similarity of random forest samples is improved, and the measurement of leaf node path distance is increased to promote the accuracy of random forest classification algorithm.
2. Improvement model of similarity calculation of random forest samples

2.1. Sample similarity matrix

An important advantage of random forest over other classifiers is that random forest can calculate the similarity between samples and obtain the similarity matrix between samples. We can measure the similarity between the two samples, or the probability that the two samples belong to the same category, by the frequency with which they occur on the same leaf node in each tree. Assuming that the number of samples is \( N \), the calculation process of the similarity matrix is as follows: first, initialize the sample similarity matrix \( \text{PROXI} \) as the all-zero matrix of \( N \) rows and \( N \) columns. All samples are identified by each tree generated already, and each sample falls on a leaf node of the tree. For samples \( i \) and \( j \), if they both fall on the same leaf node of the tree, add 1 to the corresponding row \( i \) and column \( j \) of the \( \text{PROXI} \) matrix. Repeat this process for each tree in the forest, traversing each tree to get a total sum. And then divide each element in the \( \text{PROXI} \) matrix by the total number of trees, and you will get the final \( \text{PROXI} \) matrix. The \( \text{PROXI} \) matrix is a symmetric matrix of \( N \) rows and \( N \) columns, with all diagonal elements 1. The element \( \text{PROXI}(i, j) \) of row \( i \) and column \( j \) is defined as the similarity between sample \( i \) and sample \( j \). The larger the value of \( \text{PROXI}(i, j) \), the higher the similarity between samples \( i \) and \( j \)[6,7].

2.2. Improvement of sample similarity calculation

Compared with other common learning algorithms, such as support vector machines, neural networks, decision trees, etc., random forest can calculate the similarity between training samples[8]. The random forest measures the similarity between the two samples by the frequency with which they appear on the same leaf node of each tree.

When calculating the similarity between two samples, random forest discriminates all samples with each generated tree, and each sample will fall on a leaf node of the tree. If two samples fall on the same leaf node of the tree, add 1 to the corresponding position of the similarity matrix. If the leaves fall on different nodes, add nothing. This kind of calculation ignores the relationship between the path distances among leaf nodes to the similarity degree between samples, which is easy to cause that one size fits all. When calculating the similarity degree of samples, losses will be caused, and the similarity degree between samples cannot be measured comprehensively and completely. Such as:

![Diagram of Node Relationships](image)

Figure 1. The case being considering the similarity of adjacent nodes.

Firstly, explain the concept of path distance between nodes. The path distance of two nodes is the sum of the path lengths from the nearest common ancestor of theirs to these two nodes. As shown in Figure 1, the nearest common ancestor of node 6 and node 9 is node 1, and the distance from node 1 to node 6 and node 9 is 2 and 3, respectively. Therefore, the path distance between node 6 and node 9 is 2 plus 3, which equals 5. The closer the path distance between two nodes is, the higher the probability that they split from the same branch is, the closer the distribution of samples on leaf nodes is in some attribute values, and the higher the sample similarity is.
In the random forest algorithm to calculate the sample similarity, only consider samples of the same leaf node in a tree, to fall on the degree of similarity between samples of different leaf nodes is unified as 0, such a one-size-fits-all solution not completely give the similarity relations of sample, which causes the loss of the sample similarity measurement to a certain extent. For example, in Figure 1, for samples 1 and 3, consider two cases where they fall on different leaf nodes of the tree. Case 1: sample 1 and 3 fall on node 5 and 9 respectively; Case 2: example 1 and 3 fall on nodes 6 and 10 respectively. In the above two cases, the leaf node classification standards of the samples should be consistent. The paths of nodes 5 and 9 are closer than those of nodes 6 and 10. Nodes 5 and 9 have two common ancestors, while nodes 6 and 10 have only one, node 1. From the whole tree, samples on nodes 5 and 9 have more identical distribution of attribute values than those on nodes 6 and 10, so samples on nodes 5 and 9 should have higher similarity than samples on nodes 6 and 10.

However, when the sample similarity is calculated by random forest, the treatment of the above two cases is the same without any distinction, which may cause losses to the correct calculation of sample similarity to some extent. We proposed an improved algorithm for the similarity matrix of random forest samples, taking into account the different leaf nodes. The calculation formula for the similarity between samples is as follows:

$$PROXI(i, j) = PROXI(i, j) + \frac{C}{e^n - e^{-i \alpha}}$$

Where $i, j$ is the sample number, $l$ is the path length between leaf nodes, and $n$ is any positive real number. $C$ is constant, and $e$ (2.71828…) is the Napierian base.

Input: training set $S=\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, random forest $H_{pre}=[h_1, h_2, h_3, \ldots, h_{\text{move}}]$.

Output: similarity matrix $PROXI_{pre}$.

For $i = 1$ to $n$

For $j = 1$ to $n$

For $k = 1$ to $n_{pre}$

If the sample $i, j$ is the same leaf node that falls on the tree $h_k$

$PROXI(i, j) = PROXI(i, j) + \frac{C}{e^n - e^{-i \alpha}}$

Else if the sample $i, j$ is the same classification standard as the different leaf nodes on the tree, calculate the leaf node path length $l$. $PROXI(i, j) = PROXI(i, j) + \frac{C}{e^n - e^{-i \alpha}}$

End if

End for

End for

End for

Obviously, as the parameter $n$ increases, the sample similarity measurement will get closer and closer to the similarity of the original random forest, especially when $n \rightarrow +\infty$, it is the original sample similarity. Therefore, from this perspective, the sample similarity measurement of this study is a generalization of the original random forest similarity.

3. Application results and discussion

The general k-nearest neighbour algorithm uses Euclidean distance as the similarity measurement between samples. When labelling the class standard of a test sample, it adopts the class standard of K training samples with the closest Euclidean distance to the sample to conduct the labelling by simple majority voting. In order to test the effect of the improved random forest sample similarity matrix measurement, the pre-improved and post-improved sample similarity measures were applied as the similarity measurement of the k-nearest neighbour algorithm, and their classification performance was compared[9,10]. The specific algorithm process is described as follows:

1. Build a random forest $H_{pre}=[h_1, h_2, h_3, \ldots, h_{\text{move}}]$ on the training set $S$. 

3
2. All samples on S and sample $X_i$ to be tested were put into forest $H_{tree}$ to calculate the similarity between $X_i$ and each sample in S.

3. Calculate the average similarity between $X_i$ and samples of each category according to the categories of samples in S.

4. Select the category with the maximum average similarity as the class standard of sample $X_i$ to be tested.

3.1. Experimental approach and scheme setting

The validity of the improved sample similarity is verified by data experiments and compared with the original similarity calculation method of random forest. In the comparison experiment of classification based on sample similarity, the difference between the two is mainly in the calculation of sample similarity in the second step of the above algorithm. Data needed for feeder fault prediction were obtained as sample data by means of data preprocessing, data exploration and analysis, and feature selection. Data from 168 feeders for two years were selected as training sample data. The data of the following six months were taken as the test sample data; The next three months were used as the prediction samples.

This study was organized into 4 data sets for comparison experiment. The data set information is shown in Table 1. To ensure the stability of experimental results, we conducted a total of 30 randomized trials for each data set. In each randomized experiment, 70% of the data set was stochastically selected as the training sample and 30% as the test sample. All experiments are run in the Matlab environment.

| DataSet         | Number of samples | Number of attributes | The number of categories |
|-----------------|-------------------|----------------------|-------------------------|
| Fault characteristics | 986               | 4                    | 3                       |
| External factors    | 1126              | 5                    | 3                       |
| Own factors         | 2789              | 6                    | 3                       |
| Running factors     | 1638              | 3                    | 3                       |

3.2. Experimental results and analysis

Figure 2.-Figure 5. show the comparison of the mean test accuracy of two algorithms, PROXI-based classification and PROXI-improved classification in four data sets. The vertical axis shows the average test accuracy, while the horizontal axis shows the forests of different sizes.

![Figure 2. Curve of fault characteristics.](image1)

![Figure 3. Curve of external factors.](image2)
From the experimental results of the above comparison, we can see that, except for the similar performance of Figure 3 and the classification results based on PROXI, the improved classification one performed better than the original classification in all other data sets under different forest sizes. Especially in Figure 2 and Figure 5 data sets, improved PROXI-classification can achieve better results than based classification under different forest sizes, and the average test accuracy can be improved by 8.4%, 3.1%, 3.9% and 5.2%, respectively. In all data sets, when the forest size is around 250, the improved classification prediction accuracy is similar or slightly better than the original classification. When the forest size is around 200, the improved classification shows significantly better classification prediction accuracy.

4. Conclusion

In this paper, an improved algorithm of random forest is proposed. Compared with the original random forest sample similarity calculation method, the measurement of leaf node path distance is increased. In the case that samples fall on different leaf nodes and the class standards of leaf nodes are the same, this paper believes that there is a certain degree of similarity between them, and the closer the path distance between leaf nodes is, the higher the similarity is. Then the improved sample similarity is applied to fault classification prediction and compared with the original method of similarity calculation in random forest. Experimental results on multiple data sets show that the improved method can achieve better results than the original method, which proves the effectiveness of the improved one.

References

[1] J. R. Quinlan, Induction of decision trees, Machine learning. 1986, 1(1) 81-106.
[2] S. L. Salzberg, C4.5: programs for machine learning, Machine Learning. 1994, 16(3) 235.
[3] J. H. Chen, H. M. Chen, S. Y. Ho, Design of nearest neighbour classifiers: multi-objective approach, International Journal of Approximate Reasoning. 2005, 40(1) 3-22.
[4] W. Lin, Z. Wu, L. Lin, et al., An ensemble random forest algorithm for insurance big data analysis, IEEE Access. 5(2017) 16568-16575.
[5] S. Ayesha, T. Mustafa, A. Sattar, I. Khan, Data Mining Model for Higher Education System, European Journal of Scientific Research. 2018, 43(1)24-29.
[6] N. S. Ahmed, M. H. Sadiq, Clarify of the Random Forest Algorithm in an Educational Field, 2018 International Conference on Advanced Science and Engineering. IEEE, 2018, pp. 179-184.
[7] E. A. Amr, H. Hamtini, I. Aljarah, Mining Educational Data to Predict Student's academic Performance using Ensemble Methods, International Journal of Database Theory and Application. 2018, 9(8)119-136.
[8] M. Zhu, J. Xia, X. Jin, et al., Class weights random forest algorithm for processing class imbalanced medical data, IEEE Access. 6(2018) 4641-4652.
[9] W. W. Tao, S. Liu, Y. Su, et al., A power network illegal access monitoring method based on random forests, Advances in Intelligent Systems and Computing. 905(2020)894-903.
[10] G. O. Kaya, A hybrid method based on empirical mode decomposition and random forest regression for wind power forecasting, Journal of Multiple-Valued Logic and Soft Computing. 2018, 31(1-2)123-137.