Control Engineering Practice 31 (2014) 9–23

Contents lists available at ScienceDirect

Control Engineering Practice

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

Quality prediction and analysis for large-scale processes based on multi-level principal component modeling strategy

Zhiqiang Ge *

State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027, Zhejiang, PR China

ARTICLE INFO

Article history:
Received 12 February 2014
Accepted 22 June 2014
Available online 16 July 2014

Keywords:
Large-scale processes
Multi-level modeling
Principal component regression
Quality prediction
Process analysis

ABSTRACT

This paper proposed a multi-level principal component regression (PCR) modeling strategy for quality prediction and analysis of large-scale processes. Based on decomposition of the large data matrix, the first level PCR model divides the process into different sub-blocks through uncorrelated principal component directions, with a related index defined for determination of variables in each sub-block. In the second level, a PCR model is developed for local quality prediction in each sub-block. Subsequently, the third level PCR model is constructed to combine the local prediction results in different sub-blocks. For process analysis, a sub-block contribution index is defined to identify the critical-to-quality sub-blocks, based on which an inside sub-block contribution index is further defined for determination of the key variables in each sub-block. As a result, correlations between process variables and quality variables can be successfully constructed. A case study on Tennessee Eastman (TE) benchmark process is provided for performance evaluation.

1. Introduction

It is well-known that some important quality variables in chemical processes are difficult to measure online. Instead, those variables are often determined by offline analyses in the laboratory or by some online analyzers. However, both offline analyses and online analyzers are expensive and time-consuming. To this end, an indirect method named as inference measurement or soft sensing has been proposed to predict the key variables by using some easy-to-measure secondary process variables (Joseph & Brosilow, 1978; Kresta, Marlin, & MacGregor, 1994; Tham, Morris, & Montague, 1991). Generally, these methods can be categorized into two types: the model-based method and the data based method. Compared to the model based method which strongly relies on the first-principle model of the process, the data-based method has little dependence on process knowledge. In the past years, data-based methods for quality prediction and analysis have received much attention in the chemical engineering area (Fortuna, Graziani, Rizzo, & Xibilia, 2007; Ge, Song, & Gao, 2013; Kadlec, Gabrys, & Strandt, 2009; Kano & Nakagawa, 2008; Khatibisepehr, Huang, & Khare, 2013). Many data-based methods have been developed for quality prediction or soft sensing purpose, such as principal component regression (PCR), partial least squares (PLS), multiway PLS, multiblock principal component analysis and PLS, fuzzy clustering and analysis methods, artificial neural network (ANN), support vector machine (SVM), etc. (Ge, Chen, & Song, 2011; Ge, Song, & Kano, 2014; Gonzaga, Meleiro, Kiang, & Filho, 2009; Hartnett, Lightbody, & Irwin, 1998; Jin, Wang, Huang, & Forbes, 2012; Khatibisepehr, Huang, Xu, & Espejo, 2012; Kruger, Antory, Hahn, Irwin, & McCullough, 2005; Kruger, Chen, Sandoz, & McFarlane, 2001; Lee, Joung, Lee, Park, & Woo, 2005; Yu, 2012; Zhang & Zhang, 2009).

However, due to the complexity of modern chemical processes, some of which may have a large number of process variables. For those large-scale processes, existing soft sensing methods may lack of interpreting the relationship between the quality variable and various process variables. For quality prediction and analysis of those large-scale processes, it is of particular interest and important to know which variables are critical to the final product quality, which are not, and to what extent they influence the product quality. However, with a large number of process variables, detailed process analyses are difficult to carry out. Actually, these process variables often come from different units of the process, which may have different impacts to the quality variables. In other words, different units may provide distinct effects on different quality variables. Therefore, if we can identify the critical units and their crucial variables for the corresponding quality variable, both of the control and optimization of this production

* Tel.: +86 87951442.
E-mail addresses: gezhiqiang@zju.edu.cn, gezhiqiang@gmail.com

http://dx.doi.org/10.1016/j.conengprac.2014.06.006
0967-0661/

process can be improved. Another interesting issue is how to combine the local prediction results in different units. If the quality variable is critical to a single unit, the problem is straightforward. However, if several different units have been identified as the critical units to the quality variable, they should be combined through an appropriate way, otherwise, a meaningless prediction result may be generated.

Among all of the developed data-based methods in last several decades, the multiblock based methods such as multiblock PCA, multiblock PLS has been considered as an efficient way to model the process with a large number of variables. These methods divided the process variable set into several meaningful sub-blocks, based on which different multiblock statistical models can be developed (Bras, Bernardino, Lopes, & Menezes, 2005; Cherry & Qin, 2006; Choi & Lee, 2005; Ge, 2014; Ge & Song, 2009; Jing, Cai, & Shao, 2010; Kohonen, Reinikainen, & Hoskuldsson, 2009; MacGregor, Jaeckle, Kiparissides, & Kourtoudi, 1994; Qin, Valle, & Piovoso, 2001; Smilde, Westerhuis, & Sijmen de, 2003; Westerhuis, Kourti, & MacGregor, 1998). MacGregor et al. (1994) developed monitoring and diagnosis charts for each sub-block as well as a global monitoring chart for performance enhancement. Westerhuis et al. (1998) provided a comprehensive analysis of several multiblock and hierarchical PCA and PLS algorithms, and give those algorithms in a unified notation. Qin et al. (2001) also gave a detail analysis of several multiblock and hierarchical PCA and PLS algorithms. A framework for sequential multiblock component methods has been proposed, and several sequential multiblock methods have been reviewed (Smilde et al., 2003). Several multiblock PLS models have also been proposed for modeling of spectra datasets (Cherry & Qin, 2006; Choi & Lee, 2005). However, most of these multiblock methods need some process knowledge for block division. Although several automatic block division methods have been developed, such as greedy search based methods, they are very complicated and computationally expensive. Besides, detailed analyses of the relationships between different sub-blocks and the final product quality are ignored by most existing methods.

The objective of this paper is the kind of processes with a large number of variables, which we call large-scale processes here. As mentioned, those variables always come from different units/parts of the process, which may have quite different effects on the quality variable. In order to divide the process into sub-blocks automatically, a principal component decomposition method is proposed. Generally, an effective block division strategy should simultaneously meet the diversity and accuracy requirements. Here, the diversity means the process variables should be divided into different sub-blocks which have as less correlation with each other as possible, and the accuracy means most important variables should be incorporated in their corresponding sub-block. PCA aims to decompose the original signals into different directions, which are uncorrelated with each other. Based on this important characteristic, we intend to divide the process variables into different sub-blocks through these uncorrelated directions. Therefore, the diversity of sub-blocks can be well addressed. On the other hand, if we select the most important variables through each uncorrelated direction, the accuracy of each sub-block can also be obtained. Different from existing block division methods, the principal component decomposition based method is much easier for process interpretations and practice applications.

After the process variables have been divided into different sub-blocks, the relationship between each sub-block and the quality variables should be built. Thus, a sub-PCR model is used for regression purpose in each sub-block, which is named as the second level PCR model in this paper. Therefore, while the first level PCR model divides the process variables into different sub-blocks, the second level PCR model builds the relationship between each sub-block and the quality variables. After that, the relations between different sub-blocks and quality variables can be analyzed, e.g., which sub-block is critical to the considering quality variable, and how it influences the final quality. If more than one critical-to-quality sub-blocks have been identified, a combination strategy should be used to integrate different sub-block prediction results. It is worth to notice that different sub-blocks may have overlapped variables, which means the sub-blocks may be correlated to each other. Therefore, an additional PCR model is used as the combination strategy for regression modeling between local prediction results and the final quality variable, which is named as the third level PCR model in this paper. Another important issue is to determine the key variables in each critical-to-quality sub-block, depending on which further advanced control and quality optimization can be made. This is because the key variables have more significant impact on the quality variable, if these variables are used for control or optimization, more obvious effects and less manipulating efforts can be rendered to the control system.

The main contributions of the present paper can be summarized as: (1) A new automatic block division method is proposed; (2) A multi-level PCR modeling strategy is developed for quality prediction of large-scale processes; (3) Detailed process analyses are provided to explore the relationships of the product quality over different sub-blocks, to identify sub-blocks which are critical to quality, and to determine the key variables in each sub-block. The rest of this paper is organized as follows. In Section 2, the principle of the traditional PCR method is introduced, which is followed by the detailed demonstration of the proposed multi-level PCR modeling strategy in the next section. In Section 4, a benchmark case study on the TE process is provided to evaluate the performance of the proposed method. Finally, conclusions are made.

2. Principal component regression (PCR)

Suppose the measurement matrix can be represented as $\mathbf{X} \in R^{n\times m}$, where n is the number of data sample and m is the number of measured variables. The predicted variable matrix can be given as $\mathbf{Y} \in R^{n\times l}$, where n is the number of data sample and l is the number of predicted variables. The traditional derivation of PCR can be expressed as follows (Hartnett et al., 1998):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^T + \mathbf{F} \tag{2}$$

where $\mathbf{P} \in R^{m\times k}$ is the loading matrix, $\mathbf{T} \in R^{n\times k}$ is the principal component matrix, k is the selected number of principal components, which can be selected by the cumulative percentage variance (CPV) method, $\mathbf{C} \in R^{l\times k}$ is the another loading matrix corresponding to the quality variable matrix, $\mathbf{E}$ and $\mathbf{F}$ are the residual matrices with appropriate dimensions. For a new data sample $\mathbf{x}_{new} \in R^{m\times 1}$, the predicted variables can be calculated as

$$\hat{\mathbf{y}}_{new} = \mathbf{C}\mathbf{P}^T\mathbf{x}_{new} \tag{3}$$

3. Multi-level PCR modeling strategy for quality prediction and analysis of large-scale processes

In this section, a detailed description of the proposed method is given. First, the automatic block division method is proposed based on the principal component decomposition. Second, multiple local PCR models are constructed in different sub-blocks for quality prediction. Third, an additional PCR model is developed for combination of local quality prediction results generated in