Journal of Process Control 24 (2014) 1046–1056

Contents lists available at ScienceDirect

Journal of Process Control

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

Developing a soft sensor based on sparse partial least squares with variable selection

Jialin Liu*

Center for Energy and Environmental Research, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan, ROC

ARTICLE INFO

Article history:
Received 2 October 2013
Received in revised form 20 May 2014
Accepted 21 May 2014
Available online 14 June 2014

Keywords:
Soft sensors
Process dynamic modeling
Variable selection
Partial least squares

ABSTRACT

Soft sensors are used to predict response variables, which are difficult to measure, using the data of predictors that can be obtained relatively easier. Arranging time-lagged data of predictors and applying partial least squares (PLS) to the dataset is a popular approach for extracting the correlation between data of the responses and predictors of the process dynamic. However, the model input dimension dramatically soars once multiple time delays are incorporated. In addition, the selection of variables in the dynamic PLS (DPLS) model is a critical step for the robustness and the accuracy of the inferential model, since irrelevant inputs deteriorate the prediction performance of the soft sensor. The sparse PLS (SPLS) is a variable selection method that simultaneously selects the important predictors and finds the correlation between the predictors and responses. The sparsity of the model is dependent on a cut-off value in the SPLS algorithm that is determined using a cross-validation procedure. Therefore, the threshold is a compromise for all latent variable directions. It is necessary to further shrink the inputs from the result of SPLS to obtain a more compact model. In the presented work, named SPLS-VIP, the variable importance in projection (VIP) method was used to filter out the insignificant inputs from the SPLS result. An industrial soft sensor for predicting oxygen concentrations in the air separation process was developed based on the proposed approach. The prediction performance and the model interpretability could be further improved from the SPLS method using the proposed approach.

1. Introduction

In industrial processes, operators adjust manipulated variables to maintain product qualities or exhaust gases within the specifications of the product or government regulations, according to online analyzers and laboratory tests. However, online analyzers can malfunction, and there can be significant delays during laboratory testing; therefore, soft sensors that infer the primary output from other process variables can provide useful information for regulating process operations. Soft sensor applications have attracted significant attention in the process industry [1]. There are two main categories of soft sensor development: first-principle models and data-driven models. A first-principle physical model can be obtained from the fundamental process knowledge. However, due to the complexity of the manufacturing process, such fundamental models either require a lot of effort and time to develop, or are too simplistic to be accurate in practice. On the other hand, data-driven models provide accurate information for a particular operation region by multivariate regression methods [2] such as: principal component regression (PCR), partial least squares (PLS) and canonical coordinates regression (CCR). Such models are usually linear, therefore they lack the ability to extrapolate into different operating regions. To cover a wide range of operations, nonlinear models may be used, such as artificial neural networks (ANN) [3], support vector machines (SVM) [4] and kernel partial least squares (KPLS) [5]. However, the above-mentioned soft-sensing techniques are based on the assumption that processes are operating at steady states. The static soft sensors may suffer accuracy and robustness difficulties when the process dynamics dominating; such as: transitions between operating modes, disturbance rejections, and so on. Therefore, the dynamic correlations between inputs and outputs need to be concerned for developing a reliable soft sensor. In most industrial processes, the sampling rate of primary output is usually less frequent than that of process variables; thus, it is a challenging task to estimate the process dynamics based on the data-driven approaches. Lin et al. [6] integrated the data lifting method [7] and a weighted partial least-squares (WPLS) to develop a quality estimator for predicting the amount of free lime (CaO) in the clinker based on the multiple-rate sampled data, which were collected from a cement kiln process. Lu et al. [8] modified the

* Tel.: +886 3 5735294; fax: +886 3 5725924.
E-mail address: jialin@che.nthu.edu.tw

http://dx.doi.org/10.1016/j.jprocont.2014.05.014
0959-1524/© 2014 Elsevier Ltd. All rights reserved.

differential evolution (DE) algorithm to estimate the parameters of finite impulse response (FIR) template, which is a combination of second-order transfer functions with time delays to describe the dynamic relations between inputs and output, for capturing the process dynamics from the two-rate samples. More recently, Shang et al. [9] improved the work of Lu et al. [8] based on a Bayesian framework by incorporating FIR and SVM to deal with the process dynamics and nonlinearity, respectively. In their approach, the parameters of FIR and SVM were optimized according to the Bayesian inference; meanwhile, the logarithm of total probability was referred to as the evidence to examine whether the model was overfitted; therefore, the validation dataset was not necessary. However, Shang et al. [9] assumed that the process variable dynamics can be described by a first-order model with time delay; therefore, the time delay and the time constant for each process variable should be determined. In chemical processes, using dozens of inputs to develop a soft sensor is common. In this situation, the FIR approach [9] may encounter a heavy-computational-loading difficulty.

The PLS algorithm is a popular multivariate statistical tool for modeling data of the predictor and response variables. It has been proven that the maximal covariance between two datasets can be captured by PLS [10]. However, the dynamic characteristic of a process cannot be neglected when developing a soft sensor. Dynamic PLS (DPLS) has been widely applied in the design of dynamic models for process control [11,12] and in the development of soft sensors for batch processes [13,14]. For a continuous process, the input variables of DPLS are formed using the presented data and some time-lagged data of the predictor variables. However, it is often unclear if the response variables are affected by the length of the predictors' delay, and how to determine the predictor time lags for the DPLS model remains an open question. Kano et al. [15] evaluated the model performances of the predictors using different sampling intervals, and then collected the data with better modeling performances to form the training dataset to build the inferential model. Kaneko and Funatsu [16] proposed using the time difference data for modeling a soft sensor in order to eliminate the effects of drift and gradual changes for process data. In order to capture the process dynamics, the data of predictors were prepared according to the time differences, in which the current data were differentiated with the data from the different sampling intervals. Since 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. The high-dimensional dataset often contains data of predictors that are irrelevant for predicting the variations of response variables. For example, when the data of predictors are later than the corresponding time delays to the response variables, these data are irrelevant for predicting the current outputs of response variables. Since the time delay of each predictor is usually unknown, the training dataset of DPLS inevitably contains these irrelevant data for predicting the outputs. Even if the contribution to the model is small, the prediction performance can be deteriorated by these irrelevant data of predictors. In the perspective of DPLS modeling method, these irrelevant data of predictors are called the irrelevant variables.

If the modeling data contain a massive number of irrelevant predictors, the latent variables (LVs) of PLS will tend to capture the variances of that predictors rather than those of the responses [17,18]. Since PLS captures the covariance between the predictors and the responses, it is inevitable that the latent structure will be affected by variations of the predictors. Helland [18] illustrated that the model prediction errors will be large when the number of predictors is excessive by comparing the regression parameter and the estimator, in which the former represents the real model and the latter is obtained using a limited number of data. Therefore, in order to reduce the number of predictors for enhancing the prediction performance of the PLS model, several variable selection methods have been developed, as shown in the literature.

In the PLS algorithm, the weight vector ($\mathbf{w}$) is successively found by maximizing the covariance of the deflation data of predictors and responses. The smaller values in the weight vector represent that the corresponding predictors are less correlated with the responses. Therefore, a threshold was set to filter out the insignificant predictors and the final regression model was performed using the retained predictors, named intermediate least squares (ILS) [19]. A similar concept for variable selection was applied to the regression coefficients of the PLS model [20]. In [20], the authors concluded that the wavelength selection of spectroscopic data significantly improved the model prediction ability, since only some variables are relevant for the prediction. Wold et al. [21] introduced the measure of variable importance in projection (VIP) that evaluated the captured variance of response by each predictor on the latent space. Chong and Jun [22] recommended the range of cut-off values to the VIP scores, in which the threshold may be higher than one for the high portion of irrelevant variables, and vice versa. In addition, they also reported that the variables selected using the regression coefficients (PLS-Beta) and the VIP scores (PLS-VIP) were complementary. A combination of PLS-VIP and PLS-Beta for variable selection should provide fewer variables for prediction. In the review paper [23], the above-mentioned methods were categorized to the filter methods for the variable selection of PLS.

Other than the filter methods, the input variables were classified into several subsets based on a number of criteria. For each subset, a model was built and the prediction performance was evaluated. Thus, the predictors of the final model could be collected from the subsets that have better prediction performances. For example, Arakawa et al. [24] applied the genetic algorithm (GA) to select the wavelength of near-infrared (NIR) spectral data for modeling soil properties and the sugar content of apples. The wavelength of the spectral data was divided into several subintervals. The fitness values of combining the subinterval wavelength were calculated using PLS, in which the root mean squared error of cross-validation (RMSECV) value was used to justify the model performance. It is well known that the GA search is a time-consuming randomized search. In addition, the number of subintervals or the initial population of chromosomes are the critical parameters for GA-based PLS [25]. However, the issue of the initial subintervals, which determine the resolution of the selected variables, is not addressed in the GA approaches. Reinikainen and Höskuldsson [17] proposed covariance procedures (COVPROC) for variable selection in the PLS regression. In their approach, the training dataset was divided into several subsets according to the predictors and the time periods of data collection. For each subset, the important predictors were selected according to the largest magnitudes of weights, which were the absolute values in the eigenvector of covariance between the predictors and the responses in the corresponding PLS model. Since the covariances may be different for the subsets, the selected variables for each subset may not be consistent. They concluded that the models developed using the COVPROC method should not be used for predicting the present response values and that the models could be used to understand the dynamic characteristics of processes. More recently, Fujiwara et al. [26] applied the nearest correlation spectral clustering (NCSC) method to screen the input variables for PLS modeling. The input variables with a similar correlation were classified into a subset. For each subset, a PLS model was built to evaluate the fitness using the contribution ratio, which was defined according to the measurements and the estimates of the response variables. The final PLS model was then built using the subsets that had models with higher ranking contribution ratios. The approach was based on the variables in the same subset having the same contribution to the responses. However,