A Comparative Study of the Application of Neural Networks and Decision Trees in the Classification of Incipient Faults in Power Transformers

Luciana G. Castanheira, João A. de Vasconcelos, Agnaldo J. Rocha Reis, Paulo H. V. Magalhães, Sávio A. L. da Silva

***Abstract*— The power transformer is one of the most important equipment in an electric power system. If this equipment is out of order for some reason, the damage for both society and electric utilities are very significant. In this work, we present a comparative study of the application of Multi-Layer Perceptrons trained via Rprop algorithm and Decision Trees in the classification of incipient faults in power transformers. The proposed procedures have been applied to real databases derived from chromatographic tests of power transformers. The results obtained by both techniques are compared and fully described. The classifiers discussed here can be seen as a very important component in power transformer predictive maintenance activities.**

I. Introduction

FOR many years, preventive maintenance programs in power transformers consisted of inspections, tests and actions in periodic time intervals usually suggested by the manufacturers or determined through practical experience. It was also common the application of routine tests and procedures such as: measurement of dielectric losses, insulation resistance and winding resistance; physic-chemical and chromatographic oil analysis; and manual or automatic monitoring of temperature [1].

We discuss in this work the use of Neural Networks (NN) and Decision Trees (DT) for pattern recognition as supporting tools for the diagnosis of faults in power transformers. Considering that the power transformer is crucial for the power system operation, techniques for diagnosis and fault detection are required. To be more specific, many faults that occur in power transformers are due to changes in the gas concentrations in their insulating oil. Taking in consideration that there are not efficient mathematical models to describe the relationship between the rate of evolution of these concentrations and the failures, and the process of gathering historical data is a common practice nowadays, the development of pattern classifiers based on Support Vector Machines [2,3], Neuro-Fuzzy tools [4,5], Wavenets [6], Neural Networks [7, 8] and Decision Trees [9, 10] has received a great deal of attention.

It is used in the present work a method for fault detection in power transformers proposed by [1]. The pattern classification is carried out based on the levels of the dissolved gases in the power transformer oil such as Ethylene (C_2H_4), Methane (CH_4), Acetylene (C_2H_2), Hydrogen (H_2) and Ethane (C_2H_6). The two main goals here are: 1) to present a comparative study between NN and DT for the problem of incipient faults classification in power transformers; 2) to work towards the development of an artificial intelligence-based predictive maintenance tool for power transformers. In order to validate the proposed methodologies, we have made use of real data from chromatography tests of power transformers.

This paper is divided as it follows. In Section II – Development – it is presented the database description, and the conception of the neural classifiers and DT-based classifiers are discussed as well. All six tests considered in this work are fully described in Section III and analyzed in depth in Section IV. The main conclusions of the paper and suggestions for future work appear in Section V.

II. Development

The tests for the classification of the faults described in this paper follow the process of Knowledge Discovery Database (KDD). KDD process refers to the procedure of extracting knowledge from rough data. Data mining is one of the steps of this process. Its main goal is to transform the data pre-processed into information. The data mining task requires an algorithm known as data miner. In this particular paper, we utilize both NN and DT as data miners.

A. Database Description

The preprocessing of power transformer oil databases relied on a method proposed by Duval [1], which only takes into account the relative percentage concentration of the gases acetylene, ethane and methane. In the triangle shown in Fig.1, it is represented the evolution of the produced gases to some failures. The ratio between each gas and the total amount of the produced gas is calculated in order to find the

Manuscript received in June 30th, 2011.

Luciana G. Castanheira, Agnaldo J. Rocha Reis✉, Paulo H. Magalhães and Sávio A. L. da Silva are with the Department of Control Engineering and Automation, School of Mines, Federal University of Ouro Preto, Campus Morro do Cruzeiro, Ouro Preto, MG, Brazil, 35.400-000. (e-mails: lugcastanheira@yahoo.com.br, agnreis@gmail.com, phvmag@gmail.com, saviolsil@gmail.com).

João Antônio de Vasconcelos is with the Department of Electrical Engineering, School of Engineering, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil, 31.270-010. (e-mail: joao@cpdee.ufmg.br).

coordinates. Besides, there are some tags to be considered, namely: PD (Partial Discharge), T1 (Thermal Failure for Temperature T<300°C), T2 (Thermal Failure for 300ºC<T<700°C), T3 (Thermal Failure for T>700°C), D1 (Low Energy Discharges), D2 (High Energy Discharges), DT (Mix of failures). Thus three electrical failures (D1, D2, DT) and three thermal failures (T1-T3) can be found in the Duval's triangle.

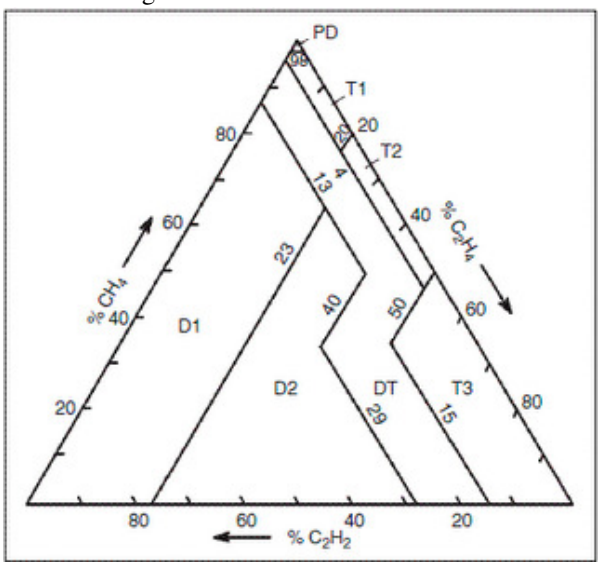

Fig.1: Duval Triangle (Source: [1]).

It was considered here three different databases containing the concentration of dissolved gases in the power transformers insulating oil (input data) and the fault diagnosis (output data) each. These data appear in [11]. The databases are composed for concentrations of the five most important gases found in the power transformer oil, namely: Hydrogen (H_2), Methane (CH_4), Ethylene (C_2H_4), Ethane (C_2H_6) and Acetylene (C_2H_2). The formation of these gases is due to:

- Hydrogen: large quantities of this gas are associated with the partial discharge conditions;
- Hydrogen, Ethane, Methane and Ethylene: the production of these gases results of the thermal decomposition of the oil due to the its contact with hot parts of the power transformer;
- Acetylene: its production is associated with the formation of electrical arc in the oil.

The first database is called 'IEC' which comprises part of a database made available by IEC TC 10 [12]. It is composed of 53 samples divided in the three following sets: (a) 16 'Normal' samples, (b) 22 samples diagnosed as 'Electrical failure' and (c) 14 samples related to 'Thermal faults'. The second database is called 'BASE1' and includes data made available by the Research Center in Electrical Power (CEPEL). This database is composed of 224 samples divided in 83 normal samples, 61 electrical failure samples and 80 thermal fault samples. Finally, the third database is named 'BASE2' and it is composed of 212 samples divided in 180 normal samples, 10 electrical failure samples and 22 thermal fault samples [11]. As a matter of fact, when a database presents a different number of instances by class, it is said that this database is 'unbalanced'. For instance, BASE2 instances are divided in three classes and each class has a different number of instances (180+10+22). Finally, all the aforementioned pattern classification tasks were carried out by experts via specific measurements.

The simulations performed in the three data are described as it follows. In the first group, both data miners were trained with 70% of the IEC's data and validated with the remaining data (the learning stage). Both NN-based classifier and DT-based classifier were used to classify BASE1 and BASE2 databases (the test stage). In the second group, other classifiers were trained with 70% of BASE1 data and validated with the remaining data. These Artificial Intelligence(AI)-based classifiers were tested with IEC and BASE2 data. Finally, the third group of data is formed by merging IEC and BASE1 databases. As in the previous groups, 70% of the data were used in the training stage and the remaining data were used for validation purposes. The AI-based classifiers were tested with BASE2 data. It is important to notice that although it was used different databases in the test stage, the same inputs as in the learning stage were taken into account. This technique is thought here as a kind of cross-validation and its main objective is to evaluate the generalization ability of the classifiers. Also, it is worth mentioning here the following codification for the power transformer failures considered in this work:

- Class A: for power transformers diagnosed as NORMAL (N);
- Class B: for power transformers presenting ELECTRICAL FAILURE (EF);
- Class C: for power transformers presenting THERMAL FAILURE (TF).

B. Simulation Details

All simulations regarding NN were performed using the software MatLab®. All databases were allocated in two different matrices: Input Matrix (IM) and Output Matrix (OM). The type of gas and its concentration appear in the IM (5 types of gases have been considered) and the codification for the power transformer, i.e., the class it belongs (N, EF or TF), can be found in OM.

Multi-Layer Perceptrons (MLP) trained with the Resilient Backpropagation algorithm [13] has been used in this work. Many different network configurations have been evaluated. Parameters such as activation functions, number of hidden neurons, and number of iterations have been set up based on the previous work of Zhang et al.[7], Lu et al. [14], and Wang et al. [15]. Three particular activation functions have been employed. 'Hyperbolic Tangent' and 'Sigmoid' activation functions have been used in the hidden layer while the linear function has been used in the output layer only. For every neural classifier, the number of iterations has been

modified (1000, 4000, and 8000 iterations) as well as the number of hidden neurons (4, 6, 8, and 10 neurons). A minimum training error equals to $1e10^{-5}$ has been defined as an additional training stop criterion. The learning rate and the momentum rate have been set up as 0.4 and 0.5, respectively. At the end, 36 different MLP schemes have been evaluated. The following results have been obtained with those classifiers that presented the best generalization ability.

On the other hand, all the simulations concerning the DT classifiers were performed using the software WEKA (Waikato Environment for Knowledge Analysis). The software is formed by a set of several machine learning algorithms and it is indicated to derive useful knowledge from databases that are too large to be analyzed by hand [16-17]. WEKA is implemented in Java language and it was developed in the University of Waikato, New Zealand, in 1999. The algorithm is based on the concepts of entropy and Information Gain (IG) in order to construct the tree. It always aims to reduce the entropy (i.e., the randomness of the objective variable), to be consistent with the database and to have the smaller number of nodes.

The classification task was carried out according the same three explainedclasses. All databases were allocated in an .arff file. In that file appear the type of gas and its concentration (the same 5 types of gases have been considered) and the codification for the power transformer, i.e., the class it belongs (N, EF or TF) represented as A, B or C, respectively (ABC labels for classes is an WEKA's requirement).

The file has the following structure: the first valid line indicates the relation name to find (i.e. @ relation relation_name). In the following lines one needs to list all the attributes and its kind (i.e. type of attribute and nature of its values). The classes' labels must be put between "{}" and be separated by commas. All numerical data appear in the sequel. Fig.2 is an example of an arff code fraction.

```
@relation IEC

@attribute H2 real
@attribute CH4 real
@attribute C2H2 real
@attribute C2H4 real
@attribute C2H6 real
@attribute diag {A, B, C}

@data
134,134,0.04,45,157,A
100,200,20,200,200,A
0.04,225,3,110,225,A
```

Fig.2: An arff code fraction.

All simulations were made changing (or not changing sometimes) the pruning parameters of the tree and the Confidence Factor (CF). The CF is a suitable form to evaluate the precision of the obtained rules in the training stage. This factor is calculated by the ratio X/Y, where X is the number of records that satisfies the predecessor and the successor of the rule, and Y is the total number of records that satisfies the predecessor of the rule [18]. The J4.8 algorithm does the pruning task by means of the post-pruning approach. [19], [20].

III. TESTS

In order to carry out the classification of incipient faults in power transformers via monitoring their gases concentrations, six tests were performed for each AI-based classifier and their results can be found in Tables 1-12. Each single test is fully explained below.

A. *TEST1*

In this test, it was considered the three rough data sets IEC, BASE1, and BASE2, which are unbalanced data sets, and a global concordance index. Here the interest was in analyzing the overall performance of the neural classifier; there is not discrimination among the classes A, B, and C.

The Global Percentage Concordance Index (GPCI) for the training and validation data sets are shown in Tables 1 and 2 as well as the GPCI for the test sets. It is important to affirm that **all** the six tests were performed in a database that did not take part of the learning process of the classifier. Indeed, in the very first analysis reported in Table 1, the classifier had been trained/validated with IEC database and tested with BASE1 and BASE2 databases.

TABLE 1 – GLOBAL PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR UNBALANCED DATABASES (NN-BASED CLASSIFIER).

First Analysis	Global Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Overall	95.4	68.8	50.9	52.6
Second Analysis	Training/Validation (BASE 1)		Test Sets	
			IEC	BASE 2
Overall	91.6	66.3	57.7	52.8
Third Analysis	Training/Validation (IEC+BASE1)		Test Sets	
			BASE2	
Overall	92.2	70.3	69.8	

TABLE 2 – GLOBAL PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR UNBALANCED DATABASES (DT-BASED CLASSIFIER).

First Analysis	Global Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Overall	97.2	68.8	67.9	66.8
Second Analysis	Training/Validation (BASE 1)		Test Sets	
			IEC	BASE 2
Overall	92.8	76.5	72.8	61.3
Third Analysis	Training/Validation (IEC+BASE1)		Test Sets	
			BASE2	
Overall	86.5	81.5	72.8	

B. TEST2

In this test it was taken into account the same three unbalanced data sets used in TEST1. However, at this time, individualized concordance indexes were used. The aim here was to analyze the performance of the neural classifier for each one of the classes A, B, and C.

The Individualized Percentage Concordance Index (IPCI) for the training and validation data sets are shown in Tables 3 and 4 as well as the IPCI for the test sets. Notice now that the performance of the classifiers for classes A (Normal), B (Electrical Failure), and C (Thermal Failure) are evaluated separately.

TABLE 3 –INDIVIDUALIZED PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR UNBALANCED DATABASES (NN-BASED CLASSIFIER).

First Analysis	Individualized Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Normal	98.8	98.6	42.8	51.1
Electrical Fault	90.5	46.8	49.9	71.5
Thermal Fault	99.0	98.0	65.6	61.1
Second Analysis	Training/Validation (BASE 1)		Test Sets	
	Trein.	Valid.	IEC	BASE 2
Normal	93.9	54.1	63.1	63.6
Electrical Fault	91.9	60.3	82.8	78.7
Thermal Fault	87.8	76.1	21.3	12.1
Third Analysis	Training/Validation (IEC+BASE1)		Test Set	
	Training	Validation	BASE2	
Normal	93.5	86.2	78.1	
Electrical Fault	96.2	87.3	48.3	
Thermal Fault	84.3	41.1	3.8	

TABLE 4 –INDIVIDUALIZED PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR UNBALANCED DATABASES (DT-BASED CLASSIFIER).

First Analysis	Individualized Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Normal	100	100	59.1	46.9
Electrical Fault	93.3	42.9	67.2	80.0
Thermal Fault	100	100	77.5	72.7
Second Analysis	Training/Validation (BASE 1)		Test Sets	
	Trein.	Valid.	IEC	BASE 2
Normal	94.8	68.0	75.0	68.2
Electrical Fault	93.1	72.3	90.9	84.6
Thermal Fault	89.5	83.3	50.0	25.9
Third Analysis	Training/Validation (IEC+BASE1)		Test Set	
	Training	Validation	BASE2	
Normal	91.2	90.3	83.3	
Electrical Fault	93.1	92.0	53.9	
Thermal Fault	75.8	64.3	11.0	

C. TEST3

In this test, the three data sets IEC, BASE1, and BASE2 were considered. Nevertheless, at this time, the data sets were balanced by duplicating those data in smaller number of samples [21, 22]. As in TEST1, a global concordance index was considered as well.

The Global Percentage Concordance Index (GPCI) for the training and validation data sets is shown in Tables 5 and 6 as well as the GPCI for the test sets.

TABLE 5 – GLOBAL PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR BALANCED DATABASES (NN-BASED CLASSIFIER).

First Analysis	Global Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Overall	95.4	77.9	59.8	60.5
Second Analysis	Training/Validation (BASE 1)		Test Sets	
			IEC	BASE 2
Overall	92.7	78.6	75.2	52.8
Third Analysis	Training/Validation (IEC+BASE1)		Test Sets	
			BASE2	
Overall	94.8	80.2	78.3	

TABLE 6 – GLOBAL PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR BALANCED DATABASES (DT-BASED CLASSIFIER).

First Analysis	Global Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Overall	97.8	83.6	75.9	69.7
Second Analysis	Training/Validation (BASE 1)		Test Sets	
			IEC	BASE 2
Overall	93.7	87.3	84.9	74.2
Third Analysis	Training/Validation (IEC+BASE1)		Test Sets	
			BASE2	
Overall	95.7	90.1	84.6	

D. TEST4

In this test the same three balanced data sets used in TEST3 were taken into account but this time individualized concordance indexes were applied. The aim here was analyzing the performance of the classifiers for each one of the classes A, B, and C.

The Individualized Percentage Concordance Index (IPCI) for the training and validation data sets are shown in Tables7 and 8 as well as the IPCI for the test sets. As in TEST2, the performance of the classifiers for classes A (Normal), B (Electrical Failure), and C (Thermal Failure) are evaluated separately.

TABLE 7 – INDIVIDUALIZED PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR BALANCED DATABASES (NN-BASED CLASSIFIER).

First Analysis	Individualized Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Normal	100	86.7	45.8	68.2
Electrical Fault	86.7	57.2	63.9	52.6
Thermal Fault	100	85.8	69.9	60.9
Second Analysis	Training/Validation (BASE 1)		Test Sets	
	Training	Validation	IEC	BASE 2
Normal	96.6	61.9	81.0	86.6
Electrical Fault	94.8	60.3	92.8	56.5
Thermal Fault	86.3	76.1	52.4	67.6
Third Analysis	Training/Validation (IEC+BASE1)		Test Set	
	Training	Validation	BASE2	
Normal	91.2	68.0	89.0	
Electrical Fault	97.1	83.6	62.6	
Thermal Fault	96.6	90.3	83.3	

TABLE 8 – INDIVIDUALIZED PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR BALANCED DATABASES (DT-BASED CLASSIFIER).

First Analysis	Individualized Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Normal	100	100	65.9	88.8
Electrical Fault	93.4	57.2	86.1	60.7
Thermal Fault	100	100	77.0	63.3
Second Analysis	Training/Validation (BASE 1)		Test Sets	
	Training	Validation	IEC	BASE 2
Normal	96.6	82.7	86.4	90.5
Electrical Fault	93.1	92.9	96.2	66.5
Thermal Fault	91.4	88.1	72.8	68.7
Third Analysis	Training/Validation (IEC+BASE1)		Test Set	
	Training	Validation	BASE2	
Normal	92.7	90.3	92.2	
Electrical Fault	97.1	93.6	79.9	
Thermal Fault	96.9	96.8	82.1	

E. TEST5

In this test, the same three balanced data sets used in TEST4 were considered, but, at this time, the percentage of Total Combustible Gas (TCG) in the power transformer insulating oil, which many times is related to overload conditions, was considered as well. Roughly, the concentration of each type of gas is divided by the sum of all concentrations for that gas which implies that some kind of normalization procedure was employed (see [23, 24] for further details). As in TEST1 and TEST3, a global concordance index was used.

The Global Percentage Concordance Index (GPCI) for the training and validation data sets are shown in Tables 9 and 10 as well as the GPCI for the test sets.

F. TEST6

In this test the same three balanced and normalized data sets used in the previous test was considered here, but, at this time, individualized concordance indexes were applied. As in TEST2 and TEST4, the interest was in analyzing the performance of the classifiers for each one of the classes A, B, and C.

The Individualized Percentage Concordance Index (IPCI) for the training and validation data sets are shown in Tables 11 and 12 as well as the IPCI for the test sets. Again, the performance of the classifiers for classes A (Normal), B (Electrical Failure), and C (Thermal Failure) are evaluated separately.

TABLE 9 –GLOBAL PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR BALANCED AND NORMALIZED DATABASES (NN-BASED CLASSIFIER).

First Analysis	Global Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Overall	98.2	81.1	66.9	69.2
Second Analysis	Training/Validation (BASE 1)		Test Sets	
			IEC	BASE 2
Overall	95.2	81.9	82.4	60.3
Third Analysis	Training/Validation (IEC+BASE1)		Test Sets	
			BASE2	
Overall	97.1	85.3	83.7	

TABLE 10 –GLOBAL PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR BALANCED AND NORMALIZED DATABASES (DT-BASED CLASSIFIER).

First Analysis	Global Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Overall	98.0	84.5	77.2	73.6
Second Analysis	Training/Validation (BASE 1)		Test Sets	
			IEC	BASE 2
Overall	95.7	89.6	87.2	75.1
Third Analysis	Training/Validation (IEC+BASE1)		Test Sets	
			BASE2	
Overall	98.1	92.6	87.8	

TABLE 11 –INDIVIDUALIZED PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR BALANCED AND NORMALIZED DATABASES (NN-BASED CLASSIFIER).

First Analysis	Individualized Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Normal	100	100	63.8	79.5
Electrical Fault	92.2	57.4	66.2	54.9
Thermal Fault	100	85.7	71.2	70.7
Second Analysis	Training/Validation (BASE 1)		Test Sets	
	Training	Validation	IEC	BASE 2
Normal	97.7	84.2	87.1	86.2
Electrical Fault	95.3	77.4	95.9	56.9
Thermal Fault	91.1	82.6	61.5	65.4
Third Analysis	Training/Validation (IEC+BASE1)		Test Set	
	Training	Validation	BASE2	
Normal	94.1	80.6	89.9	
Electrical Fault	98.5	83.8	70.5	
Thermal Fault	97.1	93.5	83.3	

TABLE 12 –INDIVIDUALIZED PERCENTAGE CONCORDANCE INDEX FOR TRAINING, VALIDATION AND TEST SETS FOR BALANCED AND NORMALIZED DATABASES (DT-BASED CLASSIFIER).

First Analysis	Individualized Percentage Concordance Index (%)			
	Training/Validation (IEC)		Test Sets	
	Training	Validation	BASE1	BASE2
Normal	100	100	67.9	81.2
Electrical Fault	96.3	61.2	87.8	57.1
Thermal Fault	98.2	100	77.6	75.1
Second Analysis	Training/Validation (BASE 1)		Test Sets	
	Training	Validation	IEC	BASE 2
Normal	97.2	84.8	88.3	90.4
Electrical Fault	95.3	96.7	98.6	63.1
Thermal Fault	93.9	90.2	77.3	68.7
Third Analysis	Training/Validation (IEC+BASE1)		Test Set	
	Training	Validation	BASE2	
Normal	100	90.6	93.3	
Electrical Fault	96.3	95.7	82.9	
Thermal Fault	98.9	97.2	84.7	

IV. DISCUSSION

In order to discuss the aforementioned results, we will split this section in two parts. In the first part, we will analyze the results with respect to processing techniques, database design, and input data representation. Yet in the second part, we will compare the performance of both AI-based classifiers.

A. Overall analysis

When TEST1 is compared to TEST3 (unbalanced data) and TEST2 to TEST4 (balanced data), it can be noticed that the balancing in the database increased the generalization ability of the classifiers. This is therefore a clear indication that the preprocessing of the database should be treated with attention to this problem.

On the other hand, three different analyses have been performed for each one of the six tests. The difference among them is related to the way that the available databases (IEC, BASE1, BASE2) are arranged in order to form the training/validation/test sets. For the first analysis, IEC database has been used to train and validate the neural classifiers, and BASE1 and BASE2 databases have been used for test purposes. In the second analysis, BASE1 has been used to train and validate the neural classifiers, and IEC and BASE2 databases have been utilized for test purposes. And, for the third analysis, IEC and BASE1 databases have been used for training and validation purposes, whereas BASE2 has been used to test the classifiers. After comparing the three analyses for each one of the six tests, it can be noticed that the best concordance indexes, both global and individualized ones, have been obtained in the third analysis.

This outcome suggests that when IEC and BASE1 databases are merged in the learning stage, the classification problem representation is enhanced, and, consequently, the generalization capacity of the classifiers is improved. Therefore, it can be inferred that the design of the database, both in terms of qualitative and quantitative point of view, positively influences the performance of the neural classifiers.

Finally, when TEST3 (five gases concentrations as inputs, global indexes) is compared to TEST5 (normalized inputs, global indexes) and TEST4 (five gases concentrations as inputs, individualized indexes) is compared to TEST6 (normalized inputs, individualized indexes), it can be noticed that the classifiers used in Tests 5 and 6 were more effective. This result highlights the importance of the input data representation for classification purposes.

B. Performance Comparison

Although both classifiers presented concordances rates over 70 % (mean values, Tables 9-12), the DT approach was superior most of the time in Tests 1-4. Nevertheless, in Tests 5-6, where the input data representation was improved, the DT approach was slightly better. These outcomes suggest that the DT approach, for this problem, is less sensitive to the input data representation than the NN approach.

V. CONCLUSION AND FUTURE WORK

The growing demand for electricity and the fact that some Electric Power Systems (EPS) have operated overloaded, make the efficient distribution task of the existing energy a crucial point for the electric utilities. The power transformer is an indispensable equipment in the EPS. If this equipment is out of order in an unplanned way, the damage for both society (load shedding) and electric utilities (fines due to unplanned interruption) are very significant. Hence, it is evident the importance of monitoring this equipment continuously.

In this work, Neural Networks and Decision Trees were used to classify incipient faults in power transformers. The presented procedures have been applied to real databases derived from chromatographic tests of power transformers. The outcomes of the best neural classifiers can be found in Tables 9 and 11. The obtained results in Table 9 show that the employed technique produced the following concordance rates: 60.3% (MIN.), 72.4% (MEAN), and 83.7% (MAX.). Yet in Table 11 it can be noticed that the neural classifiers produced figures 54.9% (MIN.), 73.5% (MEAN), and 95.9% (MAX.) as concordance rates. Yet the outcomes of the best DT-based classifiers can be found in Tables 10 and 12. The obtained results in Table 10 show that the employed technique produced the following concordance rates: 73.6% (MIN.), 80.18% (MEAN), and 87.8% (MAX.). Yet in Table 12 it can be noticed that the DT-based classifiers produced figures 57.1% (MIN.), 79.6% (MEAN), and 98.6% (MAX.) as concordance rates.

The main contributions of this paper are: 1) To provide an affordable tool for predictive maintenance of power transformers that can be used by both electric utilities and companies that run their own power systems. Although this subject has been studied for many years [14, 15], in many under-developed and semi-developed countries the corrective maintenance of the power transformers is still widely employed. 2) In many cases, industries have a rough power transformer database, but do not have the necessary information regarding the real situation of the transformer. The connection between a power transformer database and its physical interpretation has been discussed here to some extent as well. Further information regarding it can be found in [12], [25].

Differences among power transformers such as volume of the insulating oil, constructive aspects, voltage classes and environmental operation conditions, added to the inherent uncertainty in the chromatography process for power transformers, make it impractical the goal of error-free classification. Nevertheless, as it was highlighted in the previous section, tasks such as database setup, input data representation and data preprocessing, including the support of power transformers experts, can improve the generalization ability of the neural classifiers. Thus, in a future work, these topics shall be investigated more deeply.

ACKNOWLEDGEMENTS

The authors would like to thank PROPP/UFOP, and FG (Gorceix Foundation) for the financial support.

REFERENCES

[1] M. Duval, “A Review of faults detectable by gas-in-oil analysis in transformers,” *IEEE Electrical Insulation Magazine*, vol. 18, n.3, p.8-17, May/June 2002.

[2] S.-Wei Fei, Y. Sun, “Forecasting dissolved gases content in power transformer oil based on support vector machine with genetic algorithm,” *Electric Power Systems Research*, 78, 507-514, 2008.

[3] M.-Yuan Cho, Tsair-fwu Lee, Shih-wei Kau, Chin-shiuh Shieh, Chao-ji Chou. “Fault diagnosis of power transformers using svm/ann with clonal selection algorithm for features and kernel parameters selection,” *International Conference on Innovative Computing, Information and Control*, pp. 26-30, 2006.

[4] D. R. Morais and J. G. Rolim, “A hybrid tool for detection of incipient faults in transformers based on the dissolved gas analysis of insulating oil”, *IEEE Transactions on Power Delivery*, vol. 21, no. 2, April 2006.

[5] R. Naresh, Veena Sharma and M. Vashisth, “An integrated neural fuzzy approach for fault diagnosis of transformers”, *IEEE Transactions on Power Delivery*, Vol. 23, No. 4, pp. 2107-2024, Oct. 2008.

[6] W. Chen, Chong Pan, Yuxin Yun, Yilu Liu, "Wavelet networks in power transformers diagnosis using dissolved gas analysis", *IEEE Transactions on Power Delivery*, page(s): 187 - 194, Volume: 24 Issue: 1, Jan. 2009.

[7] Y. Zhang, X. Ding, Y. Liu, and P. J. Griffin, “An Artificial Neural Network Approach to Transformer Fault Diagnosis. *IEEE Transactions on Power Delivery*, 11 (4): 1836-1841, Aug. 2002.

[8] L. G. Castanheira, J. A. Vasconcelos, A. J. Rocha Reis, P. H. V. Magalhães, and S. A. L. Silva, “Application of Neural Networks in the Classification of Incipient Faults in Power Transformers: A Study of Case”, to appear in the Proceedings of the *International Joint Conference on Neural Networks 2011 (IJCNN 2011)*, 2011.

[9] F. Zhao and H. Su, “A Decision Tree Approach for Power Transform Insulation Fault Diagnosis”, 7th World Congress on Intelligent Control and Automation (WCICA), 2008.

[10] K. X. P. Lai and T. R. B. T. Blackburn, “Descriptive Data Mining of Partial Discharge using Decision Tree with Genetic Algorithm”, Australasian Universities Power Engineering Conference (AUPEC'08), 2008.

[11] D. R. Morais, “Intelligent tool for detecting incipient faults in transformers based on the analysis of dissolved gases in insulating oil,” Master Thesis, Federal University of Santa Catarina, Florianópolis, SC, Brazil., 2004 (in Portuguese).

[12] M. Duval, and A. DePabla, “The Interpretation of gas-in-oil analysis using IEC publication 60599 and IEC TC10 databases,” *IEEE Electrical Insulation Magazine*, vol.17, n.2, p.31-41, Mar./Apr. 2001.

[13] M. Riedmiller and H. Braun, “A direct adaptive Method for faster backpropagation learning: the Rprop algorithm,” *In Proc. IEEE International Conference on Neural Networks*, vol. 1, p. 586-591, 1993.

[14] H. Lu, R. Setiono, H. Liu. “NeuroRule: A Connectionist Approach to Data Mining,” *In: Proceedings of the 21st VLDB Conference*, 1995.

[15] Z. Wang, Y. Liu, P. J. Griffin, “A Combined ANN and Expert System Tool for Transformer Fault Diagnosis,” *IEEE Transactions on Power Delivery*, 13 (4): 1224- 1229, 1998.

[16] http://www.cs.waikato.ac.nz/ml/weka/ (Last access in June, 30th 2011.)

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA Data Mining Software: An Update”; SIGKDD Explorations, Volume 11, Issue 1, 2009.

[18] J. J. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann, 320 p, San Mateo, 2993.

[19] C. L. Curotto and N. F. F. Ebecken, “Decision Trees” (in Portuguese). http://curotto.com/vita/decisiontrees_2000/curotto_arvores_2000.pdf (Last access in June, 30th 2011.)

[20] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Academic Press, San Diego, 2001.

[21] H. Mannila. “Data Mining: Machine Learning, Statistics and databases.” *Proceedings of the Eighth International Conference on Scientific and Statistical Database Management* (SSDBM '96), 1996.

[22] M. Kubat and S. Matwin. “Addressing the curse of imbalanced training sets: one-sided selection.” In *Proceedings of the Fourteenth International Conference on Machine Learning* (ICML'97), pp. 179-186.

[23] B. Pahlavanpour and J. Nunes, “The most recent developments in the Dissolved Gas Analysis field. *Proceedings of the XIII ERIAC*,” May 2009.

[24] "IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers," IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991) , pp.C1-27, Feb. 2 2009.

[25] L. G. Castanheira, “Data Mining Techniques Applied to Problems of Pattern Classification.,” Master Thesis, Federal University of Minas Gerais, Belo Horizonte, MG, Brasil, 2008 (in Portuguese).C. L.