Measurement 54 (2014) 180–190

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Data set preprocessing methods for the artificial intelligence-based diagnostic module

Piotr Bilski *

Institute of Radioelectronics, Warsaw University of Technology, Ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
Department of Applied Informatics, Warsaw University of Life Sciences, Ul. Nowoursynowska 159, 02-776 Warsaw, Poland

ARTICLE INFO

Article history:
Available online 1 April 2014

Keywords:
Analog systems diagnostics
Artificial intelligence
Data processing
Statistical methods

ABSTRACT

The paper presents the application of statistical (econometrics-originated) methods to process learning and testing data sets used by the artificial intelligence (AI) methods in the diagnostics of analog systems. Before the training and evaluation of the intelligent module is performed, the measurement data are analysed to minimize the number of attributes (symptoms) required to distinguish between different states of the System Under Test (SUT). This way the knowledge extracted from the set is simplified, increasing the operation speed and minimizing the threat of overlearning. Also, elimination of unnecessary symptoms from the set allows for decreasing the set of test points where measurements are taken (which is economically desirable). Preprocessing operations include elimination of constant or quasi-stationary symptoms and finding their minimal set, allowing for the efficient fault detection or parameter identification. The paper focuses on the Hellwig and Multiple Correlation Coefficient methods adjusted to the technical diagnostics applications. They are implemented to optimize data sets obtained from simulation of the fifth order lowpass filter. Their usefulness is tested using the artificial neural network (ANN) and Rough Sets (RS) classifiers responsible for detection, and identification of parametric faults.

1. Introduction

Contemporary methods used in the diagnostics of analog systems are often AI-based. Their advantages are versatility (making them useful in the analysis of wide range of objects), high efficiency and autonomy (ability to work without the human supervision). Advances in computer technologies enabled implementation of sophisticated fault detection and identification algorithms in small microprocessor systems (microcontrollers, or programmable logical controllers) [1]. The principles of AI-based diagnostic modules (including ANN [2] or Support Vector Machines – SVM [3]) include exploitation of data sets obtained from simulations (the most typical approach) of the SUT, or the real system. They allow for extracting knowledge about dependencies between features (symptoms) observed in the SUT output signals, and its actual state. To achieve this, machine learning methods are used, delivering knowledge to the diagnostic expert systems [4]. The latter make decision about the fault existence. The module quality is usually verified using the testing data set, containing examples that describe the SUT behaviour not present during the training. The correct sets' preparation is important to ensure the knowledge generalization. The optimal selection of training examples describing parametric faults (which are the main interest of the presented research) is one of the most challenging

* Address: Institute of Radioelectronics, Warsaw University of Technology, Ul. Nowowiejska 15/19, 00-665 Warsaw, Poland. Tel.: +48 22 234 7479.
E-mail addresses: pbilski@ire.pw.edu.pl, piotr_bilski@sggw.pl

http://dx.doi.org/10.1016/j.measurement.2014.03.023
0263-2241/© 2014 Elsevier Ltd. All rights reserved.

tasks in the contemporary diagnostics. The contents of the set should allow for both fault detection and location, assuring the high system testability [5]. In many cases only the former is required (go/no-go tests in the manufacturing process). The additional information about the source and intensity of the problem is used to repair the system or compensate the influence of the faulty parameter on its work regime. In contrast to the catastrophic faults (changing the SUT structure and therefore easier to detect) [6], their parametric counterparts temporarily or permanently change the SUT's behaviour (caused, for instance, by wearing out its constituent elements). During the data set generation, the following features must be provided:

- The learning set should cover as many cases of the SUT states, as possible (including nominal and faulty situations). This increases the chance of identifying them in the actual system.
- The number of examples in the set should be small to avoid constructing too complex knowledge during the training.
- The set of symptoms (characteristic information about the SUT state measured from its responses) should be minimized. The aim is to simplify the AI method and decrease the number of SUT nodes, which must be monitored (making the diagnostic procedure cheaper).

The preparation of 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 for simulations). Selection of the analysis domain, excitation signals, testing frequencies [7] and the optimal set of symptoms is difficult and time-consuming. Usually, all available information is collected and presented to the AI method, which selects important features [8]. This leads to data sets with large number of attributes, only some being crucial for the fault detection and identification process. Therefore the preprocessing stage is needed before the knowledge extraction [9]. The general scheme of the diagnostic module application is in Fig. 1 (omitting the data acquisition-specific operations, such as signal conditioning and denoising).

The aim of the preprocessing is the elimination of redundant data and selecting the most important symptoms to distinguish between various SUT states.

The proposed approach was designed to work with single faults. Preparing data describing the SUT's behaviour affected by multiple parametric faults is impractical. It requires checking numerous (potentially infinite) combinations of simultaneous faults [10]. Therefore prepared data sets contain only experiments representing one problem, while all remaining parameters are within tolerance margins.

The Hellwig and Multiple Correlation Coefficients (MCC) methods [11] are implemented here to prepare training and testing data sets for the AI-based diagnostic algorithm. They identify the set of attributes containing the greatest amount of information about the actual SUT's state. The minimization of analysed symptoms simplifies the AI method's knowledge, decreases the training duration and the cost of the actual SUT monitoring. The latter is determined by the set of analysed nodes. Minimizing them allows for decreasing the number of required sensors (like in machines or turbines [12]) and outputs accessible through the pins in the case of the integrated circuit. The testability requirement stipulates to keep the diagnostic outcome as high as possible with the minimal number of nodes [5]. Verification of both preprocessing approaches is performed using the diagnostic module based on the selected AI methods (ANN and RS). They were trained and tested on the original and reduced data sets.

The paper is organized as follows. In Section 2 the computational problem solved in the paper is defined. Section 3 contains the state of the art in the data preprocessing methods to show alternative approaches applicable for the task. Section 4 introduces the data sets' structure, used during the experiments. In Section 5 the implemented preprocessing methods are discussed. Section 6 presents the SUT diagnosed in the presented research, i.e. the fifth order filter. In Section 7 results of the data analysis are presented. Finally, Section 8 contains conclusions and future prospects.

2. Problem statement

Finding important features in the large amount of available data is the aim of statistical methods used in classification and regression [13]. Intelligent method works on the reduced set, resulting in the higher accuracy. Approaches using the correlation matrix locate dependencies between symptoms and identify the most important ones. Their disadvantage is that the information about the relation between the symptoms and the actual SUT state is ignored. The problem solved in the paper is to find the minimal set of symptoms, describing the SUT with the highest accuracy.

The solution proposed to solve this problem considers the correlation between the fault category and the set of attributes visible in response signals, increasing the diagnostic efficiency. Two statistical methods used in econometrics for building models of the company [14] are applied here. They also exploit the correlation matrix, but additionally consider relations between the symptoms and the SUT's actual state. This way the original data set may be modified to contain only the most relevant attributes. The idea is to eliminate all symptoms linearly dependent on each other. As opposed to the correlation-based method, the proposed approaches also measure the

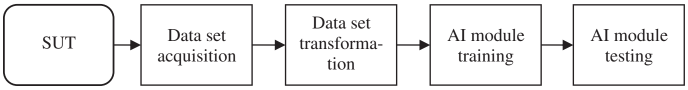

Fig. 1. Scheme of the data processing during training and testing of the AI-based diagnostic module.