Applicability Domain of Soft Sensor Models Based on One-Class Support Vector Machine

Hiromasa Kaneko 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.14010
Published online January 18, 2013 in Wiley Online Library (wileyonlinelibrary.com)

Soft sensors are widely used to estimate process variables that are difficult to measure online. By using soft sensors, analyzer faults can be detected by estimation errors. However, it is difficult to detect abnormal data and determine the reasons because estimation errors increase not only due to analyzer faults but also due to variations caused by changes in the state of chemical plants. To separate those factors, we previously proposed to construct the relationships between distances to soft sensor models (DMs) and the accuracy of prediction of the models quantitatively and estimate the prediction accuracy of new data online. In this article, we used a one-class support vector machine (OCSVM) to estimate data density and the output of an OCSVM as a DM. The proposed method was applied to real industrial data and the superiority of the proposed DM to the traditional ones was demonstrated by comparing their results. © 2013 American Institute of Chemical Engineers *AIChE J*, 59: 2046–2050, 2013
Keywords: soft sensor, predictive accuracy, applicability domain, distances to model, one-class support vector machine

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 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 of the difficulties is the degradation of soft sensor models. The predictive accuracy of soft sensors 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, a regression model is updated with new data,[3,4] a new model is constructed in prediction,[5,6] or a time difference model, which is constructed between time difference of explanatory variables, $\mathbf{X}$, and that of an objective variable, $\mathbf{y}$,[7–9] is used, for example.

Meanwhile, if the degradation is reduced and the accuracy of soft sensor models is maintained, a $\mathbf{y}$-analyzer fault can be detected by prediction errors. In actual plants, a threshold value of the prediction errors of $\mathbf{y}$ is set and fixed with training data, and when the prediction error of $\mathbf{y}$ exceeds a threshold value, it is considered to be abnormal. However, it is difficult to detect abnormal data and determine the reasons.[4] Prediction errors increase not only due to a $\mathbf{y}$-analyzer fault but also due to variations in the process variables caused by changes in the state of chemical plants. In fact, reliability of a soft sensor model is affected by process conditions.

Kaneko et al.[10] introduced applicability domains (ADs) and distances to models (DMs) concepts, which are researched mainly in the field of quantitative structure–activity relationship analysis,[11–13] and obtained the relationships between a DM and prediction accuracy quantitatively using the distances to the average of training data as a DM. False alarms could be prevented by estimating large prediction errors when the state was different from that of training data and actual $\mathbf{y}$-analyzer faults could be detected with certain accuracy. Recently, Liu et al.[14] developed interval soft sensors that estimate not only values of $\mathbf{y}$ but also their confidence values. This is the same concept with the ADs representing reliability of soft sensor models.

The distances to the average of training data, however, cannot handle a nonlinear relationship among process variables. Additionally, if data distributions are multimodal, the distances cannot represent the true DM appropriately. Improvement of the ability for estimating the prediction errors or the predictive accuracy of soft sensor models is desired for process control.

We, therefore, propose to apply data density to a DM of soft sensors. A soft sensor model will be well-trained in data regions where data density is high and the predictive accuracy of the model will be high in the same regions. On the other hand, a model will not be trained enough in data regions where data density is low and the predictive accuracy of the model will be low in the same regions. In this research, we employ a one-class support vector machine (OCSVM)[13,15] to estimate data density and the output of an OCSVM as a DM.

In addition, we also propose an index of effectiveness of a DM. The proposed method is applied to real industrial data

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

and the superiority of the proposed DM to the traditional ones is demonstrated by comparing their results.

Method

One-class support vector machine

An OCSVM is a method in which an SVM is applied to a domain description problem. Given a set of training data in a high-dimensional input space, the objective of an OCSVM is to learn a function that will take the value $+1$ in the region where the majority of the data is concentrated and the value -1 everywhere else. The function to be learned is modeled as a hyperplane in a transformed space, and the hyperplane parameters are estimated so that its margin with respect to the training data is maximized, as dictated by the data-driven distribution–free paradigm.

The maximum margin solution of the OCSVM problem is obtained by solving the following quadratic optimization problem

Minimize

$$\frac{1}{2}\|\mathbf{w}\|^2+\frac{1}{\upsilon n}\sum_i \xi_i-b \tag{1}$$

subject to

$$\begin{aligned}\mathbf{w}\cdot\Phi(\mathbf{x}_i)&\geq b-\xi_i\\ \xi_i&\geq 0\end{aligned} \tag{2}$$

where $\mathbf{w}\in \mathrm{R}^n$ denotes a weight vector; b, a bias; ξ_i, slack variables; n, the number of training data; υ, the parameter that represents the upper bound on the fraction of outliers in the data; and Φ, a nonlinear function. Finally, the decision function inferred by the learned hyperplane is as follows

$$f(\mathbf{x})=\mathbf{w}\cdot\Phi(\mathbf{x})-b \tag{3}$$

By using a kernel function K, Eq. 3 is represented as follows

$$f(\mathbf{x})=\sum_{i=1}^{n}\alpha_i K(\mathbf{x}_i,\mathbf{x})-b \tag{4}$$

where $\mathbf{x}_i$ is the i^{th} training data and $K(\mathbf{x}_i, \mathbf{x})$ can be represented as the inner products of $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x})$ as follows

$$K(\mathbf{x}_i,\mathbf{x})=\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x})\rangle \tag{5}$$

In our application, a kernel function is a radial basis function

$$K(\mathbf{x}_i,\mathbf{x})=\exp\left(-\gamma\|\mathbf{x}_i-\mathbf{x}\|^2\right) \tag{6}$$

where γ is a tuning parameter that controls the width of the kernel function. The negative reciprocal of the output of an OCSVM $f(\mathbf{x})$ is the proposed DM of soft sensors because when data density is low, the prediction accuracy, that is, the error bar, will be large and then the value of a DM should be large and vice versa.

Area under coverage and root-mean-square error curve

The relationship between the coverage and the root-mean-square error (RMSE) is used to compare the performance of DMs.[8,13] First, data are sorted in ascending order of each DM and we calculate the coverage that is the rate of the number of the data within each AD to the total number of the data N_{all}. Then, the coverage of the i^{th} data is defined as follow

$$\text{Coverage}_i=i/N_{\text{all}} \tag{7}$$

The i^{th} RMSE value is calculated with the i data as follows

$$\text{RMSE}=\sqrt{\frac{\sum_{j=1}^{i}\left(y_{\text{obs},j}-y_{\text{calc;pred},j}\right)^2}{i}} \tag{8}$$

where y_{obs} is the measured $\mathbf{y}$-value; $y_{\text{calc,pred}}$ is the calculated $\mathbf{y}$-value for the i data in training data or the predicted $\mathbf{y}$-value for the i data in test data. It is desired for ADs that the smaller the values of the coverage are, the smaller the RMSE values are and vice versa. To compare DMs quantitatively, area under coverage and RMSE curve (AUCR) is introduced in this article. The basic concept of AUCR is shown in Figure 1. The relationship between the coverage and the RMSE of A is a more desired curve than that of B, which is mentioned above and the AUCR of A is smaller than that of B. It can be said that the performance of A as a DM is better than that of B.

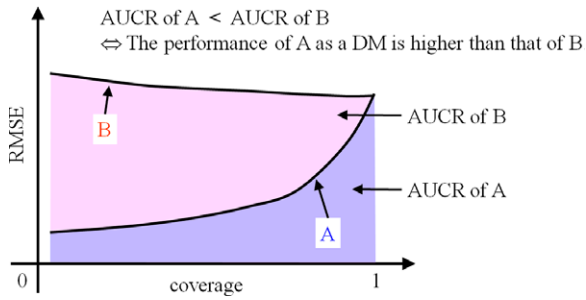

Figure 1. Basic concept of AUCR.

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

Results and Discussion

To verify the performance of the proposed method, two data sets were analyzed in this article. One data set is obtained from an operation of a distillation column at Mizushima works, Mitsubishi Chemical Corporation, and a relationship between $\mathbf{X}$ and $\mathbf{y}$ can be represented by a linear function. The other data set is obtained during an industrial polymer process at Mitsui Chemicals, and relationships between $\mathbf{X}$ and $\mathbf{y}$ are nonlinear.

Distillation column data

Figure 2 shows a schematic representation of the distillation column and Table 1 shows process variables used in this study. A $\mathbf{y}$-variable is the concentration of the bottom product having a lower boiling point; and $\mathbf{X}$-variables are that represent 19 variables such as temperature and pressure. The input variables are F3 and F4, and the operational variables are F1 and F2. The measurement interval of $\mathbf{y}$ is 30 min. For training data, we used data from monitoring that took place from January to March 2003, and for test data,

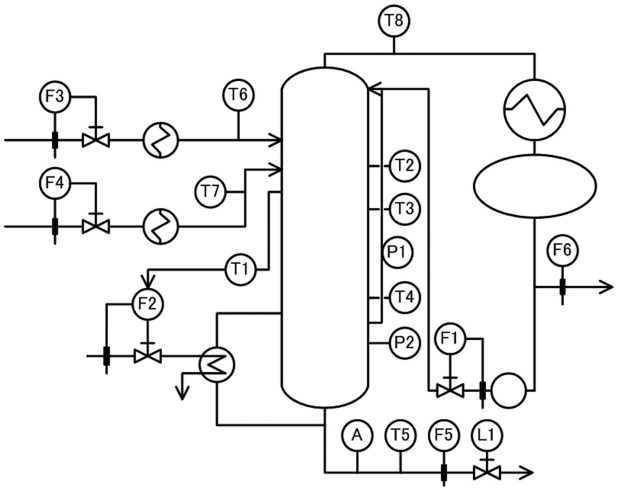

Figure 2. A schematic representation of the distillation column.

we used data from April 2003 to December 2006. Data that reflects variations caused by **y**-analyzer fault were eliminated in advance.

To incorporate the dynamics of process variables into soft sensor models, **X** included each explanatory variable that was delayed for durations ranging from 0 to 60 min in steps of 10 min, that is, **X** consists of seven time points (0, 10, 20, 30, 40, 50, and 60 min) times 19 variables ($7 \times 19 = 133$).

A soft sensor model was constructed between **X** and **y** with training data and then values of **y** were predicted by inputting values of **X** of test data into the model. We used a partial least squares (PLS) method[16] as a regression approach. Then, we calculated relationships between DMs and calculation or prediction errors. The Euclidean distance (ED) and the Mahalanobis distance (MD) to the average of training data, and the output of an OCSVM (the proposed method) were applied as DMs. In OCSVM modeling, a parameter ν was changed from 2^{-20} to 2^{-4} by squaring and from 0.1 to 0.9 in steps of 0.1; the other parameter γ was changed from 2^{-20} to 2^{10} by squaring; and accordingly 806 OCSVM models were constructed.

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	L1	Liquid level
8	P1	Pressure 1
9	P2	Pressure 2
10	T1	Temperature 1
11	T2	Temperature 2
12	T3	Temperature 3
13	T4	Temperature 4
14	T5	Bottom temperature
15	T6	Feed 1 temperature
16	T7	Feed 2 temperature
17	T8	Top temperature
18	F4/F3=R	Reflux ratio
19	F1/F6=F	Feed flow ratio

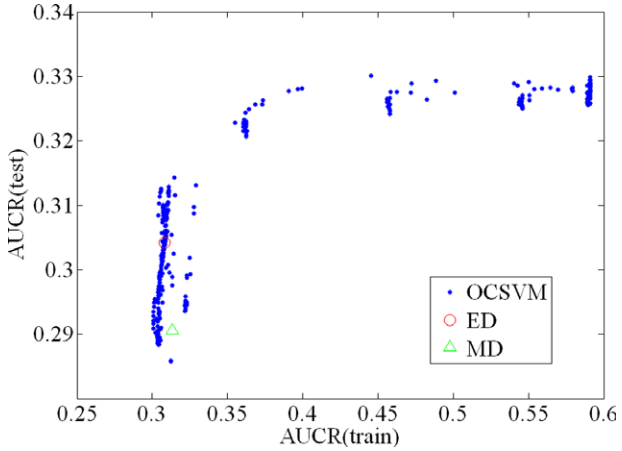

Figure 3. Relationships between AUCR of training data and that of test data for the distillation column.

The blue points represent the results of OCSVM; the red circle represents that of ED; and the green triangle represents that of MD. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 3 shows the relationship between AUCR of training data and that of test data for each DM. The blue points represent the results of OCSVM; the red circle represents that of ED; and the green triangle represents that of MD. The AUCRs of both training data and test data were relatively small when ED and MD were used as DMs. This will come from a linear relationship between **X** and **y**[4,10] and even ED and MD of **X** could represent information on ADs. Meanwhile, as shown in Figure 3, the some results of OCSVM were more left and lower than those of ED and MD, indicating the high performance of the proposed method.

The relationships between the coverage and the RMSE of test data are shown in Figure 4. The blue continuous line represents the results of OCSVM; the red dashed line represents the results of ED; and the green chain line represents the results of MD. The OCSVM parameters γ and ν were the

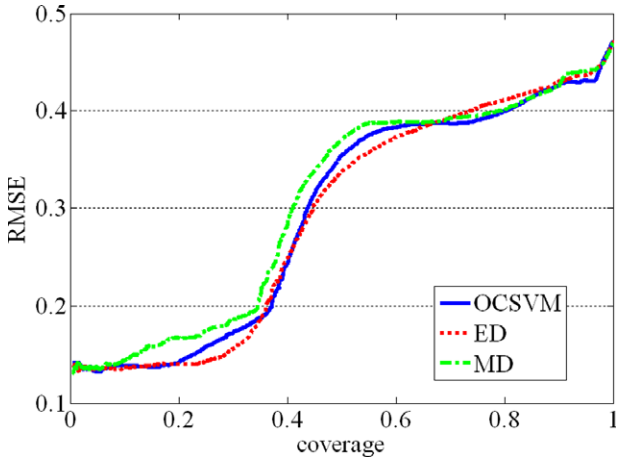

Figure 4. Relationships between the coverage and the RMSE for the distillation column.

The blue continuous line represents the results of OCSVM; the red dashed line represents the results of ED; and the green chain line represents the results of MD. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

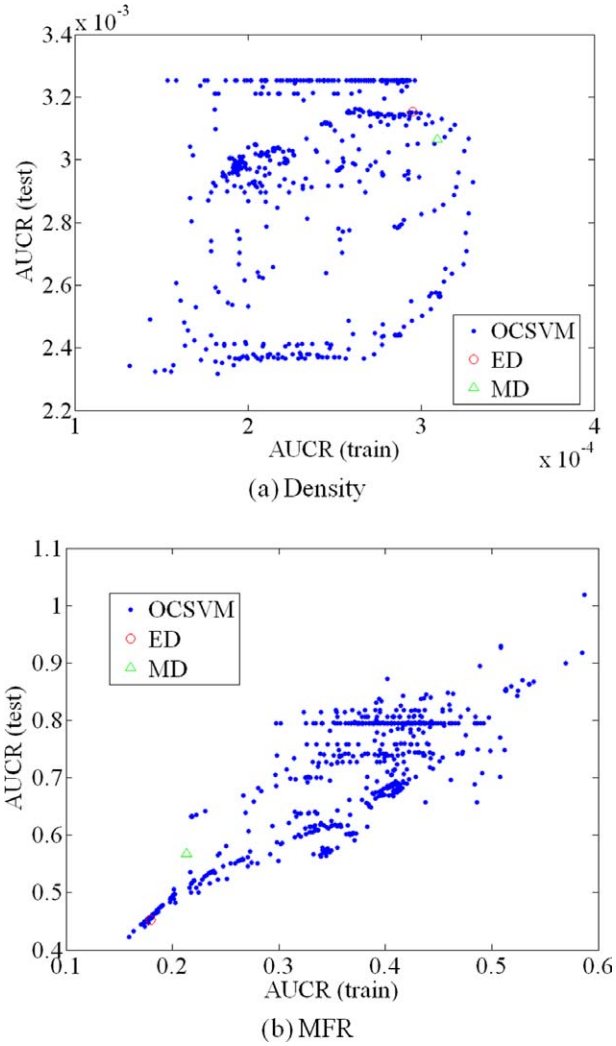

Figure 5. Relationships between AUCR of training data and that of test data for the industrial polymer process.

The blue points represent the results of OCSVM; the red circle represents that of ED; and the green triangle represents that of MD. (a) Density (b) MFR. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

ones where the AUCR value of training data was smallest. As shown in the results of OCSVM, ED, and MD, the smaller the values of the coverage are, the smaller the RMSE values are and there is little difference between the relationships obtained by OCSVM, ED, and MD. This means that the ADs calculated by OCSVM, ED, and MD appropriately represent the data region where the constructed soft sensor model has high predictive accuracy. We confirmed that the proposed method functioned well and had little difference with the traditional ones when there is a linear relationship between **X** and **y**.

Polymer process data

Then, we applied the proposed method to actual industrial data obtained during an industrial polymer process at Mitsui Chemicals to verify the performance of DMs. There can be a nonlinear relationship between a polymer quality variable (**y**) and other process variables (**X**).[9,17] Accordingly, it is difficult to estimate the predictive accuracy of the soft sensor model between **X** and **y**. We, therefore, attempted to model a relationship between a DM and the predictive accuracy of the model.

First, we collected data recorded for many grades and constructed support vector regression[18] models between **X** and **y** by considering the dynamics of process variables in this study. The **y**-variables are density and the melt flow rate (MFR), and the **X**-variables are 37 process variables such as the temperature in the reactor, the pressure, and concentrations of the monomer, comonomer, and hydrogen. Then, we calculated relationships between DMs and calculation or prediction errors. The ED and the MD to the average of training data, and the output of an OCSVM (the proposed method) were applied as DMs. The combinations of the parameters ν and γ are the same as those of the analysis of the distillation column data in OCSVM modeling and accordingly 806 OCSVM models were constructed.

Figure 5a shows the relationships between AUCR of training data and that of test data for each DM when **y** is density. The blue points represent the results of OCSVM; the red circle represents that of ED; and the green triangle represents that of MD. The distribution of the results of the proposed method locates at the bottom left compared with those of ED and MD, reflecting that the proposed DM had the better performance in most OCSVM parameters than the traditional DMs did. The OCSVM models could represent the

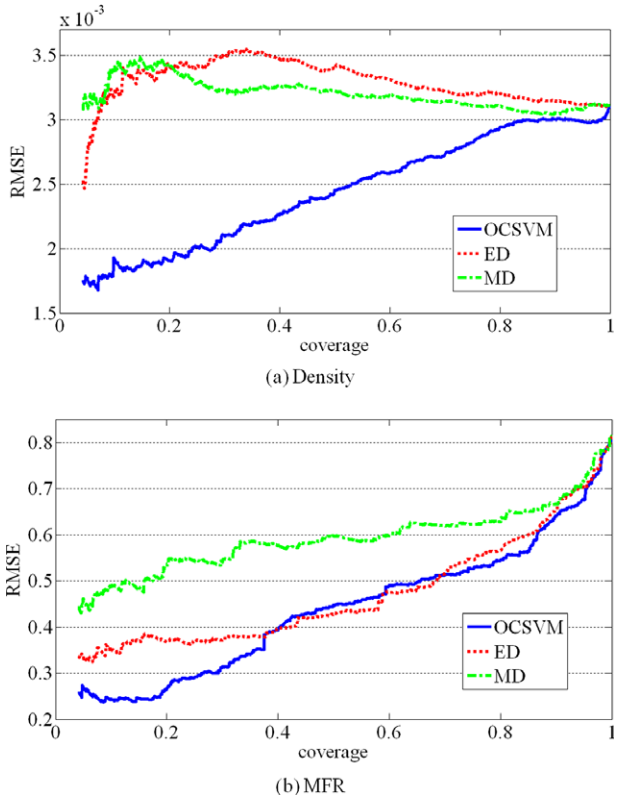

Figure 6. Relationships between the coverage and the RMSE for the industrial polymer process.

The blue continuous lines represent the results of OCSVM; the red dashed lines represent the results of ED; and the green chain lines represent the results of MD. (a) Density (b) MFR. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

nonlinearity among process variables. In addition, the OCSVM model having the smallest value of AUCR of training data in the all parameters had also high performance for test data. We can say that an optimal OCSVM parameter can be decided using training data.

Then, when **y** is MFR, the relationships between AUCR of training data and that of test data for each DM are shown in Figure 5b. In this case, the AUCRs of both training data and test data were relatively small when ED was used as a DM. This will come from a physical relationship between MFR and **X**[19–22] and even ED of **X** could represent information on MFR. Meanwhile, the some results of OCSVM were more left and lower than that of ED, indicating the high performance of the proposed method.

The relationships between the coverage and the RMSE of test data are shown in Figure 6. In each figure, the OCSVM parameters γ and ν were the ones where the AUCR value of training data was smallest. The blue continuous lines represent the results of OCSVM; the red dashed lines represent the results of ED; and the green chain lines represent the results of MD. As shown in the results of OCSVM, the smaller the values of the coverage are, the smaller the RMSE values are. By comparing the values of the RMSE for each coverage, at only small coverage for MFR, the values of the RMSE predicted with the proposed method were smaller than those of the other methods. This means that the model could predict more number of data with higher predictive accuracy by using the proposed index.

Conclusions

In this article, we proposed a DM calculated by using OCSVM to estimate prediction errors of soft sensor models accurately. Then, AUCR was introduced as a comparative indicator, and hence, the performance of DMs could be compared quantitatively. Through the analysis of two sets of real industrial data, we confirmed that the predictive accuracy of each data could be estimated with high accuracy by using the proposed method even when there is a nonlinear relationship between **X** and **y**. By applying this method to process control, industrial plants will be operated stably.

Acknowledgments

H. Kaneko is grateful for financial support of the Japan Society for the Promotion of Science (JSPS) through a Grant-in-Aid for Young Scientists (B) (No. 24760629). The authors acknowledge the support of Mitsui Chemicals, Inc. and Mizushima works, Mitsubishi Chemical Corporation, and the financial support of Mizuho Foundation for the Promotion of Sciences and the Toyota Physical and Chemical Research Institute.

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. Qin SJ. Recursive PLS algorithms for adaptive data modeling. *Comput Chem Eng*. 1998;22:503–514.
4. 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.
5. Cheng C, Chiu MS. A new data-based methodology for nonlinear process modeling. *Chem Eng Sci*. 2004;59:2801–2810.
6. Fujiwara K, Kano M, Hasebe S, Takinami A. Soft-sensor development using correlation-based just-in-time modeling. *AIChE J*. 2009;55:1754–1765.
7. Kaneko H, Funatsu K. Maintenance-free soft sensor models with time difference of process variables. *Chemom Intell Lab Syst*. 2011;107:312–317.
8. Kaneko H, Funatsu K. A soft sensor method based on values predicted from multiple intervals of time difference for improvement and estimation of prediction accuracy. *Chemom Intell Lab Syst*. 2011;109:197–206.
9. Kaneko H, Funatsu K. Development of soft sensor models based on time difference of process variables with accounting for nonlinear relationship. *Ind Eng Chem Res*. 2011;50:10643–10651.
10. Kaneko H, Arakawa M, Funatsu K. Applicability domains and accuracy of prediction of soft sensor models. *AIChE J*. 2011;57:1506–1513.
11. 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.
12. 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.
13. Baskin II, Kireeva N, Varnek A. The one-class classification approach to data description and to models applicability domain. *Mol Inf*. 2010;29:581–587.
14. Liu Y, Huang D, Li Y. Development of interval soft sensors using enhanced just-in-time learning and inductive confidence predictor. *Ind Eng Chem Res*. 2012;51:3356–3367.
15. Schoelkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge: The MIT Press, 2001.
16. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst*. 2001;58:109–130.
17. Kaneko H, Arakawa M, Funatsu K. Novel soft sensor method for detecting completion of transition in industrial polymer processes. *Comput Chem Eng*. 2011;35:1135–1142.
18. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer, 2006.
19. McAuley KB, MacGregor JF. On-line inference of polymer properties in an industrial polyethylene reactor. *AIChE J*. 1991;37:825–835.
20. Ohshima M, Tanigaki M. Quality control of polymer production processes. *J Process Control*. 2000;10:135–148.
21. Lee EH, Kim TY, Yeo, YK. Prediction of the melt index in a high-density polyethylene process. *J Chem Eng Japan*. 2007;40:840–846.
22. Oh SJ, Lee J, Park S. Prediction of pellet properties for an industrial bimodal high-density polyethylene process with Ziegler−Natta catalysts. *Ind Eng Chem Res*. 2005;44:8–20.

Manuscript received May 8, 2012, and revision received Dec. 11, 2012