Transmission Line Loss Prediction by Cross Validation and Gradient Boosting Decision Tree

Jicheng Yu¹, Feng Zhou¹, Kai Zhu¹, Changxi Yue¹⁺, Jiangchu Wang² and Congzhen Xie²

¹China Electric Power Research Institute, Wuhan 430074, China
²South China University of Technology, Guangzhou 510000, China

*Corresponding author e-mail: yuechangxi@epri.sgcc.com.cn

Abstract. To solve the problem of the accuracy and generalization of the transmission line loss prediction, a new method for transmission line loss prediction based on cross-validation (CV) and gradient boosting decision tree (GBDT) is proposed. In this method, time granularity matching and statistical feature extraction are firstly carried out to improve the information content of the data. Then the data is divided into a training set and test set by CV method to train the GBDT model. Finally, the actual transmission line loss verification test is conducted. The result shows that for 500kV transmission lines with an average daily supply of 7990MWh, the average line loss error of this model is 35MWh, and the average line loss rate error is 0.093%, which verifies the effectiveness of the method in this paper.

1. Introduction

The transmission line loss reflects the planning, production, and management level of the power grid and is an important criterion for assessing the power sector. However, the error of the theoretical transmission line loss prediction will cause the report to not accurately reflect the actual line loss, which will bring great problems to line loss management. With the advancement of line loss fine management, there is an urgent need for accurate transmission line loss prediction methods.

At present, the traditional methods used to predict transmission line loss mainly include maximum load loss time method, loss factor method, representative root-mean-square current method, electric quantity method, load curve characteristic coefficient method, equivalent resistance method, voltage loss method, and improved power flow method [1-6]. However, due to the difference in the grid structure among lines, the load variation of lines and the change of weather will affect the line loss in the same period. There is still room for improvement in the prediction accuracy of the traditional line loss prediction method that only considers the current, voltage, and other parameters, and the universality of different transmission lines remains to be studied.

Nowadays, the transmission line loss prediction method has been enriched and developed, and many new methods have appeared. For example, theoretical transmission line loss prediction based on interval mathematics, regression analysis method, neural network prediction method, prediction based on fuzzy recognition technology, based on improved core Vector machine calculations, etc. Based on the improved extreme learning machine algorithm, Liu et al. constructed the synchronous line loss prediction model [7]. The extreme learning machine model is optimized by ant colony algorithm to form
an accurate line loss fitting model. However, this method did not overcome the possible locality problems caused by the ant colony algorithm. Li et al. took into account the correlation knowledge between different line network structures, and used clustering algorithm to classify and extract line category characteristics and built BP neural network models according to different categories [8]. However, after clustering, the number of single model training data decreased, resulting in reduced generalization of the model. Zhang et al. selected some influential correlation factors through analytic hierarchy process (AHP) algorithm and adopts multi-gray model to fit the relationship between correlation factors and line loss rate [9]. However, AHP is a data fusion method with subjective weight, which inevitably has some subjective limitations.

Given the problems existing in the current transmission line loss prediction method, this paper proposes a transmission line loss prediction method based on CV and GBDT, and finally verifies the effectiveness of the method in this paper by testing the actual line data.

2. GBDT Model

GBDT model is an integrated model that integrates the prediction results of multiple tree models and continuously reduces the residual errors generated in the training process to realize data classification or regression.

Given the limited data set \( D = \{ Z, y \} \), in which \( Z = [Z_1, \ldots, Z_n, \ldots Z_N] \) for the input, \( y = [y_1, \ldots, y_N] \) for the output. The construction process of GBDT is as follows:

(1) Initialize the model. Estimate the model parameter \( \gamma \) that minimizes the loss function \( L(y, \gamma) \) and take it as the initial model \( f_0(Z_i) \), namely:

\[
f_0(Z_i) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)
\]  

(1)

(2) Set \( C \) as the number of iterations. For iteration \( c, c=1, 2, \ldots C \), execution (1)~(4).

① Calculate the loss function of the current model and the negative gradient \( r_{ic} \) of the model, namely the residual:

\[
r_{ic} = -\left[ \frac{\partial L(y_i, f(Z_i))}{\partial f(Z_i)} \right]_{f(Z)=f_{c-1}(Z)} , i = 1, 2, \ldots N 
\]  

(2)

② \( r_{ic} \) is taken as the new label of sample \( Z_i \) to obtain the new sample data set \( [(x_i, r_{ic}), i=1,2,\ldots N] \), and it is taken as the new training data to fit and get the next regression tree. The new tree model consists of leaf nodes \( R_j, (j=1, 2, \ldots J) \), in which \( J \) represents the number of leaf nodes in the regression tree model.

③ For each leaf node \( R_{jc} \), calculate the best fitting value of the sample \( j_{jc} \).

\[
\gamma_{jc} = \arg\min_{\gamma} \sum_{Z_i \in R_{jc}} L(y_i, f_{c-1}(Z_i) + \gamma)
\]  

(3)

④ Update the \( m \)-th iteration model:

\[
f_c(Z_i) = f_{c-1}(Z_i) + \sum_{j=1}^{J} \gamma_{jc} I(Z_i \in R_{jc})
\]  

(4)

(3) Output the final model \( f_c(Z) \).
The final GBDT regression model is obtained through iteration and combination of $m$ trees.

3. Transmission Line Loss Prediction Method based on CV and GBDT
The process of transmission line loss prediction method based on CV and GBDT is as follows: First of all, data pre-processing, including time granularity matching, and data statistic information extraction. Then CV algorithm was used to construct multiple sets of training sets and test sets. Finally, some GBDT models are trained by the training sets to verify the test sets, and the model set of CV results was extracted as the final transmission line loss prediction model.

3.1. Data Sources
The data comes from the actual line loss data of 31 transmission lines above 500kV in a certain province, including energy metering data, meteorological data, power flow data, and data of line ontology. The energy metering data including the date of data acquisition, power supply, and line loss. The meteorological data including site hours precipitation, temperature, humidity, wind speed, wind direction, and air pressure. Power flow data including voltage, current, active power, reactive power. Line ontology data including commissioning time, voltage level, total length of line, high calls, span, phase sequence, etc. The data range is from September 2017 to April 2019. Partial data samples are shown in the figure below.

![Figure 1. Partial data samples](image)

3.2. Data Preprocessing
The original data includes energy metering data, meteorological data, power flow data and line ontology data, among which the first three are time-related time-series data, while the line ontology data are fixed data.

For time-series data, it is necessary to match the time granularity so that it can be presented in the form of table, and then express its time change trend by mining its statistical characteristics through the construction of the time window.

As for the data of line ontology, there are line variables such as voltage level and commissioning time, as well as line tower variables such as call height, phase sequence, and tower topographic geology. The fixed-line variable of the line does not need to be processed, and for numeric line tower variables such as high calls and span extracted average values. For variables such as phase sequence and topographical geology of poles and towers, the value of the variable is the number of poles and towers of the corresponding type variable. For example, the phase-sequence data includes: ‘ABC’, ‘BCA’, ‘CAB’, ‘ACB’, ‘BAC’, ‘CAB’, 6 categories in total. If line A has a total of 58 poles and towers, of which 20 are ABC phase-sequence poles and 38 are BCA phase-sequence poles, the phase-sequence data of the line can be converted into the following form:

$$
f_C(Z_i) = f_0(Z_i) + \sum_{c=1}^{C} \sum_{j=1}^{J} \gamma_{jc} I(Z_i \in R_{jc})$$  \hspace{1cm} (5)
3.2.1. **Time Granularity Matching.** Since different data types come from different monitoring systems, the data granularity varies. Energy metering data are collected once a day, while meteorological data and power flow data are recorded once a minute. In order to ensure the accuracy of data information, under-sampling of meteorological data and power flow data is carried out to match them with energy metering data, that is, one time per day. The under-sampling method is shown in the following table:

| Data type          | Under-sampling mode |
|--------------------|---------------------|
| Wind speed         | Daily mean          |
| Wind direction     | Daily mean          |
| Humidity           | Daily mean          |
| Temperature        | Daily maximum       |
| Precipitation      | Daily maximum       |
| Air pressure       | Daily minimum       |
| Voltage            | Daily mean          |
| Current            | Daily mean          |
| Active power       | Daily mean          |
| Reactive power     | Daily mean          |

3.2.2. **Statistics Information Extraction.** Since the multi-dimensional influences of transmission line loss, such as weather, voltage, and current and so on, are time series data, mining its statistical characteristics in a certain time window can better describe the actual line situation at that time. According to the theoretical experience, the time window of this paper is set as 7 days, and the statistical features include average value, minimum value, maximum value and the average difference value. For a single influence quantity $V$, $V_w^i$ is the influence quantity data of the $i$-th day in the window $w$, $i = [1, 2... 7]$, where the average value $V_{\text{mean}}$ and the average difference $V_{\text{mean\_diff}}$ are calculated as follows.

\[
V_{\text{mean}} = \frac{\sum_{i=1}^{w} V_w^i}{w}
\]  \hspace{1cm} (6)

\[
V_{\text{mean\_diff}} = \frac{\sum_{i=2}^{w} (V_w^i - V_w^{i-1})}{w-1}
\]  \hspace{1cm} (7)

That is to say, each type of influence quantity is extended to four types of statistical features.

3.3. **Data Segmentation based on CV**

The training model data set needs to be divided into training set, validation set, and test set. Commonly used simple validation methods randomly divide the given data into two parts, part as the training set, another part of the training building model as a test set, but this approach only use some of the data, the
accuracy of verification set is closely related to the grouping of original data, so the results obtained by this method are not convincing.

In this paper, the \( k \)-fold CV method is adopted to partition data sets to adjust parameters and build an optimal model. \( k \)-fold CV randomly divided the sample set into \( k \) parts, \( k-1 \) part as the training set and 1 part as the verification set, and then rotated the training set and the verification set \( k \) times to construct \( k \) models. The prediction error was the average value of \( k \) times. Specific methods are as follows:

1. Divide the original data into \( k \) parts randomly without repeated sampling.
2. Select one of them as the test set and the remaining \( k-1 \) as the training set for model training.
3. Repeat step 2 \( k \) times, so that each subset has one chance as the test set and the rest as the training set. After training on each training set, a model is obtained, which is used to test on the corresponding test set, calculate and save the evaluation index of the model.
4. Calculate the average value of test results of \( k \) group as the estimation of model accuracy, and as the performance index of the model under the current \( k \)-folding and cross-validation. This is shown below.

![Figure 2. K-fold cross-validation](image)

This paper extracts the first 80% data of each line as training and verification data. Set the CV’s folds as 5, then extract 5 training sets and test sets in total for GBDT model training. This paper builds GBDT model based on LightGBM [7] framework in python language. The training parameters are set as follows: the maximum number of iterations is 1000, the learning rate is 0.1, early stopping rounds is 100, and the rest parameters are all default values, root mean squared error \( E_{rmse} \) (RMSE) is adopted, and formula (8) is used to calculate:

\[
E_{rmse} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\text{observed}_i - \text{predicted}_i)^2}
\]  

Where \( \text{observed} \) is the actual value of the \( i \)-th line loss sample; \( \text{predicted} \) is the line loss predicted by the model. \( N \) is the total number of samples.

The CV results are as follows: the average error of the five models is 35.8Mwh.
Table 3, 5-fold cross-validation error

| Iterations | $E_{\text{rms.e}}$/MWh |
|------------|-------------------------|
| Model 1    | 33.5                    |
| Model 2    | 37.3                    |
| Model 3    | 45.2                    |
| Model 4    | 34.3                    |
| Model 5    | 28.7                    |

4. Verification of Line Loss Calculation Model

4.1. The verification results
The last 20% data of each line is used as the test set to verify the model performance. The model prediction error is 35.5 MWh, which is close to the CV error. It shows that the verification result of this model has a strong generalization. Moreover, considering that the minimum index value of the electricity meter of 500 kV line is 200 MWh, the daily average output power is 7990 MWh, the electricity meter changes one time each time, and the line loss rate increases by 2.5%, the prediction model error in this paper is only 17.7% of the minimum index value. The line loss rate prediction of a line is shown in the figure below.

Figure 3. The result of line loss rate prediction
Figure 4. The result of the error of line loss rate prediction

It can be seen from the verification results that the model has a high degree of fit, and the maximum error of the line loss rate of the model is 0.71%, and the average error is 0.093%. It can be considered that the model satisfies the line loss prediction accuracy requirement and can better fit the multidimensional relationship between the amount of influences and the line loss.

4.2. Model performance comparison

In order to verify the advantages of the method proposed in this paper in line loss rate prediction task, the following model is constructed for test verification.

(1) Simple verification method + GBDT model
(2) 5-folds CV + BP model
(3) Methods of this paper

The test results are as follows:

| The model name                                      | RMSE/MWh |
|----------------------------------------------------|-----------|
| simple verification method + GBDT model             | 40.0MWh   |
| 5-folds CV + BP model                              | 81.5MWh   |
| methods of this paper                              | 35.5MWh   |

It can be seen from table 4 that the 5-folds CV method used in this paper reduced the line loss prediction error by 12.5%, and the GBDT method used in this paper reduced the prediction error by 43.5% compared with the commonly used BP model, proving that the method proposed in this paper is more suitable for actual line loss prediction.

5. Conclusion

In this paper, a transmission line loss prediction method based on CV and GBDT is proposed, through data pre-processing, data segmentation, and data modelling, the transmission line loss prediction model can accurately predict the actual line loss, and has the characteristics of high precision and strong generalization. Finally, the validity of the proposed method is proved by the actual line loss test.
Acknowledgments
This work was financially supported by State Grid Corporation of China Headquarters Science and Technology Project (JL71-18-002) fund.

References
[1] Cheng Huilin. Research on non-normal collection and response methods and application in line loss calculation. Yanshan university, Qinhuangdao, Hebei, China, 2013.
[2] Li Xueping. Research and analysis on quasi-online calculation of theoretical line loss of power grid. Yanshan university, Qinhuangdao, Hebei, China, 2013.
[3] Li Chaoying. Research on intelligent power grid based low and medium voltage line loss management. Tianjin university, Tianjing, China, 2012.
[4] Manjang, Salama, Y. George et al. Analysis of power losses of the 150 kV transmission using Poynting vector. International Conference on Power Engineering & Renewable Energy IEEE, 2012.
[5] Ansyari, F., C. S. Ozveren, and D. King. Allocation of transmission losses using three different proportional sharing methods. Universities Power Engineering Conference, 2007. UPEC 2007. 42nd International IEEE, 2007.
[6] Avinash, D., and B. Chalapathi. MW-Mile method considering the cost of loss allocation for transmission pricing. 2015 Conference on Power, Control, Communication and Computational Technologies for Sustainable Growth (PCCCTSG) IEEE, 2015.
[7] Liu dioxin, Bai Junfeng, Hu Hanghai, et al. Simultaneous line loss prediction based on improved limit learning machine, Computer and Digital Engineering, 46 (2018) 2528 - 2533.
[8] Li yunhao, Wang Jianxue, Wang Xiuli, et al. Power system loss assessment method based on hybrid cluster analysis, Power system automation, 40 (2016) 60 - 65.
[9] Zhang Yongjun, Shi Hui, Zhai Weifang, et al. Prediction of line loss rate based on analytic hierarchy process-grey comprehensive correlation and multi-grey model combination modelling, Power System Technology, 35 (2011) 71 - 76.
[10] Zhou Ting, Yang Jun, Zhou Qiangming, et al. Evaluation method of power system transient stability based on improved LightGBM, Power System Technology, 43 (2019) 1931 - 1940.