Diagnosing faults in power transformers with autoassociative neural networks and mean shift

Vladimiro Miranda, *Fellow IEEE,* Adriana R. Garcez Castro and Shigeaki Lima

***Abstract* – This paper presents a new approach to incipient fault diagnosis in power transformers, based on the results of dissolved gas analysis. A set of autoassociative neural networks or autoencoders are trained, so that each becomes tuned with a particular fault mode or no fault condition. The scarce data available forms clusters that are densified using an Information Theoretic Mean Shift algorithm, allowing all real data to be used in the validation process. Then, a parallel model is built where the autoencoders compete with one another when a new input vector is entered and the closest recognition is taken as the diagnosis sought. A remarkable accuracy of 100% is achieved with this architecture, in a validation data set using all real information available.**

***IndexTerms* – Transformer fault diagnosis, Dissolved Gas Analysis, autoassociative neural networks, mean shift, information theoretic learning.**

I. Introduction

This paper describes a new approach to the problem of fault detection and identification in power transformers, that reaches 100% accuracy: a diagnosis system based on a set of autoassociative neural networks. The new model gives indication of no-fault or normal condition of the transformer and, if a faulty condition is detected, it identifies the type of fault. This capacity has not been reached before.

Power transformer incipient fault diagnosis based on dissolved gas [1] analysis (DGA) has been attempted many times, due to the economic importance of potential equipment failure. It is a problem prone to be addressed by researchers since the publication an IEC norm (IEC 60599 [2]) and a seminal paper [3] that included a data base for diagnosed failures denoted IEC TC10.

A number of models have been proposed, adopting a diversity of techniques: expert systems [4], fuzzy set models [5], multi-layer feedforward artificial neural networks (ANN) [6][7], wavelet networks [8], hybrids fuzzy sets/ANN [9], radial basis function neural networks [10], Support Vector Machines (SVM) [11], Self-Organizing Maps (SOM) or Kohonen Neural Networks [12]. This listing, while not exhaustive, is quite representative.

All models addressed the problem of recognizing the type of incipient fault from the composition or ratios of dissolved gases but in many works the number of fault modes is rather limited, while in other works the validity of the result may be questioned given that notoriously few data samples (as low as 2 in some cases) were used in a testing procedure, when used at all. A problem yet remained, of discriminating between transformers with and without on-line tap changers (OLTC), whose action cause the contamination of oil: it is rather surprising that it is so difficult to find a publication addressing this distinction, especially in a unified approach with the identification of failure mode, when the publication IEC 60599 addresses this issue.

Instead of a single neural network discriminating all types of transformer condition, the idea behind the work reported in this paper is the following: assuming that vectors representing dissolved gas concentrations fall into different clusters, depending on whether there is a fault or not and on the type of fault, a set of autoassociative networks are trained to match each cluster and capture the characteristics of each data manifold. Then, when activated with a new input vector, one of the networks should be in tune and display a small error, while all the others should be out of tune and will display a larger input-output error. The healthy/faulty condition and the type of fault are thus identified by recognizing which autoencoder presents minimum error – i.e. which cluster is the input vector most similar to.

The paper also introduces a novelty to overcome the problem of lack of data: the densification of the data sets using the Information Theoretic Mean Shift algorithm. The IEC TC10 data available on DGA and failure modes are scarce, which means that any validation procedure would have to be based on a limited number of cases. With the new procedure called the *densification trick*, virtual data are created in a way that they are compatible with the original cluster of data. These virtual data can be used to train the neural networks with much more efficiency and accuracy and the scarce real data may be all used in the testing procedures.

II. Autoencoders

The new model relies on autoassociative neural networks, or simply *autoencoders*, which are special feedforward neural networks designed and trained in such a way that the output reproduces the input. With adequate training, an autoencoder stores in its weights the information about the non-linear manifold where data lie. Once trained, the autoencoder may be used as recognition machine – if a new data vector composition is compatible with the learned manifold, the autoencoder will produce an output with a small error

V. Miranda is with INESC TEC – INESC Technology and Science, coordinated by INESC Porto, and also with FEUP, Faculty of Engineering of the University of Porto, Portugal (vmiranda@inescporto.pt).

Adriana Castro is with UFPA – Federal University of Pará, Brazil (adcastro@ufpa.br).

Shigeaki Lima is with UFMA – Federal University of Maranhão, Brazil, and with INESC TEC, Portugal (shiilima@inescporto.pt)

regarding the input; however, if this vector is distinct from the global pattern of the data used for training, the autoencoder will return in the output a result not matching the input – therefore, the error will be high.

The concept of an autoencoder implies that the number n of neurons of the input and output layers is the same. The simplest autoencoder architecture has only one middle layer, with a distinct number m of neurons. As autoencoders have been firstly proposed to achieve data compression, it is traditional to find schemes in the literature with m < n – then, from input to middle layer the autoencoder achieves an effective compression from a space S to a space S'. Data can then be stored, represented by smaller encoded vectors whose components are the values of the output of the middle layer neurons. The transition from middle to output layer performs decompression and allows retrieving the stored compressed information. Fig. 1 illustrates the case where dim(S) > dim(S') or n > m.

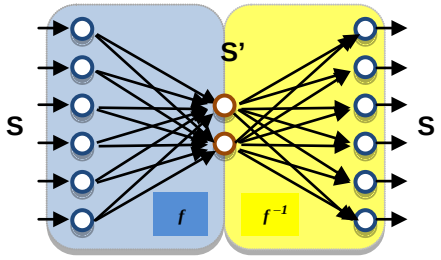

Fig. 1. An autoassociative neural network or autoencoder, with input and output layers of the same dimension and a smaller middle layer. If trained to reproduce the input variables in the output, one has in the middle layer a set of values that encode, in a different space S', the values in S. The function f achieves data compression and the function f^{-1} performs decompression.

This compression/decompression feature has been used to build data compression machines [13][14][15][16]. Also, the property that a point over the data manifold will be correctly projected back and forth by a trained autoencoder; and a point not lying on the data manifold will not be correctly re-projected back, producing a large input-output error ε, has been used in pattern recognition and classification, as well as in novelty detection [17].

It has been shown that autoencoders with linear activation functions produce an input/middle layer mapping equivalent to Principal Component Analysis (PCA) [18]. This is equivalent to say that information is projected along the direction of orthogonal axes (eigenvectors) such that variance is minimized. As with any data compression technique, there some amount of loss of information. However, when the activation functions are non-linear (sigmoidal), it has been shown that the mapping is not equivalent to PCA and has better characteristics [19].

Training an autoencoder has been treated no differently to training any other feedforward neural network, and backpropagation algorithms are currently used. Training is just an exercise on optimization and the classical cost function adopted is the Minimum Square Error. If **X** is the input vector and **Y** the output vector, then for N samples

$$\text{MSE}: \min \varepsilon = \frac{1}{N}\sum_{k=1}^{N}\left\|\mathbf{X}_k - \mathbf{Y}_k\right\|^2 \qquad (1)$$

A good interpretation of the MSE criterion is that it represents the minimization of the variance of the pdf (probability density function) of the error distribution. However, this criterion is optimal only if this distribution is Gaussian, which may be questionable in many applications. A non-parametric method should be preferred.

Because of the extremely large dimension that autoencoders may reach in problems of data compression (thousands of inputs and tens of thousands of weights to be tuned), experience has shown that achieving good convergence in training is rather difficult, requiring many attempts, a careful choice of initial weight values and many training epochs. Some researchers have proposed schemes of incremental training, especially when dealing with autoassociative networks having several middle layers, such as referred to in [13].

III. Information Theoretic Mean Shift and Cluster Densification

The Information Theoretic Mean Shift algorithm [20][21] was introduced as a means to capture the dominant structures in the data set, as embedded in its estimated probability density function (pdf).

Renyi's quadratic entropy [22] for a pdf is defined as

$$H(X) = -\log \int_{-\infty}^{+\infty} p^2(x)dx \qquad (2)$$

and the pdf p(X) can be estimated by the Parzen windows technique [23]

$$\hat{p}(X) = \frac{1}{N}\sum_{i=1}^{N} G_\sigma(x - x_i) \qquad (3)$$

where $G_\sigma(t) = e^{-\frac{t^2}{2\sigma^2}}$ is a Gaussian kernel having bandwidth $\sigma > 0$. Replacing (3) into (2) gives

$$H(X) = -\log V(X) \qquad (4)$$

with

$$V(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma'}(x_i - x_j) \qquad (5)$$

having $\sigma' = \sqrt{2\sigma}$. V(X) is called *information potential* of the pdf p(x). The derivative of this expression with respect to a single point x_i gives a quantity denoted *information force* exerted by all data particles on x_i [24][25][26].

The *cross entropy* between two pdf can be defined by

$$H(X, X_0) = -\log V(X, X_0) \qquad (6)$$

$$\text{with } V(X, X_0) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma'}(x_i - x_{0j}) \qquad (7)$$

The Cauchy-Schwartz distance measure between two pdf denoted p and q is

$$D_{CS}(X, X_0) = \log\left(\frac{\left(\int p^2(x)dx\right)\left(\int q^2(x)dx\right)}{\left(\int p(x)q(x)dx\right)} \right) \qquad (8)$$

and $$D_{CS}(X, X_0) = -\left[H(X) + H(X_0) - 2H(X, X_0)\right] \qquad (9)$$

The Information Theoretic Mean Shift algorithm aims at finding data sets X that capture structural information from a set X_0. This is achieved by a double criteria optimization, minimizing the entropy of X while keeping the Cauchy-Schwartz distance at some value k. An unconstrained optimization formulation, under a parameter λ that represents the trade-off between the two objectives, is given by

$$J(X) = \min H(X) + \lambda\left[D_{CS}(X, X_0) - k\right] \qquad (10)$$

Differentiating J(X) with respect to each x_i gives an algorithmic rule that allows the transformation of X_0 into another set at iteration t+1, making use of the information contained in the pdf of X at iteration t, estimated by (3):

$$x_i^{t+1} = \frac{c_1 \sum_{j=1}^{N} G_{\sigma'}(\| x_i^t - x_j^t \|) x_j^t + c_2 \sum_{j=1}^{N} G_{\sigma'}(\| x_i^t - x_{0j} \|) x_{0j}}{c_1 \sum_{j=1}^{N} G_{\sigma'}(\| x_i^t - x_j^t \|) + c_2 \sum_{j=1}^{N} G_{\sigma'}(\| x_i^t - x_{0j} \|)}$$

$$\text{where } c_1 = \frac{(1-\lambda)}{V(X)} \text{ and } c_2 = \frac{2\lambda}{V(X, X_0)} \qquad (11)$$

This works as if the information particles, in a field of information potential, move under the influence of the information forces such as the derivative of (5) referred to above.

When one makes $\lambda = 1$ in (11), it can be shown that the algorithm gives the modes of the pdf p(X). When λ is increased between 1 and 2, the algorithm tends to make points x to converge to the principal curve of the data, and a further increase in λ will concentrate points around the denser regions of the pdf. Furthermore, each generation of points x_i^t describe a pdf $p(X^t)$ that retains information from $p(X_0)$. Each point x_i^t along the iterations t describes a path from x_{i0} towards a mode of the pdf $p(X_0)$, or to a principal curve of the data cluster, or to a region of higher density, depending on the value of λ adopted. By path one means a succession of points $X_0, X^1, \ldots, X^t, \ldots$ that may be driven towards or away from the mode, depending on allowing points to follow the direction of the information force – $\partial V / \partial X$ as in (5) – or the reverse direction.

This property is used in this paper, in a novel way, to densify clusters X_0 by recovering the intermediate iteration points and using them as new virtual data points, compatible with the original pdf.

This *densification trick* becomes especially useful when data is scarce – and this happens with the information in IEC TC 10. For instance, for the mode of failure PD – Partial discharge the data base contains only 30 points; a usual neural network training practice would split these points into training and validation sets in a proportion 2/3 – 1/3. This would leave 10 points only to validate the neural network behavior, which is rather insufficient. As we shall see in next Section, the insufficient number of samples in transformer fault diagnosis studies is a difficulty present in many works reported – some papers report a number as low as 2 samples (!) to validate a proposed model. In particular, the solidity of models whose validation rests on such a low number of test samples may be questioned.

With the use of the Information Theoretic Mean Shift, the training set may be composed of only virtual points, keeping the totality of the real data to be used in the testing phase – this largely increases the robustness of the testing procedure and the confidence in the results it will provide.

In order to improve the chance that the network, after being trained with virtual points, may correctly generalize and have good results in the test set (composed with real data), the *densification trick* includes the generation of a few virtual points using the reverse direction of the information force vector, applied to each real data point. This means that the real data available for the test set will in no case be external points of the cluster. The results presented below and in Section VI were obtained with a single mean shift step outward, before resuming the convergence of points towards the cluster center. In the case study reported in this paper, this was enough.

IV. DGA DIAGNOSIS DATA IN POWER TRANSFORMERS

The presence of dissolved gases in the oil of power transformers is a well-recognized phenomenon, used to monitor the condition of the equipment. In the past, this procedure was used sporadically but nowadays one finds in place sensors that already provide an online continuous monitoring. The DGA – dissolved gas analysis, is therefore a powerful technique that should allow the confirmation of healthy states and the detection of incipient failures, when gas concentrations deviate from the healthy pattern. When this change is considered significant, other procedures are put in place namely to locate and deal with the fault.

A landmark publication is the norm IEC 60599 [2], introducing a fault classification summarized in Table 1. These rules represented an important advancement in spite of the fact that, when applied to the transformer data set IEC TC10 [3], they still produce a number of mistaken classifications plus a number of non-classified patterns (non-identified failures). Nevertheless, it must be said that the data set has been extremely useful in fostering research efforts towards achieving a more accurate classification.

The 6 faulty cases listed in Table 1 are associated with a diversity of cases in the data base. It is usual to find that the cases T2 and T3are lumped together in many studies, because the number of cases in the data base is too small for an adequate training of neural networks or building of knowledge bases. This data set published together with [3] included also many cases of "typical (normal) values in service", corresponding to transformers with and without a communicating OLTC (on line tap changer).

TABLE 1 – IEC 60599 FAULT DIAGNOSIS RULES

Case	Fault type	$\frac{C_2H_2}{C_2H_4}$	$\frac{CH_4}{H_2}$	$\frac{C_2H_4}{C_2H_6}$
PD	Partial discharge	NS	<0.1	<0.2
DL	Low energy discharge	>1	0.1-0.5	>1
DH	High energy discharge	0.6-2.5	0.1-1	>2
T1	Thermal fault – T<300°C	NS	>1 but NS	<1
T2	Thermal fault 300°C< T<700°C	<0.1	>1	1-4
T3	Thermal fault – T>300°C	<0.2	>1	>4

However, researchers concentrated their attention mainly in the discrimination of the fault types identified in Table 1 and ignored the possible distinction among healthy cases.

Table 2 presents a summary of sizes of databases and results published by different authors. It includes the size of the database and the size of the training set, as well as the % of success cases both in the training phase and in the test sets, when this information is available. It also identifies (last column) the number of outputs in each system, with N meaning the identification also of the normal or healthy state.

TABLE 2 – DATA AND RESULTS IN DISTINCT SYSTEMS/PUBLICATIONS

Model	Year	No. samples		% of correct diagnoses		No. faults
		Total	Test	Train	Test	
[6] Y Zhang et al	1996	40	(?)	(?)	**95**	3+N
[4] Wang	1998	188 + 22	60	99.3 to 100	**93.3 to 96.7**	5+N
[7] YC Huang et al	2003	220 + 600	0	95.12	---	4+N
[27] HT Yang, CC Liao	1999	561	280	93.88	**94.9**	4+N
[28] Guardado et al	2001	69	33	100	**100**	5+N
[29] Castro,Miranda	2005	431	139	100	**97.8**	3
[9] Miranda, Castro	2005	318	88	100	**99.4**	5
[30] G Lv et al	2005	75	25	100	**100**	3+N
[31] WH Tang et al	2008	168	(?)	(?)	**80**	3+N
[32] LX Dong et al	2008	220	60	(?)	**88.3**	3+N
[33] MH Wang et al	2009	21	0	100	---	8+N
[34] SW Fei, XB Zhang	2009	142	(?)	(?)	**94.2**	3+N
[35] NAM Isa et al	2011	160	40	100	**100**	3+N
[36] [37] Castro, Miranda	2011	318	88	100	**100**	5
[38] K Bacha et al	2012	94	30	(?)	**90**	6+N

This table does not constitute an exhaustive survey but provides a clear picture of the state of the art evolution. Different data sets were used and a direct comparison of percentage of hits/misses must be approached with care.

TABLE 3 – COMMENTS ON THE DISTINCT SYSTEMS/PUBLICATIONS

Model	Comments
[6] Y Zhang et al	ANN. Too few testing samples: presumed only 2-3 testing samples on average per mode.
[4] Wang	Expert System and ANN combined. No PD fault mode.
[7] YC Huang et al	ANN modified by Evolutionary Algorithm. No validation. Only 220 samples for fault cases, 600 for normal state.
[27] HT Yang, CC Liao	Fuzzy rule system. Use of additional 150 artificial data for 3 extra types of faults.
[28] Guardado et al	ANN trained with 5 gas ppm concentrations. Too few testing samples: only 5 testing samples av. per mode.
[29] Castro,Miranda	ANN and fuzzy rule system. No normal mode. Includes IEC TC10 data
[9] Miranda, Castro	ANN and fuzzy rule system. IEC TC10 data. No normal mode.
[30] G Lv et al	3 cascading SVMs. Data for 1 single transformer and not from a diversity of machines. Too few testing samples: only 2 samples for testing DH faults.
[31] WH Tang et al	Applies Parzen Windows and PSO.
[32] LX Dong et al	Applies a rough set classifier and the fusion of 7 wavelet neural networks. No PD mode.
[33] MH Wang et al	Couples the Extension Fuzzy Set theory with Genetic Algorithms. No validation. Too few samples: only 2 testing samples on average per mode.
[34] SW Fei, XB Zhang	Applies cascading SVM tuned with a Genetic Algorithm. No PD mode. No information on the size of test set, presumed small.
[35] NAM Isa et al	Couples a feed-forward neural network with k-means clustering.
[36] [37] Castro, Miranda	Autoencoders. No normal mode. Small number of test samples in some modes.
[38] K Bacha et al	Applies SVM. Too few samples: PD mode with only 2 samples, DL mode with only 3 samples, etc.

In order to facilitate such comparison, Table 3 includes a rough identification of the main tools used in building these diagnosis systems and some comments, especially focused on the validation procedures and dimension of test sets. It may be seen that a few works claimed 100% accuracy. However, either the number of types of faults was notoriously smaller than the number in the IEC 60599 publication, rendering the clustering exercise much simpler, or then the size of the test set was notorious smaller than advisable to confer validity/credibility to the method proposed.

The model in [36], by the same authors as this paper, was devoted to discriminating the type of fault *given that* a faulty condition is assumed. The work reported now is an extension of previous preliminary results and, while keeping as we shall see an accuracy of 100%, it also allows the distinction between healthy and faulty states, as well as making a distinction between transformers with and without OLTC (on-line tap changing)..

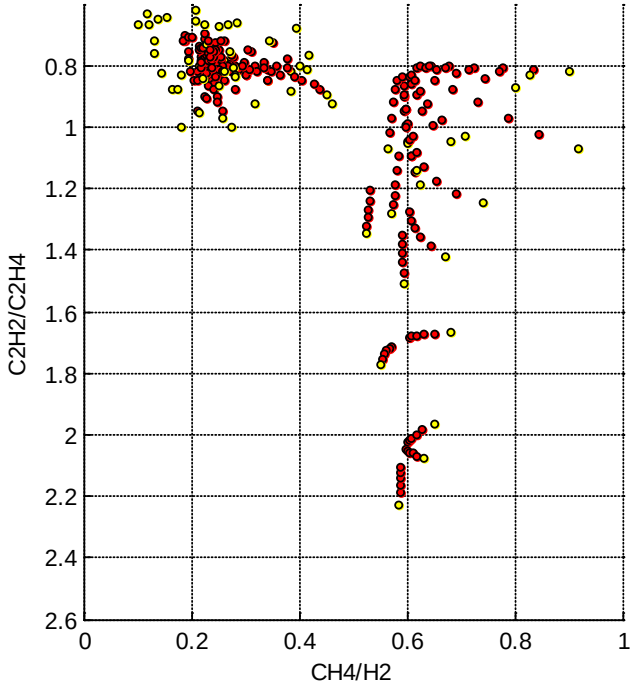

Fig. 2. Densification of the cluster for the D2 fault type (high energy discharges). Light spots: original data. Dark spots: new virtual points. Results obtained with λ=1 and σ=0.1 (kernel bandwidth), after generating 320 new points. This is a 2-D projection from 3-D data points.

In the work reported in this paper, the data base from [3] was used, complemented with data from other origin, comprising 318 cases for faulty states and 34 cases for healthy states. Each sample in the data base includes information of dissolved gas concentration of H_2 (hydrogen), CH_4 (methane), C_2H_6 (ethane), C_2H_4 (ethylene) and C_2H_2 (acetylene) as well as the verified condition of the transformer. Then, using the Information Theoretic Mean Shift algorithm, 1400 new virtual data points were generated.

Fig. 2 illustrates the densification process for a particular experiment, in a projection of 3-D data. It illustrates how the mean shift algorithm pushes data points along paths that converge to local modes of the pdf of data, as estimated by the Parzen windows technique (3) with a given kernel bandwidth. This convergence to local modes is induced by the use of coefficient λ=1 in (10).

This *densification trick* allowed one to populate the clusters and have available enough data for training the neural networks. The number of real data samples for each fault type and healthy condition is given in Table 4. The virtual data were used to form the training sets and all the real data were included in the test sets.

V. Diagnosis via Competitive Autoencoders

The most usual approach to automated fault diagnosis relies on a single classifier. Given a sample as input, it must produce an output signal indicating the proposed fault classification. The model described in this paper implements a different and perhaps more effective concept. The new idea is to match an autoassociative neural network to each cluster of data – either to each failure mode or to the healthy conditions.

Table 4 – Samples grouped for training and validation

	Types of fault/no fault	Training set (virtual)	Test set (real data)
T1	Thermal fault – T<300°C	200	77
T2	Thermal fault – T>300°C	200	71
PD	Partial discharge	200	30
DL	Low energy discharge	200	37
DH	High energy discharge	200	103
H0	Healthy states (no OLTC)	200	20
H1	Healthy states (with OLTC)	200	14

Each autoencoder is trained to store, in its weight matrix, the characteristics of a condition type. Then, when activated by an unclassified sample, only a specific autoencoder will "resonate" with it while the other will display large reconstruction errors. Arranged in a competitive setting, this cluster of autoencoders forms the new diagnosis system. Because of normal aging processes, dissolved gas concentrations are not stable and evolve with time. Transformers in a healthy condition but with different ages may present very distinct dissolved gas concentrations. This is why the best models use ratios instead of absolute concentrations, and both the IEC model and other models adopted such approach.

The same is followed in the new model: as in the IEC 60559 norm, the concentration ratios $(C_2H_2)/(C_2H_4)$, $(CH_4)/(H_2)$ and $(C_2H_4)/(C_2H_6)$ are used as characterizing vectors. The input space is therefore reduced to a space with 3 dimensions. In this case, it makes no sense to build an autoencoder with a smaller inner layer.

The new system adopts autoencoders with a 3-15-3 architecture, and sigmoid activation functions. Of course, this autoencoder does not perform data compression but this is not what is needed here. The large middle layer assures the necessary non-linear flexibility for the network to capture the features of the cluster it is meant and trained to learn.

The seven autoencoders (1 for healthy condition in transformers with no OLTC, 1 for transformers with OLTC and 5 for fault types) are then linked in a competitive parallel arrangement such as in Fig. 2. When a gas concentration ratio vector is shown to the system, each autoencoder will generate an attempt to reconstruct the input vector – but only one of these reconstructions will have a small error. A simple min box will then allow the selection of a winner. A word must be said about the concept of "error". While in this work the error is calculated using Eq. (1), in fact we should rather be talking of a similarity measure instead of error. Alternative measures to be used instead of the MSE would be based on information theory – that would translate into a numerical value the distance between the data manifold and the vector reconstruction produced by the autoencoder. Although theoretically a non-parametric method would be preferable, this has not been tested because the Square Error dissimilarity measure gave, in practice, a success rate of 100% meaning that it is a satisfactory approximation for all purposes in this problem.

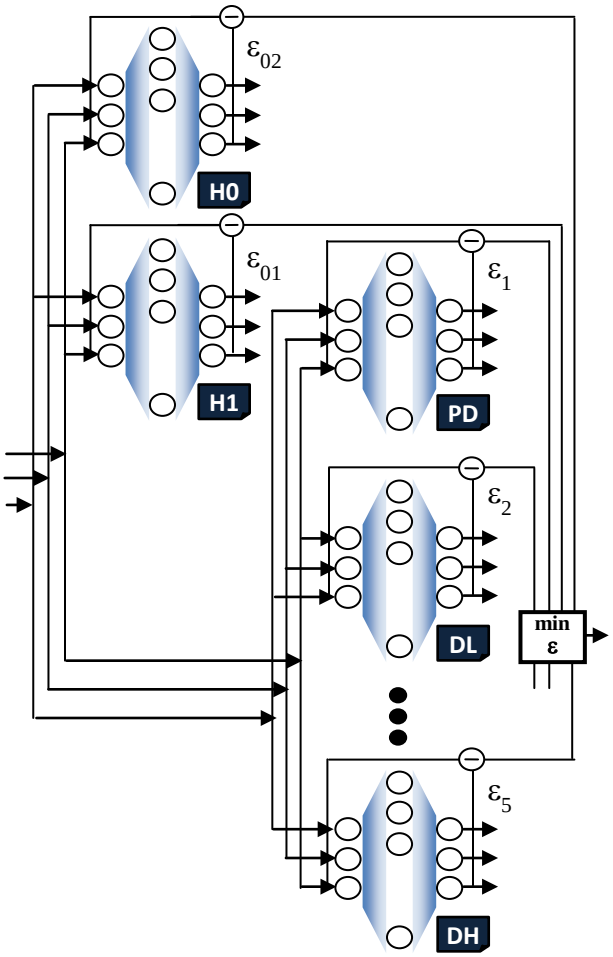

Fig. 3. General architecture of the new diagnosis system, based on a set of autoassociative neural networks activated in parallel. The top ones test for the healthy condition, with or without OLTC; the other networks in parallel are tuned each for a specific fault type. Their output errors compete to decide which network recognizes the input vector as laying on the manifold it emulates, by producing a minimal error.

VI. Training and Testing

A. Validation with real data

The error in each autoencoder was calculated as in Eq. (1), which is equivalent to a Euclidean distance between the two vectors (input and output). The activation functions used in the input and hidden layers were hyperbolic tangents, while in the output layer each neuron had a linear activation function. The training procedure adopted the Levenberg-Marquardt algorithm and was performed in Matlab. Table 4 presents the number of samples included in all training and validation sets for the 7 autoencoders.

Table 5 presents the results obtained, after training, in the discrimination of healthy/faulty condition. It is obvious that enough discriminating power was achieved, because there are no false answers: 100% accuracy was achieved. This is even more remarkable because the networks could discriminate between transformers with and without OLTC just from the oil samples.

Table 6 displays the results produced with the new Competitive Autoencoder Set, as well as the diagnosis obtained when applying IEC 60599 to the same data. A remarkable result emerges, not totally unexpected because it was already hinted in [36]: 100% accuracy in pinpointing the type of fault – or indicating a no fault condition. Notice that no errors or misclassification were produced by the new system (352 hits in 352 cases!). Furthermore, with the densification trick using the Information Theoretic Mean Shift algorithm, all real cases were tested positively! The comparative results from applying IEC 60599 indicate that the validation set was not especially easy to diagnose and that rectangular hulls (such as the ones induced by the application of the norm) are not the most convenient way to encapsulate or represent the clusters associated to the several fault types. Additionally, Table 6 also tells that the virtual data generated by the mean shift algorithm were not doctored to satisfy the IEC criteria.

To illustrate this result with a few examples, Table 7 presents some classification results, with the correct fault identification, the diagnosis produced by the autoencoder model and the result provided by IEC 60599 (NI – non-identified). The new system displays an absolute superiority over IEC 60599 and is better than any result reported and summarized in Table 2. In relation to IEC 60599, the autoencoder model was able to solve and correctly identify all undecided cases produced by this method.

The 100% hit is, indeed, a remarkable result. It can only be explained by the capacity of the autoencoders to really learn distinct manifolds for the distinct sets or clusters of data.

TABLE 5 – DIAGNOSIS ACCURACY IN DISCRIMINATING HEALTHY FROM FAULTY CONDITION

Model	% correctly identified healthy/faulty condition		No. cases with wrong diagnosis
	training set	testing set	
Miranda-Castro-Lima	100 %	100 %	0

TABLE 6 –DIAGNOSIS ACCURACY COMPARISON IN THE DISCRIMINATION OF FAULT TYPE

Model	% correctly identified faults		Total no. of non-identified faults or cases with wrong diagnosis
	Training (virtual)	Testing (real)	
Miranda-Castro-Lima	100 %	100 %	0
IEC 60599	93.00	95,28	85 in total (15 in real data)

TABLE 7 – EXAMPLES FROM THE IEC TC10 DATABASE. PERFORMANCE COMPARISON BETWEEN IEC 60599 AND THE AUTOENCODER SYSTEM

$\frac{C_2H_2}{C_2H_4}$	$\frac{CH_4}{H_2}$	$\frac{C_2H_4}{C_2H_6}$	Fault	Auto-encoder	IEC 60599
0.0417	1.1628	0.4444	T1	T1	T1
0.0198	1.8438	4.0	T2	T2	T2
0.0001	0.1102	0.0001	PD	PD	NI
1.1667	0.1065	0.1000	PD	PD	NI
0.0001	0.0476	0.0001	PD	PD	PD
1.0	0.1667	1.0	DL	DL	NI
4.0	0.1607	4.0	DL	DL	DL
0.6667	0.2250	4.0	DH	DH	DH

B. Sensitivity to noise

Given that dissolve gas concentrations are prone to experimental measuring errors, it has become usual that some sensitivity study is conducted.

Fig. 4 displays the results from applying to the database uniform noise of different bandwidths, by replicating 20 times each sample and contaminating every component of each vector with noise. It may be seen that even with a severe noise bandwidth of 20% around the real values (gross errors in all measured components) the accuracy level remained above 90%, a value comparable or above the results of many systems in Table 2 for data free of noise. In the past publications claiming 100% accuracy in their tests, this robustness to noise test could not be found.

This result adds confidence to the robustness of the new system devised.

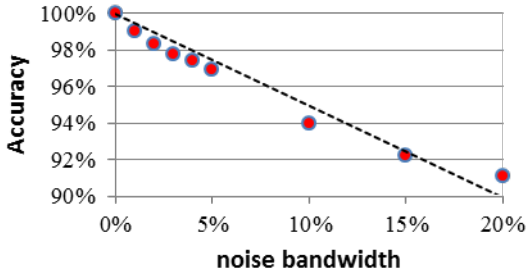

Fig. 4. Degradation of accuracy with growth in noise bandwidth

VII. CONCLUSIONS

The problem of condition diagnosis in power transformers, namely when subject to online monitoring of dissolved gases in oil, has taken a significant advancement with the work reported in this paper: a new system, composed of a set of competitive autoassociative neural networks, has achieved 100% accuracy in diagnosis – both in detecting healthy/faulty conditions and in pinpointing the exact failure mode in each case of faulty condition, while discriminating cases from transformers with and without OLTC.

The novelties in this work are threefold:

- First, it is the use of competing autoencoders to "resonate" with the correct type of condition.
- Second, it is the capacity of recognizing healthy states and discriminating them from faulty states, all in the same diagnosis system.
- And last, the idea of using the Information Theoretic Mean Shift algorithm to densify the data clusters, allowing all real data to be used in the test/validation phase of the process.

This idea, which was called the *densification trick*, is of general application. It was particularly useful in the case of the transformer fault diagnosis problem because the data sets for each failure mode or healthy condition are not large (are indeed very small). This causes problems in the correct training of neural networks.

For instance, the autoencoders adopted in this work required the tuning of 150 parameters (weights) each, but several data clusters had a low number of points – as low as 14. The application of the densification trick allowed one to enrich each cluster with many more points, while keeping the main structural information about the probability density function induced by the data, as estimated with the Parzen windows technique. We have thus used 352 real data samples to validate our model, by far the largest testing set ever used in all the publications referenced.

Furthermore, the confidence in the accuracy of the results is greatly enhanced. The elegance of the *densification trick* rests in the fact that all real data can be used for validation purposes, while in all previously proposed methods part of these data had to be employed in building the model and therefore could not be used to validate it.

The competitive architecture proposed for classification and diagnosis is completely general and its application is not restricted to transformer condition monitoring. In this case, a Euclidean similarity measure (square error) has proven effective and enough to discriminate results. However, in other applications it is possible that more sophisticated similarity measures should be necessary, both to train the autoencoders and to decide the winner among the competing results of the parallel set of autoassociative networks.

This architecture exhibits some advantages over a single neural network trying to accomplish the same diagnosis objective. Each module may be trained independently and therefore the knowledge of new data on one type of fault will only have impact on the retraining of one component, while the behavior for all other modes will remain intact. The fact that it requires the training of more neural networks is seen as negligible, given that it is made offline and the training times involved are in the order of a few minutes.

The remarkable success reported in this paper does not mean that an infallible machine has been built – for instance, there must be cases, in practice, where the simultaneous effects of more than one fault may blur the concentration ratios and render an ambiguous case for diagnosis. However, the results presented surpass, in accuracy, in number and type of fault/healthy conditions detected and in the validation effort, all that has been published so far.

ACKNOWLEDGMENT

This work is partially integrated in project LASCA PTDC/EEA-EEL/104278/2008 and in project GEMS PTDC/EEA-EEL/105261/2008, both financed by FCT, Portugal.

Shigeaki Lima acknowledges the support of CNPq (Brasil).

REFERENCES

[1] ANSI/IEEE Std C57.104.1991, IEEE Guide of gases generated in oil-immersed Transformer, IEEE Power Engineering Society, 1992.

[2] IEC Publication 60599, “Interpretation of the analysis of gases in transformers and other oil-filled electrical equipment in service”, March 1999.

[3] M. Duval and A. Pablo, “Interpretation of Gas-in-oil Analysis using new IEC Publication 60599 and IEC TC10 Databases”, IEEE Electrical Insulation Magazine, vol. 17, no. 2, pp. 31–41, March/April 2001.

[4] Z. Wang, Y. Liu, and P. J. Griffin, "A combined ANN and expert system tool for transformer fault diagnosis", IEEE Transactions on Power delivery, vol.13, pp. 1224–1229, October 1998.
[5] K. Tomsovic, M. Tapper and T. Ingvarsson, "A Fuzzy Information Approach to Integrating different Transformer Diagnostic Methods", IEEE Transactions on Power Delivery, vol. 8, no. 3, pp. 1638–1644, July 1993.
[6] Y. Zhang, X. Ding, Y. Liu and P. J. Griffin, "An Artificial Neural Approach to Transformer Fault Diagnosis", IEEE Transactions on Power Delivery, vol. 11, no. 4, pp 1836–1841, Oct. 1996.
[7] Y.-C. Huang. "Evolving Neural Nets for fault Diagnosis of Power Transformer", IEEE Transactions on Power Delivery, vol. 18, no. 3, pp 843–848, July 2003.
[8] L. Honglei, X. Dengming and C. Yazhu, "Wavelet ANN Based Transformer Fault Diagnosis Using Gas-in-Oil Analysis", in *Proc. of the 6th International Conference on Properties and Applications of Dielectric Materials*, 2000, pp. 147–150
[9] A. R. G. Castro and V. Miranda, "Improving the IEC Table for Transformer Failure Diagnosis with Knowledge Extraction from Neural Networks, IEEE Transactions on Power Delivery, vol. 20, pp. 2509–2516, Oct. 2005.
[10] J. P. Lee, D. J. LEE, P. S. Ji, J.Y. Lim and S. S. Kim, "Diagnosis of Power Transformer Using Fuzzy Clustering and Radial Basis Function Neural Network", 2006 International Joint Conference on Neural Networks, pp. 1398–1404, July 2006
[11] D-H Liu, J.-P. Bian and X.-Y. Sun, "The Study of Fault Diagnosis Model of DGA for Oil-Immersed Transformer Based on Fuzzy Means Kernel Clustering and SVM Multi-Class Object Simplified Structure", in Proc. of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, July, 2008, pp. 12–15.
[12] K. F. Thang and R. K. Aggarwal, "Analysis of Power Transformer Dissolved Gas Data Using the Self-Organizing Map", IEEE Transactions on Power Delivery, vol. 18, no. 4, pp. 1241, October 2003.
[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks", *Science*, Vol. 313, no. 5786, pp. 504 – 507, July 2006
[14] G.W. Cottrell, P. Munro and D. Zipse**r, "**Learning internal representations from gray-scale images: An example of extensional programming", *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, Seattle (WA), USA, 1987
[15] M.K. Fleming, G.W. Cottrell, "Categorization of faces using unsupervised feature extraction", *Proceedings of IJCNN - International Joint Conference on Neural Networks*, vol. 2, pp. 65-70, San Diego (CA), USA, 17-21 Jun 1990
[16] B. Golomb and T. Sejnowski, "Sex Recognition from Faces Using Neural Networks", in A. Murray (ed.), in *Applications of Neural Networks*, pp. 71-92, Kluwer Academic Publishers, 1995
[17] B. B. Thompson, R.J. Marks II, J.J. Choi, M.A. El-Sharkawi, M.Y. Huang and C. Bunje , " Implicit learning in autoencoder novelty assessment", *Proceedings of the 2002 International Joint Conference on Neural Networks*, 2002, IEEE World Congress on Computational Intelligence, Honolulu (Hawaii), USA, pp. 2878-2883, May12-17, 2002.
[18] Sanger, T.D., "Optimal unsupervised learning in a single-layer linear feedforward neural network", *Neural Networks*, Vol. 2, pp. 459-473, 1989
[19] N. Japkowicz, S.J. Hanson, M.A. Gluck, "Nonlinear Autoassociation is not Equivalent to PCA", *Neural Computation*, Vol. 12 , Issue 3, pp. 531 - 545, March 2000
[20] Sudhir Rao, Allan de Medeiros Martins, Weifeng Liu, Jose C. Principe, Information Theoretic Mean Shift Algorithm ", Intl. Work. on Neural Networks for Signal Processing", Maynooth, Ireland, pp. -, 9 2006
[21] Sudhir Rao, Allan de Medeiros Martins, Jose C. Principe, " Mean shift: An information theoretic perspective", Pattern Recognition Letters, no. 30, pp. 222–230, 2009.
[22] A. Renyi, "Some Fundamental Questions of Information Theory", *Selected Papers of Alfred Renyi*, vol 2, pp. 526-552, Akademia Kiado, Budapest, 1976.
[23] E. Parzen, "On the estimation of a probability density function an the mode", Annals Math Statistics, vol 33, 1962.
[24] J. C. Principe and D. Xu "Information-theoretic learning using Renyi's quadratic entropy", in J.-F. Cardoso, C. Jutten, and P. Loubaton, editors, Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation, Aussois, France, pages 407-412, 1999.
[25] D. Erdogmus and J. C. Principe, "Generalized Information Potential Criterion for Adaptive System Training", *IEEE Transactions on Neural Networks*, vol. 13, no. 5, September 2002, pp. 1035-1044
[26] R.A. Morejon, J.C. Principe, "Advanced search algorithms for information-theoretic learning with kernel-based estimators", *IEEE Transactions on Neural Networks*, vol. 15(4), pp. 874 - 884, July 2004
[27] H-T Yang and C-C Liao, "Adaptive Fuzzy Diagnosis System for Dissolved Gas analysis of Power Transformers", IEEE Transactions on Power Delivery, Vol 14, Nº. 4, pp. 1342-1350, October 1999.
[28] J.L Guardado, J.L Naredo, P. Moreno, adn C.R. Fuerte, "A Comparative study of neural networks efficiency in power transformers diagnosis using dissolved gas analysis", IEEE Transactions on Power delivery, Vol. 16, pp 643-647, October 2001.
[29] A. Castro and V. Miranda, "Knowledge Discovery in Neural Networks with Application to Transformer Failure Diagnosis", IEEE Transactions on Power Systems, Vol. 20, no.2, pp. 717-724, May 2005
[30] G. Lv, H. Cheng, H. Zhai, L. Dong, "Fault diagnosis of power transformer based on multi-layer SVM classifier", Electric Power Systems Research , Volume 75 , Issue 1 , Pages 9-15, July 2005
[31] W. H. Tang, J. Y. Goulermas, Q. H. Wu, Z. J. Richardson and J. Fitch, "A Probabilistic Classifier for Transformer Dissolved Gas Analysis With a Particle Swarm Optimizer", IEEE Transactions on Power Delivery, Volume 23, Issue 2, pp. 751-759, April 2008
[32] L. Dong, D. Xiao, Y. Liang, Y. Liu, "Rough set and fuzzy wavelet neural network integrated with least square weighted fusion algorithm based fault diagnosis research for power transformers", Electric Power Systems Research, Volume 78, Issue 1, pp. 129-136, January 2008
[33] M.H. Wang, Y.F. Tseng, H.-C. Chen, K.H. Chao, "A novel clustering algorithm based on the extension theory and genetic algorithm", Expert Systems With Applications , Volume 36 , Issue 4 , pp. 8269-8276, May 2009
[34] S.W. Fei, X.B. Zhang, "Fault diagnosis of power transformer based on support vector machine with genetic algorithm", Expert Systems with Applications, Volume 36 , Issue 2, pp. 11352–11357, 2009
[35] N.A.M. Isa, W.M.F.W.Mamat, "Clustered-Hybrid Multilayer Perceptron network for pattern recognition application", Applied Soft Computing Journal , Volume 11 , Issue 1 , Pages 1457-1466, January 2011
[36] A. Castro, V. Miranda and S. Lima, "Transformer fault diagnosis based on autoassociative neural networks", ISAP 2011, 16th International Conference on Intelligent Systems Applications to Power Systems , Crete, Greece, Sep 2011
[37] A. Castro and V. Miranda, "Sistema Inteligente para Diagnóstico de Faltas Incipientes em Transformadores baseado em Redes Neurais Auto-Associativas" (in Portuguese), Proceedings of SBSE 2010 – Simpósio Brasileiro de Sistemas Elétricos, Belém (PA), Brasil, May 2010
[38] K. Bacha, S. Souahlia, M. Gossa, "Power transformer fault diagnosis based on dissolved gas analysis by support vector machine", Electric Power Systems Research, Volume 83, Issue 1, pp. 73–79, February 2012 (on-line preview)

Vladimiro Miranda (M'90, SM'04, F'05) received his B.Sc. and Ph.D. degrees from the Faculty of Engineering of the University of Porto, Portugal (FEUP) in 1977 and 1982 in Electrical Engineering. In 1981 he joined FEUP and currently holds the position of Full Professor. He is also a researcher at INESC since 1985 and is currently Director at INESC Porto and INESC TEC (INESC Technology and Science), an advanced research institute in Portugal. He has authored many papers and been responsible for many international projects, namely with focus on wind power prediction in recent years, and all in areas related with the application of Computational Intelligence to Power Systems.

Adriana Rosa Garcez Castro graduated and obtained her M.Sc. degree in Electrical Engineering from UFPA (Federal University of Pará, Brazil) in 1992 and 1995. She finished her Ph.D. degree at INESC Porto and FEUP (Faculty of Engineering of the University of Porto, Portugal) in 2004. She is presently Assistant Professor at UFPA. Her interests are in Power Systems and Control and the application of Computational Intelligence techniques.

Shigeaki Lima graduated and obtained his M.Sc. degree in Electrical Engineering from UFMA (Federal University of Maranhão, Brazil). He is a PhD student at UFMA and is developing his thesis work in 2010/2011 at INESC Porto/INESC TEC, Portugal.