Development of multiple-step soft-sensors using a Gaussian process model with application for fault prognosis

CrossMark

Yiqi Liu [a], Hongjun Xiao [a], Yongping Pan [b], Daoping Huang [a,*], Qilin Wang [c,**]

[a] School of Automation Science & Engineering, South China University of Technology, Wushang Road, Guang Zhou 510640, China
[b] Department of Biomedical Engineering, National University of Singapore, Singapore Medical Drive, Singapore, 117583, Singapore
[c] Advanced Water Management Centre, The University of Queensland, St. Lucia, Brisbane, QLD, 4072, Australia

ARTICLE INFO

ABSTRACT

Predicting the degradation of working conditions and trending of fault propagation before they reach the alarm or failure control limit is significantly important to optimize the operational capacity of a chemical process. However, traditional one-step-ahead (OS) soft-sensors render such benefits inadequate. Direct, Recursive and Direct-recursive strategies are proposed to generalize the Gaussian Process Regression (GPR) model for multi-step-ahead (MS) prediction, thereby supporting the fault diagnosis and prognosis of the product qualities control for chemical processes. The proposed methodology was firstly demonstrated by applying the designed algorithm to a wastewater plant (WWTP) simulated with the well-established model, i.e., Benchmark Simulation Model 1 (BSM1), then extended to a full-scale WWTP with data collected from the field influenced by filamentous sludge bulking. Results showed that the proposed strategies significantly improved the prediction performance.

© 2016 Published by Elsevier B.V.

1. Introduction

Early detection and diagnosis of the occurrence of an abnormal event in a chemical process is very important for ensuring plant safety and maintaining product qualities. Tremendous advancements in the area of instrumentation have made measure hundreds of variables every few seconds possible. These measurements have brought in useful signatures about the status of the plant operation and also have promoted a wide variety of techniques, for detecting faults [1,2].

Typically, a fault is declared once monitored values cross over a predefined control limit. The most commonly used methodologies for fault diagnose are multivariable statistics, such as Principal Component Analysis (PCA) and Factor Analysis (FA) [3–5]. These methods explore the data collected from the process to build an empirical model, which in turn acts as a reference to justify the desired process behavior of the new data. However, both of PCA and FA are premised on the assumptions that the collected data supposed to follow specified distributions. Independent Component Analysis (ICA) and its nonlinear version provide alternatives to deal with this problem [6,7]. Nonetheless, aforementioned fault diagnosis strategies are merely carried out during or after the breakdown moment [8]. These approaches are satisfied if a fault is not such urgent and is

able to be controlled during a short time. For most of cases, a significant disaster could have occurred already, once a fault with respect to product qualities has been recognized in a chemical process. If the selected product qualities can be predicted properly, it would lead to better process control performance in advance. Soft-sensors provide an alternative to deal with this problem. Soft-sensors are widely used to estimate variables that are difficult to measure online because of technical difficulty, large measurement delays, high investment cost, and so on [9–11]. To build a proper relationship between easy-to-measure variables x and those that are hard-to-measure y, statistical methods including, but not limited to, partial least squares (PLS) [12], Principal Component Regression (PCR) [13], nonlinear PLS [14], support vector machine based regression [15] are researched as the soft sensor models. Despite potential advantages, few studies have devoted soft-sensors for fault diagnosis and prognosis [16].

Due to classical soft-sensors limiting to one-step-ahead prediction, the fault effects could have resulted in a disaster when the failure of an operation process is indicated by out-of-control monitoring statistics. Therefore, it would be too late for the engineers to make any corrective actions with the reactive detection results. One of plausible ways is to generalize one-step-ahead to multiple-step-ahead predictions. Multiple-step-ahead prediction is a difficult task not only because of predicted model construction but also the uncertainty analysis of the resulted model. Currently, the form of the multiple-step-ahead prediction depends only on the one-step-ahead model under linear assumptions regardless of nonlinear requirement of the multiple-step predictor [17]. Various numerical and Monte Carlo methods have been devoted to compute the multiple-step-ahead prediction when

---

* Corresponding author.
** Correspondence to: Q. Wang, Advanced Water Management Centre (AWMC), The University of Queensland, Australia.
E-mail addresses: aulyq@scut.edu.cn (Y. Liu), biepany@nus.edu.sg (Y. Pan), audhuang@scut.edu.cn (D. Huang), q.wang@awmc.uq.edu.au (Q. Wang).

the nonlinear relationship between the one-step-ahead predictors and responses as well as the innovation distribution are known. However, these approaches are premised on assumptions of distribution being known, and their resulting accuracy depends heavily on the adequacy of the model. Computational intelligence methods such as artificial neural networks [18–20] and nearest-neighbors techniques [21,22] have drawn the attention of the forecasting community to deal with the inadequacy of a model for multi-step ahead prediction. To derive accurate prediction of a fault, the model needs not only to capture strongly nonlinear effects of operational conditions, but also to be robust for the presence of uncertainty and noises in a chemical process. If ignorance of uncertainty or merely use of expectation of uncertainty, the prediction performance could be poor and result in deviation of fault diagnosis or prognosis, thus leading to a unreliable operation of the system. Additionally, a standard regression-based prediction is accuracy-driven, i.e., to minimize the residual errors between the prediction and training values. However, the residual error is significantly worse when compared to regression-based methods because the correlation between the predictor and response variables remain invariant, which in turn, hinders its ability to minimize residual errors [18,23]. To quantize the uncertainties properly, distribution-driven models, such as Support Vector Machine and Relevant Vector Machine, were presented to approach predicted distribution [18,24]. Also, a distribution-driven model, Gaussian Process for regression (GPR), has been used for modeling both of linear and nonlinear dynamic systems simply using different covariance matrix. Compared to neural networks and other distribution-driven models, relatively few parameters need to be estimated, which lessens the need for the complex optimization or regularization schemes. Additionally, the prediction distribution of a GPR at a set of test points is simply assumed as a multivariate Gaussian distribution [25]. This assumption will be able to serve as the basis for the uncertainty description to facilitate decision making for fault identification and control. Even though GPRs have been tested in chemical processes, few attempts are used for fault diagnosis and prognosis.

To predict the occurrence of a fault in advance, the GPR model is required to extend to multi-step-ahead prediction necessarily. Two most common strategies for multi-step-ahead forecasting are recursive and direct prediction methodologies. In the first strategy, a multiple-step-ahead prediction task with horizon H is tackled by iterating H times a one-step-ahead predictor. Once the predictor has estimated the future series value, the value is fed back as an input to the following prediction. Hence, the predictor uses estimated values as inputs, rather than the actual observations, with evident negative consequences in terms of error propagation [26,27]. Another way of performing H-step-ahead forecasting is to estimate a set of H prediction models, each of which returns a direct forecast at time. Direct methods often require a higher degree of functional complexity than iterated ones [28], in order to model the stochastic dependency between two series values at two distant instants. When a very long term prediction is at stake and a stochastic setting is assumed, the modeling neglects the existence of stochastic dependencies between future values and consequently biases the prediction accuracy. A possible way of remedying this shortcoming is to combine recursive prediction and direct predicted ways. By combining these two methodologies, stochastic dependency obtained from direct strategy could propagate over a long term horizon by the iterative strategy, hence improving the prediction performance of multi-step-ahead prediction.

The objective of this study is to develop a multiple-step-ahead GPR predicted model to monitor the evolution of hard-to-measure product qualities in the chemical processes timely. The contributions of this paper are summarized as follows: Firstly, direct, recursive and direct-recursive strategies are proposed to extend the GPR model for multiple-step-ahead prediction, and comparisons are made with ARMA and RBF models for multi-step-ahead prediction. Due to selection of different covariance for the GPR model conveniently, diverse nonlinear relationships can be approached properly; Secondly, the uncertainty

information generated from multiple-step-ahead GPR would facilitate the description of error propagation for ill-suited multi-step prediction and model parameter identification; Thirdly, the obtained uncertainty intervals are able to envelop the deviation of multi-step-ahead predicted values, thereby providing pre-caution for fault diagnosis and prognosis during multi-step-ahead prediction.

In Section 2 the basic knowledge of GPR model are introduced. Section 3 presents the multiple-step-ahead GPR soft-sensors. In Section 4, the proposed direct, recursive and direct-recursive soft-sensors are firstly validated through a WWTP benchmark with relative stable operations to monitor the evolution of effluent BOD. Then, the proposed soft-sensors are used to predict the evolution of Sludge Volume Index (SVI) and further to monitor the occurrence of filamentous sludge bulking. Section 5 performs a discussion. Finally Section 6 concludes.

2. Preliminaries

2.1. GPR model identification

GPR model is a simple and general class of models of functions. To be precise, a GPR is any distribution over functions in such way that any finite set of function values $f(x_1), f(x_2), \cdots, f(x_N)$ have a joint Gaussian distribution. GPR is usually formulated as follows: given a training set $D = \{(x_i, y_i)\}_{i=1}^{N} c$ of N pairs of inputs x_i and noisy outputs y_i, compute the predictive distribution of f at a new testing input x^*. We assume that the noise is additive, independent and Gaussian, such that the relationship between the (latent) function $f(x_i)$ and the observed noisy targets y are given by [29]

$$y_i = f(x_i) + \varepsilon_i \tag{1}$$

$$\varepsilon \sim N(0, \sigma_n^2) \tag{2}$$

$$f(\cdot) \sim GP(0, k(\cdot, \cdot)) \tag{3}$$

where $GP(0, k(\cdot, \cdot))$ represents a Gaussian process with mean and covariance matrix equaling to 0 and $k(\cdot, \cdot)$, respectively. The noise ε follows the Gaussian distribution with mean 0 and covariance σ_n^2.

To be simple, we define covariance matrix $K = k_{ij}$. By inference, it is easy to obtain that the outputs follow multivariate joint Gaussian distribution:

$$y \sim N(0, K_y) \tag{4}$$

where $K_y = K + \sigma_n^2 I$, K_y is the covariance matrix with dimension being $N \times N$, the corresponding $(i,j)^{th}$ element is

$$(K_y)_{ij} = cov(y_i, y_j) = k(x_i, x_j) + \sigma_n^2 \delta_{ij} \tag{5}$$

where δ_{ij} is the Kronecker function. The difference of K and K_y is that K is noise free but the other is not. In summary, the parameters needed to be identified are formulated as $\theta = (\theta', \sigma_n^2)$, where θ' represents the parameters of corresponding covariance matrix K_y. Assume the covariance matrix is Squared-exp. (SE), the parameters are $\theta' = (\sigma_f, l)$. More details for kernel selections can see Appendix part.

2.2. GPR model for predictions

Due to $(y_1, y_2, \ldots, y_N, f(x^*))^T$ following Gaussian distribution also, the prediction at the location x^* can be obtained with the mean and variance:

$$E(f(x^*)|D) = k^* K_y^{-1} y \tag{6}$$

$$var(f(x^*)|D) = k(x^*, x^*) - k^* K_y^{-1} k^* \tag{7}$$