Neurocomputing 150 (2015) 289–303

Contents lists available at ScienceDirect

Neurocomputing

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

Fully unsupervised fault detection and identification based on recursive density estimation and self-evolving cloud-based classifier

Bruno Sielly Jales Costa [a,*], Plamen Parvanov Angelov [b,1], Luiz Affonso Guedes [c,2]

[a] Campus Natal – Zona Norte, Federal Institute of Rio Grande do Norte – IFRN, Brazil
[b] School of Computing and Communications, Lancaster University, Lancaster, UK
[c] Department of Computing Engineering and Automation, Federal University of Rio Grande do Norte – UFRN, Brazil

ARTICLE INFO

Article history:
Received 1 December 2013
Received in revised form
5 March 2014
Accepted 1 May 2014
Available online 30 October 2014

Keywords:
Fault detection
Fault diagnosis
Fault identification
Recursive density estimation
Evolving classifiers
Autonomous learning

ABSTRACT

In this paper, we propose a two-stage algorithm for real-time fault detection and identification of industrial plants. Our proposal is based on the analysis of selected features using recursive density estimation and a new evolving classifier algorithm. More specifically, the proposed approach for the detection stage is based on the concept of the density in the data space, which is not the same as the probability density function, but is a very useful measure for abnormality/outliers detection. This density can be expressed by a Cauchy function and can be calculated recursively, which makes it memory and computational power efficient and, therefore, applicable to on-line applications. The identification/diagnosis stage is based on a self-developing (evolving) fuzzy-rule-based classifier system proposed in this paper, called the AutoClass. An important property of AutoClass is that it can start learning "from scratch". Not only do the fuzzy rules not need to be prespecified, but neither do the number of classes for AutoClass (the number may grow, with new class labels being added by the online learning process), in a fully unsupervised manner. In the event that an initial rule base exists, AutoClass can evolve/develop it further based on the newly arrived faulty state data. In order to validate our proposal, we present experimental results from a level control didactic process, where control and error signals are used as features for the fault detection and identification system, but the approach is generic and the number of features can be significant due to the computationally lean methodology, since covariance or more complex calculations, as well as storage of old data, are not required. The obtained results are significantly better than the traditional approaches.

1. Introduction

In the past few decades fault detection and identification (FDI) field of research has received extensive attention. It is an important problem in control and automation engineering and is the centre of abnormal event management (AEM) field of research [63]. Applications of FDI techniques in industrial environments are increasing in order to improve the operational safety as well as to reduce the costs related to unscheduled stoppages. The importance of the FDI research in control and automation engineering is based on the fact that prompt detection of an occurring fault, while the system is still operating in a controllable region, usually prevents or, at least, reduces productivity losses and health risks.

With the increasing complexity of the procedures and scope of the industrial activities, AEM is a challenging field of study nowadays. The human operator plays a crucial role in this matter since it has been shown that people responsible for AEM often take incorrect decisions. Industrial statistic shows that 70–90% of the accidents are caused by human errors [64].

In the industrial context, there are several different types of faults that could affect the normal operation of a plant. Among these we can list [53]:

- *Gross parameter changes*: Also known as parametric faults, which refer to disturbances to the process from independent variables, whose dynamics are not known. As examples of parametric faults one can list a change in the concentration of a reactant, a blockage in a pipeline resulting in a change of the flow coefficient and so on.

* Corresponding author. Tel.: +55 84 40069509; fax: +55 84 40069504.
E-mail addresses: bruno.costa@ifrn.edu.br (B.S.J. Costa), p.angelov@lancaster.ac.uk (P.P. Angelov), affonso@dca.ufrn.br (L.A. Guedes).
[1] Tel.: +44 1524 510391.
[2] Tel.: +55 84 32153771; fax: +55 84 32153738.

http://dx.doi.org/10.1016/j.neucom.2014.05.086

- *Structural changes*: These refer to equipment failures, which may change the model of the process. An appropriate corrective action to such abnormality would require the extraction of new modeling equations to describe the current faulty status of the process. Examples of structural changes are failure of a controller, a leaking pipe and a stuck valve.
- *Faulty sensors and actuators*: Also known as additive faults, refer to incorrect process inputs and outputs, and could lead the plant variables beyond acceptable limits. Some examples of abnormalities in the input/output instruments are constant (positive or negative) bias, intermittent disturbances, saturation, out of range failure and so on.

The entire process of AEM is often divided into a series of steps, which in fault-tolerant design is called the fault diagnosis scheme. Fault detection or anomaly detection is the first stage and it has extreme importance to FDI systems. In this stage, we are able to identify if the system is working in a normal operating state or in a faulty mode. However, in this stage, vital information about the fault, such as physical location, length or intensity, is not provided to the operator [56].

In this sense, the need of a subsequent stage arises. The detector system (first stage) continuously monitors the process variables (or attributes) looking for symptoms (deviations from the normal variables values) and sends these symptoms to the diagnosis system, second stage, which is responsible for the classification process.

The diagnosis stage presents its own challenges and obstacles, and can be handled independently from the first one. It demands different techniques and solutions, and is divided in two substages called isolation and identification. The term isolation refers to determination of the type, location and time of detection of a fault, and follows the fault detection stage [26]. Identification, on the other hand, refers to determination of the size and time-variant behavior of a fault, and follows the fault isolation.

A lot of approaches to FDI have been proposed in the literature. We can mention, for example, the observer-based [20,41,46,58], analytical redundancy-based [57,15,65], fuzzy model-based [50,66,27,36], neural network-based [39,62,18], immune system-based methods [37,38] and so on. Unfortunately, most of the above-mentioned techniques require either previous knowledge or empirical observation about the model or behaviour of the system, need extensive computational efforts or too many thresholds or problem-specific parameters to be pre-defined in advance, inhibiting/hampering their use in on-line applications. Thus, these technical features make difficult their adoption in real problems.

One group of methods which is worth to mention, and serves as a basis for comparison with our proposal, later in this paper, is the group of statistical process control approaches (SPC). SPC deals with data which are snapshot windows of moving the history of a process control system [31]. It is used for process variables monitoring and is based on statistical analysis (mean and standard deviation values), calculated in time windows and compared with pre-defined thresholds. Although, SPC is an on-line approach, most of the applications in use today were developed based on the premise that the process parameters being controlled follow Gaussian/normal distributions. Independence of the inputs and infinite number of observatories are other premises which, in reality are not satisfied. For further information on SPC methods, the reader is referred to Martin et al. [49], Cook et al. [21], Liukkonen and Tuominen [43], Kano et al. [33].

Being aware of these shortcomings, in this paper we propose a recursive fully unsupervised fuzzy rule-based (FRB) classifier for fault detection and identification in industrial processes, which can be generalised for other specific problems. The proposed FDI system does not demand neither mathematical models based on first principles nor explicit previous knowledge about the analysed process. It is based, instead, on the estimation of the density and proximity in the data space. This density can be expressed by a Cauchy function and can be calculated recursively [5], which makes it memory- and, thus, computational power-efficient and suitable for on-line applications. In this sense, it is autonomous (user-independent) and is able to perform FDI on-line and without the above-mentioned disadvantages. The proposed approach has two well-defined and sequential stages – detection and identification – with a minimum of very intuitive parameters, that can be associated with other existing approaches.

The proposed on-line detection algorithm is based on the recently introduced recursive density estimation (RDE) approach [11]. This approach allows us to build, accumulate, and self-learn a dynamically evolving information model of "normality" based on the process data for particular specific plant based on the normal/"good"/accident-free cases only. Theoretically, such an approach can start fault detection "from scratch" from the very first data sample observed.

It is important to stress that only a few techniques for data density analysis in fault detection have been previously proposed, most of them applied to software fault detection applications and based on probability density function (PDF), not data distribution density. Breunig et al. [19] present the probability density-based local outlier factor (LOF) algorithm. In this approach the anomaly score of a data sample is defined as the average local probability density of its neighbors. Similar methods based on the KNN algorithm were presented in Tang et al. [61,30,51]. However, most of the existing algorithms suffer from high complexity, therefore, are not suitable for large datasets or real-time applications.

For the identification stage, the proposed approach is based on the new self-learning (fully unsupervised) evolving classifier algorithm called the *AutoClass*. It builds upon the family of evolving clustering – eClustering [2], ELM [16], DEC [17] – and classifier – eClass [14], simpleClass [6] – algorithms. The new clustering algorithm, called the *AutoClass* differs from the eClass0 in the way clusters are defined and updated. While they are based on the concept of traditional clusters, AutoClass works with the concept of *data clouds* [12], structures with no defined boundaries or shapes. Another innovation, when compared to *eClass0*, for example, is that AutoClass can store a finite vector of points (for a limited time) which do not belong to any existing class and later create a new class from them. Like eClass0, AutoClass also can start from an empty knowledge base, from the first data sample acquired.

Among the related work, it is important to mention some of the recently presented approaches in the field of fault detection, using adaptive and evolving FRB models. The paper [55] presents an approach to FDI based on data-driven evolving fuzzy models and dynamic residual analysis for extracting fault indicators. The authors introduce a two-stage algorithm, one off-line (model identification and training) and one on-line (fault detection), where neither annotated samples nor fault patterns/models need to be available *a priori*. The FDI system is successfully applied to a power plant coal mills. Lemos et al. [40], Lughofer and Guardiola [45] present two different fully on-line FDI systems, using evolving fuzzy classifiers, based on the evolving Takagi–Sugeno (eTS) algorithm, first introduced by Angelov and Filev [9] and Angelov and Zhou [14]. The work of Lughofer [44] also worth mentioning, since the author developed an evolving image classifier, capable of sort the images into "good" (fault-free production items) and "bad" (faulty production items). Regarding the extraction of decision rules from data streams and handling time changing data, few approaches can be mentioned, e.g. Gama and Kosina [29] and Kosina and Gama [35]. In the first paper, the authors present a new algorithm to learn rule sets, designed for open-ended data streams and, in the latter, an on-line, any-time and one-pass algorithm for learning decision rules in the context of time