Journal of Process Control 24 (2014) 1454–1461

Contents lists available at ScienceDirect

Journal of Process Control

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

Active learning strategy for smart soft sensor development under a small number of labeled data samples

Zhiqiang Ge*

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

ARTICLE INFO

Article history:
Received 3 October 2013
Received in revised form 2 June 2014
Accepted 23 June 2014
Available online 14 July 2014

Keywords:
Soft sensor
Smart
Active learning
Unlabeled data samples
Data-based modeling

ABSTRACT

This contribution proposes a new active learning strategy for smart soft sensor development. The main objective of the smart soft sensor is to opportunely collect labeled data samples in such a way as to minimize the error of the regression process while minimizing the number of labeled samples used, and thus to reduce the costs related to labeling training samples. Instead of randomly labeling data samples, the smart soft sensor only labels those data samples which can provide the most significant information for construction of the soft sensor. In this paper, without loss of generality, the smart soft sensor is built based on the widely used principal component regression model. For performance evaluation, an industrial case study is provided. Compared to the random sample labeling strategy, both accuracy and stability have been improved by the active learning strategy based smart soft sensor.

1. Introduction

For prediction of key variables in the process industry, e.g. quality-related variables, various soft sensing (virtual measurement) methods have been developed, including first-principle model-based and data-based soft sensors [1–16]. Compared to those which mainly based on the knowledge of the process, the data-based soft sensing methods have been used more widely, thanks to the easy measurement and acquirement of data from industrial processes. While most process variables can be easily obtained by the distributed control system, some other variables are difficult to measure online, which are often related to the product quality, key process conditions, etc. The data-based soft sensor tries to build a regression model between difficult-to-measure variables and easy-to-measure variables. In the present paper, the difficult-to-measure variables are referred to key or quality variables and the easy-to-measure variables are referred to secondary process variables.

Conventionally, in order to build a data-based soft sensor for online prediction of the key variables in the process, the values of key variables in the training dataset should be determined manually, e.g. human expert, laboratory analysis, and so on. As a result, significant efforts will be incorporated to perform the labeling task of the key variables, which are time-consuming and costly in terms of human resource and money. Therefore, in practice, we may only have a limited number of labeled data samples and hold a large number of unlabeled data samples. If only those labeled data samples are used for construction of the soft sensing model, the performance of the model may not be guaranteed, which leads to poor estimation/prediction accuracy for online new data samples.

The motivation of this paper is to get use of the unlabeled data samples, or precisely, try to minimize the number of unlabeled data samples to be labeled by human efforts. From a modeling view, it starts from a small number of training data samples, then additional samples are selected properly from a large amount of unlabeled dataset. During this process, selecting the most significant unlabeled samples which consist of additional data information becomes particularly important. Therefore, if those selected data samples are labeled, the performance of the soft sensor model could be mostly improved. In this paper, this modeling process is termed as the active learning strategy. With the incorporation of the active learning strategy, lots of human efforts can be saved for construction of the soft sensor, while the prediction performance keeps high.

Despite the importance of this modeling problem for soft sensing in practice, it has rarely been researched to date. In this paper, without loss of generality, the active learning strategy is introduced and incorporated with the widely used principal component regression model for soft sensing of key variables in the process

* Tel.: +86 87951442.
E-mail addresses: gezhiqiang@zju.edu.cn, gezhiqiang@gmail.com

http://dx.doi.org/10.1016/j.jprocont.2014.06.015
0959-1524/

industry. However, this new soft sensor modeling idea has no limitation on the structure of the data model, which means it can be easily extended to other soft sensor modeling methods, such as Neural Network and SVMs. The first step of the new active learning PCR (ALPCR) method is to build a PCR model for the initial labeled data samples. Then, new data samples are selected from the unlabeled dataset, which will be labeled for the next learning step. The iterative learning process can be controlled by defining a stop criterion which is usually based on the converge speed of the learning process. In the ALPCR method, the most important issue is how to determine the selected data samples for labeling in the next step. In this paper, we intend to select the most significant unlabeled data samples based on the feature space of the previous PCR model. By constructing an effective selection statistic, the significance of each unlabeled data sample can be evaluated and arranged in an descend order, based on which a number of most significant unlabeled data samples are picked out for labeling, and then added to the training dataset for the next learning step.

The rest of this paper is organized as follows. In Section 2, detailed methodology of the new active learning strategy based PCR model is demonstrated for soft sensor development, followed by an industrial case study in the next section. Finally, conclusions are made.

2. Methodology

In the present paper, it is assumed that we only have a very limited number of labeled training data samples for modeling. Meanwhile, the industrial process has collected a large amount of unlabeled data samples. The aim of the active learning PCR model is to select some appropriate data samples for labeling, in order to improve the performance of the soft sensor. Simultaneously, due to expensive costs and significant efforts that may be caused for data sample labeling, keeping a low number of training samples is important to the industrial process. Here, the principal component regression model is used as the example for active learning of the data-based soft sensing method. However, the idea can be extended to other soft sensing methods, such as partial least squares, independent component regression, artificial neural networks, etc.

Given the labeled and unlabeled datasets as $\{\mathbf{X}_L\} \in R^{n_l \times m}$ and $\{\mathbf{X}_U\} \in R^{n_u \times m}$, where n_l and n_u are numbers of labeled and unlabeled data samples, usually $n_l \ll n_u$ holds, and m is the number of process variables. Suppose the dataset of the quality variables is given as $\{\mathbf{Y}_L\} \in R^{n_l \times r}$, where r is the number of quality variables. For soft sensing purpose, conventionally, the PCR model is constructed between datasets $\{\mathbf{X}_L\} \in R^{n_l \times m}$ and $\{\mathbf{Y}_L\} \in R^{n_l \times r}$, given as follows [8]

$$\mathbf{X}_L = \mathbf{T}_L \mathbf{P}_L^T + \mathbf{E}_L \tag{1}$$

$$\mathbf{Y}_L = \mathbf{T}_L \mathbf{C}_L^T + \mathbf{F}_L \tag{2}$$

where $\mathbf{P}_L \in R^{m \times k}$ is the loading matrix of the PCR model, $\mathbf{T}_L \in R^{n_l \times k}$ is the principal component matrix, k is the selected number of principal components, which can be determined by the cumulative percentage variance (CPV) method, $\mathbf{C}_L \in R^{l \times k}$ is the regression matrix corresponding to the quality variable matrix and the principal component matrix, $\mathbf{E}_L$ and $\mathbf{F}_L$ are the residuals matrices of $\mathbf{X}_L$ and $\mathbf{Y}_L$ with appropriate dimensions. Due to the small number of training data samples, the modeling and regression performance of the PCR model may not be well guaranteed. To this end, the active learning strategy is incorporated into the PCR model, in order to boost its performance by labeling an appropriate number of unlabeled data samples which contain additional information and can provide a compensation effect to the original PCR model.

Based on the feature of PCR model, the whole process variable space can be divided into two subspaces, termed as principal component subspace (PCS) and residual subspace (RS). A statistic for measuring the distance between the data samples and the modeling space can be constructed in each of the two subspaces, respectively. In the principal component subspace, the distance between a new data sample $\mathbf{x}_u \in \{\mathbf{X}_u\}$ and the model space can be measured by the Mahalanobis distance, thus

$$T_u^2 = \mathbf{t}_u^T \mathbf{\Lambda}^{-1} \mathbf{t}_u \tag{3}$$

where $\mathbf{t}_u = \mathbf{P}_L^T \mathbf{x}_u$ is the extracted principal component of the new data sample and $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the eigenvalues of the PCR model. Therefore, if a data sample stays inside of the PCR model, its T^2 statistic keeps as a low value, otherwise, it should be a big T^2 value in the principal component subspace. Compared to the data samples which stay inside of the PCR model, those which stay far away of the PCR model or have relatively big T^2 statistic values can provide more additional information for soft sensor modeling. Hence, we should set a higher priority for selection of these far away unlabeled data samples in the next learning procedure.

Similarly, in the residual subspace of the PCR model, another statistic can be constructed for measuring the distance between the new data samples and the PCR model, given as follows

$$\begin{aligned} \mathbf{e}_u &= \mathbf{x}_u - \mathbf{P}\mathbf{P}^T \mathbf{x}_u \\ SPE_u &= \mathbf{e}_u^T \mathbf{e}_u \end{aligned} \tag{4}$$

where $\mathbf{e}_u$ is the residual information of the new data sample. Therefore, if the residual of an unlabeled data sample is small, we can say that this data sample follows the PCR model, thus little additional information can be provided. On the other hand, if another unlabeled data sample shows a significant residual value, it means that this data sample may stay outside of the modeling space of the PCR model, thus could provide additional information if it is incorporated for modeling.

A confidence limit for each of the T^2 and SPE statistics can be determined as follows [17]

$$\begin{aligned} T_{\lim}^2 &= \frac{k(n-1)}{n-k} F_{k,(n-k),\alpha} \\ SPE_{\lim} &= g\chi_{h,\alpha}^2 \end{aligned} \tag{5}$$

where k is the number of PCs, α is significance level, $g = \mathrm{var}(SPE)/[2mean(SPE)]$ and $h = 2[mean(SPE)]^2/\mathrm{var}(SPE)$; $mean(SPE)$ and $\mathrm{var}(SPE)$ are mean and variance values of SPE for the training dataset. Based on these two confidence limits, a combined unlabeled data sample evaluation index can be defined as follows

$$Q_u = \frac{\sqrt{e^{-T_u^2/T_{\lim}^2}} + \sqrt{e^{-SPE_u/SPE_{\lim}}}}{2} \tag{6}$$

Based on this definition, it can be easily inferred that the value of the Q index is between zero and one. When the value approaches to one, it means the corresponding data sample has a high credit to stay inside of the previous PCR model space. Otherwise, if the credit value is very small, it means that the corresponding unlabeled sample has probably violated the PCR model, and thus should be treated as a high priority sample to be selected for the next modeling step. Based on the confidence limit of the T^2 and SPE statistics, a corresponding limit of the evaluation index Q can also be defined as follows

$$Q_{\lim} = \frac{\sqrt{e^{-1}} + \sqrt{e^{-1}}}{2} = 0.6065 \tag{7}$$

In practice, there are two sample selection strategies, one is to use a fixed number in each active learning step, the other one is to determine cut-off value of the Q index. While the determination of the cut-off value varies from process to process, it is much easier to set a fixed number of unlabeled data samples which have