Journal of Process Control 42 (2016) 90–103

Contents lists available at ScienceDirect

Journal of Process Control

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

Developing a soft sensor with online variable reselection for unobserved multi-mode operations

Jialin Liu*

Department of Chemical and Materials Engineering, Tunghai University, No. 1727, Sec.4, Taiwan Boulevard, Taichung, Taiwan, ROC

ARTICLE INFO

Article history:
Received 2 September 2015
Received in revised form 17 January 2016
Accepted 31 March 2016
Available online 25 April 2016

Keywords:
Soft sensors
Multi-mode operations
Online variable reselection
Partial least squares

ABSTRACT

Soft sensors are used to predict response variables, as these variables are difficult to measure, the prediction models use data of predictors that are relatively easier to obtain. Arranging time-lagged data of predictors and applying the partial least squares (PLS) method to the dataset is a popular approach for extracting the correlation between data of the responses and predictors of the process dynamic. Because irrelevant inputs deteriorate the prediction performance of the soft sensor, the selection of variables in the PLS-based model is a critical step for developing a robust and accurate model. Furthermore, it is necessary to reselect the important predictors of a soft sensor when the operating mode is changed. However, a switch in the operating mode may not be measured, directly. In this study, two statistics are proposed to detect a change of operating mode to enable the reselection of the predictors of the soft sensor. This work involved the development of a soft sensor based on operating data from the industrial ethane removal (de-ethane) process. The changeover of crude oil types cannot be observed from the data of process variables; however, the correlation between input and output variables is significantly affected by the different types of crude oil. The result shows that the use of a soft sensor with online variable reselection is capable of maintaining the accuracy and robustness of the inferential model, effectively.

1. Introduction

In industrial processes, operators rely on online analyzers and laboratory tests to adjust manipulated variables to maintain product qualities or exhaust gases within the specifications of the product or government regulations. As online analyzers could malfunction or laboratory testing could result in significant delays, soft sensors that infer the primary output from other process variables provide useful information for regulating process operations. In fact, soft sensor applications have attracted significant attention in the process industry [1]. The PLS algorithm is a popular multivariate statistical tool for modeling input/output data. It has been proven that the maximal covariance between two datasets can be captured by PLS [2]. However, the accuracy of static soft sensors may suffer under dominating process dynamics, i.e. the output variable depends on process variables at some delayed times. Therefore, the dynamic correlations between inputs and outputs need to be considered when developing a reliable soft sensor. A straightforward approach is to extend methods used in univariate time series analysis to the multivariate time series models. In this regard, dynamic PLS (DPLS) has been widely applied in the design of dynamic models for process control [3] and in the development of soft sensors for batch processes [4]. For a continuous process, the input variables of DPLS are obtained using the presented data and some time-lagged predictor data. Because the dimension of the input variables dramatically increases with the order of the modeling time lags, a high-dimensional dataset can easily be formed once several time delays are incorporated. Information criterions, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to select the number of variables to be included in a model [5]; however, exploring all possible lag combinations of process variables is impractical for a chemical process. Because dozens of correlated variables in the process are common, it is difficult to identify the dynamic response for each variable. The high-dimensional dataset often contains data that are irrelevant for predicting the variations of response variables; for example, when the predictor data are previous to the time delays corresponding to the response variables, these data are irrelevant for predicting the current outputs of response variables. As the time delay of each predictor is usually unknown, the

* Corresponding author.
E-mail address: jialin@thu.edu.tw

http://dx.doi.org/10.1016/j.jprocont.2016.03.007
0959-1524/

training dataset of DPLS inevitably contains such irrelevant data. Moreover, even if their contribution to the model is small, the prediction performance can be deteriorated by these irrelevant predictor data.

In a review paper [6], the methods for PLS variable selection were classified into three categories: filter methods, wrapper methods, and embedded methods. Filter methods use some indices with corresponding thresholds to filter out irrelevant predictors; for example, the regression coefficients (PLS-Beta) and variable importance in projection (PLS-VIP) are two popular indices. However, the thresholds of the indices are decided using cross-validation in most cases; i.e., subsets of variables are selected according to the different values of the threshold. The procedure for determining the threshold turns filter methods into wrapper methods. Wrapper methods use a search algorithm to extract the subsets of variables and evaluate each subset by fitting a model to the subset variable. For example, the genetic algorithm with PLS (GA-PLS) [7] is a randomized search algorithm in wrapper methods. Because the number of subsets exponentially increases with the number of variables, it is impractical to evaluate all possible subsets. Embedded methods integrate the variable selection into the modeling step; therefore, by trimming the irrelevant variables, they constitute more efficient ways to build a prediction model, contrary to their counterparts. For example, Chun and Keleş [8] reformulated the object function of PLS to find the weighting matrix **w** to maximize the covariance between predictors and responses by introducing penalty terms to shrink the elements in **w**. This procedure was named sparse PLS (SPLS).

For an industrial process, the reliable soft sensor needs to be adapted to accommodate the time-varying nature of the process. Qin [9] proposed a block-wise recursive PLS (RPLS) for adapting the inferential model. Although RPLS accounts for the time-varying nature of processes by updating models with the newest data, it leads to a reduction in the speed of adaptation as the amount of data increases. The moving window algorithm is an alternative approach to exclude the oldest data once new data become available. Qin [9] reported the computational loading of the moving window PLS is proportional to the window size. Liu et al. [10] proposed a fast-moving window algorithm to adapt the PLS model for predicting the outputs of response variables for the time-varying process, named fast-moving window PLS (FMWPLS). In their approach, the computational loading is independent of the window size, which is more practicable when updating models online.

On the other hand, the characteristics of a time-varying process can be modeled using local modeling methods, such as the just-in-time (JIT) learning technique [11]. A local model was built based on the data that were collected according to the distances of the data points to the query data under the predefined threshold. Fujiwara et al. [12] pointed out that the prediction performance of the JIT model is not always high because the variable correlation was not taken into account. They also maintained that a good model cannot be obtained using data based on a weak correlation among input-output variables, even though the distances among samples are sufficiently close. Therefore, they developed the correlation-based just-in-time (CoJIT) [12] method to collect modeling data by considering the correlation between variables. However, the index they derived to measure variable correlation only used predictor data, i.e., their use of the CoJIT approach only considered the correlation among input variables, whereas the correlation between input and output variables was omitted. More recently, Kaneko and Funatsu [13] developed an adaptive soft sensor based on a database in which only informative data was stored and proposed a database-monitoring index (DMI) to measure the similarity between two data points. Their sensor relied on the DMI being large when two data points are dissimilar, and vice versa. In addition, should the DMI of a new sample exceed the predefined DMI threshold, the datum was considered to contain new information whereupon it was collected into the modeling database. Thus, the DMI does not only measure the similarity of predictor data, it also compares the data of response variables. However, the DMI is a distance-based method, and the variable correlation is not taken into account. The above discussion suggests that a soft sensor can be built either using a local modeling method (JIT or CoJIT), which collects similar data, or by collecting dissimilar data (DMI) for global modeling. These approaches measure the similarity among data points by not taking the correlation between input and output variables into account. However, cross correlation may be the only way to discriminate between data from unobserved multi-mode operations. Therefore, the presented work proposes two statistics to measure the cross correlation of input and output variables. The variable reselection step needs to be performed, once any one of the proposed statistics is out of its control limits.

The remainder of this paper is organized as follows. Section 2 provides the preliminaries of the PLS-based algorithms that will be applied in the proposed approach. In Section 3, the unobserved multi-mode operation and the proposed approach for online variable reselection are detailed. Section 4 presents a numerical example to compare the performance of data discrimination using CoJIT, DMI, and the proposed approach. In addition, operating data from the de-ethane process is utilized to demonstrate the effectiveness of the proposed approach. Finally, conclusions are provided in Section 5.

2. Preliminaries

2.1. Partial Least Squares (PLS)

PLS regression is a popular statistical tool for modeling the predictor and response datasets, in which a set of latent variables (LVs) is solved iteratively to describe the predictor (**X**) and response (**Y**) data matrices.

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

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

where **T** represents the first k terms of the latent variables or the score vectors, **P** and **Q** are the loading vectors of the data matrices **X** and **Y**, respectively, and **E** and **F** are the residual terms of PLS. In general, each score is extracted through deflating **X** and **Y** by the nonlinear iterative partial least-squares (NIPALS) algorithm until all variance in the data structure is explained. If the original **X** was used, the score vectors can be expressed as:

$$\mathbf{T} = \mathbf{X}\mathbf{R}, \quad \mathbf{R} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1} \tag{3}$$

Rearranging Eqs. (2) and (3), the regression form of the response variables corresponding to each predictor can be written as:

$$\mathbf{Y} = \mathbf{X}\mathbf{B}_{\mathrm{PLS}} + \mathbf{F}, \quad \mathbf{B}_{\mathrm{PLS}} = \mathbf{R}\mathbf{Q}^{\mathrm{T}} \tag{4}$$