Improved DGA method based on rules extracted from high-dimension input space

M.A.B. Amora, O.M. Almeida, A.P.S. Braga, F.R. Barbosa, L.A.C. Lisboa and R.S.T. Pontes

The diagnosis of incipient faults in power system elements such as transformers is usually based on the concentrations of dissolved gases existent in the insulation oil. There are consolidated DGA-based (dissolved gas analysis) methods in the literature, such as the Duval triangle. However, they present some limitations such as the existence of non-decision areas and erroneous results. Proposed is a simple methodology to improve the analysis of incipient faults based on rules extracted from a high-dimension space (21 attributes), formed by the gases concentrations and some of their interrelations. From such input space, the C4.5 method (decision tree) is used to extract a set of interpretable rules. Databases known in the DGA technical literature such as IEC TC 10 are adopted to analyse the proposed approach. When compared with a standard method, considering all data test folders in the performed 10-folder cross-validation statistical analysis, the extracted rules show greater accuracy with an error in the diagnosis of incipient faults of 6.25%, against 18.75% for the Triangle method in the worst case.

Introduction: Incipient internal faults in power system equipments with insulation oil such as power transformers can be monitored and diagnosed through periodic analysis of dissolved gases in the oil [1, 2]. For proper diagnosis, the expert usually employs one or more standard methods, which are described in IEEE C57.104 [3] and IEC 60599 [4]. However, such methods have severe limitations: (i) existence of situations that are not covered by the criteria imposed by the standards, (ii) existence of more than one possible diagnosis, or (iii) erroneous results. One of the most widely used and accurate methods is the Duval triangle [4], which is considered as the standard method for comparison in this Letter.

Several studies [1, 2, 5–8] have used artificial intelligence (AI) techniques, e.g. artificial neural networks (ANNs) [5, 6] and fuzzy logic inference systems (FISs) [7, 8] to overcome the aforementioned limitations in methods supposed to diagnose incipient faults. This approach achieves significant improvement in accuracy, but also typically has its very own limitations: (i) ANN-based solutions have high precision, but the reasoning to justify the diagnosis is not interpretable, and (ii) FIS-based solutions can be interpretable, but dealing with a large number of rules.

This Letter proposes an alternative AI methodology for the diagnosis of faults in transformers and power system equipment that use insulation oil. It combines a compact set of interpretable rules with high accuracy in the diagnosis. For this purpose, the C4.5 method [9] is used to infer a decision tree that classifies data from a high-dimension input space, formed by the gas concentrations and some of their interrelations, considering labels determined by an expert. The C4.5 method is an extension of the ID3 algorithm [10] able to handle numeric attributes. However, C4.5 makes partitions on the input space using hyperplanes that are orthogonal to the axes. It can generate large decision trees with low capacity to generalise data that are nonlinearly separable in the attribute axes [11] as in DGA. Therefore, analogously to the representation space created by a hidden layer in an ANN [12], this work creates a high-dimensional space with some meaningful relationships between the gases, thus making the classification of incipient faults easier to be treated by hyperplanes that are orthogonal to the attribute axes.

The proposed input space has considered the following attributes: the concentration of the usual seven key gases [1–4] in parts per million (ppm) (H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO, and CO_2), the ratios between key gases (CH_4/H_2, C_2H_2/C_2H_4, C_2H_4/C_2H_6, C_2H_4/CH_4, C_2H_2/CH_4, C_2H_2/H_2, and CO/CO_2), and the relative percentages ($\%CH_4 = 100x/(x+y+z)$, $\%C_2H_2 = 100y/(x+y+z)$, $\%C_2H_4 = 100z/(x+y+z)$, $\%H_2 = 100.H_2/(H_2+C_2H_6+CO+CO_2)$, $\%C_2H_6 = 100.C_2H_6/(C_2H_6+x+y+z)$, $\%CO = 100.CO/(C_2H_6+x+y+z+CO+CO_2)$, and $\%CO = 100.CO_2/(C_2H_6+x+y+z+CO+CO_2)$, where $x = CH_4$, $y = C_2H_2$, and $z = C_2H_4$). Some of these relations are used in methods described by standards [3, 4] or several works existent in the literature [1, 8, 13, 14], although they are not used together as in this work. According to the aforementioned criteria, the idea is to use an initial large number of variables in the input space to explore the relationships of the gases, allowing more efficient fault diagnosis in transformers and other equipment with insulation oil.

Development: The decision tree was trained and tested from a database of 162 DGA cases (117 cases from the IEC TC 10 database [15], 39 cases from Tables I to III in [16], and six cases from working transformers in the Northern Power Grid of India, as presented in [13]). The following assumptions were made in the data selection: (i) if the concentration of a given gas is not available, then it is considered zero; (ii) a ratio $0/0$ is set as null; (iii) a given ratio $v/0$ is set to 20 as in [14], considering that v is not null; and (iv) a concentration indicated by '<1' is set as 0.5 [14].

The output labels for the decision tree indicates the following possible diagnoses, in accordance with the Triangle method [4]: partial discharges (PD), discharges of low energy (D1), discharges of high energy (D2), thermal faults of temperature $< 700°C$ (T1/T2), and thermal faults of temperature $> 700°C$ (T3).

The software used in the development of the decision tree is SIPINA [17]. Given the number of cases in the considered database, the cross-validation technique [18] is adopted for the statistical analysis. A 10-folder cross-validation was adopted, i.e. the database was partitioned into 10 complementary subsets (folders). While nine folders are considered as the training set, the remaining one is the validation set. To reduce variability, multiple rounds of cross-validation are performed. The results are indicated in Table 1.

Table 1: Percentage of classification error, from cross-validation

Folder / Set	1	2	3	4	5	6	7	8	9	10
Training (Extracted rules)	2.7	1.4	4.1	3.4	1.4	2.7	2.0	2.7	4.1	2.0
Training (Triangle)	13.7	13.7	13.7	13.7	13.7	13.7	13.7	13.7	13.0	13.0
Test (Extracted rules)	25.0	6.25	6.25	25.0	25.0	12.5	12.5	12.5	12.5	6.25
Test (Triangle)	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	18.75	18.75

The line 'Folder' in Table 1 represents the number of the subset considered as the validation set in the current round of cross-validation. In each round, all folders are subsets formed by 10% of the total data, where cases involving the five considered labels (PD, D1, D2, T1/T2, and T3) are uniformly distributed. Care was also taken so that a uniform number of errors related to the Triangle method in each folder is always considered. Table 1 presents the errors related to the training and test sets for each folder from cross-validation, using both extracted rules obtained from a decision tree and the Triangle method. The percentage error in data is calculated based on 146 cases in the training set, and 16 cases in the validation set. Analysing the results in Table 1, the decision tree associated with folder 2 demonstrates the best performance in both training and validation. Thus, the set of rules obtained by the decision tree is considered as the best one among those generated in the cross-validation.

Table 2 shows a comparison between the classification error when applying these rules to each folder and the error obtained by the Duval triangle in the same folders. Table 3 summarises the 13 rules extracted from this best decision tree (folder 2 in Table 1).

Table 2: Per cent error in test sets

Folder / Set	1	2	3	4	5	6	7	8	9	10
Test (Extracted rules)	0	6.25	0	0	0	0	6.25	6.25	0	0
Test (Triangle)	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5	18.75	18.75

Evaluation of proposed rules: From Table 3, one can see that, although 21 attributes were initially considered in the training of the decision trees, only eight of them remained in the generated rules. This occurs because C4.5 considers only the attributes that have more influence in the output. In fact, C4.5 determines that, from the considered training cases, the aforementioned attributes ($\%C_2H_2$, $\%C_2H_4$, C_2H_4/CH_4, CH_4/H_2, C_2H_2/H_2, C_2H_2/C_2H_4, CH_4, and C_2H_6) are the most informative ones regarding the diagnosis of incipient faults.

Table 3: Extracted rules

$\%C_2H_2$	$\%C_2H_4$	C_2H_4/CH_4	CH_4/H_2	C_2H_2/H_2	C_2H_2/C_2H_4	CH_4	C_2H_6	Diagnosis
<12.9		<0.01				> = 6.75		PD
<12.9	<23.35	<0.01				<6.75		T1 or T2
> = 12.9	<23.35						> = 2345	T1 or T2
<12.9		>1.25	<0.08					T1 or T2
<12.9		> = 0.01 and <1.25	<135	<0.02				T1 or T2
<12.9		> = 0.01 and <1.25	<135	> = 0.02	> = 0.03			T1 or T2
<12.9		> = 0.01	>135					T3
<12.9		> = 1.25	> = 0.08 and <135					T3
<12.9		> = 0.01 and <1.25	<135	> = 0.02	<0.03			T3
> = 12.9	<23.35						<2345	D1
> = 12.9	> = 23.35		<0.05					D1
> = 12.9	> = 23.35		> = 0.36				<3	D1
> = 12.9	> = 23.35		> = 0.05				> = 3	D2
> = 12.9	> = 23.35		> 0.05 and <0.36				<3	D2

Table 4 shows the percentage of success in the diagnosis of faults using the extracted rules compared with that obtained by using the Triangle method for each database used in this work.

Table 4: Percentage of success in databases

Database (Cases)	Duval Triangle (%)	Extracted Rules (%)
IET TC 10 Database (117 cases)	88.03	99.15
Tables I–III (39 cases) in [16]	82.05	94.87
Six cases described in [13]	83.33	100.0

By analysing the results in Table 4, one can conclude that the rules in Table 3 have a success rate higher than that presented by the Triangle method, despite the small number of rules and attributes. Regarding the IEC TC 10 database cited in the IEC 60599 standard, the hit rate reaches 99.15% against 88.03% for the Triangle method. Another important advantage is that there are no blank intervals in extracted rules such as in Rogers' and Dörnenburg's methods mentioned in the standards [3, 4].

Conclusions: This Letter has presented a new set of rules extracted from a decision tree for the diagnosis of incipient faults in electric equipment that uses insulation oil. From the methodological point of view, the idea of using a large number of input attributes that explores some relationships among the gases has allowed the conception of a diagnosis method based on a decision tree that produces interpretable rules. The major advantage of the diagnosis based on such rules if compared with the Triangle method (IEC 60599 standard) is the improved precision. Other AI-based DGA methods mentioned in the technical literature [5–8] also have larger hit rate than that achieved with the rules existing in the standards. However, methods using ANNs and FISs have limitations regarding non-interpretable rules or a large set of rules. The present work has achieved great accuracy in the diagnosis (less than 7%, i.e. 6.25% for the worst case in Table 2, and less than 6%, i.e. 5.13% for the worst case in Table 4) with a reduced set of interpretable rules (13 rules considering eight attributes, as seen in Table 3).

19 April 2012
doi: 10.1049/el.2012.1363

M.A.B. Amora, O.M. Almeida, A.P.S. Braga, F.R. Barbosa, L.A.C. Lisboa and R.S.T. Pontes (*Department of Electrical Engineering, Research Group on Automation and Robotics, Universidade Federal do Ceará, Fortaleza–CE, Brazil*)

E-mail: marcio@dee.ufc.br

References

1 Amora, M.A.B., Almeida, O.M., Braga, A.P.S., Barbosa, F.R., Lima, S.S., and Lisboa, L.A.C.: 'Extraction of knowledge from artificial neural networks to application in analysis of transformers'. Intelligent 15th Int. Conf. on System Applications to Power Systems, ISAP '09, Curitiba, Brazil, November 2009

2 Barbosa, F.R., Almeida, O.M., Braga, A.P.S., Amora, M.A.B., and Cartaxo, S.J.M.: 'Application of an artificial neural network in the use of physicochemical properties as a low cost proxy of power transformers DGA data', *IEEE Trans. Dielectr. Electr. Insul.*, 2012, **19**, (1), pp. 239–246

3 IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers, IEEE Standard C57.104-2008, September 2008

4 IEC Mineral Oil-Impregnated Electrical Equipment in Service – Guide to the Interpretation of Dissolved and Free Gases Analysis, IEC 60599–2008

5 Wang, M.H.: 'Extension neural network for power transformer incipient fault diagnosis', *IEE Proc., Gener. Transm. Distrib*, 2003, **150**, (6), pp. 679–685

6 Esp, D.G., and McGrail, A.J.: 'Data mining applied to transformer oil analysis data', *IEE Digest*, **1999**, (119), pp. 9/1–9/7

7 Spurgeon, K., Tang, W.H., Wu, Q.H., Richardson, Z.J., and Moss, G.: 'Dissolved gas analysis using evidential reasoning', *IEE Proc., Sci. Meas. Technol.*, 2005, **152**, (3), pp. 110–117

8 Miranda, V., and Castro, A.R.G.: 'Improving the IEC table for transformer failure diagnosis with knowledge extraction from neural networks', *IEEE Trans. Power Deliv.*, 2005, **20**, (4), pp. 2509–2516

9 Quinlan, J.R.: 'C4.5: Programs for machine learning' (Morgan Kaufmann, 1993)

10 Quinlan, J.R.: 'Induction of decision trees', *Mach. Learn.*, 1986, **1**, (1), pp. 81–106

11 Rokach, L., and Maimon, O.: 'Top-down induction of decision trees classifiers – a survey', *IEEE Trans. Syst. Man Cybern.*, 2005, **35**, (4), pp. 476–487

12 Haykin, S.: 'Neural networks: a comprehensive foundation' (Prentice Hall, 1999, 2nd edn)

13 Naresh, R., Sharma, V., and Vashisth, M.: 'An integrated neural fuzzy approach for fault diagnosis of transformers', *IEEE Trans. Power Deliv.*, 2008, **23**, (4), pp. 2017–2024

14 Li, X., and Wu, H.: 'DGA interpretation scheme derived from case study', *IEEE Trans. Power Deliv.*, 2011, **26**, (2), pp. 1292–1293

15 Duval, M., and de Pablo, A.: 'Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases', *IEEE Electr. Insul. Mag.*, 2001, **17**, (2), pp. 31–41

16 Duval, M.: 'A review of faults detectable by gas-in-oil analysis in transformers', *IEEE Electr. Insul. Mag.*, 2002, **18**, (3), pp. 8–17

17 Rakotomalala, R. (http://eric.univ-lyon2.fr/~ricco/sipina.html, accessed 12nd April 2012)

18 Mitchell, T.: 'Machine learning' (McGraw Hill, 1997)