Expert Systems With Applications 63 (2016) 134–144

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An evolving approach to unsupervised and Real-Time fault detection in industrial processes

Claubér Gomes Bezerra[a,*], Bruno Sielly Jales Costa[b], Luiz Affonso Guedes[c], Plamen Parvanov Angelov[d,e]

[a] Federal Institute of Rio Grande do Norte - IFRN, Campus EaD, Av. Senador Salgado Filho 1559, Tirol, CEP: 59015-000, Natal, RN, Brazil
[b] IFRN - Campus Natal Zona Norte, Rua Brusque 2926, Potengi, CEP 59112-490, Natal, RN, Brazil
[c] Federal University of Rio Grande do Norte - UFRN, Department of Computer Engineering and Automation - DCA, Campus Universitrio, Lagoa Nova, CEP: 59078-900, Natal, RN, Brazil
[d] Lancaster University, Data Science Group, School of Computing and Communications, Lancaster LA1 4WA, United Kingdom
[e] Chair of Excellence, Carlos III University, Madrid, Spain

A R T I C L E I N F O

Article history:
Received 30 December 2015
Revised 15 June 2016
Accepted 16 June 2016
Available online 27 June 2016

Keywords:
Fault detection
Industrial processes
Typicality
Eccentricity
TEDA
Autonomous learning

A B S T R A C T

Fault detection in industrial processes is a field of application that has gaining considerable attention in the past few years, resulting in a large variety of techniques and methodologies designed to solve that problem. However, many of the approaches presented in literature require relevant amounts of prior knowledge about the process, such as mathematical models, data distribution and pre-defined parameters. In this paper, we propose the application of TEDA – Typicality and Eccentricity Data Analytics – , a fully autonomous algorithm, to the problem of fault detection in industrial processes. In order to perform fault detection, TEDA analyzes the density of each read data sample, which is calculated based on the distance between that sample and all the others read so far. TEDA is an online algorithm that learns autonomously and does not require any previous knowledge about the process nor any user-defined parameters. Moreover, it requires minimum computational effort, enabling its use for real-time applications. The efficiency of the proposed approach is demonstrated with two different real world industrial plant data streams that provide "normal" and "faulty" data. The results shown in this paper are very encouraging when compared with traditional fault detection approaches.

1. Introduction

Nowadays, industries from a variety of production sectors increasingly seek to meet the market requirements, such as production increase, continuity and reliability of the processes, in addition to safety and environmental restrictions. In order to cope with these challenges, industries have been investing more and more in automation of the production processes, increasing the general complexity of the systems. Thus, process maintaining becomes a complex task due to the large number of equipment and variables that need to be monitored.

Therefore, there is a growing demand for robust and reliable industrial control and monitoring systems. The industrial process should be able to perform a specified function, under determined conditions, in a given period of time, while remaining safe for people, equipment and the environment (Isermann, 2006). Moreover, these systems should be efficient in the sense of being able to handle large amounts of variables and data provided by the equipment of the plant.

One of the approaches for tackling both problems is to increase quality, safety and robustness of the sensors, actuators and controllers, in addition to the structure of the plant itself. However, over time, the industrial equipment are likely to show a number of signs of degradation, such as exhaustion, dirt, corrosion, cracks, damage caused by operators, among others. The appearance of such signs turns the plant susceptible to fault occurrences during its operation.

A fault consists of an unpermitted deviation of at least one characteristic property or variable in a system from its acceptable, usual or standard condition (Isermann, 1997). In an industrial process, a fault can be defined as an unexpected change on the functioning of one or more process components that can lead it to a critical situation. Sometimes, a fault may cause a number of problems, such as unexpected stoppages, production losses, reduction

* Corresponding author.
E-mail addresses: clauber.bezerra@ifrn.edu.br (C.G. Bezerra), bruno.costa@ifrn.edu.br (B.S.J. Costa), affonso@dca.ufrn.br (L.A. Guedes), p.angelov@lancaster.ac.uk (P.P. Angelov).

http://dx.doi.org/10.1016/j.eswa.2016.06.035
0957-4174/

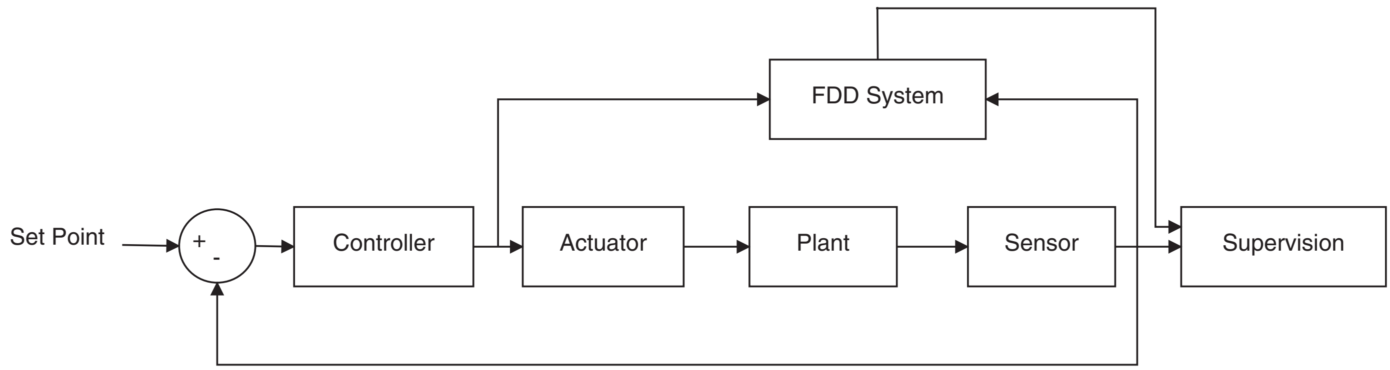

Fig. 1. FDD system scheme.

of equipment lifespan, or even accidents with severe consequences to the environment and human life (Venkatasubramanian, 2003).

Very often, a fault-free process is not feasible. Thus, the use of a fault detection and diagnosis (FDD) system becomes crucial (Ding, 2008). FDD systems usually are responsible for the increase of process availability, reliability and safety, in addition to cost reduction and more efficient maintaining. A FDD system is often integrated to the traditional supervision and control systems, as shown in Fig. 1.

The FDD systems work by monitoring process variables and analyzing their behaviors. Therefore, they should be able to determine the occurrence of a fault – *fault detection* – , its location and cause – *fault diagnosis* – , by analyzing process inputs/outputs and sending information regarding the fault to the supervisory system. Therewith, the operator is able to decide how and when to act in order to avoid a critical state of the process. With this strategy, it is possible to avoid unnecessary stoppages and accidents.

High demands for monitoring and fault detection in industrial systems resulted in research and development of many FDD techniques in the last few decades using different data analytics methods. These methods are often classified as model-based and process history-based (Katipamula and Brambley, 2005; Venkatasubramanian, Rengaswamy, Kavuri, (2003).

Model-based methods use the concept of residual analysis. In this type of approach, the residual error, which consist of the difference between a value measured on the output and a value estimated from a previously defined quantitative or qualitative model, is calculated and considerable difference between the estimated and measured values might indicate the presence of a fault.

On the other hand, process history-based methods do not required pre-defined models of the system. These methods, also known as data-driven, analyze the temporal evolution of data from the system in order to detect anomalies in its behavior.

Many different approaches have been used to tackle FDD problems, including fuzzy systems (Mendonça, Sousa, & Sá da Costa, 2009; Oblak, Skrjanc, & Blazic, 2007; Yang, Xia, & Liu, 2011), state observers (Chen & Saif, 2007; Li & Yang, 2012; Sobhani & Poshtan, 2011; Zhou, Liu, & Dexter, 2014), neural networks (Leite, Hell Jr, & Gomide, 2009; Mrugalski & Korbicz, 2007; Yuan, Lu, Ma, & han Chen, 2015; Zhou, Pang, Lewis, & Zhong, 2011), principal component analysis (Cui, Li, & Wang, 2008), support vector machines (Zeng et al., 2013), parity equations (Zakharov, Tikkala, & Jms-Jounela, 2013), analytical redundancy (Anwar & Chen, 2007; Halder & Sarkar, 2007; Serdio, Lughofer, Pichler, Buchegger, & Efendic, 2014; Serdio et al., 2014; Xu & Tseng, 2007) and immune system-based methods (Laurentys, Palhares, & Caminhas, 2010; Laurentys, Ronacher, Palhares, & Caminhas, 2010). One of the main disadvantages of most of these approaches is that they require a predefined model (quantitative or qualitative) of the system, mathematically defined or estimated by offline training.

However, most of the mentioned approaches are limited in the sense that they require some kind of previous knowledge about the characteristics of the process. Therefore, the availability of mathematical, physical or behavioral models or the non-intuitive definition of parameters and thresholds are required. Moreover, large databases and extensive training are often mandatory.

Recently, methods for outlier detection have been applied to different problems, including fault detection in industrial problems (Chandola, Banerjee, & Kumar, 2007; Hodge & Austin, 2004; Singh & Upadhyaya, 2012). An outlier consists of an element from a data set that is significantly distinct from the other elements. Considering a signal obtained from an industrial plant, an outlier might indicate an anomaly or fault in the process.

Generally, the data in an industrial process is obtained continuously, in real time and, thus, outlier detection methods must be able to handle the data in the form of data streams. Therefore, each sample analyzed has a temporal aspect and is only available at the instant of the acquisition. In this context, an outlier is detected from the observation of a sequence of data samples analyzed over time.

Accordingly, other important aspects should be considering when choosing an outlier detection method, such as computational effort when handling high dimensional streaming data. Hence, information about past data samples must be stored and analyzed without compromising memory and execution time.

Many authors address such problem with time series analysis Hu, Dong, (2015) and outlier detection methods, thoroughly discussed in Chandola et al. (2007) and Hodge and Austin (2004), which include Statistical Modeling (Ma, Hu, & Shi, 2013; Yan, Chen, Yao, & Huang, 2016), Neural Networks (King et al., 2002; Li, Pont, & Jones, 2002), Spectral Decomposition (Fujimaki, Yairi, & Machida, 2005) and Rule-based Systems (Ramezani & Memariani, 2011).

In this work, we deal solely with the fault detection stage, omitting, then, the diagnosis stage. This is an application of the anomaly detection field of study, consisting of a "one-class" classification problem, by deciding whether a data sample belongs to the "normal" class or not (fault).

In order to solve this problem, we will make use of a recently proposed approach to anomaly detection within a data stream. Typicality and Eccentricity Data Analytics (TEDA) is based on the spatial proximity among the data samples and has been successfully applied to anomaly detection (Bezerra, Costa, Guedes, & Angelov, 2015), clustering, classification, regression, among other problems (Kangin & Angelov, 2015).

This paper presents a practical application of TEDA algorithm to two different real world industrial fault detection problems. The first application uses the well known DAMADICS fault detection benchmark, that provides real data (not simulated) from the operation of a sugar factory plant. The second application consists of a laboratory pilot plant for process control, equipped with real industrial instruments.