Chemometrics and Intelligent Laboratory Systems 151 (2016) 201–218

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Enhancing quality of statistic monitoring models by training set design with active learning approach

Zhengbing Yan [a,b], Junghui Chen [b,*]

[a] College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou, 325035, China
[b] Department of Chemical Engineering, Chung Yuan Christian University, Chungli, Taoyuan, Taiwan, 32023, R.O.C.

ARTICLE INFO

Article history:
Received 19 July 2015
Received in revised form 19 December 2015
Accepted 21 December 2015
Available online 30 December 2015

Keywords:
Active learning
Gaussian process model
Latent structure statistical model
Process monitoring

ABSTRACT

The objective of this paper is to enhance the quality of the process monitoring models by designing a training set through an active learning approach. Although conventional process monitoring models are effective in many manufacturing processes, these models falter when confronted by a set of training data with poor quality or a small volume of training data. As the limitations of the monitoring models become increasingly obvious in face of even more complex manufacturing processes, in this work, the active learning process monitoring (AL-PM) model is developed. To design a good training set, Gaussian process (GP) models are first used to construct the relationships between the score variables of the latent structure model and the designable process variables because the GP model is capable of providing the accurate predictive mean and variance. The variance can quantify its prediction uncertainty. Second, the uncertainty index is presented and utilized to adequately explore for which regions the new data samples should be used to enhance the quality of the monitoring model. The proposed AL-PM model can be applied to any types of latent structure-based monitoring models. Its effectiveness and promising results have been demonstrated by its applications to a numerical example and a penicillin benchmark process.

1. Introduction

Process monitoring is one of the key techniques that improves the competitiveness in industries. The objective of process monitoring is to early detect and reduce the process variability in delays, process upsets, equipment malfunctions, improper operation procedures, incorrect estimates, etc., for quality control of an operating process because these variabilities can cause deterioration of products [1]. Process monitoring methodology is developed based on control charts. Control charts are the graphical tools for continuously monitoring a process in order to maintain the process in control [2]. Univariate control charts were first developed by Shewhart based on the concept of statistical hypothesis testing. In practice, a process usually involves a number of process variables correlated with each other. Monitoring a multivariate process using several univariate control charts would be inefficient. Conventional multivariate control charting procedures are reasonably effective as long as the number of variables to be monitored is not very large. However, as the number of variables increase, detecting and interpreting out-of-control situations becomes more difficult. To overcome this weakness, a number of methods addressing multivariate statistical process control (MSPC) have been developed. These methods are often referred to as projection methods (or latent structure methods). The basic idea of projection methods is that a highly dimensional space, spanned by a number of measured variables, is projected onto a model space of fewer dimensions [3]. The model space is spanned by linear or nonlinear combinations of the original variables so that "new fused" variables can be formed. The new fused variables are often referred to as principal components or latent variables. The main purpose of MSPC is to reduce the original variables to fewer latent variables independent of each other and make them sufficient to characterize the information contained in the data.

As discussed before, the different control charts developed based on different characteristics of the operating process are routinely adapted. In general, there are two steps to construct MSPC models. In Step I, the in-control data are collected from the operating process, and in Step II, the control charts are built upon the collected data. Then they are used for continuous monitoring of the process over time. In Step I, in-control data are picked up from the historical data in an existing process. The trial procedures would be performed repeatedly until clean in-control data are obtained. In Step II, control charts are constructed by the clean in-control data set obtained in Step I; then they are used to correctly and quickly detect out-of-control observations and to keep the process in control. Most MSPC models and control charts were developed for Step II applications in the past, and relatively few attempts have been made to improve Step I applications. Thus, in most applications, it is assumed that high volume data with good quality are available in industrial processes. This pragmatic and straightforward approach is used in many applications although it is not ideal. In practice, the number of data for Step 1 would be either large or small.

* Corresponding author. Tel.: +886 3 2654107; fax: +886 3 2654199.
E-mail address: jason@wavenet.cycu.edu.tw (J. Chen).

http://dx.doi.org/10.1016/j.chemolab.2015.12.016
0169-7439/© 2016 Elsevier B.V. All rights reserved.

When big data are available in the database, the designs based on the whole data set are usually viewed by practitioners as monolithic and inflexible. As a result, the computational load is not reduced. The ability to construct the MSPC model based on the large data set is limited to the capacity of the computer. Difficulties occur in the analysis of the data sets with very large observations. Reduction of the size of the training set is necessary if the observations are huge. The size of the data selected to correctly estimate the parameters of an MSPC model would vary from one problem to another. The process of the sample reduction is also called sample selection. Many sample selection methods have been developed, such as probability sample selection and nonprobability sample selection. In probability sample selection methods, every sample in the population has a chance of being selected, and this probability can be accurately determined. Several methods that determine the probability have been proposed for different applications, such as simple random, systematic random, stratified random, and cluster random methods [4]. Nonprobability sample selection methods are based on the population of interest. The concept of the condensed nearest neighbor algorithm and its extensions [5,6] is based on the predictive performance or the error of a model. The pattern by the ordered projections method [7] is used to select only some border samples and eliminate the samples that are not on the boundaries of the regions where they belong. Generalized-modified Chang's algorithm [8] split an original data set into several clusters and then centers of the clusters were selected. The best known uniform selection in chemometrics is the Kennard–Stone (KS) algorithm [9]. It maximizes the minimal Euclidean distances between the selected samples and the remaining samples. It selects a subset of samples uniformly distributed in the sample space. However, although the sample selection methods can select the good representative samples from the historical data set, the MPSC model based on the samples still cannot represent the current operating process well because there may not be enough favorable samples in the current historical set to describe the required process characteristics.

In reality, the story is quite different for some manufacturers in polymer, pharmaceutical, semiconductor, and biochemical industries though. Data volumes are low, and the branch of knowledge is extremely high-tech. On-line quality measurements, like infrared analyzers, ultraviolet, and visible-radiation analyzers, are generally much more expensive. Because of their big costs and inspecting time, the use of the on-line analyzer is normally decided based on process economics. Thus, product qualities are often analyzed off-line and infrequently. In this situation, the collected data may not contain the entire process characteristics. The statistical monitoring models constructed directly upon the collected data cannot function well. The detections using the established monitoring models are likely to get false alarms.

Traditionally, trial-and-error approaches were used to find new samples from the database. If data were not available, it was further required to design and execute the experiment. The trial-and-error methods were time consuming, expensive, and even self-defeating. The samples are expected to provide the maximum amount of information so that the control charts can be constructed effectively. To overcome the problem of the traditional methods, an active learning strategy should be proposed to obtain efficient samples. It has been proved useful in the planning of experiments. Such an approach, in fact, has received considerable attention from the engineering and statistic communities for its advantages of flexibility and adaptability. For example, the efficient global optimization algorithm was used to derive sequential designs for the optimization of deterministic simulation models. It chose the data points at each step to maximize the improvement [10]. An adaptive strategy based on an explicit trade-off between reduction in global uncertainty and exploration of region of interest was proposed to accurately approximate the target region [11]. Among these designs of experimental approaches, Bayesian inference attracted more attention. A Bayesian approach was used to derive sequentially integrated mean square error designs [12]. Bayesian inference based on the probability theory treats variables as stochastic variables. That is, if the predictive distribution at a testing sample point is tightly packed, the model quality at this point is high; on the other hand, the prediction distribution spreading widely over a range of values indicates that the model quality at this point is highly uncertain. Thus, to enhance the model quality, Bayesian inference can be used to select data in the region of high uncertainty.

In this paper, the active learning process monitoring (AL-PM) model is proposed. This model integrates the conventional MSPC with a supervised Gaussian process (GP) model. The rest of this paper is organized as follows. Section 2 revisits the conventional MSPC models with the latent structure and gives a uniform expression of some most popular models. A numerical example is used to describe the limitations when the training data are directly applied to MSPC models. In Section 3, the concept of active learning is first introduced, and the GP model definition and its training method are reviewed. Then the AL-PM model is proposed. The GP model is used to construct the relationships between the score variables of the latent structure model and the designable process variables. The uncertainty index utilizing prediction uncertainty of the GP model can adequately explore for which regions the new data samples should be used to enhance the quality of the monitoring model. The results of a numerical example and a penicillin benchmark process are presented and discussed in Section 4. In Section 5, concluding remarks are made.

2. Revisit of the latent structure-based statistical monitoring models and their characteristics

The basis of the MSPC methods is to simplify a monitoring system into a latent structure method by projecting data onto the lower dimensional subspace and then further analysis would be made. The dimensionality reduction or the decoupling process is a natural way in multivariate process monitoring. Several optimal dimensionality reduction techniques for latent structure models have been heavily studied and applied to chemical process monitoring over the past two decades [13,14]. The latent structure method leverages the interaction between the variables and the monitoring changes in the correlation structure of the variable. The identification of a latent structure method involves finding the fused variables that best describe the major features of the data set constituted by the measured variables, and the fused variables span only the true dimension of the process. Thus, correlated data are not a difficulty but a necessity for latent structure methods. The investigation of sample covariance matrices ($\mathbf{A}$) can be unified as

$$\mathbf{A} = \mathbf{T}\mathbf{P}^T + \tilde{\mathbf{T}}\tilde{\mathbf{P}}^T \tag{1}$$

and $\mathbf{A} \in \mathfrak{R}^{N \times M}$ is a transformation of $\mathbf{X}$,

$$\mathbf{A} = \begin{cases} \mathbf{X} & \text{for PCA, PPCA, ICA} \\ \overline{\mathbf{K}} & \quad\text{for KPCA} \end{cases} \tag{2}$$

where $\mathbf{X} \in \mathfrak{R}^{N \times M}$ is a sampling matrix in which N is number of samples and M is the number of variables. Assume that $\mathbf{X}$ is scaled to zero mean. PCA is mainly used in the processes with linear and steady state. PPCA models are the probabilistic formulation of PCA models, which provide a single statistic for fault detection [15,16]. Independent component analysis (ICA) decomposes observed data into linear combinations of statistically independent components, which can characterize non-Gaussian characteristics of processes [17,18]. For nonlinear processes, kernel-based PCA (KPCA) can be used to improve the nonlinear ability of normal PCA through using the kernel function [19]. The basic idea of KPCA is to first map the nonlinear input space into a linear feature space and then to compute PCA in that feature space [20]. In Eq. (2), $\overline{\mathbf{K}} = \mathbf{K} - \mathbf{KE} - \mathbf{EK} + \mathbf{EKE}$ and $\mathbf{E}_{ij} = 1/N$. $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is the entry in the kernel matrix $\mathbf{K}$. The application of MSPC in monitoring batch process data was also developed, like multiway PCA (MPCA) [21], multiple-subspace PCA [22], etc.