Project Classifymeister End-term Evaluation

Team Members: Diptansu, Harshit, Rutvik

Objective:- Report on a study on the application of Decision Tree and Random Forest algorithms in the main transformer fault evaluation.

Introduction:

1. The detection and monitoring methods of power equipment are relatively complex.
   Since power system equipment data has rapidly expanded, the drawbacks of traditional data modeling and analysis approaches are there.
2. This study deals with monitoring, fault, and test data for the main transformer equipment of Guangxi Province in China. The authors preprocessed the data set before designing decision trees and random forest evaluation models.
3. The evaluation results of the two algorithms were obtained and compared, which proved the effectiveness of the fault evaluation algorithm and selected a more accurate fault evaluation algorithm.

Feature engineering:

1. Various data sources were fused in data pre-processing using the main transformer's unique ID. This integrated monitoring, inspection, and environmental data, as well as high-frequency monitoring data. Inherent data, such as equipment specifications and operation time, were also incorporated.
2. A regressor x was built using "time + unique ID of the device + inherent attributes + grid data + device monitoring data + environmental data" based on the fused panel data. The dependent variable y recorded the status of the power transformer, indicating whether a defect or failure existed. Missing values were addressed to further process the fused data due to discrepancies in recording time across different systems. The average of each attribute was used to fill in the missing values.

Table.1. Variable settings

SYMBOL	MEANING	ATTRIBUTE
TEMP	Sampling oil temperature (℃)	numeric/regressor
H_2	Hydrogen by oil chromatography	numeric/regressor
C_2H_2	Acetylene by oil chromatography	numeric/regressor
C_2H_4	Ethylene by oil chromatogram	numeric/regressor
CH_4	Methane by oil chromatographic	numeric/regressor
C_2H_6	Ethane by oil chromatography	numeric/regressor
CO	CO by oil chromatography	numeric/regressor
CO_2	Carbon dioxide by oil chromatography	numeric/regressor
ACID	Oil acid ester (mgKOH/g)	numeric/regressor
PH	Ph of oil	numeric/regressor
V	Oil breakdown voltage (kV)	numeric/regressor
FLASH_POINT	Oil closing flash point (℃)	numeric/regressor
MACRO_WATER	Oily moisture content (mg/L)	numeric/regressor
STATUS	110kV main transformer condition	categorical /dependent

Algorithm:

1. Decision trees are tree-like nonlinear supervised classification models. Discrimination conditions are represented by branch connection points, while record types are represented by branch leaf nodes. When classifying with the decision tree, we choose branches based on whether the data matches the conditions stated at the node and repeat the aforementioned processes until we reach a leaf node.
2. A decision tree is randomly selected from a big data set. Repeating the above technique yields many decision trees with varied nodes and shapes. This is a random forest.

Evaluation:

The decision tree algorithm has a comprehensive accuracy of 0.863, while the random forest performs slightly better at 0.882. However, the decision tree algorithm is superior in predicting defective main transformers due to its simplicity and faster speed. The results highlight the importance of oil temperature and methane/carbon dioxide content in evaluating transformer defects.

Table 2 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

Table3 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

Conclusion:

The study compares the defects and fault evaluation of 110kV main transformers using decision trees and random forest models. In assessing transformer faults, the decision tree model achieves greater accuracy and emphasizes the importance of sample oil temperature and methane/carbon dioxide content. It introduces the fusion of grid-equipment-environment data for fault evaluation and offers insights for accurate power equipment status assessment.