Journal of Process Control 32 (2015) 25–37

Contents lists available at ScienceDirect

Journal of Process Control

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

Robust semi-supervised mixture probabilistic principal component regression model development and application to soft sensors

Jinlin Zhu, Zhiqiang Ge*, Zhihuan Song

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

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form 25 March 2015
Accepted 29 April 2015
Available online 22 May 2015

Keywords:
Soft sensor
Outliers
Semi-supervised learning
Student's *t* distribution
Mixture latent variable models

ABSTRACT

Traditional data-based soft sensors are constructed with equal numbers of input and output data samples, meanwhile, these collected process data are assumed to be clean enough and no outliers are mixed. However, such assumptions are too strict in practice. On one hand, those easily collected input variables are sometimes corrupted with outliers. On the other hand, output variables, which also called quality variables, are usually difficult to obtain. These two problems make traditional soft sensors cumbersome. To deal with both issues, in this paper, the Student's *t* distributions are used during mixture probabilistic principal component regression modeling to tolerate outliers with regulated heavy tails. Furthermore, a semi-supervised mechanism is incorporated into traditional probabilistic regression so as to deal with the unbalanced modeling issue. For simulation, two case studies are provided to demonstrate robustness and reliability of the new method.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

A soft sensor is an inferential model that characterizes quantitative relationships between process variables that are easy to measure and those that are not [1]. The main advantage of soft sensor is that it can provide reliable, fast and low-priced estimations for important variables [2,3]. For these reasons, soft sensors have been widely used in industrial processes to estimate those quality and key variables that are difficult to measure online [4–8].

Traditional soft sensors are typically constructed upon first principal strategies that require process knowledge and expert experiences [9]. Besides time-consuming analysis, such strategy can also suffer from the lack of sophisticated process kinetic knowledge which can be commonly encountered for chemical processes [10]. As an alternate, data-based soft sensors can be effectively built by requiring little process knowledge and expert experiences [11]. In this sense, data-based soft sensors such as multivariate regression, principal component regression (PCR) and partial least squares (PLS) have been popular over the past few decades [6,12]. Originally, PCR and PLS are constructed regardless of the underlying uncertainty introduced by data noise [13]. To overcome this issue, the probabilistic definitions on the framework of latent variable models have been developed, the derived model are probabilistic PCR (PPCR) and probabilistic PLS (PPLS) [14,15]. By assigning a Gaussian distribution for each variable, the probabilistic methods should be more extendable and elegant in modeling with the Bayesian inference mechanism [16]. For example, the PPCR can be either extended as a mixture model for multimode modeling or modified as a fully Bayesian method so as to conduct the model selection for latent space dimensionality [17,18].

Although there are many potential benefits, a main practical problem is the fact that all these methods are designed with Gaussian based assumptions. So when non-Gaussian variations such as outliers are introduced, they can be susceptible by outliers since the Gaussian distribution with weak outlier-tolerant mechanisms can be easily skewed by modeling layouts [19,20]. Unfortunately, it is well known that most industrial datasets contain outliers due to incorrectly observed or recorded process measurements [11,21]. On the other hand, manual evaluation and discard can be time consuming and inefficient [22]. Moreover, simply case deletion of those outliers can lose important information since one has to discard the whole sample just because only one 'dirty' entry record. Recently, some studies have made attempt to develop the Student's *t* based probabilistic models so as to deal with modeling outliers [23–25]. Compared with the Gaussian distribution

* Corresponding author. Tel.: +86 87951442.
E-mail address: gezhiqiang@zju.edu.cn (Z. Ge).

http://dx.doi.org/10.1016/j.jprocont.2015.04.015
0959-1524/© 2015 Elsevier Ltd. All rights reserved.

which is commonly appeared for constructing PPCR and PPCA, the Student's t counterpart shows more stability and compatibility since the heavy tail part that explains the noise is usually adjusted by the parameter called degree of freedom which can be naturally adapted from the training procedure [26,27]. In this sense, the Student's t distribution is more robust to outliers than the Gaussian one [28,29]. Followed by this idea, this work tries to introduce the Student's t mechanism for robust probabilistic soft sensor development.

Another common problem for data-based soft sensors is that they are designed with the demand that modeling data of output variables should have the same length as input variables. In this paper, the dataset with both input and output samples are annotated as labeled data, while those with no output assignments are regarded as unlabeled ones. Therefore, most data-based soft sensors are developed upon completely labeled situations. However, as mentioned above, output variables such as quality variables can be hardly observed or sampled online and one needs to take offline lab endeavors for numerical details. As a matter of fact, for modern processes with large volumes of measured input dataset, one can hardly provide the same amount of output counterpart and only a small subset of input samples can undergo lab efforts and are attached with output labels. In this case, modeling performance cannot be guaranteed when a small portion of labeled samples are employed (learning with labeled data alone also refer to the down sampling mechanism [30]). However, if the omitted unlabeled samples are incorporated, the soft sensor performance can be improved [31]. Such procedures by learning with partially labeled dataset are usually denoted as semi-supervised learning [32,33]. Recently, the semi-supervised strategy has been proposed for PCR and successfully applied into chemical processes [3,31]. However, such models are still limited to Gaussian assumptions, and can be susceptible by those outliers. In fact, both input outliers and low-rate output sampling mechanism do exist simultaneously in practical industrial systems. In such case, all aforementioned schemes like outlier case deletion and input down sampling mechanisms can be unreliable since none of them can make use of modeling information effectively.

The motivation of this article is to propose a novel regression model that can cope with the modeling outliers as well as the uneven length of output variables. First, the conventional mixture PPCR (MPPCR) is modified with the Student's t distribution so as to conduct the robust modeling with potential outliers. Notice that the developed robust method is designed with mixture form so as to deal with the multimode process data. Based on the robust mixture model, a semi-supervised learning mechanism is further incorporated so that the proposed model can deal with the unbalanced output samples. During the online soft sensing procedure, for each new coming measurement, we first estimate the output values from each local model and then align them softly with the corresponding weight. The global estimation is considered as the current time production quality. Notice that the local weight is calculated as the posterior of the measurement with respect to each local model, which can be achieved by Bayes rule.

The rest of paper is organized as follows. In Section 2, the conventional mixture PPCR is revisited, followed by the introduction of robust mixture PPCR model with the EM algorithm. Based on that, the robust soft sensor is developed on the basis of semi-supervised mechanism. Followed by modeling, the online soft sensing mechanism is developed based on the proposed model. After that, two case studies are used to validate the proposed method. Finally, conclusions are made.

2. Preliminaries

2.1. Mixture PPCR

Given N input dataset $\{\mathbf{x}_n|\mathbf{x}_n \in R^{D_x}\}_{n=1}^N$ and output $\{\mathbf{y}_n|\mathbf{y}_n \in R^{D_y}\}_{n=1}^N$, assume that a set of M mixture local components are combined with each local component a single PPCR. The generative model for MPPCR seeks to find the relationship between input and output which is given as [17]:

$$\mathbf{x}_{n,m} = \mathbf{P}_m\mathbf{t}_{n,m} + \boldsymbol{\mu}_{x,m} + \mathbf{e}_m, \quad m = 1, 2, \ldots, M \tag{1}$$

$$\mathbf{y}_{n,m} = \mathbf{Q}_m\mathbf{t}_{n,m} + \boldsymbol{\mu}_{y,m} + \mathbf{f}_m, \quad m = 1, 2, \ldots, M \tag{2}$$

$$\mathbf{x}_n = \sum_m p(m)\mathbf{x}_{n,m} \tag{3}$$

$$\mathbf{y}_n = \sum_m p(m)\mathbf{y}_{n,m} \tag{4}$$

where $\mathbf{P}_m \in R^{D_x\times d}$ and $\mathbf{Q}_m \in R^{D_y\times d}$ are the projection matrixes for input space and output space, D_x and D_y refer to the dimensionalities for input and output variables, d is the dimensionality for all latent spaces, $p(m)$ is the mixture weight that satisfies $\sum_{m=1}^M p(m) = 1$. $\mathbf{t}_{n,m} \in R^{d\times 1}$ is the latent variable for the employed data spaces, additional terms like $\mathbf{e}_m \in R^{D_x\times 1}$ and $\mathbf{f}_m \in R^{D_y\times 1}$ denote the noise for each space. The probability distributions for the above model are defined by Gaussian ones as $p(\mathbf{t}_{n,m}) = N(0, \mathbf{I}_d)$, $p(\mathbf{e}_{n,m}) = N(0, \tau_{x,m}\mathbf{I}_{D_x})$, $p(\mathbf{f}_{n,m}) = N(0, \tau_{y,m}\mathbf{I}_{D_y})$. Let $\mathbf{z}_n = \begin{pmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{pmatrix}$, $\mathbf{W}_m = \begin{pmatrix}\mathbf{P}_m\\ \mathbf{Q}_m\end{pmatrix}$, $\boldsymbol{\mu}_m = \begin{pmatrix}\boldsymbol{\mu}_{x,m}\\ \boldsymbol{\mu}_{y,m}\end{pmatrix}$, $\Phi_m = \begin{pmatrix}\Phi_{x,m} & 0\\ 0 & \Phi_{y,m}\end{pmatrix}$, $\Phi_{x,m} = \tau_{x,m}\mathbf{I}_{D_x}$, $\Phi_{y,m} = \tau_{y,m}\mathbf{I}_{D_y}$, then $p(\mathbf{z}_n|\mathbf{t}_{n,m}) = N(\mathbf{W}_m\mathbf{t}_{n,m} + \boldsymbol{\mu}_m, \Phi_m)$. Therefore, undetermined parameters for PPCR are $\boldsymbol{\Theta} = \{p(m), \mathbf{P}_m, \mathbf{Q}_m, \boldsymbol{\mu}_{x,m}, \boldsymbol{\mu}_{y,m}, \tau_{x,m}, \tau_{y,m}\}_{m=1}^M$, which can be iteratively estimated by EM algorithm, one can refer to many literatures for more mathematical details, such as [34].

2.2. Robust MPPCR

In robust MPPCR (RMPPCR), the generative model structure is the same as MPPCR, except for the utilization of Student's t distribution for model assumption. Specifically, the priors and likelihoods can be given as [28]:

$$p(\mathbf{t}_{n,m}) = \mathrm{S}(\mathbf{t}_{n,m}|0, \mathbf{I}_d, \nu_m) \tag{5}$$

$$p(\mathbf{x}_n|\mathbf{t}_{n,m}) = \mathrm{S}(\mathbf{x}_n|\mathbf{P}_m\mathbf{t}_{n,m} + \boldsymbol{\mu}_{x,m}, \Phi_{x,m}, \nu_m) \tag{6}$$

$$p(\mathbf{y}_n|\mathbf{t}_{n,m}) = \mathrm{S}(\mathbf{y}_n|\mathbf{Q}_m\mathbf{t}_{n,m} + \boldsymbol{\mu}_{y,m}, \Phi_{y,m}, \nu_m) \tag{7}$$

Download English Version:

https://daneshyari.com/en/article/688701

Download Persian Version:

https://daneshyari.com/article/688701

Daneshyari.com