Proceedings of the 9th International Symposium on
Dynamics and Control of Process Systems (DYCOPS 2010),
Leuven, Belgium, July 5-7, 2010
Mayuresh Kothare, Moses Tade, Alain Vande Wouwer, Ilse Smets (Eds.)

Correlation-based Spectral Clustering for Flexible Soft-Sensor Design

Koichi Fujiwara * **Manabu Kano** * **Shinji Hasebe** *

* *Department of Chemical Engineering, Kyoto University*
Nishikyo-ku, Kyoto 615-8510, Japan
(e-mail: manabu@cheme.kyoto-u.ac.jp)

Abstract: The current issues concerning soft-sensors are how to cope with changes in process characteristics and how to cope with parallelized, slightly different, multiple processes. To make soft-sensors adaptive and flexible, the development of practical design techniques, instead of impracticable ideas, is crucial; this is the motivation of the present research. In practice, it is difficult to successfully apply a single soft-sensor to parallelized production devices due to their individual difference. Since the individual difference is expressed as difference of the correlation among variables, it is useful to classify samples into multiple clusters according to the correlation in order to adopt a multi-model approach. In the present work, a new correlation-based clustering method, referred to as NC-spectral clustering, is proposed by integrating the nearest correlation (NC) method and spectral clustering. The NC method can detect samples that are similar to the query from the viewpoint of the correlation. In the proposed method, the NC method is used for constructing the weighted graph that expresses the correlation-based similarities between samples and the constructed graph is partitioned by using spectral clustering. In addition, a new soft-sensor design method is proposed on the basis of the proposed NC-spectral clustering. The superiority of the proposed method over conventional methods is demonstrated through a numerical example and a case study of parallelized batch processes.

Keywords: Software sensing, Spectral clustering, Graph theory, Pattern recognition, Classification, Estimation, Principal component analysis, Partial least squares

1. INTRODUCTION

Soft-sensors or virtual sensors have been widely used for estimating product quality or other key variables in various industries. However, there is a gap between academic research and industrial practice. For example, artificial neural network (ANN) has been actively investigated in the literature even after the year 2000 (Kano and Nakagawa (2008)), but the number of its industrial applications is far fewer than that of linear regression (Kano and Ogawa, 2009). Moreover, ANN has lost popularity and has been replaced by linear regression at least in the Japanese chemical and petroleum refining industries. Among linear regression, multiple regression analysis is the most popular and partial least squares (PLS) is the second. Application of PLS to distillation processes has long history (Mejdell and Skogestad (1991) and Kano et al. (2000)), and more recently PLS is integrated with another method to solve practical problems (Kamohara et al. (2004) and Kaneko et al. (2009)). In addition, various extension has been reported in the literature. For example, Amirthalingam and Lee (1999) and Kano et al. (2008) have proposed to apply subspace identification to soft-sensor design.

Even if an accurate soft-sensor is developed successfully, its estimation performance deteriorates as process characteristics change. In chemical processes, process characteristics are changed by catalyst deactivation, scale adhesion and so on. In semiconductor manufacturing processes, periodic cleaning of equipment changes the process characteristics dramatically. Therefore, maintenance of soft-sensors is very important to keep their estimation performance. On the basis of questionnaire survey results, Ogawa and Kano (2009) conclude that soft-sensors should be updated as process characteristics change, and also manual, repeated construction of them should be avoided due to its heavy workload.

To cope with changes in process characteristics and to update statistical models automatically, recursive PLS (Qin (1998)) and Just-In-Time (JIT) modeling (Bontempi et al. (1999) and Cheng and Chiu (2004)) have been proposed. Recursive PLS updates a PLS model recursively, and JIT modeling builds a local model from neighbor samples around the query only when an estimate is requested. Shigemori et al. (2009) reported a successful application of locally weighted regression, which is a kind of JIT modeling, in the steel industry. Although these methods can adapt models to new operating conditions, they cannot always achieve high estimation performance. Recently, Fujiwara et al. (2009a) has proposed a new JIT modeling method, referred to as Correlation-based JIT (CoJIT) modeling. Since difference of process characteristics is expressed as difference of the correlation among variables, CoJIT modeling builds a local model from samples whose correlation can properly describe the query. An industrial application shows the usefulness of the CoJIT modeling.

In addition, the individual difference of production devices should be taken into account. In semiconductor processes,

for example, tens of parallelized production devices are used, and they have different characteristics even if their catalog specifications are the same. In such a case, a soft-sensor developed for one particular device is not always applicable to another device, and it is very laborious to construct soft-sensors for each device. Therefore, a practical soft-sensor design method that can cope with such individual difference should be developed.

When there are some devices whose characteristics are similar to each other, a common soft-sensor may be applicable to them. That is, to construct soft-sensors that can cope with the individual difference of production devices, it is useful to classify operation data of parallelized devices into fewer classes according to their characteristics and to construct models for each class.

The k-means method has been widely used for sample classification. Although it can cluster samples on the basis of the distance, it does not take into account the correlation among variables. Recently, self-organizing map (SOM) has been proposed (Kohonen (2001)). SOM is a machine learning process that imitates the brain learning process, and it can visualize high dimensional data as a map on the basis of similarities between samples. However, SOM does not always give clear boundaries between clusters on the map. In addition, it requires high computational load, and its parameter tuning and data preprocessing are complicated.

In the present work, a new clustering method, referred to as NC-spectral clustering, is proposed. In the proposed method, the nearest correlation (NC) method, propose by Fujiwara et al. (2009b), that can detect samples whose correlation is similar to the query and spectral clustering, proposed by Ding et al. (2001) and Ng et al. (2001), that can partition a weighted graph are integrated. The proposed NC-spectral clustering can classify samples according to their correlation among variables without teacher signals. In addition, a new soft-sensor design method based on the NC-spectral clustering is developed. The usefulness of the proposed method is demonstrated through a case study of parallelized batch processes.

2. SPECTRAL CLUSTERING

Spectral clustering is a clustering method based on the graph theory. It can partition a weighted graph, whose weights express affinities between nodes, into subgraphs through cutting some of their arcs.

Although several spectral clustering algorithms have been proposed, the Max-Min Cut (Mcut) method proposed by Ding et al. (2001) is described in this section.

Given a weighted graph G and its adjacency matrix (affinity matrix) $\boldsymbol{W}$, G is partitioned into two subgraphs A and B. The affinity between A and B is defined as

$$\mathrm{cut}(A,B) \equiv W(A,B) \tag{1}$$

$$W(A,B) = \sum_{u\in A, v\in B} W_{u,v} \tag{2}$$

$$W(A) \equiv W(A,A). \tag{3}$$

where u and v denote nodes of subgraphs A and B, respectively. The affinity between subgraphs $\mathrm{cut}(A,B)$ is

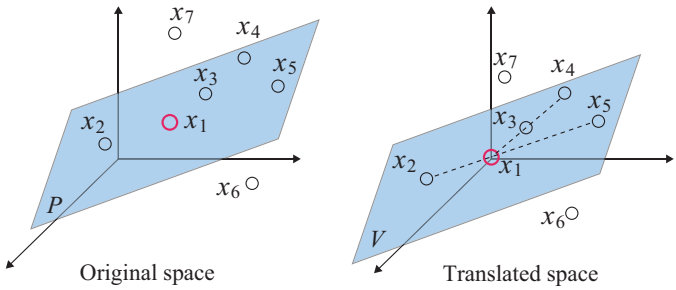

Fig. 1. An illustrative example of the NC method

the sum of the weights of the arcs between subgraphs. The Mcut method searches subgraphs A and B that can minimize $\mathrm{cut}(A,B)$ and maximize $W(A)$ and $W(B)$ simultaneously. The optimization problem of the Mcut method is as follows:

$$\min\ J = \frac{\mathrm{cut}(A,B)}{W(A)} + \frac{\mathrm{cut}(A,B)}{W(B)} \tag{4}$$

This minimization problem results in the eigenvalue problem (Ding et al. (2001)).

In spectral clustering, the definition of an affinity is arbitrary and affects results. Ng et al. (2001) defined the affinity between samples $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ by using the Gaussian kernel:

$$(\boldsymbol{W})_{i,j} = \exp\left(\frac{-d^2(\boldsymbol{s}_i,\boldsymbol{s}_j)}{2\sigma^2}\right) \tag{5}$$

where $d(\cdot,\cdot)$ is a distance function and σ is a tuning parameter.

3. NC-SPECTRAL CLUSTERING

In the present work, a new clustering method based on the correlation among variables is proposed. In the proposed method, the correlation-based affinities between samples are calculated by using the nearest correlation (NC) method to construct a weighted graph, and the constructed weighted graph is partitioned by spectral clustering. This method is referred to as NC-spectral clustering.

3.1 Nearest correlation method

The NC method can detect samples whose correlation is similar to the query without any teacher signals (Fujiwara et al. (2009b)).

The concept of the NC method is as follows. Suppose that the affine subspace P in Fig. 1 (left) shows the correlation among variables and all the samples on P have the same correlation. Although $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_5$ have the same correlation, $\boldsymbol{x}_6$ and $\boldsymbol{x}_7$ have a different correlation from the others. The NC method aims to detect samples whose correlation is similar to the query $\boldsymbol{x}_1$. In this example, $\boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_5$ on P should be detected.

At first, the whole space is translated so that the query becomes the origin as shown in Fig. 1 (right). That is, $\boldsymbol{x}_1$ is subtracted from all the other samples $\boldsymbol{x}_i$ $(i = 2, 3, \cdots, 7)$. Since the translated affine subspace contains the origin, it becomes the linear subspace V.

Next, a line connecting each sample and the origin is drawn. Suppose another sample is found on this line. In this case, $\boldsymbol{x}_2$-$\boldsymbol{x}_5$ and $\boldsymbol{x}_3$-$\boldsymbol{x}_4$ satisfy such a relationship as shown in Fig. 1 (right). The correlation coefficients of these pairs must be 1 or -1. On the other hand, $\boldsymbol{x}_6$ and $\boldsymbol{x}_7$ that are not the elements of V cannot make such pairs. Therefore, the pairs whose correlation coefficients are ± 1 are thought to have the same correlation as $\boldsymbol{x}_1$.

In practice, the threshold of the correlation coefficient γ ($0 < \gamma \leq 1$) has to be used, since there are no pairs whose correlation coefficient is strictly ± 1. Therefore, the pairs should be selected when the absolute values of their correlation coefficients are larger than γ.

Using the above procedure, the pairs whose correlation is similar to the query can be detected.

3.2 NC-spectral clustering

The correlation-based affinity matrix for spectral clustering can be constructed by using the NC method. Assume that samples $\boldsymbol{x}_n \in \Re^M$ ($n = 1, 2, \cdots, N$) are stored in the database. The procedure of the proposed NC-spectral clustering is as follows:

(1) Set the zero matrix $\boldsymbol{S} \in \Re^{N \times N}$, γ ($0 < \gamma \leq 1$), and $L = 1$.
(2) Set the zero matrix $\boldsymbol{S}_L \in \Re^{N \times N}$.
(3) $\boldsymbol{x}'_n = \boldsymbol{x}_n - \boldsymbol{x}_L$ for $n = 1, 2, \cdots, N$ ($n \neq L$).
(4) Calculate the correlation coefficients $C_{k,l}$ between all possible pairs of $\boldsymbol{x}'_k$ and $\boldsymbol{x}'_l$ ($k \neq l$).
(5) Select all the pairs of $\tilde{k}$ and $\tilde{l}$ satisfying $|C_{k,l}| \geq \gamma$.
(6) $(\boldsymbol{S}_L)_{\tilde{k},\tilde{l}} = (\boldsymbol{S}_L)_{\tilde{l},\tilde{k}} = 1$.
(7) $\boldsymbol{S} = \boldsymbol{S} + \boldsymbol{S}_L$.
(8) If $L = N$, output $\boldsymbol{S}$ as the affinity matrix. Otherwise, $L = L + 1$ and return to 2.
(9) Partition the graph expressed by $\boldsymbol{S}$ through spectral clustering.

In the above procedure, steps 3∼5 correspond to the NC method.

3.3 Illustrative example

The detailed function of the proposed method is illustrated through a simple example. An objective data set consists of nine samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_9 \in \Re^2$; $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_4$ and $\boldsymbol{x}_5, \boldsymbol{x}_6, \cdots, \boldsymbol{x}_8$ are on the lines l and k, respectively, as shown in Fig. 2. On the other hand, $\boldsymbol{x}_9$ is an outlier. That is, the data set consists of three classes $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_4\}$, $\{\boldsymbol{x}_5, \boldsymbol{x}_6, \cdots, \boldsymbol{x}_8\}$ and $\{\boldsymbol{x}_9\}$. In addition, samples $\boldsymbol{x}_2$, $\boldsymbol{x}_7$, and $\boldsymbol{x}_9$ are arranged in line by chance.

First, the zero matrix $\boldsymbol{S} \in \Re^{9 \times 9}$ is defined, and the correlations of all possible pairs of samples are checked by the NC method. For example, when $\boldsymbol{x}_1$ is the query, $\boldsymbol{x}_2$-$\boldsymbol{x}_3$, $\boldsymbol{x}_2$-$\boldsymbol{x}_4$, and $\boldsymbol{x}_3$-$\boldsymbol{x}_4$ are detected as pairs whose correlation is similar to $\boldsymbol{x}_1$, and $(\boldsymbol{S})_{2,3} = (\boldsymbol{S})_{3,2} = 1$, $(\boldsymbol{S})_{2,4} = (\boldsymbol{S})_{4,2} = 1$, and $(\boldsymbol{S})_{3,4} = (\boldsymbol{S})_{4,3} = 1$.

In the same way, since $\boldsymbol{x}_1$-$\boldsymbol{x}_3$, $\boldsymbol{x}_1$-$\boldsymbol{x}_4$, $\boldsymbol{x}_3$-$\boldsymbol{x}_4$, and $\boldsymbol{x}_7$-$\boldsymbol{x}_9$ are detected when $\boldsymbol{x}_2$ is the query, one is added to the elements of $\boldsymbol{S}$ corresponding to these pairs. As a result, $(\boldsymbol{S})_{3,4} = (\boldsymbol{S})_{4,3} = 2$ because the pair $\boldsymbol{x}_3$-$\boldsymbol{x}_4$ is detected again. This procedure is repeated so that all samples become the query. Finally, the affinity matrix $\boldsymbol{S}$ becomes as follows:

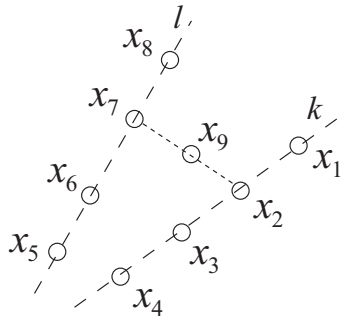

Fig. 2. An objective data set

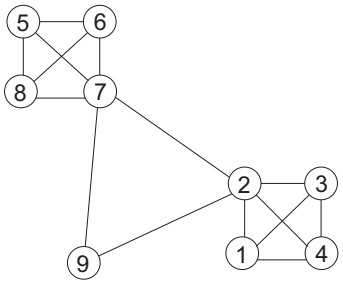

Fig. 3. A weighted graph expressing the affinity matrix $\boldsymbol{S}$

$$\boldsymbol{S} = \begin{bmatrix} 0 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & 2 & 0 & 0 & 1 & 0 & 1 \\ 2 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 2 & 2 & 0 \\ 0 & 1 & 0 & 0 & 2 & 2 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}. \quad (6)$$

In $\boldsymbol{S}$, $(\boldsymbol{S})_{2,7} = (\boldsymbol{S})_{7,2} = 1$, $(\boldsymbol{S})_{2,9} = (\boldsymbol{S})_{9,2} = 1$, and $(\boldsymbol{S})_{7,9} = (\boldsymbol{S})_{9,7} = 1$ since samples $\boldsymbol{x}_2$, $\boldsymbol{x}_7$, and $\boldsymbol{x}_9$ are arranged in line by chance and the pairs of these samples are detected by the NC method. However, the weights of these pairs in $\boldsymbol{S}$ are smaller than those of the pairs that have the true correlation.

Figure 3 shows an example of the graph expression of the calculated affinity matrix $\boldsymbol{S}$. In Fig. 3, the length of each arc is inversely proportional to their weights. By partitioning this graph using NWJ algorithm that is a revised spectral clustering algorithm proposed by Ng et al. (2001), the nodes are classified into $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_4\}$, $\{\boldsymbol{x}_5, \boldsymbol{x}_6, \cdots, \boldsymbol{x}_8\}$, and $\{\boldsymbol{x}_9\}$. This example clearly shows that the proposed NC-spectral clustering can classify samples according to the correlation among variables.

3.4 Numerical examples

The discrimination performance of the proposed NC-spectral clustering is compared with that of the k-means method and spectral clustering through numerical examples.

Two-dimensional case. The discrimination performances are compared through a two-dimensional example. The ob-

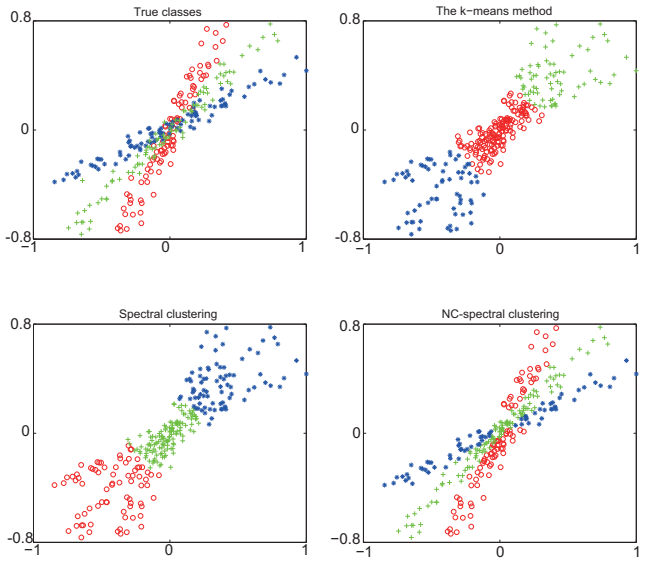

Fig. 4. Classification results of the 2-dimensional samples: true classes (top-left), the k-means method (top-right), spectral clustering (bottom-left), and NC-spectral clustering (bottom-right)

jective data set consists of three classes that have different correlation. 100 samples are generated in each class by

$$\boldsymbol{x} = s\boldsymbol{a}_i + \boldsymbol{n} \quad (i = 1, 2, 3) \tag{7}$$

where $s \sim N(0, 10)$, $\boldsymbol{n} = [n_1 \ n_2]^T$, $n_i \sim N(0, 0.1)$, and $N(m, \sigma)$ is a random number following the normal distribution whose mean is m and standard deviation is σ. The coefficient matrices $\boldsymbol{a}_i \in \Re^2$ are $\boldsymbol{a}_1 = [\ 1 \ \ 2\ ]^T$, $\boldsymbol{a}_2 = [\ 2 \ \ 2\ ]^T$, and $\boldsymbol{a}_3 = [\ 2 \ \ 1\ ]^T$.

In the conventional spectral clustering, the affinities between samples are defined by using the Gaussian kernel in (5), the Euclidean distance is used as a distance function $d(\cdot, \cdot)$, and $\sigma = 1$. In the proposed method, the parameter of the NC method is $\gamma = 0.999$.

The generated samples and the clustering results of the k-means method, spectral clustering, and the NC-spectral clustering are shown in Fig. 4. The conventional methods cannot classify samples correctly. On the other hand, the proposed method can classify samples accurately in most regions except around the origin.

Five-dimensional case. The discrimination performances are compared through a five-dimensional example. The objective data set consists of three classes that have different correlation. 100 samples are generated in each class by

$$\boldsymbol{x} = \boldsymbol{A}_i\boldsymbol{s} + \boldsymbol{n} \quad (i = 1, 2, 3) \tag{8}$$

where $\boldsymbol{s} = [s_1 \ s_2]^T$, $s_i \sim N(0, 10)$, and $\boldsymbol{n} = [n_1 \ n_2]^T$, $n_i \sim N(0, 0.1)$. The coefficient matrices $\boldsymbol{A}_i \in \Re^{5\times 2}$ are as follows:

$$\boldsymbol{A}_1 = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 1 \\ 2 & 3 \\ 1 & 3 \end{bmatrix} \quad \boldsymbol{A}_2 = \begin{bmatrix} 3 & 3 \\ 2 & 1 \\ 3 & 1 \\ 3 & 2 \\ 2 & 0 \end{bmatrix} \quad \boldsymbol{A}_3 = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 1 & 3 \\ 0 & 4 \\ 3 & 1 \end{bmatrix} \tag{9}$$

The discrimination rate is defined as

Table 1. Discrimination performances in the 5-dimentional example

	Discrimination rate [%]		
	Class 1	Class 2	Class 3
The k-means method	43	19	75
Spectral clustering	40	35	31
NC-spectral clustering	85	84	93

$$\text{Discrimination rate}[\%] = \frac{L}{K} \times 100 \tag{10}$$

where K is the number of detected samples and $K = 100$ in this example. L ($L \leq K$) is the number of samples that belong to the true class, out of K detected samples. The settings of spectral clustering and NC-spectral clustering are the same as the previous example.

Table 1 shows the discrimination performances of the k-means method, spectral clustering, and NC-spectral clustering. The proposed NC-spectral clustering can achieve the higher discrimination performance than the other methods. These results clearly show that NC-spectral clustering can discriminate the correlation among variables.

4. SOFT-SENSOR DESIGN BASED ON NC-SPECTRAL CLUSTERING

The operation data are clustered according to their characteristics to construct soft-sensors that can cope with the individual difference of production devices. Since the individual difference affects the correlation among variables, the proposed NC-spectral clustering can classify the operation data according to their characteristics.

In the present work, a new soft-sensor design method based on NC-spectral clustering is proposed. In the proposed method, the operation data are clustered by using NC-spectral clustering and models are constructed for each class.

Assume that the input samples $\boldsymbol{x}_n \in \Re^M$ $(n = 1, 2, \cdots, N)$ and output samples $\boldsymbol{y}_n \in \Re^L$ are stored in the database. The proposed soft-sensor design procedure is as follows:

(1) Classify the input samples $\boldsymbol{x}_n$ to P classes using NC-spectral clustering, and $\Omega_j = \{n \mid \boldsymbol{x}_n \in K_j\}$, $(j = 1, 2, \cdots, P)$.
(2) Construct models $f_j : X \longrightarrow Y$ from $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ $(i \in \Omega_j)$ for K_j, where X is the set of input and Y is that of output.
(3) Classify a new sample $\tilde{\boldsymbol{x}}$ to class $K_{\tilde{j}} = h(\tilde{\boldsymbol{x}})$ when $\tilde{\boldsymbol{x}}$ is measured, where $h : X \longrightarrow K$ is a classifier and K is the set of class.
(4) Calculate an estimate $\hat{\boldsymbol{y}} = f_{\tilde{j}}(\tilde{\boldsymbol{x}})$.

In the above algorithm, any modeling method can be used for building a local model f. In the present work, PLS is used to cope with the collinearity problem.

The Q statistic (Jackson and Mudholkar (1979)) can be used as the evaluation function of a classifier h. The Q statistic is derived from principal component analysis (PCA), and it is the distance between the sample and the subspace spanned by principal components. In other words, the Q statistic is a measure of dissimilarity between the sample and the modeling data from the viewpoint of the correlation among variables.

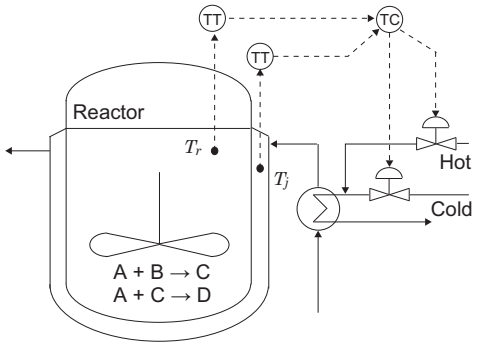

Fig. 5. Schematic diagram of the batch reactor with multivariable control system

The class that minimizes the Q statistic of sample $\boldsymbol{x}$ should be selected as its class. The classifier h is described as

$$h(\boldsymbol{x}) = \arg\min_{K_j} \ Q_j \tag{11}$$

$$Q_j = \boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{V}_R^{\{j\}}\boldsymbol{V}_R^{\{j\}T})\boldsymbol{x} \tag{12}$$

where $\boldsymbol{V}_R^{\{j\}}$ is the loading matrix of PCA derived from the matrix $\boldsymbol{X}^{\{j\}}$ whose rows are samples belonging to class K_j.

5. CASE STUDY

In this section, the proposed NC-spectral clustering is compared with the k-means method through their applications to operation data of parallelized batch processes. In addition, soft-sensors for product composition are designed on the basis of their clustering results. The detailed batch process model used in this case study is described in Cott et al. (1989).

5.1 Problem setting

A schematic diagram of the batch reactor is shown in Fig. 5. In this process, a well-mixed, liquid-phase reaction system is considered, and two reactions take place:

$$\begin{array}{l} \text{reaction 1}: \ \mathrm{A} + \mathrm{B} \longrightarrow \mathrm{C} \\ \text{reaction 2}: \ \mathrm{A} + \mathrm{C} \longrightarrow \mathrm{D} \end{array} \tag{13}$$

The component C is the desired product while D is the byproduct. The objective is to achieve a good conversion of product C. The initial reactor temperature is 20 °C, and the initial amount of the row materials A and B changes as the random numbers following $N(20, 0.1)$.

Since reaction 1 proceeds at 90 °C or higher, the reactor temperature should be raised as fast as possible after operation starts. As reaction 1 proceeds, the reactor temperature increases due to reaction heat. However, reaction 2 proceeds at the reactor temperature exceeding 100 °C, and product C is converted to byproduct D. Therefore, a rise in the reaction temperature has to be controlled after it reaches around 90 °C. In this process, the reactor temperature T_r and the jacket temperature T_j are measured every one minute and controlled through the multivariable control system. The termination time of operation is 120 minutes.

In this case study, five reactors $R_1, R_2, \cdots, R_5$ are operated in parallel. In addition, their heat transfer coefficients are unchanged during the batch operation, and they

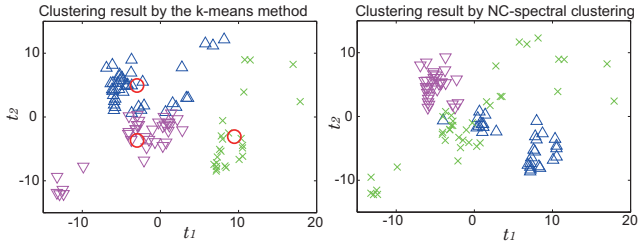

Fig. 6. Classfication results: the k-means method (left) and NC-spectral clustering (right)

change every batch operation as the random number. The heat transfer coefficients U_i $(i = 1, 2, \cdots, 5)$ are

$$U_i = \begin{cases} U(40.60, 40.62), & i = 1, 2 \\ U(40.57, 40.59), & i = 3, 4 \\ U(40.54, 40.56), & i = 5 \end{cases} \tag{14}$$

where $U(a, b)$ denotes the uniform random numbers on the closed interval $[a, b]$. That is, there are three types of the correlation among variables although there are five reactors. In this case study, the operation data of 20 batches of each reactor are stored in the database. The objective is to construct soft-sensors that can estimate the amount of the product C accurately at the end of the batch.

5.2 Clustering

Before constructing soft-sensors, the operation data of all 100 batches stored in the database are clustered into three classes using the k-means method and the proposed NC-spectral clustering. As preprocessing of the operation data, its dimension is reduced by Multiway PCA (Nomikos and MacGregor (1994)); the number of the retained principal components is two. That is, the input variables of these clustering methods are the scores t_1 and t_2. The parameter of NC-spectral clustering is $\gamma = 0.99$.

The clustering results of the k-means method and NC-spectral clustering are shown in Fig. 6, where the sample distribution on the t_1-t_2 plane is shown. In the result of the k-means method, the center of each class $\boldsymbol{c}_j$ $(j = 1, 2, 3)$ is designated by a circle, and samples are certainly classified based on the distance. On the other hand, in the case of NC-spectral clustering, samples are classified regardless of the distance.

5.3 Soft-sensor design

A soft-sensor is constructed to estimate the amount of product M_{C} [kmol] at the end of the batch. A model is built for each of three classes clustered by NC-spectral clustering. The input variables of soft-sensors are the time series of T_r and T_j, and PLS is used for modeling.

In addition, another soft-sensor is designed on the basis of the clustering result of the k-means method. In this case, the distance from the class center $\boldsymbol{c}_j$ is used for sample discrimination when an output is requested. That is, the classifier g is described a

$$g(\boldsymbol{x}) = \arg\min_{K_j} \ ||\boldsymbol{x} - \boldsymbol{c}_j||_2. \tag{15}$$

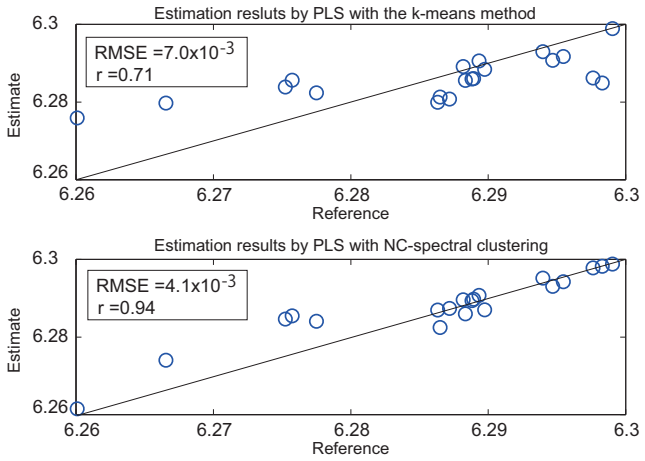

Fig. 7. Estimation results by PLS with the k-means method (top) and NC-spectral clustering (bottom)

In this case study, M_C of the new reactor R_6 is estimated. The heat transfer coefficient of R_6 changes as the random number following $U(40.60, 40.62)$, which is the same as that of R_1 and R_2. The number of the validation batches is 20.

The estimation results of soft-sensors are shown in Fig. 7. In these figures, the horizontal line and the vertical line express the measurement and the estimates, respectively. RMSE is the root-mean-squared error and r denotes the correlation coefficient between measurements and estimates. These results clearly show that the proposed method can achieve the higher estimation performance than the k-means method-based soft-sensor, and RMSE is improved by 43%.

6. CONCLUSION

A new clustering method is proposed by integrating the NC method and spectral clustering. The proposed NC-spectral clustering can accurately discriminate the correlation among variables. In addition, a new soft-sensor design method based on NC-spectral clustering is proposed. Since NC-spectral clustering can discriminate the individual difference of production devices, the proposed soft-sensor design method can improve the estimation performance. The superiority of the proposed method over conventional methods is demonstrated through a numerical example and a case study of parallelized batch processes.

REFERENCES

Amirthalingam, R. and Lee, J. (1999). Subspace Identification Based Inferential Control Applied to a Continuous Pulp Digester. *J. Process Control*, 9, 397-406.

Bontempi, G., Birattari, M., and Bersini, H. (1999). Lazy Learning for Local Modeling and Control Design. *Int. J. Control*, 72, 643-658.

Cheng, C., and Chiu, MS. (2004). A New Data-Based Methodology for Nonlinear Process Modeling. *Chem. Eng. Sci.*, 59, 2801-2810.

Cott, BJ., and Macchietto, S. (1998). Temperature Control of Exothermic Batch Reactors Using Generic Mode Control. *Ind. Eng. Chem. Res.*, 28, 1177-1184.

Ding, C., He, X., Zha, H., Gu, M., and Simon, HD. (2001). A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering. *Proc. of IEEE Int'l Conf. Data Min.*, San Jose.

Fujiwara, K., Kano, M., and Hasebe, S. (2009a). Soft-Sensor Development using Correlation-Based Just-In-Time modeling. *AIChE J.*, 55, 1754-1765.

Fujiwara, K., Kano, M., and Hasebe, S. (2009b). Correlation-Based Pattern Recognition Algorithm and Its Application to Adaptive Software Sensing. *Proc. of ADCHEM*, Istanbul, Turkey.

Jackson, JE., and Mudholkar, GS. (1979). Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, 21, 341-349.

Kamohara H., Takinami A., Takeda M., Kano M., Hasebe S., and Hashimoto I. (2004). Product Quality Estimation and Operating Condition Monitoring for Industrial Ethylene Fractionator. *J. Chem. Eng. Japan*, 37, 422-428.

Kaneko, H., Arakawa, M., and Funatsu, K. (2009). Development of a new soft sensor method using independent component analysis and partial least squares. *AIChE J.*, 55, 87-98.

Kano, M., Lee, S.,, and Hasebe, S. (2008). Two-Stage Subspace Identification for Softsensor Design and Disturbance Estimation. *J. Process Control*, 19, 179-186.

Kano, M., Miyazaki, K., Hasebe, S., and Hashimoto, I. (2000). Inferential Control System of Distillation Compositions Using Dynamic Partial Least Squares Regression. *J. Process Control*, 10, 157-166.

Kano, M. and Nakagawa Y. (2008). Data-Based Process Monitoring, Process Control, and Quality Improvement: Recent Developments and Applications in Steel Industry. *Comput. Chem. Eng.*, 32, 12-24.

Kano, M. and Ogawa, M. (2009). The State of the Art in Advanced Chemical Process Control in Japan. *Proc. of ADCHEM*, Istanbul, Turkey.

Kohonen, T. (2001). Self-organizing maps. Springer, New York, 3rd edition.

Mejdell, T., and Skogestad, S. (1991). Estimation of Distillation Compositions from Multiple Temperature Measurements Using Partial-Least-Squares Regression. *Ind. Eng. Chem. Res*, 30, 2543-2555.

Ng, AN., Jordan, MI., and Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm. *Proc. of Neural Inf. Process. Syst.*, Vancouver.

Nomikos, P., and MacGregor, JF. (1994). Monitoring Batch Processes Using Multiway Principal Component Analysis. *AIChE J.*. 40, 1361-1375.

Qin, JS. (1998). Recursive PLS Algorithms for Adaptive Data Modeling. *Comput. Chem. Eng.*, 22, 503-514.

Shigemori H., Kawamura S., Kano M., and Hasebe S. (2009) Optimum Quality Design System for Steel Products through Locally Weighted Regression Model. *J. Process Control*, submitted.