Research on Transformer Life Forecast Based on Random Forest Algorithm

Fei Liu*, Shili Liu, Xiang Gao and Xiaohu Zhu
State Grid Anhui Electric Power Co. LTD Economic Research Institute, Hefei, China

*Corresponding author e-mail: liuf2022@ah.sgcc.com.cn

Abstract. Accurately assessing the life and operating status of transformers has important guiding significance for the formulation of maintenance strategies for power grid companies, and at the same time plays a key role in the risk management of power grid companies. However, the traditional methods for predicting the remaining life of the equipment have the problems of insufficient accuracy or long data training time. In order to achieve a more accurate assessment of the life and status of the transformer, a random forest-based transformer life prediction method is constructed in this paper. Relying on the theory of big data analysis, by mining and analyzing the accumulated data of massive transformers, the life prediction model of the transformer is established and the characteristic parameters affecting the life of the transformer are extracted to predict the life of the transformer. The experimental data research demonstrates that the model can be accurate and effective. Predicting the life of transformers has higher prediction accuracy than traditional methods, providing method references for asset management and risk management of power grid companies.

Keywords: Predicted Lifetime, Random Forest, Transformer, Prediction Accuracy

1. Introduction
Power transformer is a key asset in power grid equipment and plays a vital role in the reliability of power system operation. During the operation and maintenance period, the investment of overhauls technical innovation project would not only affect the safe and stable operation of the power grid, but also affect the profitability output of the company. Based on the goal of technical feasibility and the best economic benefits, realizing the precise decision of capital investment in equipment overhauls technical innovation project was an important research direction at the moment [1]. At the same time, the continuous increase in energy demand and the increase in the number of operating transformers have made the operating transformers close to or have exceeded their expected technical life, leading to the problem of high failure rates of operating transformers [2]. The failure of an in-service transformer will have catastrophic consequences for the economy and operation of the power grid. Therefore, it is necessary to conduct regular condition monitoring of the transformer to predict the life of the transformer to plan equipment maintenance and replacement and reduce the risk of equipment failure. How to use the advantages of big data to find a data-supported analysis method for the life
management of transformers is a hot research issue for smart grids [3-4].

The main purpose of data mining technology is to explore the relationship between hidden variables from big data sets. Data mining technology involves three aspects: statistical learning, artificial intelligence and machine learning. In addition, data mining techniques are also used to analyze and predict research objects. Most of the data mining techniques used for analysis use clustering algorithms and association rules, while the data mining techniques used for prediction mainly use classification and regression algorithms. These include decision trees, artificial neural networks, genetic algorithms, K nearest neighbors, and naive Bayes. Generally speaking, the process of data mining on large data sets includes 7 steps. These steps can be defined as "data cleansing", "data integration", "data selection", "data conversion", "data mining", "model evaluation" and "analysis report".

So far, most of the research work in this field has focused on the discussion of the first two procurement strategies, operation and maintenance strategies, and LCC cost calculation methods, while the research on how to formulate decommissioning strategies, especially determining the economic decommissioning point of equipment, is very limited. Literature 5 proposes a method to determine the economic retirement point of the main equipment of the power grid by constructing an equipment failure model [5]. A large number of researchers have used data mining technology in power-related fields [6], For example, in document 6 and 7 [7].Applying the random forest algorithm to the field of power load has ideal results. Use big data processing to mine the relevant variables of each transformer, find out how it affects the life of the transformer, and predict the life of the transformer.

Some scholars have proposed to use the massive data of electricity meters for correlation analysis, and use data mining techniques to find out the hidden relationship between big data and performance status of electricity meters that are difficult to explore with traditional statistical methods. In literature 8, the characteristic parameters that affect the life of the transformer are extracted, and these characteristic parameters are learned through an adaptive fuzzy neural network, and the back propagation algorithm is used to solve the adaptive dynamic adjustment of the weights, and the life prediction model of the transformer is constructed [8]. In Literature 9, a dynamic failure rate model of a transformer suitable for short-term and medium- and long-term prediction is established based on the Markov model, and the remaining life of the transformer is modeled according to the calculated failure rate [9]. In Literature 10, a deep belief network is used to extract and classify the multi-dimensional data of power transformer faults, and combined with D-S evidence theory to solve the uncertainty problem in fault diagnosis, a multi-level decision fusion model for power transformer fault diagnosis is constructed [10]. However, the above methods have problems such as insufficient prediction accuracy and long data training time.

This paper proposes a transformer model based on random forest (RF). By mining and analyzing the basic attribute information of the transformer, the service life of the transformer is used as the output label of the model to predict [11]. Each meter can be obtained before the transformer is put into use. The estimated life of the transformer provides an analysis basis for the monitoring and rotation cycle of the transformer, and verifies the accuracy of the model proposed in this article. The transformer life prediction based on the random forest algorithm has extremely high accuracy and is currently a new prediction method in terms of transformer life prediction. It has great advantages over the previous prediction methods and has improved prediction accuracy and training duration.

2. Introduction to Random Forest Model Algorithm
Random forest is one of the extension methods of decision trees. It is an integrated learning algorithm composed of multiple decision trees. Because decision trees are prone to overfitting, in order to improve this shortcoming, the prediction result of random forest is composed of multiple decision trees. The trees are independently voted on, and the combination of decision trees makes parallel training of data sets possible. When the data set is large in scale and complex in nature, a single decision tree is not enough to obtain the data information in the synchronized phasor data, and a single
tree requires more time to classify the entire data set, and multiple decision trees are used to work in parallel. The speed and accuracy of classifying data sets are very efficient. The classification principle of random forest is shown in Figure 1.

Random forest does not use all variables to split tree nodes, but selects a random subset of variables at each node to obtain the best split of the node. The main purpose of such randomization is to remove related decision trees, so that the set of all trees has a lower variance. The method of constructing a random forest generally includes the following steps:

1) Extract n-tree sample subsets from the original data.
2) Use each sample subset to generate a decision tree. At each node of the tree, a variable M is randomly selected to split. Continue to grow the tree so that the number of nodes at each terminal node is not less than the size of the node.
3) Use voting mechanism to count the results of n-tree decision trees for classification. Random forest adopts the mode of multiple decision trees working in parallel, sampling the data randomly with replacement, and its predictive ability is relatively better than the single classification model. It is suitable for large data sets. Its classification model is generally considered to have high precision.

![Random forest classification principle](image)

Figure 1. Random forest classification principle

3. Construct a Life Prediction Model for Transformer Equipment

This paper preprocesses the collected transformer data and divides it into a training set and a test set, and then filters out multiple variables that may be associated with the life of the transformer, establishes a random forest model through computer programming and training set data, and substitutes it into the test set Data and adjust parameters to obtain different prediction results, and compare with the prediction results of the SVM method.

3.1. Data Preprocessing

The transformer data used in this article is provided by Anhui Power Grid. The transformer data in the data includes 14 basic information for analysis, including: "equipment classification", "start date", "asset manufacturer", "model", "Asset manufacturing country", "maintenance factory", "maintenance factory", "factory area", "cost center", "physical management department", "use storage department", "use custodian", "voltage level", "equipment Attribute data of various types of transformers: increase method" and "transformation capacity". These data are converted into corresponding digital labels, which are respectively used as the input feature quantities of the prediction model, and the life expectancy can be predicted by establishing the model.

When constructing the equipment life prediction model, the service life of 2000 transformers is counted and classified according to the service life, which is divided into "A", "B", "C", "D", "E" There are 7 levels of "F" and "G", where A level represents the equipment service life of 0-5 years, the B level represents the equipment service life of 5-10 years, and the C level represents the equipment service life of 10-15 years. D Class E means the service life of the equipment is 15-20 years, Class E
means the service life of the equipment is 20-30 years, Class F means the service life of the equipment is 30-40 years, and Class G means the service life of the equipment is more than 40 years. The remaining 14 items of basic attribute information are combined and converted into input feature quantities composed of 15 numbers.

3.2. Establishment of the Life Prediction Model of Transformer Equipment

There are more than 2,065 electricity meters used in this article. Among them, 2000 pieces of data are used as training data, and 65 pieces of data are used as prediction data to test the model. The prediction idea is shown in Figure 2.

![Figure 2. Flow chart of transformer life prediction](image)

In this article, the collected equipment data will be preprocessed to obtain a transformer data set composed of many influencing factors. The transformer data set is divided into two parts, one part is used for model training and consists of 2000 sets of data. This part is used to test the model and consists of 65 sets of data. As a result, analyze the predictions of the two models and compare them to get the final conclusion.

4. Analysis of Transformer Life Prediction Results

In order to verify the validity of the prediction model proposed in this article, the transformer data is verified experimentally according to the previous retained data and the verification method of model reliability.

4.1. Transformer Life Prediction Results

In the life prediction model of the electric meter, the 14 feature vectors mentioned above are input, and the error of the prediction result is shown in Figure 3. Among them, the data used for training the prediction model is 2000 transformer equipment data, and the data used for testing is 65 equipment data. Bring in different n-tree values to train the experimental data and the results are shown in the figure below.
Figure 3. RF prediction error map of different transformer life classes

In the figure, the n-tree value corresponding to RF1 is 25, the n-tree value corresponding to RF2 is 20, the n-tree value corresponding to RF3 is 15, the n-tree value corresponding to RF4 is 10, and the n-tree value corresponding to RF5 is 5.

It can be seen from the table that using the random forest prediction model constructed in this article, the accuracy of the equipment lifetime predicted by the basic information of 15 transformer equipment is above 95% except for the experimental result with an n-tree value of 5. And the larger the n-tree value, the lower the prediction error, and the better the prediction effect, and there are a small amount of prediction error for transformers with a service life of 15-20 years. However, the training time required for prediction will continue to increase with the increase of the n-tree value. If the amount of data is large, a lower n-tree value can be used for life prediction to balance the prediction time and the prediction effect. According to different needs make changes.

4.2. Random Forest Prediction Vs. SVM Prediction

Similarly, use the SVM algorithm to predict the life of the equipment using the data of 2000 transformer equipment and 65 equipment data. The results are shown in Figure 4 below. In most life classes, the number of scrapped equipment predicted by the SVM algorithm is greater than the number of true scraps, and the predicted number of scrapped equipment for the transformer life class of some equipment is less than 50% of the real data.

Figure 4. The result of life prediction of transformer equipment based on SVM algorithm

In Figure 5, the prediction result of the random forest prediction model with an n-tree value of 25, except that there are a small amount of error in the prediction of the life level of 15-20 years, the prediction accuracy of the rest of the life level is 100%, and the overall prediction is accurate. The rate is 98.45%, indicating that random forest can predict the life of equipment well based on the given data.

The comparison between the prediction results of the random forest prediction model with an n-tree value of 25 and the life prediction results of transformer equipment based on the SVM algorithm is shown in Figure 6 below.

It can be seen from the figure that the prediction accuracy of the random forest model under the transformer life level is higher than that of the SVM model, except that the sample number of the equipment life level of 30-40 years is 0. It shows that the prediction accuracy of the random forest model is higher than that of the SVM model. Under the condition of 15 items of data, the prediction effect of the random forest model in the prediction of transformer life is much higher than that of the SVM model. Moreover, the data training time of the SVM model is much longer than that of the random forest model. From the overall accuracy analysis, the accuracy of the SVM model is 63.08%, which is much lower than the 98.45% of the random forest model.
5. Conclusion

Based on the random forest algorithm, this paper constructs a transformer life prediction model. Through the provided 2065 transformer data, the experimental test is carried out, and the following conclusions can be obtained:

1) The random forest model constructed in this paper can effectively predict the life of the transformer, and through comparison with other algorithms, it is verified that the model has a high prediction accuracy.

2) Changing the value of the model's only parameter n-tree, the prediction accuracy will increase as the value of n-tree increases, but the prediction time will also increase as the value of n-tree increases; therefore, change n-tree according to different needs. The value of tree to balance the prediction time and prediction accuracy.

3) The method is based on the idea of big data, and the operation is simple and practical, which is convenient for the asset management and risk management of power grid enterprises.

References

[1] Li Na, Wang Xiaoliang, Li Chengqi, Zhang Zhenjun, Zhang Weiwei. The overhauls technical innovation project optimization method of power grid device based on Life Cycle Asset Management. Energy Reports, 2020,6(S9).

[2] L. Xue, Y. Liu, Y. Xiong, Y. Liu, X. Cui, G. Lei, A data-driven shale gas production forecasting method based on the multi-objective random forest regression, Journal of Petroleum Science and Engineering 196 (2021) 107801.

[3] Wu Moxuan. Analysis of assets' life cycle cost and benefits of technical overhaul. Jiangxi Electric Power 2017; 77–94.

[4] Liang Gang, Li Shengwei, Guo Tiejun, et al. Assistant decision - making method for transformer - replacement based on equivalent annual cost in life cycle. Proc CSU - EPSA 2017; 29(6):130–4.

[5] XIE Ning, WANG Chengmin, XIAO Dingyao, SONG Xiaoran. Economic Disposal Time of Primary Electricity Equipments . Electric Power Construction, 2014,35(06):165-168.

[6] Da Liu, Kun Sun. Random forest solar power forecast based on classification optimization. Energy, 2019, 187.

[7] Huang Jitao, Fan Bo,Zhou Yuanfeng, Hu Tingting, Liang Fei, Zeng Xiaodong. Fault and life prediction model of smart meter based on random forest . Ordnance Industry Automation, 2019,38(10):57-60.

[8] Hu Biwei, Deng Xiangli, Jia Shenghao. Transformer life estimation and state assessment based on ANFIS. Electrical Measurement & Instrumentation: 1-8[2021-02-08]. https://kns-cnki-
[9] F. Lo, C.M. Bitz, J.J. Hess, Development of a Random Forest model for forecasting allergenic pollen in North America, Science of The Total Environment 773 (2021) 145590.

[10] E. Mussumeci, F. Codeço Coelho, Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression, Spatial and Spatio-temporal Epidemiology 35 (2020) 100372.