Journal of Process Control 26 (2015) 90–101

Contents lists available at ScienceDirect

Journal of Process Control

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

Improving classification-based diagnosis of batch processes through data selection and appropriate pretreatment

Geert Gins, Pieter Van den Kerkhof, Jef Vanlaer, Jan F.M. Van Impe*

Chemical and Biochemical Process Technology and Control (BioTeC), Department of Chemical Engineering, KU Leuven, W. de Croylaan 46, B-3001 Leuven, Belgium

ARTICLE INFO

Article history:
Received 26 August 2014
Received in revised form 5 January 2015
Accepted 12 January 2015
Available online 31 January 2015

Keywords:
Batch processes
Fault detection/isolation
Process control
Mathematical modeling
Fault classification

ABSTRACT

This work considers the application of classification algorithms for data-driven fault diagnosis of batch processes. A novel data selection methodology is proposed which enables online classification of detected disturbances without requiring the estimation of unknown (future) process behavior, as is the case in previously reported approaches.

The proposed method is benchmarked in two case studies using the `Pensim` process model of Birol et al. (2002) implemented in RAYMOND. Both a simple k Nearest Neighbors (k-NN) and complex Least Squares Support Vector Machine (LS-SVM) are employed for classification to demonstrate the generic nature of the proposed approach. In addition, the influence of different data pretreatment methods on the classification performance is discussed, together with a motivation for selecting the correct pretreatment steps. Finally, the influence of the number of available training batches is studied.

The results demonstrate that a good classification performance can be achieved with the proposed data selection method even with a low number of faulty training batches by exploiting knowledge on the nature of the to-be-diagnosed faults in the data pretreatment. This provides a proof of concept for classification-based batch diagnosis and demonstrates the importance of incorporating process insight in the construction of data-driven process monitoring and diagnosis tools.

1. Introduction

Batch processes play an important role in the chemical and biochemical industries for the production of goods with a high added value (e.g., pharmaceuticals, food products, polymers, semiconductors) owing to their lower capital cost and higher flexibility to produce multiple products or grades. As any industrial process, batch processes can be prone to a number of disturbances such as impurities in the raw materials, fouling of heat exchangers, sensor failures, plugged pipes, etc. Since online product quality measurements are rarely available, close monitoring of batch processes and fast Fault Detection & Isolation (FDI) are absolute requirements to avoid unnecessary variations and ensure the final products are within specifications. The dynamic nature of batch processes further complicates FDI.

Today's process plants dispose of large historical databases containing measurements from hundreds of online sensors. Statistical Process Monitoring (SPM; sometimes referred to as Statistical Process Control) aims to exploit these historical data for FDI.

Most recent research in the field of FDI/SPM for batch processes has been devoted to fault detection and identification using *latent variable* approaches based on Principal Component Analysis (PCA) or Independent Component Analysis (ICA). While progress has been made in improving fault *detection* by including process dynamics (e.g., dynamic PCA [1], auto-regressive PCA [2], dynamic ICA [3]) or nonlinear extensions of PCA (e.g., kernel PCA [4], kernel ICA [5]), correct *isolation* (diagnosis) of the detected disturbance remains a difficult issue [6].

Contribution plots [7,8], which chart the contribution of each variable to an out-of-control signal, are by far the most popular tool to identify the root cause of an alarm signal in batch monitoring. The generation of contribution plots requires no prior knowledge about disturbances. However, process insight is necessary for interpreting the contribution pattern and finding the root cause. Moreover, Westerhuis et al. [9] and Van den Kerkhof et al. [10] illustrated that contribution plots can yield misleading results due to the so-called *fault smearing effect*.

* Corresponding author.
E-mail addresses: geert.gins@cit.kuleuven.be (G. Gins), pieter.vandenkerkhof@cit.kuleuven.be (P. Van den Kerkhof), jef.vanlaer@cit.kuleuven.be (J. Vanlaer), jan.vanimpe@cit.kuleuven.be (J.F.M. Van Impe).

http://dx.doi.org/10.1016/j.jprocont.2015.01.006
0959-1524/© 2015 Published by Elsevier Ltd.

If historical data of the different *known* faults types are available, diagnosis reduces to classification. A classifier is trained on faulty data and, upon detection, it assigns the detected fault to the class it most resembles. The time-consuming interpretation of contribution plots to find the root cause is eliminated, which significantly reduces the response time between fault detection and subsequent corrective action(s). For continuous processes in steady state, various classification-based diagnosis methods using, e.g., discriminant partial least squares [11], Fisher discriminant analysis [12], support vector machines [13,14] and neural networks [15] were reported in literature. For batch processes however, reports on the successful application of classification techniques are scarce. Since process plants are monitored and controlled to achieve satisfactory product quality and prevent process faults, the number of available faulty batches is limited. This is an important consideration for the design of a data-driven fault diagnosis scheme and forms an important bottleneck in the development of data-driven FDI.

Cho and Kim [16] proposed a Fisher Discriminant Analysis (FDA)-based classifier, but required a number of past faulty batches greater than the dimensionality of the fault data. In their case study, they simulated as many as 3500 faulty training batches for a process where 11 variables are monitored over 241 time points. Cho and Kim [17] generated pseudo-batches to deal with the data insufficiency. Cho [18] handled the data insufficiency more efficiently by extending linear FDA to nonlinear problems by employing *kernel* FDA. Li and Cui [19] further reduced the need for pseudo-batch generation. Recently, Support Vector Machines (SVMs) were utilized as a learning algorithm for fault classification of batch processes [20]. SVMs are based on statistical learning theory developed by Vapnik [21] and have shown to exhibit a large generalization performance, especially when the number of training samples is small [22]. Hence, SVMs are well suited for classification-based FDI.

The aforementioned methods [16–20] are essentially offline diagnosis methods since the classifier is trained on entire faulty batches, not faulty episodes. Hence, before fault classification, the new (faulty) batch needs to be completed or its future variable trajectories estimated. This is a major drawback. Moreover, the similarity between faulty batches with exactly the same fault at the same time instance decreases over time due to nonlinear process dynamics, different corrective actions and additional faults.

Therefore, this paper proposes a new fault diagnosis methodology which focuses on the onset of the fault rather than the entire batch history by sending only a small data window at the time of detection to the classifier. This eliminates the need for estimating future process behavior. The proposed method is validated on data generated from the `Pensim` model developed by Birol et al. [23], a widely used benchmark for data-driven FDI. To demonstrate the general applicability of the proposed method, both a simple k Nearest Neighbors (k-NN) [24] and complex Least Squares Support Vector Machine (LS-SVM) [25] classifier are used. Several data pretreatment steps are proposed to improve classifier performance, together with guidelines for selecting the appropriate steps. Additionally, the influence of the number of available training batches on the method's performance is studied.

The remainder of the paper is organized as follows. First, the proposed fault diagnosis methodology is outlined in Section 2. Section 3 describes the case study on which the fault diagnosis method is validated. Sections 4 and 5 briefly summarize the basics of standard PCA-based batch monitoring, which is used as the fault detection method in this work, and fault identification using k-NN or LS-SVM classifiers, respectively. They are followed by a discussion of the validation procedure in Section 6 and the obtained results in Section 7. Finally, conclusions and future research directions are provided in Section 8.

2. Proposed fault diagnosis method

Section 2.1 presents an overview of the proposed classification-based diagnosis method while Section 2.2 delves deeper into the critical issue of data pretreatment.

2.1. Overview

The proposed diagnosis methodology entails two phases: (i) an offline model building phase and (ii) an online diagnosis phase. A general scheme of the method is presented in Fig. 1.

2.1.1. Offline model building

The first step of the offline model building phase consists of scanning past batches for faulty episodes by means of an appropriate fault detection model (e.g., an NOC PCA model and corresponding fault detection statistics). By combining process knowledge and operator experience, the root cause of each detected fault is investigated. Based on this study, fault classes are defined and the detected faults are assigned to one of these fault classes. The definition of fault classes is an important step. A large number of fault classes reduces classification performance due to a decreasing amount of training batches per class and an increasing similarity between different classes. On the other hand, coarsely defined classes are unhelpful for straightforward corrective actions. Therefore, in practice the number of classes is a trade-off between classifier performance and practical use of the diagnosis results.

During the second step, the raw data are converted into a form amenable for learning a classification model (e.g., a k-NN or LS-SVM classifier). The goal of data pretreatment is to obtain a uniform characteristic fault pattern for each fault class. Data pretreatment is crucial for the classifier's diagnosis performance.

After collecting the data and converting these into a form suited for classification, the classifier is trained.

2.1.2. Online fault diagnosis

During the online or application phase, the current batch is monitored using the existing fault detection system. If a fault is

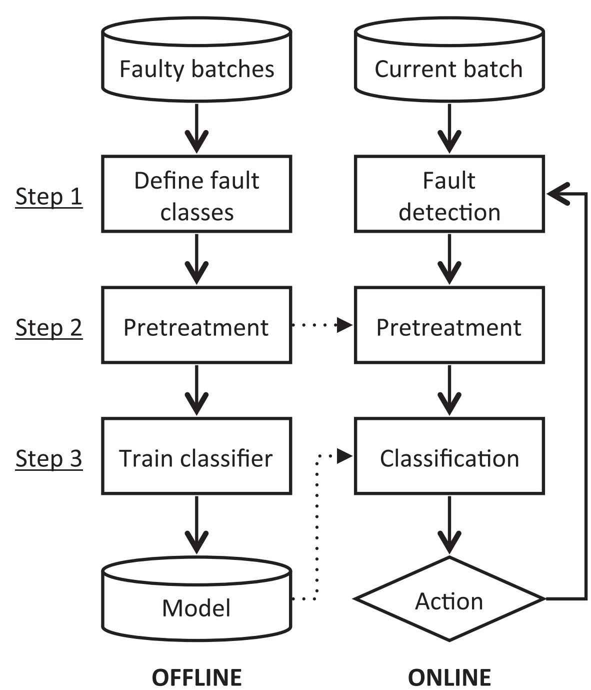

Fig. 1. Scheme of the proposed batch diagnosis method.