Journal of Process Control xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Process Control

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

Semi-supervised online soft sensor maintenance experiences in the chemical industry

Bo Lu, Leo Chiang*

Analytical Technology Center, The Dow Chemical Company, United States

ARTICLE INFO

Article history:
Received 15 April 2016
Received in revised form 7 January 2017
Accepted 23 March 2017
Available online xxx

Keywords:
Soft sensor
Model maintenance
Multivariate statistics
Inferential sensors
Online models
Process monitoring
Adaptive sensors

ABSTRACT

With the increasing availability of spectral, vibrational, thermal and other sensors, the challenge of "Big Data" in chemical processing industry is not only to analyze the data offline, but also to make use of the data online to improve process operation. To this end, accurate and reliable measurements or estimations of product quality are critical in today's demanding manufacturing environments. Data-driven soft sensors based on Projection to Latent Structure (PLS) methods are often used to model key quality variables using measureable inputs. However, most processes do not operate around a true steady state due to changes in equipment, feedstock, sensor and operating strategy. Therefore, soft sensor models need to be updated periodically. Current model maintenance approaches such as moving window update, recursive update in industry center around rebuilding the model using more recent process data. This approach is not robust enough in scenarios where process data is contaminated with outliers, downtime and other non-steady state transients. In this study, an alternative model update approach is developed. First, we adapted two key performance indicators (KPIs) for assessing the performance of the current soft sensor model. The Hotelling's T^2 based KPI is a predictive KPI that monitors for model extrapolations against future process data; the prediction residual based KPI then detects long term prediction degradation trends using a filtered prediction error. Second, we developed an update strategy using the robust mean and variance estimators of the inputs and outputs. Through case studies using industrial process data, this update method was demonstrated to be effective in improving prediction performance without rebuilding the PLS model from scratch. Lastly, the model update mechanism can be combined with both KPIs indicators. Through simulation of online behavior using industrial data, we showed that this update strategy effectively improved the prediction performance of the PLS soft sensor. In cases where the initial model was suboptimal, the update strategy also allowed for timely identification of underlying problems and alerted engineers of the need to rebuild the model.

1. Introduction

Accurate and reliable measurements and prediction of quality variables ensure critical success in today's demanding manufacturing environments. In the absence of an economical or effective online measurement, inferential or soft sensors could serve as an alternative solution [1–4]. Over 300 soft sensors are currently deployed globally across Dow. These soft sensors infer important quality variables such as concentration or impurity levels from existing process variables such as temperature, pressure, and flow. These quality variables are usually measured offline in a lab which offers limited number of samples with significant time delay, making it difficult to make timely process adjustments. With the onset of "Big Data" movement (larger storage, higher frequency data, diverse sensors, and more computational power), larger and more diverse datasets will be available for soft sensor development. However, as Reis et al. pointed out in [5], there remains various research challenges that need to be addressed before the full potential of "Big Data" can be utilized.

One of the challenges in soft sensors is that the life spans of most data-driven soft sensors are limited. This is because most processes do not operate around a true steady state. Changes in equipment, feedstock, sensor and operating strategy are often on a much slower time-scale than the available training data. In addition, outliers and nonlinearity are also factors that affect soft sensors performance and longevity, many works have been published in literature that deal with these issues exclusively [6–8]; and thus these topics are not discussed in this paper. In Kano and Fujiwara's report of indus-

* Corresponding author.
E-mail addresses: bo.lu@utexas.edu (B. Lu), hchiang@dow.com (L. Chiang).

http://dx.doi.org/10.1016/j.jprocont.2017.03.013
0959-1524/

trial survey results on soft sensor implementations [9], issues of "Accuracy deterioration due to changes in process characteristics" accounted for 29% of the problems in soft sensing. As a result, data-driven soft sensors need to be updated periodically to maintain their performance in the most efficient and least intrusive way possible. To this end, an important question that need to be answered in the era of "Big Data" is whether using more data will lead to better model performances [5].

One way to utilize all the available data for model update is through recursion or moving window updates. These methods have been popular in recent academic literatures [10–14]. In these approaches, the regression methods are repeatedly applied whenever new data are obtained. While this adaptive update process can be automated, challenges for industrial implementation exist. For instance, the convergence and stability of these update algorithms depends on the data being free of faults or outliers, which is rarely the case in industrial systems. In addition, these adaptive techniques usually only update the coefficients and do not consider potential changes in the underlying causal variables. This limitation creates a dilemma where we need to consider the trade-off between simplicity of the model structure versus the robustness during model adaptation. There is also difficulty in assessing the performance of an adaptive model system in closed-loop configuration [2]; in other words, how can we detect fundamental disagreements between model and process data, so that we avoid delaying imminent human intervention? Lastly, to apply these methods online also requires modern algorithm execution platforms, which has slowed their adoption industrywide. It is clear that we need to be more selective about how to use the additional data. The update scheme should be robust, simple and effective. The model maintenance should fail gracefully when the differences between the trained model and process data are too big to reconcile.

Fujiwara and Kano has proposed a new paradigm through the use of just-in-time modeling [1]. In their approach, the model maintenance step is bypassed since a new model is trained each time a prediction is made. The trained model uses data that is most relevant for that scenario. However, the use of a database to store all the past training data inherently assumes that the past data is representative of the future. This means that this approach is not designed for processes with drifts or new operating conditions. In this study, we attempt to address model maintenance of existing PLS soft sensors in a robust and simple manner without the use of closed-loop adaptive updates. The overall objectives of this study are the following:

1. Determine a set of key performance indicators (KPI) to provide guidelines on when should the model be updated vs. when the model should be completely overhauled.
2. Develop a method for performing a quick update on soft sensor models preserving the underlying correlation from the original models.
3. Derive a set of update rules that combines the KPI andmodel update mechanisms to form a unifying model maintenance framework.

1.1. Preliminaries

PLS soft sensors rely on the partial least squares (PLS) regression, which belongs to a class of latent projection based regression methods. Latent projection methods have a proven record of success in various soft sensor applications [2–4,15–17]. Here we give a very brief overview of PLS regression, readers can refer to [15]; for a more detailed explanation. The PLS methods decompose the mean-centered matrices into the following form:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F}$$

where $\mathbf{T} \in \mathrm{R}^{n\times A}$ and $\mathbf{U} \in \mathrm{R}^{n\times A}$ are the **X** and **Y** scores respectively, $\mathbf{P} \in \mathrm{R}^{m\times A}$ and $\mathrm{Q} \in \mathrm{R}^{p\times A}$ are the loadings for **X** and **Y**, respectively. The number of components in the PLS model is typically determined through cross-validation or through information criterion such as the Akaike Information Criterion (AIC) [18]. The PLS algorithm maximizes the covariance between the **X**-scores and **Y**-scores; this property leads to PLS requiring fewer components when compared to principal component regression models [18].

To apply the PLS latent structures in regression, given unfolded input data matrix $\mathbf{x}_0$, the output predictions $\hat{y}_0$can be calculated linearly using $\hat{y}_0 = \boldsymbol{x}_0\beta_{PLS}$. The β_{PLS} can be expressed as a function of the latent variables as follows:

$$\hat{\beta}_{PLS} = \boldsymbol{R}\left(\boldsymbol{T}^T\boldsymbol{Y}\right) = \boldsymbol{R}\boldsymbol{R}^T\boldsymbol{X}\boldsymbol{Y}$$

where **R** is the loading weight matrix following the notation in [19].

In addition to numerical stability and capability to handle high dimensionality data, another key advantage of PLS/PCA are the additional diagnostic indices (The Hotelling's T^2 and Squared Prediction Error (SPE)) that can reveal additional information about the process. The multivariate statistics Hotelling's T2 and the SPE can be calculated as follows:

$$T^2 = \boldsymbol{t}_0^T\boldsymbol{\Lambda}^{-1}\boldsymbol{t}_0 \sim \frac{A\left(n^2-1\right)}{n\left(n-A\right)}F_{A,n-A}$$

$$SPE = \|\boldsymbol{x}_0 - t_0 p_0\|^2 \sim g\chi_h^2$$

where $\mathbf{t}_0$ and $\mathbf{p}_0$ are the PLS decomposed scores and loadings for the data batch being tested, and A is the number of components in the PLS model. $\boldsymbol{\Lambda}$ is defined by $\boldsymbol{\Lambda} = \frac{1}{n-1}\boldsymbol{T}^T\boldsymbol{T}$. Given significance level α, control limits for T^2 and SPE can be calculated from the Fischer and the χ^2 distributions respectively. The Hotelling's T^2 statistic detect mean shifts from PLS/PCA score vectors as an indicator of process operation normality. The SPE (sometimes also referred to as Distance to Model for the X inputs (DMODX)) estimates the magnitude of model residual for incoming data, where a deviation would indicate degrading model performance or abnormal incoming data. An in-depth introduction of these diagnostic indices can be found in [20].

Soft sensor degradation refers to the decrease in the prediction performance of the soft sensor when compared against the nominal performance when the soft sensor was first developed. To account for occasional unaccounted for process variations such as sensor failures or upstream/downstream disturbance, the prediction performance is monitored using aggregate statistics (indices) instead of singular prediction errors. Most of the indices are based on prediction error or variance explained, such as the root mean squared prediction error (RMSPE), sum of squared residual (SSR), or coefficient of determination (R^2). In [17], the most popular indices in evaluating prediction performance have been provided.

The underlying causes of performance degradation can be grouped into three primary contributing factors [2] have been summarized into three categories in Table 1.

Process fault refers to changes in the actual process condition that causes a change in the input-output relationship from the time of soft sensor model training. The duration of the process fault can be long or short depending on the actual change that took place. For example, a process upset could be a relatively short upset that requires no change to the soft sensor model, while a catalyst change or a plant turn-around would result in completely different operating characteristics that requires soft sensor model update.

The model fault refers to all numerical and statistical challenges associated with developing the actual regression model used in