Preparation of Comprehensive Data from Huge Data Sets for Predictive Soft Sensors

Hiromasa Kaneko and Kimito Funatsu

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

Corresponding author: Kimito Funatsu

Tel: +81-3-5841-7751

Fax: +81-3-5841-7771

E-mail: funatsu@chemsys.t.u-tokyo.ac.jp

Abstract

Soft sensors predict values of process variables that are difficult to measure in real time. Predictive ability of adaptive soft sensors depends on databases. However, there is no way to construct initial databases from huge data sets that are measured in plants and that are data rich, but information poor. Therefore, we propose a method to select comprehensive data from huge data sets to build soft sensors with high predictive ability. A genetic algorithm and the Kennard-Stone algorithm are modified for data selection considering predictive ability of regression models and data distribution. Through the analyses of numerical simulation data and real industrial data, we confirm that initial databases could be appropriately constructed from huge data sets and predictive accuracy of soft sensors subsequently increased.

Keywords: Process control; Soft sensor; Database; Data selection; Genetic algorithm; Kennard-Stone algorithm

1 Introduction

Soft sensors are widely used to predict values of process variables that are difficult to measure online [1,2]. An inferential model is constructed between process variables that are easy to measure online, which are called X-variables, and process variables that are difficult to measure online, which are called y-variables. Values of y can then be predicted using that model. Through the use of soft sensors, values of y can be predicted with high degree of accuracy.

One of the crucial difficulties of soft sensors is that predictive accuracy drops because of changes in states of chemical plants. This is called soft sensor model degradation [3]. To reduce degradation, soft sensor models are reconstructed using the newest data. For instance, moving window (MW) models [4–9] are constructed with the data that were measured most recently, and just-in-time (JIT) models [10–14] are constructed with data that are more similar to a query than other data. Models such as MW, JIT, and time difference [15–17] models that have adaptive mechanisms are called adaptive models [18].

To construct adaptive models with high predictive accuracy for wide data ranges, database monitoring is a crucial problem [19]. To reduce the size of databases, JIT models select new measurement data to be stored based on prediction errors of y-variables [20,21]. Jin *et al.* proposed a method in which a new sample should replace the most similar sample in a database [22]. However, relationship between overall information in a database and that in a new measurement sample was not considered in data selection. Kaneko *et al.* proposed a database monitoring index (DMI) for database management (DBM) that monitors the amount of information in a new measurement datum [23] and achieves a maintenance-free DBM and highly predictive soft sensors [24]. Only data with high DMI values are stored in databases, and thus the total amount of information increases while the amount of data stored is maintained. Lv *et al.* defined reversible and irreversible process variations based on changes in X and y as well as rapidity of these changes, and developed different strategies for database updates based on them [25]. DBM can also be applied to process monitoring [26–32], where models are updated or reconstructed with a data set that includes new measurement data.

However, there is no way to construct initial databases to start prediction from huge data sets that have been measured in plants. These data sets are data rich, but information poor. When there is a huge number of process variables that are variable rich but information poor, latent variable modeling approaches such as principal component analysis and partial least squares work well. However, our target is not huge process variable sets but huge data sets.

Therefore, we propose a method to select comprehensive data from huge data sets to create adaptive soft sensor models with high predictive ability. A genetic algorithm (GA) [33] and the Kennard-Stone algorithm (KS) using Euclidean distance [34,35] are applied to select data sets from which predictive soft sensor models can be constructed and whose distribution is widely spread in data space. The fitness function in GA is based on predictive ability of regression models. Indexes of distance in KS are the Euclidean distance, the Mahalanobis distance [36], a normalized X and y distance [37] and DMI [23,24]. KS combined with DMI is a new method proposed in this paper. We also propose a new KS, which includes initial data comprising the highest and lowest y-values, to consider extreme data.

Through the analyses of numerical simulation data and real industrial data, we confirm that initial databases can be appropriately constructed from huge data sets and the predictive ability of adaptive soft sensors increases by using the proposed methods.

2 Method

In this study, GA is employed since sample selection is a combinatorial optimization problem, and KS is applied because KS can select a subset of samples while considering data distribution. After a brief explanation of GA and KS, we describe the proposed data selection methods, based on the proposed KS with initial data (KSI). In this study, we denote the number of the first data as N and the number of the final data selected by the proposed methods as m .

2.1 Genetic Algorithm (GA)

GA is a well-developed optimization method that is used to model the principles of natural evolution in biology [33]. Species with high levels of fitness under certain environmental conditions will prevail in the next generation, and the best species may be reproduced by crossover together with the random mutation of chromosomes in those species that survive. The solution space around superior individuals is searched for preferentially, which leads to the discovery of solutions that are close to the optimum. GA is applied to various fields such as variable selection [38].

The flow of GA is shown in Figure 1. In this study, the values in the chromosomes are real numbers ranging from 0 to 1 (*i.e.*, real-coded GA). Each value in chromosomes corresponds to the possibility of selecting a sample. The length of a chromosome is the number of data in an original database. m data that have the highest values of a chromosome are selected. Fitness values of chromosomes are calculated using m data, as explained in Section 2.4. Chromosomes with sufficiently high fitness

values can be obtained using the flow in Figure 1. The number of generations must first be set as the termination condition.

2.2 Kennard-Stone algorithm (KS)

KS aims to obtain m data that are widely distributed in data space from N data. It follows a stepwise procedure, in which a new sample is selected so as to be far from the already-selected data. In classical KS, the Euclidean distance (ED) between two data $\mathbf{z}^{(i)} \in \mathbf{R}^{1 \times n}$ (n is the number of variables) and $\mathbf{z}^{(j)} \in \mathbf{R}^{1 \times n}$, $ED^{(i,j)}$, is used as an index of distance, which is given as follows:

$$ED^{(i,j)} = \sqrt{(\mathbf{z}^{(i)} - \mathbf{z}^{(j)})(\mathbf{z}^{(i)} - \mathbf{z}^{(j)})^T} \quad (1)$$

In this study, $\mathbf{z}^{(i)}$ consists of $\mathbf{x}^{(i)} \in \mathbf{R}^{1 \times (n-1)}$ and $y^{(i)}$ and is represented as follows:

$$\begin{aligned} \mathbf{z}^{(i)} &= \begin{bmatrix} \mathbf{x}^{(i)} & y^{(i)} \end{bmatrix} \\ &= \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & \dots & x_n^{(i)} & y^{(i)} \end{bmatrix}, \end{aligned} \quad (2)$$

where $\mathbf{x}^{(i)}$ is the i^{th} data point of $\mathbf{X}$, $x_p^{(i)}$ is the value of the p^{th} $\mathbf{X}$ -variable of the i^{th} data point, $y^{(i)}$ is the y -value of the i^{th} data point, and n is the number of $\mathbf{X}$ -variables.

In KS, the sample whose distance is the largest from the average is selected first. Then, for each sample that has not been selected yet, the distance from the already-selected data is calculated and the minimum distance is saved. The sample whose minimum distance is the largest is newly selected. The minimum distance calculation and data selection are repeated until the number of data required have been selected.

2.3 Proposed Kennard-Stone algorithm with initial data (KSI)

To consider extreme data, we propose a method KSI that uses initial data in KS. Initial data are the data set having the p highest and p lowest y -values and the number of initial data is $2p$. The data selection procedure from the $(2p+1)^{\text{th}}$ sample is the same as that of KS.

2.4 Database selection method

In this study, we assume that there are a huge number of data, including y-values, and the quality of y-values and that of X-values are equal for each sample.

It is necessary in data selection for regression models predicting y-values of wide data ranges with high accuracy to be constructed with a small amount of limited data. With that point in mind, we introduce a coefficient of determination for all N data predicted with a k -nearest neighbor (kNN) algorithm using m training data as the fitness value in GA. This is called GA- kNN . By selecting m data based on the proposed coefficient, the y-values of all N data will be predicted well using only selected m data. Although support vector regression (SVR) [39] can be used as a regression method, kNN is employed because regression has to be done many times in GA and SVR modeling has a high computational cost. In this paper, k is set to 5, which is a commonly used value in kNN . Data with the p highest and p lowest y-values are always included in m data.

KSI can be used to select m data from N data. This approach is called KSI-ED. The Mahalanobis distance (MD) is an alternative to the Euclidean distance in KSI when variables are correlated [40]. This method is called KSI-MD. The MD between two data $\mathbf{z}^{(i)}$ and $\mathbf{z}^{(j)}$, $MD^{(i,j)}$, is given as follows:

$$MD^{(i,j)} = \sqrt{(\mathbf{z}^{(i)} - \mathbf{z}^{(j)})^T \mathbf{S}^{-1} (\mathbf{z}^{(i)} - \mathbf{z}^{(j)})}, \quad (3)$$

where $\mathbf{S}$ is the covariance matrix. In this study, both X-variables and a y-variable are considered in distance calculation. Hence, a normalized X and y distance (ND) is developed to equally handle the different distribution of data in X and y [37]. When ND is used in KSI, the algorithm is called KSI-ND. The ND between two data $\mathbf{z}^{(i)}$ and $\mathbf{z}^{(j)}$, $ND^{(i,j)}$, is given as follows:

$$ND^{(i,j)} = \frac{ED_X^{(i,j)}}{\max_{i,j \in N} ED_X^{(i,j)}} + \frac{ED_Y^{(i,j)}}{\max_{i,j \in N} ED_Y^{(i,j)}}, \quad (4)$$

where $ED_X^{(i,j)}$ and $ED_Y^{(i,j)}$ are respectively given as follows:

$$ED_X^{(i,j)} = \sqrt{(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}) (\mathbf{x}^{(i)} - \mathbf{x}^{(j)})^T}, \quad (5)$$

$$ED_y^{(i,j)} = |y^{(i)} - y^{(j)}| \quad (6)$$

To consider both the similarity in X-variables and that in a y-variable appropriately, DMI [23,24] can be used in KSI, called KSI-DMI. DMI^(i,j) between two data ($\mathbf{x}^{(i)}, y^{(i)}$) and ($\mathbf{x}^{(j)}, y^{(j)}$) is given as follows:

$$DMI^{(i,j)} = \frac{|y^{(i)} - y^{(j)}|^a}{\text{sim}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}, \quad (7)$$

where $\text{sim}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is the similarity between $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$, and a is a constant. In this study, we use the inverse Euclidean distance as a similarity index, and thus DMI^(i,j) is given as follows:

$$DMI^{(i,j)} = |y^{(i)} - y^{(j)}|^a \sqrt{\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2} \quad (8)$$

The set of candidate values for a is $\{-2, -1.9, -1.8, \dots, 1.8, 1.9, 2.0\}$. a is selected to maximize the coefficient of determination for all N data that are predicted with a regression model constructed with m data. The optimal value could be found within this set in our case studies.

By using GA- k NN, KSI-ED, KSI-MD, KSI-ND, and KSI-DMI, a small but comprehensive data set can be selected from an original database from which highly accurate soft sensor models can be constructed.

3 Results and Discussion

After the effectiveness of the proposed methods is demonstrated using numerical simulation data, the proposed methods are applied to real industrial data sets. The final soft sensor models are constructed using SVR because m data are already selected. The hyperparameters in SVR modeling are optimized quickly using both 5-fold cross-validation and theoretical analysis on the training data [41]. The number of chromosomes and generations in the GA are both 1000. The values of k in k NN and the candidate values for a did not affect the results significantly.

3.1 Nonlinear numerical simulation data

We verified the performance of the proposed methods when relationship between X and y is nonlinear. We considered two X -variables related by

$$\begin{aligned} \exp(y) = & \left(1 + (x_1 + x_2 + 1)^2 (19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2)\right) \\ & \times \left(30 + (2x_1 - 3x_2)^2 (18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)\right). \end{aligned} \quad (9)$$

x_1 and x_2 are X -variables. Equation (9) has been previously used as a test problem [42]. Data of X -variables were generated as 81 grid points whose intervals of both x_1 and x_2 were $-2, -1.5, -1, \dots, 1.5, \text{ and } 2$. The same 81 grid points were generated 150 times, giving a total of 12,150 ($= 81 \times 150$) data of X -variables. Equation (9) outputs y -values from the 12,150 X -values. Random numbers from the normal distribution with a mean of 0 and standard deviation of 0.01 were added to both the y -variable and X -variables. The x_1 vs. x_2 plot is shown in Figure 2.

In this case study, we attempted to select 100 data from the 12,150 data ($N = 12,150$ and $m = 100$). When GA is used, the final results are different depending on the seed state of a random number generator. Therefore, ten GA calculations were performed for GA- kNN and all results were checked. As a comparison, 100 data were randomly selected from the 12,150 because random selection can reflect the amount of information within the data, to some extent.

Data selection results are shown in Figure 3. For KSI-DMI, α was optimized as 0.7. Figure 3(a) is a randomly selected data set shown in an x_1 vs. x_2 plot. The selected data were biased, and thus the areas where no data were selected are large. In Figure 3(b), one example of the ten GA- kNN results is shown. The data distribution is almost as sparse as that of random selection (see Figure 3(a)). In contrast, selected data are widely distributed in x_1 vs. x_2 plots in Figures 3(c)–(f) when KSI is employed in data selection. Irrespective of the distance indexes, widely distributed and small data can be selected by considering similarity or distance between data using the proposed methods.

For each method, an SVR model was constructed with the selected 100 data. It was then used to predict the y -values of all 12,150 data. Table 1 shows the calculation results including the root-mean-square error (RMSE) of the 12,150 data and computation time. For GA- kNN , the RMSE and computation time are the averages of the ten calculations. Compared with random selection, the RMSE values of the proposed methods are low. Using GA- kNN , although the data distribution was not diverse (see Figure 3(b)), the prediction ability improved. A data set with which an SVR

model was constructed with highly predictive ability could be selected using GA-*k*NN. For KSI-DMI, the RMSE value is lower than those of KSI-ED, KSI-MD and KSI-ND, which is because the predictive ability for all 12,150 data is considered with respect to the optimization of a , and thus the nonlinearity between x_1 , x_2 , and y is appropriately reflected in data selection. Data sets that are informative and widely distributed in data space can be selected using these methods, and the SVR models constructed using such data sets can accurately predict y -values over a wide data range.

Figure 4 shows a plot of the measured and predicted y for all 12,150 data for each method. The data for the results of the proposed methods are closer to the diagonal than those of random selection. The results show that the proposed methods can select small data sets with which highly predictive SVR models can be constructed. In particular, the SVR model for KSI-DMI has a very high predictive ability. A moderate reduction of the predictive ability of the GA-*k*NN-based SVR with respect to the KSI-DMI-based SVR could be the result of a local solution in GA and the difference between the regression methods in the sample selection and final model construction, *i.e.* *k*NN and SVR.

The computation time of GA-*k*NN is far greater than that of KSI-ED, KSI-MD, KSI-ND, and KSI-DMI. This is because GA needs high computational cost and the *k*NN modeling takes a large amount of time to calculate fitness values. Hence, KSI-ED, KSI-MD, KSI-ND, and KSI-DMI are preferable to GA-*k*NN in terms of calculation cost. KSI-DMI requires more time than KSI-ED, KSI-MD, and KSI-ND because of the optimization procedure of a .

3.2 Real industrial data in a sulfur recovery unit (SRU)

We applied the proposed methods to data obtained from the operation of an SRU. Two gases, monoethanolamine (MEA) gas rich in H_2S and gas from a sour water stripping (SWS) plant rich in H_2S and ammonia, are input into an SRU and environmental pollutants are removed from these gases. Gases are burnt in reactors and H_2S is decomposed into pure sulfur and water vapor through a partial oxidation reaction with air. The chemical reactions are given as follows:

Output gas from the SRU includes residual H_2S and SO_2 . Although H_2S and SO_2 concentration can be measured every minute, soft sensors are required because H_2S gas

and SO₂ gas damage hardware sensors, which need frequent maintenance as a result. In this paper, the y-variable is the H₂S concentration at the outlet of the SRU, and the X-variables are the MEA gas flow, first air flow, second air flow, gas flow in the SWS zone, and air flow in the SWS zone. The measurement intervals of the y-variable and X-variables are both one per minute. Details of the data set can be found in [43]. We used the 30 explanatory variables included in each X-variable, which was delayed for a duration ranging from 0 to 10 min in steps of 2 min. Thus, we prepared 10,000 data (10,000 min) and attempted to select 500 data ($N = 10,000$ and $m = 500$). Systematic sampling was used as a comparison for the proposed methods. For each data selection method, an SVR model was constructed with the selected 500 data to predict the y-values of all 10,000 data.

Table 2 shows the RMSE values of all 10,000 data and the computation time for each method. For KSI-DMI, α was optimized as 0.3. Using the proposed GA-kNN, KSI-ED, KSI-ND, and KSI-DMI methods, the RMSE was decreased with respect to the results of systematic sampling, and thus the prediction accuracy of the SVR model was improved. In the proposed methods, KSI-DMI had the highest predictive ability, although the difference between the RMSE values was slight. In contrast, KSI-MD could not select an appropriate data set, and its RMSE value was high, which could be because the covariance matrix $\mathbf{S}$ in the Mahalanobis distance was calculated using all the data, and thus any local correlation of X and y was ignored.

GA-kNN required a large amount of computation time. Data selection could be performed in less than 1 min for KSI-ED and KSI-ND. KSI-DMI needed more time than KSI-ED and KSI-ND because of the optimization of α .

Relationships between actual and estimated y for all 10,000 data are shown in Figure 5. For GA-kNN, KSI-ED, KSI-ND and KSI-DMI, data points are closer to the diagonal than those for the systematic sampling, which indicates that the predictive ability of an SVR increases using the proposed methods. KSI-ED achieved the least estimation errors (the highest prediction accuracy) for high H₂S concentration values (see Figure 5(C)). Prediction results for KSI-DMI depend on α . When α is set to 1, high prediction accuracy for high H₂S concentration values are achieved, as shown in Figure 6, although the RMSE value decreased a little. The KSI-DMI method is hence able to design soft sensors flexibly by changing α and selecting a different data set.

These results confirm that appropriate data selection could be achieved and as a result, high predictive nonlinear regression models could be constructed using GA-kNN, KSI-ED, KSI-ND and KSI-DMI.

3.3 Real industrial data in an exhaust gas denitration process

To verify the performance of a real soft sensor constructed with a selected data set, we applied the proposed methods to data from an exhaust gas denitration process at Mitsui Chemicals, Inc., which were also used in [9]. NH_3 is injected into a denitration reactor, where exhaust gas passes through a catalytic layer, and NO_x is decomposed into harmless N_2 and water vapor. The reactions are as follows:

Soft sensors can be used to control the remaining NH_3 concentration at the outlet of the denitration reactor (denitration outlet NH_3), NO_x concentration at the outlet of the denitration reactor (denitration outlet NO_x), and NO_x concentration at the outlet of the outlet gas duct (outlet gas duct NO_x). In this study, the y-variable is the denitration outlet NH_3 , and the X-variables consist of 23 process variables such as temperature, pressure, and flow rates at the gas mixer and denitration reactor.

We prepared 10,000 data (10,000 min) from November 2012 and attempted to select 1000 data ($N = 10,000$ and $m = 1000$). Systematic sampling was used for comparison with the proposed methods.

The calculation results and plots of measured and predicted y are shown in Table 3 and Figure 7, respectively. Compared with systematic sampling, the RMSE values decreased when GA-kNN, KSI-ED, KSI-ND and KSI-DMI were used, indicating the increase of predictive accuracy in the SVR models. Additionally, although predicted y-values are biased for high and low measured y-values in the systematic sampling, wide ranges of y-values were predicted accurately using GA-kNN, KSI-ED, KSI-ND and KSI-DMI (see Figure 7). However, the predictive ability decreased when using KSI-MD. This is because $\mathbf{S}$ was calculated using all 10,000 data, and it hence could not represent the change in the relationship between X and y caused by the change of process states.

In terms of calculation cost, GA-kNN needed a large amount of time, and thus, KSI-ED, KSI-ND, and KSI-DMI are preferable. Although a more comprehensive data set was obtained using KSI-DMI than by using KSI-ED or KSI-ND, it should be noted that KSI-DMI requires more calculation time than KSI-ED or KSI-ND.

In our case studies, we did not have to normalize X and y when distance was calculated because there was little difference between the results of KSI-ED and KSI-ND. This was because y could be represented by X to some extent, and information

on y in X was reflected in the distance in X -space and y -space. However, when there are many unnecessary X -variables, the use of KSI-ND is preferable. By using KSI-DMI, a comprehensive data set could be flexibly selected by optimizing a , considering the nonlinearity between X and y . However, it takes more time to use KSI-DMI than it does to use KSI-ED or KSI-ND.

Comprehensive and small data datasets can be selected in a short time by using the KSI-ED, KSI-ND, and KSI-DMI methods.

4 Conclusions

We proposed a GA-based method and several KS-based methods to select comprehensive and small data from huge data sets. The fitness function in GA considers predictive ability of regression models (GA- k NN). Additionally, the possible indexes of distance in KSI (in which the initial data with the p highest and p lowest y -values are always included) are the Euclidean distance (KSI-ED), the Mahalanobis distance (KSI-MD), a normalized X and y distance (KSI-ND) and DMI (KSI-DMI). Through case studies using a numerical simulation data set and two real industrial data sets, we confirmed that KSI-ED, KSI-ND and KSI-DMI could consider a diversity of data sets and select appropriate data from which highly accurate nonlinear regression models could be constructed within a reasonable time. KSI-DMI selected a more comprehensive data set than KSI-ED and KSI-ND. However, KSI-DMI requires more time than KSI-ED or KSI-ND to calculate. In general, KSI-DMI is preferable, but either KSI-ED or KSI-ND would be appropriate when the number of data is too large to practically compute KSI-DMI.

After the initial database has been constructed by the proposed methods, plant engineers will be able to manage the database and adaptive soft sensors simultaneously online with a database monitoring method [24].

Acknowledgements

The authors acknowledge the support of the Core Research for Evolutionary Science and Technology (CREST) project “Development of a knowledge-generating platform driven by big data in drug discovery through production processes” of the Japan Science and Technology Agency (JST). The authors acknowledge the support of Mitsui Chemicals, Inc.

References

- [1] M. Kano, Y. Nakagawa, Data-based process monitoring, process control, and

- quality improvement: Recent developments and applications in steel industry, *Comput. Chem. Eng.* 32 (2008) 12–24.
- [2] P. Kadlec, B. Gabrys, S. Strandt, Data-driven soft sensors in the process industry, *Comput. Chem. Eng.* 33 (2009) 795–814.
 - [3] H. Kaneko, K. Funatsu, Classification of the degradation of soft sensor models and discussion on adaptive models, *AIChE J.* 59 (2013) 2339–2347.
 - [4] S.J. Qin, Recursive PLS algorithms for adaptive data modeling, *Comput. Chem. Eng.* 22 (1998) 503–514.
 - [5] H. Kaneko, M. Arakawa, K. Funatsu, Development of a new soft sensor method using independent component analysis and partial least squares, *AIChE J.* 55 (2009) 87–98.
 - [6] P. Kadlec, B. Gabrys, Local learning-based adaptive soft sensor for catalyst activation prediction. *AIChE J.* 57 (2010) 1288–1301.
 - [7] H. Kaneko, K. Funatsu, Application of online support vector regression for soft sensors, *AIChE J.* 60 (2014) 600–612.
 - [8] Z.Q. Ge, Z.H. Song, M. Kano, External analysis-based regression model for robust soft sensing of multimode chemical processes, *AIChE J.* 60 (2014) 136–147.
 - [9] H. Kaneko, K. Funatsu, Adaptive soft sensor based on online support vector regression and bayesian ensemble learning for various states in chemical plants, *Chemom. Intell. Lab. Syst.* 137 (2014) 57–66.
 - [10] C. Cheng, M.S. Chiu, A new data-based methodology for nonlinear process Modeling, *Chem. Eng. Sci.* 59 (2004) 2801–2810.
 - [11] K. Fujiwara, M. Kano, S. Hasebe, A. Takinami, Soft-sensor development using correlation-based just-in-time modeling, *AIChE J.* 55 (2009) 1754–1765.
 - [12] S. Schaal, C.G. Atkeson, S. Vijayakumar, Scalable techniques from onparametric statistics for real time robot learning, *Appl. Intell.* 17 (2002) 49–60.
 - [13] S. Kim, M. Kano, H. Nakagawa, S. Hasebe, Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection, *Int. J. Pharm.* 421 (2011) 269–274.
 - [14] Q.C. Jiang, X.F. Yan, Just-in-time reorganized PCA integrated with SVDD for chemical process monitoring, *AIChE J.* 60 (2014) 949–965.
 - [15] H. Kaneko, K. Funatsu, Maintenance-free soft sensor models with time difference of process variables, *Chemom. Intell. Lab. Syst.* 107 (2011) 312–317.
 - [16] H. Kaneko, K. Funatsu, 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.* 109 (2011) 197–206.

- [17] H. Kaneko, K. Funatsu, Development of soft sensor models based on time difference of process variables with accounting for nonlinear relationship, *Ind. Eng. Chem. Res.* 50 (2011) 10643–10651.
- [18] P. Kadlec, R. Grbic, B. Gabrys, Review of adaptation mechanisms for data-driven soft sensors, *Comput. Chem. Eng.* 35 (2011) 1–24.
- [19] H. Kaneko, K. Funatsu, Development of a new index to monitor database for soft sensors. *J. Comput. Aided Chem.* 14 (2013) 11–22. (in Japanese)
- [20] A.G. Kalmukale, M.S. Chiu, Q.G. Wang, Partitioned model-based IMC design using JITL modeling technique, *J. Process Contr.* 17 (2007) 757–769.
- [21] Y. Kansha, M.S. Chiu, Adaptive generalized predictive control based on JITL technique, *J. Process Contr.* 17 (2009) 1067–1072.
- [22] H.P. Jin, X.G. Chen, J.W. Yang, L. Wu, Adaptive soft sensor modeling framework based on just-in-time learning and kernel partial least squares regression for nonlinear multiphase batch processes, *AIChE J.* 71 (2014) 77–93.
- [23] H. Kaneko, K. Funatsu, Database monitoring index for adaptive soft sensors and the application to industrial process, *AIChE J.* 60 (2014) 160–169.
- [24] H. Kaneko, K. Funatsu, Adaptive database management based on the database monitoring index for long-term use of adaptive soft sensors, *Chemom. Intell. Lab. Syst.* 146 (2015) 179–185.
- [25] Y. Lv, T. Yang, J. Liu, An adaptive least squares support vector machine model with a novel update for NO_x emission prediction, *Chemom. Intell. Lab. Syst.* 145 (2015) 103–113.
- [26] J.F. Macgregor, C. Jaeckle, C. Kiparissides, M. Koutoudi, Process monitoring and diagnosis by multiblock PLS methods, *AIChE J.* 5 (1994) 826–838.
- [27] B.R. Bakshi, Multiscale PCA with application to multivariate statistical process monitoring, *AIChE J.* 44 (1998) 1596–1610.
- [28] R. Dunia, S.J. Qin, Subspace approach to multidimensional fault identification and reconstruction, *AIChE J.* 44 (1998) 1813–1831.
- [29] M. Kano, S. Tanaka, S. Hasebe, I. Hashimoto, H. Ohno, Monitoring independent components for fault detection, *AIChE J.* 49 (2003) 969–976.
- [30] R. Dunia, T.F. Edgar, M. Nixon, Process monitoring using principal components in parallel coordinates, *AIChE J.* 59 (2013) 445–456.
- [31] J. Mori, J. Yu, A quality relevant non-Gaussian latent subspace projection method for chemical process monitoring and fault detection, *AIChE J.* 60 (2014) 485–499.
- [32] J.L. Zhu, Z.Q. Ge, Z.H. Song, Robust modeling of mixture probabilistic principal component analysis and process monitoring application. *AIChE J.* 60 (2014) 2143–

2157.

- [33] D. Whitley, A genetic algorithm tutorial, *Stat. Comput.* 4 (1994) 65–85.
- [34] R.W. Kennard, L.A. Stone, Computer aided design of experiments, *Technometrics*, 11 (1969) 137–148.
- [35] M. Daszykowski, B. Walczak, D.L. Massart, Representative subset selection, *Anal. Chim. Acta*, 468 (2002) 91–103.
- [36] R.D. Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, *Chemom. Intell. Lab. Syst.* 50 (2000) 1–18.
- [37] R.K.H. Galvão, M.C.U. Araujo, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, A method for calibration and validation subset partitioning, *Talanta* 67 (2005) 736–740.
- [38] V. Galyanin, A. Melenteva, A. Bogomolov, Selecting optimal wavelength intervals for an optical sensor: A case study of milk fat and total protein analysis in the region 400–1100 nm, *Sensor Actuat. B-Chem.* 218, 97–104.
- [39] C.M. Bishop, *Pattern recognition and machine learning*, Springer, New York, 2006.
- [40] A. Saptoro, O. Moses, M.O. Tadé, H. Vuthaluru, A modified KS algorithm for optimal data division of data for developing artificial neural network models, *Chemical Product and Process Modeling*, 7 (2012) 1–14.
- [41] H. Kaneko, K. Funatsu, Fast optimization of hyperparameters for support vector regression models with highly predictive ability, *Chemom. Intell. Lab. Syst.* 142 (2015) 64–69.
- [42] G. Li, V. Aute, S. Azarm, An accumulative error based adaptive design of experiments for offline metamodeling, *Struct. Multidiscip. O.* 40 (2010) 137–155.
- [43] L. Fortuna, S. Graziani, A. Rizzo, M.G. Xibilia, *Soft sensors for monitoring and control of industrial processes*. London: Springer-Verlag; 2007.

List of Figure Captions

Figure 1. Flow of GA.

Figure 2. Numerical simulation data: x_1 vs. x_2 plot.

Figure 3. Data selected when the numerical simulation data were used.

Figure 4. Measured and predicted y of all 12,150 data when the numerical simulation data were used.

Figure 5. Measured and predicted y for all 10,000 SRU process data.

Figure 6. Measured and predicted y for KSI-DMI when a equals 1 for the SRU data (RMSE = 0.026).

Figure 7. Measured and predicted y of all 10,000 data for the exhaust gas denitration process.

Table 1. Calculation results when the numerical simulation data were used. The RMSE was calculated for all 12,150 data.

		Random	GA- k NN	KSI-ED	KSI-MD	KSI-ND	KSI-DMI
RMSE	Average	0.913	0.528	0.585	0.783	0.596	0.105
	Std.	0.145	0.196	-	-	-	-
Computation time [min]	Average	-	1291	0.01	0.03	0.06	15.0
	Std.	-	5.4	-	-	-	-

Table 2. Calculation results for the SRU process. The RMSE was calculated for all 10,000 data.

		Systematic sampling	GA- <i>k</i> NN	KSI-ED	KSI-MD	KSI-ND	KSI-DMI
RMSE	Average	0.028	0.026	0.025	0.116	0.024	0.022
	Std.	-	0.0003	-	-	-	-
Computation time [min]	Average	-	2494	0.34	8.3	0.72	43.7
	Std.	-	13	-	-	-	-

Table 3. Calculation results for the exhaust gas denitration process. The RMSE was calculated for all 10,000 data.

		Systematic sampling	GA- k NN	KSI-ED	KSI-MD	KSI-ND	KSI-DMI
RMSE	Average	0.226	0.179	0.174	0.450	0.176	0.172
	Std.	-	0.002	-	-	-	-
Computation time [min]	Average	-	1058	1.5	15.6	2.8	190
	Std.	-	3.3	-	-	-	-

Figure 1. Flow of GA.

Figure 2. Numerical simulation data: x_1 vs. x_2 plot.

Figure 3. Data selected when the numerical simulation data were used.

Figure 4. Measured and predicted y of all 12,150 data when the numerical simulation data were used.

Figure 5. Measured and predicted y for all 10,000 SRU process data.

Figure 6. Measured and predicted y for KSI-DMI when a equals 1 for the SRU data (RMSE = 0.026).

(a) Systematic sampling

(b) GA-kNN

(c) KSI-ED

(d) KSI-MD

(e) KSI-ND

(f) KSI-DMI

Figure 7. Measured and predicted y of all 10,000 data for the exhaust gas denitration process.