Study on Line Loss Status Classification Based on Decision Tree

Tan Haibo1,2*, Luo Bingxiang 1,2, Liu Mouhai1,2, Guo Guang1,2 and Xie Xiong1,2

1State Grid Hunan Electric Power Limited Company, Changsha, Hunan, 410004, China
2Hunan Province Key Laboratory of Intelligent Electrical Measurement and Application Technology, Changsha, Hunan, 410116, China
*Corresponding author’s e-mail: 1376466188@qq.com

Abstract. Line loss rate is the ratio of power loss to power supply in a certain period of time, and it is an important index to assess the technical economy of power grid. In recent years, with the increasing scale of power grid construction and the aging of power lines, the traditional monitoring methods with simple logic judgment can not meet the accurate judgment status of line loss, so it is urgent to effectively improve the lean management level of transmission and transformation line loss abnormality. In order to realize the automatic classification of line loss anomalies machine learning algorithm is proposed, the paper takes 1500 transmission line loss data as training samples, builds a line loss anomaly classification model based on decision tree, and realizes the accurate classification of line loss anomaly causes. Meanwhile, the paper uses KNN, support vector machine and decision tree algorithm to identify the anomaly for comparing the accuracy of different machine learning algorithms, and uses MATLAB software to verify the algorithm. The experimental results show that the comprehensive the recognition accuracy of decision tree for line loss anomaly is the highest, reaching 88.9%, 60% for support vector machine and 82% for KNN. This is of great significance to the realization of lean line loss management and has a high engineering value.

1. Introduction
Line loss refers to the active power consumption during electric transportation and using due to resistance and conductance [1]-[3], it is the most important index of power grid. For a long time, line loss has being researched by domestic and foreign scholars and achieved many results. The line loss is divided into five categories: management line loss, theoretical line loss, statistical line loss, economic line loss and fixed line loss, and related loss reduction measures have been taken from the perspective of line loss management and technology [4]-[7]. With the development of informatization and big data, many scholars use new technology to reduce line loss. A scholar uses GPRS technology to establish a line loss monitoring system for low-voltage power distribution networks, it realizes the remote monitoring of line loss and improves line loss management [8]. Some research collects the electric energy meter data from time to time through the concentrator, establishes the online real-time monitoring and analysis system of the distribution network, analyzes the collected data to monitor real-time line loss and propose targeted loss reduction measures [9-10]. Some research improves the power factor and the terminal voltage by reactive power compensation device, and it can reduce line loss [11-12]. And a new line loss reduction method is proposed, it uses big data technology to accurately analyzes the line loss and obtain an accurate model for calculating, this strategy reduces the
calculation error and accurately reflects the actual line loss[13]-[16]. In the above papers, they mainly carry out research on line loss reduction strategies, but ignore the classification of abnormal line loss. Accurate classification of abnormal line loss is a prerequisite for effective loss reduction. It is necessary to research on the classification of abnormal line loss.

In this paper, a line loss anomaly recognition model based on machine learning algorithms is established by training through a large amount of historical electricity data at the gate. It can determine the line loss status automatically and warn of line loss efficiently. It solves the problem of long-term manual monitoring of line loss.

2. Line Loss Data Pre-processing
The difference electric quantity between energy metering’s selling side and generating side can reflect the line loss accurately and objectively. This article selects 1500 gateway measurement points line loss data from the marketing system of a network province company. Each set of data has 6 kinds of electrical information including selling side and generating side voltage, current, power factor, transformer transformation ratio and two types of electricity data on selling side and generating side electric quantity and the electric meter.

2.1. Data Preprocessing
Data pre-processing is an important part of machine learning. It includes data cleaning, aggregation, feature creation, discretization, missing and outlier processing, etc. In the research, different data have different characteristics, so there will be different ways to pre-process the data.

(1) Data cleaning. As shown in Figure 1, the energy metering data of substation selected in this paper may be abnormal, missing, duplicate due to statistics, system errors and other reasons. In order to improve the prediction accuracy of the machine learning model, removing these abnormal data.

![Figure 1. Hybrid metering system configuration scheme](image)

(2) Unbalance treatment. In the training data, the proportion of abnormal data is very small, most of them are normal data. In order not to affect the training results, this paper uses some abnormal data features to generate a part of abnormal data to ensure that the amount of abnormal data is similar to the amount of normal data.

(3) Feature selection. The line loss rate of each line and the energy meter measurement are obtained by deriving and sorting out two types of electricity data, the energy metering device records, the selling side and generating side voltages, currents, power factors, transformer transformation ratios, electric quantity, and meter start and end codes.
3. Line Loss Anomaly Classification and Recognition Model

Machine learning algorithms are divided into two types: supervised learning and unsupervised learning. Supervised learning is mainly used when the results of training data are known, it includes nearest neighbours, support vector machines, decision trees, and integrated learning algorithms. Unsupervised learning is used when there is no data result label. It includes K-means, Gaussian mixture and other algorithms. The training data in this paper has data labels, so this paper uses a supervised learning algorithm.

3.1. Classifier selection

At present, the mature classification algorithms in supervised learning mainly include support vector machine, decision tree, nearest neighbour, naive Bayes, etc[17]. As shown in Table 1, in order to compare the training effects of different algorithms, this paper builds training models for different classification algorithms in MATLAB, and carried out a comparative analysis from the two dimensions of learning time and the prediction accuracy.

| Algorithm                           | Prediction Accuracy(%) | Learning Time(s) |
|-------------------------------------|------------------------|------------------|
| Nearest Neighbor                    | 83.7                   | 1.263            |
| Decision Tree                       | 88.6                   | 0.473            |
| Support Vector Machine              | 78.6                   | 1.537            |
| Integrated Learning (bag-packed, promoted) | 88.7               | 34.182           |

According to Table 1, the prediction accuracy of the decision tree reached 89.9%, which is significantly higher than the other three classification algorithms. The prediction accuracy of integrated learning is not far from that of the decision tree, but the learning time is much higher than that of the decision tree. The AUC values of various machine learning models are shown in Fig. 2. The comprehensive analysis of this paper selects decision trees as the classification algorithm.

As shown in Fig. 3, it is a frame diagram of line loss anomaly and recognition model based on decision tree.
4. Decision Tree Classification Algorithm Introduction and Model Building

4.1. Decision Tree Classification Algorithm

At present, there are three commonly used decision tree algorithms: ID3, C4.5, and CART. In 1993, J.R. Quinlan introduced the concept of information entropy in information theory into the decision tree algorithm, proposed the ID3 (Iterative Dichotomizer3) algorithm. In 1984, L Breiman, J Fried Man proposed the CART classification algorithm, this algorithm uses the Gini index as a test attribute, and finally generates a binary tree form. In 1994, in order to improve the deficiencies of the ID3 algorithm itself, J.R. Quinlan proposed the C4.5 algorithm.

The ID3, C4.5 and CART algorithms all have their own characteristics and advantages. Their difference is shown as Table 2. The C4.5 algorithm is an improved algorithm of ID3, it not only can avoid the tendency to choose the attribute with the most value when the information gain is used in the ID3 algorithm to select attributes, but also improve the algorithm of decision tree pruning. The C4.5 algorithm can be pruned in the process of building a tree. This paper chooses C4.5 as the decision tree algorithm, because the data studied contains continuous values and consider the decision tree pruning function.

| Algorithms | Support model  | Tree structure | Feature selection | Continuous value processing | Pruning |
|------------|----------------|----------------|-------------------|----------------------------|---------|
| ID3        | classification | Polytree       | Information gain  | Not support                | Not support |
| C4.5       | classification | Polytree       | Information gain ratio | support                 | support |
| CART       | Classification, return | Binary tree | Gini Coefficient | support                 | support |

4.2. Model of Decision Tree

This paper selects the C4.5 algorithm, its establishment process is shown in Fig. 3, establishing process node feature selection (information gain), decision tree construction, decision tree pruning (anti-overfitting). And use information gain rate to judge the feature selection of C4.5 algorithm, the calculation formula is as formula (1), (2), (3).

Information entropy:

\[ \text{Ent}(D) = - \sum_{k=1}^{N} p(k) \log_2 p(k) \]  

(1)

In the formula (1), \( p_k \) represents the probability of selecting the category in the current sample set \( D \), and \( N \) is the number of categories.

Information gain:

\[ \text{Gain}(D, a) = \text{Ent}(D) - \sum_{v \in \mathcal{V}} \left| \frac{D^v}{|D|} \right| \text{Ent}(D^v) \]  

(2)

In the formula (2), \( v \) represents the value of discrete attribute \( a \): use attribute \( a \) to divide the set \( D \), then \( v \) branch nodes will be generated, among them, the \( v \)-th branch node contains all samples in \( D \) that have values on attribute \( a \), and this sample is denoted as \( D^v \).

Information gain rate:

\[ \text{Gainratio}(A) = \frac{\text{Gain}(A)}{\text{Gain}(D, a)} \]  

(3)


4.3. Building a decision tree model with MATLAB

MATLAB includes a machine learning toolbox, it contains a wealth of machine learning algorithms. MATLAB is convenient to use and has a friendly interface.

After data pre-processing, import the 1500 groups of line loss data introduced in Chapter I into the machine learning toolbox as training data. And divide 1500 groups of data into test group and training group, the test group accounts for 10% of the total data, test group is used to cross-validate the trained model. Each group of data output has three attribute labels: abnormal line loss, normal line loss and low load, which are represented by ‘B’, ‘G’, and ‘L’ respectively. After importing the data, choose decision tree algorithm to train, the tree shape of the decision tree is shown in Fig 5.

![Decision Tree Diagram of Abnormal Line Loss](image)

In Fig. 5, there are 18 hierarchical decision trees. In order to further improve the accuracy of the prediction model and prevent over-fitting, the pruning process is performed to obtain the pruned decision tree as shown in Fig 6.
5. Analysis of Line Loss Abnormal Recognition Results by Decision Tree Model

According to the decision tree model built in chapter IV, the line loss analysis results for 1500 groups data are shown in Fig 7, Fig 8, and Fig 9.

Figure 6. Decision tree diagram of abnormal line loss after pruning

Figure 7. Scatter plot of forecast results

Figure 8. Model confusion matrix
Fig. 7 is a scatter plot of line loss analysis results. The wrong analysis is the "×" symbol, and the correct one is the coloured 'ꞏ' symbol. Obviously, the result is less crossed, that means the training classification result is correct. As shown in Fig. 8 that the model's classification accuracy rates for low line loss, normal, and abnormal load are 92%, 88%, and 84%, the comprehensive prediction is 88% correct. In the ROC curve in Fig. 9, larger area under the curve means better model performance. That is, the larger AUC (area under curve) value, the more accurate the line loss analysis, as shown in Fig. 9, the AUC value of the model reached 0.93, indicating that the model is accurate for line loss analysis.

In order to verify the accuracy of training model for line loss analysis, this paper extracts 200 groups of data with known results, export the trained model to Workspace, predict new ‘unknown’ data. Use two different algorithms, KNN and SVM to predict, and the results are compared with the known results, as shown in Table 3.

| No. | Training Data | True Result | Forecast Result |
|-----|---------------|-------------|-----------------|
|     | X₁           | X₂          | Tree KNN SVM    |
| 1   | 0.2044       | -0.0072     | g L × L × L ×   |
| 2   | 163.0150     | -0.3712     | b b √ b √ b √   |
| 3   | 1.7773       | 0.0054      | g g √ g √ g √   |
| 4   | 25.7000      | 0.0039      | g g √ g √ g √   |
| 5   | 0.9189       | 0.0096      | g g √ g √ g √   |
| 6   | 25.5200      | 0.0043      | g g √ g √ g √   |
| 7   | 21.9000      | 0.0098      | g g √ g √ g √   |
| 8   | 0.2154       | 0.0025      | L L √ g × L √   |
| 196 | 45.2400      | -1.0000     | b b √ b √ g ×   |
| 197 | 30.2500      | 0.7558      | b b √ b √ b √   |
| 198 | 22.7600      | 0.0087      | g g √ g √ g √   |
| 199 | 1.2272       | 0.0046      | g g √ g √ g √   |
| 200 | 16.1580      | 0.0914      | b b √ b √ g ×   |

Total errors: 6 18 68

Note: The correct prediction results in the table are indicated by "√", and the prediction errors are indicated by "×".

It can be seen from the table that when the model is used to predict new data, the number of prediction errors of the decision tree is only 6, which is the least of other models. The correct rates of Tree, KNN, and SVM are 97%, 91%, and 66% respectively. It can be seen that the prediction rate of...
the decision tree model is the highest, which further confirms the correctness of the above experimental results.

6. Conclusion
This paper aims at the problems of high repetition rate and heavy workload of the current power grid line loss abnormal identification manually, introduces the decision tree classification algorithm into the line loss analysis and establishes a line loss abnormal recognition model based on the decision tree algorithm. The result shows that the line loss anomaly classification model of the decision tree based on historical data can accurately identify and classify line loss anomalies, with a prediction accuracy rate of 88%, it can provide early warning of line loss in time. And compared the prediction accuracy of the nearest leading algorithm, support vector machine algorithm, and bag integration algorithm, this demonstrates the correctness and feasibility of the decision tree model. This method will greatly improve the line loss management of State Grid, and has important practical application value.

References
[1] Fu S.H. (2019) Analysis of abnormal line loss in power supply station area of ×× power supply department and research on treatment method, Electrical Equipment and Economy, vol. 02, pp. 50-52.
[2] Tao J, Li Y, Ynag L (2017) Power grid line loss management and loss reduction measures,” China Power Enterprise Management, vol. 36, pp. 48-49.
[3] Chen Y.H, Chen F.J, Wang M.T. (2013) Analysis of Line Loss Causes and Research on Control Methods, Mechanical and Electrical Information, vol. 24, pp. 13-15.
[4] Yin W.J. (2019) Simple Discussion on the Line Loss Management of 10kV Distribution Network and the Loss Reduction Measures, Telecom World, vol. 26, pp. 125-126.
[5] Chen Y.F. (2019) Research on Line Loss Management and Loss Reduction Measures of Distribution Network, Telecom World, vol. 26, pp. 221-222.
[6] Tan M, (2019) Analysis of Line Loss Management and Loss Reduction Measures in 10kV Distribution Network, Shandong Industrial Technology, vol. 10, pp. 173.
[7] Zeng Z.X, (2017) Marketing Line Loss Management Research Based on Power Utilization Information Collection, North China Electric Power University.
[8] Li X, (2007) The Research and Design of Low Voltage Distribution Loss Measure System based on GPRS, Hunan University.
[9] Tantai X.H, (2017) Line Loss Analysis and Treatment of 10kV Distribution Network Based on Real-time Monitoring, North China Electric Power University.
[10] Wang P, Liu Y, Song C, (2019) Problems and Countermeasures of Line Loss Management in Tranformer Area, Electric Engineering, vol. 08, pp. 83-84+86.
[11] Guo X.Q, Guo D.Q, Zhang Y.X. Li W.D, Zhang M.P. (2014) Investigation on rating of shunt APF in Unified Power Quality Conditioner, Power Capacitor & Reactive Power Compensation, Chennai, vol. 35, pp. 43-49.
[12] Song P.X, Hao L.L, Lou B.L, Ma J.C, Huang H.Y, Lu C.Y, Wang C.M, (2017) Voltage Control Method Based on the Coordinated Optimization of Energy Saving and Loss Reduction in Distribution Network, Power Capacitor & Reactive Power Compensation, vol. 40, pp. 101.
[13] Huang W.S, Xu Y.B, Zhou B.P, Lu X, (2015) Research on line loss calculation analysis based on big data, Electrotechnical Application, vol. 34, pp. 79-84.
[14] Lv S.X, (2017) Estimation Model Simulation on the Loss-Reduction Effect of Smart Grid Based on Big Data Analysis, Electric Power, vol. 50, pp. 172-175.
[15] Xu F, Sun M.Z, Lin S, Cheng C. (2019) Synchronous Line Loss Management of Distribution Network Based on Power Big Data,” Anhui Electric Power, vol. 36, pp. 43-46.
[16] Zhou H.L. (2017) Research on Reactive Power Optimization and Loss Reduction of Network Based on Data Technology, Hunan University.
[17] Dan Y.F, Tao J.W, Xu H.T. (2020) Semi-Supervised Classification Method of Possibilistic Clustering Assumption, Computer Engineering and Applications, vol. 56, pp. 65-74.