Applicability Domains and Accuracy of Prediction of Soft Sensor Models

Hiromasa Kaneko, Masamoto Arakawa, and Kimito Funatsu
Dept. of Chemical System Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku,
Tokyo 113-8656, Japan

DOI 10.1002/aic.12351
Published online July 30, 2010 in Wiley Online Library (wileyonlinelibrary.com).

Soft sensors are used to estimate process variables that are difficult to measure online. However, the predictive accuracy gradually decreases with changes in the state of chemical plants. Regression models can be updated, but if the model is updated with abnormal data, the predictive ability deteriorates. In practice, when the prediction error of an objective variable exceeds a threshold, an abnormal situation is detected. However, no effective method exists to decide this threshold. We have proposed a method to estimate the relationships between applicability domains and the accuracy of prediction of soft sensor models quantitatively. The larger the distances to models (DMs), the lower the estimated accuracy of prediction. Hence, the model between DMs and accuracy can separate variations in process variables and y*-analyzer fault. This method was applied to real industrial data. The fault detection ability of the proposed method was better than that of the traditional one.* © 2010 American Institute of Chemical Engineers *AIChE J,* 57: 1506–1513, 2011

Keywords: soft sensor, process control, fault detection, prediction error, applicability domain

Introduction

During the operation of chemical plants, an operator has to monitor the operating conditions of the plants and control process variables, such as temperature, pressure, liquid level, and concentration of products. Therefore, these variables have to be measured online; however, it is not easy to measure all variables online because of technical difficulties, large measurement delays, and high investment costs. Therefore, in chemical plants, soft sensors have been widely used to estimate the 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 is estimated by the model. In particular, the partial least squares (PLS) method[3,4] has been used as the modeling method for soft sensors. In addition, various methods, such as a nonlinear PLS method,[5,6] methods using artificial neural network,[7,8] and support vector machine-based regression methods[9–11] have been researched for use as soft sensor methods. By using soft sensors, the values of objective variables can be estimated with high accuracy.

However, there are some practical difficulties with soft sensors. A crucial difficulty is that the predictive accuracy gradually decreases due to changes in the state of chemical plants, catalyzing performance loss, sensor and process drift, and so on. To reduce the degradation of the soft sensor model, regression models are updated with new data.[12–17] By updating the models, they can follow gradual changes in process variables. However, these methods also have certain problems.

First, if soft sensor models are updated with any abnormal data, their predictive ability might deteriorate. Any abnormal data have to be detected with high accuracy. In actual plants, a threshold value of the prediction errors of an objective

Correspondence concerning this article should be addressed to K. Funatsu at funatsu@chemsys.t.u-tokyo.ac.jp.

variable (y) is set and fixed with training data, and when the prediction error of y exceeds a threshold value, it is considered to be abnormal. However, it is difficult to detect abnormal data and determine the reasons.[18] Prediction errors increase not only due to a y-analyzer fault but also due to variations in the process variables caused by changes in the state of chemical plants.

Second, when soft sensor models are updated, information on previous important variations is gradually lost. If models are updated only with steady-state data for a long time, they begin to specialize in predictions over a narrow data range. Subsequently, when variations in the process variables occur, these models cannot predict the variations in data with high accuracy. In such cases, the accuracy of prediction is lower than that in a steady state. However, if these variations of data can be predicted with the same threshold value as that in a steady state, these will be wrongly judged as being a y-analyzer fault when the threshold value is exceeded. In addition, the accuracy of prediction must be low when unknown variations occur in chemical plants.

To solve these problems, we have proposed a method to quantitatively estimate the relationships between applicability domains (ADs) and the accuracy of prediction of soft sensor models. The larger the distances to models (DMs) are, the lower the estimated accuracy of prediction would be. In this manner, variations in the process variables and a y-analyzer fault can be separated. Studies on the ADs of statistical models have been performed primarily in the field of quantitative structure-activity relationship analysis.[19–23] Some excellent results were obtained in these studies. In this article, the distance to the average of training data and the distance to the nearest neighbor (NN) of training data are used as DMs. We analyze the relationships between these DMs and the standard deviation of the prediction errors quantitatively. To verify the availability of the proposed method, we apply this method to real industrial data.

First, the relationships between the DMs of the data monitored in various states of the plant and the standard deviation of prediction errors were obtained. Then, we obtained a prediction for the test data and attempted to detect y-analyzer faults by using this relationship. We compared the results of fault detection obtained using both the proposed and traditional methods—the standard deviation of prediction errors was set as constant—and verified the availability of the proposed method.

Method

The proposed method predicting the predictive accuracy on soft sensors is explained in this section. In addition, we briefly introduce the 3-sigma method as a traditional fault detection method.

3-Sigma method

In chemical plants, the 3-sigma method is widely used as statistical quality control. The sigma is the standard deviation and this means variability or dispersion. The 3-sigma accounts for 99.7% of the sample population, assuming the distribution is normal. Prediction errors of soft sensors (y_{err}) are as follows:

$$y_{\mathrm{err}} = y_{\mathrm{obs}} - y_{\mathrm{pred}} \tag{1}$$

where y_{obs} denote the actual y value; and y_{pred}, the predicted y value. Then, the standard deviation of the prediction errors, that is, the sigma is defined as follows:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_{\mathrm{err},i} - \overline{y_{\mathrm{err}}}\right)^2} \tag{2}$$

where n denotes the number of samples. Therefore, upper control limit and lower control limit are as follows:

$$\begin{aligned} UCL &= \overline{y_{\mathrm{err}}} + 3\sigma \\ LCL &= \overline{y_{\mathrm{err}}} - 3\sigma \end{aligned} \tag{3}$$

If a prediction error of test data exceeds these limits, it is diagnosed as abnormal. In this method, the sigma is set as constant.

Proposed method predicting the predictive accuracy on soft sensors

We have proposed a method to quantitatively estimate the relationships between ADs and the accuracy of prediction of soft sensor models. First, the relationships between the DMs of training and/or validation data and absolute prediction errors were calculated on the basis of soft sensor modeling results. In this study, we use the distance to the average of training data (DM1) and the distance to the NN of training data (DM2) as the DMs. DM1 and DM2 of kth sample are defined as follows:

$$\begin{aligned} \mathrm{DM\,1} &= \sqrt{\sum_{i=1}^{d}\left(x_{k,i} - \overline{x_i}\right)^2} \\ \mathrm{DM\,2} &= \sqrt{\sum_{i=1}^{d}\left(x_{k,i} - x_{\mathrm{nearest},i}\right)^2} \end{aligned} \tag{4}$$

where d denotes the number of variables; and x_{nearest}, the NN of training data. The absolute prediction errors will increase with the DMs, and their distributions will become wider.

Subsequently, the relationships between DMs and the standard deviation of prediction errors are constructed. First, the data are sorted in ascending order of DMs. Then, we take every n samples from the top and calculated the standard deviation of the prediction errors and the average of the DMs. This calculation is maintained moving n samples.

For estimating standard deviations of the prediction errors for test data, the relationships between the DMs and the standard deviation of prediction errors are complemented with a line. Then, the standard deviations are calculated by using these relationships and the DMs of the test data.

Statistics

To compare the accuracy and the predictive accuracy of statistical models, r^2 and q^2 values and RMSE (root mean square error) are used as the measure in this study and defined as follows:

$$r^2 = 1 - \frac{\sum\left(y_{\mathrm{obs}} - y_{\mathrm{calc}}\right)^2}{\sum\left(y_{\mathrm{obs}} - \bar{y}\right)^2} \tag{5}$$

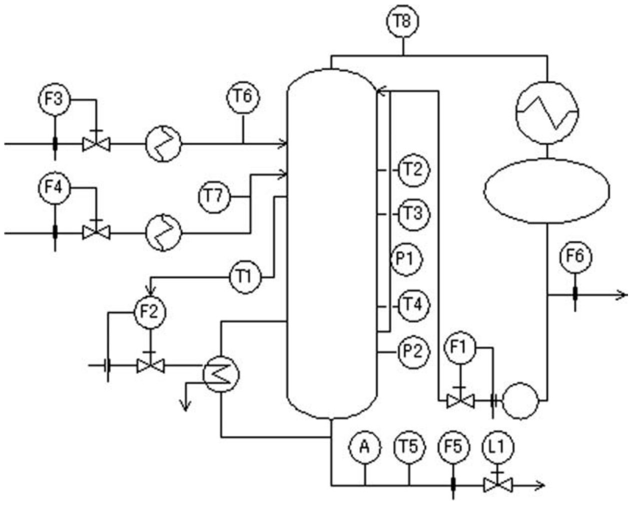


Figure 1. A schematic representation of the distillation column.

$$q^2 = 1 - \frac{\sum \left(y_{\mathrm{obs}} - y_{\mathrm{pred}}\right)^2}{\sum \left(y_{\mathrm{obs}} - \bar{y}\right)^2} \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum \left(y_{\mathrm{obs}} - y_{\mathrm{calc,pred}}\right)^2}{n}} \tag{7}$$

where y_{obs} denotes the actual y value; y_{calc}, the calculated y value; y_{pred}, the predicted y value in the procedure of cross-validation, such as leave-one-out; and n, the number of samples. The r^2_{pred}value is an r^2 value that is calculated with the validation data.

Results and Discussion

We analyzed the data obtained from an operation of a distillation column at Mizushima works, Mitsubishi Chemical Corporation. Figure 1 shows a schematic representation of the distillation column and Table 1 shows the process variables. An objective variable (y) represents the concentration of the bottom product having a lower boiling point, and explanatory variables (X) represent 18 variables, such as temperature and pressure. The input variables are $F3$ and $F4$, and the operational variables are $F1$ and $F2$. The sampling interval is 1 h, and we used the data monitored in 2002 and 2003. Figure 2 shows a plot of y. The objective variable y was transformed with an average of zero and a standard deviation of one. Variations that appeared at around A, B, C, and D are caused by disturbances, plant inspections, concentration analyzer faults, and plant tests, respectively. The variation C has to be detected as a y fault.

The variations A, B, and C were eliminated from 2002, and only steady-state data was retained. Training data was sampled from the steady-state data every 10 samples. For the validation data, we used the data of the variations in A, B, and the steady-state data omitting the training data. We then obtained relationships between DMs and the standard deviation of prediction errors. We made a prediction for data from 2003 and attempted to detect any trouble with the y analyzer as a fault by using these relationships; we then updated the regression model.

Table 1. Process Variables

	Symbol	Objective variable
	A	Bottom product concentration
No.	Symbol	Explanatory variables
1	F1	Reflux flow
2	F2	Reboiler flow
3	F3	Feed 1 flow
4	F4	Feed 2 flow
5	F5	Bottom flow
6	F6	Top flow
7	P1	Pressure 1
8	P2	Pressure 2
9	T1	Temperature 1
10	T2	Temperature 2
11	T3	Temperature 3
12	T4	Temperature 4
13	T5	Bottom temperature
14	T6	Feed 2 temperature
15	T7	Feed 1 temperature
16	T8	Top temperature
17	F1/F6	Feed flow ratio
18	F4/F3	Reflux ratio

Construction of soft sensor models

In this article, soft sensor models were constructed between y and 25 variables, which are 12 variables (No. 2, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18) in Table 1, and the 12 variables and y whose time lags are 1 h. The method that can be used to construct a PLS model with the dynamics of process variables is called a dynamic PLS (DPLS).[24,25] We constructed the DPLS model using the training data. In addition, the support vector regression (SVR) model was constructed with the dynamics of process variables for comparison with a nonlinear regression method. The details of PLS and SVR are shown in Appendix A and B. Table 2 shows the results of modeling by using these methods.

In PLS modeling, the number of components was 21, r^2 was 0.840, and q^2 calculated using a leave-one-out cross-validation method was 0.821. These values were considered to be sufficiently high. The r^2 value calculated using the validation data, except for the variations A and B, was 0.841, and this value was approximately equal to that of the training data. The RMSEs of training data and validation data, except

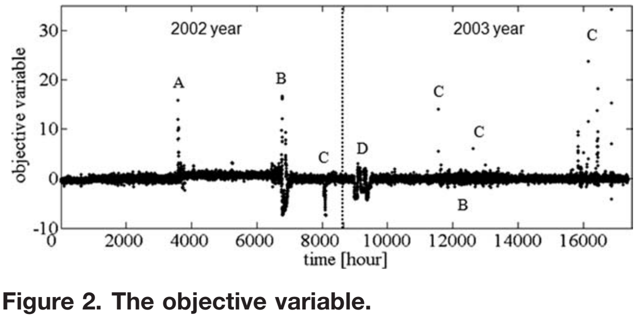


Figure 2. The objective variable.

The values were monitored in 2002 and 2003 and transformed with an average of zero and a standard deviation of one. Variations that appeared at around A, B, C, and D are caused by disturbances, plant inspections, concentration analyzer faults, and plant tests, respectively.

Table 2. Modeling Results

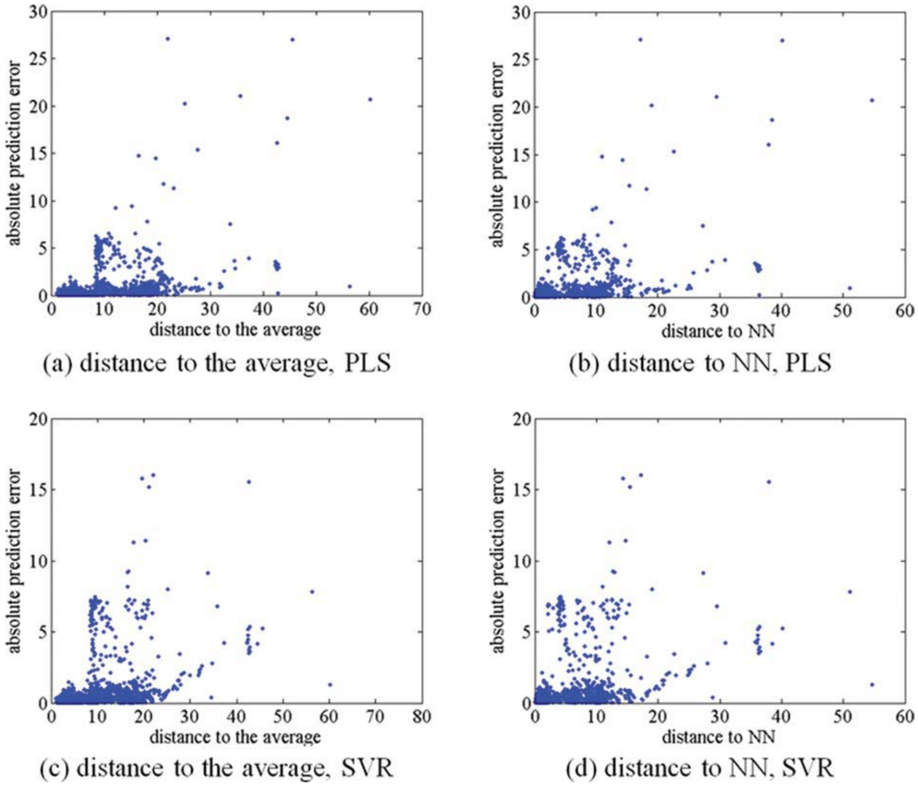
Data		PLS	SVR
Training	r^2	0.840	0.930
	RMSE	0.199	0.131
	q^2	0.821	0.822
	RMSE	0.210	0.210
Validation other than A and B	r^2_{pred}	0.841	0.869
	RMSE	0.201	0.182
Validation A and B	r^2_{pred}	−0.109	0.068
	RMSE	3.11	2.85

for the variations A and B, were also almost equal. However, the r^2 value calculated with the variations A and B was low, and hence, the predictive accuracy for the data whose state is different from that of the training data would be low. The tendencies in SVR modeling were identical to those in PLS. The prediction accuracy of a steady state is high, whereas that of an unsteady state is low.

Subsequently, the calculation results of the relationships between DMs of training and validation data and the absolute calculation errors are shown in Figure 3. The tendency is identical overall. The absolute prediction errors increased with the DMs, and their distributions became wider. In particular, when the distance to the average was more than 10, wherein considerable data on the variations A and B were included, the absolute prediction errors varied widely. Then, Figure 4 shows the relationships between DMs and standard deviations of prediction errors. In this study, the number of samples was set as 500 to calculate the standard deviation of the prediction errors. When the DMs were small, the standard deviations of prediction errors were small and almost constant. When the distances to the average became greater than ∼7, and the distances to the NN became more than ∼1.3, wherein considerable data on the variations in A and B were included, the standard deviations of prediction errors increased. Overall, the standard deviations of prediction errors increased with the DMs.

Prediction and fault detection

We predicted the values of y in 2003 to verify the prediction and fault detection accuracy of the proposed method. In this article, it is shown that Figure 4a was extensively researched. The other results are not shown here, but these were almost identical. First, the threshold values of the prediction errors of y for detecting the y-analyzer fault were set to i times the estimated standard deviations of the prediction errors, and i was increased from 0.1 to 5 in steps of 0.1. Then, a receiver operating characteristic (ROC) curve was drawn. An ROC curve is a plot whose horizontal axis represents the false alarm rate and the vertical axis represents the detection rate. The top-left area has high fault detection ability. For comparison, an ROC curve was constructed using the traditional method, in which a threshold value of the prediction error of y was i times the constant standard deviations of the prediction errors obtained using the training data. When i equals 3, this is the 3-sigma method. Figure 5



(a) distance to the average, PLS

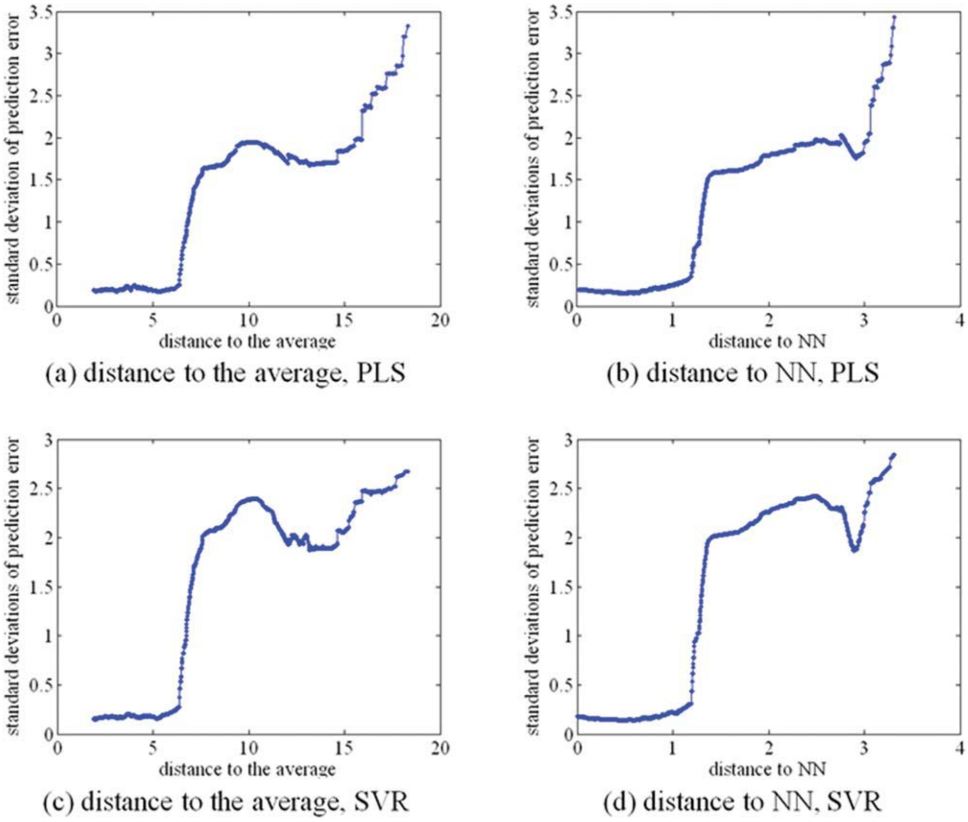
(b) distance to NN, PLS

(c) distance to the average, SVR

(d) distance to NN, SVR

Figure 3. The relationships between DMs and the absolute prediction errors.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]



(a) distance to the average, PLS

(b) distance to NN, PLS

(c) distance to the average, SVR

(d) distance to NN, SVR

Figure 4. The relationships between DMs and standard deviations of prediction errors.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

shows the results of the ROC curves. Open circles represent results using the proposed method, whereas dots represent those from the traditional one. The area under the ROC curve of the proposed method was 0.86, whereas that in the case of the traditional one was 0.83. From these results, the proposed method could detect abnormal data with higher detection rates and lower false alarm rates than the traditional method.

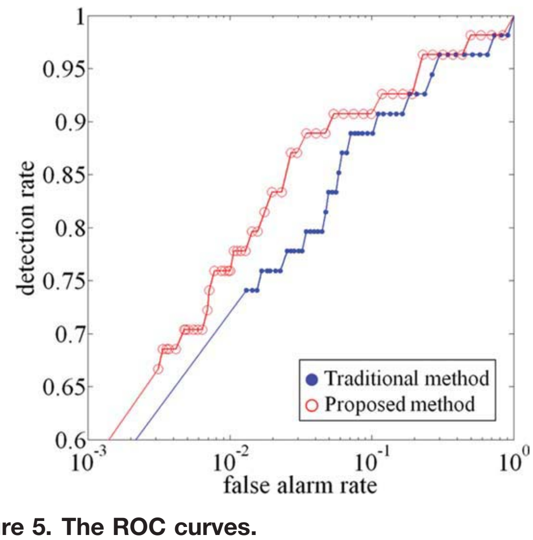


Figure 5. The ROC curves.

The horizontal axis is a common logarithmic scale. Open circles represent results using the proposed method, whereas dots represent those from the traditional one. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

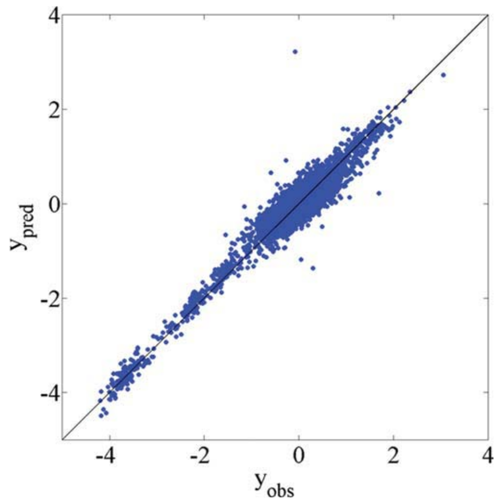
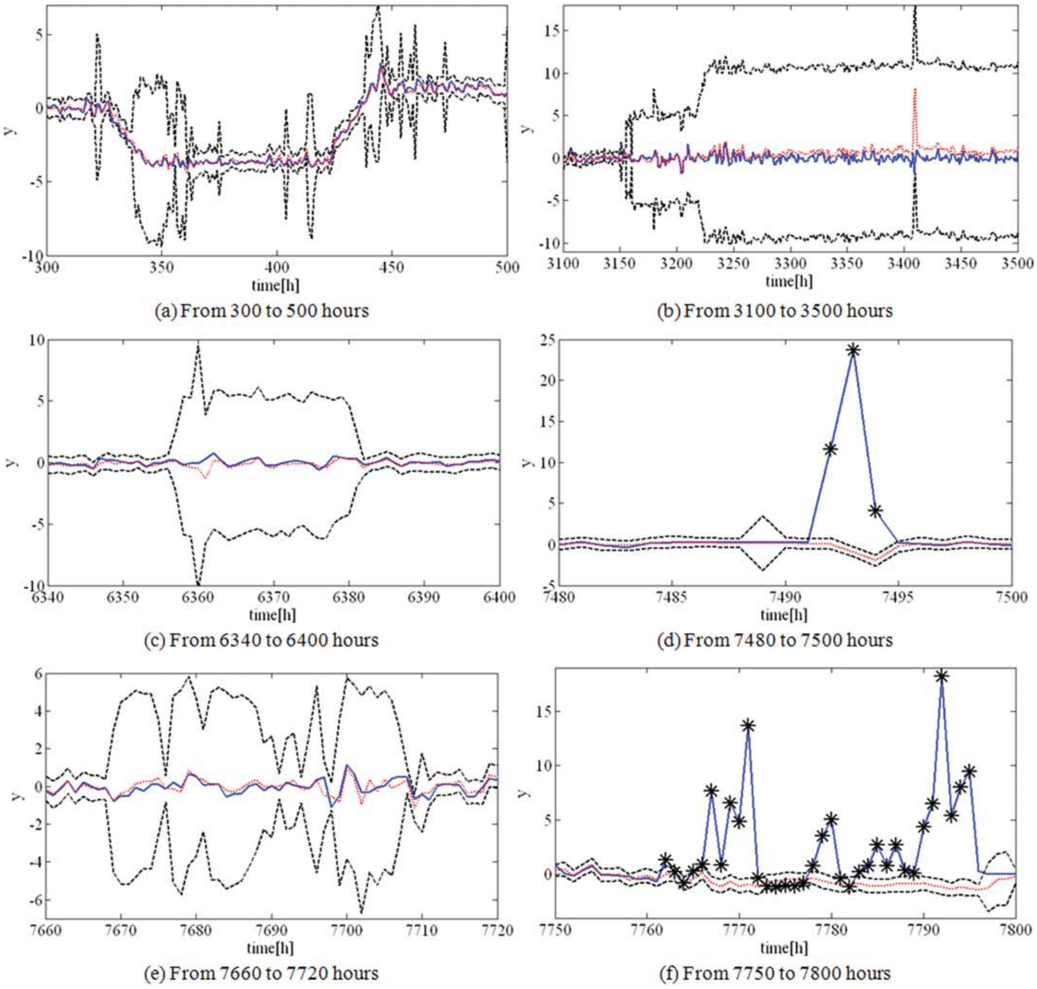


Figure 6. The relationship between measured and predicted y of 2003.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]



(a) From 300 to 500 hours

(b) From 3100 to 3500 hours

(c) From 6340 to 6400 hours

(d) From 7480 to 7500 hours

(e) From 7660 to 7720 hours

(f) From 7750 to 7800 hours

Figure 7. Prediction examples.

Black line, gray dashed line, black broken line, and black asterisks represent the measured values, predicted values, upper and lower limits of prediction errors, and the actual *y* analyzer, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

When *i* was set as 3, the detection rate and false alarm rate of the proposed method were considered well-balanced. In this case, the detection rate was 0.78 and false alarm rate was 0.01. However, the value of *i* did not affect the results largely. A plot of the measured and predicted *y* values after omitting the abnormal data is shown in Figure 6. The r^2 value and RMSE were 0.918 and 0.193, respectively. These values were sufficiently high and low, and the plot showed a considerably tight cluster of predicted values along the diagonal. This shows that a model with high predictive accuracy was constructed.

Figure 7 shows prediction examples. The values of the objective variable *y* predicted using the proposed method are shown for some periods. Black line, gray dashed line, and black broken line represent the measured values, predicted values, and upper and lower limits of prediction errors, respectively. When the black line is outside the black dashed line, the status is diagnosed as abnormal. Black asterisks represent the actual *y*-analyzer fault.

Figure 7a shows the results of the prediction for variation D in Figure 2. The predicted values followed the measured values well, even though the measured values changed largely. However, as all these variation data were not included in the latest data used to update the model, the predicted values for these variations may be unreliable. In fact, the difference from training data was estimated to be around 350 h, and the standard deviations of prediction errors were estimated to be large.

Figure 7b shows the results of predictions for the variation B in Figure 2. It is possible that the accuracy of prediction for B is low in 2002. When the variations increased beyond around 3150 h, the estimated standard deviations of the prediction errors increased. Thus, a variation of around 3400 was not mistaken as the *y*-analyzer fault.

The states shown in Figures 7c and e differ from those of training data. Unknown variations are considered to occur in these cases. In the case of around 6360 and 7670 h, the prediction errors were large. Hence, the variations around these hours were not mistaken as the *y*-analyzer fault.

Figures 7d and f show fault detection. When the prediction errors were larger than thrice that of the estimated standard deviations of prediction errors, they were diagnosed as a fault. Although all the abnormal data were detected in

Figure 7d, some were not detected in Figure 7f. However, the first fault could be detected in Figure 7f. In such a case, subsequent faults need not be considered if an operator of a plant corrects the fault appropriately. If the proposed method is used in the operations of chemical plants, the detection rate would increase, whereas the fault alarm rate would decrease.

Conclusions

To estimate the prediction accuracy of *y* faults in chemical operations, we have attempted to obtain the relationships between DMs and prediction accuracy quantitatively. Subsequently, by using real industrial data, we verified that absolute prediction errors increased with the DMs. Hence, false alarms can be prevented by estimating large prediction errors when the state is different from that of training data; further, actual *y*-analyzer faults can be detected with high accuracy. In this article, the distance to the average of training data and the distance to the NN of training data are used as DMs. However, considerable research has been conducted on ADs; an optimal DM of the soft sensor model increases the detection rate and decreases the fault alarm rate. In addition, the proposed method can be applied to any method used for constructing soft sensor models.

Acknowledgments

The authors acknowledge the support of Mizushima works, Mitsubishi Chemical and the financial support of Japan Society for the Promotion of Science.

Literature Cited

1. Kano M, Nakagawa Y. Data-based process monitoring, process control, and quality improvement: recent developments and applications in steel industry. *Comput Chem Eng.* 2008;32:12–24.
2. Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Comput Chem Eng.* 2009;33:795–814.
3. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst.* 2001;58:109–130.
4. Lin B, Recke B, Knudsen JKH, Jorgensen SB. A systematic approach for soft sensor development. *Comput Chem Eng.* 2007;31: 419–425.
5. Baffi G, Martin EB, Morris AJ. Non-linear projection to latent structures revisited (the neural network PLS algorithm). *Comput Chem Eng.* 1999;23:1293–1307.
6. Zhao SJ, Zhang J, Xu YM, Xiong ZH. Nonlinear projection to latent structures method and its applications. *Ind Eng Chem Res.* 2006; 453:843–3852.
7. Radhakrishnan VR, Mohamed AR. Neural networks for the identification and control of blast furnace hot metal quality. *J Process Control.* 2000;10:509–524.
8. Dai XZ, Wang WC, Ding YH, Sun ZY. Assumed inherent sensor inversion based ANN dynamic soft-sensing method and its application in erythromycin fermentation process. *Comput Chem Eng.* 2006;30:1203–1225.
9. Vapnik VN. *The Nature of Statistical Learning Theory*. New York: Springer, 1999.
10. Yan WW, Shao HH, Wang XF. Soft sensing modeling based on support vector machine and Bayesian model selection. *Comput Chem Eng.* 2004;28:1489–1498.
11. Lee DE, Song JH, Song SO, Yoon ES. Weighted support vector machine for quality estimation in the polymerization process. *Ind Eng Chem Res.* 2005;44:2101–2105.
12. Dayal BS, MacGregor JF. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *J Process Control.* 1997;7:169–179.
13. Qin SJ. Recursive PLS algorithms for adaptive data modeling. *Comput Chem Eng.* 1998;22:503–514.
14. Mu SJ, Zeng YZ, Liu RL, Wu P, Su HY, Chu J. Online dual updating with recursive PLS model and its application in predicting crystal size of purified terephthalic acid (PTA) process. *J Process Control.* 2006;16:557–566.
15. Fu YF, Su HY, Chu JA. MIMO soft-sensor model of nutrient content for compound fertilizer based on hybrid modeling technique. *Chin J Chem Eng.* 2007;15:554–559.
16. Liu JL. On-line soft sensor for polyethylene process with multiple production grades. *Control Eng Pract.* 2007;15:769–778.
17. Kaneko H, Arakawa M, Funatsu K. Development of a new soft sensor method using independent component analysis and partial least squares. *AIChE J.* 2009; 55:87–98.
18. Ookita K. Operation and quality control for chemical plants by soft-sensors. *CICSJ Bull.* 2006;24:31–33. (in Japanese).
19. Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov Today.* 2006;11:700–707.
20. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD. Assessing applicability domains of toxicological QSARs: definition, confidence in predicted values, and the role of mechanisms of action. *QSAR Comb Sci.* 2007;26:238–254.
21. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J Chem Inf Model.* 2008;48:766–784.
22. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Öberg T, Todeschini R, Fourches D, Varnek A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model.* 2008;48:1733–1746.
23. Horvath D, Marcou G, Varnek A. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J Chem Inf Model.* 2009;49:1762–1776.
24. Kano M, Miyazaki K, Hasebe S, Hashimoto I. Inferential control system of distillation compositions using dynamic partial least squares regression. *J Process Control.* 2000;10:157–166.
25. Kamohara H, Takinami A, Takeda M, Kano M, Hasebe S, Hashimoto I. Product quality estimation and operating condition monitoring for industrial ethylene fractionator. *J Chem Eng Jpn.* 2004;37:422–428.
26. Chang CC, Lin CJ. LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Appendix A: PLS

PLS is a method for relating **X** and **y**, by a linear multivariate model, but goes beyond traditional regression methods in that it models also the structures of **X** and **y**. In PLS modeling, the covariance between score vector $\mathbf{t}_i$ and **y** is maximized. Generally, PLS models have higher predictive power than those of multiple linear regression.

A PLS model consists of the following two equations:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \tag{A1}$$

$$\mathbf{y} = \mathbf{Tq}' + \mathbf{f} \tag{A2}$$

where **P** is an **X**-loading matrix, **q** is a **y**-loading vector, **E** is a matrix of **X** residuals, and **f** is the vector of **y** residuals. The PLS-regression model is as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{const} \tag{A3}$$

$$\mathbf{b} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{q}' \tag{A4}$$

where **W** is an **X**-weight matrix and **b** is a vector of regression coefficients. To construct a highly predictive model, the number of components must be appropriately decided. q^2 values are used as the measure. In this study, the optimum number of components is determined by the first local maximum of q^2.

Appendix B: SVR

SVR is a method applying support vector machine (SVM) to a regression analysis and can construct nonlinear models by applying the kernel trick as well as SVM. Primal form of SVR can be shown to be a following optimization problem:

Minimize

$$\frac{1}{2}\|\mathbf{w}\|^2+C\sum_i |y_i - f(\mathbf{x}_i)|_e \tag{B1}$$

subject to

$$|y_i - f(\mathbf{x}_i)|_e = \max(0, |y_i - f(\mathbf{x}_i)| - e) \tag{B2}$$

where y_i and $\mathbf{x}_i$, are training data; is a weight vector; e is a threshold and C is a penalizing factor, which controls a trade-off between a training error and a margin. By minimization of Eq. (B.1), we can construct a regression model, which has a well balance between adaptive ability to the training data and generalization capability. A kernel function in our application is a radial basis function:

$$K(x, x') = e^{-\gamma * |x - x'|^2} \tag{B3}$$

where γ is a tuning parameter controlling width of the kernel function. In this study, an LIBSVM[26] is used for constructing the SVR models, and other calculations are carried out using in-house programs written in MATLAB.

Manuscript received Nov. 19, 2010, revision received Apr. 27, 2010, and final revision received Jun 22, 2010.