Chemical Engineering Science 193 (2019) 394–410

Contents lists available at ScienceDirect

Chemical Engineering Science

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

Quality variable prediction for chemical processes based on semisupervised Dirichlet process mixture of Gaussians

Weiming Shao, Zhiqiang Ge, Zhihuan Song *

State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

HIGHLIGHTS

- A novel semisupervised soft sensing approach for chemical processes is proposed.
- A Dirichlet process mixture of Gaussians is proposed for regression application.
- A VI-based learning algorithm is developed for the proposed fully Bayesian model.

ARTICLE INFO

Article history:
Received 3 April 2018
Received in revised form 3 August 2018
Accepted 19 September 2018
Available online 20 September 2018

Keywords:
Soft sensor
Semisupervised learning, Dirichlet process mixture models
Variational inference
Nonlinear chemical processes
Non-Gaussian data

ABSTRACT

Data driven soft sensors have found widespread applications in chemical processes for predicting those important yet difficult-to-measure quality variables. In the vast majority of chemical processes, relationships among primary and secondary variables are nonlinear, and process data inherently contain uncertainties and present strongly non-Gaussian characteristics. In addition, labeled samples are often scarce due to certain technical or economical difficulties. These process and data characteristics impose challenges on high-accuracy soft sensors. To deal with these issues, this paper proposes a soft sensing approach referred to as the semisupervised Dirichlet process mixture of Gaussians (SsDPMG). In the SsDPMG, a fully Bayesian model structure is first designed to enable semisupervised tasks that are suitable for regression applications. Subsequently, a Bayesian learning procedure for the SsDPMG is developed based on variational inference framework, where information contained in both labeled and unlabeled samples are extracted. Case studies are carried out on one numerical example and two real-life chemical processes to evaluate the performance of the proposed approach. The results demonstrate that the SsDPMG is an effective soft sensing approach with promising application foreground.

1. Introduction

In chemical processes, in additional to those easy-to-measure variables such as pressure, temperature, flow rate and liquid/material level, there are a class of product quality-related variables called primary variables such as concentration, melt index, and octane number. These primary variables are usually of significant importance for quality control and operation safety, but are difficult to measure. Acquisitions of them are conventionally realized via offline laboratory analysis or online analyzer, which may introduce significant delay (hours or even longer) or tremendous investment cost, raising challenges to real-time process monitoring and closed-loop control (Ge et al., 2017; Yu, 2012; Wang et al., 2010).

Soft sensors, which are also known as virtual sensors or inferential sensors, are able to resolve the above mentioned difficulties associated with the lab analysis or hardware analyzer. They can economically provide delay-free estimations for primary variables using secondary variables (i.e., those easy-to-measure variables) and mathematical models (Shao and Tian, 2012; Yan et al., 2017). Owing to less dependence on in-depth expert knowledge and availability of amounts of process data, compared to model-driven soft sensors, data-driven ones have received much attention and found increasingly wide applications in many industrial fields such as chemical engineering, biological engineering, metallurgical engineering, and so forth (Ge et al., 2017; Yuan et al., 2018). During past decades, a variety of data-driven soft sensors have been developed, and one can refer to Kadlec et al. (2009, 2016) for comprehensive reviews of popular data-driven soft sensing algorithms as well as their industrial applications.

Due to reasons such as complex process mechanisms, multiple manufacturing phases and/or operating conditions, the vast majority of chemical processes are nonlinear, and process data present strongly non-Gaussian characteristics (Yu, 2012; Xie et al., 2014;

* Corresponding author.
E-mail address: songzhihuan@zju.edu.cn (Z. Song).

https://doi.org/10.1016/j.ces.2018.09.031
0009-2509/

Cai et al., 2017). In addition, chemical processes are inherently stochastic, which results from certain factors like noisy measurement environments and transmission disturbances (Yuan et al., 2018; Zhou et al., 2014; Yuan et al., 2017). It is desirable to account for these process characteristics when developing soft sensor models, which however raises challenges to traditional soft sensing approaches. For example, some popular multivariate statistical models such as those based on principal component analysis (PCA) and partial least squares (PLS) could not capture the nonlinear relationships among primary and secondary variables. Although kernelized PCA/PLS and other nonlinear function approximators such as support vector regression (SVR) and back propagation neural networks (BPNN) can deal with process nonlinearities, they are not good at modeling process uncertainties. Specifically, the deterministic SVR and BPNN do not treat process variables as random variables, and thereby they could provide only point estimations of primary variables. However, it is desirable to provide not only point estimations but also estimation uncertainties, which could be very useful for many purposes such as abnormal sample classification, reliable online model update, and online hardware analyzer calibration Liu et al. (2010),Kaneko et al. (2011). In addition, it is difficult for SVR and BPNN to deal with the issue of missing value, which can be dealt with more easily by probabilistic modeling approaches based on expectation-maximization (EM) algorithm or variational EM algorithm (Yuan et al., 2017).

In contrast, Gaussian mixture models (GMM) is capable of simultaneously dealing with those process characteristics (i.e., the non-Gaussianity, nonlinearity as well as randomness) by approximating any complex non-Gaussian distributions with finite number of Gaussian distributions (Yu and Qin, 2008). Therefore, GMM and its variations have recently established themselves as widely adopted soft sensing approaches (Yu, 2012; Grbić et al., 2013; Yuan et al., 2014; Xiong et al., 2016; Wang et al., 2016; Chan and Chen, 2017; Mei et al., 2017; Zhu et al., 2017). To construct a GMM-based soft sensor, two tasks should necessarily be completed. The first one is to learn model parameters including mixing coefficients, mean vectors and covariance matrices for each Gaussian component (GC). The EM (Dempster et al., 1977) algorithm is commonly used to obtain point estimates of model parameters for the standard GMM where model parameters are treated as deterministic variables rather than random variables. However, the EM algorithm aims to maximize the likelihood function and thus easily gets caught into local minimum and suffers from overfitting. For Bayesian GMM (BGMM) which randomizes model parameters and penalizes model complexity by integrating them out, the Monte Carlo Markov Chain (MCMC) (Rasmussen, 2000) and variational inference (VI) (Bishop, 2006) are usually used to learn posterior distributions over model parameters instead of their point estimates. Even though the MCMC provides a systematical way for learning of Bayesian models, it can be prohibitively slow and its convergence is difficult to detect (Bishop, 2006; Blei and Jordan, 2006). As a result, the MCMC is often limited to small scale problems. Such drawbacks of the MCMC can be overcome by the VI, which transforms the parameter learning task into a functional optimization problem. Note that some relaxation such as factorization approximation is usually necessary for the VI to find tractable solutions.

The other task that has to be finished for GMM is model selection, i.e., to select the number of Gaussian components (GCs). Insufficient GCs result in underfitting while excessive GCs lead to overfitting. Therefore, appropriately determining the number of GCs is of crucial significance for GMM-based soft sensors to achieve satisfying performance. There are various methods that can perform model selection for mixture models, which can generally be classified into two categories, namely criterion-based and Bayesian methods. In the first group, the optimal number of GCs is determined as the one that can minimize some criteria such as Akaike information criterion (AIC) (Yan et al., 2017), Bayesian information criterion (BIC) (Bourouis et al., 2014), absolute increment log-likelihood (AIL) (Yuan et al., 2014), Bayesian Ying-Yang index (BYYI) (Choi et al., 2005), minimum message length (MML) (Bouguila, 2012), pseudolikelihood information criterion (PLIC) (Stanford and Raftery, 2002), etc. This kind of model selection approaches could be highly computationally demanding, as they need to traverse all candidate numbers of GCs. In Bayesian methods, the mixing coefficients of GCs are treated as random variables, and their posterior distributions are learned and used for model selection. There are two representative strategies within this type of methods, i.e., parametric and nonparametric ones. In parametric methods, for example the variational mixture of Gaussians (VMG) presented in Bishop (2006), the number of GCs can be set as relatively large value, and contributions of superfluous GCs are driven to be sufficiently tiny. In contradiction to parametric methods, nonparametric ones assume the data are generated from a Dirichlet process, resulting in the Dirichlet process mixture models (DPMM) that are composed of infinite number of components (Zhu et al., 2017; Blei and Jordan, 2006; Lai et al., 2018). Based on studies on the VMG and DPMM (Zhu et al., 2017; Bishop, 2006; Blei and Jordan, 2006; Lai et al., 2018), we can see that the traversal of numbers of GCs is no longer necessary, as model selection and parameter learning can be finished within one training round. Moreover, compared to the VMG, the DPMM theoretically doesn't need to know the number of GCs.

In soft sensor applications, collected samples are usually partially labeled as labeling samples could be expensive due to high investment of mass spectrometer, or with large delay introduced by time-consuming laboratory analysis. Therefore, data-driven soft sensor modeling is essentially a semisupervised task with rare labeled samples and large amounts of unlabeled samples (Yan et al., 2016), which may lead to difficulties for GMM-based soft sensors in both parameter learning and model selection in the scenario of insufficient labeled samples. Specifically, the EM-based parameter learning by maximizing the likelihood function would suffer from overfitting and numerical issues. For model selection, without the support of sufficient samples, the commonly used criteria BIC is prone to penalize model complexity unduly and leads to biased results, while the best model suggested by the AIC tends to overfit (Bishop, 2006; Zhu et al., 2015; Burnham and Anderson, 2004). Despite that the VI-based learning schemes such as the VMG (Bishop, 2006) and DPMM (Zhu et al., 2017; Blei and Jordan, 2006; Lai et al., 2018) are able to alleviate the issue of overfitting, insufficient labeled samples may still prevent them from achieving satisfying performance.

Semisupervised soft sensors that make use of both labeled and unlabeled data have been proven effective in remedying the limitation of insufficiency of labeled samples. However, at the learning stage, conventional GMM-based soft sensors are either unsupervised (Yu, 2012; Grbić et al., 2013; Xiong et al., 2016; Wang et al., 2016; Chan and Chen, 2017) or supervised (Yuan et al., 2014; Mei et al., 2017; Zhu et al., 2017), failing to perform semisupervised task due to the structure of conventional GMM. Even though some semisupervised GMM and its extensions, such as semisupervised GMM (Yan et al., 2017; Xing et al., 2013), semisupervised variational GMM (Yang et al., 2017) and semisupervised DPMM (Kimura et al., 2009), have been developed, they are used for classification purpose and unable to develop predictive soft sensors which are regression models. Therefore, in this paper we propose a novel soft sensor-oriented SsDPMG (i.e., semisupervised Dirichlet process mixture of Gaussians) to resolve the issues discussed in the above paragraph which result from insufficient labeled samples. Our main contributions are summarized as follows: