Short-term Transmission Line load and State Evaluation method based on improved gradient lifting Stochastic Forest algorithm

Xie Peng1, *, Duan Yuanchuan2, Zhang Da3

1School of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, Hunan 410000
2School of Information and Computer Engineering, Northeast Forestry University, Harbin, Heilongjiang 150036
3School of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang 150001

*Corresponding author: xiepeng2020@csuft.edu.cn

Abstract. In order to accurately predict the short-term load change of power system and provide guidance for the safe, economical and efficient operation of power system, a power system load forecasting method which combines fuzzy clustering and random forest regression algorithm is proposed. The rough set is used to construct compensation rules to modify and compensate the prediction results. Based on the analysis of the factors affecting the state of overhead transmission lines, the state quantity system of overhead transmission lines containing 79 state variables is constructed based on the state variables selected by 8 units of the lines and the state quantities of special cases. In the random forest regression prediction, the same kind of data after clustering is used as the training set sample to construct the decision tree. Considering that the stochastic forest regression prediction is conservative and the load fluctuates greatly at the peak, the rough set theory is used to generate compensation rules to modify the load forecasting. Accurate state evaluation of overhead transmission lines in the power grid can effectively reduce the failure rate and improve the power supply performance of the system. Finally, an example is given to analyze the actual data of 66 kV voltage grade overhead transmission lines in a city. The results show that the accurate evaluation of the state of overhead transmission lines can be realized by using this method, so as to provide decision support for the operation and regulation of regional power grid.

Keywords: Gradient lifting tree, Random forest algorithm, State evaluation, Overhead transmission lines, Power load.

1. Introduction

Power is the resource base of building a country and a nation. Accurate power load forecasting is of great significance to maintain the stable operation of power grid and draw up power dispatching plan. Due to the influence of many factors, such as production level, population density, residents' demand
and climate change, power load forecasting has the characteristics of high randomness and difficult modeling. In recent years, it has become a hot topic studied by scholars at home and abroad.

With the popularization of artificial intelligence technology, machine learning and deep learning methods have become important technical means to solve the problem of power prediction. Among them, the BP neural network load prediction model [1], which takes into account the related factors of daily characteristics, further reduces the negative influence of uncertain factors such as weather on the model. It is also proved that the time series model considering the influence of time factors on the global distribution of data is a powerful tool to solve the problem of power prediction. Feedforward neural networks after [2] will enter a valid mapping as the basic load component, weather sensitive components, special events and random component of the four parts, through the implementation of the neural network on each component connection operation, all the study out of the proportion of each component in the prediction process, greatly improve the prediction accuracy. Then, based on the emergence of the wavelet neural network model [3], Fourier transform was used in the training process to convert data to the time domain space for iterative calculation, and the prediction effect was very close to the actual electricity consumption.

The load forecasting model based on GBDT is trained with actual historical load data. Then, the importance of load influencing factors and the partial dependence of load on influencing factors are calculated [4]. Experimental results show that this method can analyze the nonlinear effects of various influencing factors on power loads.

2. Load forecasting model

2.1. Load prediction with gradient lifting algorithm

When selecting state variables to establish state variables system, it is necessary to consider each component unit of the circuit in detail. According to the structure of overhead transmission line, the overhead transmission line is divided into 8 units, such as foundation, pole tower, ground conductor, insulator string, metal fittings, grounding device, auxiliary facilities and channel environment.

For a given data set \( X=\{x_1,x_2,\ldots,x_n\} \), to determine its c-means fuzzy clustering, need to input number category \( C \), each cluster center \( m_j(j=1,2,\ldots,c) \). For each sample, \( x_k \), that is, the degree of membership of the KTH sample to Class IX [5]. Then the clustering loss index function based on the membership function can be expressed as Equation (1).

\[
J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left[ \frac{\left( \mu_j(x_i) \right)^b}{\sum_{j=1}^{c} \left( \mu_j(x_i) \right)^b} \right] |x_i - m_j|^2
\]  

(1)

\( B \) stands for the weighted index, also known as the smoothing factor, which represents the degree of sharing of a sample among fuzzy categories. There is still controversy about the optimal value of \( b \) in the academic circle. In consideration of the calculation amount and the calculation accuracy, the weighted index is usually 2.

Make and \( (\partial J/\partial m) = 0 \) partial derivative is zero, we can obtain the necessary condition to make the minimum \( J \) for type (2), (3).

\[
m_j = \frac{\sum_{i=1}^{n} \left( \frac{\left( \mu_j(x_i) \right)^b}{\sum_{j=1}^{c} \left( \mu_j(x_i) \right)^b} \right) x_i}{\sum_{i=1}^{n} \left( \frac{\left( \mu_j(x_i) \right)^b}{\sum_{j=1}^{c} \left( \mu_j(x_i) \right)^b} \right)}
\]  

(2)
\( \mu_j(x_i) = \frac{|x_i - m_j|^2}{\sum_{i=1}^{k} |x_i - m_j|^2} \)  

(3)

2.2. Load forecasting

Several weak prediction models are trained and a strong prediction model is formed through certain combination strategies.

According to the principle of empirical risk minimization in machine learning, \( \hat{F}_x \) can minimize the average loss is obtained on the training set. Iterative algorithm is generally used to optimize the model. At the beginning, the model consists of a constant function, and then the greedy algorithm is used to iterate continuously to generate the final model.

\[
F_0 = \arg \min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)
\]

(4)

\[
F_m(x) = F_{m-1}(x) + \arg \min_{h \in H} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x_i))
\]

(5)

When using C-means fuzzy clustering to perform clustering analysis, the number of categories C must be set in advance. The value of C has a profound influence on clustering. If the number of clustering is too large, the samples that should be of the same kind will be divided into different categories. If the number of clusters is too small, different types of data may be grouped into the same category. Setting the wrong number of clusters will lead to wrong clustering results and even make iteration unable to converge.

2.3. Improve the load prediction of the decision tree

The weak prediction model in the gradient lifting algorithm is usually composed of decision trees. Each decision tree is fitted by decision tree algorithm. Algorithm input include: the factors influencing the history data of \( x \) and \( y \) historical load training set \( D = \{x_i, y_i\}_{i=1}^{n} \). Loss function \( L(y, F(x)) \), the number of iterations, M, vector \( \eta \).

\[
f_j = \frac{1}{K} \sum_{k=1}^{K} S_{k,j} - s
\]

(6)

Calculation process is as follows: First, based on loss function \( L(y, F(x)) \), \( s \) to calculate the initial loss.1. Secondly, for each influence factor, \( J \). Then Random scrambles the attribute \( j \), and other attributes remain unchanged. Generate a new training set, \( D_{k,j} \). And adopt a new training set, \( D_{k,j} \) loss calculation, \( k \) j s.

The importance calculation of load influencing factors only reflects the relative importance of the factors affecting the load, and can not quantify the influence of the changing trend of load influencing factors on the load. Based on the prediction model, the concerned influencing factors are set to different values within the value range of this factor, while other factors remain unchanged. In this case, the average predicted load of the model on all training sets is the function of the influencing factors of the concerned load, which is called the partial dependent function of the model predicted load on the influencing factors.
3. Analysis of calculation examples
There are many factors affecting the load change of power system, but it can be seen from the analysis in Chapter 2 that meteorological factors and human activity factors are more important factors affecting the load of power system. Therefore, this paper considers the selected daily meteorological factors and real-time meteorological factor parameters. The daily meteorological information mainly includes maximum temperature, minimum temperature, average temperature, average pressure, precipitation, average wind speed, illumination time and humidity. The real-time meteorological information mainly includes real-time temperature, real-time humidity, real-time air pressure, real-time rainfall, real-time visibility, real-time cloud cover, real-time wind speed and real-time sunshine. The daily weather information is taken as the feature, including the daily maximum temperature, daily minimum temperature, daily average pressure, daily precipitation, daily average wind speed, daily illumination time and daily humidity.

![Fig. 1](image1.png)
**Fig. 1** Relationship between comprehensive index and cluster number

![Fig. 2](image2.png)
**Fig. 2** Correlation between power load and influencing factors
The importance is used to measure the influence of load factors on load in this area, and the order of load is sorted according to the magnitude. The results are different from the Pearson correlation coefficient method. When the importance is used to calculate the impact on the load, the maximum impact is the time; When Pearson’s correlation coefficient is used, the biggest effect is on temperature. Because Pearson correlation coefficient only reflects the linear relationship between factors and loads, but does not reflect the nonlinear relationship between the two. The correlation of the influence of time on load is different in different time periods. Some time periods are positively correlated.

It can be seen from the above prediction results that for the short-term load prediction of power system, there are indeed errors of peak value and abrupt value to improve the accuracy of power system load prediction.

4. Conclusions
In view of the shortcomings of the existing correlation analysis methods, on the basis of training the load forecasting model based on gradient lifting tree, it is proposed to use the importance and partial dependence of influencing factors to measure the nonlinear influence of influencing factors on load. It is verified by the actual power load data, and the nonlinear effects of time, air temperature, current date, working day and other factors on the load are analyzed in detail. First of all, this paper obtains the factors that affect the load change of power system through analysis and induction, which provides the basis for the establishment of random forest network load forecasting model. Secondly, the clustering algorithm is used to divide the similar days in the historical data, which makes the predicted sample data more targeted and discusses the selection of forecasting parameters. The load forecasting model based on random forest is established and the rough set algorithm is used to modify the forecasting results. Finally, by introducing the data from Northern Ireland as an example, the average absolute error percentage between the predicted results and the actual values directly obtained by using the random forest algorithm model is 2.32%. After rough set compensation, the average absolute error percentage is reduced to 2.09%. In a word, the effectiveness of the proposed algorithm is verified by an example algorithm, and the prediction result is more accurate after using rough set compensation algorithm.

The experimental results show that: (1) by using the importance of influencing factors to measure the influence of various factors on load, the nonlinear relationship between influencing factors and power load can be taken into account, identify the important factors affecting the load; (2) the partial dependence can be used to measure the nonlinear influence of the change of influencing factors on the load trend. Using this method to measure the impact of various factors on the load, for the power grid dispatching operators, we can adopt different forecasting methods to improve the forecasting accuracy according to different loads; for the power market personnel, we can understand the impact of different factors on the load and guide their behavior to participate in the power market.

References
[1] Wang Liwei, Zhang Qilong. Short-term load forecasting of regional distribution network based on BP neural network [J]. Electronic production, 2021 (01): 49-51 100.
[2] Yuan Xiaokai, Li Guo, Huang Shiping. Research on multivariable short-term power demand forecasting based on big data technology [J]. Power Grid and Clean Energy, 2020 and 36 (12): 30-34 years.
[3] Fan Haihong. Research on short-term load forecasting method based on firefly algorithm [J/OL]. China Electric Power: 1-11 [2021-01-12]. Http://kns.cnki.net/kcms/detail/11.3265.TM.20201221.1537.021.html.
[4] Zhang Suxiang, Zhao Bingzhen, Wang Fengyu, Zhang Dong. Short-term load forecasting under massive data [J]. Chinese Proceedings of Electrical Engineering, 2015 Magi 35 (01): 37-42.
[5] Tai Nengling, Hou Zhijian, Li Tao, Jiang Chuanwen, Song Jiong. Power system short-term load forecasting method based on wavelet analysis [J]. Chinese Proceedings of Electrical Engineering, 2003 (01): 46-51.