Chemometrics and Intelligent Laboratory Systems 137 (2014) 128–139

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Batch process monitoring based on just-in-time learning and multiple-subspace principal component analysis

Zhaomin Lv, Xuefeng Yan *, Qingchao Jiang

Key Laboratory of Advanced Control and Optimization for Chemical Processes of Ministry of Education, East China University of Science and Technology, Shanghai 200237, PR China

ARTICLE INFO

Article history:
Received 22 April 2014
Received in revised form 12 June 2014
Accepted 16 June 2014
Available online 24 June 2014

Keywords:
Batch process monitoring
Just in time learning
Multiple subspace
Principal component analysis

ABSTRACT

Batch or fed-batch process monitoring is a challenging task because of its characteristics such as batch-to-batch variations, inherent time-varying dynamics, and multiple operating phases. Thus, a new batch process monitoring method based on just-in-time learning (JITL) and multiple-subspace principal component analysis (MSPCA) is developed. Based on offline one batch normal data, the division algorithm of multiple subspace is proposed, in which mutual information (MI) and K-means are employed to derive the segmentation rule of variable subspace and then the variables are divided into several subspaces according to the segmentation rule of variable subspace. At online monitoring, the training data set for modeling is obtained by JITL and separated into each subspace according to the segmentation rule of variable subspace. Principal component analysis is employed to build the model in each subspace, and all components are retained to calculate T^2 statistics. A unique probability index is obtained by Bayesian inference (BI) as the decision fusion strategy of T^2 statistics of all subspaces. A simple numerical example is used to show the advantages of the proposed MSPCA method. The feasibility and effectiveness of JITL–MSPCA is demonstrated by fed-batch penicillin fermentation.

1. Introduction

Batch and fed-batch processes play an important role in the production of high-quality, low-volume products, such as special chemicals, food, pharmaceutical, and semiconductor. Quality consistency of products and safe operation of batch process must be ensured. The characteristics of batch processes, such as finite duration, batch-to-batch variations, inherent time-varying dynamics, and multiple operating phases, differ from that of a continuous process. Hence, traditional continuous process monitoring methods cannot be used in batch process monitoring, and such characteristics of batch processes make it difficult to establish a monitoring model for them.

Multi-way principal component analysis (MPCA) and multi-way partial least squares (MPLS) are two of the most popular methods used for batch process monitoring [1–4]. Conventional multivariate statistical process monitoring methods based on MPCA or MPLS rely on the assumption that batch process data come from a single operating phase. However, batch processes are usually conducted in a sequence of steps, which are called multiple operating phases. Significantly diverse variable correlation structures exist at different phases. Traditional MPCA and MPLS methods consequently fail to detect fault in batch processes. Numerous stage-based approaches have been developed to deal with the multiple operating phases of batch processes [5–8]. However, the number of operational stages needs to be specified by user, and a biased estimation may affect the monitoring accuracy.

Aside from the multiple-operating phase problem, the inherent time-varying dynamics of batch process has posed difficulties to batch process monitoring methods. An adaptive methodology, which mainly includes recursive strategy and moving window (MW) technique, is generally integrated in various monitoring methods to trace the inherent time-varying dynamics [9–12]. Lee used MPCA with variable-wise unfolding and time-varying score covariance structures to preserve the dynamic relations of data [9]. Rännar proposed recursive hierarchical adaptive principal component analysis (HPCA) algorithm [10]. Both methods use local data structures, which can solve the inherent time-varying dynamics and multiple-operating phase problems. However, local data structures require even phase durations in batch process data. Given that phase durations differ from batch to batch, the two monitoring methods may not function well. As an alternative solution, just-in-time learning (JITL) strategy has attracted increasing attention in soft sensor modeling and process monitoring fields [13–19]. Cheng [14] and Ge [15] applied multivariate statistical methods to analyze the residuals between predicted outputs and process outputs of JITL. Hu [19] integrated an MW strategy with JITL to analyze the selected relevant samples for process monitoring directly. Unlike traditional local modeling methods, which are built offline, JITL-based local models are constructed online, which only choose the most relevant points and use them for modeling. Therefore, the current state

* Corresponding author at: P.O. BOX 293, MeiLong Road NO. 130, Shanghai 200237, PR China. Tel.: +86 21 64251036.
E-mail address: xfyan@ecust.edu.cn (X. Yan).

http://dx.doi.org/10.1016/j.chemolab.2014.06.010
0169-7439/

of the process can be well tracked by the JITL method. Meanwhile, the inherent time-varying dynamics and uneven length phase duration problems in batch processes can be solved.

Multiple-variable subspace methods that focus on continuous processes have been introduced in recent years [20–26]. Ge proposed a linear subspace method for nonlinear process monitoring [24]. PCA decomposition method was used for linear subspace construction, and the original process variables were divided into $k + 1$ overlapped variable subspace, where k is the number of retained principal components. Tong proposed a four-subspace construction method [26], which uses PCA to derive four distinct and explicable subspaces from the original process variables; each subspace serves as a low-dimensional representation of the original data space. Multiple-variable subspace methods can reduce the complexity of process analysis. However, the characteristics of batch process are distinctly different from those of continuous processes, such as their inherent time-varying dynamics and multiple operating phases. Therefore, further application of multiple-subspace methods is hindered to batch processes.

In this study, JITL is introduced into multiple-subspace principal component analysis (MSPCA) for batch process monitoring. In comparison to the strategies of Tong and Ge, which use PCA to derive multiple-variable subspace, a new division algorithm of multiple subspace is proposed. A relevant matrix is defined according to the mutual information (MI) of any two variables based on the normal data of batch process. According to the clustering result of the relevant matrix by K-means clustering algorithm [27], a variable subspace segmentation rule is developed, and the variables are divided into several subspaces (two subspaces are employed in this study.). At online monitoring, relevant samples are selected from the historical data set by JITL, which can well track the state of the process; thus, the inherent time-varying dynamics and multiple operating phases can be effectively handled. The relevant samples are separated into each variable subspace according to the variable subspace segmentation rule, with each variable subspace serving as a low-dimensional representation of the original data space. PCA is employed to build the model in each variable subspace, and all components are retained to calculate T^2 statistics. Thus, all local information can be sufficiently utilized. Bayesian inference (BI) strategy is used to combine T^2 statistics generated from each subspace. By integrating JITL and MSPCA, the proposed batch process monitoring method, i.e., JITL–MSPCA, can accurately detect different types of batch process faults.

The rest of this paper is divided into several sections. Section 2 presents PCA, MPCA [9], and HPCA [10]. Section 3 elucidates JITL–MSPCA and the monitoring procedure. Section 4 discusses the numerical process used to show the advantage of the proposed MSPCA method and the fed-batch penicillin fermentation employed to demonstrate the feasibility and effectiveness of JITL–MSPCA. Finally, Section 5 concludes.

2. Preliminaries

This section reviews PCA, MPCA [9], and HPCA [10] methods for process monitoring. PCA is compared with the proposed MSPCA method in a simple numerical process. MPCA and HPCA methods are compared with JITL–MSPCA in fed-batch penicillin fermentation.

2.1. PCA

Considering the matrix $\boldsymbol{X}(m \times n)$ with m samples and n variables, PCA is defined as follows:

$$\boldsymbol{X} = \boldsymbol{T}\hat{\boldsymbol{P}}^T + \boldsymbol{E}, \tag{1}$$

where $\hat{\boldsymbol{P}}(n \times z)$ is the loading matrix, $\boldsymbol{T}(m \times z)$ is the score matrix, $\boldsymbol{E}(m \times n)$ is the residual matrix, and z is the number of principal components. $\hat{\boldsymbol{P}}(n \times z)$ can be obtained from the singular value decomposition of the covariance matrix $\boldsymbol{C}$, which is expressed as follows:

$$\boldsymbol{C} = \frac{\boldsymbol{X}^T\boldsymbol{X}}{m-1} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T, \tag{2}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2 \cdots, \lambda_J)$ is the eigenvalue matrix. Eigenvalues are arranged in descending order. $\boldsymbol{P}$ is separated into two parts, i.e., $\hat{\boldsymbol{P}}(n \times z)$ and $\tilde{\boldsymbol{P}}(n \times (n-z))$, which are called PCS and RS, respectively.

T^2 and SPE are two statistics for online monitoring that correspond to PCS and RS, respectively. They can respectively be calculated as follows:

$$T^2 = \boldsymbol{x}\hat{\boldsymbol{P}}\boldsymbol{\Lambda}_z^{-1}\hat{\boldsymbol{P}}^T\boldsymbol{x}^T \tag{3}$$

and

$$SPE = \boldsymbol{e}\boldsymbol{e}^T; \quad \boldsymbol{e} = \boldsymbol{x} - \boldsymbol{x}\hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^T, \tag{4}$$

where $\boldsymbol{\Lambda}_z = diag(\lambda_1, \lambda_2 \cdots, \lambda_z)$, and $x(1 \times n)$ is the monitored sample. Proper control limits are defined to determine whether the process is operated under normal conditions as follows:

$$T^2 \le \frac{z\left(m^2-1\right)}{m(m-z)} F_{z,(m-z),\alpha} \tag{5}$$

and

$$SPE_\alpha \le a\chi^2_{b,\alpha}; \quad a = \frac{v}{2h}, b = \frac{2h^2}{v}, \tag{6}$$

where $F_z(m - z)$, α is F-distribution with z and $m - z$ degrees of freedom under confidence limit α; h and v are the estimated mean and variance of the squared prediction error from the modeling data, respectively.

2.2. MPCA [9]

The dataset of a batch process is a three-way array, i.e., $\ddot{\boldsymbol{X}}(I \times J \times K)$, where I represents the number of batches, J denotes the number of measurement variables, and K is the number of sampling instants within each batch. Combination of batch-wise and variable-wise unfolding is

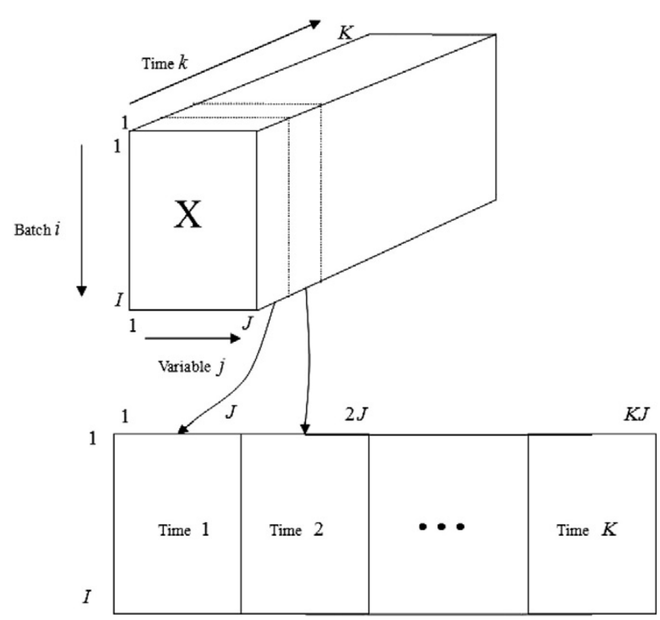

Fig. 1. Batch-wise unfolding.