Chemometrics and Intelligent Laboratory Systems 109 (2011) 197–206

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A soft sensor method based on values predicted from multiple intervals of time difference for improvement and estimation of prediction accuracy

Hiromasa Kaneko, Kimito Funatsu *

Department of Chemical System Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113–8656, Japan

A R T I C L E   I N F O

Article history:
Received 2 May 2011
Received in revised form 23 August 2011
Accepted 12 September 2011
Available online 17 September 2011

Keywords:
Soft sensor
Time difference
Prediction error
Applicability domain
Distance to model
Ensemble prediction

A B S T R A C T

Soft sensors are widely used to estimate process variables that are difficult to measure online. However, their predictive accuracy gradually decreases with changes in the state of the plants. We have been constructing soft sensor models based on the time difference of an objective variable, $\boldsymbol{y}$, and that of explanatory variables (time difference models) for reducing the effects of deterioration with age such as the drift without model reconstruction. In this paper, we have attempted to improve and estimate the prediction accuracy of time difference models, and proposed to handle multiple $\boldsymbol{y}$-values predicted from multiple intervals of time difference. A weighted average is a final predicted value and the standard deviation is an index of its prediction accuracy. This method was applied to real industrial data and then, could predict more number of data with higher predictive accuracy and estimate the prediction errors more accurately than traditional ones.

1. Introduction

Soft sensors have been widely used to estimate process variables that are difficult to measure online [1,2]. An inferential model is constructed between those variables that are easy to measure online and those that are not, and an objective variable, $\boldsymbol{y}$, is then estimated using that model. Through the use of soft sensors, the values of objective variables can be estimated with a high degree of accuracy. Their use, however, involves some practical difficulties. One crucial difficulty is that their predictive accuracy gradually decreases due to changes in the state of chemical plants, catalyzing performance loss, sensor and process drift, and the like. In order to reduce the degradation of a soft sensor model, the updating of regression models [3,4] and Just-In-Time (JIT) modeling [5,6] have been proposed. While many excellent results have been reported based on the use of these methods, there remain some problems for the introduction of soft sensors into practice [7–9].

First of all, if soft sensor models are reconstructed with the inclusion of any abnormal data, their predictive ability can deteriorate [4,10]. Though such abnormal data must be detected with high accuracy in real time, under present circumstances it is difficult to accurately detect all of them. Second, reconstructed models have a high tendency to specialize in predictions over a narrow data range [11,12]. Subsequently, when rapid variations in the process variables occur, these models cannot predict the resulting variations in data with a high degree of accuracy. Third, if a soft sensor model is reconstructed, the parameters of the model, for example, the regression coefficients in linear regression modeling, are dramatically changed in some cases. Without the operators' understanding of a soft sensor model, the model cannot be practically applied. Whenever soft sensor models are reconstructed, operators check the parameters of the models so they will be safe for operation. This takes a lot of time and effort because it is not rare that tens of soft sensors are used in a plant [13]. Fourth, the data used to reconstruct soft sensor models are also affected by the drift. In the construction of the model, data must be selected from a database which includes both data affected by the drift and data after correction of the drift.

In order to solve these problems, it was proposed to construct 'time difference models' that are soft sensor models based on the time difference of explanatory variables, $\boldsymbol{X}$, and that of $\boldsymbol{y}$ for reducing the effects of deterioration with age such as the drift and gradual changes in the state of plants without reconstruction of the models [7,8]. In other words, models which are not affected by these changes must be constructed using not the values of process variables, but the time difference in soft sensor modeling. The problems inherent in model reconstruction as described earlier can be avoided because time difference models do not have to be reconstructed and the data is represented as the time difference that cannot be affected by the drift. Besides, time difference models were applied in case that there exists the nonlinearity of process variables [9] showed through the analysis of actual industrial data that the time difference model maintained its predictive accuracy for a period of three years, even when the model was never reconstructed. However, its predictive accuracy was lower than that of the updated model [8]. One of the reasons would be that time difference model could not account for variation changing over time.

* Corresponding author. Tel.: +81 35841 7751; Fax: +81 5841 7771.
E-mail address: funatsu@chemsys.t.u-tokyo.ac.jp (K. Funatsu).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.09.003

On the one hand, it is important to separate variations in the process variables and a $\boldsymbol{y}$-analyzer fault for process monitoring, fault detection, and so on [4,13]. Kaneko et al. introduced applicability domains (ADs) and distances to models (DMs) [14–19] concepts, which is researched mainly in the field of quantitative structure–activity relationship analysis [20], to soft sensors and then, proposed to construct the relationships between ADs and the accuracy of prediction of soft sensor models quantitatively and estimated the prediction accuracy of new data online. The larger the DMs are, the lower the estimated accuracy of prediction would be. Kaneko et al. used the distances to the average of training data and to the nearest neighbor of training data as DMs, obtained the relationships between the DMs and prediction accuracy quantitatively, and then, false alarms could be prevented by estimating large prediction errors when the state was different from that of training data; further, actual $\boldsymbol{y}$-analyzer faults could be detected with certain accuracy [11]. Improvement of the ability for estimating the prediction errors or the predictive accuracy of soft sensor models is desired for process control.

It is expected that we can estimate prediction accuracy without the effects of deterioration with age such as the drift and gradual changes in the state of plants by using time difference, but DMs cannot be simply modified to time difference since time difference is an amount of change and thus, values of time difference in unsteady-states can be the same values in steady-states. It is impossible for simple DMs with time difference to estimate prediction accuracy in this situation.

Therefore in this paper, our objectives are set as follows:

1. To improve the prediction accuracy of time difference models
2. To estimate the prediction accuracy precisely

Then, we have focused attention on an interval of time difference. In terms of prediction by using a time difference model, when an interval is small, the model could not account for difference in process variables between a state of a plant before a variation and a state after that. On the other hand, when an interval is large, the difference could be accounted for. Thus, we have proposed an ensemble prediction method handling multiple $\boldsymbol{y}$-values predicted by inputting multiple intervals of time difference of $\boldsymbol{X}$ into a time difference model.

We introduce three types of weight averages as a final predicted value; an average, a linear weighted average, and an exponentially weighted average of the multiple predicted values. For the two latter weighted averages, the larger a time interval is, the less or exponentially less a weight is, because time difference from the older value would have less influence on a target $\boldsymbol{y}$-value. Besides, predictive accuracy of a final predicted value will be high if variation in the multiple predicted values is small and vice versa. Therefore, we have proposed to use the standard deviation of the multiple predicted values (SD) as a DM, that is, an index of prediction accuracy of a final predicted value. By using the proposed ensemble prediction method, we can estimate both a final predicted value of $\boldsymbol{y}$ and its predictive accuracy.

The proposed method is not an ensemble learning method such as bagging [21] and boosting [22] but an ensemble prediction method. While it is expected that variance of final prediction errors could be small from the point of view of ensemble prediction concept [23], we construct a single model for one data set including one objective variable and use multiple $\boldsymbol{y}$-values predicted by inputting multiple intervals in prediction. Therefore, in practice, the proposed method can save a lot of time and effort for development and maintenance of soft sensors.

2. Method

We explain the proposed ensemble prediction method handling multiple $\boldsymbol{y}$-values predicted by inputting multiple intervals of time difference of $\boldsymbol{X}$ into a time difference model. Before that, we briefly introduce the time difference modeling method and traditional DMs as compared methods.

2.1. Time difference modeling method [8]

In a traditional procedure, modeling relationship between explanatory variables, $\boldsymbol{X}(t)$, and an objective variable, $\boldsymbol{y}(t)$, is done by regression methods after preparing data, $\boldsymbol{X}(t)$ and $\boldsymbol{y}(t)$, related to time t. In terms of prediction, the constructed model predicts the value of $y(t')$ with the new data $\boldsymbol{x}(t')$.

In time difference modeling, time difference of $\boldsymbol{X}$ and that of $\boldsymbol{y}$, $\Delta\boldsymbol{X}(t)$ and $\Delta\boldsymbol{y}(t)$, are first calculated between the present values, $\boldsymbol{X}(t)$ and $\boldsymbol{y}(t)$, and those in some time i before the target time, $\boldsymbol{X}(t-i)$ and $\boldsymbol{y}(t-i)$.

$$\Delta X(t) = X(t) - X(t-\boldsymbol{i}) \tag{1}$$

$$\Delta y(t) = y(t) - y(t-\boldsymbol{i}) \tag{2}$$

Then, relationship between $\Delta\boldsymbol{X}(t)$ and $\Delta\boldsymbol{y}(t)$ is modeled by regression methods.

$$\Delta y(t) = \mathrm{f}(\Delta X(t)) + e \tag{3}$$

where f is a regression model and $\boldsymbol{e}$ is a vector of calculation errors. In terms of prediction, the constructed model, f, predicts the time difference of $y(t')$, $\Delta y(t')$, using the time difference of the new data, $\Delta\boldsymbol{x}(t')$, calculated as follows:

$$\Delta x(t') = x(t') - x(t'-i) \tag{4}$$

$$\Delta y_{\mathrm{pred}}(t') = \mathrm{f}(\Delta x(t')) \tag{5}$$

$y_{\mathrm{pred}}(t')$ can be calculated as follows:

$$y_{\mathrm{pred}}(t') = y(t'-i) + \Delta y_{\mathrm{pred}}(t') \tag{6}$$

because $y(t'-i)$ is given previously. This method can be easily expanded to a case that an interval i is not constant. By constructing time difference models, the effects of deterioration with age such as the drift and gradual changes in the state of plants can be accounted for, because data is represented as time difference that cannot be affected by these factors.

2.2. Traditional DM

Previously, we proposed a method to estimate the relationships between DMs and the accuracy of prediction of soft sensor models quantitatively [11]. For example, the Euclidean distance to the average of training data (ED) is used as a DM. The ED of explanatory variables of data, $\mathbf{x}$, is defined as follows:

$$\mathrm{ED} = \sqrt{(x-\mu)(x-\mu)^{\mathrm{T}}} \tag{7}$$

where $\boldsymbol{\mu}$ is a vector of the average of training data. When there is correlation among the variables, the Mahalanobis distance [24] is often used as the distance. The MD of $\boldsymbol{x}$ is defined as follows:

$$\mathrm{MD} = \sqrt{(x-\mu)\sum\nolimits^{-1}(x-\mu)^{\mathrm{T}}} \tag{8}$$

where $\sum$ is the variance–covariance matrix of training data. The absolute prediction errors will increase with the DMs, and their distributions will become wider. By quantifying the relationships between these distances and an index of prediction errors, we can estimate the index of the prediction errors for test data.