Decision Tree Approach for Fault Type Identification of Transmission Line

Wancai Zhang 1, *, Yi Wang 1, Xingxun Wang 1, Jianyi Wang 2, Kun Zhang 1 and Xuri Sun 3

1 Beijing State Grid Fuda Science & Technology Development Co. Ltd., Beijing 100070 China;
2 China Electric Power Research Institute, Haidian District, Beijing 100192, China;
3 Qingdao Power Supply Company of State Grid Shandong Electric Power Company, Qingdao, 266000, China.

* Corresponding author e-mail: lyjtxb@163.com

Abstract. In order to solve the limitation of artificial identification of fault causes and insufficient application of artificial intelligence due to fewer data samples in transmission line faults, a decision tree-based model and method for fault cause identification of transmission lines are studied. Expert support and manual intervention are introduced to verify and confirm the diagnosis conclusions on the basis of automatic diagnosis. It is recognized that the automation of line fault diagnosis and the visualization of the whole process of fault diagnosis can be realized, thus reducing the workload of operation and maintenance staff. Based on the defect data of power company operation and maintenance, the validity and validity of the method are analyzed and verified, which can provide theoretical basis for equipment operation and maintenance decision.

1. Introduction
Fault type identification of transmission lines is of great significance to power system fault analysis, fault location and relay protection. At present, the common fault types of transmission lines mainly include lightning strike, icing, external force damage, pollution flashover and so on. The development of power Internet of things technology improves the depth and breadth of information perception in each link of power grid [1-2], and the data are multi-source and heterogeneous. Traditionally, the main method of line fault detection is manual inspection. When tripping occurs, the information is scattered and can not be collected in time. A lot of alarm information or abnormal information is difficult to determine which factor or multi-factor coupling effect caused the accident. The speed and correctness of fault diagnosis based on the experience of operators are limited.

In recent years, the application of artificial intelligence algorithm to fault identification of transmission lines has become a research hotspot of researchers at home and abroad. The model of fault identification can be trained independently by machine learning [3-4]. However, due to the small number of fault samples and fewer samples reflecting the change of data in the process of fault occurrence, it is difficult to use a small number of samples to train in-depth learning model, which restricts the
application of artificial intelligence technology [5] in condition evaluation and fault diagnosis of transmission line equipment.

By analyzing the logical relationship between fault samples of transmission line operation and maintenance, this paper establishes different fault classification models by using decision tree, matches the optimal prediction method according to real-time data characteristics, and makes prediction. It provides a technical means to make discrimination by learning expert experience through artificial intelligence.

2. System Overall Framework

In this paper, anomaly diagnosis model is established by deep mining and analysis of feature attributes. Based on the sample data to be processed, the diagnostic model should not only ensure accuracy, but also have short data processing time and less memory. Therefore, the key to construct the diagnostic model is to select a reasonable algorithm.

Based on the above requirements, this paper adopts decision tree algorithm to realize the identification of transmission line fault types. Decision tree is a traditional modeling method of data mining, which has been applied in many fields and achieved good results. For example, in reference [6], a two-step algorithm, Pat HT, is proposed to generate decision tree for variable data stream classification. This method can significantly improve the accuracy or reduce the training time when processing steady data stream, and has a good classification effect [6].

Because the decision tree algorithm has a good classification effect, the algorithm is relatively simple and easy to implement. It does not need equipment investment and reduces the cost of system construction. Therefore, this paper uses decision tree algorithm to generate system anomaly diagnosis model. The structure of anomaly diagnosis system is shown in Figure 1. Different decision tree group anomaly diagnosis models are established through model training of sample data. Then Hadoop data processing cluster uses this model to distinguish anomaly diagnosis and presents the diagnosis information to the analysis center. Maintenance personnel adjust the diagnosis model by returning the prediction accuracy and adding new abnormal diagnosis status to realize efficient identification of transmission line fault types.

3. Design of Decision Tree Approach for Fault Type Identification of Transmission Line

3.1. Segmentation strategy of feature attributes

The key of building anomaly diagnosis model of decision tree group is the problem of node splitting, that is, selecting appropriate feature attributes to classify data sets. In view of the power consumption data provided by the power consumption information acquisition system, the feature attributes involved contain a large number of continuous features, so the information gain rate is used to select the segmentation features. Assuming that the number of abnormal operation watt-hour meters and normal operation watt-hour meters in the training set are p and N respectively, a decision tree can usually classify a class of abnormal situations with the required amount of information:

$$I(p, n) = - \frac{p}{p + n} \log_2 \frac{p}{p + n} - \frac{n}{p + n} \log_2 \frac{n}{p + n}$$  \hspace{1cm} (1)

If the root of the decision tree is the classification A of different voltage levels, A has V values ($v_1, v_2, \ldots, v_v$). The training set is divided into v subsets ($H_1, H_2, \ldots, H_v$). Suppose that the subset $H_i$ contains $P_i$ fault transmission lines and $N_i$ fault-free transmission lines. The information entropy $E(H_i)$ of the subset $H_i$ is:

$$E(H_i) = - \frac{P_i}{P_i + N_i} \log_2 \frac{P_i}{P_i + N_i} - \frac{N_i}{P_i + N_i} \log_2 \frac{N_i}{P_i + N_i}$$ \hspace{1cm} (2)

The information entropy based on attribute A is as follows:

$$E(A) = \sum_{i=1}^{v} \frac{P_i + N_i}{P + N} E(H_i)$$ \hspace{1cm} (3)
So the information gain based on A is as follows:

\[ \text{Gain}(A) = I(p, n) - E(A) \]  

The information gain rate is:

\[ \text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{Split}(A)} \]  

Among them, Split (A) is:

\[ \text{Split}(A) = -\sum_{i=1}^{n} E(H_i) \]

The decision tree traverses the information gain rate of all the feature attributes mentioned above, chooses the feature attributes that maximize Gain-Ratio (A) as the root node, and recursively calls the above process to the subset corresponding to the different values of the root node to generate the sub-nodes of the decision tree.

Taking the years of operation of continuous feature attributes and the weather of discrete feature attributes as examples, the following process is adopted to realize feature attribute segmentation:

1) Calculate the information gain rate of the continuous feature years. For the continuous feature attributes in a larger numerical range, this paper calculates the information gain of the continuous feature attributes by piecewise multi-splitting points, and divides the information gain rate equally, and then obtains the information gain rate. Because there are fewer values in this example, only one is the best splitting point.

2) The midpoint between normal and fault types corresponding to the number of years of operation is taken as possible splitting points (i.e. 13 and 14.5), which divides the data set into four parts and calculates the information gain of each possible splitting point.

3) As shown in Table 1, the optimal splitting point presented in this table is 14.5, and the information gain of formula (1) ~ formula (5) is 0.318 with 13 as the splitting point and 0.459 with 14.5 as the splitting point.

4) The information gain of each splitting point is revised, i.e. subtracted, where \( N \) is the number of possible splitting points, i.e. 2, \( |D| \) is the size of the data set, i.e. 6, so the revised value is 2.584. Because the two splitting points have the same correction value, the revised information gain of 13 as the splitting point is 2.902, and the revised information gain of 14.5 splitting point is 3.0. 43;

5) By comparison, the optimal split point of the operation year is 14.5, and the information gain rate of the optimal split point is calculated as the information gain rate of the operation year. According to formula (5), the information gain rate of the segmentation is 0.918, so the information gain rate of the consecutive feature attribute operation year is 3.314.

(2) Calculating Discrete Characteristic Weather

As shown in the table above, from Formula (1) ~ Formula (5), the amount of user classification information is 0.918, the information entropy is 0, the information gain is 0.918, and the segmentation information rate is 0, so the information gain rate is infinite.

(3) Comparing the continuous feature attribute with the information gain rate calculated by the discrete feature, the feature with the largest information gain rate is chosen as the split feature, and the weather is chosen as the root node by the comparison of the above values.

| number | years of operation | weather     | operation status |
|--------|-------------------|-------------|-----------------|
| 1      | 13                | Thunder storms | thunderstrike   |
| 2      | 13                | Cloudy      | normal          |
| 3      | 14                | moderate rain | thunderstrike   |
| 4      | 15                | Cloudy      | normal          |
| 5      | 15                | Cloudy      | normal          |
| 6      | 15                | Cloudy      | normal          |
The decision tree model is constructed by checking all the characteristic attributes based on the information gain rate, selecting the characteristic attributes with the greatest information gain rate to generate the decision tree nodes, establishing tree branches from the different values of the tree nodes, and recursively calling the above algorithm on the training subsets of each branch. With this method, the nodes and branches of the decision tree are established. Until the stop condition of decision tree generation is satisfied.

3.2. Pruning optimization strategy
After the decision tree is generated, many branches of the decision tree reflect the anomalies in the training set because of the noise in the collected data and the special conditions of transmission line operation and maintenance. Complexity pruning algorithm is used to prune the fully growing decision tree. By deleting the branches of nodes, unreliable branches are gradually pruned. Thus, fast classification can be achieved and the ability of correct selection of decision tree can be improved [8].

For the fully growing decision tree generated by the feature attribute segmentation strategy, the surface error rate gain of each subtree $T$ in the tree is calculated:

$$
\alpha = \frac{R(t) - R(T_t)}{|N_t| - 1} \quad (7)
$$

In the formula $|N_T|$ is the number of leaf nodes for the growth of a subtree, and $R(t)$ is the error cost of leaf nodes $t$, whose value is as follows:

$$
R(t) = R(t)p(t) \quad (8)
$$

$R(t)$ is the error rate of leaf node $t$, $P(t)$ is the proportion of data classified by leaf node $T$ to all data, and $R(T_t)$ is the error cost of subtree $T_t$, whose value is the sum of the error cost of all leaf nodes contained in subtree $T_t$.

Select the sub-tree with the smallest alpha value and prune the branches generated by it. When the alpha value of multiple sub-trees reaches the smallest at the same time, $|N_T|$ is the largest for pruning until the evaluation needs of the model are met.

3.3. Decision tree anomaly diagnosis model
Through the construction of decision tree, this section summarizes the anomaly diagnosis model of decision tree established in this paper. In each feature attribute segmentation and selection, greedy algorithm is adopted, that is, when each selection and classification is made, only the current interests are maximized, and the decision tree is generated from top to bottom by recursive method. The anomaly diagnosis of decision tree adopted in this paper is based on the recursive method. The model mainly diagnoses the types of transmission line faults. The specific steps are as follows:

(1) Determine the decision attribute categories and characteristic attributes of the decision tree for anomaly analysis, which starts with a single node;

(2) Preliminary processing of feature attributes is carried out so that the information gain of continuous feature attributes is calculated in segments, and the maximum value at $n$ is taken as the optimal segmentation point, so that the continuous feature attributes are discretized. Under different classifications, 8 optimal segmentation points are selected according to the aggregation of number.

(3) Entropy measure of information gain rate is used as heuristic information of feature attributes to calculate information gain rate of continuous attributes and discrete attributes.

(4) Choose the feature attributes with the largest information gain rate as the feature attributes of sample classification. The feature attributes become the decision attributes of nodes, create a branch for each known value of the decision attributes, and classify the training set accordingly.

(5) Determine whether the decision tree satisfies the following stopping growth conditions

1) Determine that all training sample subsets of nodes belong to the same class;
2) If there is no residual feature attribute to further classify the subset of training samples, the node is regarded as a leaf node, and the class with the largest subset of training samples is classified as the leaf node.

If the decision node does not satisfy any of the above conditions, the algorithm returns (1) to form the sample decision tree branches on each classification recursively from top to bottom. Once an attribute appears on a node, the sub-node of the node eliminates this characteristic attribute until all nodes satisfy the above stop growth conditions.

(6) Based on cost complexity pruning algorithm, the fully growing decision tree is pruned and optimized, and the sub-trees with the lowest surface error rate gain are pruned in turn.

(7) After the pruned tree is generated, an independent test set is used to evaluate the accuracy of the decision tree. If there are still more noise data, step 6 is returned. Finally, a set of pruned decision trees satisfying the evaluation needs of the model is obtained.

3.4. Simulation

The annual transmission line operation and maintenance fault data of a regional power company are analyzed. The sample information includes line name, cause classification, voltage level, operation life, maximum temperature, minimum temperature, wind direction, wind power and weather. There are 4262 fault data samples. Data are stored in databases according to their different categories. 70% of the sample data are randomly extracted from the training set and the remaining 30% of the samples are test sets. The training set data were randomly divided into 10 copies. In addition, the overall training set was 1 copy. A decision tree group anomaly diagnosis model with 6 decision trees was established for different types of anomaly diagnosis. All the experimental results showed that the majority of voters won.

![Figure 1. Single Tree Test Results in Decision Tree Groups](image)

As can be seen from Fig. 1, in the decision tree group, the correct rate is about 62%-80%, and the recall rate is about 70%-90%. According to the test results, the results of fault cause discrimination test for transmission line based on decision tree are as follows:

$$\text{accuracy} = \frac{9 + 1000}{1242} = 81.23\%$$

$$\text{recall rate} = \frac{9}{9 + 3} = 75\%$$

From the above data, the actual operation status of the watt-hour meter can be effectively predicted based on the cause discrimination of transmission line faults based on decision tree. The model can be based on the relevant data of transmission line operation and maintenance provided by existing platforms, and relies on parallel and distributed computation of large data for data processing, so as to satisfy the cause discrimination of transmission line faults.
4. Conclusion
By integrating multi-source information such as online monitoring, ledger information and historical fault database, this paper studies and constructs the transmission line fault cause discrimination model and method based on decision tree, so as to realize the automation of line fault diagnosis and visualization of the whole process of fault diagnosis, thus reducing the workload of operation and maintenance staff. This model system has been used in a provincial company, and will provide powerful technical support for transmission line fault diagnosis through more data accumulation and case diagnosis in the future.

Acknowledgments
This work was supported by the science and technology projects of State Grid Corporation of China (GYB17201700204 Fault Diagnosis and Maintenance of Electrical Equipment based on Big Data Technology).

References
[1] WANG Si-qi. Research on smart grid monitoring system based on Internet of Things[J]. Chinese Journal of Power Sources, 2018,42(01):125-127.
[2] LÜ Jun, LUAN Wenpeng, LIU Rilian , et al. Architecture of Distribution Internet of Things Based on Widespread Sensing & Software Defined Technology[J]. Power System Technology, 2018,42(10):3108-3115.
[3] XU Shuwei, QIU Caiming, ZHANG Dongxia, et al. A Deep Learning Approach for Fault Type Identification of Transmission Line. Proceeding of the CSEE, 2019, 39(1): 65-74.
[4] ZHOU Zandong,GONG Hao,WU Nian.Fault Diagnosis of Transmission Line Icing Based on Decision Tree[J].High Voltage Apparatus,2019,55(01):165-169
[5] DAI Yan, WANG Liuwang, LI Yuan, et al. A Brief Survey on Applications of New Generation Artificial Intelligence in Smart Grids[J]. Electric Power Construction, 2018, 39(10): 1-11.
[6] Han Meng, Wang Zihai, Ding Jian. Efficient decision tree for evolving data streams based on frequent patterns[J]. Chinese Journal of Computers, 2016, 39(8): 1541-1554.
[7] Liu Yuxun, Xie Niuniu. Improved ID3 algorithm. Proceedings of 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT 2010) . 2010.
[8] Mohammad Reza Aghamohammadi, Mohammad Reza Salimian. A three Stages Decision Tree Based Intelligent Blackout Predictor for Power Systems using Brittleness Indices[J]. IEEE Transactions on Smart Grid. 2017, (99): 1-1.