12th IMEKO TC10 Workshop on Technical Diagnostics
New Perspectives in Measurements, Tools and Techniques for Industrial Applications
June 6-7, 2013, Florence, Italy

Data set processing for the optimization of the artificial intelligence-based diagnostic methods

Piotr Bilski[1,2]

[1] *Institute of Radioelectronics, Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, +48 22 234-7479, pbilski@ire.pw.edu.pl*
[2] *Department of Applied Informatics, Warsaw University of Life Sciences, ul. Nowoursynowska 159, 02-776, piotr_bilski@sggw.pl*

***Abstract*-**The paper presents the application of selected statistical methods to process the training and testing data sets and prepare them for the artificial intelligence-based method used in the diagnostics of analog systems. The size of the set (especially the number of processed attributes – stamps) determines the efficiency of the selected algorithm and minimizes the amount of information to be measured in the actual system. The preprocessing operations include elimination of constant or quasi-stationary stamps and selection of their most important set, allowing for the efficient fault detection or parameter identification. The paper presents two methods from the econometrics domain adjusted to the technical diagnostics applications. Their implementation is tested on the electronic analog filter. Also, efficiency of the artificial neural network (ANN) working with the original and preprocessed data is verified.

I. Introduction

The modern diagnostic methods often use Artificial Intelligence (AI) approaches. Their advantages are flexibility, high efficiency and autonomy. Because of the advancement in the computer technologies it is now possible to implement even sophisticated fault detection and identification methods in the microprocessor systems (such as microcontrollers or programmable logical controllers). As AI-based methods (such as ANN or support vector machines) rely on data sets, containing information about the System Under Test (SUT) behaviour, they must be properly prepared. This includes the following features:

- the set of learning examples should cover as many cases of the SUT states, as possible (including nominal and faulty situations). This allows for the generalization of the AI method;
- the number of examples should be small, but large enough to cover most faulty cases. The minimal size of the set helps avoiding overlearning by the diagnostic algorithm;
- the set of stamps (characteristic information measured in the SUT responses) should be also minimized. This is important for the AI method's efficiency and for the economy of the SUT measurement, minimizing the number of nodes to be analysed.

The preparation of the training data sets is the crucial part of the automated diagnostic method design. It usually requires a deep knowledge of the SUT work regime (to select the most representative examples). Selection of the analysis domain, excitation signals and set of stamps is a difficult and time-consuming process. Therefore the designer usually gathers all the available information, further relying on the AI algorithm. During the training, the most important stamps can be selected to ensure the maximum efficiency of the classification or regression task, used in the fault detection and identification or the parameter identification of the SUT. To minimize the knowledge structure stored by the AI method, it is reasonable to prepare data sets by eliminating redundant data and selecting the most important stamps to distinguish between various SUT states.

There are multiple approaches to select the most important stamps in the data set. They were used in many applications of data analysis problems, such as econometrics or bioinformatics. The most widely used is the Principal Component Analysis (PCA) approach [1], which was applied to the analysis of transformers or gas turbines. The PCA transforms original stamps into the "virtual" ones. Although they are simpler to analyse, their physical interpretation is difficult. The rough sets approach [2] during the knowledge extraction from the data set creates reducts, i.e. minimal groups of attributes allowing for the complete distinguishing between all fault categories. They are calculated on the discretised data, therefore the proper discretization method must be applied. The disadvantage of reduction is its high computational complexity, making the exact approaches impractical. Instead, approximate solutions (using genetic or Johnson's algorithms [3]) are found.

The alternative approaches are statistical methods used in econometrics for building models of the company [4]. They exploit the correlation matrix, which is already used to find stamps dependent on each other. Analysis of such a matrix allowed for optimizing the ANN operation in regression task [5]. This approach considers only the relation between the observed stamps, disregarding their relation with the diagnosed SUT parameters.
The aim of the paper is to present the statistical data processing approaches to prepare the data sets for the AI-based diagnostic algorithm. They are Hellwig method and multiple correlation coefficients [6], searching for the set of attributes containing the greatest amount of information about the actual SUT's state. The minimization of the analysed stamps should decrease the number of stamps and simplify the AI method's structure.
The paper is organized as follows. First, the data set structure used in the experiments is presented. In section III the exemplary SUT, i.e. the fifth order filter is introduced. Section IV covers the implemented methods. In section V results of the data analysis are discussed. In section VI conclusions and future prospects are included.

II. Data set description

The content of the data set is crucial in training the AI-based diagnostic module. The aim of the intelligent algorithm is to find dependencies between the stamps (extracted from the SUT response signals) and the actual value of the SUT parameter, influencing its behaviour. Therefore examples and types of information they contain must be carefully selected. The presented approach uses the supervised learning methodology, where the actual SUT state is known to the designer and can be used to label each example. The data set D [2] contains n examples, each with m stamps (1). The latter are extracted from the SUT's model simulation and are the source of knowledge about its behaviour. The form of stamps depends on the analysis domain (time, frequency, DC, etc.) and the diagnosed system. For instance, in the time domain the stamp can be the maximal value of the response signal or the time instant of zero-crossing. Each example is supplemented with the information about the SUT's state. Usually it is the number and the actual value of the faulty parameter. In the fault-free state (which must also be present to make the identification of the nominal SUT possible) this value is "0". In the presented research the additional columns c_k contain the number and the real value of the faulty parameter. Each example considers a single parametric fault (only one source of the incorrect SUT behaviour is assumed).

$$D = \begin{bmatrix} s_{11} & \cdots & s_{1m} & c_1 \\ \vdots & \ddots & \vdots & \vdots \\ s_{n1} & \cdots & s_{nm} & c_k \end{bmatrix} \tag{1}$$

III. System UnderTest

The fifth-order analog filter is a good example to implement the data processing methods. It is complex (because of the large number of nodes and SUT elements) and difficult to diagnose. Therefore multiple stamps are needed to allow for identification of all elements (resistances and capacitances). The circuit in Fig. 1 contains ten analysed elements of the following nominal values: $R_1=R_2=R_3=R_4=R_5=1k\Omega$, $C_1=16nF$, $C_2=19nF$, $C_3=13nF$, $C_4=51nF$ and $C_5=49nF$. The cutoff frequency of the filter for such values is 10kHz. The model of the filter was implemented in the Simulink environment. Simulations were performed to obtain examples of the SUT behaviour for different values of elements (up to 90 percent of the nominal value). The excitation signal provided at node No. 1 was a sinusoid with 9kHz frequency (i.e. close to the cutoff frequency). The filter was analysed in the time domain. The constructed data sets varied in the number of examples (from 70 to 180). The number of stamps was 54, including first three maximal and minimal values of the output sinusoids with their time instants and time instants of zero crossings, measured at nodes 2, 3, 5, 6, 8 and 9.

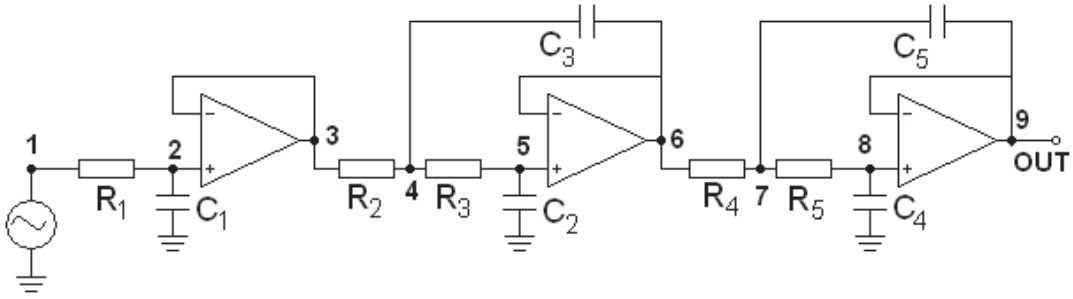

Figure 1. Scheme of the 5th order lowpass filter

IV. Description of data processing methods

The presented approaches are well known in the econometrics field. They are used to determine, which variables in the set of linear equations (describing the analysed object) are not important for explaining the output variable. Here they will be used to determine, which attributes in the set are the most important for describing the change of the faulty SUT parameter. The data processing is executed in two stages, presented below.

A. Determining the quasi-stationary stamps

The constant or almost-constant stamps have values in all examples close to each other. As the data set is prepared to cover multiple SUT states, they are not important to the fault identification and can be eliminated, decreasing the number of features to be covered by the AI method. To detect such stamps, their variation coefficient [4] is calculated (2).

$$\upsilon_i = \frac{\sigma_{si}}{\hat{s}_i} \tag{2}$$

Here $\hat{s}_i$ is the arithmetic mean value of the i–th stamp in the set D and σ_{si} is its standard deviation. Columns having the variation coefficient smaller than the predefined threshold θ (for instance, *0.12*), are eliminated from the set as not important for describing the SUT state. The proper selection of the value of θ is the task for the designer and requires experience or multiple trials.

B. Finding the most important sets of stamps

After eliminating quasi-stationary stamps, the p remaining ones form the simplified data set *D'*, for which the optimal set of features, i.e. bearing the maximum information about the SUT state, is calculated. The selection is made from all possible combinations of stamps, which number is 2^p*-1*. For instance, three stamps form seven combinations, such as $\{s_3\}$, $\{s_1, s_2\}$, $\{s_1, s_2, s_3\}$, etc. Two approaches were used here. The Hellwig's method measures the information capacity for each combination of stamps from *D*'. The second approach uses the multiple correlation coefficient. It selects the combination of stamps correlated strongest with the faulty parameter. Both approaches identify only one variable, while the c_k part in *D* and *D'* considers various parameters for the identification. Therefore the sets must be divided into subsets depending on the identified SUT parameter. Each of them is then processed separately (so the correlation between stamps and the source of the fault is found exclusively for each parameter). Both methods use data structures (3) and (4).

$$R_0 = \left[r_1,\ldots,r_p\right]^T \tag{3}$$

where R_0 is the vector of the linear correlation (Pearson) coefficients between each of p stamps and the value of the diagnosed parameter. The second structure is the correlation matrix between pairs of stamps:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & 1 & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{4}$$

The Hellwig method for the q-th stamp combination calculates the information capacity H_q (5):

$$H_q = \sum_{i \in N} h_{si}; \quad h_{si} = \frac{r_i^2}{\sum_{i,j \in N} \left|r_{ij}\right|} \tag{5}$$

where h_{si} is the partial information capacity of the i-th stamp from the q-th combination and N is the vector of stamps' numbers. The combination with the greatest value of H_q is the one describing the faulty parameter best.

$$c_s^* : H_q^* = \max\left\{H_q : q = 1,2,\ldots,2^p - 1\right\} \tag{6}$$

On the other hand, the multiple correlation coefficients K_q are calculated for the q-th combination of stamps:

$$K_q = \sqrt{1 - \frac{\det(W_q)}{\det(R_q)}} \tag{7}$$

where $\det(R_q)$ is the determinant of the correlation matrix (4) for the stamps existing in the q-th combination and $\det(W_q)$ is the determinant of the matrix constructed of the correlation vector (3) and the correlation matrix (4):

$$W_q = \begin{bmatrix} 1 & R_{0q}^T \\ R_{0q} & R_q \end{bmatrix} \tag{8}$$

The combination with the greatest value of K_q is selected for training the AI diagnostic method. Both presented methods were implemented in the Matlab environment.

V. Experimental results

This section presents the experimental results. First, results of data processing using both methods are discussed. Then, efficiency of the artificial neural networks learning from different data sets is compared.

A. Reducing the number of stamps

The data sets of the filter from section III were processed using both presented approaches. In the first stage the quasi-stationary stamps were identified. For $\theta=0.12$ only nine stamps remained in the data set (from signals measured at all nodes except 2, which proves analysis here is not necessary). Note that the elimination of quasi-stationary stamps is required, otherwise 2^54 combinations of stamps must be created, which is out of range of most computing toolboxes. For each SUT parameter both stamp selection methods were applied separately. In Table 1 exemplary results for the multiple correlation coefficient approach are present. Values "1" in the row describe stamps existing in the combination referring the particular parameter. For instance, the element R_1 is best described by stamps number *20*, *29*, *33*, *38*, *42*, *47* and *51*.

Table 1. Results of the stamp analysis using the multiple correlation coefficient

Stamp No. / Parameter	11	15	20	29	33	38	42	47	51
R_1	0	0	1	1	1	1	1	1	1
R_2	1	0	0	1	1	1	1	1	1
R_3	0	0	1	1	0	1	1	1	1
R_4	0	1	0	1	1	1	1	1	1
R_5	1	0	1	1	1	1	1	1	1
C_1	0	0	0	1	1	1	1	1	1
C_2	0	1	0	1	1	1	1	1	1
C_3	0	1	1	1	0	1	1	1	1
C_4	0	0	1	1	1	0	1	1	1
C_5	1	0	1	1	1	1	1	1	1

All nine stamps are needed to correctly identify values of the parameters, although stamps *11* and *15* are used only in three cases. Therefore it is possible to eliminate them and make the set smaller (although this step decreases the accuracy of the AI method). The Hellwig method for the same set of stamps generated smaller optimal group, containing only five features: *11, 20, 38, 42* and *51*. The rest of stamps did not bear the maximal information for any parameter, therefore they were disregarded.

B. Testing the artificial neural network for the reduced data sets

Based on the reduced stamps the AI method was implemented for the classification task. The one-directional artificial neural network was selected for the task, being the popular method used for regression and classification tasks. The diagnostic procedure focused on determining the discrete information about the number of the faulty parameter and its intensity. These data were obtained from the original columns c_k by discretising the actual value of the parameter. The discretization rule involved the nominal c_{nom} and actual c_{act} value of the faulty parameter in the example. If the relative difference is greater than ten per cent of the nominal value (which is the assumed tolerance margin for the element) in the positive direction, it is assigned the "large" fault code. If

this difference is greater than the per cent of the nominal value in the negative direction, it is assigned the "small" fault code. Otherwise, the "nominal" fault code is set ("0") For instance, if the first parameter has the values $c_{nom}=1k\Omega$ and $c_{act}=1.3k\Omega$, the fault code would be "11". If $c_{act}=0.8k\Omega$, the code is "-11". For each of ten parameters, three fault codes were assigned (two for faults and one common for the nominal state).
Such a modified data sets were used to train and test the ANN-based diagnostic module. The task was to check the optimal structure of the network for each set, compare efficiency of modules and training duration. The standard one-directional classification network was used. Each fault code was assigned one configuration of output neurons. Although more complex coding of the category is possible, in this case the "*1 of k*" rule was used, where *k* is the number of outputs (here, 21). Considering results from section V.A, three module configurations were tested (Fig. 2), differing in the number of inputs and hidden layers. Among available training methods, Bayesian regulation combined with Levenberg-Marquardt optimization was selected. Although the most time-consuming, it ensures the highest fault identification outcomes. The ANN was simulated using the Matlab toolbox.

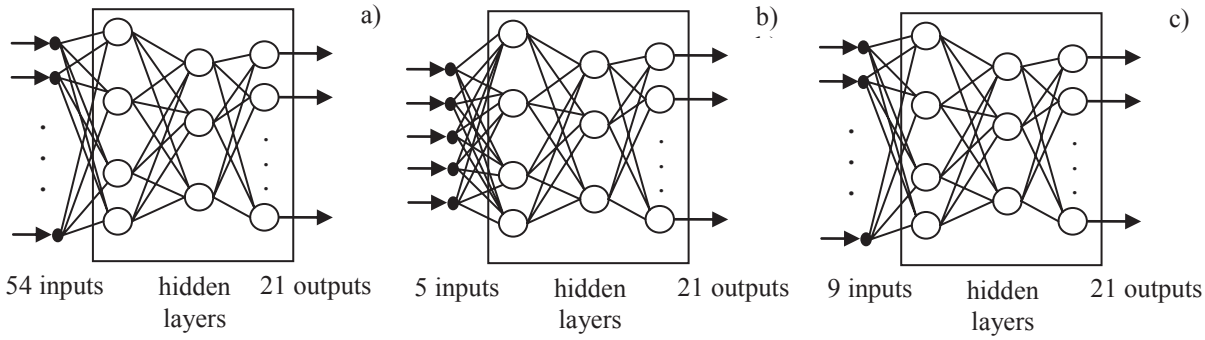

Figure 2. Configurations of ANN diagnostic modules.

Diagnostic outcomes for various configurations of the ANN are in Table 2. For each number of stamps three factors are presented: the optimal number of neurons in the network configuration (respectively, for one and two hidden layers), the percentage of correctly classified examples and the Mean Square Error (MSE). The second parameter measures the training accuracy, while the third one shows the generalization ability of the module. The number of neurons in both layers was changed up to 50. Simulating the module for each configuration was time consuming, causing the main problem during the experiments. The dependency between the obtained MSE and the number of neurons in the hidden layer is in Fig. 3. In most cases two hidden layers are better than one layer, although the exact number of neurons depends on the structure of the data set. Five stamps selected by the Hellwig method are not enough to maintain the original diagnostic quality. The multiple correlation coefficient-based set of stamps is better processed by ANN, leading to the similar results as for the original set. The worse value of MSE for the testing than the training data set are caused by small number of examples to learn from about the SUT states (no more than three for each category except the nominal one) and the high complexity of the object. This also explains relatively low diagnostic efficiency of the applied ANN modules.

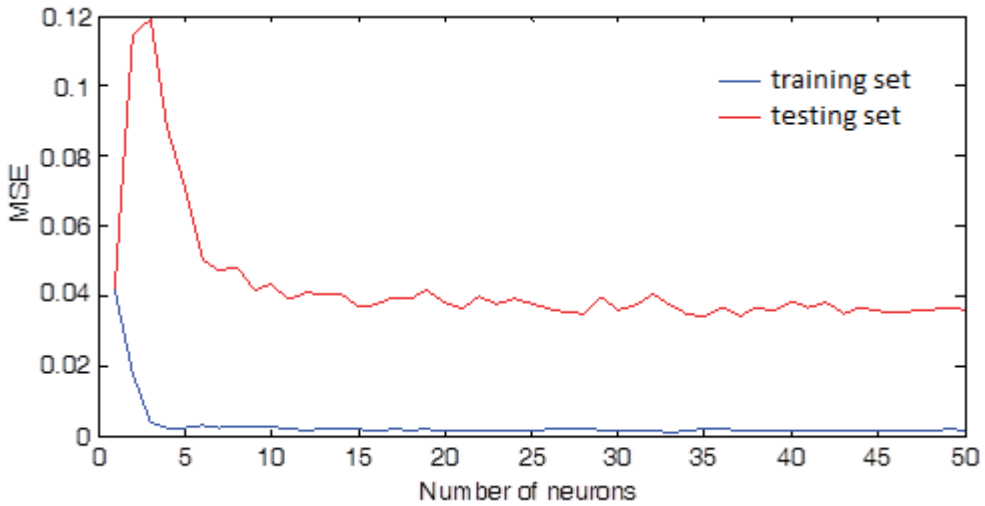

Figure 3. Mean Square Error for ANN with one hidden layer, depending on the number of neurons.

Table 2. Quality of the tested ANN configurations

	5 stamps			9 stamps			54 stamps		
Layers	ANN conf.	Quality [%]	MSE	ANN conf.	Quality [%]	MSE	ANN conf.	Quality [%]	MSE
1	[16]	48.5	2.61E-14	[40]	62.85	2.67E-14	[30]	68.5	6.6e-15
2	[15:22]	55.7	3.78E-14	[50:31]	71.42	1.97E-14	[60:44]	65.7	1.7E-14

The duration of computations depends on the complexity of the network (Fig. 4). Here results of simulating six network configurations are presented. Each simulation consisted in training and testing the network with the number of neurons in the layer (indicated as "x") changed from 1 to 50. Reduction of the number of stamps greatly accelerates computations, although the diagnostic quality is still the most important factor.

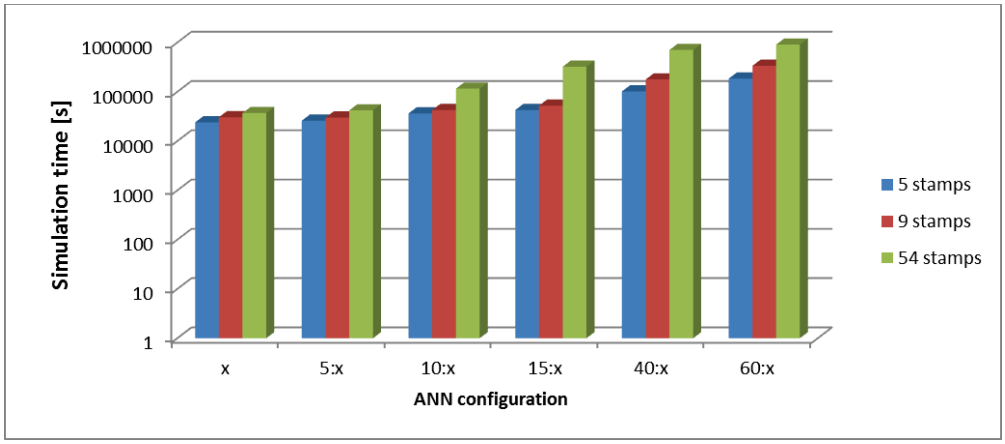

Figure 4. Comparison of simulation duration for ANN diagnostic modules

VI. Conclusions

The presented methods can be used to optimize the content of the learning data sets for the intelligent diagnostic algorithms. Elimination of quasi-stationary stamps is the more important operation, influencing the subsequent stage. The difficulty in applying these methods is the need to properly select parameters such as the threshold θ. Comparison between approaches selecting the optimal set of stamps shows that the Hellwig method leaves smaller number of stamps, which degrades the diagnostic efficiency, although decreases the simulation duration. Multiple correlation coefficients tend to use almost all available stamps, therefore requiring their greater number. Further research must be aimed at verifying the influence of the data set size to the diagnostic quality and testing other objects to find the optimal network configuration and threshold θ for the stamp reduction procedure.

This work is supported by the Polish National Centre of Science, grant No. 2011/03/D/ST8/04309.

References

[1] J. Mina, C. Verde, "Fault detection using dynamic principal component analysis by average estimation", *2nd International Conference on Electrical and Electronics Engineering*, 7-9 Sept. 2005, pp. 374-377, 2005.
[2] P. Bilski, J. Wojciechowski, "Rough-sets-based reduction for analog systems diagnostics," *IEEE Transactions on Instrumentation and Measurement*, Vol. 60, Issue 3, pp. 880-890, 2011.
[3] F. Li, T. Guan, X. Zhang, X. Zhu, "An Aggressive Feature Selection Method based on Rough Set Theory," Second International Conference on Innovative Computing, Information and Control, 2007. ICICIC '07, 5-7 Sept. 2007, pp. 176-179.
[4] F. Macho, F. Javier, "Wavelet multiple correlation and cross-correlation: A multiscale analysis of euro zone stock markets," http://ideas.repec.org/p/ehu/biltok/201104.html, 2011.
[5] Z Gou, C Fyfe, "A canonical correlation neural network for multicollinearity and functional data," Neural Networks, 17, 2004, pp. 285-293.
[6] R. A. Ashley, "*Fundamentals of Applied Econometrics*," Wiley, 2012.