Research on the application of Decision Tree and Random Forest Algorithm in the main transformer fault evaluation

This research paper focused on the difficulties caused by rising volume and variety of data produced by technologies used in power systems. Traditional methods of data handling are unable to handle this enormous data and thus provide low data utilization and faulty analysis of equipment. This paper suggests that by using machine learning techniques we can draw meaningful conclusions and can develop flexible evaluation plans. The study focused on monitoring, defect and test information related to the main transformer equipment in Guanxi Province.

The market for equipment used in electricity transmission and transformation has seen extensive use of big data technology. Significant potential are presented by the variety of data kinds, which includes monitoring, environmental, and construction data. This study focuses on comparing classification algorithms and data fusion strategies to assess main transformer status. The goal of this technique is to improve maintenance procedures and overall equipment performance by utilising the wealth of power grid data. The extensive application of big data analysis is crucial for the intelligent management of power equipment and extending its lifespan.

The process begins with data pre-processing, where data from different sources is fused using the unique ID of the main transformer. This allows for the association of monitoring and inspection data with transformer defects and faults, expanding the grid data and completing the data preprocessing. Additionally, environmental data is fused with corresponding device data based on the area and time, establishing a connection between transformer data and environmental data. High-frequency monitoring data, such as low voltage alarms and oil chromatography readings, are obtained from the latest data. The fused data, including inherent attributes like equipment manufacturer, voltage level, and operation time, are used to form the regressor x, while the status of the power transformer (whether there is a defect or failure) is recorded as the dependent variable y. The processed data is further analyzed, and missing values are filled using the average of each attribute.

Finally, the dataset is divided, with the last 100 records serving as the test data set and the remaining data as the training data set. This division allows for the evaluation and validation of the developed fault evaluation model. Overall, this research paper explores the application of machine learning algorithms and big data analysis techniques to improve the evaluation and analysis of power system equipment, specifically main transformers, in the face of increasing data challenges.

A decision tree is a supervised classification model that allows nonlinear classification. It has a tree like structure where every node is a condition used to reach two different child categories. The leaf nodes in a decision tree represents the various categories in which the data can be classified.

The category for a particular input can be found out by traversing the tree until a leaf node is reached representing the respective category. The main goal while building a decision tree is the maximization of the difference between different categories. The classification and regression (CART) decision tree uses Gini coefficient as a loss function.

$$Gini(p) = \sum_{=k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

Here pk represents the probability of a category and K is the number of categories in the decision tree. While building a decision tree we try to minimize the Gini coefficient.

If the dataset is very large we randomly split the data into different parts and create decision tree for all different parts. The different decision trees after combining creates a random forest.

Analysis of model results

We applied the decision tree and the random forest model to the training set, and obtained the following experimental results.Then we applied the trained model to the test data set to evaluate the accuracy of the two modelsin evaluating the state of main transformers.

Test results of the decision tree model on the test data set

	PREDICTED			ACCURACY
ACTUAL		**No**	**Yes**	
	No	130	8	0.942
	Yes	14	9	0.391
				0.863

Test results of random forest model on test data set

	PREDICTED			ACCURACY
ACTUAL		**No**	**Yes**	
	No	138	0	1
	Yes	19	4	0.174
				0.882

From the result table that the comprehensive classification accuracy of the decision tree algorithm is 0.863, and that of the random forest is slightly higher, reaching 0.882. The decision tree algorithm performed much better than random forest. We believe that the decision tree algorithm is more suitable for the evaluation of 110kV main transformer defects and faults. From the results of the decision tree algorithm, we could see that the temperature of the sampled oil and the content of methane and carbon dioxide in the sampled oil chromatography have great significance for evaluating the defects and faults of the main transformers.