Study on Chromatographic Condition Assessment of Transformer Oil Based on Random Forest Model

Xiu Zhou, Yunlong Ma*, Yan Luo, Tian Tian, Weifeng Liu, Xiuguang Li, Ninghui He, Zhenghua Yan, and Hui Ni

Power Research Institute of the State Grid Ningxia Power Company Limited, 750002, Yinchuan, China

Abstract. In order to solve the problem that the accuracy of transformer condition assessment based on oil chromatographic data is not high, this paper adopts random forest to evaluate transformer state. In order to prevent the attribute values carried by the decision tree of the basic unit in the evaluation model from being too few and thus affecting the evaluation accuracy, the ratio of \(H_2, CH_4, C_2H_6, C_2H_4, C_2H_2\), total hydrocarbon and various characteristic gases is used as the characteristic quantity, a stochastic forest optimization model for oil chromatography state assessment was constructed immediately following the heading. It is compared with BP neural network and SVM, the test results of field oil chromatogram data show that the accuracy of random forest state assessment is 88.5%, which is better than BP and SVM, and the random forest assessment model has faster assessment speed. This paper provides technical support for the fast and accurate condition assessment of transformer.

1 Introduction

Power transformer is the infrastructure of power system, which plays an irreplaceable role. Its stable operation is not only the premise of power system security and stability, but also the lifeblood of ensuring the stable development of national economy [1-2]. It is convenient to evaluate the operation state of power transformer and to carry out the maintenance of transformer, so as to improve the reliability of transformer. Therefore, it has important theoretical significance and practical application value. Nowadays, a large number of scholars are studying how to accurately evaluate the state of power transformer. In reference [3-4], a condition assessment model of transformer is established based on membership function evaluation, but the boundary of fuzzy set is difficult to be determined [3-4]. In reference [5], neural network is used to evaluate the state of transformer. However, when the characteristics of training and learning data set are not obvious, neural network has the problem of low evaluation accuracy [5]. In reference [6-8], the condition assessment of power transformer is based on SVM, but the accuracy of SVM is high only in small sample data set [6-8]. In reference [9], transformer fault diagnosis is based on random forest, but its structural parameters are fuzzy and parameters are not optimized [9].

163
addition, there are also literature from support vector machine regression \cite{10}, Bayesian network \cite{11-12} and other methods to evaluate the operation state of transformer. However, the actual power transformer oil color spectrum data set is messy, and the training samples need to be screened before the condition assessment. Therefore, it is urgent to put forward a better evaluation model, which can accurately evaluate the state of the transformer.

The contents of hydrogen (H\textsubscript{2}), methane (CH\textsubscript{4}), ethane (C\textsubscript{2}H\textsubscript{6}), ethylene (C\textsubscript{2}H\textsubscript{4}), acetylene (C\textsubscript{2}H\textsubscript{2}) and other gases are closely related to the operation status of the transformer. The operation status of the transformer is analyzed according to the dissolved gas in the oil, which is called dissolved gas analysis (DGA)\cite{13-18}. In machine learning algorithm, random forest algorithm is a good big data algorithm, which needs few parameters, fast running speed, high prediction accuracy and good robustness, and does not need to filter and reduce the original data \cite{19}. In this paper, dissolved gases H\textsubscript{2}, CH\textsubscript{4}, C\textsubscript{2}H\textsubscript{6}, C\textsubscript{2}H\textsubscript{4}, C\textsubscript{2}H\textsubscript{2} and their possible ratios C\textsubscript{2}H\textsubscript{2}/C\textsubscript{2}H\textsubscript{4}, CH\textsubscript{4}/H\textsubscript{2}, C\textsubscript{2}H\textsubscript{4}/C\textsubscript{2}H\textsubscript{6}, C\textsubscript{2}H\textsubscript{6}/H\textsubscript{2}, C\textsubscript{2}H\textsubscript{4}/H\textsubscript{2}, C\textsubscript{2}H\textsubscript{6}/CH\textsubscript{4}, C\textsubscript{2}H\textsubscript{4}/CH\textsubscript{4}, C\textsubscript{2}H\textsubscript{2}/CH\textsubscript{4}, C\textsubscript{2}H\textsubscript{4}/C\textsubscript{2}H\textsubscript{6} in oil are used as characteristic quantities. The transformer is divided into four operation states: good, attention, abnormal and serious. The random forest is used as the transformer condition assessment model. After training 300 groups of data and testing 75 groups of data, the accuracy of evaluation is 88.5%; compared with the evaluation results of BP neural network and support vector machine (SVM), the performance of random forest evaluation model is better. The related methods proposed in this paper can provide technical support for fast and accurate condition assessment of transformers.

2 Decision tree and random forest

2.1 Decision tree algorithm

Decision tree algorithm is the basic unit of random forest algorithm. Its idea comes from human decision-making process. It is a very intuitive model, which is easy to understand and realize. There are only three elements in decision tree: root node, non-leaf node and leaf node. Each node extends from top to bottom until each leaf node corresponds to a target, and its classification model is shown in Figure 1.

![Decision tree classification model](image)

Figure 1. Decision tree classification model.

The key of constructing decision tree is to choose the optimal attribute, and the methods to divide the optimal attribute are information gain, information gain rate and Gini index. In the decision tree of the basic unit of random forest, Gini index is used to divide the optimal attributes. If there are n samples in sample set D, the proportion of class I samples is pi, and the proportion of class j samples is pj, the Gini index is defined as:
\[ Gini(D) = \sum_{i=1}^{n} \sum_{j \neq i}^{n} p_i p_j = 1 - \sum_{i=1}^{n} p_i^2 \] (1)

2.2 Bagging algorithm

Bootstrap sampling refers to taking a sample that has been put back to the original sample set D, taking a total of N times to form a new sample D. Through this sampling method, the impact of outliers in the original sample on the training model can be reduced.

Bagging algorithm (bootstrap aggregation) is a kind of integrated learning framework, which can improve the performance of the model. The main idea is to use bootstrap method to sample in the original sample set. After sampling, a new sample set is trained to generate a weak learner. After this step is repeated N times, n weak learners are formed by voting mode.

2.3 Random forest algorithm

In the framework of bagging, the decision tree algorithm model is used as a weak learner, and randomness is added to the attribute value of the decision tree, which is called random forest algorithm. The algorithm flow is as follows:

1. The bootstrap method is used to extract n sample sets from the sample set. The n sample sets are used as the training set of decision tree, which is called tree number ntree in random forest;

2. If the sample has M characteristics, \( m \leq M \) is selected as the sub attribute set. In the whole process of "forest" growth, the value of m remains unchanged. The m sub attribute value is used to divide the optimal attribute, which is called mtry. The decision tree selected in this paper uses Gini index as the basis of dividing the optimal attributes;

3. According to step 1 and step 2, n decision trees are made up of which are different from each other. Then, the N decision trees are used to form the learning machine, and the random forest model is made up of the method of voting mode.

![Figure 2. Flow chart of random forest assessment.](image)
The flow chart of random forest model assessment is shown in Figure 2. Random forest still has a high accuracy rate in dealing with the data set with disorderly data and weak representativeness, and it can process the high-dimensional data directly. Therefore, before the training of the original sample set in the random forest model, there is no need for data screening, dimensionality reduction and other operations, which greatly increases the operability of the evaluation model.

3 Evaluate coding rules and select characteristic quantity

3.1 Evaluation coding rules

There are 375 sets of existing 110kV transformer oil color spectrum data sample sets, each set of samples contains the content of six characteristic gases, namely H$_2$, CH$_4$, C$_2$H$_6$, C$_2$H$_4$, C$_2$H$_2$ and total hydrocarbon. The samples have four transformer States, 300 sets are randomly used as training data, the remaining 75 sets are used as test data, and the sample data set is shown in Table 1.

<table>
<thead>
<tr>
<th>Condition type</th>
<th>Training set (Group)</th>
<th>Test set (Group)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>66</td>
<td>14</td>
</tr>
<tr>
<td>Attention</td>
<td>75</td>
<td>13</td>
</tr>
<tr>
<td>Abnormal</td>
<td>58</td>
<td>18</td>
</tr>
<tr>
<td>Serious</td>
<td>101</td>
<td>30</td>
</tr>
<tr>
<td>total</td>
<td>300</td>
<td>75</td>
</tr>
</tbody>
</table>

In order to facilitate the training of sample set in the model, four states are coded, and the coding rules and their meanings are shown in Table 2.

<table>
<thead>
<tr>
<th>Code</th>
<th>Condition</th>
<th>Stable working performance, very low probability of failure The test data is normal or the individual state quantity indicates that the reliability is slightly reduced, it can continue to operate, and the failure probability is low The probability of failure is increasing The overall function of the transformer is not good, and most of the monitored state quantities exceed the standard. There may be faults, or the possibility of faults is very high</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Good</td>
<td>Stable working performance, very low probability of failure</td>
</tr>
<tr>
<td>1</td>
<td>Attention</td>
<td>The test data is normal or the individual state quantity indicates that the reliability is slightly reduced, it can continue to operate, and the failure probability is low</td>
</tr>
<tr>
<td>2</td>
<td>Abnormal</td>
<td>The probability of failure is increasing</td>
</tr>
<tr>
<td>3</td>
<td>Serious</td>
<td>The overall function of the transformer is not good, and most of the monitored state quantities exceed the standard. There may be faults, or the possibility of faults is very high</td>
</tr>
</tbody>
</table>

2.2 Select characteristic quantity

When the characteristic quantity is H$_2$, CH$_4$, C$_2$H$_6$, C$_2$H$_4$, C$_2$H$_2$ and total hydrocarbon in 6 dimensions, the decision tree of the random forest contains too few sub attribute values, which makes the performance of the decision tree model worse, and the random forest model will not reach a high precision.
Therefore, it is necessary to increase the characteristic quantity correspondingly. In order to prevent the linear relationship of the characteristic quantity, the possible ratio of the characteristic gas is taken as the characteristic quantity by imitating the three ratio method of IEC. The increased characteristic quantity is: H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, C_2H_2 / C_2H_4, CH_4 / H_2, C_2H_6 / C_2H_4, C_2H_4 / H_2, C_2H_2 / H_2, C_2H_4 / CH_4, C_2H_2 / CH_4, C_2H_2 / CH_6.

The 6-D data and 15-D data are respectively imported into the random forest model for training. When the parameters of random forest adopt the default value, random forest still has high accuracy. At this time, the random forest parameters adopt the default values (ntree = 500, mtry = \sqrt{M} , where m is the quantity value of the characteristic quantity, and mtry takes the integer downward). The comparison of evaluation accuracy is shown in figure 3, 4. The accuracy is 83.8% in 6-D time and 87.9% in 15-d time. Obviously, after the data is upgraded, the performance of the model is significantly improved, which proves the feasibility of increasing the feature quantity.

![Figure 3. Performance of 6-D feature model.](image)

![Figure 4. Performance of 15-D feature model.](image)
4 Performance comparison between random forest and other models

In order to verify the unique advantages of random forest model in the condition assessment of power transformer, under the same training sample, BP neural network optimized by genetic algorithm (GA) and support vector machine (SVM) optimized by grid search algorithm are used for training, and the same test set is used for prediction.

3.1 Transformer condition assessment based on GA-BP neural network

![Figure 5. Performance of GA-BP neural network evaluation model.](image)

Under the same training set and test set, the transformer state is evaluated by the method in [20], and the evaluation simulation results are shown in Figure 5. It can be seen from the figure that the accuracy of GA-BP neural network for transformer condition assessment is 68.711%.

3.2 Transformer condition assessment based on gs-svm

In the same training set and test set, the condition assessment of transformer is carried out by the method in reference [10]. The radial basis function (RBF) kernel function is used, and the expression of RBF is shown in formula (2):

$$k(a,b) = \exp(-\gamma \|a-b\|^2)$$  

(2)

![Figure 6. Performance of GA-BP neural network evaluation model.](image)

The gs-svm evaluation model is used to train the above training set, and the test sample is used to test. The test accuracy is shown in Figure 6. According to the figure, the accuracy of the evaluation model is 76%. 

168
3.3 Performance comparison and result analysis

Compare the accuracy and evaluation time of the above proposed models, and the performance comparison results of each model are shown in Figure 7. Time each training model with TIC and TOC functions in MATLAB, and the training time of each model is shown in Table 3.

<table>
<thead>
<tr>
<th>Model</th>
<th>BP</th>
<th>SVM</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>training time</td>
<td>4.37s</td>
<td>34.5s</td>
<td>1.27s</td>
</tr>
</tbody>
</table>

![Comparison of SVM prediction results of test set (RBF kernel function)](image)

**Figure 6.** Gs-svm evaluation model performance.

**Figure 7.** Evaluation accuracy of each model.
As shown in the above chart, the random forest assessment model is not only more accurate, but also faster in training. The main reason is that the random forest is an integrated algorithm, which has better performance than the model with a single algorithm; and every basic unit (decision tree) in the model runs in parallel, each decision tree is different from each other and does not interfere with each other, and the running speed is fast.

5 Example verification

In order to verify the reliability of the random forest model, the oil chromatogram data of a transformer is evaluated by the random forest model. The oil chromatogram data and evaluation results of a transformer on August 31, 2019 are shown in Table 4.

Table 4. Oil chromatogram data and evaluation results of a transformer.

<table>
<thead>
<tr>
<th>Characteristic gas</th>
<th>H2</th>
<th>CH4</th>
<th>C2H6</th>
<th>C2H4</th>
<th>C2H2</th>
<th>Total hydrocarbon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content(ppm)</td>
<td>31.5</td>
<td>52</td>
<td>14.5</td>
<td>44.4</td>
<td>1.3</td>
<td>112.2</td>
</tr>
<tr>
<td>Assessment results</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It can be seen from table 4 that the evaluation result is "2". Compared with the coding rule in Table 2, the transformer status is "abnormal", that is, the possibility of failure is high. It is confirmed by the on-site maintenance personnel that although the transformer can continue to operate in this period of time, abnormal phenomena such as continuous heating occur, which is consistent with the assessment results. Several other transformers are evaluated respectively, and the evaluation results are consistent with the field test, which proves the reliability of the model.

6 Conclusion

The conclusion of this paper is as follows: For the condition assessment of transformer oil chromatogram, the random forest evaluation model has high evaluation accuracy, and does not need data pre-processing operations such as data screening. In this paper, the dissolved gases in transformer oil such as H₂, CH₄, C₂H₆, C₂H₄, C₂H₂, total hydrocarbon and its possible ratio C₂H₂/C₂H₄, CH₄/H₂, C₂H₆/C₂H₄, C₂H₄/H₂, C₂H₂/H₂, C₂H₄/CH₄, C₂H₂/CH₄, C₂H₂/C₂H₆ are used as the characteristics of the random forest assessment model Eigenvector, the transformer is divided into four operation states: good, attention, abnormal and serious. The evaluation accuracy of the model reaches 87.9%. The performance of the model is better than that of BP neural network and SVM. The evaluation results of the model are consistent with those of the field maintenance personnel. The method proposed in this paper can provide technical support for the condition assessment of power transformer.

References


