G Model
JMSY-540; No. of Pages 10

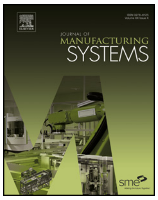
ARTICLE IN PRESS

Journal of Manufacturing Systems xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

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



Technical Paper

Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs

Wei Guo[a], Ashis G. Banerjee[a,b,*]

[a] Department of Industrial & Systems Engineering, University of Washington, Seattle, WA 98195, USA
[b] Department of Mechanical Engineering, University of Washington, Seattle, WA 98195, USA

ARTICLE INFO

Article history:
Received 1 October 2016
Received in revised form 22 February 2017
Accepted 24 February 2017
Available online xxx

Keywords:
Topological data analysis
Feature selection
Yield prediction
Fault detection

ABSTRACT

Topological data analysis (TDA) has emerged as one of the most promising approaches to extract insights from high-dimensional data of varying types such as images, point clouds, and meshes, in an unsupervised manner. To the best of our knowledge, here, we provide the first successful application of TDA in the manufacturing systems domain. We apply a widely used TDA method, known as the Mapper algorithm, on two benchmark data sets for chemical process yield prediction and semiconductor wafer fault detection, respectively. The algorithm yields topological networks that capture the intrinsic clusters and connections among the clusters present in the data sets, which are difficult to detect using traditional methods. We select key process variables or features that impact the system outcomes by analyzing the network shapes. We then use predictive models to evaluate the impact of the selected features. Results show that the models achieve at least the same level of high prediction accuracy as with all the process variables, thereby, providing a way to carry out process monitoring and control in a more cost-effective manner.

© 2017 The Society of Manufacturing Engineers. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Sensors play an essential role in carrying out product feasibility assessment, yield enhancement, and quality control in modern manufacturing systems such as vehicle assembly, microprocessor fabrication, and pharmaceuticals development [1]. A large number of sensors of many different types are typically employed in such systems to measure a variety of process variables ranging from operating conditions and equipment states to material compositions and processing defects over extended time periods. Thus, the volume of acquired data is so vast and heterogeneous that the contribution of individual sensor measurements in predicting the overall system outputs gets obscured. This prediction is made more challenging by the fact that the measurements are often noisy and replete with missing or outlier values. Furthermore, there is significant redundancy among the sensor measurements, leading to the presence of numerous false correlations in the recorded data. It is, therefore, necessary to perform an analysis using statistical methods that are specifically suited to identifying and filtering out existing correlations in erroneous, heterogeneous, and high-dimensional data sets.

Historically, multivariate statistical process control (MSPC) methods, such as principal component analysis (PCA) and partial least-squares (PLS), have served as the dominant mode of addressing this problem [2]. The common idea behind these methods is to define a new set of variables (known as *latent variables*) through linear combinations of the original variables that describe the sensor measurements. The set of latent variables may be reduced in some cases by performing subsequent dimensionality reduction techniques. However, these methods do not work particularly well when there are a large number of input process variables, and they share highly non-linear relationships with the system outputs that cannot be modeled using Gaussian distributions. The methods also encounter difficulties in removing the false correlations among the measurements particularly when they are erroneous. More recently, several non-linear prediction methods have been developed based on response surface fitting as well as kernelized and robust variants of the MSPC techniques [3,4]. While these methods may achieve high prediction accuracy, they do not provide any direct way of quantifying the contribution or impact of the individual process variables.

Here, we present an alternative method that leverages the emerging topic area of topological data analysis (TDA) [5] to select the important variables that are subsequently used in both linear and non-linear prediction models. More specifically, we employ a well-established TDA method known as the Mapper algorithm developed by Singh et al. [6]. It is based on the core idea of understanding the unknown topology of the high-dimensional manifold

[*] Corresponding author at: Department of Industrial & Systems Engineering, Department of Mechanical Engineering, University of Washington, Seattle, WA 98195, USA.
E-mail addresses: weig@uw.edu (W. Guo), ashisb@uw.edu (A.G. Banerjee).

http://dx.doi.org/10.1016/j.jmsy.2017.02.015
0278-6125/© 2017 The Society of Manufacturing Engineers. Published by Elsevier Ltd. All rights reserved.

Please cite this article in press as: Guo W, Banerjee AG. Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs. J Manuf Syst (2017), http://dx.doi.org/10.1016/j.jmsy.2017.02.015

in which the data resides to extract hidden patterns. In particular, it clusters all the level sets of the data (defined using a projection of the high dimensional data to a lower dimensional space) to generate a topological network that represents the inherent clusters and connections among the clusters in the actual data.

This Mapper algorithm has already enjoyed immense popularity in fields such as bioinformatics and machine vision. For example, it has been used to reveal unique and subtle aspects of the folding patterns of RNA [7] and to unlock previously unidentified relationships in immune cell reactivity between patients with type-1 and type-2 diabetes [8]. Another influential example occurs in personalized breast cancer diagnosis, in which a novel subgroup of tumors with a unique mutational profile and 100% survival rate has been discovered [9]. Additionally, its deformation invariant property has been used to detect 3D objects from point cloud data with intrinsically different shapes [6].

Despite the potential of TDA in general and the Mapper algorithm in particular, there has been no prior application in the manufacturing domain to the best of our knowledge. Inspired by the success in biomedical and vision problems, we employ the Mapper algorithm and show that it facilitates the analysis of the impact of each process variable on system outputs through direct visualization. It also determines whether particular subgroups of the data are selectively responsive to different process variables, which helps to monitor and diagnose processes effectively.

We first apply the Mapper algorithm on a benchmark chemical processing data set to predict product yield [10]. Specifically, the shape of the generated topological network is used to select key features that explain the observed differences in the process measurements in a statistically significant manner. Second, we investigate the role of individual process variables in causing wafer failures in another publicly available semiconductor manufacturing data set. Although it has been recognized that k-nearest neighbor methods can identify faulty wafers effectively [11–14], the actual process variables that result in the wafer anomalies have never been identified. To this end, we demonstrate how the Mapper algorithm rapidly traces the causality hidden in this high-dimensional data set.

The rest of the paper is organized as follows. Section 2 gives an overview of the general characteristics of manufacturing data and the types of predictor (feature) and response variables that are of interest to us. In Section 3, we review the Mapper algorithm and its application in feature selection. We demonstrate the applicability of the Mapper algorithm for feature selection on two benchmark manufacturing data sets in Section 4. The effectiveness of the selected features is further assessed through predictive models. We conclude the paper with remarks and future research topics.

2. Problem formulation

In real-world manufacturing systems, data is collected using a large number of sensors that are affixed to or embedded within different machines and equipment, resulting in a high-dimensional body of heterogeneous data. The data is usually in the form of *time series measurements* of different process variables such as temperature, pressure, density, humidity, voltage, chemical or material composition including the relative proportions of various constituents of alloys or mixtures, material removal or deposition rate, number and severity of processed part flaws and defects, and so on. The sensors, thus, come in myriad forms ranging from thermocouples, pressure gauges, hydrometers, hygrometers, and voltmeters to optical cameras, spectrometers, laser scanners, and ultrasonic transducers.

Consequently, manufacturing sensor data is prone to noise terms, missing values, and outliers. These measurement errors depend on the sensitivity of the sensors to the operating conditions based on their underlying physical principles of actions. For example, it is not at all uncommon for temporary sensor hardware malfunction to result in missing values. A further problem is that of co-linearity, which is usually caused by partial redundancy in the sensor arrangement such as the placement of multiple sensors in close proximity to one another. The net result of these complications is that manufacturing systems are often “data-rich but information-poor”.

Consequently, there is a strong need to effectively select a minimal number of process variables that primarily affect the output variables of interest such as product quality and yield of a manufacturing system comprising several processes of varying types. As discussed earlier in Section 1, this form of selection facilitates process monitoring and diagnostics through targeted sensor data acquisition, storage, and processing. Even if it is cheap or convenient to manage data from all the sensors, knowing which measurements of what variables matter the most makes it feasible to rapidly regulate out-of-control processes or adapt them to manufacture high quality products at desired rates.

To formulate the problem mathematically, we suppose there are m process variables (features) and N sensor measurements recorded at different time instants. Each measurement is, thus, represented by an m-dimensional vector $\mathbf{x}_i \in \mathbb{R}^m$, $i = 1, 2, \ldots, N$. The data is then assembled into a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N\times m}$. Each column denotes a process variable, which is measured by one sensor operating alone or by the concurrent operation of several sensors that function in unison. The latter case is known as data fusion [15], which provides a wide range of sensed parameters, and is, hence, more reliable for data analysis.

In a batch process with batch length L, a 3-D data array $\bar{\mathbf{X}} \in \mathbb{R}^{N\times m\times L}$ is often unfolded batch-wise into a 2-D matrix $\mathbf{X} \in \mathbb{R}^{N\times mL}$. In this case, each measurement $\mathbf{x}_i \in \mathbb{R}^{mL}$ is a batch and each process variable is measured L times throughout the batch, hence, corresponding to L columns. For each row, the measurement is either spatially-sampled or temporally-sampled. For instance, in the semiconductor manufacturing environment, electronic wafer map data collected from in-line measurements are sampled spatially across the surface of the wafer for defect inspection [16]. Usually, there will also be one or more response variables to reflect the output quality or quantity. We write the output with r response variables into a matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]^T \in \mathbb{R}^{N\times r}$, where each response variable is represented by one column. Response variables are commonly seen as continuous variables denoting production yields or binary variables indicating pass or fail outcomes.

3. Technical approach

We now present the framework of the Mapper algorithm and outline the typical pipeline of feature selection using the Mapper algorithm. For more details about the Mapper algorithm and concrete examples of real applications, we refer the reader to [6,17].

3.1. Mapper algorithm

The Mapper algorithm can be considered as a partial clustering algorithm inspired by the classical discrete Morse theory [6]. In topology, discrete Morse theory enables one to characterize the topology of high dimensional data via some functional level sets [18]. More specifically, given a topological space $\mathcal{X}$, when $h : \mathcal{X} \to \mathbb{R}$ is a smooth real-valued function (Morse function), topological information of $\mathcal{X}$ is inferred from the level sets $h^{-1}(c)$ for some real c.

Please cite this article in press as: Guo W, Banerjee AG. Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs. J Manuf Syst (2017), http://dx.doi.org/10.1016/j.jmsy.2017.02.015

Download English Version:

https://daneshyari.com/en/article/5469498

Download Persian Version:

https://daneshyari.com/article/5469498

Daneshyari.com