Application of Naive Bayes and Decision Tree in the Prediction of Power Transformers Faults based on DGA

Yassine Mahamdi[1], Ahmed Boubakeur[2], Abdelouahab Mekhaldi[3]andYoucef Benmahamed[4]

[1,2,3,4]*Ecole Nationale Polytechnique (ENP),B.P 182 EL-HARRACH, Algiers, 16200, Algeria*

Abstract

Power transformers are the basic elements of the power grid, the state of which is directly related to the reliability of the electrical system. Many techniques were used to prevent power transformers failures, but the Dissolved Gas Analysis(DGA) remains the most effective one. Based on the DGA technique, we describe in this paper the use of two of the most effective machine learning algorithms: Naive Bayes (NB) and Decision Tree (DT) to identify power transformers faults. In our investigation, we developed 9 different input vectors from widely known DGA techniques. We used 481 samples and considered 6 types of faults. The implementation of the proposed methods has achieved an effectiveness of 86.25% in power transformers faults diagnosis.

Keywords

DGA, Decision Tree, Naive Bayes, Input vectors, faults diagnosis, Accuracy rate.

1. Introduction

Dissolved Gas Analysis is the most common and effective method for detecting transformer faults. It can immediately predict internal transformer failures, which generally avoids huge economic losses.

A transformer in service is exposed to two types of stresses: electrical and thermal [1]. Due to these stresses, the transformer oil and paper decompose, releasing a set of gases that reduce their dielectric strength. The nature and quantity of each dissolved gas produced in transformer oil can indicate the internal condition of the transformer.

The most common gases produced by the decomposition of oil are: ethane (C2H6), ethylene (C2H4),acetylene (C2H2), methane (CH4) and hydrogen (H2)[2], these differ mainly in the intensity of the energy which is dissipated by the fault [1], [3]. In addition tocarbon dioxide (CO2) and carbon monoxide (CO) that are formed as a result of the decomposition of paper[4], while, the nitrogen (N2) and the Oxygen (O2) are the non-fault gases.

There are many approaches developed for the analysis of dissolved gases in transformer oil and interpret their meaning including IEC Ratio, DORNENBURG Ratio, Rogers Ratio, Duval Triangle and Pentagon, and,Key Gas method. However, these techniques have certain limitations such as the existence of non-decision areas and erroneous results [5]. To overcome this situation, several artificial intelligence techniques have been used to improve the diagnostic accuracy of power transformers, such as fuzzy logic inference systems [6], artificial neural networks [7], hybrid grey wolf optimization [4], support vector machines and K-nearest neighbors [8-9], and have impressive performance [10-12].

In this paper, we examine the use of the Naive Bayes and the Decision Tree algorithms in faults identification. The originality comes from the introduction of several input vectors formed using

International Conference on Emerging Technologies: AI, IoT, and CPS for Science & Technology Applications, September 06–07, 2021, NITTTR Chandigarh, India

EMAIL: yassine.mahamdi@g.enp.edu.dz (A. 1); ahmed.boubakeur@g.enp.edu.dz (A. 2); abdelouahab.mekhaldi@g.enp.edu.dz (A. 3); youcef.benmahamed@g.enp.edu.dz (A. 4);
ORCID: 0000-0002-3777-7994 (A. 1); 0000-0001-9984-266X (A. 2); 0000-0001-6194-1363 (A. 3); 0000-0001-7179-3448 (A. 4);

widely known DGA techniques in order to identify the most suitable input data which gives the best performance of each algorithm and achieves the best prediction of fault in power transformers.

This article is arranged as follows, in the second section, we describe the collection of DGA data then the construction of the proposed input space followed by a brief presentation of the two classification algorithms used; Decision Tree (DT) and Naïve Bayes (NB). The results of implementing the two algorithms using our proposed input vectors are discussed in the third section where, the best input vector for each technique has been identified. Finally, the conclusions from this work were summarized and potential future work was mentioned.

2. Methodology

2.1. Data collection

The construction of our proposed input space needs gas concentration values. For this purpose, samples of transformer oil are taken periodically to check the gasesformed[12].Generally, mixtures of all gases are present in an oil sample, where the relative amount of each, could be an indicator of the existing faults, such as, partial discharges (PD), thermal faults > 700 °C (T3), thermal faults of 300 °C to 700 °C (T2), thermal faults < 300 °C (T1), high energy discharges (D2) and low energy discharges (D1)[4].

In this work, a database of 481 samples has been used in training and testing the proposed methods. This database has been extracted from the literature [13].The distribution of the training and the testing samples according to their fault type is shown in Table 1.

Table 1
Samples distribution

Fault types	Abbreviations	Samples for training	Samples for testing
Partial Discharge	PD	32	16
Thermal Faults $>$ 700 °C	T3	57	28
Thermal Faults of 300 °C to 700 °C	T2	32	16
Thermal Faults $<$ 300 °C	T1	63	32
High Energy Discharges	D2	84	42
Low Energy Discharges	D1	53	26
TOTAL		321	160

2.2. Proposed Input vectors:

The following attributes have been considered in the construction of our proposed input vectors:

1. Using the concentration of the usual five key gases in ppm:

$$\text{X=}[H_2, \text{CH}_4, \text{C}_2\text{H}_2, \text{C}_2\text{H}_4, \text{C}_2\text{H}_6] \tag{1}$$

2. Using the ratios between key gases (The IEC Ratios):

$$\text{X=}[\frac{\text{CH}_4}{H_2}, \frac{\text{C}_2\text{H}_2}{\text{C}_2\text{H}_4}, \frac{\text{C}_2\text{H}_4}{\text{C}_2\text{H}_6}] \tag{2}$$

3. Using the relative percentages of gases:

$$X = [\%C_2H_6, \%C_2H_4, \%C_2H_2, \%CH_4, \%H_2] \tag{3}$$

4. Using ROGER's four-ratio:

$$X = [\frac{C_2H_6}{CH_4}, \frac{C_2H_4}{C_2H_6}, \frac{C_2H_2}{C_2H_4}, \frac{CH_4}{H_2}] \tag{4}$$

5. Using DORNENBURG's four-ratios:

$$X=[\frac{\mathrm{CH_4}}{H_2},\ \frac{\mathrm{C_2H_2}}{\mathrm{C_2H_4}},\ \frac{\mathrm{C_2H_4}}{\mathrm{C_2H_6}},\frac{\mathrm{C_2H_2}}{\mathrm{CH_4}}] \tag{5}$$

6. Using Duval's triangle coordinates:

$$X = [C_a, C_b] \tag{6}$$

Where

$$C_a = \frac{1}{3}\frac{\sum_{i=0}^{k-1}(a_i+a_{i+1})(a_i b_{i+1}-a_{i+1}b_i)}{\sum_{i=0}^{k-1}(a_i b_{i+1}-a_{i+1}b_i)} \tag{7}$$

And

$$C_b = \frac{1}{3}\frac{\sum_{i=0}^{k-1}(b_i+b_{i+1})(b_i a_{i+1}-b_{i+1}a_i)}{\sum_{i=0}^{k-1}(b_i a_{i+1}-b_{i+1}a_i)} \tag{8}$$

The a_i are calculated by the equations:

$$\begin{aligned} a_0 &= \%CH_4 \cos\left(\frac{\pi}{2}\right) \\ a_1 &= \%C_2H_4 \cos\left(\frac{\pi}{2}+\varphi\right) \\ a_2 &= \%C_2H_2 \cos\left(\frac{\pi}{2}+2\varphi\right) \end{aligned} \tag{9}$$

And the b_i could be obtained by replacing ''cos'' with ''sin'' in the last equations with α = 2π/3

7. Using Duval's pentagon coordinates:

$$X = [C_a, C_b] \tag{10}$$

Where

$$C_a = \frac{1}{6}\frac{\sum_{i=0}^{k-1}(a_i+a_{i+1})(a_i b_{i+1}-a_{i+1}b_i)}{\sum_{i=0}^{k-1}(a_i b_{i+1}-a_{i+1}b_i)} \tag{11}$$

And

$$C_b = \frac{1}{6}\frac{\sum_{i=0}^{k-1}(b_i+b_{i+1})(b_i a_{i+1}-b_{i+1}a_i)}{\sum_{i=0}^{k-1}(b_i a_{i+1}-b_{i+1}a_i)} \tag{12}$$

The a_i are calculated using the following equations:

$$\begin{aligned} a_0 &= \%H_2 \cos\left(\frac{\pi}{2}\right) \\ a_1 &= \%C_2H_6 \cos\left(\frac{\pi}{2}+\varphi\right) \\ a_2 &= \%CH_4 \cos\left(\frac{\pi}{2}+2\varphi\right) \\ a_3 &= \%C_2H_4 \cos\left(\frac{\pi}{2}+3\varphi\right) \\ a_4 &= C_2H_4 \cos\left(\frac{\pi}{2}+4\varphi\right) \end{aligned} \tag{13}$$

Also, the b_i could be obtained by replacing ''cos'' with ''sin'' in the last equations with α = 2π/5.

8. In this case, a combination of two of the previously mentioned input vectorshas been done, Roger's and DORNENBURG's ratios:

$$X=[\frac{\mathrm{CH_4}}{H_2},\ \frac{\mathrm{C_2H_2}}{\mathrm{C_2H_4}},\ \frac{\mathrm{C_2H_4}}{\mathrm{C_2H_6}},\frac{\mathrm{C_2H_2}}{\mathrm{CH_4}},\frac{C_2H_6}{CH_4}] \tag{14}$$

9. To further improve fault recognition by expanding the proposed input space , another combination was made in the case of this input vector, Duval's triangle-pentagon coordinate'scombination:

$$X = [C_{a1}, C_{b1}, C_{a2}, C_{b2}] \tag{15}$$

Where {Ca1, Cb1} are calculated using the triangle method, while {Ca2, Cb2} are calculated according to the pentagon one.

2.3. AI techniques:

2.3.1. Naive Bayes

The NAIVE BAYES algorithm is a simple probabilistic classifier that uses Bayes theorem,which is given by the following equation [14]:

$$P(x|y) = \frac{P(y|x) \times P(\mathrm{x})}{P(y)} \tag{16}$$

Where $P(x|y)$refers to the subsequent possibility of the hypothesis x conditioned by some evidence y and$P(\mathrm{x})$ is the prior probability of x.

2.3.2. Decision tree

The decision tree algorithm is a non-parametric supervised machine learning's classifier used to split data into a set of branches. The construction of the tree is conducted from top to bottom in a recursive divide-and-conquer manner. The Decision Tree classifier training is based on finding the best split at each node as long as the full data set is not analyzed [15]. The said principle leads to the idea of partitioning the feature space until the interrupt criterion is satisfied in each list, or until all points in a given leaf belong to one class. Figure 1 illustrates the basic structure of a decision tree.

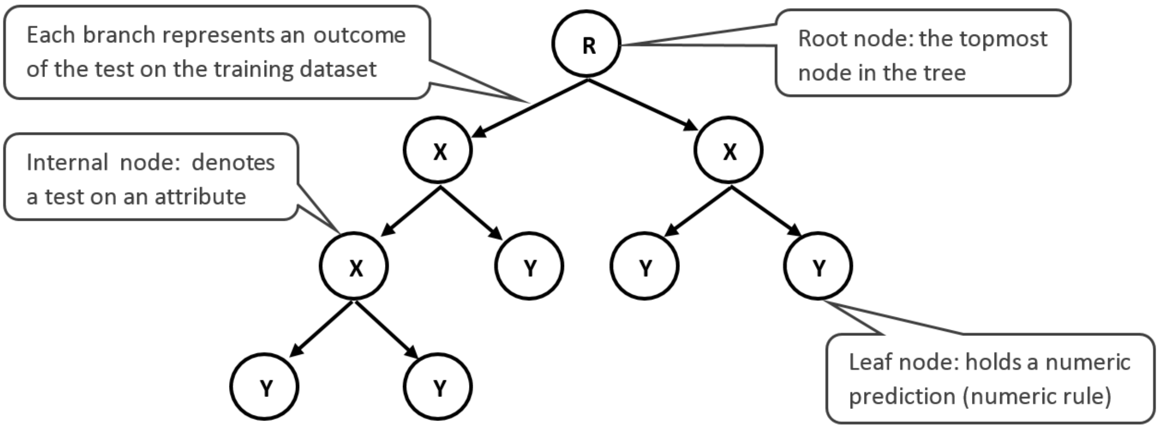
Each branch represents an outcome of the test on the training dataset
R
Root node: the topmost node in the tree
X
X
Internal node: denotes a test on an attribute
X
Y
Y
Y
Y
Y
Leaf node: holds a numeric prediction (numeric rule)

Figure 1: Decision Tree general structure

Among other classification algorithms, Decision Tree have the following advantages:

- Good performance with large data sets
- Requires little data preparation
- Easy to display graphically
- Easy to understand and interpret

<u>Construction of decision tree:</u>

In order to select the best variable to split, the Decision Tree uses the information gain. The equation for calculating information gain is as follows:

$$Gain(T, A) = Entropy(T) - \sum_{i=1}^{n} \frac{T_i}{T} Entropy(T_i) \tag{17}$$

Where $Gain(T, A)$ is the information gain of set T (training data) on an attribute A and T_i is a subgroup of T for which: A has value i.

The Entropy of node T is defined as:

$$Entropy(T) = -\sum_{i=1}^{n} p(i) \log p(i) \tag{18}$$

Where $p(i)$ is the proportion of T belonging to a class i.

3. Results and discussion:

To evaluate the performance of Naïve Bayes and Decision tree algorithms using our proposed input vectors according to six types of transformer faults, a set of 481 samples has been used to train and test the two methods; 67% of the dataset were used for the training and 33% for the testing, using the MATLAB software. Table 2 shows the results of the implementation of the two classifiers using the proposed input vectors.

Table 2

Faults diagnostic results in percent using the Naïve Bayes and the Decision Tree algorithms with all the proposed input vectors

Input vector	1	2	3	4	5	6	7	8	9
Naïve Bayes	25.62	81.87	13.75	11.25	28.75	58.25	42.50	28.75	86.25
Decision tree	75.62	80.62	83.12	83.75	77.50	45	78.75	76.25	78.75

From Table 2, it is easy to see that the highest prediction accuracy is obtained using vector 9 (combined Duval’s pentagon and triangle) with the Naïve Bayes algorithm (86.25%). Whereas, in the case of the Decision Tree, the input vector 4 (Roger's four-ratio method) gives the highest prediction accuracy, up to 83.75%.

In order to deepen the study, the performance of each algorithm with its appropriate input vector was evaluated based on the accuracy of each fault type diagnosis (Figure 2).

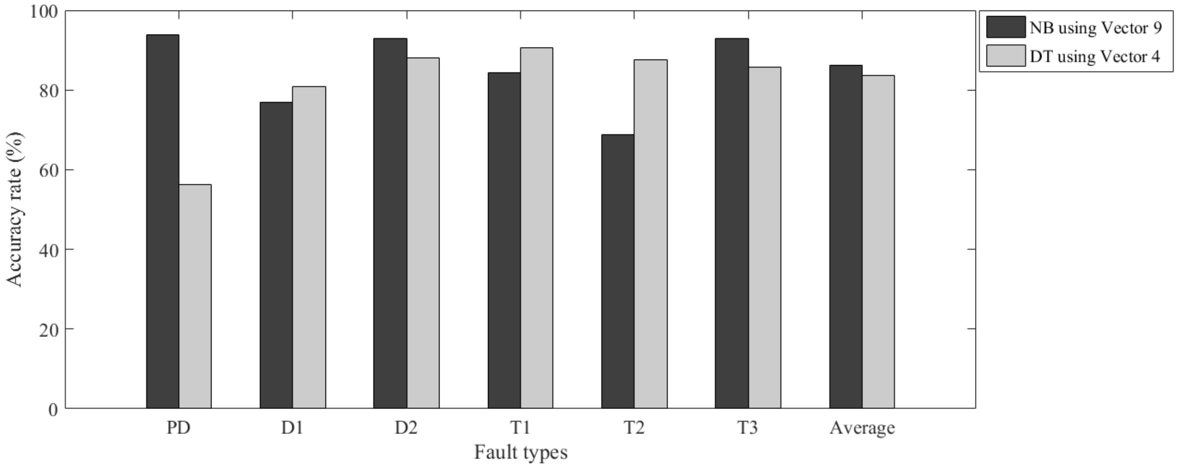

Figure 2: Histogram of accuracy rate

From Figure 2, it is clear that the performance of each algorithm differs depending on the type of fault. For example, in the case of the partial discharges (PD), the Naïve Bayes has the best performance, while, in the case of medium thermal fault (T2), the Decision Tree has the superiority in such fault recognition. Overall, the Naïve Bayes algorithm remains the one with the greatest precision.

4. Conclusion:

The Naïve Bayes and the Decision Tree classification algorithms were used to identify power transformer faults. A dataset of 481 samples was employed and 9 different input vectors were considered. The Naive Bayes algorithm achieved a diagnostic accuracy of 86.25% when using the 9th input vector (Duval’s triangle-pentagon coordinates combination), compared to 83.75% in the case of the Decision Tree using the 4th input vector (ROGER's four-ratio). These diagnostic results show an improvement in the identification of transformer faults over other traditional DGA methods. Significant differences in diagnostic accuracy were obtained when using the same classification algorithm with different input vectors, this investigation shows the appropriate input vector for the diagnosis of power transformers using the Naive Bayes and the Decision Tree algorithms.

In a future work, we will extend the proposed input space using other input vectors with an improved machine learning algorithm.

5. References

[1] ‘’Mineral Oil-Filled Electrical Equipment in Service - Guidance on the Interpretation of Dissolved and Free Gases Analysis’’, IEC Standard IEC 60599, IEC, Geneva, Switzerland, Edition 2.1, May 2007.

[2] F. Jakob and J. J. Dukarm, “Thermodynamic estimation of transformer fault severity”, IEEE Trans. on Power Delivery, vol. 30, no. 4, pp. 1941–1948, 2015.

[3] M. Duval and A. dePabla, “Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases”, IEEE Electrical Insulation Magazine, vol. 17, no. 2, pp. 31–41, Mar. 2001.

[4] A. Hoballah, D. A. Mansour and I. B. M. Taha, “Hybrid grey wolf optimizer for transformer fault diagnosis using dissolved gases considering uncertainty in measurements”, IEEE Access, vol. 8, pp. 139176–139187, 2020.

[5] S. S. M. Ghoneim and I. B. M. Taha, ‘‘A new approach of DGA interpretation technique for transformer fault diagnosis’’, Int. J. Electr. Power Energy Syst., vol. 81, pp. 265–274, Oct. 2016.

[6] Islam, S.M.; Wu, T.; Ledwich, G. ‘’A novel fuzzy logic approach to transformer fault diagnosis’’. IEEE Trans. Dielectr. Electr. Insul. 2000, 7, 177–186.

[7] S. Souahlia, K. Bacha and A. Chaari, “MLP neural network-based decision for power transformers fault diagnosis using an improved combination of Rogers and Doernenburg ratios DGA”, Int. Journal of Electrical Power & Energy Systems, vol. 43, no. 1, pp. 1346–1353,

[8] Benmahamed, Y.; Teguar, M.; Boubakeur, A. ‘’Application of SVM and KNN to Duval Pentagon 1 Transformer Oil Diagnosis’’. IEEE Trans. Dielect. Electr. Inst. 2017, 24, 3443–3451.

[9] Kherif, O.; Benmahamed, Y.; Teguar, M.; Boubakeur, A and Ghoneim, S. S. M. "AccuracyImprovement of Power Transformer Faults Diagnostic Using KNN Classifier With Decision TreePrinciple," in IEEE Access, vol. 9, pp. 81693-81701, 2021.

[10] Yang, M.-T.; Hu, L.-S. ‘’Intelligent fault types diagnostic system for dissolved gas analysis of oil-immersedpower transformer’’. IEEE Trans. Dielectr. Electr. Insul. 2013, 20, 2317–2324.

[11] J. I. Aizpurua, V. M. Catterson, B. G. Stewart, S. D. J. McArthur, B. Lambert, B. Ampofo, G.Pereira, and J. G. Cross, ‘‘Power transformer dissolved gas analysis through Bayesian networksand hypothesis testing’’, IEEE Trans. Dielectrics Electr. Insul., vol. 25, no. 2, pp. 494–506, Apr.2018.

[12] Mirowski, P.; LeCun, Y. ‘’Statistical Machine Learning and Dissolved Gas Analysis: A Review’’. IEEE Trans. Power Deliv. 2012, 27, 1791–1799.

[13] Taha, I.B.M.; Hoballah, A.; Ghoneim, S.S.M. ‘’Optimal ratio limits of Roger’s four-ratios and IEC 60599 code methods using particle swarm optimization fuzzy-logic approach’’. IEEE Trans. Dielect. Electr. Inst. 2020, 27, 222–230.

[14] Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). ‘’Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability’’. arXiv preprint arXiv:1206.1121

[15] John Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993