Journal of Process Control xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Process Control

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

Data mining and clustering in chemical process databases for monitoring and knowledge discovery

Michael C. Thomas, Wenbo Zhu, Jose A. Romagnoli*

Cain Department of Chemical Engineering, Louisiana State University, Baton Rouge, LA 70808, United States

ARTICLE INFO

Article history:
Received 26 April 2016
Received in revised form
24 November 2016
Accepted 7 February 2017
Available online xxx

Keywords:
Data mining
Data clustering
Dimensionality reduction
Knowledge discovery

ABSTRACT

Modern chemical plants maintain large historical databases recording past sensor measurements which advanced process monitoring techniques analyze to help plant operators and engineers interpret the meaning of live trends in databases. However, many of the best process monitoring methods require data organized into groups before training is possible. In practice, such organization rarely exists and the time required to create classified training data is an obstacle to the use of advanced process monitoring strategies. Data mining and knowledge discovery techniques drawn from computer science literature can help engineers find fault states in historical databases and group them together with little detailed knowledge of the process. This study evaluates how several data clustering and feature extraction techniques work together to reveal useful trends in industrial chemical process data. Two studies on an industrial scale separation tower and the Tennessee Eastman process simulation demonstrate data clustering and feature extraction effectively revealing significant process trends from high dimensional, multivariate data. Process knowledge and supervised clustering metrics compare the cluster results against true labels in the data to compare performance of different combinations of dimensionality reduction and data clustering approaches.

1. Introduction

Advancements in computing power and data storage at modern chemical plants have led to the build-up of large amounts data in historical databases which store sensor measurements from past process behavior. Recent research has led to process monitoring strategies which use the large output of process data to improve process safety and quality enhancement [1–3]. Data based process monitoring requires minimal process knowledge to perform this task, in contrast to model based approaches that require detailed mechanistic models.

Unfortunately, many of the best methods for data-driven fault detection and diagnosis are “supervised”, meaning training these algorithms require data organized into labelled groups, such as “faulty” or “normal”. In real plants labelled data is seldom available and creating properly labelled databases for training process monitoring algorithms can be a time consuming task. This task requires an engineer to assess multiple operating states, a large amount of sensors, and data from months or years of operations. This task also requires familiarity with the process to judge which measurements are abnormal under differing operating regimes. Reducing the difficulty of this initial step could lower the time and money required to create advanced fault detection and diagnosis systems and expand their application in industrial settings.

“Unsupervised” learning strategies can help discover groups of data automatically that might otherwise be buried in the sheer volume of data. Approaches to unsupervised learning include dimensionality reduction and data clustering. Learning patterns and extracting information about a process from data clusters or dimensionally reduced data can be called knowledge discovery or data mining. In order to expand the application of supervised process monitoring algorithms, a software framework must be constructed to: a) separate fault data from normal data, b) train a model based on statistics or supervised learning techniques for fault detection and c) assist with the identification and management of new faults. Each of these tasks must be performed in a way that is simple to understand for non-experts in data science and easy to deploy on multiple units around a plant with low overhead. The normal-faulty knowledge extracted using unsupervised learning can then be exploited to train supervised learning approaches for process monitoring.

Unsupervised learning is a widely studied topic in computer science [4,5] and chemometrics [6], but many clustering techniques beyond K-means have seen relatively limited application in process

* Corresponding author.
E-mail address: jose@lsu.edu (J.A. Romagnoli).

http://dx.doi.org/10.1016/j.jprocont.2017.02.006
0959-1524/

monitoring situations. Process data clustering has been previously shown to be effective in semiconductor manufacturing [7], high speed milling [8], and other applications [9,10]. Research in chemical process monitoring has also used data clustering concepts. Wang and McGreavy [11] performed an early study in clustering chemical process data from a fluid catalytic cracker simulation with a Bayesian automatic classification method. Bhushan and Romagnoli [12] utilized a self-organizing map for unsupervised pattern classification and with an application on a CSTR model for a fault diagnosis problem. Strategies integrating principal components analysis (PCA) and data clustering have also seen success. Maestri et al. [13] developed a fault detection strategy for multiple operating states based on PCA supported by data clustering. Zhu et al. [14] used a k-ICA-PCA modelling method to capture relevant process patterns with an application to monitoring the Tennessee Eastman process. Singhal and Seborg [15] developed a modified K-means methodology to cluster multivariate time-series data from similarity factors based on PCA. Barragan et al. [16] used a clustering strategy based on a wavelet transform and novel similarity metric to cluster data from the Tennessee Eastman process, but only study one process fault. Thornhill et al. [17] studied an approach for visualizing and clustering data based on PCA and hierarchical clustering.

This study uses traditional techniques for dimensionality reduction (DR) and data clustering from the computer science literature to extract faulty data and knowledge about process states from chemical process databases. Instead of focusing on how to detect and diagnose faults, this research focuses on how to create the data sets used to train conventional supervised process monitoring algorithms. We compare how effectively combinations of DR and data clustering techniques recreate fault labels on two case studies: the benchmark Tennessee Eastman process and an industrial separation tower.

An advantages of the workflow we propose is that it is relatively simple to use because each DR and clustering combination requires the specification of only one or two parameters and techniques. Additionally, we expand on previous research by considering innovative and proven clustering techniques such as DBSCAN, BIRCH, mean shift clustering that have been widely applied in computer science but are untested on fault data discovery. We also study the role of DR because of the prominent role it plays in visualization and feature extraction. As an example, Ding [18] explores the close relationship between unsupervised learning and DR and provides a theoretical basis for the use of PCA to enhance K-means clustering. The DR techniques considered include not only several techniques already adapted to fault detection and process monitoring (principal components analysis (PCA) [19], independent component analysis [20], kernel PCA [21]), but also non-linear manifold preserving techniques like Isomap and spectral embedding.

This paper is organized as follows: Section 2 summarizes our overall approach to data mining; Sections 3 and 4 introduce dimensionality reduction and data clustering respectively, providing a brief introduction to the techniques used in this study; Section 5 discusses how we decided the parameters of the DR and clustering techniques used; Section 6 considers a case study on the Tennessee Eastman process where unsupervised learning is leveraged to discover faults from sets of data; Section 7 studies the clustering of a real event on an industrial scale separation tower; Section 8 contains a brief review of the challenge of time series clustering; and Section 9 concludes and summarizes this research.

2. Data mining approach

Fig. 1 outlines the data mining approach used. First, DR techniques project the raw process data, removing redundant, correlated sensor measurements and combining them into lower dimensional scores. DR may project data to two or three dimensions to enable visualization, or the technique may simply remove redundant information from raw process data. In some cases DR may not be necessary if the data are of good quality.

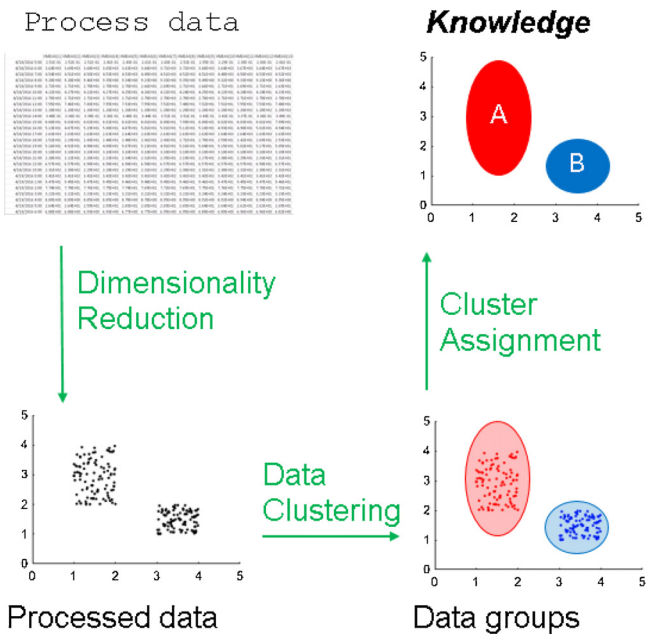

Fig. 1. Schematic of data mining approach.

After projection by a DR technique, data clustering algorithms partition the data using any number of clustering techniques. Xu and Wunsch [8] present a survey of data clustering techniques, but clustering is a subjective process and there is no universal definition of a cluster except that they consist of groups with members more similar to each other than data from different groups. Depending on the data and the parameters used to calculate the clustering, clusters found may or may not correspond to significant structures, therefore, cluster evaluation metrics are important to help the user judge the quality of the clusters extracted before more detailed analysis.

Finally, in the Cluster Assignment step, the user analyses the data in the clusters to relate them to meaningful process events such as faults or operating procedures. When labelled according to process events, the data can be used by machine learning or other supervised fault detection or diagnosis algorithms for training and fitting. Extracting information in this way from databases is called knowledge discovery. The data mining algorithms used in this study were drawn from the Python Scikit-learn module [22], which provides a rich environment of supervised and unsupervised machine learning algorithms.

3. Dimensionality reduction

Dimensionality reduction is an important data mining step because it addresses the “curse of dimensionality ”. High dimensional spaces lead to problems such as the empty space phenomena (increasing dimensionality increases volume such that available data becomes sparse), the weaker discrimination power of metrics like Euclidean distance, and correlations between variables [25]. The dimensionality reduction methods considered here were chosen based on their characteristics and computational costs. PCA is the most commonly used dimensionality reduction technique and has numerous successful applications in statistical process monitoring [19,23]. ICA and KPCA have been successfully adapted to

Please cite this article in press as: M.C. Thomas, et al., Data mining and clustering in chemical process databases for monitoring and knowledge discovery, J. Process Control (2017), http://dx.doi.org/10.1016/j.jprocont.2017.02.006