Expert Systems With Applications 95 (2018) 368–383

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Fault diagnosis in industrial chemical processes using interpretable patterns based on Logical Analysis of Data

Ahmed Ragab[a,b,c], Mohamed El-Koujok[b], Bruno Poulin[b], Mouloud Amazouz[b], Soumaya Yacout[a,*]

[a] Applied Mathematics and Industrial Engineering Department, École Polytechnique de Montréal, 2500 Chemin de Polytechnique, Montréal, Québec H3T 1J4, Canada
[b] CanmetENERGY-Natural Resources Canada (NRCan), 1615 Lionel-Boulet Blvd., P.O. Box 4800, Varennes, Québec J3X 1S6, Canada
[c] Industrial Electronics and Control Engineering Department, Faculty of Electronic Engineering, Menoufia University, Menouf, Menoufia 32952, Egypt

ARTICLE INFO

Article history:
Received 2 June 2017
Revised 16 November 2017
Accepted 17 November 2017
Available online 23 November 2017

Keywords:
Fault detection and diagnosis
Industrial chemical processes
Tennessee Eastman Process
Logical analysis of data
Machine learning and pattern recognition
Black liquor recovery boilers

ABSTRACT

This paper applies the Logical Analysis of Data (LAD) to detect and diagnose faults in industrial chemical processes. This machine learning classification technique discovers hidden knowledge in industrial datasets by revealing interpretable patterns, which are linked to underlying physical phenomena. The patterns are then combined to build a decision model that serves to diagnose faults during the process operation, and to explain the potential causes of these faults. LAD is applied to two case studies, selected to exemplify the difficulty in interpreting faults in complex chemical processes. The first case study is the Tennessee Eastman Process (TEP), a well-known benchmark problem in the field of process monitoring and control that uses simulated data. The second one uses a real dataset from a black liquor recovery boiler in a pulp mill. The results are compared to those obtained by other common machine learning techniques, namely artificial neural networks (ANN), Decision Tree (DT), Random Forest (RF), k nearest neighbors (kNN), quadratic discriminant analysis (QDA) and support vector machine (SVM). In addition to its explanatory power, the results show that LAD's performance is comparable to the most accurate techniques.

1. Introduction

Early and accurate fault detection and diagnosis (FDD) in chemical plants has been shown to minimize downtime, improve safety, and reduce manufacturing costs (Maurya, Rengaswamy, & Venkatasubramanian, 2007). For large-scale chemical plants, it is suggested to build a data-driven FDD scheme based on the historical observations that are collected from a controlled process (Russell, Chiang, & Braatz, 2012). Most of the monitoring models in such plants have been constructed almost entirely from process data (Eslamloueyan, 2011). The main concept is to use the knowledge categorization principle to exploit a large number of data observations in order to create an inferential model for the process's operator (Gao, Cecati, & Ding, 2015).

Researchers on data-driven FDD are focusing on designing outlier detection techniques to monitor faulty and abnormal situations in complex plants (Bezerra, Costa, Guedes, & Angelov, 2016; Reis & Gins, 2017; Sadeghi & Hamidzadeh, 2016; Xiao, Wang, Xu, & Zhou, 2016; Yiakopoulos, Gryllias, Chioua, Hollender, & Antoniadis, 2016; Yin & Hou, 2016). Recent review papers on data-driven FDD methods in chemical processes were published by Ming and Zhao (2017), Tidriri, Chatti, Verron, and Tiplica (2016), Yin and Hou (2016), Yin, Ding, Xie, and Luo (2014), Yin, Li, Gao, and Kaynak (2015). According to the literature, data-driven FDD methods are grouped mainly into two categories; multivariate statistical process monitoring (MVSPM) and FDD classification methods. In what follows, we briefly review the literature on the common methods in both categories, in the context of chemical process FDD.

The MVSPM methods include principal component analysis (PCA), projection to latent structures (PLS), independent component analysis (ICA) and others. The principal component analysis (PCA) has been widely applied in FDD methods in industrial processes monitoring (Jiang, Yan, & Zhao, 2013). A PLS model was proposed in Yin, Zhu, and Kaynak (2015) for the online prediction of key performance indicators (KPIs) in chemical processes. Extensions to PLS were used in online process monitoring such as the total PLS (TPLS) (Zhou, Li, & Qin, 2010), concurrent PLS (CPLS)

* Corresponding author.
E-mail addresses: ahmed.ragab@polymtl.ca, ahmed.ragab@canada.ca (A. Ragab), mohamed.elkoujok@canada.ca (M. El-Koujok), bruno.poulin@canada.ca (B. Poulin), mouloud.amazouz@canada.ca (M. Amazouz), soumaya.yacout@polymtl.ca (S. Yacout).

https://doi.org/10.1016/j.eswa.2017.11.045
0957-4174/

(Qin & Zheng, 2013) and modified orthogonal PLS (MOPLS) (Yin, Wang, & Gao, 2016). Variants like kernel PCA (KPCA) (Deng, Tian, & Chen, 2013) and kernel PLS (KPLS) (Zhang, Zhou, Qin, & Chai, 2010) were developed to tackle the nonlinearity issues of industrial processes. Gharahbagheri et al. combined the KPCA with causality analysis for diagnosing faults in the fluid catalytic cracking (FCC) unit of chemical processes (Gharahbagheri, Imtiaz, & Khan, 2017a, 2017b). Yu proposed a nonlinear kernel Gaussian mixture model (NKGMM) based on inferential monitoring in a wastewater treatment batch process (Yu, 2012). Rashid proposed an FDD method based on the hidden Markov model (HMM) and ICA (HMM-ICA) to monitor faults in chemical processes (Rashid & Yu, 2012). Odiowei and Cao proposed the so-called state space independent component analysis (SSICA) to deal with nonlinear dynamic process monitoring (Odiowei & Cao, 2010).

In FDD classification methods, the knowledge is discovered from the data by using machine learning classification techniques (Liao, Chu, & Hsiao, 2012). These techniques include artificial neural network (ANN), Decision Tree (DT), Random Forest (RF), support vector machine (SVM) and others. The classification patterns in each of these techniques are extracted from the data in different ways. The patterns of SVM are obtained by finding the optimal hyperplanes separating the classes of data points, by lifting them from their original input space to a higher dimensional feature space by using different kernel functions (Yang & Widodo, 2008). In that higher dimensional space, straight hyperplanes can be possibly found, by solving constrained optimization problems that aim at maximizing the margins between the hyperplanes and the training data of different classes (Gunn, 1998). The classification patterns of ANN are represented in the form of connection weights between the network layers. The weights are updated during the learning process based on the backpropagation error signal that represents the difference between the desired output and the predicted outputs of the network. The Decision Tree is a flow-chart-like structure where each branching node represents a test on a certain variable, and each branch represents an outcome of the test. The values of the class variable are placed at the leaves of the tree (Witten, Frank, Hall, & Pal, 2016). The Random Forest is an ensemble strategy that combines the results of many decision trees, thus reducing the effects of overfitting and improves generalization ability of the final classification model (Liaw & Wiener, 2002).

The SVM has been applied to diagnose faults in chemical processes by Yin, Gao, Karimi, and Zhu (2014). Recently, Yin et al. reviewed the development of FDD methods based on SVM (Yin & Hou, 2016). Yu proposed the support vector clustering (SVC) method for unsupervised chemical process monitoring (Yu, 2013). The so-called support vector data decomposition (SVDD) was applied by Beghi et al. to diagnose faults in industrial chillers (Beghi et al., 2014). The ANN was used to diagnose faults in chemical processes by Nashalji, Shoorehdeli, and Teshnehlab (2010). The self-organizing map (SOM), a type of ANN, was applied as a diagnostic tool in non-Gaussian processes with intuitive visualization capability (Yu, Khan, Garaniya, & Ahmad, 2014). Lau et al. purposed an online fault diagnosis framework incorporating adaptive neuro-fuzzy inference system (ANFIS) and multi-scale principal component analysis (MSPCA) for dynamic processes (Lau, Ghosh, Hussain, & Hassan, 2013).

The Bayesian network (BN) is a probabilistic graphical model that was used in chemical processes fault diagnosis by Verron, Li, and Tiplica (2010). A BN approach was developed by Yu et al. to detect abnormal events and to identify the fault propagation paths in chemical processes (Yu & Rashid, 2013). Wang et al. combined the BN and semiparametric PCA to diagnose faults in nonlinear and non-Gaussian processes (Wang, Liu, Khan, & Imtiaz, 2017). Gharahbagheri et al. used the Bayesian belief network and KPCA for FDD (Gharahbagheri, Imtiaz, & Khan, 2017a, 2017b). The Decision Tree classifier was combined with other methods and used for fault diagnosis in chemical processes with incomplete observations (Askarian et al., 2016). The Random Forest classifier which is an ensemble of decision trees was used as an efficient FDD model in chemical processes in (Auret & Aldrich, 2010; Gajjar & Palazoglu, 2016; Shrivastava, Mahalingam, & Dutta, 2017).

Recently, researchers in the domain of deep learning have been focusing on the interpretability of their models (Haufe et al., 2014; Krell & Straube, 2017). However, interpretable diagnostic schemes for chemical processes are seldom seen in the literature. Despite the efforts of some authors to obtain interpretable diagnostic models, a certain level of human experience is needed for building those models (Chiang, Jiang, Zhu, Huang, & Braatz, 2015; Dong, Zhang, Huang, Li, & Peng, 2015; Lee, Tosukhowong, Lee, & Han, 2006; Maurya et al., 2007; Singhal & Seborg, 2006). Moreover, little attention has been paid to automatic development of interpretable fault diagnosis schemes that can analyze the fault's causes in complex chemical processes (Askarian et al., 2016; Lau et al., 2013; Li, Yuan, Qin, & Chai, 2015; Sajid, Khan, & Zhang, 2017; Yu & Rashid, 2013).

Due to the interactions between process variables in large-scale processes, it is difficult or even impossible to identify the relationships between the faults' causes and their effects. Complexity means that when an incipient fault appears in any part, it can propagate throughout the other parts, causing severe abnormal situations to the whole plant. Therefore, industrial chemical processes require an interpretable FDD method that can identify and analyze the type of fault and find its root causes. This will enable plant operators to take the right actions to address the source of abnormality, in order to avoid any serious or catastrophic events. Recently, the concept of applying Logical Analysis of Data (LAD), an interpretable diagnostic scheme, using simulated data, was presented in (Ragab, El-Koujok, Amazouz, & Yacout, 2017).

This paper applies the LAD as an interpretable pattern-based classification technique for FDD in industrial chemical processes. The aim is to interpret and diagnose faults by discovering knowledge in the form of explanatory patterns. Two examples of chemical processes are used to show the effectiveness of the proposed LAD diagnostic scheme. The first one is the simulated Tennessee Eastman Process (TEP), which is a well-known benchmark. The second is a real dataset from a black liquor recovery boiler from the pulp and paper industry. Our objective is to exploit the interpretability of LAD's patterns, in order to obtain a transparent diagnostic model for chemical plants, which is based on hidden rules that are discovered in the dataset. This diagnostic model helps the process expert to deeply understand hidden phenomena with minimal efforts.

The paper is structured in six sections. Section 2 presents and discusses the basic concept of LAD, as well as its notations and steps. The first case study of the TEP dataset is presented in Section 3. A real case study from the industry is presented in Section 4. The research extensions and future directions are discussed in Section 5, followed by concluding remarks in Section 6.

2. Logical Analysis of Data (LAD)

Logical Analysis of Data is a knowledge discovery approach, based on certain concepts from the fields of optimization and the theory of Boolean functions (Boros et al., 2000). It is a rule-based machine learning classification method that uses knowledge categorization, providing a systematic procedure for pattern extraction in multi-class problems (Mortada, Yacout, & Lakis, 2014). Like many supervised learning techniques, LAD assumes the analogy of the structural characteristics and the hidden knowledge of both the training and testing datasets (Mortada, Yacout, & Lakis, 2011).