Information Sciences 259 (2014) 380–392

Contents lists available at ScienceDirect

Information Sciences

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

Comparison of data mining tools for significance analysis of process parameters in applications to process fault diagnosis

Marcin Perzyk *, Andrzej Kochanski, Jacek Kozlowski, Artur Soroczynski, Robert Biernacki

Institute of Manufacturing Technologies, Faculty of Production Engineering, Warsaw University of Technology, Narbutta 85, 02-524 Warszawa, Poland

ARTICLE INFO

Article history:
Available online 26 October 2013

Keywords:
Fault diagnosis
Data mining
Input variable significance
Manufacturing industries

ABSTRACT

This paper presents an evaluation of various methodologies used to determine relative significances of input variables in data-driven models. Significance analysis applied to manufacturing process parameters can be a useful tool in fault diagnosis for various types of manufacturing processes. It can also be applied to building models that are used in process control. The relative significances of input variables can be determined by various data mining methods, including relatively simple statistical procedures as well as more advanced machine learning systems. Several methodologies suitable for carrying out classification tasks which are characteristic of fault diagnosis were evaluated and compared from the viewpoint of their accuracy, robustness of results and applicability. Two types of testing data were used: synthetic data with assumed dependencies and real data obtained from the foundry industry. The simple statistical method based on contingency tables revealed the best overall performance, whereas advanced machine learning models, such as ANNs and SVMs, appeared to be of less value.

1. Introduction

Materials and manufacturing are generally recognized as the main cost components of products. Fault diagnosis in manufacturing systems is therefore an extremely important issue and has been a subject of great interest among scientists and practitioners for many years. A large number of advanced techniques, including those based on machine learning, are currently employed in research and new developments in this field. They can be applied to the control and fault diagnosis of hardware elements and systems which are utilized in many areas (see e.g. [1,6,7,12,18,20,41–43]). Similarly, they can be useful in the fault diagnosis of the whole production process, which is the scope of the present work. As indicated in [34], a fault or problem in the process does not have to be the result of equipment failure or does not even have to involve specific hardware. A problem might be defined as a non-optimal operation or an off-spec product. For example, in a process plant the root causes of non-optimal operation might be hardware failures, but the problems might also be the result of a poor choice of operating targets, poor feedstock quality, poor controller tuning, sensor calibration errors, or human errors.

Models that are used in production process control and fault diagnosis can be of various types, including qualitative models such as the Ishikawa Cause-and-Effect diagrams used in Statistical Quality Control (SPC) as well as quantitative casual empirical models where the inputs are process parameters, material properties, organization variables, and human factors. The latter can be used to check whether a particular combination of values of the input variables is likely to trigger off abnormal behavior.

* Corresponding author. Tel.: +48 509093935; fax: +48 228499797.
E-mail address: M.Perzyk@wip.pw.edu.pl (M. Perzyk).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.10.019

Building empirical casual models that are suitable for process fault diagnosis requires intelligent analysis of production data that is based on the data mining approach. Since the year 2000 a large growth in the number of various applications of data mining aimed at supporting and improving manufacturing processes has been observed. Some important reviews or systemic approaches can be found in [8,13,14,16,24,37–39].

A reasonable approach to detecting the root causes of various types of process faults is the application of significance analysis of the process parameters. Those variables which are found to be the most significant to a given process fault (e.g. an increasing percentage of defective parts) could be regarded as the number one candidates for the root causes of the fault. Also, the diagnosis of machine or equipment break-downs can be based on finding the most significant operation parameters for the output variable which defines the occurrence of the break-down. Various examples of using significance analysis in process parameters can be found in many works, e.g. [3,4,10,19,23,32,33,38,40]. Significance analysis can also be helpful in building optimal fault diagnosis models by selecting only important input variables. This applies both to qualitative models (e.g. cause-and-effect diagrams) and quantitative empirical models (e.g. neural networks).

Finding the most significant variables may not only be useful in fault diagnosis, but it also provides the means for establishing optimal inspection procedures in the process, thus enabling engineers to concentrate on selected process variables and avoiding unnecessary costs. Similarly, finding the most significant variables allows to select them as the most efficient process variables in Engineering Process Control, i.e. by having the largest gain coefficients.

It is important to note that fault diagnosis based on parameter significance analysis can be applied to discrete processes, which are characteristic of such industries as electronics, cars, aircrafts, and household products, as well as to process industries, such as chemical, textile, and food, in which continuous or batch-type processes are typical.

The significance of an input variable can be understood in a variety of ways. One of these is based on sensitivity analysis, which returns changes of the output variable due to small variations in the input variable, calculated at particular levels of the input. However, in the opinion of the authors, practitioners would rather be interested in finding the potentially greatest overall effect of a process variable on the process results or equipment behavior. The latter approach as been adopted in the present work. Nevertheless, particular definitions of variable significances, as presented in Section 2.2, are not directly based on the maximum difference of the output which can be obtained by changing the value of the analyzed input.

In order to perform significance analysis a suitable input–output model based on observations (past production data) should be built. For the purposes of fault diagnosis models suitable for classification-type tasks seem to be the most appropriate, as the outputs are usually in the categorical form, e.g. ‘faulty’ or ‘acceptable’ for product quality or ‘failure’ or ‘no failure’ for the equipment state. The purpose of the present work was to conduct a comparative analysis of some classification models from the viewpoint of their precision and robustness of determining the relative significances of the input variables as those playing the role of process parameters.

The authors’ experience indicates that introducing data mining techniques, especially in the manufacturing industry, is often difficult due to a lack of deep insight into the characteristics of particular methods and algorithms. Comparative analyses of different methods made from the point of view of their performance in carrying out specific tasks are therefore very important. Those aimed at assessing tools that would be suitable for determining relative significances of the process variables are relatively rare. Hence, the selection of appropriate methodology is often casual.

In a previous work [25], a study of the regression-type models was made. It revealed important characteristics of some types of the models and methodologies that have been applied for the extraction of information concerning input variable significances, including possible interactions among them. The usefulness of this significance analysis, as applied in some real foundry processes as well as in generally nonlinear processes, was also shown. The present study addresses some important issues related to the discrete types of both input and output variables. In consequence, the types of applied models and the computational methodologies are essentially different. The current work is, therefore, novel, and, as the authors believe, could help practitioners in selecting the right data mining tools to identify the process parameters that are responsible for production quality and effectiveness of control of the processes.

At this stage the analysis is limited to single variables only.

2. Methodology

2.1. Description of the data sets

As in the previous works, two types of data sets were used: simulated (synthetic) data, with assumed hidden dependencies between inputs and output as well as real, industrial data. The synthetic data were obtained by assuming analytical formulas of the type $Y = f(X1, X2, \ldots)$, from which for random values of continuous-type input variables $X1$, $X2, \ldots$ the continuous-type dependent variable Y was first calculated. A Gaussian-type noise with a zero mean was then imposed on the input variables, with maximum deviation at ±20%; this value was found to be characteristic of many real manufacturing processes. Finally, all of the continuous values were converted to categorical ones by using the equal intervals method. In most cases the assumed number of categories was 5, thus reflecting the popular verbal scale ranging from ‘very low’, through ‘small’, ‘medium’ and ‘high’, to ‘very high’. Sets comprising 1000 records (further referred to as ‘large’) and 100 records (further referred to as ‘small’) were generated in this way in order to compare the performances of the models for data with different representativeness of variable values (classes). The ‘large’ sets were used to evaluate the precision of the variable