Relatório Técnico

Núcleo de Computação Eletrônica

Neuro-Fuzzy System for Diagnosis of Engines, Based on Oil Samples Analysis

Gizelle K. Vianna
Antonio G. Thomé

NCE - 12/00

Universidade Federal do Rio de Janeiro

NEURO-FUZZY SYSTEM FOR DIAGNOSIS OF ENGINES, BASED ON OIL SAMPLES ANALYSIS

Gizelle K. Vianna
AEP-NCE/IM, Universidade Federal do Rio de Janeiro
Cidade Universitária – Ilha do Fundão, Caixa Postal 2324 – CEP: 20001-970
Rio de Janeiro, RJ, BRASIL
E-mail: kupac@novell.nce.ufrj.br

and

Antonio G. Thomé
AEP-NCE/IM, Universidade Federal do Rio de Janeiro
Cidade Universitária – Ilha do Fundão, Caixa Postal 2324 – CEP: 20001-970
Rio de Janeiro, RJ, BRASIL
E-mail: thome@novell.nce.ufrj.br

1. ABSTRACT

The present paper describes a neuro-fuzzy hybrid system applied to the diagnosis of automobile engines, based on the analysis of oil samples. A relevance analysis was done to select the most significant variables among the available ones, in order to classify the samples. Such relevance analysis is described in details along the paper. Four different systems were implemented: one pure neural system, and three different neuro-fuzzy systems. A detailed description of the neural and fuzzy systems is also presented, as well as the performance obtained by each one of them.

Keywords: Artificial intelligence, Automatic recognition, Classification, Fuzzy hybrid systems, Fuzzy modeling, Functions of Membership, Neural networks.

2. INTRODUCTION

The analysis of the oil of an automobile engine can be compared to the analysis of a blood sample of a human being. If oil samples collected from an engine are periodically submitted to chemical examination, mechanical problems might be diagnosed and timely treated. In companies with large fleet of vehicles, the adoption of a routine check-up on the engine oil status may accomplish significant cost reduction, by reducing the number of breakdowns, or actually preventing engines from breaking down, as well as by enhancing the oil exchange period.

Therefore, it is a preventive service, which comprises the following steps:

1) Periodic collecting of oil samples;
2) Chemical analysis of oil samples;
3) Identification of the samples that reveal indications of problems in the laboratorial analysis, with regard to the corrosion, combustion or contamination;
4) Problem diagnosis;
5) Release of a technical report.

The goal of a automated system in this procedure is to improve the step related to the identification of samples, by segregating them in two sets: the set of oil samples collected from engines in good shape, and the oil samples that demand a more detailed analysis, to be done by an expert staff. This step represents a major bottleneck in the process, since most oil samples do not show any problems whatsoever.

The present paper describes a survey of the problem, by using a hybrid computer model, in which both neural networks and fuzzy logic technologies are associated [1, 2, 3]. A set of samples with results obtained after the conclusion of step 2 was used for the training, validation and testing of the system. Section 2 presents the description of the samples, while section 3 contains the relevance analysis of the variables, for the purpose of selecting only the most significant ones for classification. In section 4, the definition of the selection criteria of the variables used is presented. Section 5 describes the neural model developed, and section 6 brings the final hybrid model.

3. DATA SAMPLES

The data set available for the present survey achieves a total of 725 samples, each one of them containing 27 results of oil chemical analysis. Each of the chemical analysis will be called variable, and shall be represented by the symbol V_i.

Each samples receives 3 different diagnoses: one related to the corrosion, other related to combustion, and the third one related to contamination. These diagnoses will be identified by the symbols D_1, D_2 e D_3, respectively. In the system implemented, each diagnosis D_i may receive one of two values: 0 or 1. A value 0 means that the sample presents no problems, while a value 1 means that the sample indicates the presence of problems in the corresponding engine.

In the set of historical data received from the company that provides support to the present survey, diagnoses statistics presented the distribution described on *Table 1*.

It can be noted that the percentage of diagnoses equal to 1 is quite smaller than the other, what makes the set very uneven. Such uneven distribution turns significantly harder the training and the validation of the system. In order to minimize such problem, an investigation was carried out on the hypothesis of artificially generating an additional set of samples, as described further.

Table 1.Distribuition of the diagnoses in the historical set.

D_i	Total of $D_i = 1$	% of $D_i = 1$	Total of $D_i = 0$	% of $D_i = 0$
D_1	83	12.93%	642	87.07%
D_2	18	2.55%	707	97.45%
D_3	30	4.14%	695	95.86%

It was noted during the analysis that the order of magnitude of each variable is quite heterogeneous. Observing *Table 2*, it can be also noticed that the difference between maximum and minimum values of each column is very significant. In order to homogenize the interval between possible values in a set of variable, the variables were normalized by applying a linear normalization process. The Eq. (1) was used for this purpose, being attributed, to each variable, a linearly distributed value between 0 and 1:

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

Normalization was carried out before the relevance analysis of variables commenced. Only after the normalization process the data set was made available for the training of the neural networks. *Table 2* shows additional information regarding the to original distribution of variables.

Table 2. Distribution of variables in C_1.

COLUMN	MAX.	MIN.	μ	σ
V_1	542.00	11.00	97.24	48.82
V_2	31.00	0.00	3.63	3.13
V_3	127.00	0.00	11.58	9.18
V_4	74.00	1.00	6.88	6.50
V_5	27.00	0.00	4.18	2.11
V_6	43.00	1.00	8.50	4.05
V_7	43.00	1.00	8.50	4.05
V_8	5.00	0.00	0.75	0.88
V_9	45.00	4.00	10.31	3.99
V_{10}	64.00	0.00	0.23	2.51
V_{11}	400.00	1.00	168.22	108.76
V_{12}	1037.00	15.00	471.39	300.39
V_{13}	3828.00	452.00	1676.80	704.07
V_{14}	30.00	0.00	0.20	2.16
V_{15}	1429.00	268.00	894.34	225.80
V_{16}	2004.00	340.00	1134.50	292.94
V_{17}	63.00	0.00	6.84	5.41
V_{18}	1.00	0.00	0.00	0.04
V_{19}	1.00	0.00	0.00	0.04
V_{20}	24.00	0.00	4.49	3.76
V_{21}	0.38	0.01	0.08	0.04
V_{22}	0.61	0.01	0.10	0.05
V_{23}	0.43	0.01	0.15	0.05
V_{24}	2.50	0.00	0.61	0.31
V_{25}	2.00	0.10	0.11	0.10
V_{26}	0.23	0.08	0.08	0.01
V_{27}	10.00	2.00	2.07	0.62

4. RELEVANCE ANALYSIS

The process of analysis of variables was done on the diagnosis basis. That is, the entire process was repeated three times, one for each diagnosis.

The first step was to separate the samples in two sets: one set containing samples with diagnoses equal to 1, and another formed by samples with value 0. The sets will be identified as G_1 and G_0, respectively.

The next step was to analyze each variable distribution, in each one of the sets 1and 0 [4, 5]. An amount of 27 graphics were generated, showing the distribution of values of each variable V_i within each set G_i. Two of these graphics are shown in figures 1 and 2. The distribution curve related to G_1 is exposed in a full line, while the distribution curve related to G_0. is in a dotted line.

Further on, the graphics were analyzed in order to check the discrimination capacity of each variable, with regard to diagnoses 0 and 1.

5. SELECTION CRITERIA

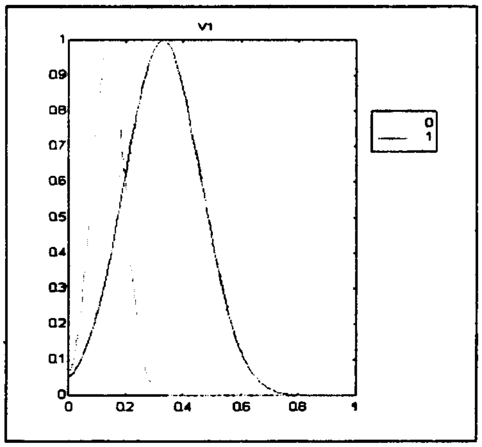

Figure 1. Distribution of variable V_1.

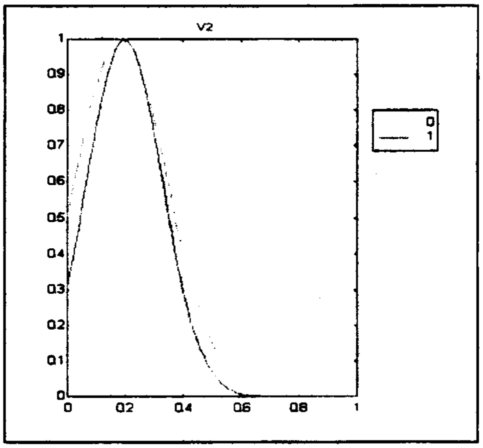

Figure 2. Distribution of variable V_2.

Figures 1 and 2 show the Gaussian curves that represent the distribution of variables V_1 and V_2 with respect to diagnoses 1 (continuous line) and to diagnoses 0 (dotted line). In figure 1, the intersection between the curves is relatively small, mainly in comparison with the intersection in figure 2, which is much bigger. A greater separation of the curves suggests a major discrimination capacity. The chosen variables were those that presented smaller intersections between the distribution curves.

After concluding the analysis previously described, the variables listed on *table 3* were pre-selected for the training of the neural networks.

After pre-selection, a verification of the cross correlation between the selected variables was performed, and the variances of each variable were compared.

The analysis of these correlations aims the identification of identifying those variables that present little additional information in comparison with the others. For this purpose, couples of variables with high correlation must be identified in the set of pre-selected variables. The selected variable is the one that presents the highest variance, the other being discarded without significant loss of information. At the end of this analysis, a smaller set of variables was sorted out, as listed on *table 4*.

Table 3. Variables selected according to the distribution analysis.

D_1	D_2	D_3
Selected	Selected	Selected
V_1	V_1	V_1
V_2	V_2	V_2
V_3	V_3	V_3
V_4	V_4	V_4
V_5	V_5	V_5
V_6	V_6	V_6
V_9	V_9	V_9
V_{11}	V_{14}	V_{14}
V_{12}	V_{17}	V_{17}
V_{13}	V_{21}	V_{18}
V_{17}	V_{22}	V_{25}
V_{18}	V_{23}	V_{27}
V_{23}	V_{24}	
V_{24}	V_{27}	
V_{27}		

Table 4. Variables selected according to the correlation and variance analysis

D_1	D_2	D_3
Selected	Selected	Selected
V_1	V_1	V_1
V_4	V_4	V_4
V_{12}	V_5	V_5
V_{13}	V_{14}	V_{14}
V_{18}	V_{21}	V_{18}
V_{24}	V_{24}	V_{25}
V_{27}	V_{27}	V_{27}

6. THE NEURAL MODEL

First Experiments

The set of samples was divided in three parts: the first part contained the samples used for the networks training sections, the second part held the samples used for validation, during training, and the third for the performance evaluation.

Initial experiments demonstrated that neural networks did not achieve a satisfactory performance during its training. The reason for such behavior was identified as a consequence of scarce examples with diagnoses equal to 1. In order to cope with this limitation, some value 1 samples were generated at random, with the same statistical characteristics, moments of 1st and 2nd order, of the set of available samples.

The operation did not alter the distributions at each level and, as a consequence, previously selected variables remain valid. Eq. (2) was used in the creation of artificial samples:

$$x_{new} = x_{ori} + rand \times v(x_{ori}) \qquad (2)$$

where:

x_{ori} is a sample pertaining to the original set,

x_{new} is a new sample

$v(x_{ori})$ is the variance of the original set of samples

With the application of Eq. (2), the number of examples for the training and validation processes was increased, in each diagnosis. The initial and final quantities obtained after this operation are described on the table 5.

Table 5. initial and final quantities of diagnoses 1 used in the network training. T= training set e V= validation set.

	D_1		D_2		D_3	
	Initial	Final	Initial	Final	Initial	Final
T	41	123	9	153	15	135
V	17	121	4	131	6	127

Description of the neural model

In the approach to the problem, three feedforward neural networks were used. The networks were trained according to the Backpropagation method, using one network for each type of diagnosis. Each network has two hidden layers of neurons, though the number of neurons in each layer varies according to the diagnosis that is being done. The training parameters were the same in all three networks.

Not every available sample was used in the training process, as the percentage difference between diagnoses equal to 1 and equal to 0 would lead once more to an insufficient performance of the system. The quantity of samples used in each diagnosis is describe on the table below:

Table 6. Distribution of diagnoses within the training, validation and testing sets.

D_i	TRAINING		VALID.		TESTING		TOTAL	
	0	1	0	1	0	1	0	1
1	230	123	138	121	92	17	460	165
2	230	153	138	131	92	4	460	162
3	230	135	138	127	92	6	460	150

Results achieved

Performance tests were carried out separately for each diagnosis, by using the testing set described on *table 6*. 182 different network architectures were analyzed, and 50 training epochs were performed with each one of them. To each architecture, an average of performances was calculated, and the best architectures were therefore selected.

The performance of the best network for each diagnosis is demonstrated in details on tables *7* to *9*. On these tables, recognitions correspond to diagnoses 1, or 0, which were correctly classified by the corresponding neural network. Errors are diagnoses 1 that were classified as 0 by the neural network, or diagnoses 0 that were classified as 1.

For each model, thresholds were defined for the diagnoses. The choice of the thresholds was done with basis on the distribution of the output provided by the networks, when presented with the validation sets of each diagnosis (*figures 3* to *5*). Each curve corresponds to the distribution of the output generated by the network for the types 0 and 1 diagnoses.

Based on the distribution analysis, each threshold was defined as being a point inside the intersection area between the distribution curves that would best select the sets. Several simulations were done for each diagnosis, so as to find out the best thresholds. Simulations were always restricted to the validation set.

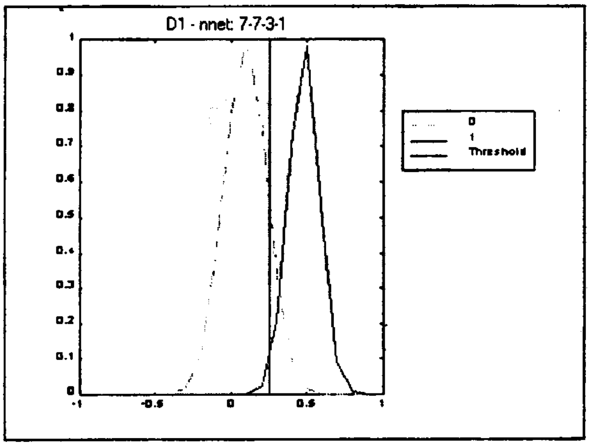

Figure 3. Distribution curves of output generated by the neural network, in the diagnosis D_1. The vertical line indicates the threshold position.

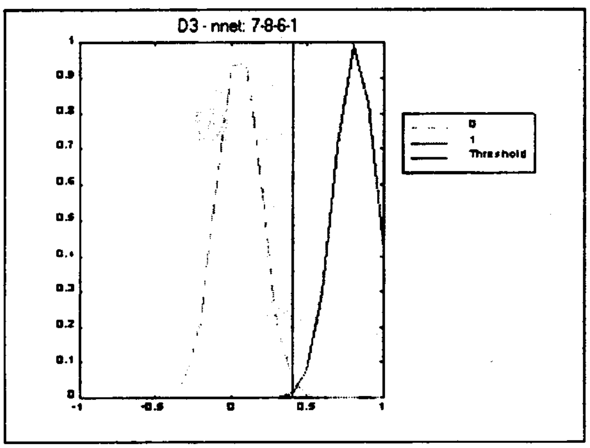

Figure 4. Distribution curves of the output generated by the neural network, in the diagnosis D_2, with threshold indicated by the vertical line.

For the first diagnosis (D_1), the experiment carried out achieved a global performance of 86.45% (table 7). The neural network that achieved the best performance presents configuration 7-7-3-1 and was trained on an average of 430 epochs.

The threshold used was 0.25. For processing elements, a linear propagation function was used, in every layer. Experiments with the same network were also carried out, using, however, the logarithmic linear propagation function. The performance was quite inferior.

For the second diagnosis, the experiment carried out obtained a global performance of 96.94% (table 8). The neural network that showed the best performance holds configuration 7-8-6-1, and was trained in an average of 300 epochs. The threshold used was 0.4. Once again training was done by using a linear activation function.

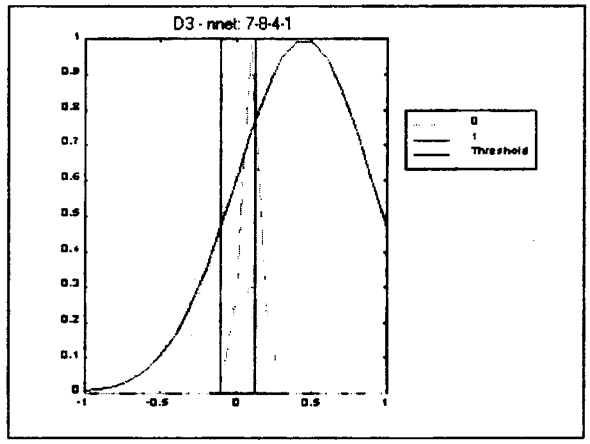

Figure 5. Distribution curves of the output generated by the neural network, in the diagnosis D_3, with threshold indicated by the vertical line.

Table 7. Performance of the network responsible for the analysis of the diagnosis 1.

TARGET	% RECOGN.	% ERRORS	TOTAL
0	85.62%	14.38%	138
[illegible]	91.04%	8.96%	25

Table 8. Performance of the network responsible for the analysis of the diagnosis 2.

TARGET	% RECOGN.	% ERRORS	TOTAL
0	97.52%	2.48%	138
[illegible]	80.80%	19.20%	5

Table 9. Performance of the network responsible for the analysis of the diagnosis 3.

TARGET	% RECOGN.	% ERRORS	TOTAL
0	82.06%	17.94%	138
[illegible]	89.11%	10.89%	9

In the third diagnosis, the experiment carried out achieved a global performance of 82.49% (table 9). The neural network that achieved the best performance held configuration 7-8-4-1, with the average of 240 training epochs. The threshold used was the interval [-0,1; 0,13]. Training was also done by using a linear activation function.

7. THE HYBRID MODEL

Description of the Fuzzy Model

Afterwards, a fuzzy model addition was considered to the neural model previously described, in order to investigate the possibility of improving the performance of the system.

The fuzzy model built is quite simple, and is based on the results provided by the neural networks. In order to build the fuzzy model, only the validation sets were used, as described on *table 6*.

Each validation set was presented to the corresponding neural network. The output generated by each network, for each sample presented, were grouped in two subsets: the first set contained the output of the network when the corresponding diagnoses were equal to 1, and the second contained the output when the diagnosis were equal to 0. The subsets will be called S_1 e S_0, respectively.

From each subset S_i, the maximum, minimum, mean, and standard deviation values were computed. One single fuzzy variable was defined, *Diagnosis*, which has only two functions of membership: OK and not-OK.

Three different formats for the functions of membership of the variable *Diagnosis* [1, 6] were tested. The three different hybrid systems, defined by the three formats of functions of membership are referred to as A, B and C systems, respectively.

All three hybrid systems used two trapezoid functions. In system A, the *OK* function is defined by the points [0, 0, min0 and max0], where m0 and max0 are the mean and the maximum values of the subset S_0, respectively. The *not-OK* function is defined by the points [min0, max0, 1, 1], where min1 and m1 are the minimum and mean values of the subset S_1, respectively.

Table 10. Membership function definition, for each system.

OK MEMBERSHIP FUNCTION				
SYS	Left-Down Point	Left-Top Point	Rigth-Top Point	Right-Down Point
A	0	0	min0	max0
B	0	0	m0	max0
C	0	0	m0	m0+2*std0
NOT-OK MEMBERSHIP FUNCTION				
SYS	Left-Down Point	Left-Top Point	Rigth-Top Point	Right-Down Point
A	min0	max0	1	1
B	min1	m1	1	1
C	m1-2*std1	m1	1	1

In system B, the *OK* e *not-OK* functions are defined by the sets of points [0, 0, m0, max0] and [min1, m1, 1, 1], respectively, and in system C, these functions were defined as: [0, 0, m0, m0+2*standard_deviation0] and [m1-2* standard_deviation1, m1, 1, 1], respectively. Figures *6* to *8* show the graphics corresponding to each one of the types of functions of membership used for diagnosis D_1, while Table *10* shows their definition points. The performance presented by the hybrid system is, in each case, described on *tables 10* to *12*.

Hybrid System Functioning

Each sample in the data set is presented to each one of the three neural networks, corresponding to each diagnosis that should be performed. As the function of the hybrid system is identical for all three diagnoses, only one of them will be shown, since the other will be the same.

The neural network that performs diagnosis D_i, after receiving the input of a sample, generates a value between 0 and 1 as output. In the pure neural model, this value is compared to a threshold and the sample is diagnosed as 1 or 0, as the case may be.

In the hybrid model, the network output is used as input for the fuzzy system. The output value of an sample, given by a neural nerwort, is then considered as the value of the fuzzy variable *Diagnosis*, and the degrees of membership in each one of the fuzzy functions of the variable are evaluated thereat. The fuzzy set that achieves the highest degree of membership, given by the corresponding function, will be the winner, and the sample will be allocated in the corresponding class.

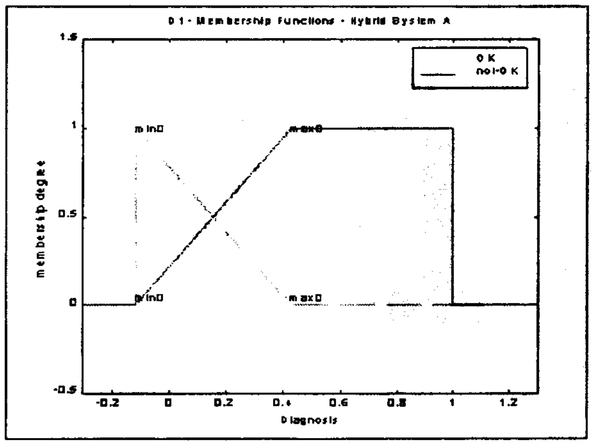

Figure 6. Hybrid system A - functions of membership of the fuzzy variable *Diagnosis*, for diagnosis D_1.

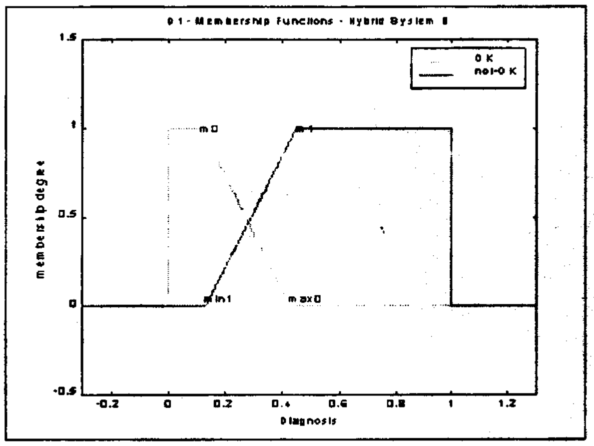

Figure 7. Hybrid system B – functions of membership of the fuzzy variable *Diagnosis*, for diagnosis D_1.

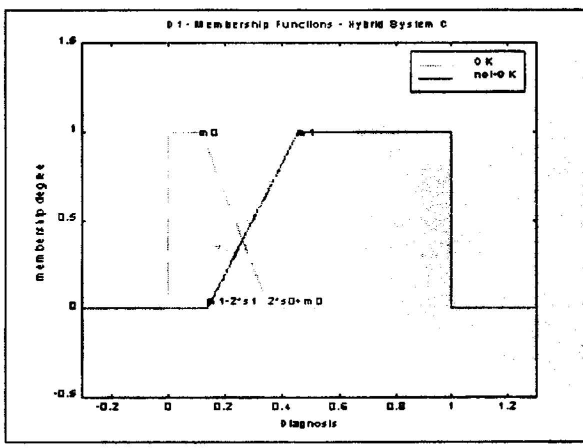

Figura 8. Hybrid system C - functions of membership of the fuzzy variable *Diagnosis*, for diagnosis D_1. Where: m_i = mean of C_i and s_i= standard deviation of C_i.

Results achieved

Tables 11 to *13* compare the performance of the neural system with the performance of the hybrid system, in each diagnosis. The performance of the neural and hybrid systems is also shown, when the systems are trained with two different sets of variables. The first set was chosen according to the statistic procedure previously described in *sections 3* and *4*. The second set is the same used by the experts of the company that uses the oil analysis. On tables *11* to *13*, these sets of variables are described in column *Var*, and are identified as E_1 and E_2, respectively.

Table 11. Comparison of performance of the different systems with respect to diagnosis D_1.

	Data set C_1 – Diagnosis I	
Var	Neural System	
	% Recognition 0	[redacted]
E_1	85.62%	91.04%
E_2	89.87%	91.04%
	Hybrid System A	
	% Recognition 0	[redacted]
E_1	57,13%	100%
E_2	53,26%	100%
	Hybrid System B	
	% Recognition 0	[redacted]
E_1	65.22%	100%
E_2	83.70%	100%
	Hybrid System C	
	% Recognition 0	[redacted]
E_1	57,13%	100%
E_2	71,46%	100%

Table 12. Comparison of the performances of the different systems with respect to diagnosis D_2.

	Data set C_1 – Diagnosis II	
Var	Neural System	
	% Recognition 0	[redacted]
E_1	97.52%	80.80%
E_2	96.99%	79.20%
	Hybrid System A	
	% Recognition 0	[redacted]
E_1	32.35%	75%
E_2	35,43%	100%
	Hybrid System B	
	% Recognition 0	[redacted]
E_1	86.96%	75%
E_2	88.04%	100%
	Hybrid System C	
	% Recognition 0	[redacted]
E_1	53,30%	75%
E_2	56,17%	100%

Comparison of performances

The main objective of the automated system is to eliminate the maximum of samples that do not indicate the existence of problems, therefore reducing to a minimum, ideally zero, the exclusion of samples that do indicate the existence of problems, which are false negative diagnoses.

The performance varies according to the system (hybrid or neural) in use. It can also be noted that, in most cases, the difference between the performance of the pure neural system, trained by set E_2, is reasonably small. This result confirms the validity of the method of variables selection applied in the experiment, as previously described.

8. CONCLUSIONS

In general terms, a conclusion that can be drawn is that the role of the hybrid is to perform a correction of the output provided by the neural network. Initially, the classification of a sample is done with basis on a predetermined threshold. When a fuzzy model is added, some samples previously classified as type 0 are classified as type 1. Such correction allows samples that were wrongfully classified by the neural network to be correctly identified in the system.

Table 13. Comparison of the performances of the different systems with respect to diagnosis D_3.

	Data set C_1 – Diagnosis I	
Var	Neural System	
	% Recognition 0	[redacted]
E_1	82.06%	89,11%
E_2	44.30%	65,33%
	Hybrid System A	
	% Recognition 0	[redacted]
E_1	13.96%	100%
E_2	0.13%	16,67%
	Hybrid System B	
	% Recognition 0	[redacted]
E_1	81.52%	100%
E_2	91.30%	16,67%
	Hybrid System C	
	% Recognition 0	[redacted]
E_1	83%	100%
E_2	65.22%	33,33%

Taking into consideration the fact that all three diagnoses present intrinsically different characteristics, the functions of membership that convey the best performance are not necessarily of the same type. In every diagnosis, the hybrid systems behave differently. Thus, for each one of the diagnoses, the most adequate hybrid system must be chosen.

Difficulties and future implementations

The set of samples is not large enough, and contains few samples of the sets G_1, the diagnoses of which are equal to 1. For this reason, there was the need to create new artificial samples.

For future implementations, it is expected that a more significant set of samples will be made available, portraying the problem with richer details and information.

9. REFERENCES

[1] Cox, E. (1998). "The Fuzzy Systems Handbook – 2nd Edition", AP Professional, pp. 1-165.

[2] Haykin, S. (1999), "Neural Networks: a comprehensive Foundation – 2nd Edition", Prentice-Hall, *Inc.*, pp. 156-252.

[3] Hines, J. W. (1997). "Fuzzy and Neural Approaches in Engineering", John Wiley & Sons, Inc., pp. 6-52.

[4] Spiegel, M. R. (1978). "Probabilidade e Estatística - Coleção Schaum", McGraw-Hill do Brasil, pp. 435-449.

[5] Wonnacott, R. J, Wonnacott, T. H. (1985). "Fundamentos de Estatística", LTC Editora S.A., pp. 234-267.

[6] Cruz, A. (1998). "Apostila sobre Lógica Nebulosa", Curso de Inteligência Computacional, Mestrado em Informática, NCE/DCC-IM UFRJ.