Faculdade de Engenharia da Universidade do Porto

FEUP

Diagnosing Faults in Power Transformers With Autoassociative Neural Networks and Mean Shift

Rafael Paiva Tavares

VERSÃO FINAL

Dissertação realizada no âmbito do
Mestrado Integrado em Engenharia Eletrotécnica e de Computadores
Major Energia

Orientador: Professor Vladimiro Miranda

Junho 2012

Resumo

Desde a Primeira Guerra Mundial que o diagnóstico de transformadores de potência é uma preocupação dos fabricantes e empresas do sector elétrico.

Por isso, vários métodos de diagnóstico foram propostos ao longo do tempo, sendo um dos mais conhecidos e utilizados o método de análise dos gases dissolvidos no óleo do transformador.

Hoje em dia, com a crise económica, a possibilidade de reduzir o custo de manutenção e também os custos de aquisição de novas máquinas é do agrado de todas as empresas do setor energético. Esta redução de custos permite que estas empresas tenham a possibilidade de reduzir o custo final da energia para o consumidor final, algo que é de extrema importância.

Nesta tese, vários métodos de diagnóstico utilizando o princípio da análise dos gases dissolvidos no óleo e diferentes ferramentas matemáticas são desenvolvidos e testados.

Devido aos dados disponíveis serem poucos, recorreu-se ao algoritmo de Mean Shift com o objetivo de criar dados virtuais para treinar as redes neuronais, sendo os dados reais utilizados apenas para o processo de validação do treino.

Palavras-chave: transformadores de potência, diagnóstico de avarias, análise de gases dissolvidos, redes neuronais auto associativas, algoritmo de mean shift

Abstract

Diagnosing power transformers as been since at least the First World War, a concern of both utilities and machine manufacturers. Therefore, several methods were proposed, being one of the most known the Dissolved Gas Analysis method.

Nowadays, with the economical crisis, the possibility to save money in maintenance and in the acquisition of new machines pleases the utilities. These savings also allow selling energy at a lower cost to the final consumer, being this one of the utilities objective.

In this thesis, several diagnosis methods are developed and tested, using different mathematical tools and the Dissolved Gas Analysis method principle.

Because of the sparse data, Information Theoretic Learning Mean Shift algorithm is used in order to create virtual points to train neural networks, leaving the real data only to its validation.

Keywords: Power transformers, fault diagnosis, dissolved gas analysis, autoassociative neural networks, mean shift algorithm.

Acknowledgements

First of all, I would like to present my most sincere 'Thank You' to my parents. They were the most supportive people in the last five years. I also know that financially supporting a kid living away from home during five years is a pretty hard task and I love you for that effort.

To my brother, with whom I shared a house again in the last two years, I know putting up with me can be a hard task, but you accomplished it well.

I also know that, in the last four or five months I have been impatient, bad-humoured and a stressed guy, but my girlfriend never showed the least sign of being tired of me. For that, I must thank you.

To all my friends who also had to deal with me in a regular basis, I admire your patience.

A very special thanks needs to go EFACEC, a Portuguese power transformer manufacturer, and to Eng. Jácomo Ramos as General Manager – Technology, for the interest in cooperating with this thesis work and authorizing the use of EFACEC's data. A warm and kind word is addressed to Dr. Maria Cristina Ferreira who was always extremely supportive and with generous availability to help me.

A very kind word needs to go to Professor Vladimiro Miranda. Most of the ideas in this thesis are his.

Table of Contents

Figures Index

Tables Index

Abbreviations and symbols

Abbreviations list:

ANN	Artificial Neural Network
DGA	Dissolved Gas Analysis
DH	High-energy Discharge
DL	Low-energy Discharge
GBMS	Gaussian Blurring Mean Shift
GMS	Gaussian Mean Shift
IEC	International Electrotechnical Commission
ITLMS	Information Theoretic Learning Mean Shift
MAE	Mean Absolut Error
OLTC	On Load Tap Changer
PD	Partial Discharge
pdf	Probability Density Function
std	Sandard Deviation
T1	Thermal Fault 150ºC<T<300ºC
T2	Thermal Fault 300ºC<T<700ºC
T3	Thermal Fault T>700ºC
TC	Technical Committee
TDCG	Total Dissolved Combustible Gas
UFPA	Universidade Federal do Pará

Symbols list

C	Carbon
CO	Carbon monoxide
C_2H_2	Acetylene
C_2H_4	Ethylene
CH_4	Methane
C_2H_6	Ethane
H	Hydrogen (radical)
H_2	Hydrogen (molecular)
T	Temperature
ºC	Celsius degrees
λ	Lagrange Multiplier
β	Quasi-random Gaussian number

Chapter 1.

Introduction

A power transformer is an electric machine generally used to change the voltage level in an electric power system.

Power transformers are a key component in any electric system. They are very expensive machines and, when there is a severe fault, the consequences can affect not only the machine itself but also surrounding facilities, equipment and people. Replacing or repairing a power transform, besides being very expensive, can also take a lot of time, which can make the consequences even worse. There are thousands of these machines in any generation, transmission and distribution system; therefore, their reliability is extremely important to maximize the energy sold and the global effectiveness and efficiency of the electric system.

Thus, any tool that can prevent a transformer to go out of service, minimize its repair cost or prevent accidents is very important and useful to utilities and transformer manufacturers.

It is known that when a fault occurs inside a power transformer, in its initial state, the consequences are very small, allowing the machine to work, and can be neglected. However, as time passes, those small faults can evolve to a more severe state that may not be reparable or may lead to the destruction of the machine. The main goal of any diagnosis system is to detect faults in their initial state and identify which type of fault occurred, if any. This allows the machine owner to analyse the situation and take preventive and corrective measures to maximize the power transformer lifetime. It is also important that these diagnosis methods can be applied while the machine is kept in service, because its disconnection may be very expensive and last for a long time. This latter point is related to the costs of the non-supplied energy.

This thesis has as main objective the studying, building and testing of new methods to diagnose power transformers. All these methods are based on the Dissolved Gas Analysis technique and the use of gas concentration ratios. Some other mathematic tools used are the Information Theoretic Mean Shift algorithm – ITLMS and neural networks, mainly a special kind of these networks often called auto associative neural networks or just autoencoders.

All these concepts and tools will be explained in following chapters of this document. It should be noticed that this work continues what has already been published in [1]. To develop this thesis, the most used software was the MATLAB R2011b version from Mathworks.

Chapter 2.

State of the Art

2.1. Dissolved Gas Analysis

The dissolved gases in power transformer oil are known to contain information about the condition of the machine. Utilities and manufacturers have been using this tool for several decades, since the First World War, to diagnose power transformers using a technique called Dissolved Gas Analysis (DGA) [1].

The power transformer oil is often mineral. Its role is insulation, thermic energy dissipation and constituting a dielectric environment. Since an oil sample can be retrieved without turning off the machine, a DGA technique allows diagnosing the condition of a power transformer at any time. This allows maximizing the power system reliability and maintenance scheduling before a small fault (incipient fault) evolves to a more severe state, like a non-reparable one. Because of this aptitude to expand the lifetime of power transformers, DGA has been recognized as such a powerful tool that, nowadays, is a standard to the electric industry worldwide.

The typical oil chemical composition is a mix of hydrocarbon molecules. Those are linked together with carbon-carbon and carbon-hydrogen bounds that can be disrupted by thermic or electric faults. When this occurs, some ions come free and recombine with other molecules, adding new chemical elements to the oil. Therefore, with machine usage and its consequent material and component degradation, the oil absorbs the gases released. This allows the inspection of the transformer condition. [2]

Different chemical compounds are created when the energy released by the fault varies, since low energy faults break the C-H bound, because they are weaker, while faults with higher energy break the C-C bounds. This means that the different gas concentrations provide information about the type of fault and its severity.

There are a lot of DGA methods using a mix of mathematic tools and different indicators. Single gas concentrations and total volume of gas in oil can be used, however there is also a key gas method, where each fault is related with one gas concentration or, the most used technique, which applies gas ratios to diagnose the transformer. In this last method there are two different techniques, the Doernenburg ratio method and the Rogers ratio method. Even though they follow the same general principle, they differ in the ratios used and the number of faults detected [3, 4].

Some methods like the International Standard IEC60599 set boundaries to gas concentration in order to classify different faults. To account for the limitations of rigid boundary definition, fuzzy set approaches have been proposed [5, 6]. The results achieved by these systems are promising, however the tuning process of the diagnosis rules may be difficult to handle. Because of the superior learning ability and the built-in power to handle data with error, Artificial Neural Networks (ANNs) have been used in DGA. These systems can be in continuous learning with new samples. However, the neural network training is often a slow process because it is sensitive to local minima presence and the backpropagation methods have trouble in dealing with this feature. [7]

2.1.1. IEC60599 Standard

The international standard IEC60599 - Mineral oil-impregnated electrical equipment in service – Guide to the interpretation of dissolved and free gases analysis [2, 8], is a milestone in the DGA methods. The last version of this document was released in 1999 and distinguishes six different faults:

- Partial Discharges (PD)
- Low Energy Discharges (D1)
- High Energy Discharges (D2)
- Thermal Faults, T<300ºC (T1)
- Thermal Faults, 300ºC<T<700ºC (T2)
- Thermal Faults, T>700ºC (T3)

And defines them as

- Partial Discharge – electric discharge where only a small part of the insulation is bridged with small perforations
- Discharge – electric discharge with total insulation bridge through carbonization (low energy discharge) and metal fusion (high energy discharge)
- Thermal fault – excessive temperature in the insulation. This fault can turn the insulation (T<300ºC), carbonize it (300ºC<T<700ºC), melt the metal and carbonize the oil (T>700ºC).

This publication assumes that every fault can be diagnosed using three gas ratios:

$$\frac{C_2H_2}{C_2H_4} \qquad \frac{CH_4}{H_2} \qquad \frac{C_2H_4}{C_2H_6}$$

where:

C_2H_2 is acetylene

C_2H_4 is ethylene

CH_4 is methane

H_2 is hydrogen

C_2H_6 is ethane

As said before, this standard sets intervals to these ratios, as shown in the table 2.1:

Table 2.1 - IEC60599 gas ratio intervals [2]

Fault	$\frac{C_2H_2}{C_2H_4}$	$\frac{CH_4}{H_2}$	$\frac{C_2H_4}{C_2H_6}$
PD	-	<0,1	<0,2
D1	>1	0,1 - 0,5	>1
D2	0,6 - 2,5	0,1 - 1	>2
T1	-	-	<1
T2	<0,1	>1	1 – 4
T3	<0,2	<1	>4

This way, the space is divided in parallelepipeds as can be seen in figure 2.1:

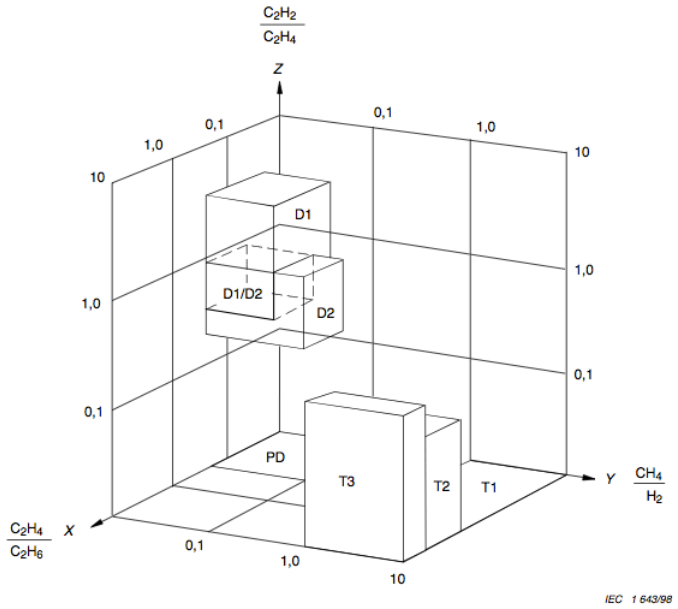


Figure 2.1 - IEC60599 graphical representation of gas ratios [2]

This standard also states that gas ratios should only be calculated if any of the gas concentrations is larger than the typical healthy values or the rate of gas increase is larger than the usual values, also published.

In addition to these faults, three more gas ratios are introduced:

$$\frac{C_2H_2}{H_2};\quad \frac{O_2}{N_2};\quad \frac{CO_2}{CO};$$

where the first one is related with the possibility of contamination of the OLTC compartment, the second one with an unusual heating of the gas and the third one with the cellulose degradation.

Diagnosing power transformers using this method has 93,94% of correct diagnoses [1].

2.1.2. IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers

The IEEE published a guideline paper in power transformers diagnosis, the IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers [3]. Its scope is very similar to the IEC standard, however the diagnosis according to this publication can be done in several ways: using individual and total dissolved combustible gas analysis (TDCG), by the key gas method and Doernenburg and Rogers ratios methods.

The first method defines sampling time intervals and operation procedures that depend of TDCG increase per day and TDCG value.

The key gas method uses the larger gas concentration in the oil to diagnose the machine. For example, when larges quantities of CO (carbon monoxide) are found in the transformers oil, this method diagnoses this fault as a thermal one.

As said before, the Doernenburg and Rogers method are very similar, however they differ in the ratios used. While the Doernenburg method uses five gas ratios, the Rogers method uses only three, not considering the Hydrogen concentration to diagnose the machine.

The Rogers method is the most similar to the IEC method because the number of ratios (and two of these) are the same, however one of them and the number of faults diagnosed differ, with the IEC method recognizing one more fault.

2.1.3. Other methods

Beside the two standards referred to above, there are other methods to diagnose power transformers using a DGA technique and several mathematical tools. These methods were developed before the publication of the standards and were used as basis to them or were developed after those milestones in order to improve the results of diagnosis made using these standards.

There are methods that use artificial neural networks and hybrid fuzzy sets [9, 10], expert systems [11], Support Vector Machines [12], self-organizing maps or Kohonen neural networks [13], fuzzy set models [14], wavelet networks [15], radial basis function neural network [6], multi-layer feedforward artificial neural networks [7, 16].

One of those methods was published by Wang et all in [11]. In this method a combined expert system and neural network is used. This way, the advantages of neural networks are combined with the human expertise.

In 1999, Yang et all [5], published a paper suggesting the use of an adaptive fuzzy system to diagnose power transformers. The system was presented as a self-learning one, using also the rules of the Doernenburg.

Other method using neural networks was published by Zhang in [16]. This method, even though a single neural network is used to diagnose the major fault types; it is also used an independent neural network to diagnose the cellulose condition.

Huang, in [7], also proposed the use of neural networks to diagnose power transformers. However, to train these neural nets an evolutionary algorithm is used. This way, the optimal connection weights and bias can be found easily and the disadvantages of the steepest-decent method are avoided.

In order to make explicit the implicit knowledge stored in neural networks after the training, Castro et al. published in [17] an algorithm to transform a neural network black box action in a set of explicit rules. After that, this method was applied in a power transformer diagnosis method.

In [1], Shigeaki suggests the use of seven autoencoders that work in parallel, that is, in a competitive way. This method will be explained later in this thesis because it will be used. In the following table, one can find a summary of all methods referred here, as well as some others methods and the corresponding results. This table is the one below (Table 2.2):

Table 2.2 – Data, results and comments of distinct systems / publications [1]

Model	Year	No. Samples		% of correct diagnoses		No. Faults	Comments
		Total	Test	Training	Test		
Y Zhang et al [16]	1996	40	?	?	95	3+N	ANN. Too few testing samples: presumed only 2-3 testing samples on average per mode.
Wang [11]	1998	188+22	60	99,3 to 100	93,3 to 96,7	5+N	Expert System and ANN combined. No PD fault mode.
YC Huang et al [7]	2003	220+600	0	95,12	--	4+N	ANN modified by Evolutionary Algorithm. No validation. Only 220 samples for fault cases, 600 for normal state.
HT Yang, CC Lao [5]	1999	561	280	93,88	94,9	4+N	Fuzzy rule system. Use of additional 150 artificial data for 3 extra types of faults.
Guardado et al [18]	2001	69	33	100	100	5+N	ANN trained with 5 gas ppm concentrations. Too few testing samples: only 5 testing samples av. Per mode.
Castro, Miranda [17]	2005	431	139	100	97,8	3	ANN and fuzzy rule system. No normal mode. Includes IEC TC10 data.
Miranda, Castro [9]	2005	318	88	100	99,4	5	ANN and fuzzy rule system. IEC TC10 data. No normal mode.
G LV et al [19]	2005	75	25	100	100	3+N	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.
WH Tang et al [20]	2008	168	?	?	80	3+N	Applies Parzen windows and PSO.
LX Dong et al [21]	2008	220	60	?	88,3	3+N	Applies a rough set classifier and the fusion of 7 wavelet neural networks. No PD mode.
MH Wang et al [22]	2009	21	0	100	--	8+N	Couples the Extension Fuzzy Set theory with Genetic algorithms. No validation. Too few samples: only 2 testing samples on av. per mode.
SW Fei, XB, Zhang [23]	2009	142	?	?	94,2	3+N	Apples cascading SVM tuned with a Genetic Algorithm. No PD mode. No information on the size of test set, presumed small.
NAM Isa et al [24]	2011	160	40	100	100	3+N	Couples a feed-forward neural network with k-mean clustering.
Castro, Miranda [1] [25]	2011	318	88	100	100	5	Autoencoders. No normal mode. Small number of test samples in some modes.
K Bacha et al [26]	2012	94	30	?	90	6+N	Applies SVM. Too few samples: PD mode with only 2 samples, DL mode with only 3 samples, etc.

When analysing the table above, one must take into account that different data sets were used in all the methods. Therefore, a comparison between the methods must be done carefully. Nevertheless, this table gives an idea how each method behaves when diagnosing a power transformer.

2.1.4. Alternatives offered in the market

Since dissolved gas analysis is an industrial standard to diagnose power transformers, there are several companies that provide this service. These companies can be power transformers manufacturers that supply their costumers with the possibility of, as a regular procedure, analyse the transformers' oil to allow early problem detection. This is the most usual approach do analyse the gas concentrations dissolved in oil and is done by companies like EFACEC. However there are monitoring systems that can be installed in the power transformer itself in order to maintain a continuous surveillance of the machine health state. This way, there is no need to periodically get an oil sample to analyse because this is done online and in situ. Companies like SIEMENS offer this service to costumers [27].

There are also companies that don't produce transformers but sell diagnosis of power transformers to its owners even though the manufacturers offer this service. Some examples are DOBLE Engineering [28] and POWERTECH Labs [29].

Almost all of these companies make use of one or a mix of both of the standard described before, however, they make small changes based in know-how acquired with experience. Usually these changes are made in the gas concentrations limits or the ratios values. In spite of this, there are companies who have also developed their own diagnosis methods or complemented the standards with more analysis like the dissolved metals in oil.

2.2. Kernel density estimation

Knowing that the probability P of a vector y fall in a region Θ is given by:

$$P = \int_{\Theta} p(x')\partial x' \qquad (2.1)$$

it is possible to estimate a smoothed version of p estimating the value of the probability P [30].

Kernel density estimation, namely the Parzen window technique [31], is a popular non-parametric method for estimating the probability density function of a data set [32]. The idea behind this estimation is very simple: placing a kernel over the samples and interpolating, with the proper normalization, gives the density estimation in each point. Therefore, the contribution of each sample in the estimation of the pdf is done in accordance to its distance to the point where the kernel is centred. If this is done to all the data set, one is able to estimate the complete pdf.

The kernel density estimation for a point of the data set in a d-dimensional space is given by:

$$\boldsymbol{p}(\boldsymbol{x},\boldsymbol{\sigma}) = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{G}\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x_i}}{\boldsymbol{\sigma}}\right\|^2\right) \qquad (2.2)$$

where G is the kernel, σ is the kernel bandwidth and N is the number of data points.

The kernel bandwidth has serious implications in the results given by this method. When σ is large, the closest samples have a very small weight, therefore, it is assumed that the pdf is a smooth, slowly changing function and the result will be a function with little resolution. On the other hand, when σ is smaller, the resulting pdf is a noisy one, with peaks centred in the samples. In this case the resolution is bigger, however, it can be affected by statistical variability. This way, one needs to seek some compromise between both cases [30, 33].

The estimation of the probability density function is a very useful mathematical tool to deal with discrete data sets because it allows transforming them in a continuous probability density function.

2.3. Mean shift algorithm

The Mean Shift algorithm is a mathematic formulation that allows analysing arbitrarily structured feature spaces. It can be used to find the modes or the principal curves of the datasets. In this algorithm the dataset is represented by its probability density function, where the modes can be found in the maximum of this function. It is a very versatile and robust algorithm in feature space analysis [32] and is often used in image segmentation [34, 35], denoising, tracking objects [36] and a several other computer vision tasks [37, 38].

Fukunaga and Hostler firstly introduced a Mean Shift algorithm in 1975 [39]. In this paper they showed that this algorithm is a steepest descent technique where the points of a new dataset are moving in each iteration towards the modes of the original dataset.

This first version of mean shift algorithm says that, if we consider a dataset $X_0 = (X_i)_{i=1}^N \in R^D$, using a Gaussian kernel given by $G(t) = e^{-\frac{t}{2}}$ and the Parzen window technique, we are able to estimate the probability density function using:

$$\boldsymbol{p(x,\sigma) = \frac{1}{N}\sum_{i=1}^{N} G\left(\left\|\frac{x - x_i}{\sigma}\right\|^2\right)} \quad \textbf{(2.3)}$$

where σ is the kernel bandwidth that is always bigger than 0. As mentioned before, the objective of this algorithm is to find the modes of the dataset where $\nabla p(x) = 0$. With that in mind, the iterative fixed-point equation is:

$$m(x) = \frac{\sum_{i=1}^{N} G\left(\left\|\frac{x - x_i}{\sigma}\right\|^2\right) x_i}{\sum_{i=1}^{N} G\left(\left\|\frac{x - x_i}{\sigma}\right\|^2\right)} \quad \textbf{(2.4)}$$

The difference $m(x) - x$ is known as mean shift.

This algorithm is known as Gaussian Blurring Mean Shift (GBMS) and it is unstable because the actual solution is a single point that minimizes the overall entropy of the data set – therefore, the modes cannot be confidently discovered.

In spite of this important development, the Mean Shift idea was forgotten until 1995, when Yizong Cheng [40] introduced a small change in the algorithm. While in Fukunaga's algorithm the original dataset is forgotten after the first iteration, $X^{(0)} = X_0$, Cheng algorithm keeps this dataset in memory. This initial dataset is used in every iteration to be compared with the new dataset Y. However Y is initialized the same way, $Y^{(0)} = X_0$. This also introduces a small change in the iterative equation:

$$m(x) = \frac{\sum_{i=1}^{N} G\left(\left\|\frac{x - x_{0i}}{\sigma}\right\|^2\right) x_{0i}}{\sum_{i=1}^{N} G\left(\left\|\frac{x - x_{0i}}{\sigma}\right\|^2\right)} \quad \textbf{(2.5)}$$

In literature, Fukunaga's algorithm is known as Gaussian Mean Shift Algorithm (GMS).

In 2006, Sudhir Rao, Jose C. Principe and Allan de Medeiros Martins [41-43] introduced a new formulation of mean shift known as Information Theoretic Mean Shift (ITLMS) and showed that GBMS and GMS are special cases of this one.

The idea in this algorithm was to create a cost function that minimizes the cross entropy of the data while the Cauchy-Schwartz distance is kept at a given value.

Knowing that a Gaussian kernel is given by:

$$G_\sigma = e^{\frac{-x^2}{2\sigma^2}} \quad (2.6)$$

an estimation of a probability density function (pdf), using the Parzen window technique, is:

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} G_\sigma \|x - xi\| \quad (2.7)$$

Renyi's quadratic entropy of a probability density function can be calculated using:

$$H(x) = -\log \int_{-\infty}^{+\infty} p^2(x)dx \quad (2.8)$$

So,

$$H(x) = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma'}\|x - x_i\|\right) \quad (2.9)$$

where $\sigma' = \sqrt{2}\sigma$.

In the literature, $H(x)$ is known as the information potential of a probability density function. This name was given because $H(x)$ resembles a potential field and its derivative reminds of forces between particles in physics. Hence, the derivate of $H(x)$ in each point gives the information force.

To measure the cross entropy between two pdf, one has:

$$H(x, x_0) = -\log\big(V(x, x_0)\big) = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma'}\|x_i - x_{0j}\|\right) \quad (2.10)$$

The Cauchy-Schwartz distance can be calculated using:

$$D_{CS}(x, x_{0j}) = \log\left[\frac{(\int p^2(x)dx).(\int q^2(x)dx)}{(\int p(x)q(x)dx)}\right] \quad (2.11)$$

where p and q are pdfs.

Using all the definitions above, it is possible to formulate a cost function to minimize the cross entropy between the two pdfs while keeping the Cauchy-Schwartz distance at a value k:

$$\mathrm{F(x)} = \big(\min H(x)\ and\ D_{CS}(x, x_0)\big) = k \quad (2.12)$$

Using a Lagrange multiplier to transform a constrained cost function in an unconstrained one:

$$\mathrm{F(x)} = \min[H(x) + \lambda(D_{CS}(x, x_0) - k)\,] \quad (2.13)$$

and differentiating at each point:

$$\frac{dF}{dx_i} = x_i^{t+1} = \frac{c_1 * S_1 + c_2 * S_2}{c_1 * S_3 + c_2 * S_4} \quad (2.14)$$

where:

$$c_1 = \frac{1-\lambda}{V(x)} \quad , \quad c_2 = \frac{1-\lambda}{V(x, x0)} \quad (2.15)$$

and:

$$S_1 = \sum_{j=1}^{N} G_\sigma \left(\frac{\left\| x_i^t - x_j^t \right\|^2}{\sigma'} \right) * x_j^t \quad \textbf{(2.16)}$$

$$S_2 = \sum_{j=1}^{N} G_\sigma \left(\left(\frac{\left\| x_i^t - x_{0j}^t \right\|^2}{\sigma'} \right) \right) * x_{0j} \quad \textbf{(2.17)}$$

$$S_3 = \sum_{j=1}^{N} G_\sigma \left(\frac{\left\| x_i^t - x_j^t \right\|^2}{\sigma'} \right) \quad \textbf{(2.18)}$$

$$S_4 = \sum_{j=1}^{N} G_\sigma \left(\frac{\left\| x_i^t - x_{0j}^t \right\|^2}{\sigma'} \right) \quad \textbf{(2.19)}$$

As shown by Sudhir Rao, Weifeng Liu, Jose C. Principe and Allan de Medeiros Martins, adjusting the λ parameter changes the data properties sought by the algorithm:

- When λ=0 the algorithm minimizes the data entropy, returning a single point. This is the GBMS algorithm.
- When λ=1, the algorithm is a mode seeking method, the same as GMS.
- When λ>1, the principal curve of the data is returned (1<λ<2). A higher value of λ makes the algorithm seek to represent all the characteristics of the pdf.

2.3.1. Iterative algorithm

With Rao's formula for x_i^{t+1}, it is possible to build an iterative algorithm where each point of the data set 'travels' in the data domain until it reaches a stable point where the information force is zero. This is the solution of the method and it can be a mode or a point that belongs to the principal curve of the data, depending on which value of λ is been used.

To stop this algorithm, the simplest way, is to calculate the distance (d) between the points given in t and $t+1$ iterations and interrupt the algorithm when d is smaller than a tol level for k consecutive iterations.

2.3.2. Steepest descent algorithm

Another way to build a mean shift algorithm is to apply the rules of steepest descent algorithms, building the equation 2.20:

$$x_i^{t+1} = x_i^t + \eta \frac{d}{dx_i} J(x) \quad \textbf{(2.20)}$$

In this case $J(x)$ is the formula of the unconstrained optimization problem that can be obtained using a Lagrange multiplier:

$$J(x) = \min H(x) - \lambda (D_{CS}(x, x_0) - k) \quad \textbf{(2.21)}$$

Differentiating $J(x)$:

$$\frac{d}{dx_i} J(x) = c_1 * F(x_i) - c_2 F(x_i, x_0) \quad \textbf{(2.22)}$$

Using this technique, one is able to adjust the η parameter in the equation 2.20, which is the step of the iteration. With this capability it is possible to control how the points travel in the data's domain, slower or faster, and the direction of the movement.

2.4. Autoassociative neural networks

Autoassociative neural networks or just autoencoders are a subtype of feedforward neural networks. These are built and trained in such way that the output vector is the same, or almost the same, as the input vector. In other words, autoencoders work as recognition machines. Since the target output vector is equal to the input vector, the number of output and input neurons is always the same. In the simplest autoencoders, there is but one hidden layer however there can be more. This inner layer usually has a smaller number of neurons than the output and input layers, however it does not need to be always like that. When the hidden layer has a smaller number of neurons, called a bottleneck, the data is compressed between the input and hidden layers and decompressed between the hidden layer and the output layer, since the vector's size is change to a smaller one. This kind of neural network is often used to compress data [44-47], restore missing sensor information [48, 49] and several other tasks [50, 51].

One interesting property of autoencoders is that when the network is trained to recognize a pattern, if an input vector with different characteristics is shown to the network, the error between the output and input tends to be high. This is extremely useful in pattern recognition tasks as detecting and restoration missing sensor information [52] and it is very important to the work done in this thesis.

Training an autoencoder is, in everything, similar to training any other neural network. The most used method to do it is known as backpropagation algorithm. In this algorithm the connection weights are adjusted in order to minimize a cost function, usually the minimum square error between the input and output vectors [53]. To train a neural network, two independent data sets are needed: one to train the network and other to validate its results. While the training set is used to adjust de connection weights, the validation set is used to verify if the network is generalising in a proper way. Generalising is the neural network ability to recognize points with the same properties of the ones in the training set but that didn't belong to it.

Below, in the figure 2.2, a diagram showing the typical autoencoder architecture is shown:

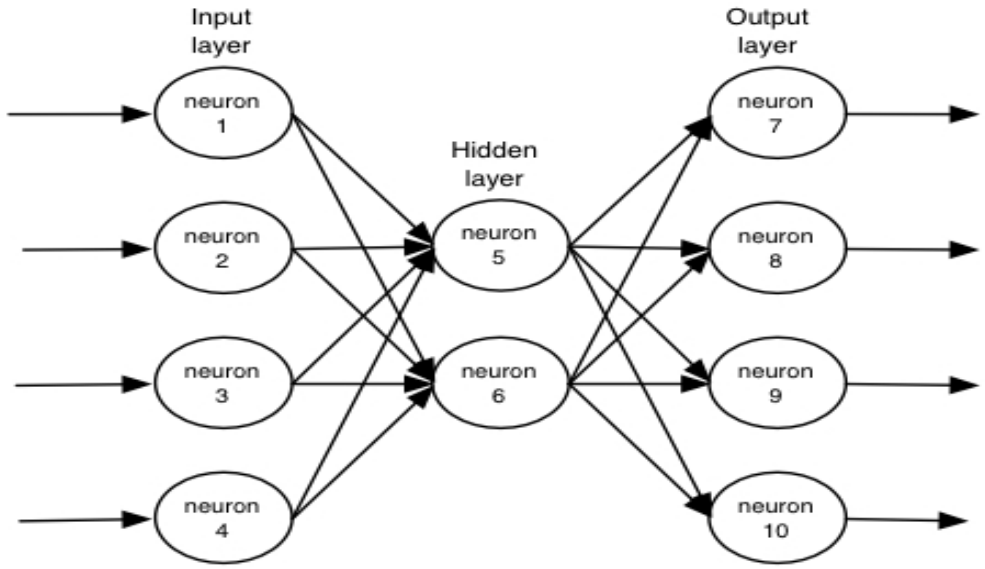


Figure 2.2 - Example of a possible autoencoder architecture

Chapter 3.

Densification of data sets

3.1. The database

To proceed with the study, was used a database containing actual dissolved gas analysis data and the corresponding diagnoses. It is a compilation of data from the IEC TC10 database and from several other origins and it was kindly provided by Prof. Adriana Castro from UFPA.

The real cases database consists in:

Table 3.1 - Number of real cases per fault in the database

Case	Fault/State	No. of samples
PD	Partial Discharge	30
DH	High Energy Discharge	103
DL	Low energy Discharge	37
T1	Thermal Fault (T<700ºC)	77
T2	Thermal Fault (T>700ºC)	71
OK	Healthy State without OLTC	20
OK with OLTC	Healthy State with OLTC	10
	Total:	348

These seven faulty/healthy states shown in table 3.1 are the ones that all the diagnosis methods studied in this thesis are built to distinguish. That is the reason why there are always seven neural networks in the systems with neural networks.

When the three gas concentration ratios are calculated, some modifications are made in order to normalize the data: ratios impossible to calculate because infinite is returned are set as 0,0001 and ratios bigger than 4 are limited to the value 4. This allows one to keep all the ratios inside limit values in order make easier the neural network train and the ITL mean shift algorithm application.

The data in this database were only used in the validation of the neural networks training. As it will be explained, the data used in the neural network training were virtual data created using the ITLMS.

3.2. Information Theoretic Learning Mean Shift algorithm applications

In this thesis, the ITL Mean Shift algorithm was used to several different tasks.

The main use given to the method was to achieve the generation of new items of information sharing some statistical properties with the original cluster of real data: this will be called the 'densification trick'; i.e., to increase de number of data in the database. These virtual points were used to train neural networks.

Mean Shift algorithm was also used to find the modes of the probability density functions associated to the different data clusters, corresponding to each healthy or faulty state, and to find the modes of the pdf of the whole dataset. Below all these procedures are explained and illustrated with some representative images. Other images can be found in the appendixes section.

3.2.1. Using the ITLMS as a mode seeking tool

As explained before, to use the mean shift algorithm as a mode seeking tool, one has to use $\lambda = 1$. This way, if the number of iterations is enough and the parameter σ is well set, the algorithm will converge the modes of the probability density function.

The parameter σ can be seen as the window around the data point where the 'neighbours' have influence in the force applied to the point. When a large value is used, there are a large number of other points that influence the first one. In this case, the probability of getting one single mode is high. On the other side, when the value is small, there can be just a few or no points influencing the force applied to the new point; hence, local modes can be found or, in the uttermost situation, there will be so many modes as points in the dataset and they will be exactly the same. This way, adjusting this parameter can be a little tricky and it was done by trial and error. Using a σ equal to the mean standard deviation of the cluster proved to return very accurate results and was used most of the time.

This algorithm only stops when the distance between the outputs of three consecutive iterations is smaller than a toll value. The value used was 10^{-10}.

The figure 3.1 represents the result of the mean shift algorithm applied to the low-energy discharge cluster where the original data points are in solid red and the blue circle indicates the mode. This was obtained using σ equal to the mean standard deviation of the original points.

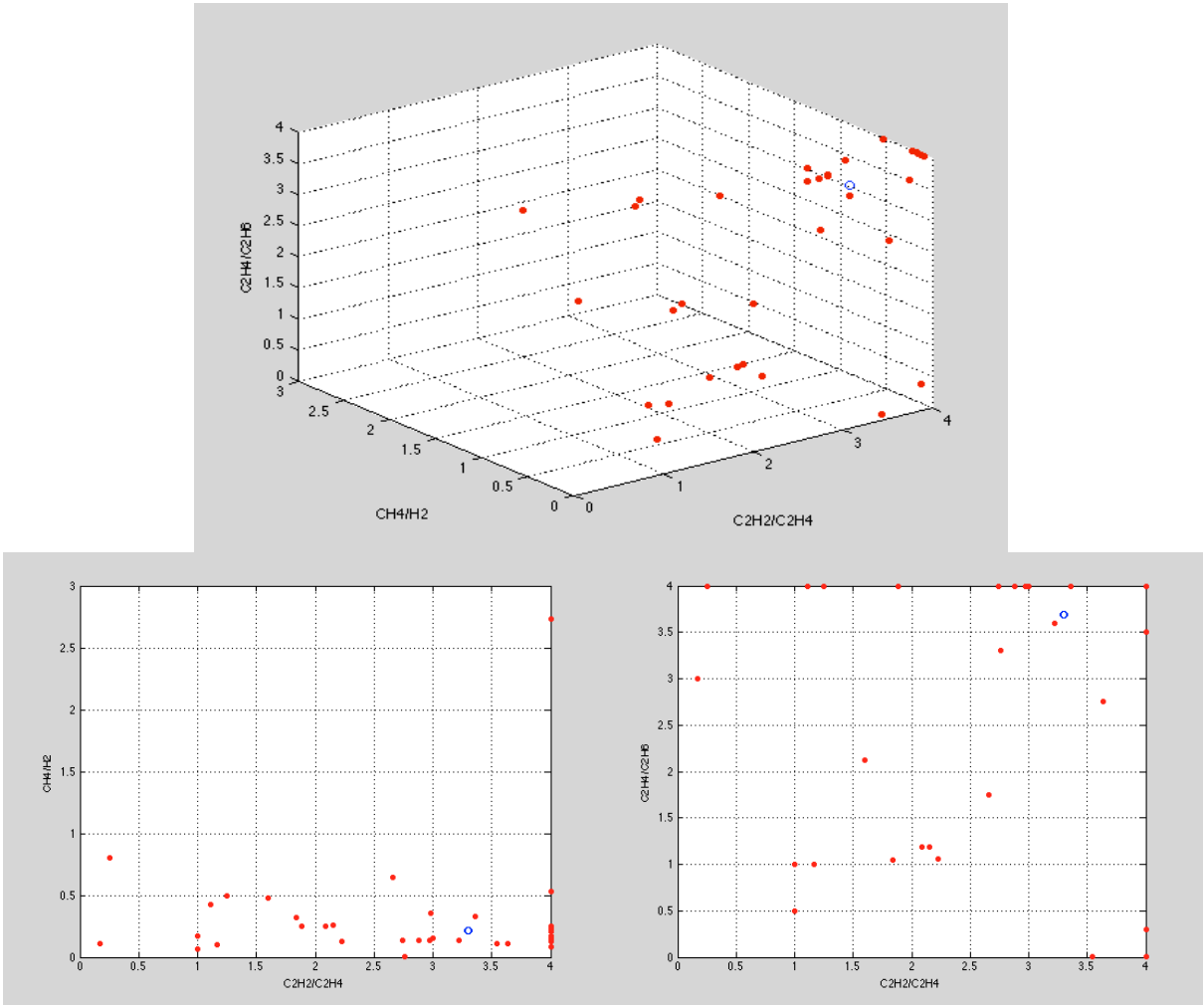


Figure 3.1 - Low-energy cluster fault ITLMS mode seeking output $\sigma = mean\ (std)$

As can be seen from the figure, the algorithm converges for a single point and this point can be found in the area with higher density.

To understand the influence of the σ value in the algorithm behaviour, if this parameter has its value reduced to half, the width of the window where the points influence each other is smaller; therefore, the number of points returned is expected to be higher. This consequence can be seen in the figure 3.2, below:

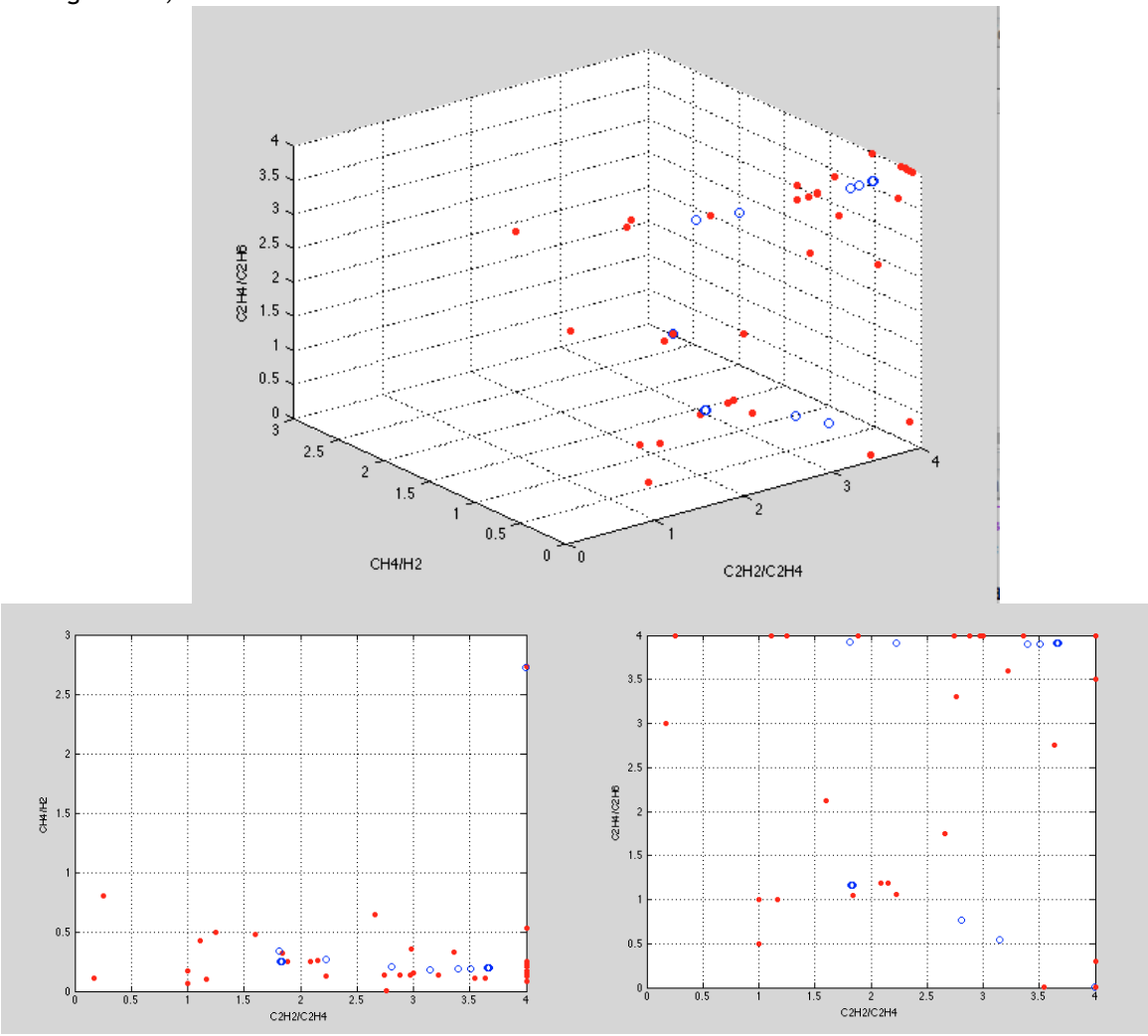


Figure 3.2 - Low-energy fault cluster ITLMS mode seeking output $\sigma = 0.5 * mean\ (std)$

In this image, more blue circles can be observed wich means the algorithm returns more local modes. One attentive analysis can conclude that one of the points, in this image, is almost the same as the one in the image before. This happens because of the higher number of points in that space zone.

As it was already mentioned, this procedure was done to all the seven clusters (partial discharge, low-energy discharge, high-energy discharge, thermal fault T<700ºC, thermal fault T>700ºC and healthy states with and without on load tap changer) in order to test the mean shift algorithm and to build a diagnosis method as will be seen after.

It was also tried to find the modes of all the fault clusters at the same time. The best result of this algorithm was obtained using $\sigma = 0{,}35 *$ standard deviation, as can be seen in the image 3.3:

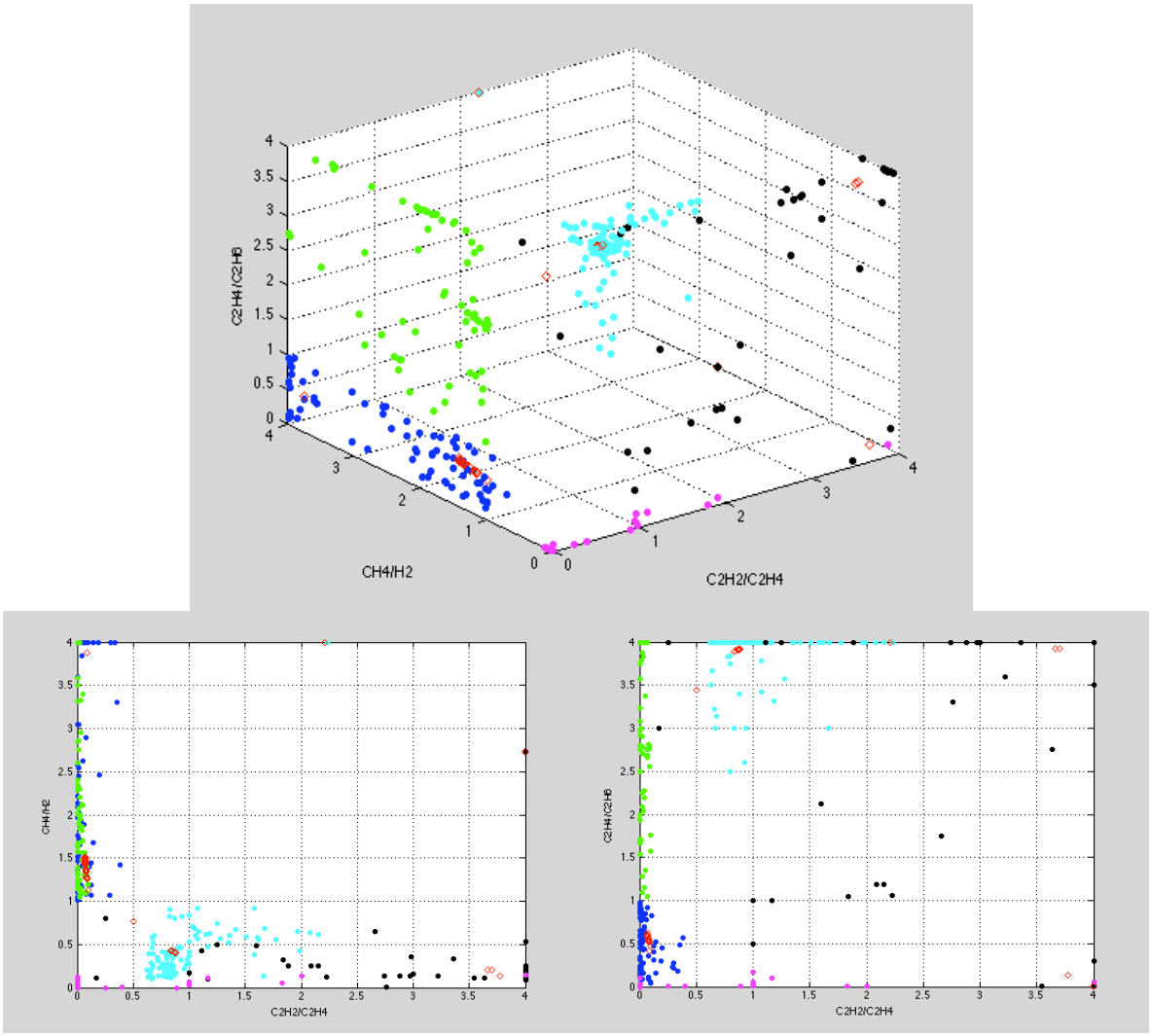


Figure 3.3 - Database mode seeking output using ITLMS ($\sigma = 0,35 * mean(std)$)

In the figure 3.3, in red are the outputs, in cyan are the high-energy discharge points, in blue the thermal fault with T<700ºC ones, the thermal fault with T>700ºC in green, in black the partial discharge representatives and in pink the low-energy discharge ones.

It can be seen that the algorithm couldn't find a mode to the thermal fault with T>700ºC cluster and several ones are found to the thermal fault with T<700ºC, other problem happened in the low energy discharge cluster, where the mode found is highly influenced by the partial discharge points, being very inaccurate.

The different density of each cluster and the proximity between points of different clusters can justify these inaccuracies. Therefore, this result wasn't used in any of the diagnosis methods.

3.2.2. Densification trick using ITLMS

As mentioned before, one of the tasks done using the ITL Mean Shift algorithm was to create virtual data points to train neural networks, keeping the real data points to validate the results given by those networks. To do this, the same algorithm as the one used to find mode can be used, however with a slight difference: the results of all the iterations are kept. This means that the trajectory of any point until it reaches the mode, or any other cluster feature, is stored. This way, one point is used to create so many points as iterations. Due to the mean shift capacity of capturing the intrinsic properties of a data cluster, the virtual data points will be consistent with the real ones.

To create virtual data, a combination of the mean shift algorithm and the steepest descent one was used. While the mean shift algorithm was used to detect the path to the modes, the steepest descent algorithm was used in order to do the first iteration of the global algorithm. In this first iteration, the direction of the movement was the inverse direction of the attraction. This way, it is guaranteed that the real data does not form the frontier of the cluster. The iteration step that was used in the steepest descent algorithm was equal to -10. This was a value big enough to guarantee that the new point wasn't too close to the original one, and small enough in order to keep the new point not too far from the original one. Once again, to adjust this parameter several tries were done.

All of the faults and healthy states were submitted to this algorithm in order to train the neural networks.

The result of this algorithm applied to the low-energy discharge, with the same parameters as the single mode-seeking algorithm ($\lambda = 1$, $\sigma = mean\,(std)$), can be seen in the figure 3.4. Once again, the red dots are the original data and the blue circles the virtual data.

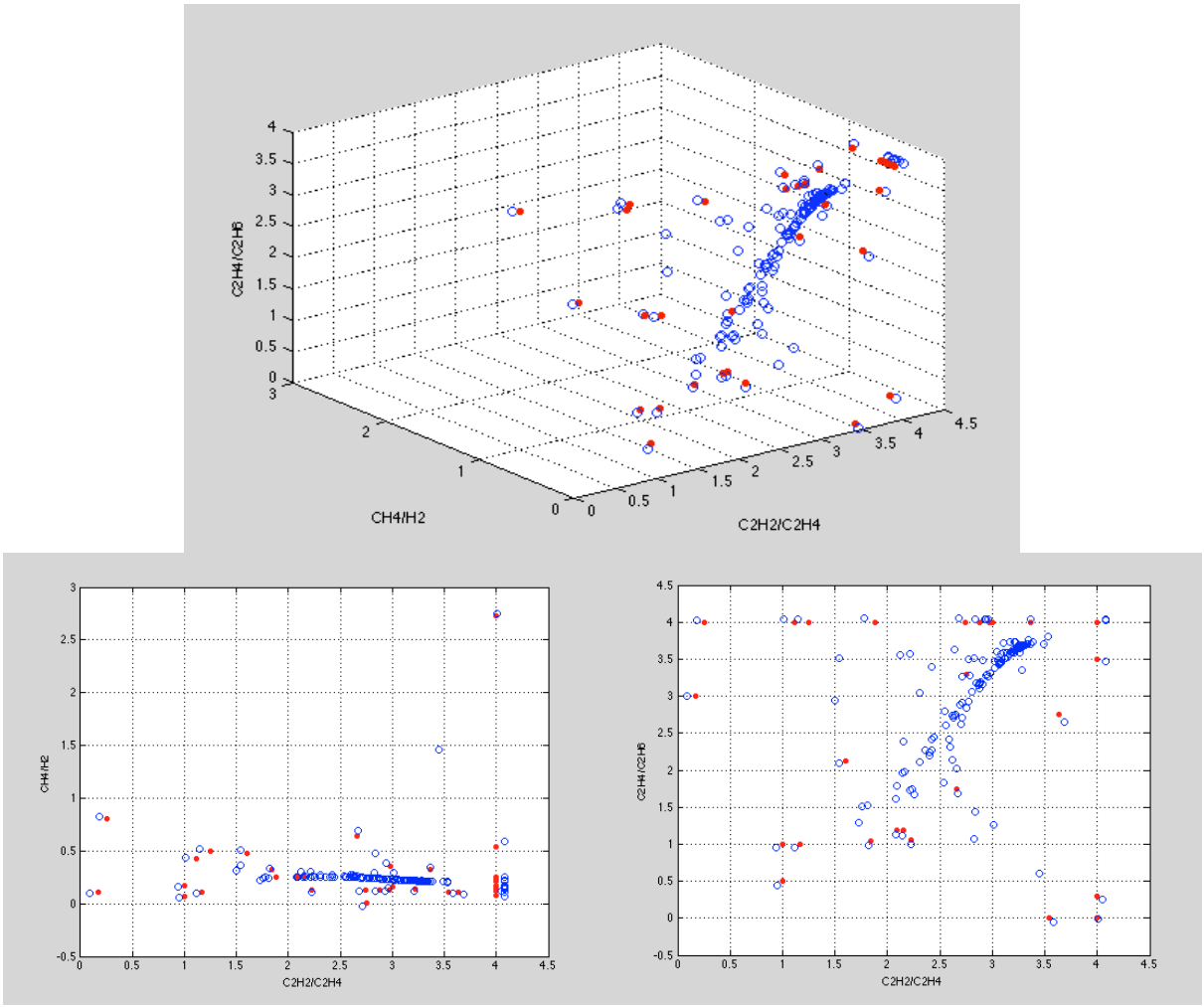


Figure 3.4- Low-energy discharge ITLMS consecutive application ($\lambda = 1$, $\sigma = mean\ (std)$)

Doing a simple analysis to the figure 3.4, the first iteration can be clearly seen. These points are the new frontier of the cluster and guarantee that the real points are always inside, and never on the cluster limits. The points' trajectory until the mode can be also observed. This new data set is the one used to train the networks.

Once again, this procedure was done to all the seven of transformers' health states.

In order to compare the neural networks behaviour when trained with $\lambda = 2$ it was also done the densification trick with this value. When one uses $\lambda = 2$ the algorithm converges to the principal curve of the pdf. Principal curves give information about the cluster formed by the data and can be defined as non-parametric, smooth and one-dimensional curves that pass through the *middle* of a d-dimensional probability distribution or data set. [42, 54].

Since the algorithm is searching for different properties of the cluster, it returns different points, which can result in different neural network connection weights after training.

In figure 3.5, the output of the algorithm applied to the low energy discharge with $\lambda = 2$, $\sigma = mean\ (std)$ and the iteration step equal to -10, is shown.

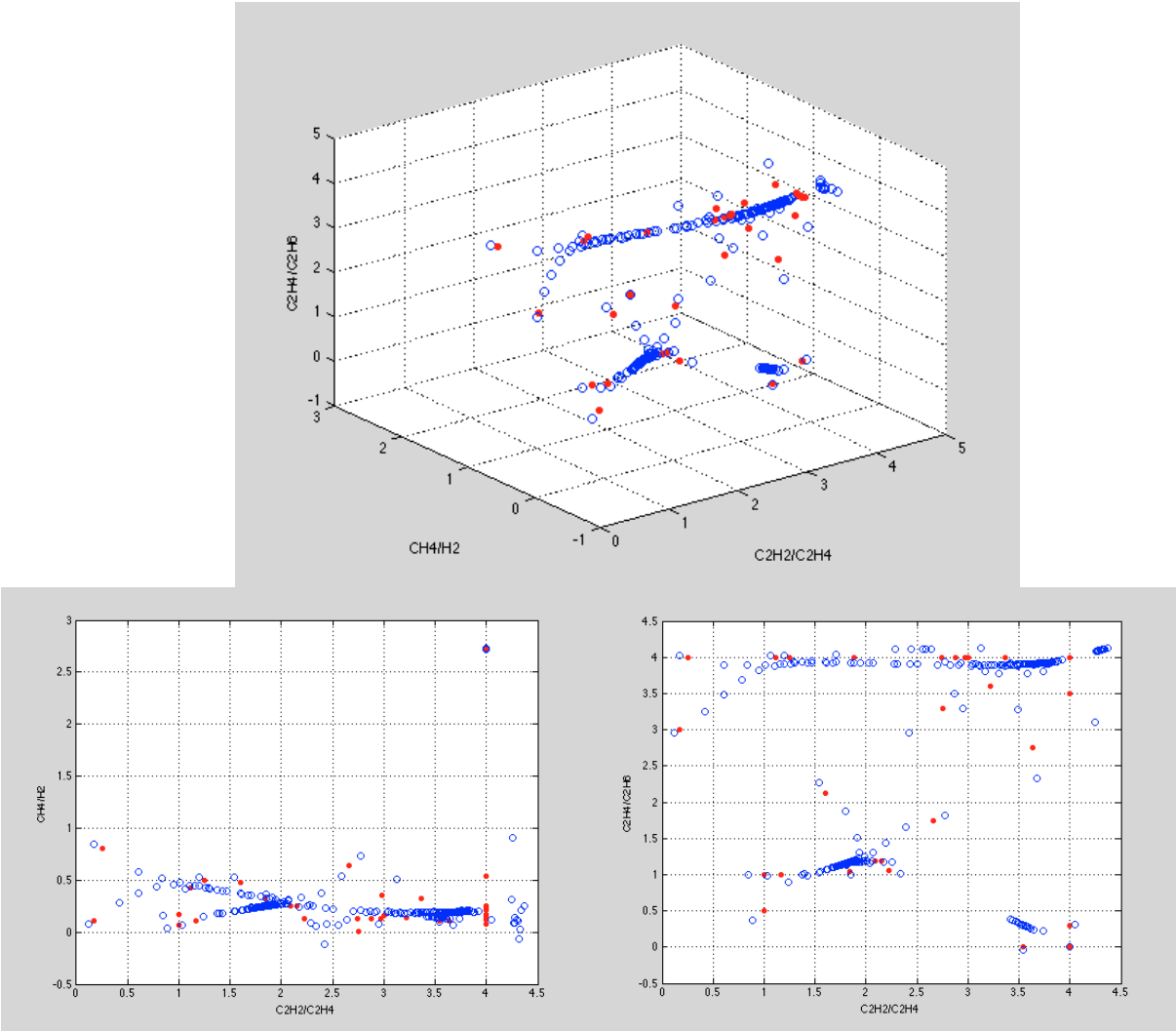


Figure 3.5 - Low-energy discharge fault ITMLS consecutive application ($\lambda = 2$, $\sigma = mean\ (std)$)

If one compares this figure with the one with $\lambda = 1$, it is easy to notice that both figures are very different. This is the result of seeking different properties of the pdf.

The output of this algorithm was used to train neural networks and the results are analysed in the proper chapter. Therefore, this procedure was done to all the faults/healthy states.

The use of virtual data points in the neural network training allows the use of all the real data in the validation stage. Thus, the results given by the validation process are more accurate, allowing one to have a better idea how the algorithm will behave when a new, completely unrelated data point is diagnosed.

In table 3.2 is the full description of the database obtained using $\lambda = 1$ and $\sigma = mean(std)$. The different values of virtual data can be justified with the number of the original real data and the number of iterations needed by the mean shift algorithm to converge to a single mode. Besides, for different ITLMS values the database can differ because of the number of iterations needed to the algorithm convergence.

Table 3.2 - Database complete description (using $\lambda = 1$ and $\sigma = mean(std)$)

Case	Fault/State	Real Data	Virtual Data
PD	Partial Discharge	30	330
DH	High Energy Discharge	103	618
DL	Low energy Discharge	37	444
T1	Thermal Fault (T<700°C)	77	1078
T2	Thermal Fault (T>700°C)	71	639
OK	Healthy State without OLTC	20	140
OK with OLTC	Healthy State with OLTC	10	90
	Total:	348	3339

3.2.1. Other ITLMS applications

As mentioned before, when one uses $\lambda > 1$ in the ITLMS algorithm, it seeks the cluster finer structures and these structures are more complex and have more information as λ increases. This way, and in order to study how the ITLMS and the diagnosis methods behave with this kind of data, several λ values were tested and it was concluded that the value where the best result were obtained was $\lambda = 7$. When the ITLMS algorithm is applied to the real data using $\lambda = 7$, the repulsion forced between the points is big enough to allowing the seeking of more structures intrinsic to the clusters, however this value isn't too big to return only the real points (local modes) of the cluster.

The application of ITLMS with $\lambda = 7$ was done using several σ values and it was chosen the value which produced the most accurate results, being this value different between clusters. However, all the σ values chosen were between $0{,}5 * mean(std)$ and $0{,}75 * mean(std)$.

The results of the application of the ITLMS to the thermal fault (T>700°C) cluster can be seen below, in image 3.6:

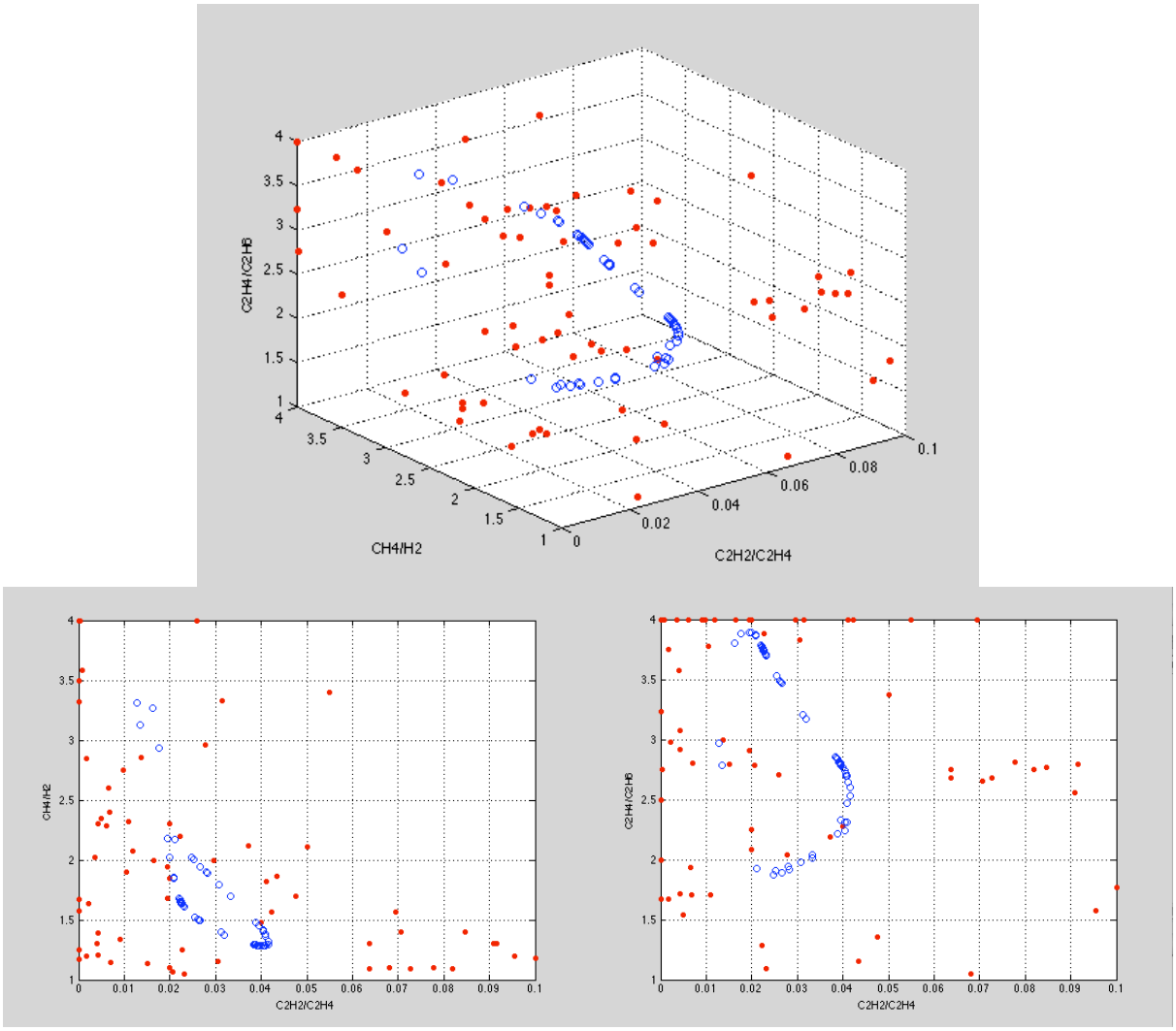


Figure 3.6- Thermal fault (T>700°C) ITMLS finer structures seeking ($\lambda = 7$, $\sigma = 0{,}75 * mean\ (std)$)

Observing the figure it is possible to see some cluster structures which was the objective of this test.

With the same objective as before, seeking data that could improve the results given by the diagnosis methods, the ITLMS algorithm was applied to the real data with $\lambda = 1$ and $\lambda = 2$ using several σ values. The variation of σ values to smaller values allows the algorithm to seek for local modes, when $\lambda = 1$, or the principal curve in smaller spaces when $\lambda = 2$.

The output of the application of the ITLMS to the low-energy discharge data can be seen in the figure 3.7:

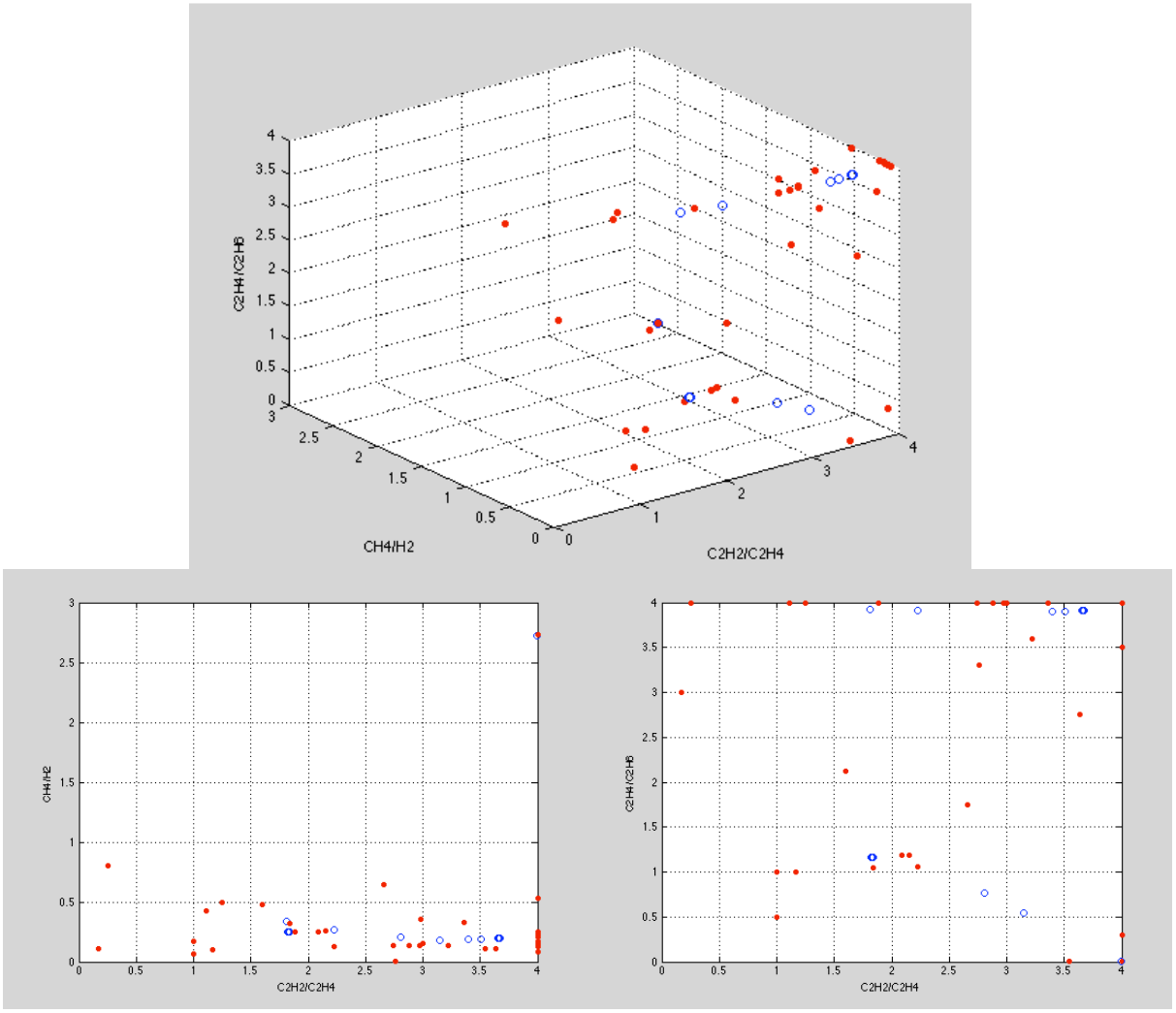


Figure 3.7 - Low-energy discharge fault ITLMS local modes seeking output ($\lambda = 1$, $\sigma = 0,5 * mean\ (std)$)

In figure 3.7 it is possible to see that most of the outputs (blue circles) of the algorithm are near the denser areas, as expected.

Chapter 4.

Incipient fault diagnosis systems

4.1. A diagnosis system using autoencoders

One of the first methods to diagnose power transformers that was studied as preparation for this thesis work was the one in [1].

In this method three dissolved gas concentration ratios are used as input, the same ones used in the IEC 60599 standard:

$$\frac{C_2H_2}{C_2H_4} \qquad \frac{CH_4}{H_2} \qquad \frac{C_2H_4}{C_2H_6}$$

In spite of the fact that there are only six different states to the transformer condition, this method uses seven autoencoders, because it distinguishes between healthy transformers with and without on load tap changers (OLTC). Each of the networks is trained to recognize one fault and when a new input vector is shown to the system, all of the networks work in parallel, trying to replicate the input vector. However, just one of the autoencoders will produce an output very similar to the input one, i.e., it will resonate: it will show a very small error between the input and output, while in all the other autoencoders this error will be very high [52]. Thus, the method assumes that the unclassified data point belongs to the fault that this networks represents.

This method assumes that a data cluster with unique properties represents each one of the faults and that's why there is one network per fault.

To build this diagnosing system, autoencoders with 3-15-3 architecture and neurons with sigmoid activation function were used. It has been shown that neurons with sigmoid activation functions induce better properties in feature reduction with autoencoders than the ones with linear activation functions [55]. When the middle layer of an autoencoder has neurons with linear activation functions it is equivalent to a Principal Component Analysis [56].

It must be said that these aren't classical autoencoders because the neural networks' hidden layers aren't smaller than the input and output ones, so no data compressing is achieved. This isn't one of the objectives, though

With systems architectures like this one, all networks are competing against the others when a new data point is presented. This is a great advantage because there are no unclassified samples after the diagnosis; however there can be wrong classifications. In the classic approach to this problem, where one single neural network is trained to recognize all the transformers' health states, there can be misclassifications and unclassified data after diagnosing.

Other advantage of competitive architectures is that there is no need to set arbitrary thresholds. These parameters are used in order to decide if a point belongs or not to a certain faulty/healthy state. Having less parameters set in an arbitrary way means that there is no need to adjust those by trial and error and the methods adjustment to the data reality is easier and more accurate.

To train the neural networks, the neural network toolbox from MATALAB was used. The training algorithm used was the Levenberg-Marquardt one [57, 58] with the minimization of the mean square error as cost function. This cost function minimizes the variance in the probability density function of the error distribution, giving optimal results when the error distribution is Gaussian.

In order to have a better view of this system, a representative block diagram (figure 4.1) can be seen below:

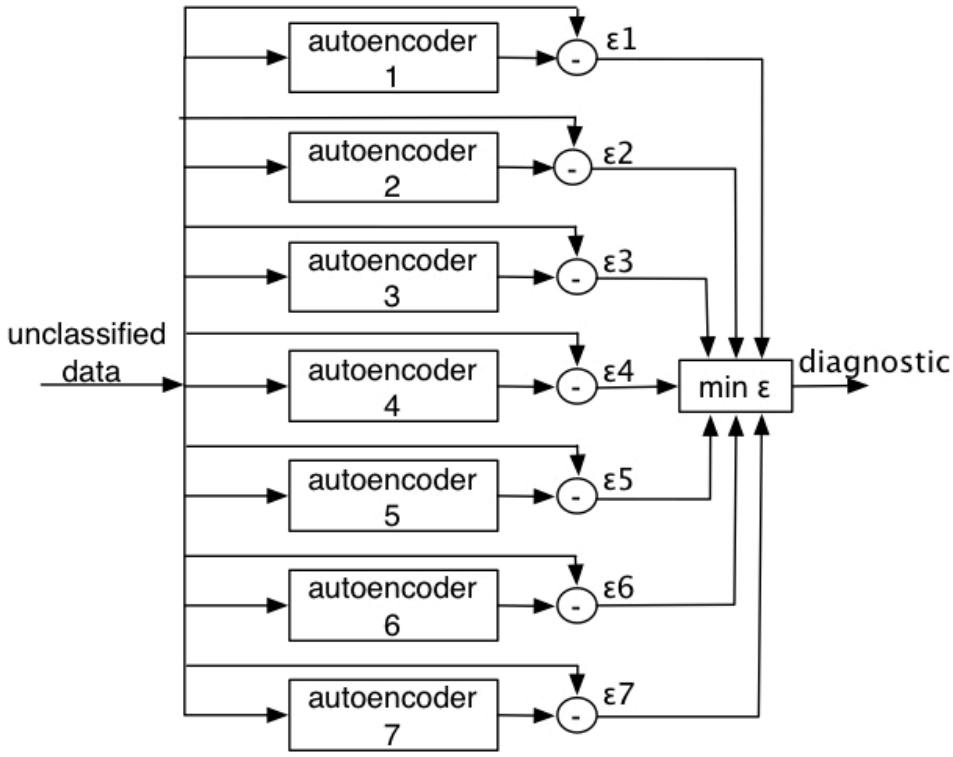


Figure 4.4.1 - Autoencoders diagnosis system architecture

The results obtained with this system were satisfactory since 95,69% of correct diagnosis was achieved when the neural networks training was done using the ILTMS virtual data created using $\lambda = 1$ and σ adapted to retrieve a single mode. This means that 333 of 348 diagnoses were correct.

Also, using $\lambda = 2$ and $\sigma = mean(std)$ in the mean shift algorithm in order to create virtual data and training autoencoders with these points retrieves 93,39% of correct diagnosis.

It must be said that training neural networks and especially autoassociative networks can be tricky because of the random initial weights MATLAB assigns. It took a long time to achieve this result and there is no guarantee that a better one can't be accomplished; however a lot of tries were made.

Training neural networks with data created with different λ values was done in order to compare which neural networks were more adapted to the data, i.e, which ones had a smaller input-output error. To do this study, the neural network that is supposed to recognize a given fault did the diagnosis for the corresponding fault data set (training and validation data) and the diagnosis error was stored. This diagnosis error was defined in this work as the mean absolute error between the input and output vectors. The interval between the maximum and the minimum error of each case was divided in 20 smaller intervals, allowing the building of the histograms of the figure 4.2:

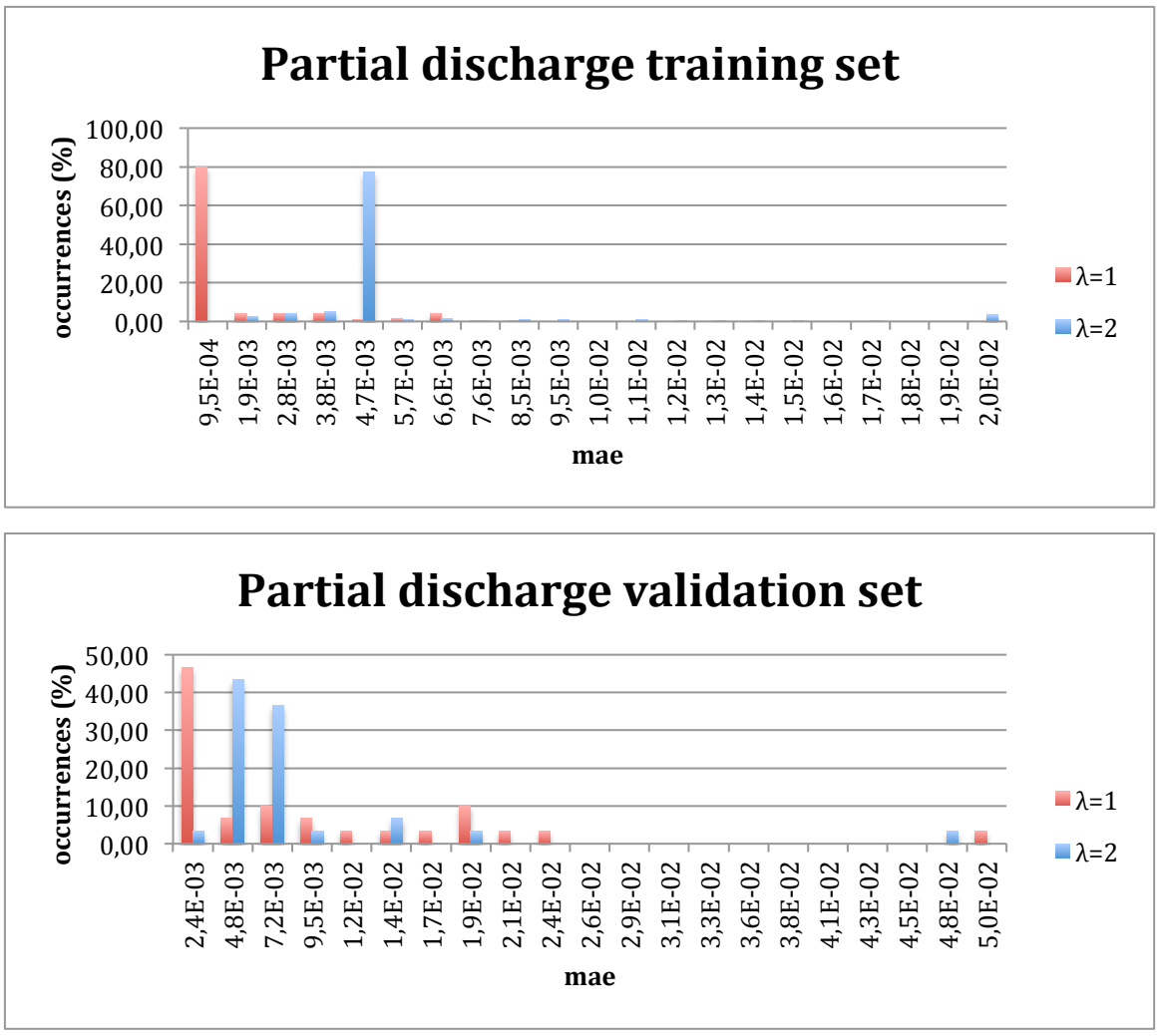


Figure 4.2 - Partial discharge autoencoders diagnosys error comparison

In the partial discharge case, the results in the training set are fairly better when $\lambda = 1$, however in the validation set, the results obtained by using the data with $\lambda = 2$ slightly are better, with the first 3 columns summing 83,33% against 63,33% obtained by the neural networks trained with $\lambda = 1$ data.

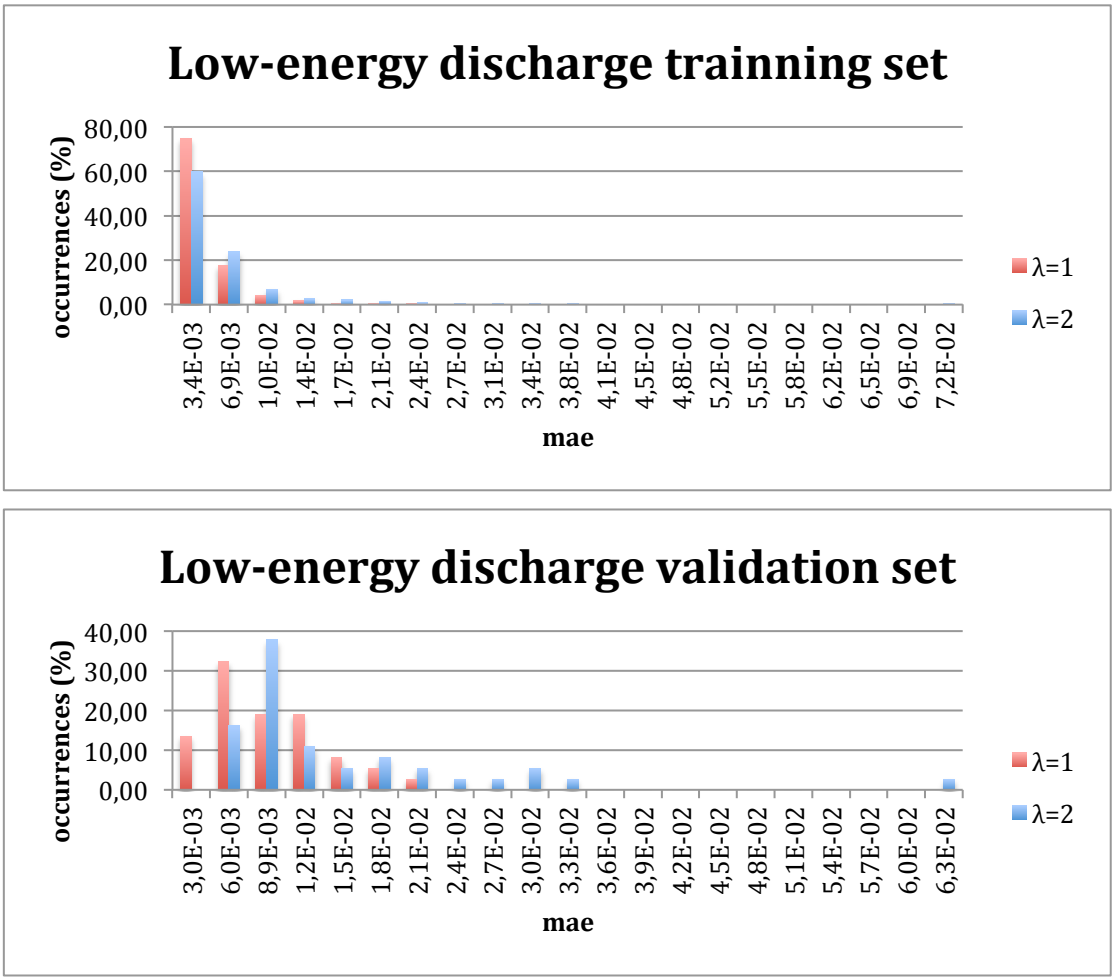


Figure 4.3 - Low-energy discharge autoencoders diagnosys error comparison

In what concerns the low-energy discharge data, presented in figure 4.3, the diagnosis of both sets are better when the neural networks are trained with the λ equal to one. This result is similar in the remaining faults/healthy states. The networks give better results when trained with virtual data created using $\lambda = 1$.

The results show that neural networks can have better results when trained with data created using the ITLMS with $\lambda = 1$ or $\lambda = 2$. This is related with the cluster pdf shape, i.e. a cluster with a pdf with a paraboloid shape doesn't have a principal curve and will have better results with $\lambda = 1$. When a neural network is trained with virtual data created using different λ values, one is adapting the method to the shape and intrinsic properties of the cluster; therefore, the results obtained should be better. When the pdf of a given cluster doesn't have a principal curve and $\lambda = 2$ is used in the ITLMS, the points will spread in the surroundings of the mode, however, because repulsion between the points is bigger, they do not converge to a single point/mode.

Due to this study results, an attempt to create a diagnosis method using neural networks trained with different λ values was made. The partial discharge autoencoder was trained with $\lambda = 2$ data and the other six faults/healthy states were trained with data created by the ITLMS algorithm using $\lambda = 1$. This way, 95,98% of correct diagnosis was achieved. This is a slight improvement of 0,3% in comparison with the method with all the neural networks trained using $\lambda = 1$ data.

The neural networks constituents of this diagnosis method were also trained with the ITLMS virtual data created using $\lambda = 1$, $\lambda = 2$ and $\lambda = 7$ and σ and various σ values for the different clusters. The results for each one were 94,54%, 94,83% and 79,6%. The diagnosis result obtained with $\lambda = 2$ and various σ values, where the ITLMS returns the principal curve of smaller zones in the space, is better than the one obtained using $\lambda = 2$ and $\sigma = mean(std)$.

The results obtained with this method are summarized in the following table:

Table 4.1 – Autoencoder method results summary

$\boldsymbol{\lambda}$	$\boldsymbol{\sigma}$	**Results (%)**
1	mean(std)	95,69
2	mean(std)	93,39
1 and 2	mean(std)	95,98
1	various	94,54
2	various	94,83
7	various	79,6

In table 4.1, where σ appears as various is because the value was adapted to each cluster in order to the ITLMS return the characteristic sought.

4.2. Diagnosis using neural networks with binary outputs

Another tested method to diagnose gas ratio samples uses a set of non-associative neural networks with 3 layers, where the middle layer contains 15 neurons and the output layer is one single neuron trained to retrieve '1' or '0', indicating the recognition (or not) of a given type of failure.

The inputs of this system are exactly the same as in the autoencoders one.

This method is also a competitive one. Therefore, seven neural networks, each trained to recognize one of the possible healthy/faulty states, work in parallel diagnosing the same input. Thus, as in the autoencoder method, only one neural network will resonate when an unclassified data point is shown. This way, when a new unclassified data point is shown to the diagnosis system, all neural networks will diagnose it, however, only one of the neural networks will retrieve a number very similar to the value one, while the others will have an output very similar to zero. Therefore, the diagnosis of the new point will be the fault that the neural network with the output similar to one represents. Again, arbitrary thresholds are avoided due to the system competitive architecture.

The advantage of this system, compared to the autoencoders one, is the smaller number of connections between neurons. While in the autoencoder system there are 3*15+15*3=90 connections, in this one there are only 3*15+10*1=55. This way, the diagnosis routine is much faster, but the training of these neural networks proved to be very slow, because all database needs to be used in the training of each one of the neural networks, i.e. the neural networks need to be trained to recognize a fault and return the value one and to return the value 0 to all the other faulty/healthy states.

In this method the neural networks were trained in the same way as the autoencoders in the former method, with Levenberg-Marquardt algorithm and the minimization of the mean square error as cost function. Regarding the data, the same approach used in the autoencoders system was followed: Mean Shift data points were used in the ANN training and the real data in their validation. The neurons activation functions were also sigmoidal ones.

While, theoretically, the output of the neural networks should be always 1 or 0, this doesn't happen all the time, except for the training set. For that reason, the diagnosis of an unclassified data point will be the one that the neural network's output is more similar to unitary one. As in the autoencoders method, the error between the output and the expected value (one) is known using the mean absolute error.

In figure 4.4, is a block chart where this algorithm is shown.

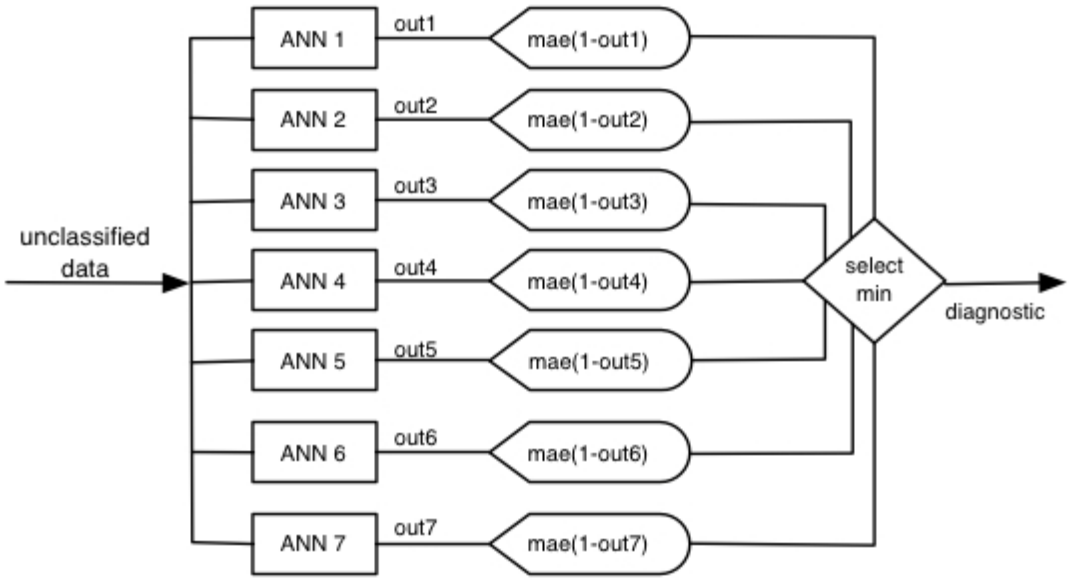


Figure 4.4 - Binary neural networks diagnosys system architecture

With this algorithm 97,70% of correct diagnoses were achieved when the neural networks are trained with virtual data created using the ITLMS with $\lambda = 1$ and $\sigma = mean\ (std)$. This means that 340 of 348 cases where diagnosed correctly, 7 more than in the system with autoencoders.

Like it was done in the autoencoders diagnosis system, in order to study the behaviour of the neural networks, they were also trained using virtual data created by the ITLMS using $\lambda = 2$ and $\sigma = mean\ (std)$. The output error of all the neural networks was then studied with the intention of discover which ones where best adapted to the database. In this system, the output error isn't the MAE between the output and the input, like in the autoencoders system, but the MAE between the output and the unitary value. In the next page, histograms resulting from this study can be seen:

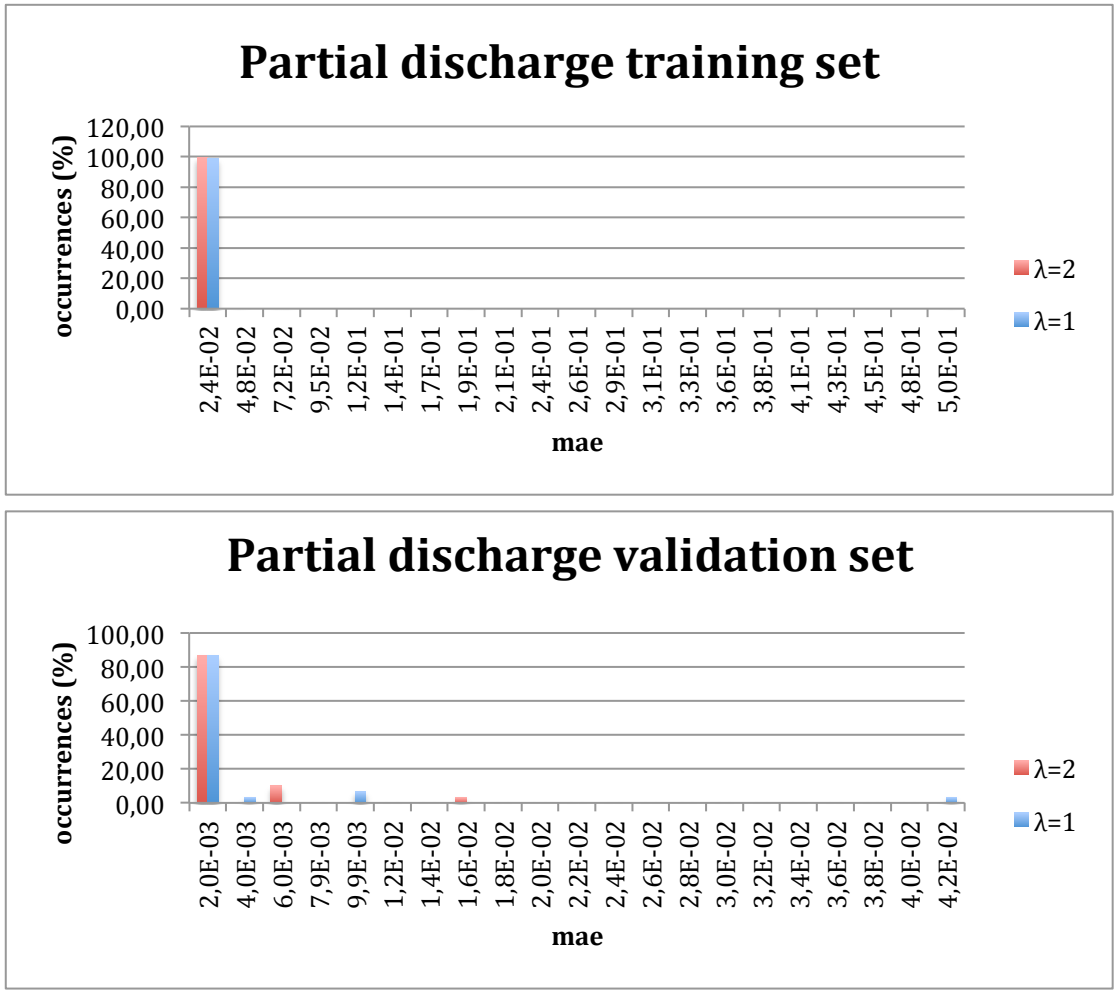


Figure 4.5 - Partial discharge binary neural networks error comparison

Regarding the partial discharge fault, both neural networks give very similar results. Therefore, the system was tested with both of them and the best results were achieved with the neural network trained with $\lambda = 1$.

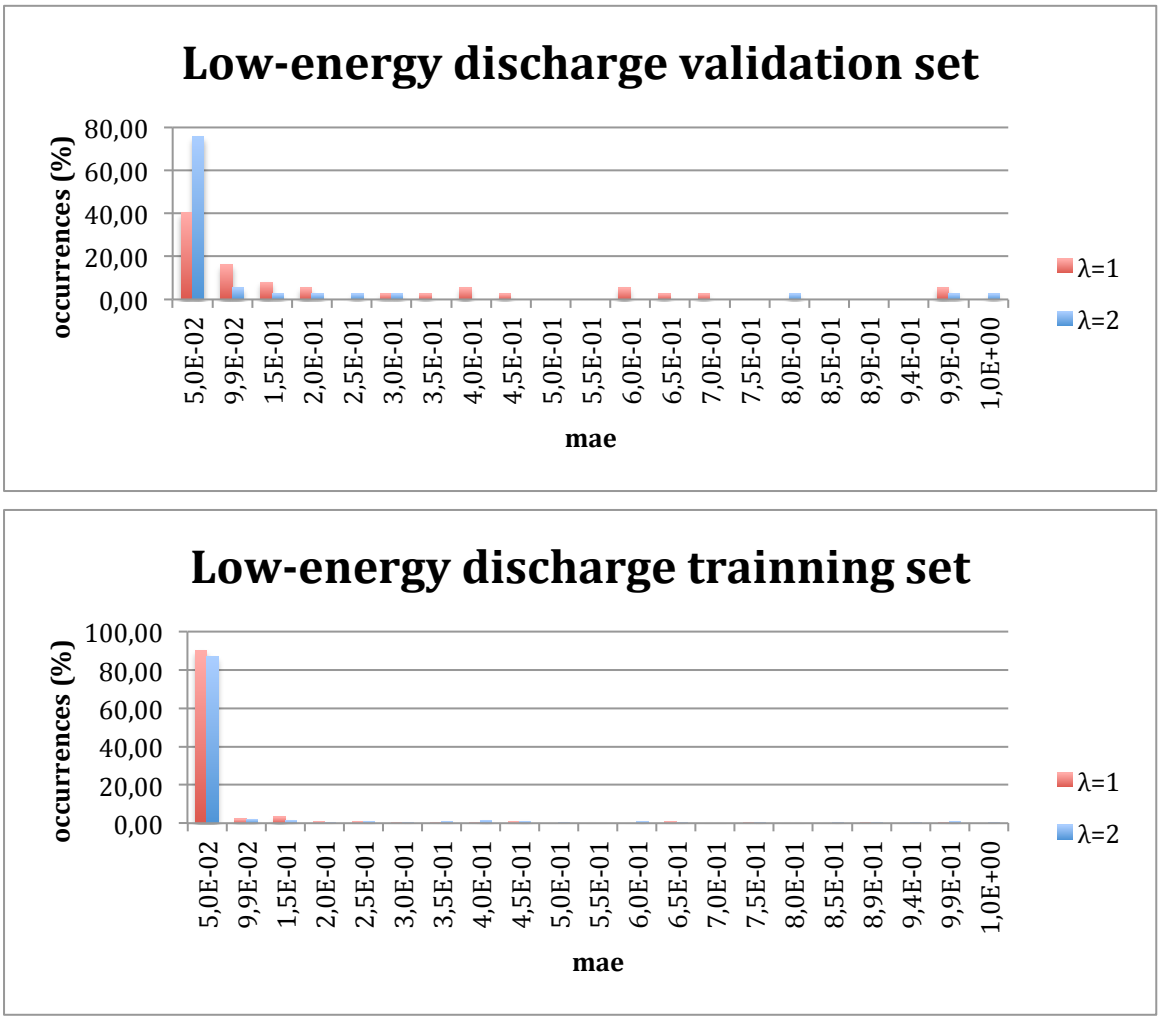


Figure 4.6 - Low-energy discharge binary neural networks error comparison

In the case of the low-energy discharge fault, data created with $\lambda = 1$ has a slightly better result in the training set, however, in the validation set, the neural network trained with $\lambda = 2$ has more data with a smaller error but it has also more data with a bigger error. Once again, the system was tested with both neural networks and the best results were obtained with the neural network trained with $\lambda = 1$ data. This happens because this is a competitive method and the diagnosis is selected from the neural network with the smaller error, therefore, if a neural network has more data with bigger error, the probability of a wrong diagnosis appearing is also bigger. A similar situation happens in the thermal fault with T<700ºC and in the healthy state in transformers without OLTC cases.

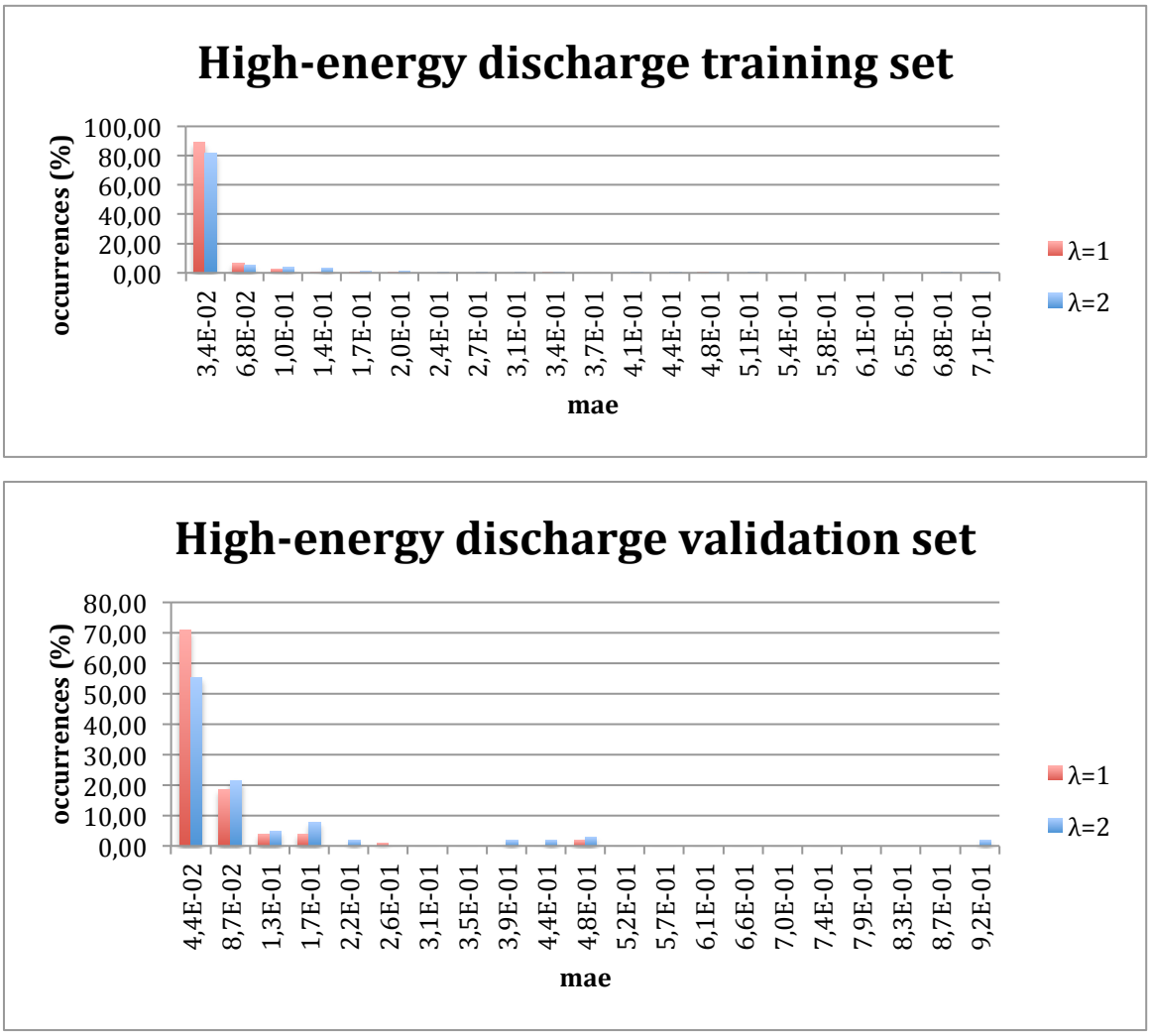


Figure 4.7 - High-energy discharge binary neural networks error comparison

Regarding the high-energy discharge fault, and the thermal fault with T>700ºC is very similar, both data sets obtain better results when the neural networks are trained with virtual data obtained using $\lambda = 1$ in the mean shift algorithm.

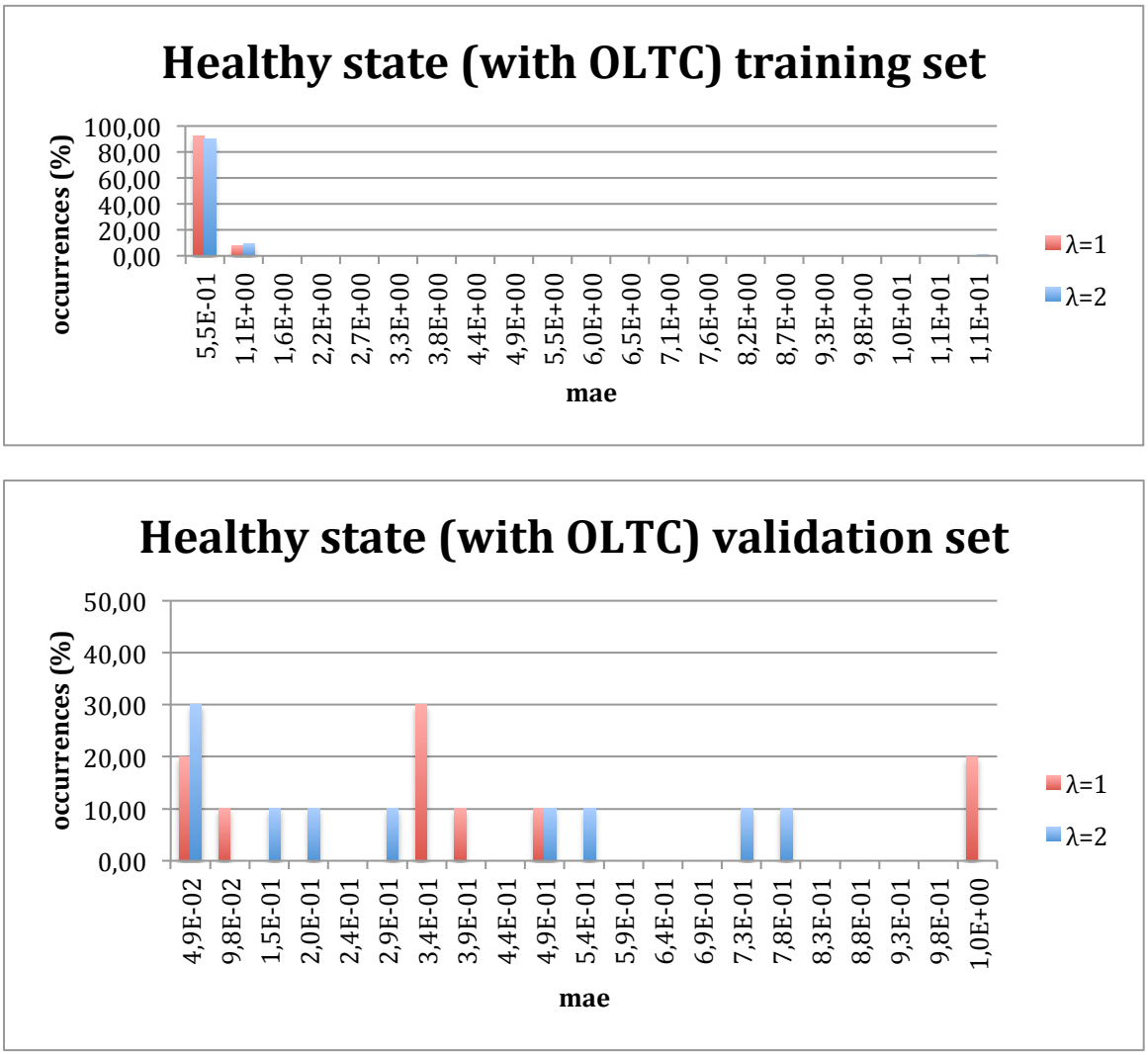


Figure 4.8 - Healthy state of transformers with OLTC binary neural networks error comparison

In the matter of healthy transformers with OLTC, in the training set, neural networks trained with $\lambda = 1$ Mean Shift data have better results, however, in the validation data set diagnosis the best results are obtained when using neural networks trained with virtual data obtained with $\lambda = 2$.

Therefore, the final diagnosis method was built using all the neural networks trained with $\lambda = 1$ data, except the neural network supposed to recognize healthy transformers with on load tap changers. This way, 97,99% of correct diagnosis were achieved. This is an improvement of 0,29% against the 97,70% obtained by the method where all the neural networks used $\lambda = 1$ data, and an improvement of 2,01% in comparison with the autoencoders method.

4.3. Mean absolute error and modes method

This is the simplest diagnosis method used in this thesis. It uses the modes, or other representatives, of each fault cluster retrieved by the ITLMS algorithm and, when an unclassified sample is presented, the mean absolute error (MAE) between this point and all the modes is determined. The diagnose is the fault which MAE is the smaller. Below, figure 4.9 presents a block scheme of this method.

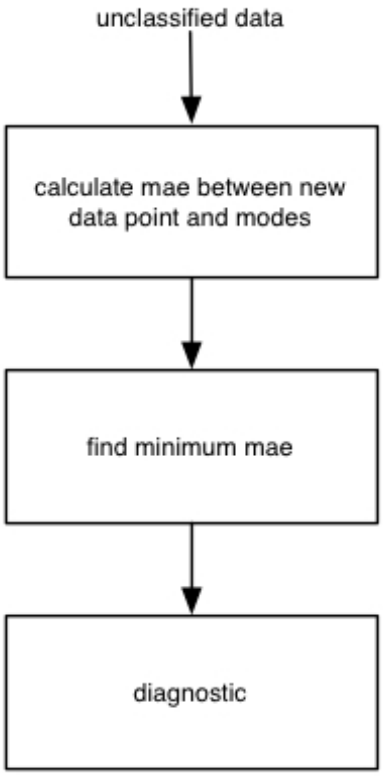


Figure 4.9 - mean absolute error and modes method architecture

Because of being so simple, the expectations on this algorithm weren't very high. However, 81,61% of all database was diagnosed correctly when a single mode obtained with the ITLMS algorithm is used to represent each one of the clusters. Several other cluster representations were tried, like a set of local modes to represent each cluster, obtaining 77,30% of correct diagnoses and ITLMS data created using $\lambda = 2$ and $\lambda = 7$ with the σ value tuned to retrieve the characteristics sought. In each case the percentage of correct diagnoses was 80,74% and 84,77%.

In order to test other methods to compare the new undiagnosed data point and the cluster representatives, the Euclidean distance was used. The diagnoses produced using this similarity measure were worse when the representatives of the clusters was $\lambda = 1$ or $\lambda = 2$ data. However, when $\lambda = 7$ data was used, 85,05% of correct diagnoses were achieved, being the best result of this method. These results don't allow taking any conclusion of which one of the similarity measures is better because in some cases the Euclidean distance is worse and in others it is better.

The table below summarizes the results obtained:

Table 4.2 – results obtained with MAE and modes method summary

λ	σ	Similarity measure	Results (%)
1	mean(std)	MAE	81,61
1	various	MAE	77,30
2	various	MAE	80,74
7	various	MAE	84,77
1	mean(std)	Euclidean dist.	81,32
1	various	Euclidean dist.	75,86
2	various	Euclidean dist.	80,46
7	various	Euclidean dist.	85,05

4.4. Steepest Descent and mean absolute error method

One of the diagnosis methods makes use of the steepest descent algorithm, presented before, and the mean absolute error (MAE) between the output and the points used to represent each fault.

One more time, the inputs for this algorithm are the three ratios used in the IEC 60599 standard.

In this method, the output points of the clustering Mean Shift algorithm were used as representatives of each fault. These points were obtained using several λ values and the algorithm used had a step in the inverse direction of the mode. This means that one isn't using the real data points as a cluster's frontier. When a new unclassified data point is presented, it is subject to the steepest descent algorithm. Therefore, the data point moves towards the properties sought of the data. After, the mean absolute error between this new point and each of mean shift points is calculated. The minimum is found and the unclassified point is diagnosed as the fault which has the minimum MAE.

This method seeks to find attraction bays intrinsic to the data, i.e. if $\lambda = 1$ is being used, the point should move to zones where the probability density function is higher, to one of the modes of the function, because of the information force applied to it.

Below is a block chart that represents this algorithm (figure 4.10).

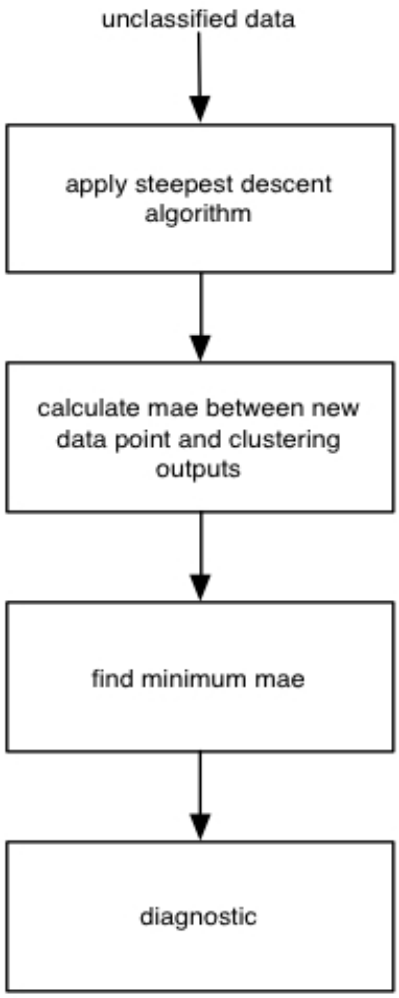


Figure 4.10 - Steepest descent and mean absolute error method architecture

As the mean shift algorithm, in this procedure adjusting the σ was needed. Once again this was done by trial and error with different values of σ. In the case of the algorithm recognizing only the five faulty states, the best results were obtained with $\sigma = 50 * mean(std)$. Adjusting the iteration step was also necessary, with the best outcome appearing with $\rho = 2$. In the table 4.3, several tries are shown:

Table 4.3 - Steepest descent and MAE method results (when recognising only faulty states) (ITLMS data created with $\lambda = 1$ and $\sigma = mean(std)$)

λ	ρ	σ	Results (%)
1	1	std	24
1	2	3*std	82,4
1	2	3,5*std	85,2
1	2	4*std	86,16
1	2	4,5*std	88,05
1	2	7*std	88,36
1	2	10*std	89,31
1	2	20*std	90,88
1	2	25*std	91,2
1	2	50*std	91,51
1	5	50*std	91,51
1	2	70*std	91,51

In spite of being a very simple algorithm to understand, it involves a lot of calculus, and, therefore, it takes a lot of time to diagnose the entire database even when only the faulty states are recognized (around 45 minutes), being slightly slower when all the seven healthy/faulty states are diagnosed. This is a serious handicap of this method.

When all the seven healthy/faulty sates are recognized by the diagnosis method, the results aren't so good, in spite of the fact that several tries with different parameters in the algorithm itself or in the virtual data creation using ITLMS were tried. A summary of the results obtained can be seen in the table 4.4. This table presents the parameters used in the ITLMS algorithm and in the diagnosis method.

Table 4.4 - Steepest descent and MAE method results (when recognising seven healthy/faulty states)

λ (ITLMS)	σ (ITLMS	λ	ρ	σ	Results (%)
1	mean(std)	1	35	25*mean(std))	71,84
1	mean(std)	1	35	40*mean(std)	71,26
1	mean(std)	1	35	15*mean(std)	63,79
1	mean(std)	1	15	25*mean(std)	71,26
1	mean(std)	1	30	20*mean(std)	72,13
1	mean(std)	1	25	20*mean(std)	71,84
1	mean(std)	1	30	25*mean(std)	71,26
1	mean(std)	1	30	22*mean(std)	71,84
1	mean(std)	1	25	25*mean(std)	72,26
1	mean(std)	1	17	20*mean(std)	71,84
1	various	1	20	20*mean(std)	70,69
1	various	1	20	20*mean(std)	70,12
1	various	1	5	5*mean(std)	70,12
1	various	1	20	20*mean(std)	70,69
1	various	1	35	35*mean(std)	70,97
7	various	7	20	50*mean(std)	1,44
7	various	1	20	50*mean(std)	1,44
7	various	1	20	5*mean(std)	8,62
7	various	7	20	5*mean(std)	8,62
7	various	7	20	3*mean(std)	1,72

λ (ITLMS)	σ (ITLMS	λ	ρ	σ	Results (%)
7	various	1	5	3*mean(std)	8,62
7	various	7	5	3*mean(std)	8,62

Once again, and in order to study another similarity measure, it was tried to use the Euclidean distance. However, the results were worse than the ones obtained with the mean absolute error.

For an algorithm so simple, the results obtained are very good, with a maximum of 91,51%. However, this result its only valid when only the five faulty states are recognized by the method. When the faulty states are also included, the maximum percentage of correct diagnoses is 72,26%. This can be related with the number of points in the ratios space with the steepest descent method not converging to the correct points.

4.5. Method Comparison

According to what was said before, it is easy to conclude that the method with the best results is the one that uses neural networks with binary outputs. However, the neural network training of this method is slower than the autoencoders one. This happens because, in spite of having fewer connections between neurons, the data used in training is bigger than the one used in autoencoders, because the neural networks need to be trained to the entire database with an output vector that varies between 1 and 0 according to the fault that should be recognized by the neural network. In the autoencoder diagnosis system, a neural network is trained only with the data that represents the fault that neural network must recognize.

It must be said that the results achieved in this thesis are very good, with the binary output neural networks system diagnosing 97,99% of the database in a correct way, and the autoencoders system 95,98%. Remembering the table published in the DGA state of the art chapter, there are only a few methods with results better than 95% of correct diagnosis, with the IEC 60599 achieving 93,94%. Still, the 100% claimed in [1] could not be reached. It must also be underlined that the database used in this work is one of the biggest, with 438 cases.

The results achieved are very good also because the training procedure of the neural networks were done using only virtual data, obtained using the ITL Mean Shift algorithm. This proves this method is a valid one and can be used when the data is sparse and that was the case in this work.

Below, the table 4.5 contains several diagnoses using the IEC 60599 standard and the two best methods of this work can be seen. This table covers the most distinct cases in the database, and it isn't representative of the complete database.

Table 4.5 - Diagnosis methods comparison

$\frac{C2H2}{C2H4}$	$\frac{CH4}{H2}$	$\frac{C2H4}{C2H6}$	Fault	IEC	Autoencoder system	Binary neural network system
9,5	0,151515152	0,052631579	PD	T1	PD	PD
0	0,102272727	0,0001	PD	T1	PD	PD
0,0001	0,075	0	PD	OK	PD	PD
1	0,166666667	1	DL	OK	DL	DL
5,779230769	0,534978201	12,26415094	DL	SD	DL	DL
1,166666667	0,104761905	1	DL	DL	DL	DL
8,077324974	0,15	30,87096774	DL	DL	DL	DL
1,882352941	0,25	17	DL	DH	DH	DH
2,230769231	0.0001	6,842105263	DH	OK	DH	DH
0,012820513	2,545454545	0,906976744	T1	T1	T1	T1
0,007142857	1,147058824	2,8	T2	OK	T2	T2
0,068181818	1,1	1,047619048	T2	OK	T2	T2
0,1	1	0,2	OK	T1	OK	OK
0,5	0,5	1	OK OLTC	DL	OK OLTC	OK OLTC
DL	0,166666667	0,109589041	DH	OK	DL	OK

In table 4.5, only the ratios are published, however, where the IEC60599 standard diagnoses a case as being an 'ok' one, it happens because the concentrations of the dissolved gases aren't bigger than the limits published. Where ND is shown, the IEC60599 isn't able do diagnose because

the values of the gas ratios do not fit in any of the intervals defined by this standard. It is also shown a case where the diagnosis of the two methods developed is different, with the method with neural networks with binary output guessing the diagnosis right.

Regarding the method that applies the mean shift algorithm to do the diagnosis and the one that use the modes, they were an attempt to study the potential of the Mean Shift algorithm as a stand-alone diagnosis method. It must be said that the results were pretty good, with the steepest descent method diagnosing 91,51% of the faulty cases with a correct diagnosis and the distance one obtaining 84,77% of correct diagnoses. Still, the accuracy of these methods is much smaller than the accuracy of the methods with neural networks, and the room for improvement is smaller. However, these diagnosis methods can be embedded in others more complex and improve them. The bigger disadvantage of the methods based on the ITLMS is that they are much slower and the computational force needed is a lot bigger when compared with those that use neural networks.

Chapter 5.

New Industrial Data

As the available data is sparse, the mean shift algorithm was used to get more data to train neural networks. However, if some new gas analyses and corresponding diagnosis could be obtained, the confidence in the diagnosis methods would be bigger because of the independence between the two datasets: the one used to create virtual data and in the validation of the neural networks training and one new, independent database. Therefore, some efforts were made with Portuguese industry in order to get this data. EFACEC, a Portuguese power transformer manufacturer, showed its availability to provide it.

After getting all the dissolved gas concentrations and the IEC diagnosis, it was needed to get the true diagnosis of the machines. When a fault is detected, not all the machines are opened and repaired, thus, the number of the gas analysis needed was pretty large.

After all this process, a new database was built and was used only to validate the methods created in this thesis in order to compare the diagnosis results in both of the databases available.

It is needed to say that the real diagnoses made by EFACEC are based on experience and merely indicatives. Most faults are diagnosed with the indication of more than one fault, for example, if a transformer has signals of an electric discharge, it is classified as a low-energy discharge and high-energy discharge. The same happens in thermal faults. There is also an advantage to the IEC60599 diagnosis method in this comparison, because only cases with some gas limit violation were gathered. This way, a big limitation of this standard isn't taken into account. Besides, there are no cases where the power transformer is healthy because in this case the machine isn't opened or repaired. This can be another advantage to the IEC60599 method.

In the next page, table 5.1 contains a list with the cases gathered and the corresponding diagnosis.

Table 5.1 - New industrial data diagnoses

$\frac{C2H2}{C2H4}$	$\frac{CH4}{H2}$	$\frac{C2H4}{C2H6}$	EFACEC Diagnosis	IEC 60599 Diagnosis	Binary NN diagnosis	Autoencoders diagnosis
0,0007	4,000	0,7545	T1 or T2	T1	T1	T1
0,0005	4,000	1,6645	DL or DH	T2	T2	T2
0,0004	4,000	3,8253	T1 or T2	T2	T2	T2
0,1975	0,068	0,0641	PD or DL or DH	PD	PD	PD
0,0160	4,000	0,8747	T1 or T2	T1	T1	T1
0,0466	0,242	2,4668	PD or DL or DH		OK	T2
2,5579	0,374	4,0000	PD or DL or DH	DL	DL	DL
0,0281	4,000	1,6772	T1 or T2	T2	T2	OK
0,0000	4,000	1,8622	T1 or T2	T2	T2	T2
1,5193	0,529	3,9473	DL or DH	DH	DH	DH
1,5712	0,346	3,8847	DL or DH	DH	DH	DH
0,8876	0,023	0,5057	DL or DH	T1	DL	DL
0,0326	4,000	2,2181	PD or DL or DH	T2	T2	T2
2,7112	0,644	2,9912	DL or DH		DL	DL
2,8907	0,208	3,8738	T1 or T2	DL	DL	DL
0,9436	0,297	4,0000	DL or DH	DH	DH	DH
0,0748	2,614	2,8218	T1 or T2	T2	T2	T2
0,0555	4,000	0,1824	PD or DL or DH	T1	T1	T1
1,0074	0,336	0,0034	T1	T1	PD	DL
0,2026	4,000	2,6192	T1 or T2		T2	OK
2,6803	0,894	4,0000	T1 or T2		DH	DL
1,2877	0,400	1,3866	DL or DH	DL	DL	DL
0,1351	2,185	3,4940	T1 or T2	T2	T2	OK

Observing the table above, the easiest conclusion is that there are no classified points in the diagnosis methods used in this thesis, while the IEC60599 method returns 4. This is a big limitation of the IEC standard because, when an unclassified output appears, the user doesn't know how to proceed with the reparation. In the methods developed in this thesis this never happens, however misclassifications can appear.

If one reckons that the probability of the EFACEC diagnosis being correct is very high and assumes them as the corrects ones, comparing all the other methods with those and counting the correct diagnosis gives 14 correct diagnosis to the IEC60599 method, 16 to the method with binary output neural networks and 13 to the autoencoders method.

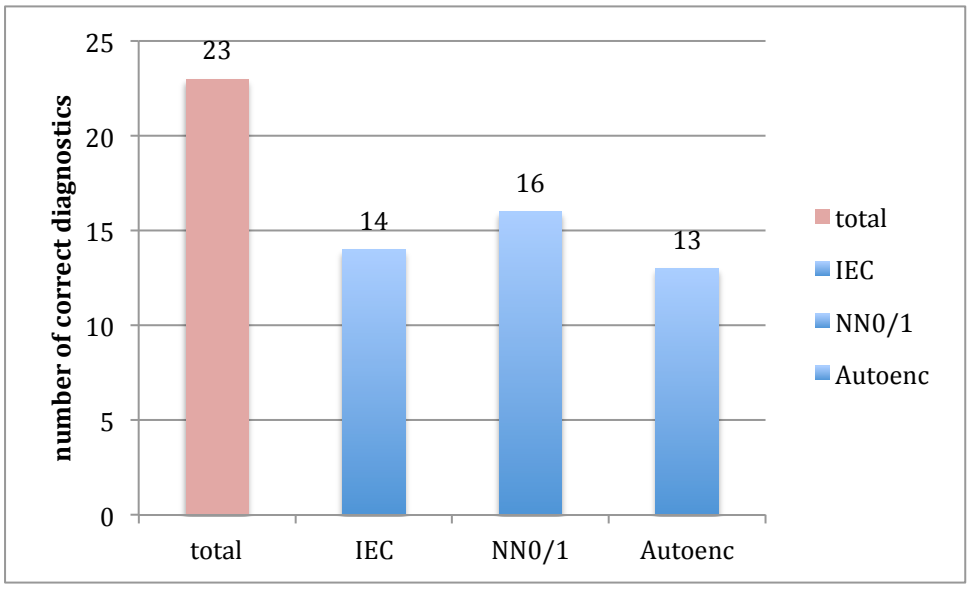


Figure 5.1 - Number of correct diagnoses of EFACEC data by method

Observing the number of correct diagnoses of the different methods it is possible to see that there is a small advantage to the binary neural networks method. Also, the number of correct diagnoses given by the IEC60599 standard method is high; nevertheless, one has to remember that a serious limitation of this method is not taken into account when only cases with violation of gas concentration limits are considered. Besides that, the diagnosis given by EFACEC can be wrong in some cases because, if the power transformer had two or more faults when the reparation was made, a fault could be seen and the other didn't.

Doing a more attentive analysis to the results given by the methods with neural networks, it can be noticed that there are some cases where the output is a healthy state in a transformer with and without on load tap changers. This can be seen as a problem of these methods, where there is room for improvement. Maybe it can be solved with a better training of the neural networks or with more data in the healthy states so as to assure the clusters are well represented.

One recommendation that may be drawn from this study is the interest in the normalization of the description of faults diagnosed, in such a way that it becomes compatible with the classifications in the IEC standard. This will be useful in assessing the merits of any diagnosis system implemented, because it will then allow comparisons with other methods, industrial or in the literature.

Another recommendation goes towards establishing a stronger feedback between observations of the state of transformers and the diagnoses produced. This will be useful in assessing the actual merits of the diagnosis method in use, including the field and laboratory procedures in dealing with the dissolved gas analyses.

Chapter 6.

Robustness tests

6.1. Introduction

After diagnosing the database, and in order to study how the methods with neural networks react to noise in the samples, a new algorithm was built where the gas concentration values are changed using a quasi-random number generator inside a Gaussian distribution with mean equal to 0 and variance equal to 25. The gases are changed in percentage, like in equation 6.1:

$$C^{new} = C^{original} * (1 + \beta/100) \quad \textbf{(6.1)}$$

where β is the quasi-random number. β was generated by MATLAB, using a built-in Gaussian quasi-random number generator with mean equal to 0 and standard deviation equal to twenty five.

Figure 6.1 is a block diagram of the algorithm used.

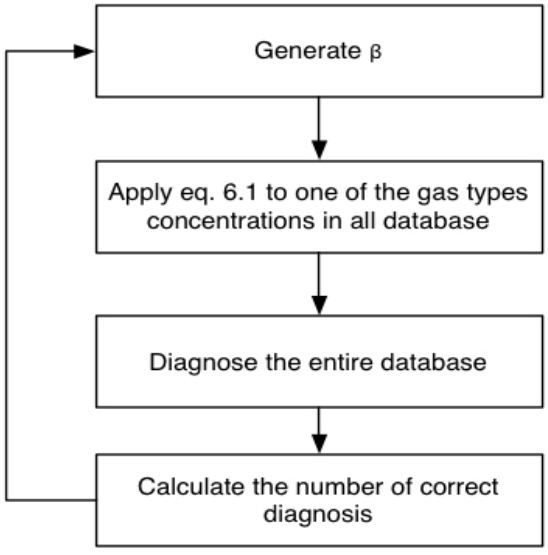


Figure 6.1 - Robustness tests algorithm

The noise simulation was done for all the five gases used to calculate the ratios, and was done gas by gas. Five hundred simulations were done for each gas and, in each iteration, the entire database was diagnosed.

This method allows doing a sensibilities test, which can be associated to the mathematical partial derivatives tool. Also, it simulates problems on the samples, like contamination, or in the gas analysis method, like a wrong machine calibration. With these tests, the confidence in the methods is bigger because they show how the diagnosis methods will behave in case of a bad oil sample or other problem with the sample or analysis.

It is needed to say that the methods' behaviour before noise with small amplitude is the most important to analyse because the probability of making a big error in analysis is very small, while a small error is easily made. This is the main reason to use a normal distribution to randomly create noise, because in a normal distribution, small values are more probable to appear than big ones.

This procedure was done to the diagnosis methods with best results, the one using autoencoders and the one with neural networks with binary outputs.

Due to lack of time, some other important robustness tests like the simultaneous variation of all the gas type concentrations are not presented in this thesis.

6.2. Autoencoders method

This test was applied first to the autoencoders diagnosis method. The results are shown in the graphics below. These figures represent the number of correct diagnoses (Y axle) for a certain β value (X axle).

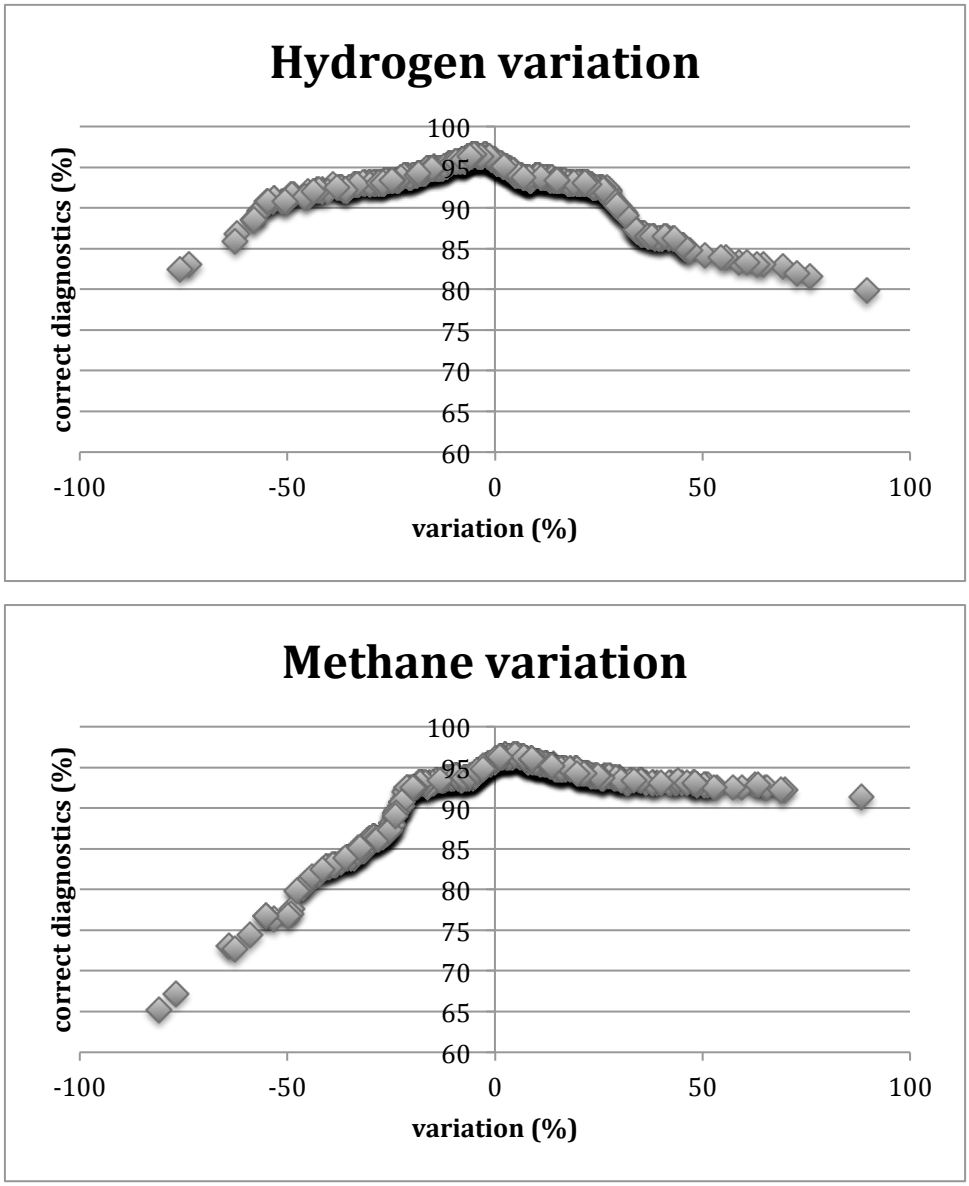
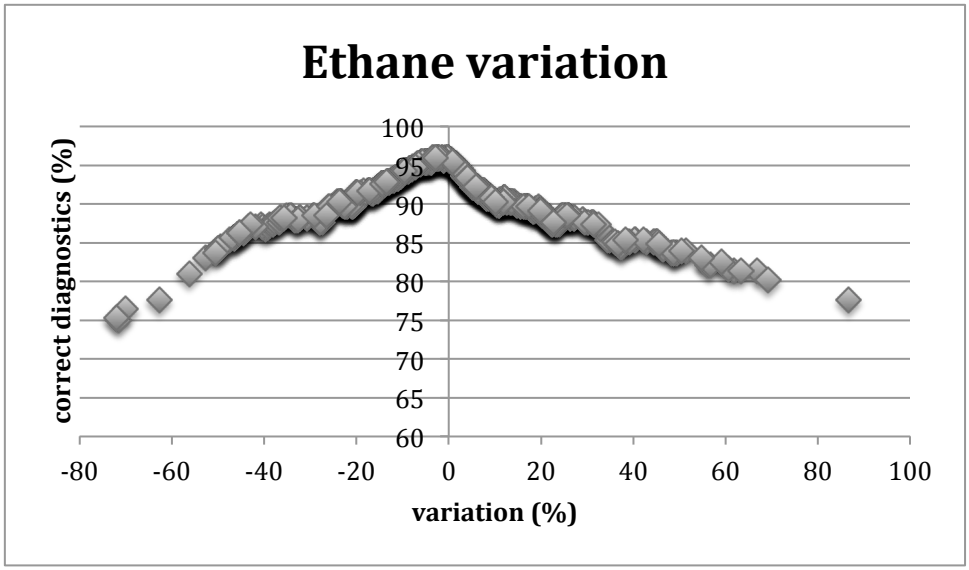
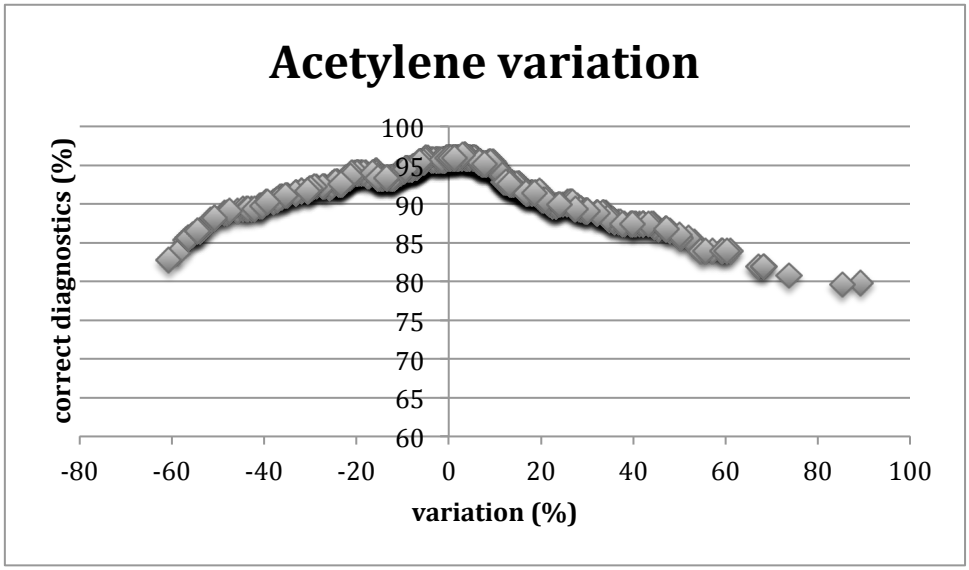


Figure 6.1a - Robustness testes to autoencoders method results





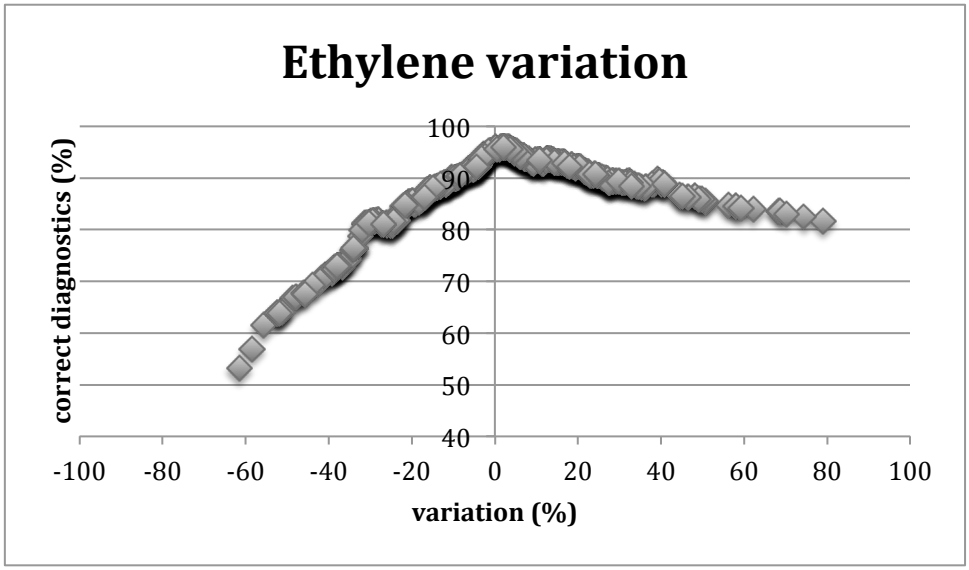


Figure 6.1b - Robustness testes to autoencoders method results

Analysing the graphics of the figure 6.1, it is possible to conclude that bigger variations meaning a smaller number of correct diagnosis. Also, the implications of noise in the analysis is variable between gases, namely, for the same β, the number of correct diagnosis is smaller in the case of ethylene than in the case of methane. However, this can happen because the ethylene is used to calculate two of the three ratios used in this diagnosis method. Therefore it is secure to say that the analysis of this gas concentration is critical, since errors can easily cause a wrong diagnosis.

Additionally, the methane and ethylene negative variations are much more inconvenient than the positives ones, while the opposite occurs in the acetylene and hydrogen variations. The curve of ethane variation is almost symmetric, meaning that both negative and positive variations produce the same effect in the diagnosis method.

Other conclusion of this test is that the diagnosis using autoencoders is pretty robust for small gas variations (±10%) if those are applied in one of the gases only, since the percentage of correct diagnosis is still very high when this happens. Moreover, a careful observation of the variation graphic allows deducing that noise by excess has a smaller implication in the diagnosis than errors by default, being the methane variation graphic the one that represent this property in a clearer way.

6.3. Neural networks with binary outputs method

The diagnosis method that uses neural networks with binary outputs was also submitted to this test. The results are shown in figure 6.2, below:

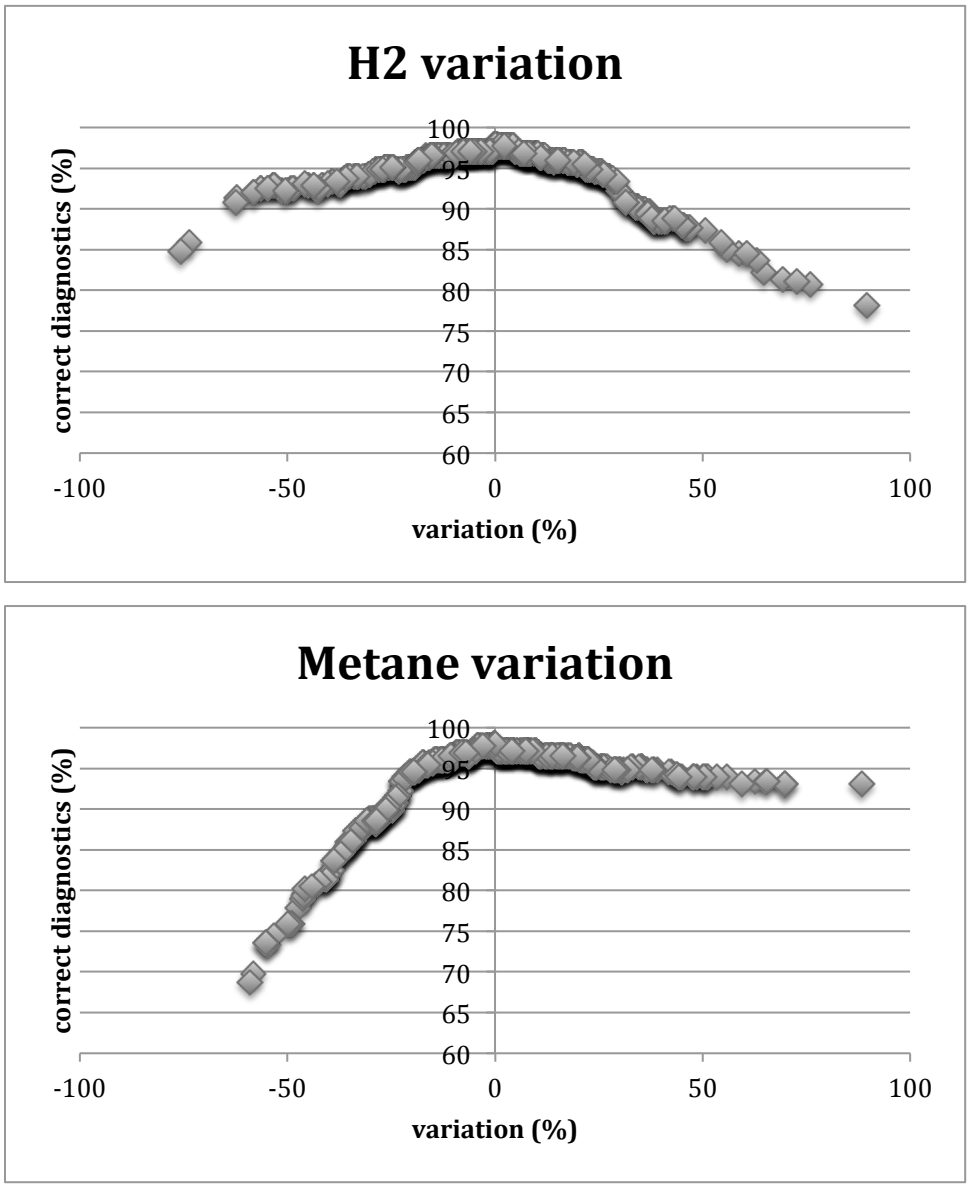


Figure 6.2a - Robustness tests to binary neural network method results

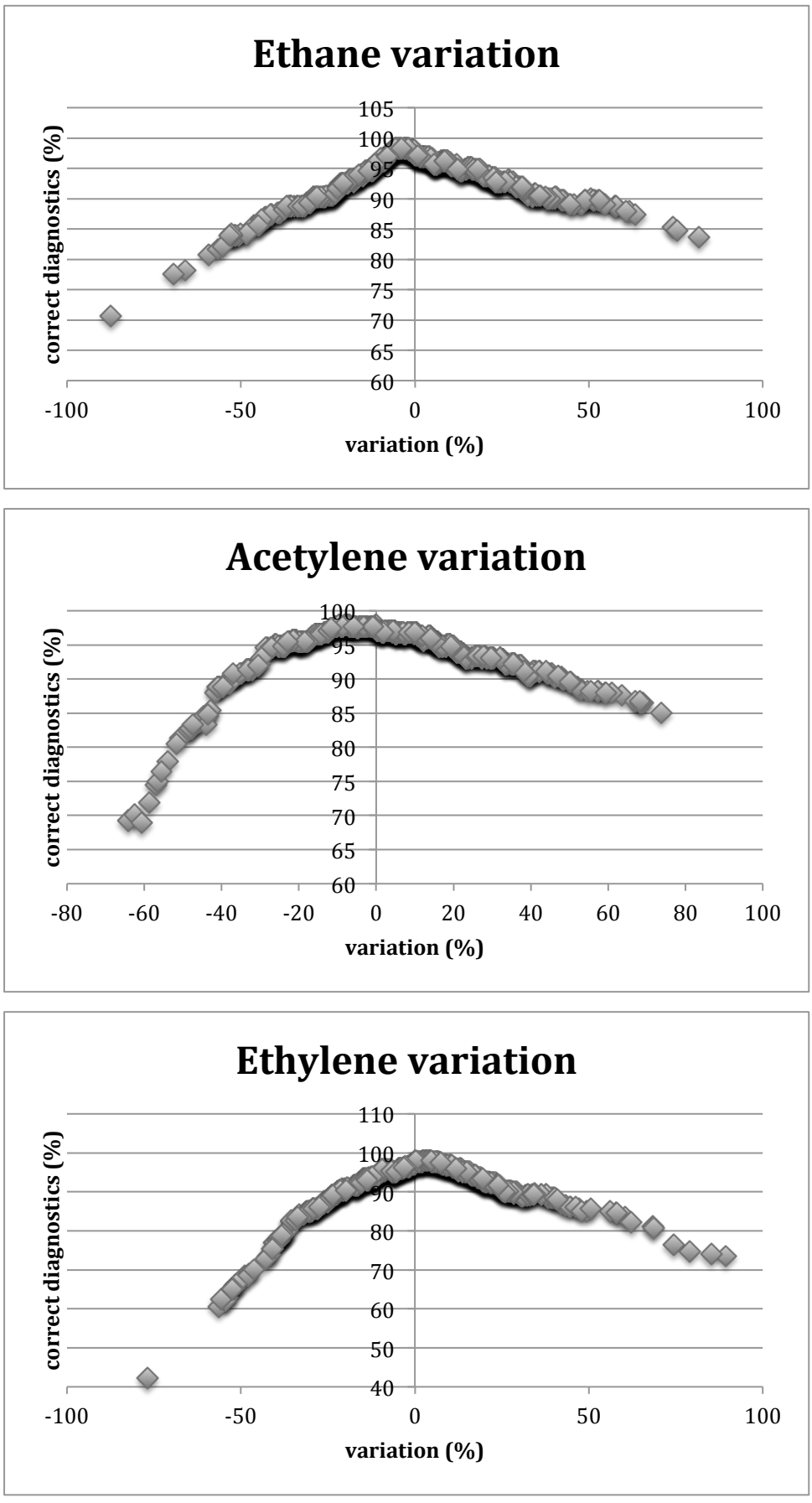


Figure 6.2b - Robustness tests to binary neural network method results

With a simple observation to the graphics above, it is possible to conclude that, once again, the worst results appear when the variation of the gas concentrations are bigger. Also, the behaviour of the diagnosis method before noise depends of the affected gas, with the ethylene being the critical gas, this is, the one where noise has biggest implications, while the variations in hydrogen have the smaller influence in the diagnosis results.

Furthermore, in this method, variations by default in all the gases concentration, unless in the hydrogen concentration, produce worse effects in the diagnosis methods than the ones produced by positive variation in the gases concentration. Regarding the hydrogen case, the opposite occurs, i.e. positive variations of the gas concentrations are worse than negative ones.

In spite of this fact, this method is very robust when small variations occur, with the number of correct diagnosis being very stable. It is possible do say this method is more robust before smaller variations in gases concentration than the autoencoders one, with curves showing a more stable behaviour in small variations. When the noise has bigger amplitude both methods behave in a very similar way.

Chapter 7.

Conclusions

In the end of this thesis several important conclusions can be drawn. First of all, and regarding the mean shift algorithm, it was proven that training neural networks with virtual data created using the Information Theoretic Learning Mean Shift algorithm is valid and returns very good results. However, these results depend of several factors, like the ITLMS parameters λ and σ. Analysing what was done, the σ can have big effects in the results obtained, and one needs to be careful when adjusting this parameter. The approach used in this thesis was a trial and error one. The training of neural networks was done to several different databases obtained with the ITLMS algorithm and the results were compared until there was no improvement.

In what concerns the power transformers diagnosis using dissolved gas analysis data, it was once again proven that this is a valid method, where good results can be achieved. The biggest problem of this method is the lack of data, which 'forced' the use of the Mean Shift algorithm in this thesis to create virtual data. With better archives and databases created by utilities and manufacturers building a diagnosis method could be easier and the results obtained validated in a more representative database.

In the diagnosis system using autoencoders the expected results, 100% of correct diagnoses, could not be achieved. This can be related with the mean shift parameters used or calculus accuracy. However, this method obtained pretty accurate results, with 96% of 438 real cases being diagnosed in a correct way. This result becomes even more important because it was achieved using autoencoders trained with virtual data only. Furthermore, the robustness tests applied to this method allowed to have a bigger confidence in its results because it was proven that the method isn't too sensitive to small gas concentration variations. This conclusion is very important if the method is to be applied in the industry, where errors can be made while capturing the oil sample or doing its analysis.

Despite of being a very slow method when the training of neural networks is being performed, the neural networks with binary outputs method accomplished the best results of all the methods studied, with 98% of the database being diagnosed correctly. Applying the same robustness tests to this method proved that it is even less sensitive to small variation in the gas concentrations than the autoencoders one. This is important because errors of small amplitude are the most likely to happen in the industry.

The good results obtained in both diagnosis methods referred above is clearly related with the fact of both being competitive methods, unlike most of the methods referred in literature. Using this kind of methods allow a better recognition of a fault by one of the neural networks, allowing the information about each fault to be stored in one of the neural networks. While in the methods which use only one neural network all the information about each one of the faulty/healthy states must be stored in its connection weights.

Regarding the other two methods studied in this work: the distance method and the steepest descent, the work done, besides not being exhaustive, showed that these methods can be interesting, not as standalone methods but to improve more complex ones. The more interesting method is the steepest descent one when recognising only faulty states. However, there is room for improvement and more studies are needed.

The new data that EFACEC provided to this thesis gave even more confidence to both methods with neural networks. Although this database is clearly favourable to the diagnosis using the IEC standard because most of the data in it contains gas ratio limits violations, both methods had similar or better results when compared to the IEC 60599.

Chapter 8.

Suggestions of work to do in the future

In spite of the advances done in this work, there is still work to do in the power transformer diagnosis.

Probably, the most important gap of the methods developed is that they are not capable of diagnosing the cellulose contained in the power transformer. This wasn't done because the data was even sparser than the data of other faults: the IEC TC10 database doesn't contain any of these cases and EFACEC just recently started doing it. However, once the data is available, adding this faulty state to one of the methods with neural networks is very simple, one only needs to add one more neural network in order to compete with the other seven and train them.

Besides several tries were made, more tests with other ITLMS λ and σ can allow the achievement of better results. However, first, one must develop a more efficient algorithm to the neural networks training because the backpropagation one is pretty slow and sensitive to local minima. There are several ones using evolutionary algorithms in the literature. Trying more λ and σ values can additionally be done in the diagnosis methods that don't use neural networks. These methods have potential and a deeper study can allow the achievement of better results. Other methods to calculate the 'distance' or similarity between a new, undiagnosed faulty/healthy state and the clusters representative can also be applied and studied. One of these methods can be the mutual information principle.

References

1. Castro, A.R.G., V. Miranda, and S. Lima. *Transformer fault diagnosis based on autoassociative neural networks*. in *Intelligent System Application to Power Systems (ISAP), 2011 16th International Conference on*. 2011.
2. IEC, *IEC-60599 - Mineral oil-impregnated electrical equipment in service - Guide to the interpretation of dissolved and free gases analysis*, 1999.
3. *IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers - Redline*. IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991) - Redline, 2009: p. 1-45.
4. Muhamad, N.A., B.T. Phung, and T.R. Blackburn. *Comparative study and analysis of DGA methods for mineral oil using fuzzy logic*. in *Power Engineering Conference, 2007. IPEC 2007. International*. 2007.
5. Hong-Tzer, Y. and L. Chiung-Chou, *Adaptive fuzzy diagnosis system for dissolved gas analysis of power transformers*. Power Delivery, IEEE Transactions on, 1999. **14**(4): p. 1342-1350.
6. Lee, J.P., et al. *Diagnosis of Power Transformer using Fuzzy Clustering and Radial Basis Function Neural Network*. in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*. 2006.
7. Yann-Chang, H., *Evolving neural nets for fault diagnosis of power transformers*. Power Delivery, IEEE Transactions on, 2003. **18**(3): p. 843-848.
8. Duval, M. and A. dePablo, *Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases*. Electrical Insulation Magazine, IEEE, 2001. **17**(2): p. 31-41.
9. Miranda, V. and A.R.G. Castro, *Improving the IEC table for transformer failure diagnosis with knowledge extraction from neural networks*. Power Delivery, IEEE Transactions on, 2005. **20**(4): p. 2509-2516.
10. Castro, A.R.G. and V. Miranda, *An interpretation of neural networks as inference engines with application to transformer failure diagnosis*. International Journal of Electrical Power & Energy Systems, 2005. **vol.27**: p. pp.620-626.
11. Zhenyuan, W., L. Yilu, and P.J. Griffin, *A combined ANN and expert system tool for transformer fault diagnosis*. Power Delivery, IEEE Transactions on, 1998. **13**(4): p. 1224-1229.
12. Dong-Hui, L., B. Jian-Peng, and S. Xiao-Yun. *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 *Machine Learning and Cybernetics, 2008 International Conference on*. 2008.
13. Thang, K.F., et al., *Analysis of power transformer dissolved gas data using the self-organizing map*. Power Delivery, IEEE Transactions on, 2003. **18**(4): p. 1241-1248.
14. Tomsovic, K., M. Tapper, and T. Ingvarsson, *A fuzzy information approach to integrating different transformer diagnostic methods*. Power Delivery, IEEE Transactions on, 1993. **8**(3): p. 1638-1646.
15. Li, H., D. Xiao, and Y. Chen. *Wavelet ANN based transformer fault diagnosis using gas-in-oil analysis*. in *Properties and Applications of Dielectric Materials, 2000. Proceedings of the 6th International Conference on*. 2000.
16. Zhang, Y., et al., *An artificial neural network approach to transformer fault diagnosis*. Power Delivery, IEEE Transactions on, 1996. **11**(4): p. 1836-1841.
17. Castro, A.R.G. and V. Miranda, *Knowledge discovery in neural networks with application to transformer failure diagnosis*. IEEE Transactions on Power Systems, 2005. **20**(2): p. 717-724.
18. Guardado, J.L., et al., *A comparative study of neural network efficiency in power transformers diagnosis using dissolved gas analysis*. Power Delivery, IEEE Transactions on, 2001. **16**(4): p. 643-647.
19. Lv, G., et al., *Fault diagnosis of power transformer based on multi-layer SVM classifier*. Electric Power Systems Research, 2005. **75**(1): p. 9-15.

20. Richardson, Z.J., et al., *A Probabilistic Classifier for Transformer Dissolved Gas Analysis With a Particle Swarm Optimizer*. Power Delivery, IEEE Transactions on, 2008. **23**(2): p. 751-759.
21. Dong, L., et al., *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, 2008. **78**(1): p. 129-136.
22. Wang, M.-H., et al., *A novel clustering algorithm based on the extension theory and genetic algorithm*. Expert Systems with Applications, 2009. **36**(4): p. 8269-8276.
23. Fei, S.-w. and X.-b. Zhang, *Fault diagnosis of power transformer based on support vector machine with genetic algorithm*. Expert Systems with Applications, 2009. **36**(8): p. 11352-11357.
24. Mat Isa, N.A. and W.M.F.W. Mamat, *Clustered-Hybrid Multilayer Perceptron network for pattern recognition application*. Applied Soft Computing, 2011. **11**(1): p. 1457-1466.
25. Castro, A. and V. Miranda, *Sistema inteligente para diagnóstico de faltas incipientes em transformadores baseado em redes neurais auto-associativas*, in *Sistema inteligente para diagnóstico de faltas incipientes em transformadores baseado em redes neurais auto-associativas*May 2010: PA, Brazil.
26. Bacha, K., S. Souahlia, and M. Gossa, *Power transformer fault diagnosis based on dissolved gas analysis by support vector machine*. Electric Power Systems Research, 2012. **83**(1): p. 73-79.
27. Siemens. *Online DGA (Dissolved Gas Analysis) Monitoring*. Available from: http://www.energy.siemens.com/us/en/services/power-transmission-distribution/transformer-lifecycle-management/online-dga-monitoring.htm.
28. DobleEngineering. *Doble Laboratory Testing*. Available from: http://www.doble.com/services/lab_services_testing.html.
29. PowertechLabs. *Oil Quality Testing*. Available from: http://www.powertechlabs.com/engineering-consulting/chemical-analysis/oil-quality-testing/.
30. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification*. 2nd ed. ed2001: Wiley.
31. Parzen, E., *On estimation of a probability density function and mode*. Annals of Mathematical Statistics, Sep. 1962. **vol. 33**(issue 3).
32. Comaniciu, D. and P. Meer, *Mean shift: a robust approach toward feature space analysis*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002. **24**(5): p. 603-619.
33. Duda, R.O. and P.E. Hart, *Patter classification and scene analysis*1973: Wiley.
34. Szeliski, R., *Computer Vision: Algorithms and Applications*2010: Springer.
35. Pooransingh, A., C.A. Radix, and A. Kokaram. *The path assigned mean shift algorithm: A new fast mean shift implementation for colour image segmentation*. in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. 2008.
36. Zhi-Qiang, W. and C. Zi-Xing. *Mean Shift Algorithm and its Application in Tracking of Objects*. in *Machine Learning and Cybernetics, 2006 International Conference on*. 2006.
37. Shah, K.A., et al. *Application of Mean-Shift algorithm for license plate localization*. in *Engineering (NUiCONE), 2011 Nirma University International Conference on*. 2011.
38. Pengfei, L., W. Shaoru, and J. Junfeng. *The segmentation in textile printing image based on mean shift*. in *Computer-Aided Industrial Design & Conceptual Design, 2009. CAID & CD 2009. IEEE 10th International Conference on*. 2009.
39. Fukunaga, K. and L. Hostetler, *The estimation of the gradient of a density function, with applications in pattern recognition*. Information Theory, IEEE Transactions on, 1975. **21**(1): p. 32-40.
40. Yizong, C., *Mean shift, mode seeking, and clustering*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1995. **17**(8): p. 790-799.
41. Rao, S., A. de Medeiros Martins, and J.C. Príncipe, *Mean shift: An information theoretic perspective*. Pattern Recognition Letters, 2009. **30**(3): p. 222-230.
42. Rao, S., et al. *Information Theoretic Mean Shift Algorithm*. in *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*. 2006.
43. Principe, J.C., *Information Theoretic Learning*. Information Science and Statistics2010: Springer.

44. Hinton, G.E. and R.R. Salakhutdinov, *Reducing the Dimensionality of Data with Neural Networks*. Science, 2006. **313**(5786): p. 504-507.
45. Golomb, B. and T. Sejnowski, *Sex Recognition from Faces Using Neural Networks*. Applications of Neural Networks, ed. K.A. Publishers1995. 71-92.
46. Fleming, M.K. and G.W. Cottrel, *Categorization of faces using unsupervised feature extraction*. Proceeding of the IJCNN - International Joint Conference on Neural Networks, 1990. **Vol.2**: p. 65-70.
47. Cottrel, G.W., P. Munro, and D. Zipser, *Learning internal representations from gray scale images: an example of extensional programming* Proceeding of the IJCNN - International Joint Conference on Neural Networks, 1990.
48. Miranda, V., et al., *Reconstructing Missing Data in State Estimation With Autoencoders*. Power Systems, IEEE Transactions on, 2012. **27**(2): p. 604-611.
49. Thompson, B.B., R.J. Marks, II, and M.A. El-Sharkawi. *On the contractive nature of autoencoders: application to missing sensor restoration*. in *Neural Networks, 2003. Proceedings of the International Joint Conference on*. 2003.
50. Tai-Ning, Y. and W. Sheng-De, *Fuzzy auto-associative neural networks for principal component extraction of noisy data*. Neural Networks, IEEE Transactions on, 2000. **11**(3): p. 808-810.
51. Kamimura, R. and S. Nakanishi. *Information maximization for feature detection and pattern classification by autoencoders*. in *Neural Networks, 1995. Proceedings., IEEE International Conference on*. 1995.
52. Thompson, B.B., et al. *Implicit learning in autoencoder novelty assessment*. in *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on*. 2002.
53. Jain, A.K., M. Jianchang, and K.M. Mohiuddin, *Artificial neural networks: a tutorial*. Computer, 1996. **29**(3): p. 31-44.
54. Hastie, T. and W. Stuetzle, *Principal curves*. Journal of the American Statistical Association, 1989. **Vol. 84**: p. 502-516.
55. Japkowiczz, N., S.J. Hanson, and M.A. Gluck, *Nonlinear Autoassociation is not Equivalent to PCA*. Neural Computation, 2000. **Vol. 12**(3): p. 531-545.
56. Sanger, T.D., *Optimal unsupervised learning in a single layer linear neural network*. Neural Networks, 1989. **vol. 2**: p. 450-473.
57. Hagan, M.T. and M.B. Menhaj, *Training feedforward networks with the Marquardt algorithm*. Neural Networks, IEEE Transactions on, 1994. **5**(6): p. 989-993.
58. Tang, J.-l., Y.-j. Liu, and F.-s. Wu. *Levenberg-Marquardt neural network for gear fault diagnosis*. in *2nd International Conference on Networking and Digital Society (ICNDS)*.

Appendixes

Appendix A – ITLMS cluster features seeking

The following figures illustrate the final arrangement of data particles after a mean shift procedure.

Appendix A.1 – Modes seeking ($\lambda = 1$, $\sigma = mean\ (std)$)

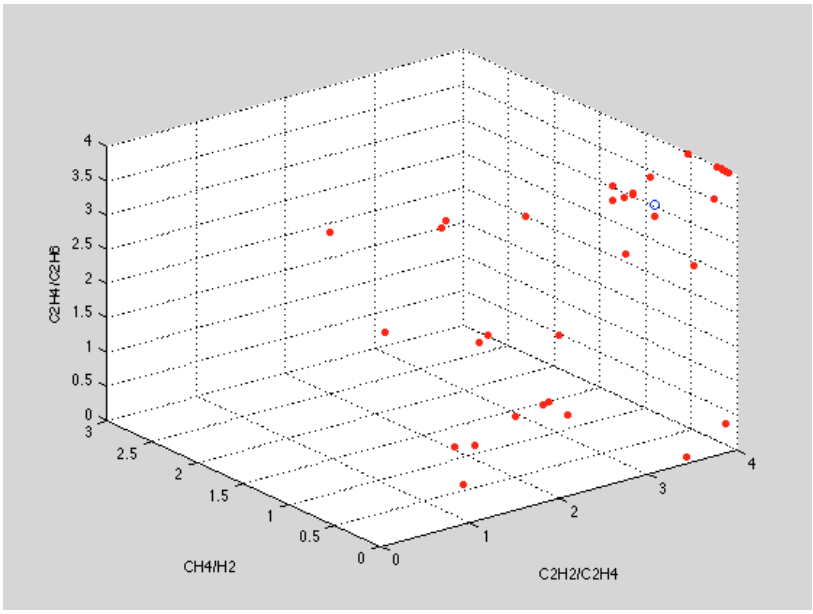


Figure A.1.1 – Low-energy discharge modes seeking

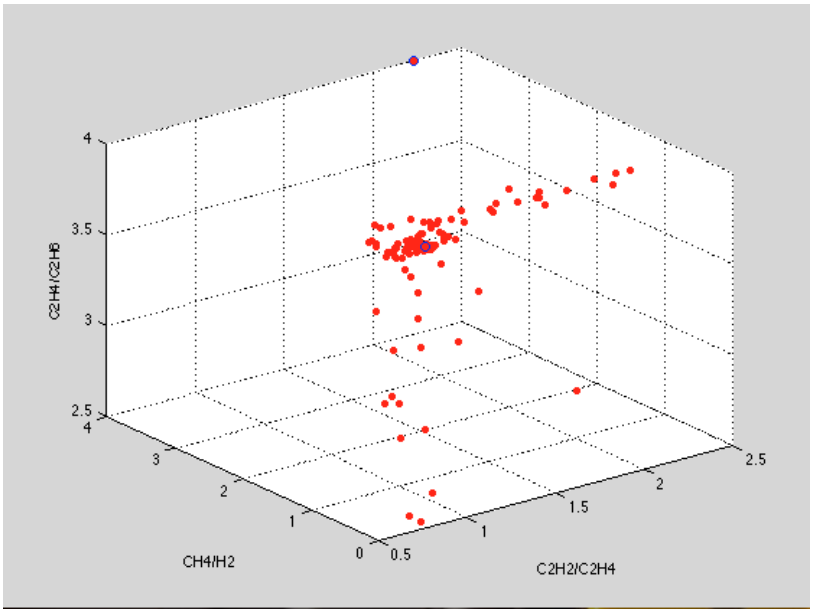


Figure A.1.2 – High-energy discharge modes seeking

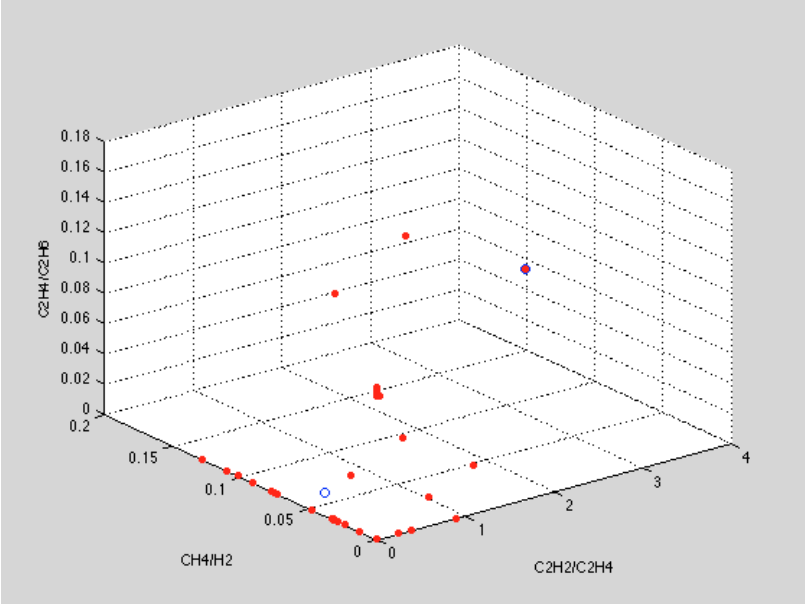


Figure A.1.3 – Partial discharge modes seeking

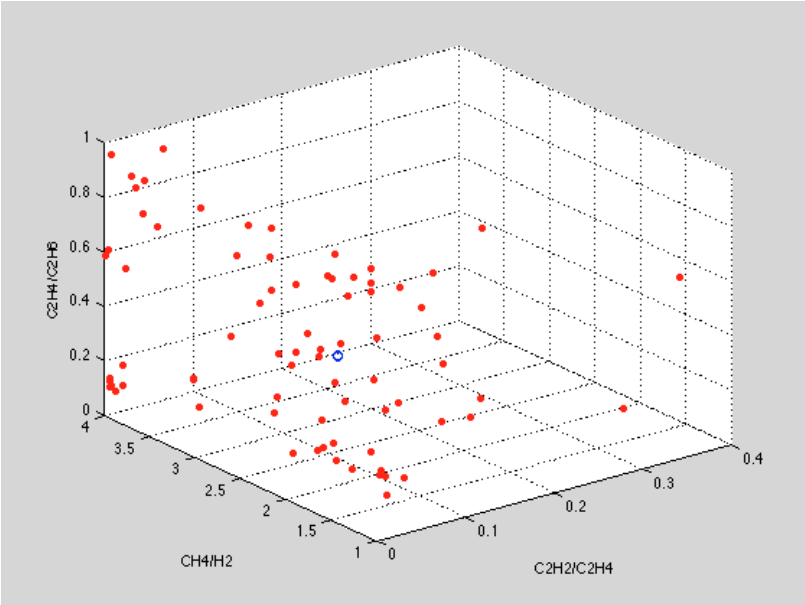


Figure A.1.4 – Thermal fault (T<700°C) modes seeking

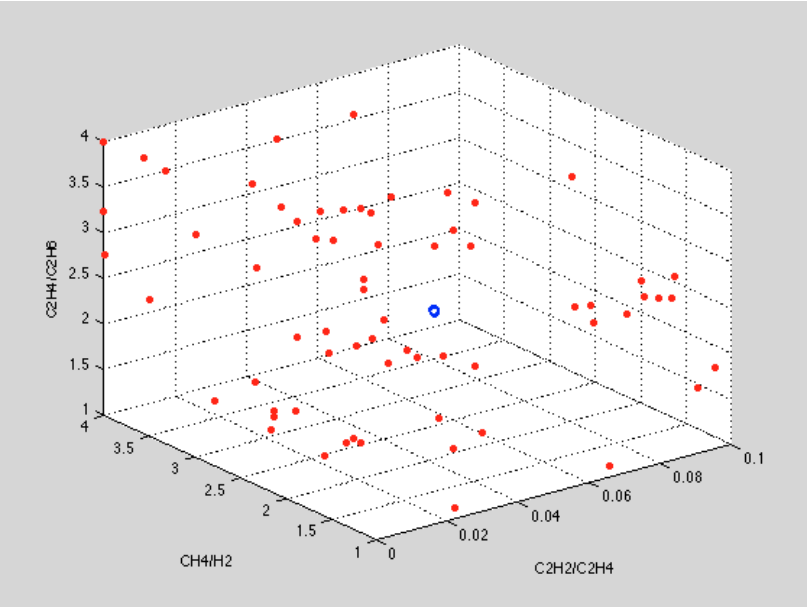


Figure A.1.5 – Thermal fault (T>700°C) modes seeking

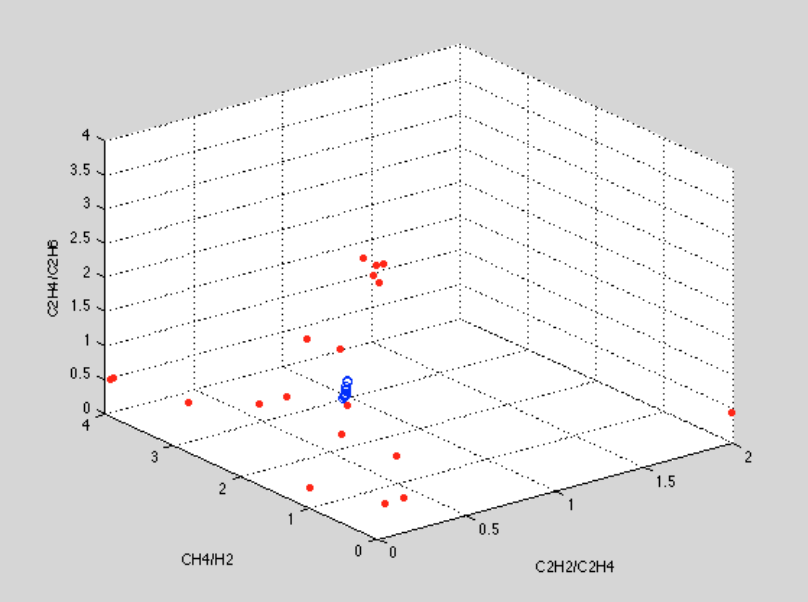


Figure A.1.6 – Healthy transformer without OLTC modes seeking

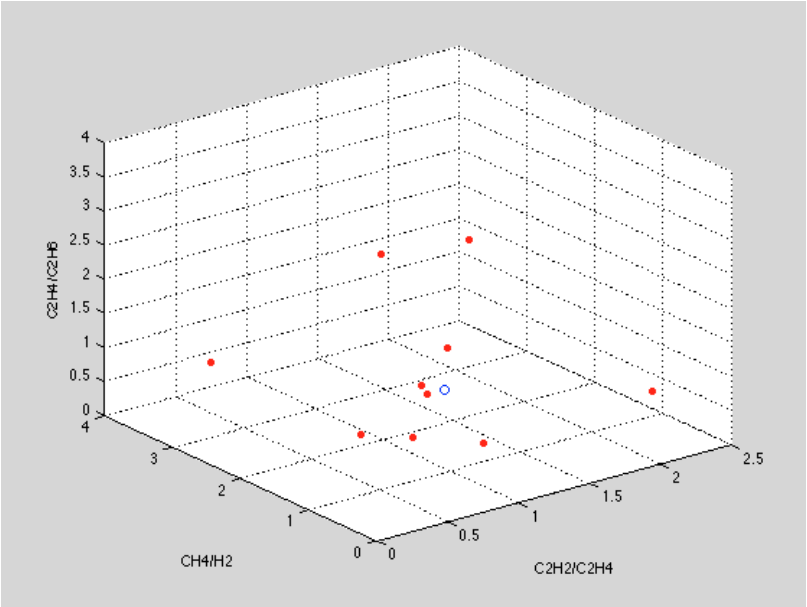


Figure A.1.7 – Healthy transformer with OLTC modes seeking

Apendix A.2 – Principal curve seeking ($\lambda = 2$, $\sigma = mean\ (std)$)

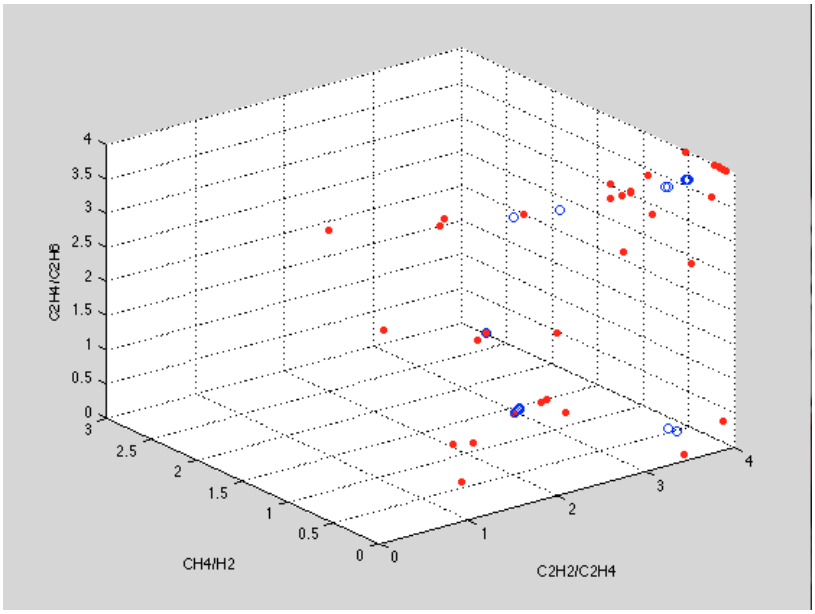


Figure A.2.1 – Low-energy discharge principal curve seeking

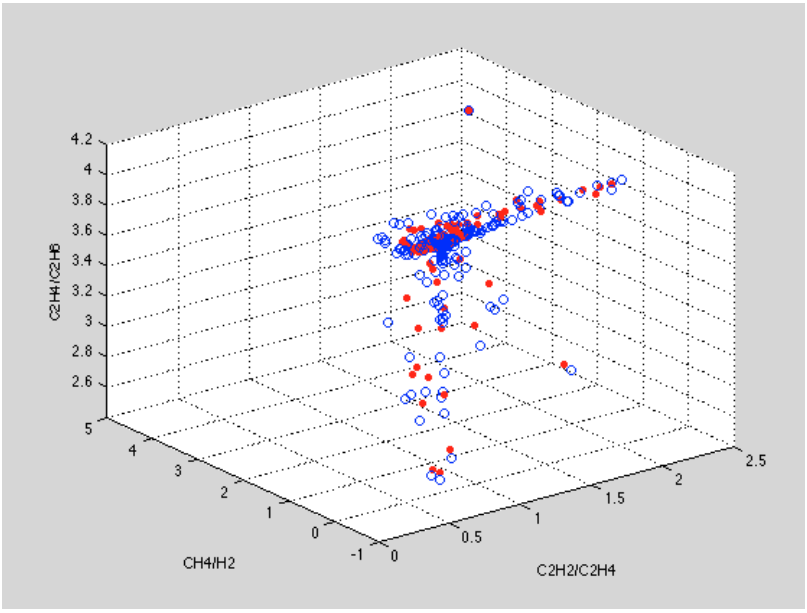


Figure A.2.2 – High-energy discharge principal curve seeking

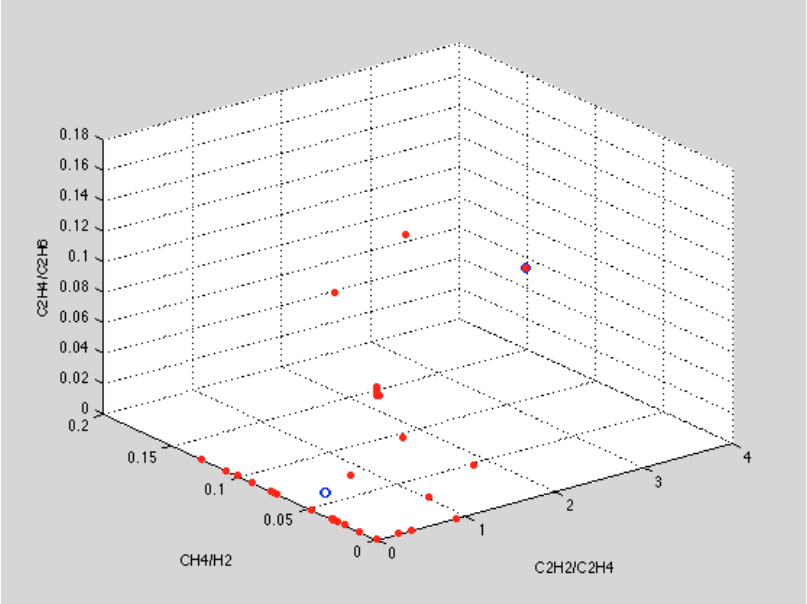


Figure A.2.3 – Partial discharge principal curve seeking

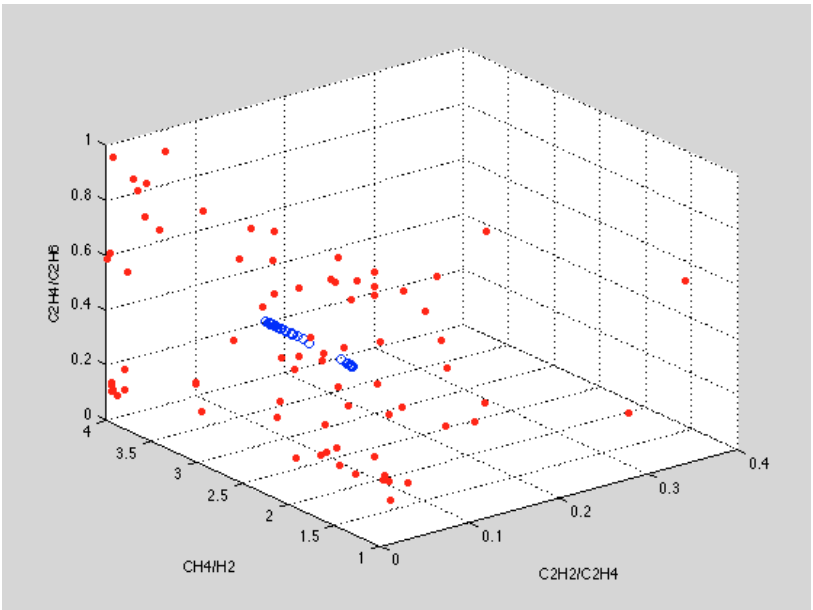


Figure A.2.4 – Thermal fault (T<700°C) principal curve seeking

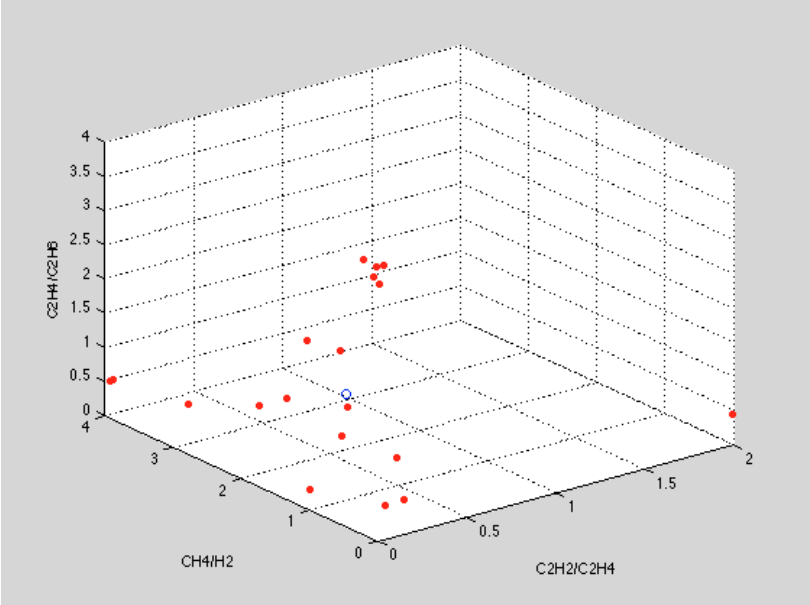


Figure A.2.5- Thermal fault (T>700°C) principal curve seeking

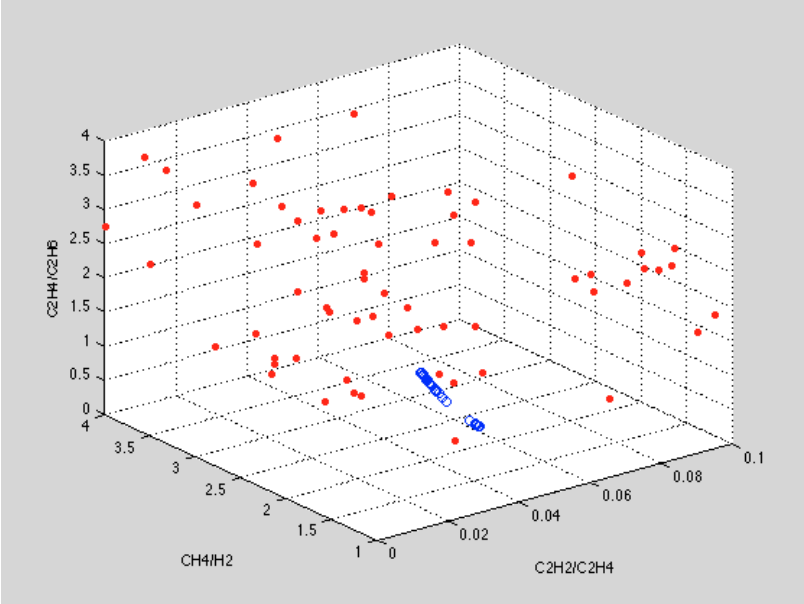


Figure A.2.6 – Healthy transformer without OLTC principal curve seeking

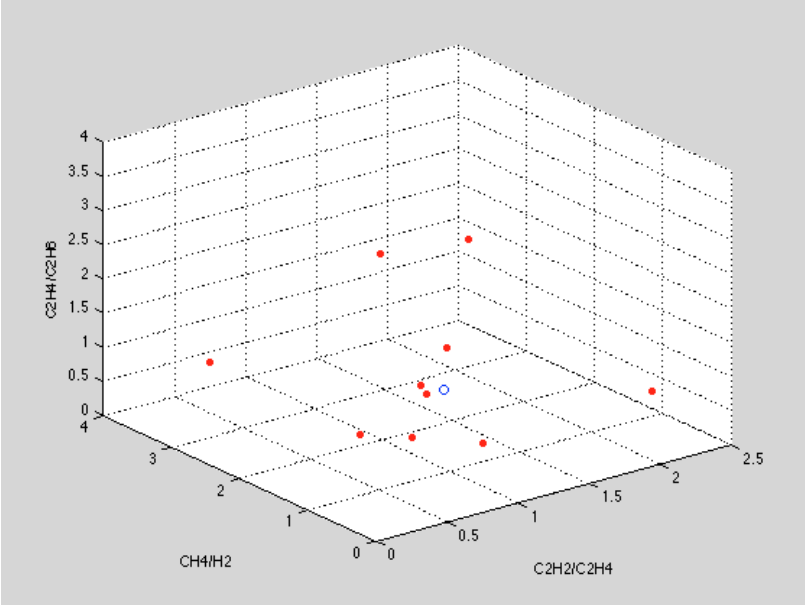


Figure A.2.6 – Healthy transformer with OLTC principal curve seeking

A.3 – Local modes seeking ($\lambda = 1$, $\sigma = various$)

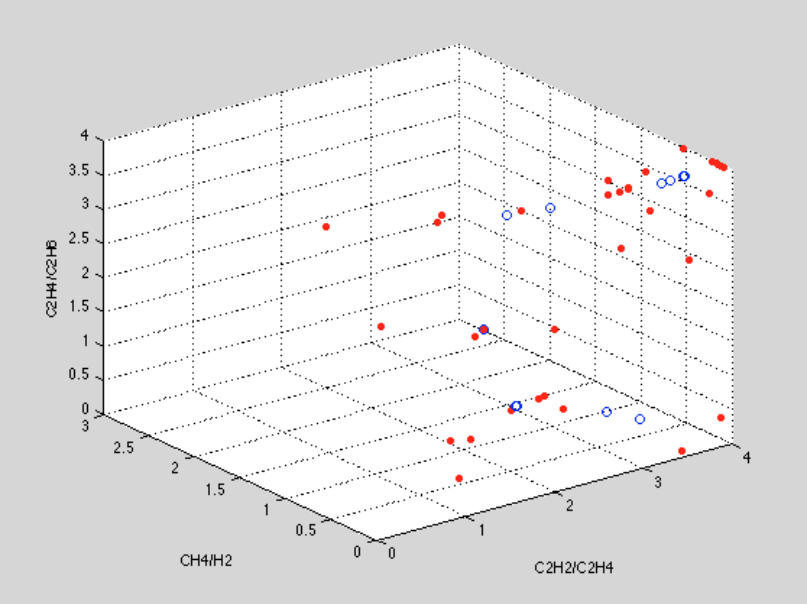


Figure A.3.1 – Low-energy discharge local modes seeking

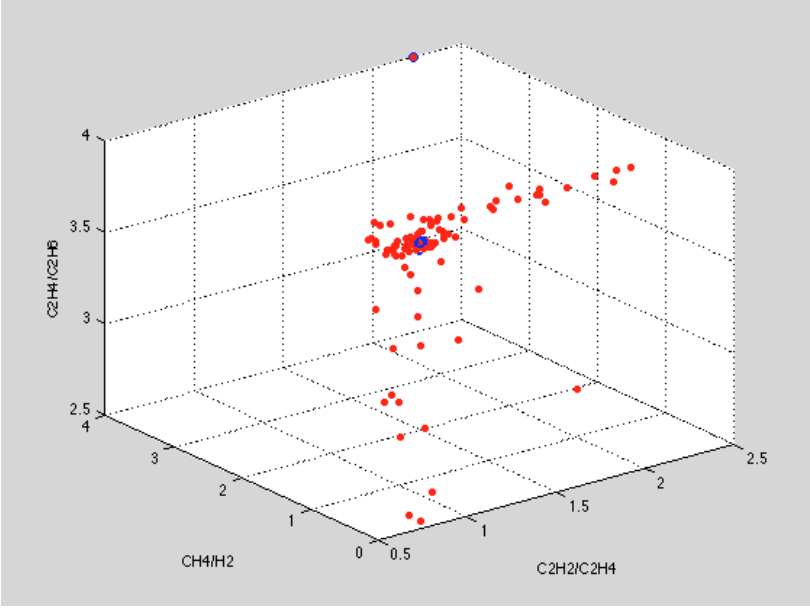


Figure A.3.2 – High-energy discharge local modes seeking

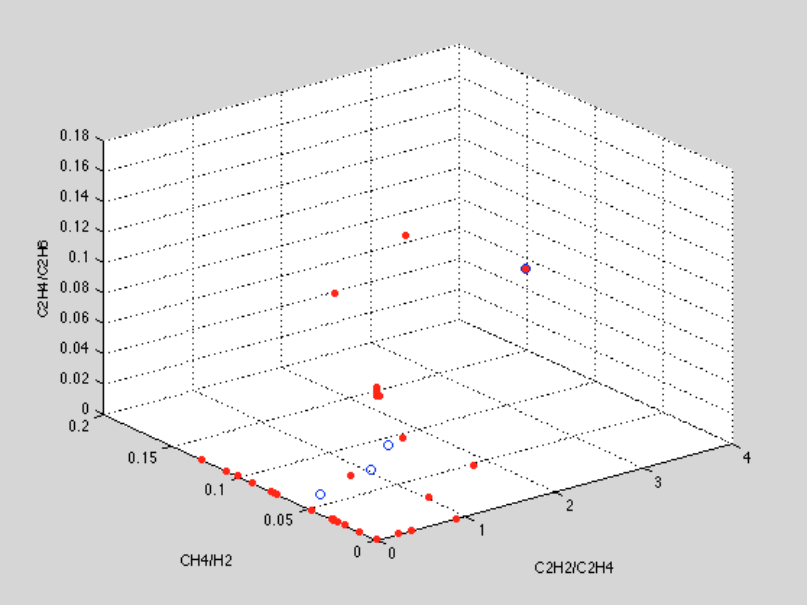


Figure A.3.3 – Partial discharge local modes seeking

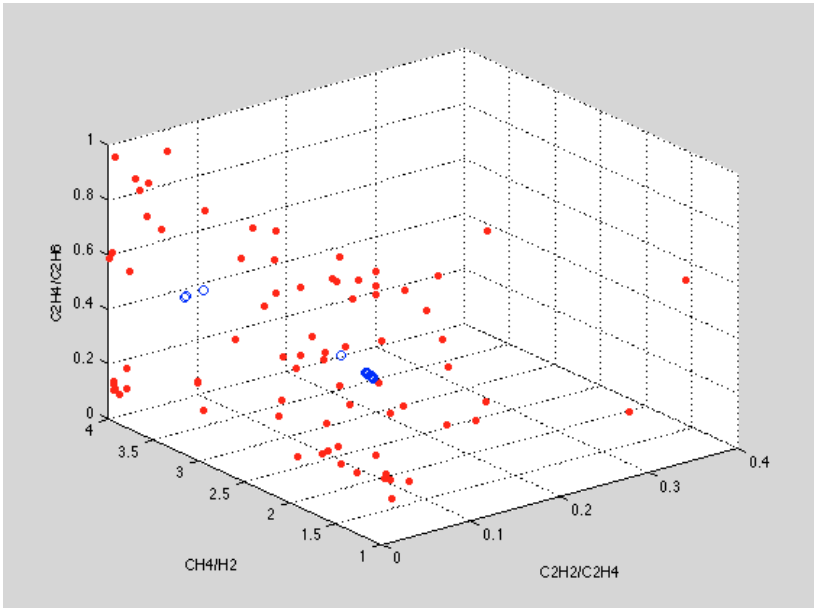


Figure A.3.4 – Thermal fault (T<700°C) local modes seeking

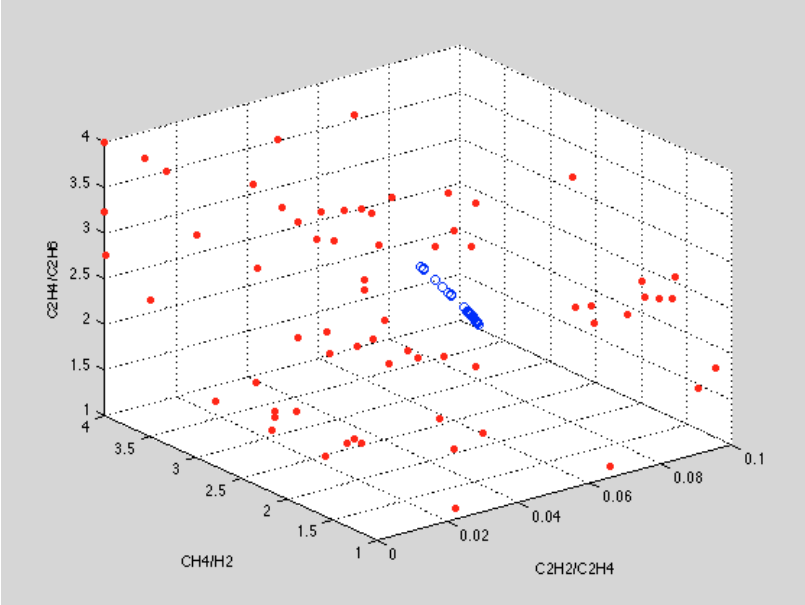


Figure A.3.5 – Thermal fault (T>700°C) local modes seeking

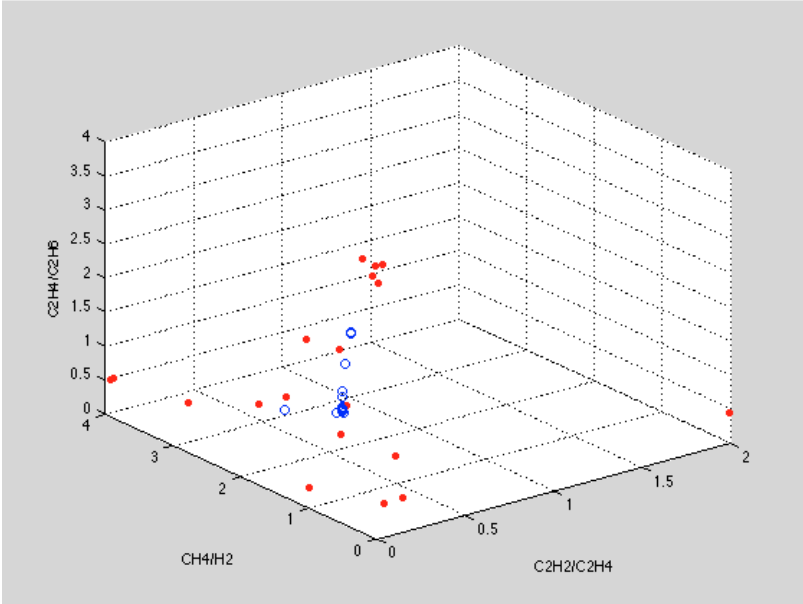


Figure A.3.6 – Healthy transformers without OLTC local modes seeking

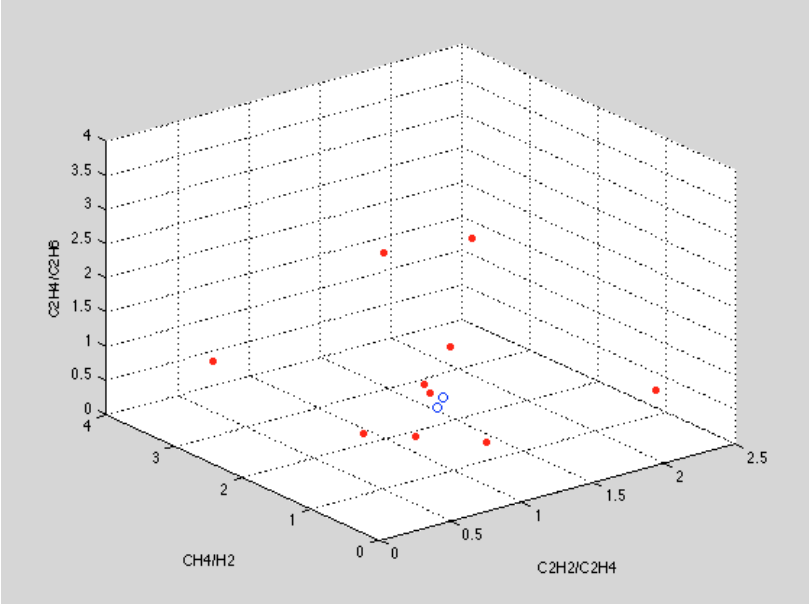


Figure A.3.7 – Healthy transformers with OLTC local modes seeking

A.4 – Finer cluster structures seeking ($\boldsymbol{\lambda} = 7$, $\sigma = various$)

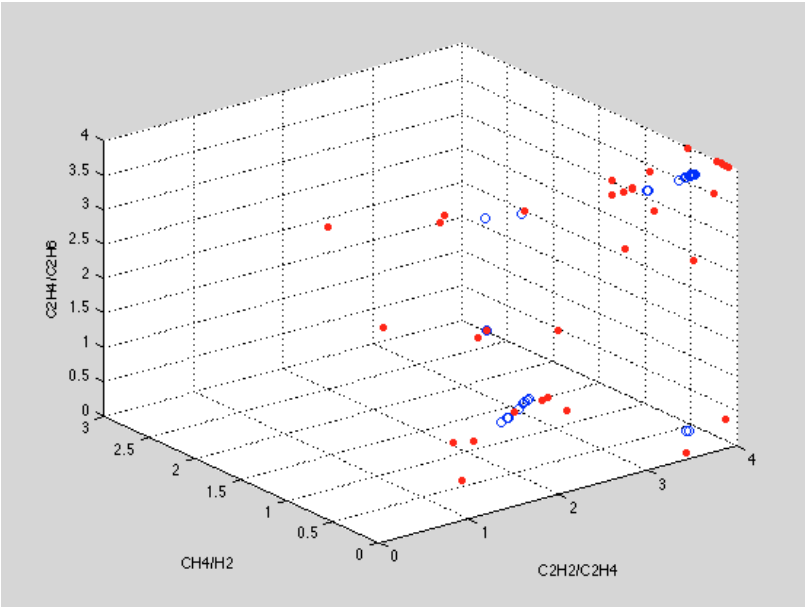


Figure A.4.1 – Low-energy discharge finer structures seeking

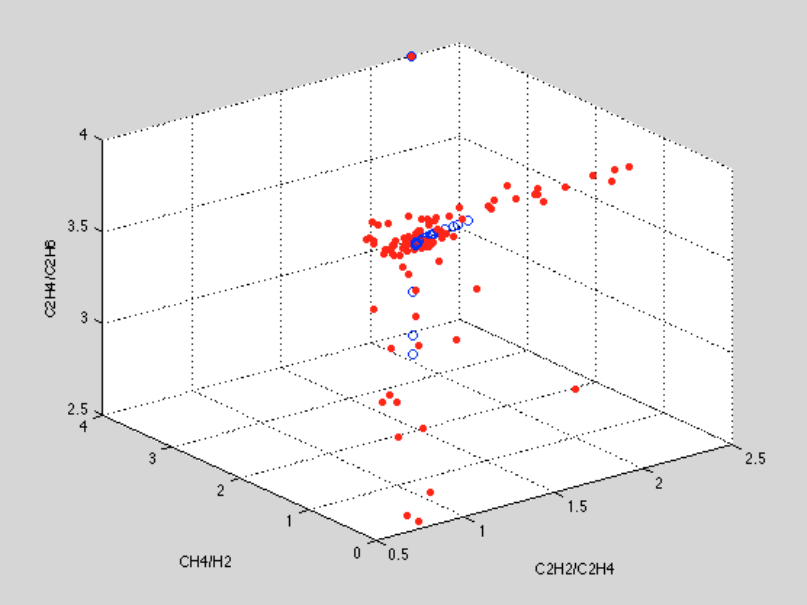


Figure A.4.2 – Low-energy discharge finer structures seeking

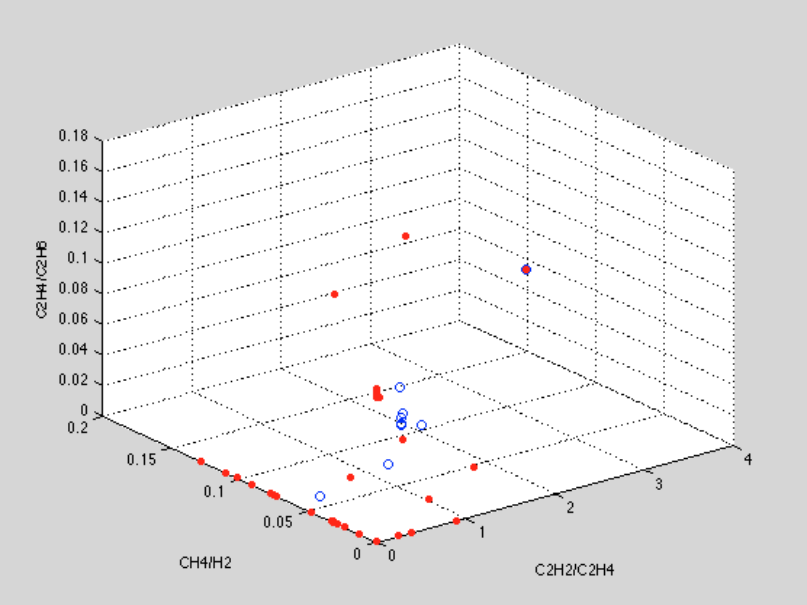


Figure A.4.3 - Partial discharge finer structures seeking

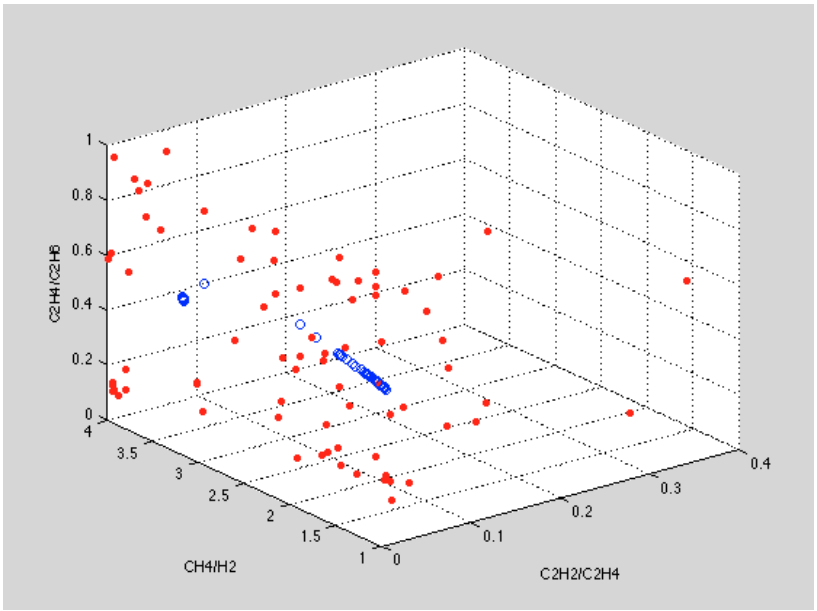


Figure A.4.4 – Thermal fault (T<700°C) finer structures seeking

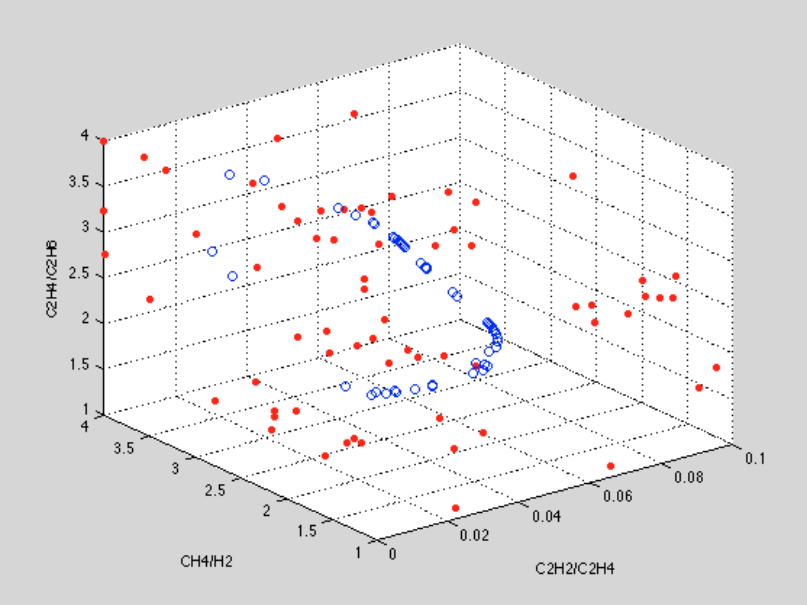


Figure A.4.5 – Thermal fault (T>700°C) finer structures seeking

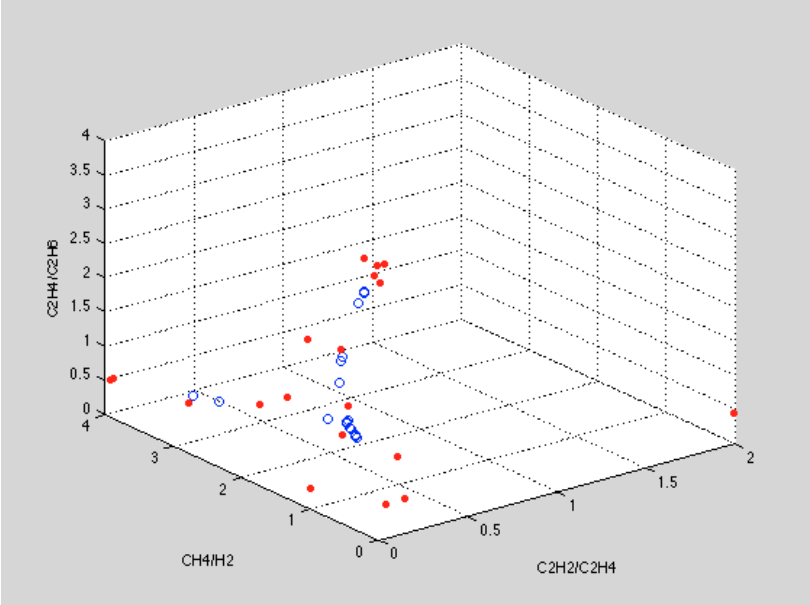


Figure A.4.6- Healthy transformer without OLTC finer structures seeking

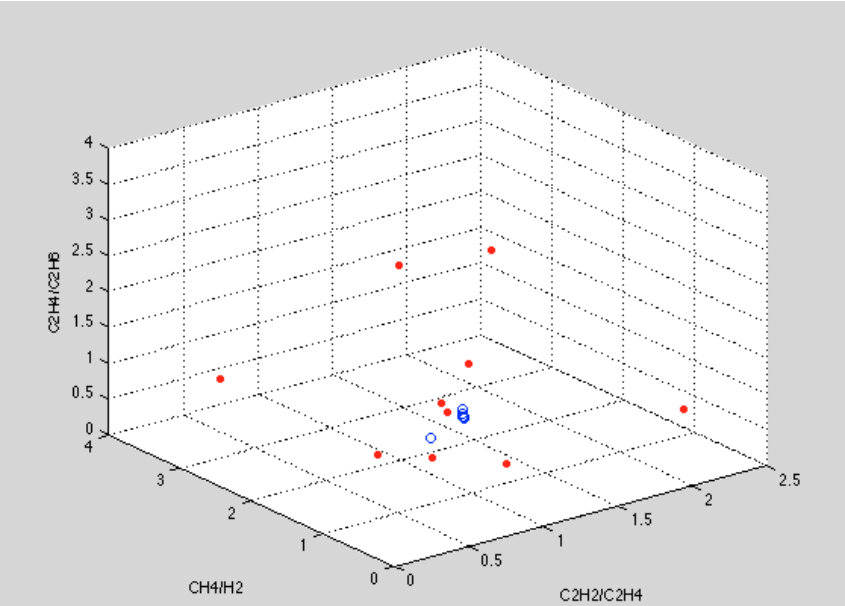


Figure A.4.7- Healthy transformer with OLTC finer structures seeking

Appendix B – Densification trick

The following figures illustrate the arrangement of data particles of successive mean shift iterations.

B.1 – Using $\lambda = 1$, $\sigma = mean\ (std)$

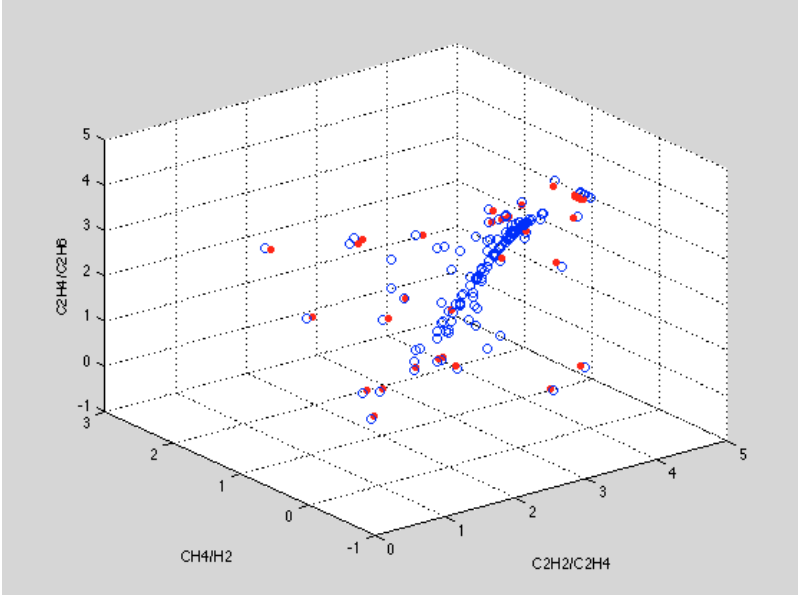


Figure B.1.1 – Low-energy discharge densification trick ($\lambda = 1$, $\sigma = mean\ (std)$)

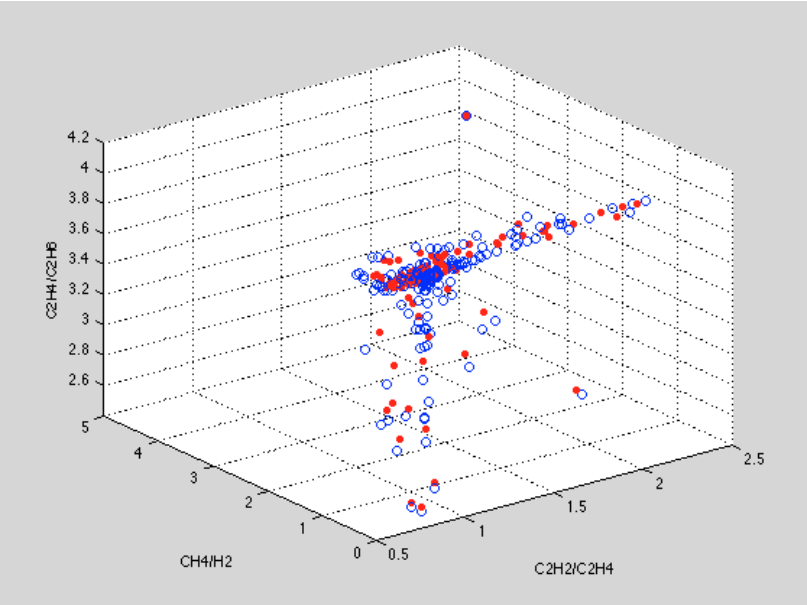


Figure B.1.2 – Low-energy discharge densification trick ($\lambda = 1$, $\sigma = mean\ (std)$)

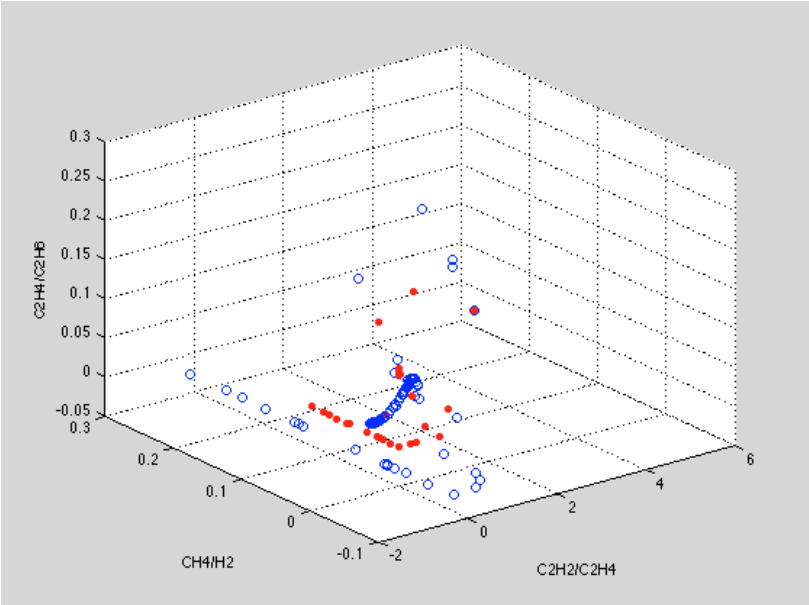


Figure B.1.3 - Partial discharge densification trick ($\lambda = 1$, $\sigma = mean\ (std)$)

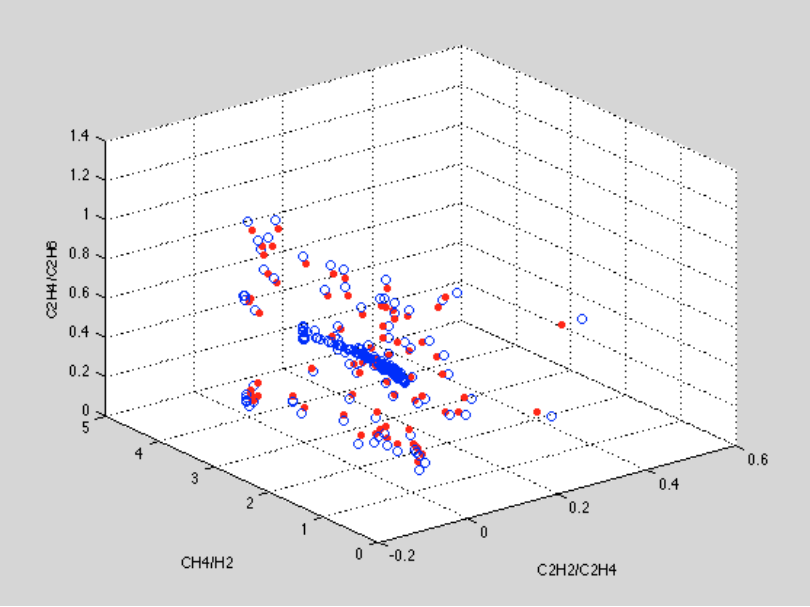


Figure B.1.4 - Thermal fault (T<700°C) densification trick ($\lambda = 1$, $\sigma = mean\ (std)$)

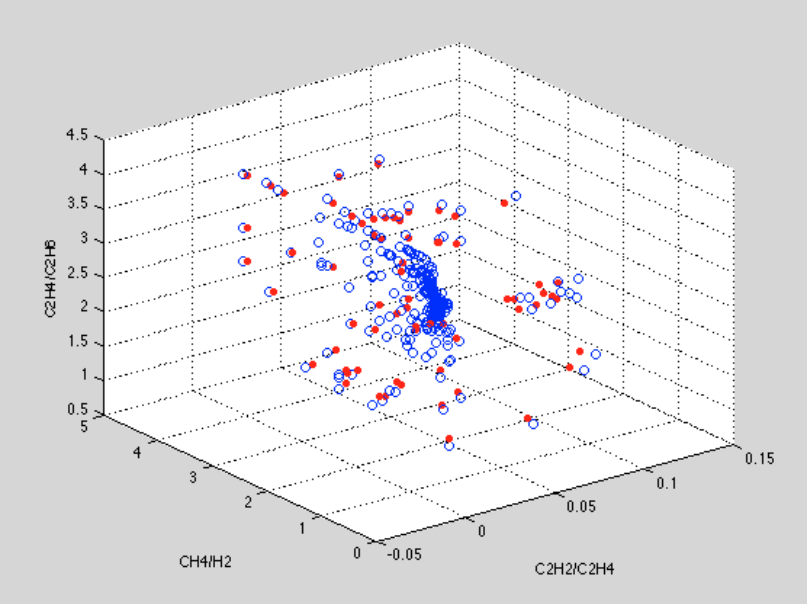


Figure B.1.5 - Thermal fault (T>700°C) densification trick ($\lambda = 1$, $\sigma = mean\ (std)$)

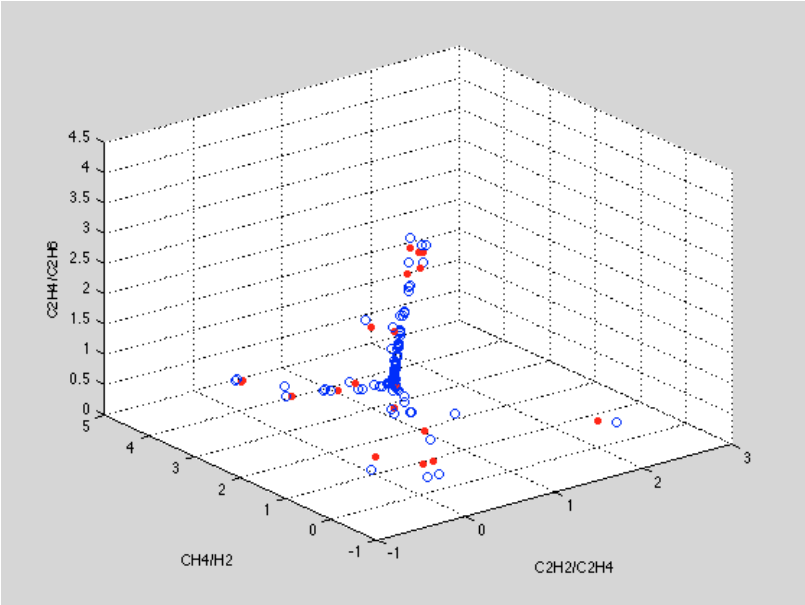


Figure B.1.6 - Healthy transformer without OLTC densification trick ($\lambda = 1$, $\sigma = mean\ (std)$)

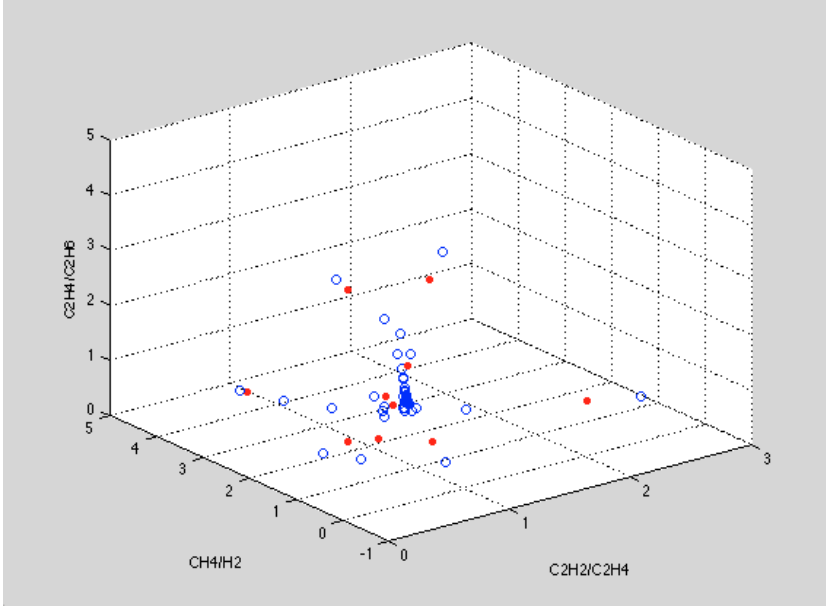


Figure B.1.6 – Healthy transformer with OLTC densification trick ($\lambda = 1$, $\sigma = mean\ (std)$)

Appendix C – Paper "Discovering structures in DGA clusters with applications in several methods for fault diagnosis"

Provas de Dissertação do MIEEC – Julho de 2012

Discovering structures in DGA clusters with applications in several methods for fault diagnosis

R. Tavares and V. Miranda, *Fellow IEEE*

***Abstract* — This paper in the form of *long abstract* presents the use the Information Theoretic Learning Mean Shift (ITLMS) algorithm and autoassociative and binary output neural networks in order discover the intrinsic data properties of the cluster used to diagnose power transformers.**
Because of the sparse data, an information theoretic learning mean shift algorithm was used in order to create virtual points to train neural networks, leaving the real data only to its validation.

***Index Terms*— Power transformers, fault diagnosis, dissolved gas analysis, autoassociative neural networks, mean shift algorithm.**

I. Introducton

The Dissolved Gas Analysis, often know has DGA, is a technic to diagnose power transformers used for several decades. It is such a powerful tool that has become the industry standard and several norms were published [1, 2].

Power transformers are a key component in any electric system. They are very expensive machines and, when there is a severe fault, the consequences can affect not only the machine itself but also surrounding facilities, equipment and people. Replacing or repairing a power transform, besides being very expensive, can also take a lot of time, which can make the consequences even worse. There are thousands of these machines in any generation, transmission and distribution system; therefore, their reliability is extremely important to maximize the energy sold and the global effectiveness and efficiency of the electric system.

Thus, any tool that can prevent a transformer to go out of service, minimize its repair cost or prevent accidents is very important and useful to utilities and transformer manufacturers.

It is known that when a fault occurs inside a power transformer, in its initial state, the consequences are very small, allowing the machine to work, and can be neglected. However, as time passes, those small faults can evolve to a more severe state that may not be reparable or may lead to the destruction of the machine. The main goal of any diagnosis system is to detect faults is their initial state and identify which type of fault occurred, if any. This allows the machine owner to analyze the situation and take preventive and corrective measures to maximize the power transformer lifetime. It is also important that this diagnosis method can be done while the machine is kept in service, because its disconnection may be very expensive and last for a long time. This latter point is related to the costs of the non-supplied energy.

In this document several diagnosis methods are presented. These methods always recognize seven healthy/faulty states and use the same dissolved gas ratios used in the IEC60599 standard.

The seven healthy/faulty states are:

Case	Faulty/Healthy State
PD	Partial Discharge
DH	High Energy Discharge
DL	Low energy Discharge
T1	Thermal Fault (T<700ºC)
T2	Thermal Fault (T>700ºC)
OK	Healthy State without OLTC
OK with OLTC	Healthy State with OLTC

II. Problem and Concept Definition

Discovering the intrinsic data properties of a cluster can be done in two different ways: using methods and algorithms that turn these properties explicit and using neural networks, where the knowledge obtained about the clusters properties is stored in the connection weights and bias and, therefore, implicit.

In order to make the cluster properties explicit, the ITLMS was used. In this algorithm several tries were made as a mean to study how the clusters were better represented. Properties like single or local modes, the principal curve or even finer structures were used as representative of the clusters.

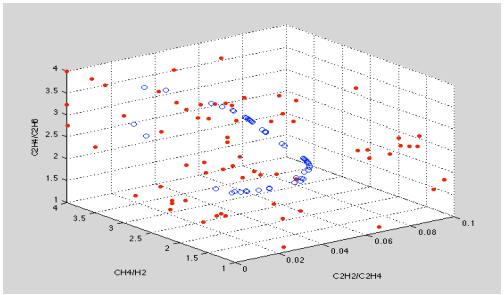


Figure 1 – Thermal fault (T>700ºC) ITMLS finer structures seeking

After getting these representatives, two different algorithms were used, the simplest one uses the representatives of each fault retrieved by the ITLMS algorithm and, when an unclassified sample is presented, the mean absolute error (MAE) between this point and all the modes is determined. The diagnose is the fault which MAE is the smaller.

The second one was a little more complex: when a new unclassified data point is presented, it is subject to a steepest descent algorithm based on the ITLMS. Therefore, the data point moves towards the properties sought of the data. After, the mean absolute error between this new point and each of mean shift points is calculated. The minimum is found and the unclassified point is diagnosed as the fault which has the minimum MAE.

In these methods the properties of the clusters are well defined and the algorithms just try to measure the similarity between a new unclassified data point and the representatives of each cluster.

Regarding the algorithms which use neural networks, and because the DGA data is sparse, the ITLMS was used to do a 'densification trick', i.e. the generation of new items of information sharing some statistical properties with the original cluster of real data in order to increase the data in the database. This virtual data was then used in the neural network training. This way the real data is used only in the validation of the neural networks, and because this way the validation data set is bigger, the results are more similar to the ones the method will retrieve when in industrial use.

Once again, two different methods using neural networks were tested. The first one used a set of seven competitive autoassociative neural networks, or just autoencoders, each one trained to recognize one of the healthy/faulty states of the power transformers. In this method, when a new, unclassified data point is shown, all the autoencoders will try to replicate it in the output, however, just one will 'resonate', i.e. have a small input-output error. This way, the diagnosis of the new unclassified data point is the healthy/faulty state that the autoencoder with the smallest input-output error represents. This diagnosis method isn't new and was firstly introduced in [3].

The second method that uses neural networks as a similar architecture, however the autoencoders have been replaced with neural networks trained to retrieve '1' when a member of the cluster that they must recognize is presented and '0' in all other cases. When a new unclassified data point is presented to this system it is diagnosed with the faulty/healthy state which neural network has the output most similar to the unitary value.

With systems architectures like these, all the networks are competing against the others when a new data point is presented. This is a great advantage because there are no unclassified samples after the diagnosis. All samples are classified, however there can be wrong classifications, while in the classic approach to this problem, where one single neural network is trained to recognize all the transformers' health states, there can be misclassifications and unclassified data after diagnosing.

III. DIAGNOSIS RESULTS

When diagnosing a database with data from the IEC TC10 database and from several other origins, with 348 real DGA data, the diagnosis method with the best results is the one with binary output neural networks. With this method 97.99% of correct diagnosis was achieved while in the autoencoders method 95,98% was obtained. However, training the binary neural networks is much slower than training autoencoders.

Regarding the methods that apply the mean shift algorithm to do the diagnosis, they were an attempt to study the potential of the mean shift algorithm as a stand-alone diagnosis method. It must be said that the results were pretty good, with the steepest descent method diagnosing 91,51% of the faulty cases with a correct diagnosis and the simplest method diagnosing 84,77% of the entire database correctly. Still, the accuracy of these methods is much smaller than the accuracy of the methods with neural networks, and the room for improvement is smaller. However, these diagnosis methods can be embedded in others more complex and improve them. The bigger disadvantage of the methods based on the ITLMS is that they are much slower and the computational force needed is a lot bigger when compared with the ones that use neural networks. Also, the results obtained by the steepest descent method when all the seven healthy/faulty stated are diagnosed drop to 71,84% of correct diagnosis.

Above, only the best results achieved are referred. However, several tries with different ITLMS parameters were made.

The results achieved, mostly with the neural networks methods, are pretty good. If one compares them with the IEC60599 standard, which obtained 93,94% of correct diagnosis [3], the new methods shows are clearly an improvement.

REFERENCES

1. IEC, *IEC-60599 - Mineral oil-impregnated electrical equipment in service - Guide to the interpretation of dissolved and free gases analysis*, 1999.
2. *IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers - Redline.* IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991) - Redline, 2009: p. 1-45.
3. Castro, A.R.G., V. Miranda, and S. Lima. *Transformer fault diagnosis based on autoassociative neural networks*. in *Intelligent System Application to Power Systems (ISAP), 2011 16th International Conference on*. 2011.