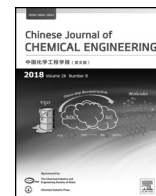




Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

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

Article

Total plant performance evaluation based on big data: Visualization analysis of TE process[☆]Mengyao Li, Wenli Du^{*}, Feng Qian^{*}, Weiming Zhong

Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

ARTICLE INFO

Article history:

Received 4 October 2017

Received in revised form 29 April 2018

Accepted 11 June 2018

Available online 19 June 2018

Keywords:

Performance evaluation

Structured logic rules

Hierarchical framework

Multidimensional visualization

KPCA–SOM

ABSTRACT

The performance evaluation of the process industry, which has been a popular topic nowadays, can not only find the weakness and verify the resilience and reliability of the process, but also provide some suggestions to improve the process benefits and efficiency. Nevertheless, the performance assessment principally concentrates upon some parts of the entire system at present, for example the controller assessment. Although some researches focus on the whole process, they aim at discovering the relationships between profit, society, policies and so forth, instead of relations between overall performance and some manipulated variables, that is, the total plant performance. According to the big data of different performance statuses, this paper proposes a hierarchical framework to select some structured logic rules from monitored variables to estimate the current state of the process. The variables related to safety and profits are regarded as key factors to performance evaluation. To better monitor the process state and observe the performance variation trend of the process, a classification-visualization method based on kernel principal component analysis (KPCA) and self-organizing map (SOM) is established. The dimensions of big data produced by the process are first reduced by KPCA and then the processed data will be mapped into a two-dimensional grid chart by SOM to evaluate the performance status. The monitoring method is applied to the Tennessee Eastman process. Monitoring results indicate that off-line and on-line performance status can be well detected in a two-dimensional diagram.

© 2018 The Chemical Industry and Engineering Society of China, and Chemical Industry Press. All rights reserved.

1. Introduction

The performance evaluation has many advantages to the process industry, for example, assessing the ability of a system, identifying the weakness, adjusting and optimizing, and verifying the resilience and reliability of the process. It has become a significant issue to seek out the appropriate method to assess the process performance accurately.

At present, much research has been conducted on the performance assessment of the process industry, most of which are concentrated on two aspects: one on the evaluation of factory management and the other on the assessment of controllers. Studies on factory management are mainly focused on improving the sustainability of factories [1–4]. In these articles, a performance index is first proposed, and some factors related to the performance index are listed; the sub-factors of the related factors are listed step by step, namely, the famous analytical hierarchy process (AHP) [5]. For example, RK Singh *et al.* [6] presented a methodology for developing a composite sustainability performance index (CSPI), which considered economy, environment, society,

governance, and technology as indexes to assess the sustainable performance of steel industries. The proposed index is typically used to compare the performance among different companies [7, 8] and provide some efficient information to decision makers. Several articles on the performance assessment of controllers also exist [9–12]. They use different methods to improve the performance of controllers, which will enhance the performance of systems at the same time. For instance, F Xu *et al.* [13] evaluated the MPC economic performance by solving benefit potentials through either variability reduction of quality output variables or tuning of constraints. NA And *et al.* [14,15] assessed the MPC performance by considering economic performance, variability reduction, and their relationships based on constraint tuning. Few studies are concerned on the total plant performance of the process.

During the course of operation, many unexpected variable variations may occur, such as the variations of feed composition, temperature, pressure, and product quality. Some of these variable variations do not have much impact on the process, whereas others may have a significant impact on the process. Therefore, determining whether these variable variations are noteworthy, that is, finding the suitable structured logic rules from data of variable variation that play pivotal roles in the total plant performance, is important.

The structured logic rules based on big data should be determined initially to evaluate the total plant performance of the process. The

[☆] Supported by the National Natural Science Foundation of China (61590923, 61422303, 21376077).

^{*} Corresponding authors.

E-mail addresses: wldu@ecust.edu.cn (W. Du), fqian@ecust.edu.cn (F. Qian).

main purpose of a process industry is to obtain the benefits; thus, the interest of the factory is a good criterion to evaluate the current operating status. In addition, safety is a worth considering criterion to guarantee normal production. Therefore, the monitored variables related to profit and safety can be regarded as the logic rules to assess the total plant performance of the process.

The paper is organized as follows. In Section 2, the hierarchy framework of performance assessment is provided and the method is used in TE process. The algorithm of KPCA–SOM and the entire flow of the on-line monitoring of the process are introduced in Section 3, followed by the experimental results in Section 4. Conclusion and discussion are provided in Section 5.

2. Hierarchy Framework for Performance Assessment

2.1. Establishment of hierarchy framework

This paper proposes a hierarchy framework to choose the performance indexes of the process industry based on the history data and the operation requirement, which is shown in Fig. 1. Three criteria determine the current performance of a process, namely, hard safety indexes, soft safety indexes, and profit indexes. Influence factors indicate the factors related to the three criteria. Influence factors and three criteria constitute the entire logic rules of the process. Hard safety indexes are factors that may lead to serious failures or even shut down and represent several indexes that the process must satisfy. The influence factors of hard safety indexes are referred to as constraints in this paper. Generally, the process that has grave accident will be regarded as a poor case if the process data exceed the boundary of constraints. If needed, the poor case may be classified into poor case 1, 2, ..., k_1 . k_1 is less than or equal to the number of constraints. Soft safety indexes represent several indexes that have limited effects on the process;

sometimes, these effects can recover automatically. The influence factors of soft safety indexes are expressed as conditions herein. If the process data belong to the range of conditions, which means that the process has a small fault, it is deemed to be less poor and also can be divided into less poor case 1, 2, ..., k_2 . Similarly, k_2 is less than or equal to the number of conditions. The constraints and conditions are determined according to the actual process. The entire productive process is safe if the hard and soft safety indexes are within normal range. Therefore, the profit of the whole process can be used as another judging criterion. Many factors are concerned with profit, such as raw material costs, operating costs (electric charge, water rate, material wastage, etc.), and product profits. For reason that when the process produces pollutant, some actions should be taken to reduce pollution which will also cause the increase of costs, environment factor is considered as a part of profit as well. As long as the factor generates costs, it will be taken into account in the profit index. Hence, the profit of the process is a summation function, and it is divided into several cases 1, 2, ..., k_3 on the basis of the historical data of profit values in diverse operating conditions. $k_1 + k_2 + k_3 = N$. Therefore, there are now N cases.

The realization process of the proposed hierarchy framework method for performance evaluation is described in detail in Section 2.2 based on TE process.

2.2. Establishment of hierarchy framework in TE process

2.2.1. Description of TE process

The TE process based on an actual industrial process was provided by J.J. Downs and E. F. Vogel [16], which has been used in many different fields, such as fault diagnosis and monitoring [17,18] and optimization [19,20]. In this paper, the TE process is used for performance assessment in accordance with the method proposed above.

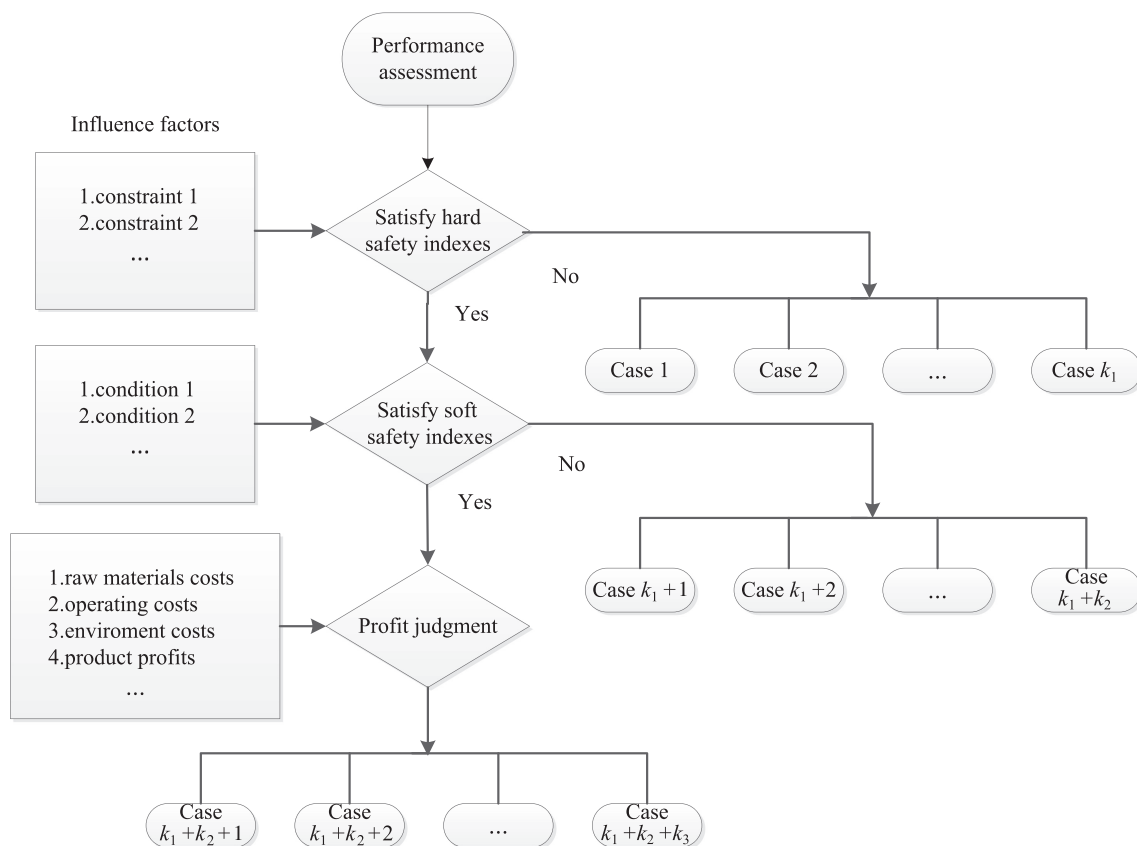
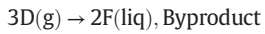
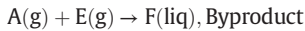
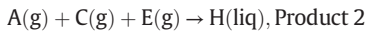
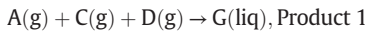


Fig. 1. Hierarchy framework for performance assessment.

The TE process has a total of eight components: A, B, C, D, E, F, G, and H. A, C, D, and E are reactants, and G and H are products. B is an inert and F is a byproduct. The reactions are as follows:



The process has five operation units, namely, the reactor, product condenser, vapor–liquid separator, recycle compressor, and product stripper shown in Fig. 1 of paper [16].

The TE process has 12 manipulated variables and 41 measurements. There are five hard safety indexes in the process: reactor pressure, reactor level, reactor temperature, product separator level, and stripper base level (see Table 1). The soft safety indexes of the process are the flowrate changes of product, composition variability of G, and flow variability of streams 1, 2, and 4 (see Table 2). Tables 1 and 2 refer to the paper ‘A Plant-wide Industrial process control problem’ written by Downs and E. F. Vogel [16]. The operating costs of the process [21] is

$$\begin{aligned} c_1(t) = & 0.0536y_{20}(t) + 0.0318y_{19}(t) \\ & + 0.44791y_{10}(t)[2.209y_{29}(t) + 6.177y_{31}(t) \\ & + 22.06y_{32}(t) + 14.56y_{33}(t) + 17.89y_{34}(t) \\ & + 30.44y_{35}(t) + 22.94y_{36}(t)] \\ & + 4.541x_{46}(t)[0.2206y_{37}(t) + 0.1456y_{38}(t) \\ & + 0.1789y_{39}(t)] \end{aligned} \quad (1)$$

Table 1
Hard safety indexes

Process variable	Normal operating limits		Shut down limits	
	Low limit	High limit	Low limit	High limit
Reactor pressure	None	2895 kPa	None	3000 kPa
Reactor level	50% (11.8 m ³)	100% (21.3 m ³)	2 m ³	24 m ³
Reactor temperature	None	150 °C	None	175 °C
Product separator level	30% (3.3 m ³)	100% (9.0 m ³)	1 m ³	12 m ³
Stripper base level	30% (3.5 m ³)	100% (6.6 m ³)	1 m ³	8 m ³

Table 2
Soft safety indexes

Production variability	Flowrate changes	Less than $\pm 5\%$ in the range 8–16 h ⁻¹
	Composition variability	Less than $\pm 5\text{mol}\%$ G in the range 6–10 h ⁻¹
Feed flow variability	Stream 1	Minimize flow variability in the range 8–16 h ⁻¹
	Stream 2	Minimize flow variability in the range 8–16 h ⁻¹
	Stream 4	Minimize flow variability in the range 12–80 h ⁻¹

Table 3
Classes of TE process

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Profit/USD · h ⁻¹	≥ 5200	5150–5200	5100–5150	5000–5100	4900–5000	4700–4900	≤ 4700

where $c_1(t)$ is in USD · h⁻¹. The term $4.541x_{46}(t)$ in Eq. (1) is the product rate in kmol · h⁻¹. $y_i(t)$ is the process measurements. The profit function of the process is

$$\text{profit}(t) = p_G(t) + p_H(t) - c_1(t) - c_2(t) \quad (2)$$

where $p_G(t)$ is the profit of product G, $p_H(t)$ is the profit of product H, and $c_2(t)$ is the cost of raw materials. Given that the sale prices of G and H are not given in the paper [16], it is supposed that their sale prices are the twice of the costs. Because the paper [16] does not discuss the effects of byproduct, exhaust gases, separate costs of G and H etc., these costs are ignored. The specific computational process is shown in Appendix. The profit function can be adjusted according to actual working conditions.

2.2.2. Selections of judgment conditions

Simulink built by N.L. Richer [22] is adopted in this process. The data of the TE process can be obtained by altering the 12 manipulated variables and the sampling time by this Simulink. In each hour, 100 samples with 53 dimensions (12 manipulated variables and 41 measurements) are collected. To acquire the data of different statuses, alter one manipulated variable while keeping other manipulated variables the same as the base case. Since the regulation of one manipulated variable will affect other manipulated variables and 41 measurements, the performance statuses are unknown before the analysis of the data. The sampling time of each status is 48 h and each manipulated variable is changed four times. Thus, 48 sets of 4800×53 data are obtained. From these sets of data, it can be easily found that when any constraint exceeds the boundary, the process will be shut down, which means that a serious fault occurred. For simplicity, we only choose one poor case here, which means that when the process is in poor case, there are five possibilities (see Table 1). The production variability of flowrate changes and composition variability have limited impact on the process, and their effectiveness will decrease with time even when they are out of the range (see Table 2). Also, composition variability of five possibilities is chosen. When the hard and soft safety indexes are in the specified range, which means that the process is under normal operating condition, the profit of the process can be concerned. Solving Eq. (5) according to 48 sets of data, 7 levels of profits are chosen and the range of 7 classes is listed in Table 3. The profit range is not of equal distance because the profits of most statuses are in the range of 5100–5200 USD · h⁻¹. The composition variability of G more than $\pm 5\text{mol}\%$ in the range 6–10 h⁻¹ is defined as case 8, whereas flowrate change of more than $\pm 5\%$ in the range 8–16 h⁻¹ is case 9. The data exceed the boundaries of constraints, namely, poor case, is regarded as case 10. The 10 cases are now chosen from the 48 sets of data. On the basis of 10 cases, 10 sets of data are elected as training data.

3. The Description of KPCA–SOM Algorithm and Application in Performance Assessment

This section briefly reviews KPCA and SOM algorithm and describes the combination of KPCA–SOM method with the hierarchy framework for performance assessment.

3.1. KPCA

The PCA [23, 24] is a typical way of feature extraction and dimension reduction. However, PCA can only deal with linearly correlated data. KPCA [25, 26] is a non-linear expansion algorithm of linear PCA, which uses a nonlinear method to extract the principal components, that is, KPCA maps the original vector to the high-dimensional space F and performs PCA analysis in space F . Many papers combine KPCA with other algorithms, such as SVD [27], LDA [28], and SVM [29], to improve its performance.

The input matrix is $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in R^{n \times m}$ and the corresponding mapping from data space to feature space is $\Phi(\cdot)$, which is defined as

$$\begin{aligned} \Phi: R^m &\rightarrow F^h \\ \mathbf{x}_i &\mapsto \xi = \Phi(\mathbf{x}_i) \end{aligned} \quad (3)$$

where $\Phi(\mathbf{x}_i)$ meet the conditions of centralization:

$$\sum_{i=1}^n \Phi(\mathbf{x}_i) = 0$$

The covariance matrix is then

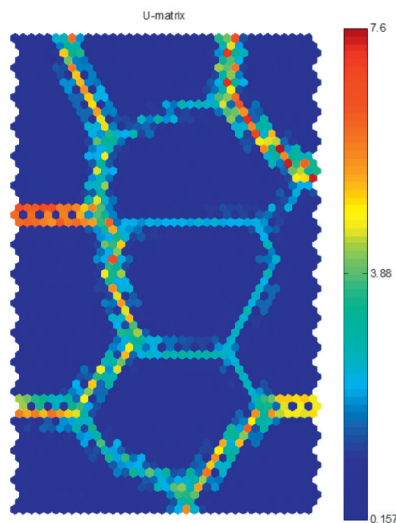
$$C^F = \frac{1}{n} \sum_{i=1}^n \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_i)^T \quad (5)$$

Solving the following eigenvalue decomposition problem:

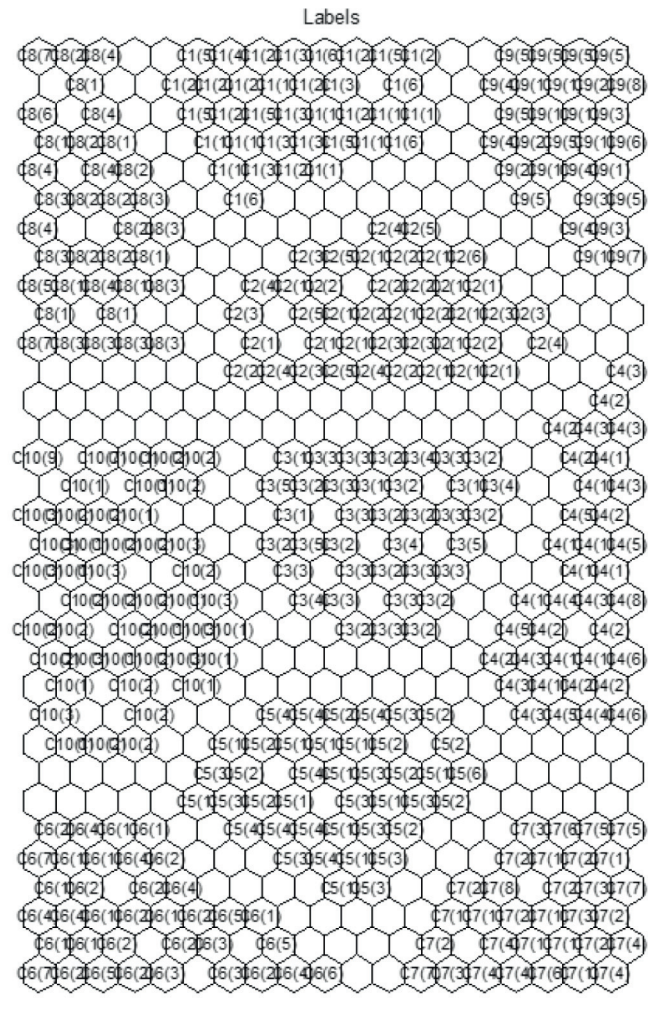
$$\begin{aligned} \lambda \mathbf{v} &= C^F \mathbf{v} \\ &= \left(\frac{1}{n} \sum_{i=1}^n \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_i)^T \right) \mathbf{v} = \frac{1}{n} \sum_{i=1}^n \langle \Phi(\mathbf{x}_i), \mathbf{v} \rangle \Phi(\mathbf{x}_i) \end{aligned} \quad (6)$$

where λ is the eigenvalues satisfying $\lambda \geq 0$; $\mathbf{v}$ is the eigenvectors satisfying $\mathbf{v} \in F \setminus \{0\}$ and $\langle \cdot, \cdot \rangle$ denotes the inner product. Considering that all solutions $\mathbf{v}$ with $\lambda \neq 0$ can be expressed as the span of $\Phi(\mathbf{x}_1)$, $\Phi(\mathbf{x}_2)$, ..., $\Phi(\mathbf{x}_n)$, that is

$$\mathbf{v} = \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) \quad (7)$$



(a) U-matrix diagram



(b) Mapping diagram

Fig. 2. Training diagrams of SOM.

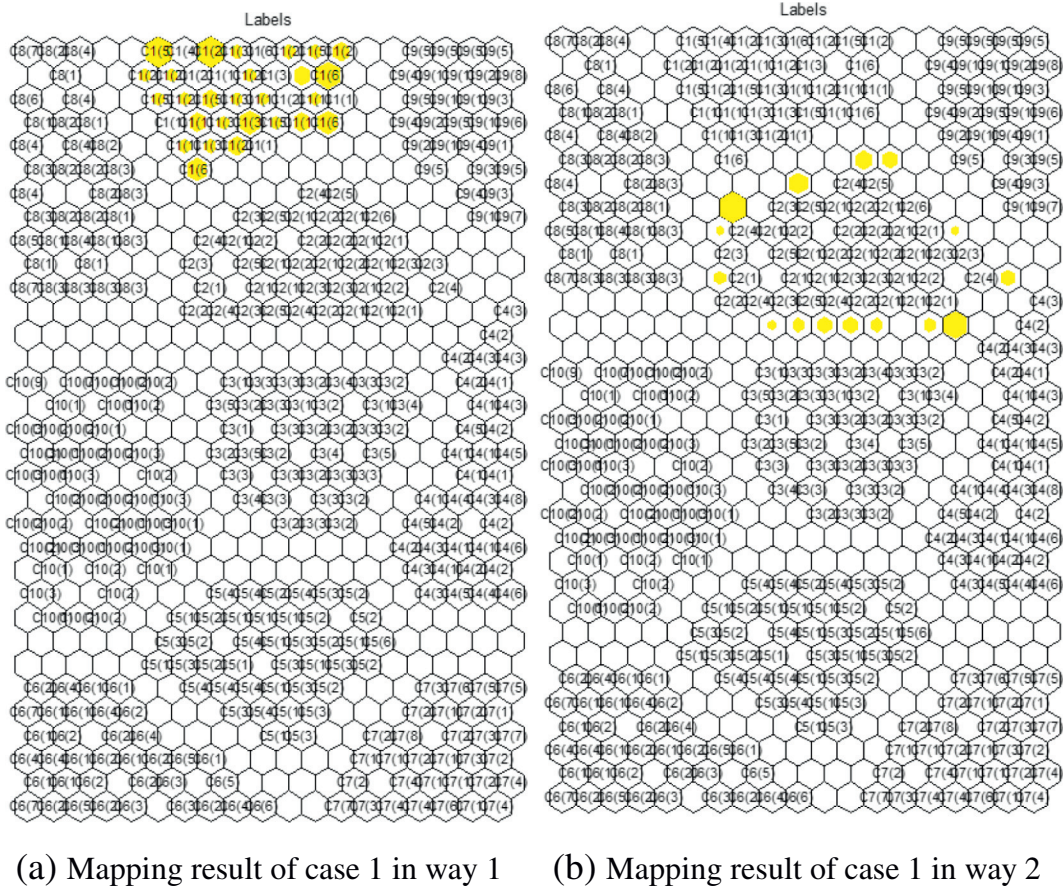


Fig. 3. Test diagrams of case 1.

By combining Eqs. (5), (6), and (7), the equation can be rewritten as

$$\lambda \sum_{i=1}^n \alpha_i \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_v) \rangle = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \alpha_j \langle \Phi(\mathbf{x}_j), \Phi(\mathbf{x}_i) \rangle \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_v) \rangle \quad (8)$$

Defining kernel matrix $\mathbf{K}_{ij} \in R^{n \times n}$ as

$$\mathbf{K}_{ij} = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \quad (9)$$

There are three kinds of kernel matrix, namely, Gaussian, polynomial, and sigmoid kernels. In this paper, the Gaussian kernel is adopted. Thus,

$$\mathbf{K}(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \in R^{n \times n} \quad (10)$$

where $\|x - y\|^2$ is the Euclidean distance matrix between $(\mathbf{X}, \mathbf{X})$ and σ is constant. The matrix $\mathbf{K}$ must be normalized with

$$\mathbf{K} \leftarrow \mathbf{K} - \mathbf{I}_n \mathbf{K} - \mathbf{K} \mathbf{I}_n + \mathbf{I}_n \mathbf{K} \mathbf{I}_n \quad (11)$$

where $\mathbf{I}_n = \frac{1}{n} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} \in R^{n \times n}$. Hence, Eq. (8) can be simplified as

$n\lambda \mathbf{K}\alpha = \mathbf{K}^2\alpha$, that is,

$$n\lambda \alpha = \mathbf{K}\alpha \quad \alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]^T \quad (12)$$

Solving Eq. (12) can derive the eigenvalues and eigenvectors. The kernel principal components (PCs) are selected according to λ . PCs with large λ should be placed into PC space, whereas PCs with small λ should be placed into residual space. The j th PC is calculated as follows:

$$\mathbf{t}_j = \langle \mathbf{v}^k, \Phi(\mathbf{x}) \rangle = \sum_{i=1}^n (\alpha_i)^j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle, \quad j = 1, 2, \dots, k \quad (13)$$

where k is the number of PCs. Compute the k largest eigenvectors of $\mathbf{K}$; pick up the corresponding eigenvector matrix $[\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_k]$; and set the global coordinates as $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_k]$.

After the above-mentioned algorithm, we can obtain a matrix $\mathbf{T} \in R^{n \times k}$. However, for each new example, relearning is necessary to obtain its low-dimensional embedding. To solve this problem, the base mapping matrix is defined as

$$\mathbf{MAP} = \mathbf{X}^T \times \mathbf{T} \quad (14)$$

With the base mapping matrix, the input matrix is mapped into a low dimensional space through Eq. (15).

$$\mathbf{X}' = \mathbf{X} \times \mathbf{MAP} \quad (15)$$

3.2. SOM

SOM is an unsupervised neural network proposed by Kohonen in 1989 [30], which is capable of projecting high-dimensional input data to a two-dimensional grid in such a way that input data close to each

other will be mapped to nearby neurons on the output map, which can preserve the topological structure. Therefore, the SOM can serve as a visualization tool. The SOM is used to reduce the dimensionality of the data and to effectively visualize multistate and transient operations in the literature [31, 32].

The SOM has only two layers: the input layer and the output layer. The input layer is a dataset $\mathbf{X}' \in R^{n \times k}$, $\mathbf{x}'_i = [x'_{i1}, x'_{i2}, \dots, x'_{ik}] (i = 1, 2, \dots, n)$, where n and k are the number and the dimension of input samples. The output layer is a set of ordered neurons, which are usually arranged in hexagonal or rectangular lattices. A hexagonal lattice is generally preferred because it provides better visualization. The connection between input and output layers is established via weight vector, which has the same dimension of the input vector. Hence,

$$\mathbf{w}_j = [w_{j1}, w_{j2}, \dots, w_{jk}] \quad j = 1, 2, \dots, J \quad (16)$$

where J is the number of neurons in the output layer. Since the number of the map neurons (J) has influence on the accuracy and generalization capability of the SOM, it should be specified according to [33]

$$J = 5\sqrt{n} \quad (17)$$

where n is the sample size.

To determine the activation degree of the neurons, some distance measures (usually Euclidean distance) are used to compare the input vectors to the weight vectors of the output neurons. The neuron with the smallest difference in reference vector will be chosen as the best matching unit (BMU), and the selection function is defined as

$$\mathbf{b}_i = \arg \min_j \|\mathbf{x}'_i - \mathbf{w}_j\| \quad (18)$$

$$i = 1, 2, \dots, n, j = 1, 2, \dots, J$$

When the winning neuron $\mathbf{b}_i$ is selected, the weight vector of the winning neuron will be updated slightly to bring it closer to the input vector. The weight vectors of the neighborhood neurons will also be updated, but the variations are smaller. The update function of weight vector is given below:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t)h_{bj}(t)\|\mathbf{x}'_i(t) - \mathbf{w}_j(t)\| \quad (19)$$

$$j = 1, 2, \dots, J$$

where $\alpha(t)$ is the learning rate parameter and $h_{bj}(t)$ is the neighborhood function.

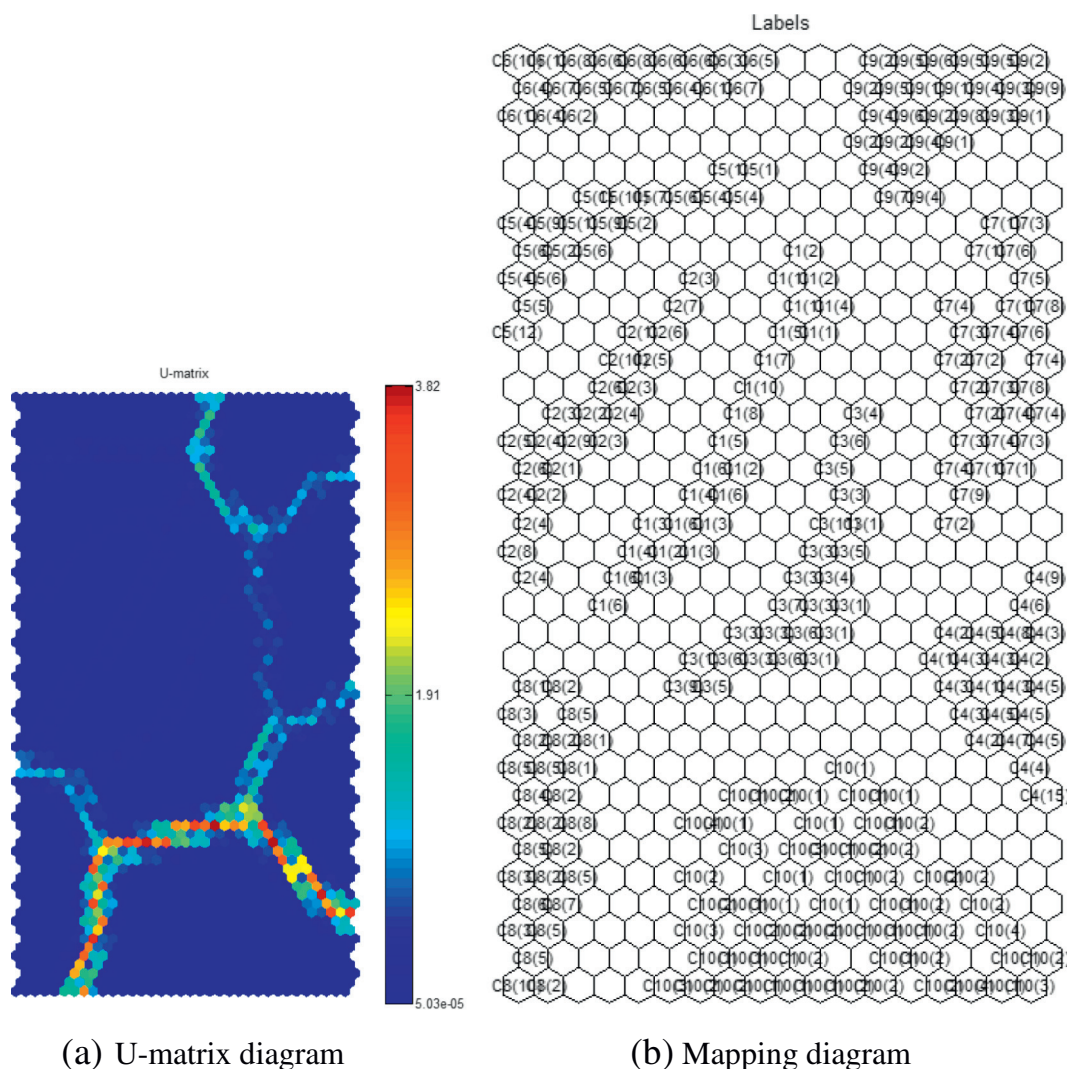
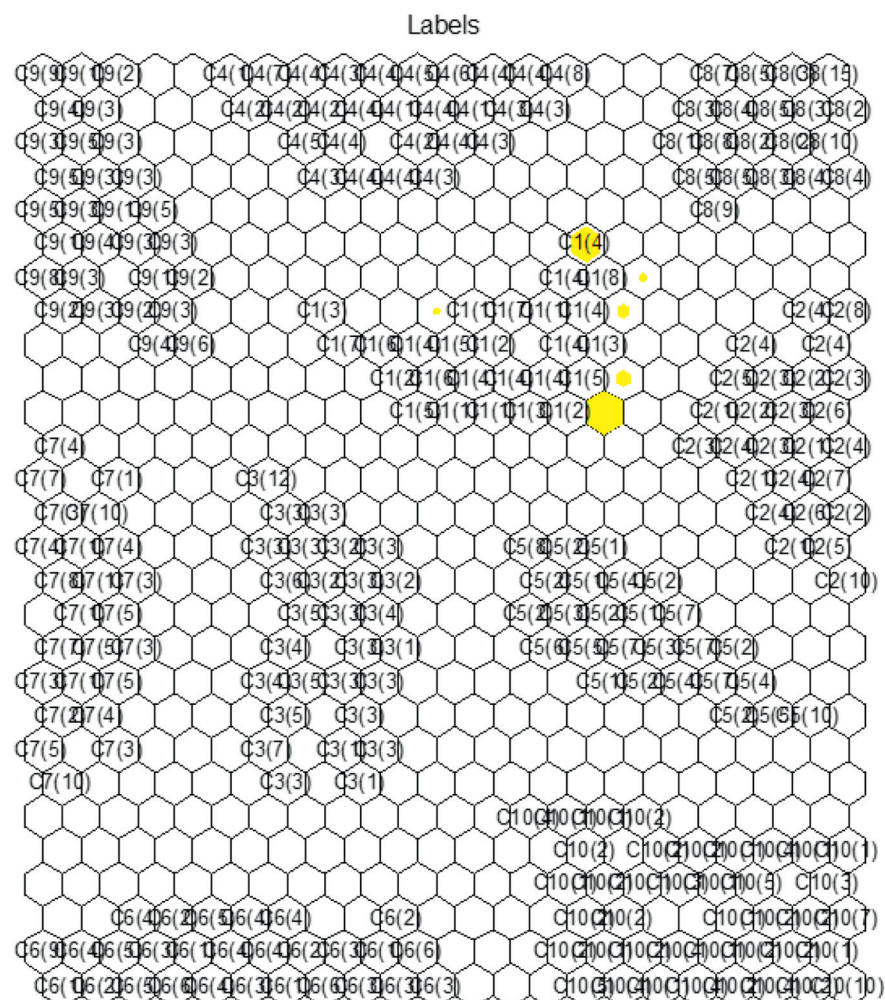
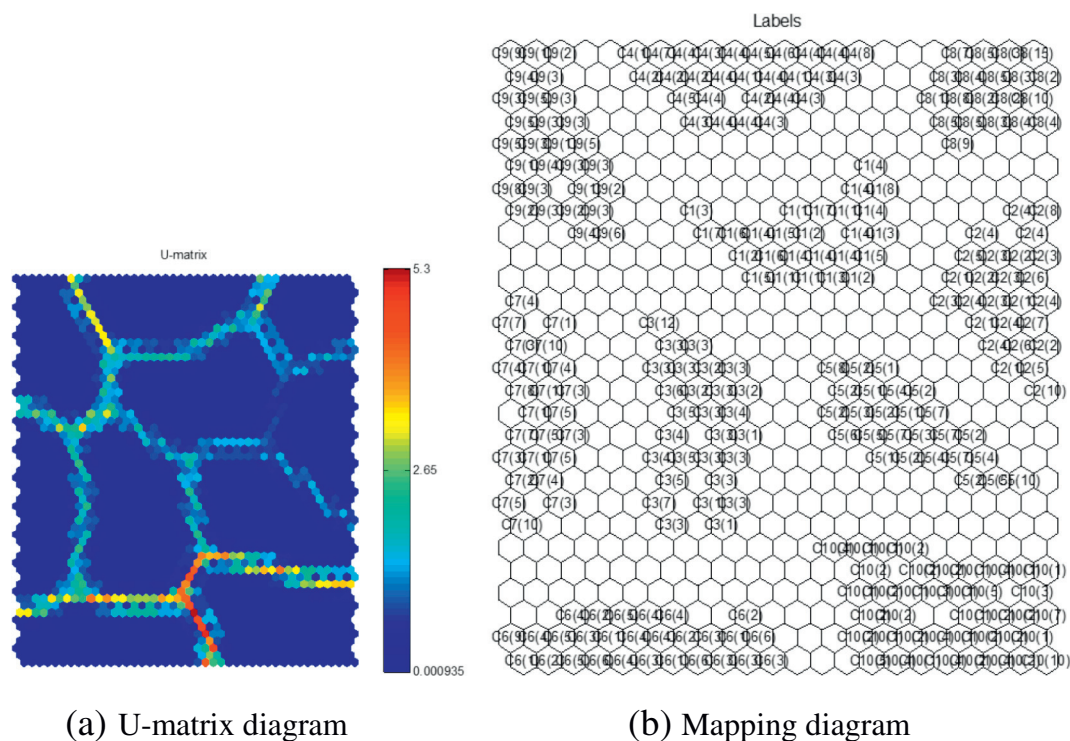


Fig. 4. Training diagrams of PCA-SOM (PCs 4).



Generally, the neighborhood function $h_{bj}(t)$ is a Gaussian function. Let the location of the j th neuron on the output map be $\mathbf{r}_j \in R^2$ and $\mathbf{r}_w$ is the position of the winning neuron, then

$$h_{bj}(t) = \exp\left(-\frac{\|\mathbf{r}_w - \mathbf{r}_j\|^2}{2\sigma(t)^2}\right) \quad (20)$$

where $\sigma(t)$ represents the neighborhood width of the neurons. To achieve convergence, $\sigma(t)$ and $\alpha(t)$ are normally assigned a large value and then gradually decrease with t . When $t \rightarrow \infty$, $\alpha(t) \rightarrow 0$ and $\sigma(t)$ approaches a small value (typically 1).

3.3. Performance assessment based on KPCA-SOM algorithm

Huge amount of data will be produced because the process industry has a large number of variables, and the data are updated in real-time. Therefore, it is a big problem to obtain the data needed and to achieve real-time evaluation of chemical process performance. This paper presents a method based on KPCA-SOM to visualize real-time changes of chemical process performance. The algorithm is divided into two parts, namely, data training and data testing.

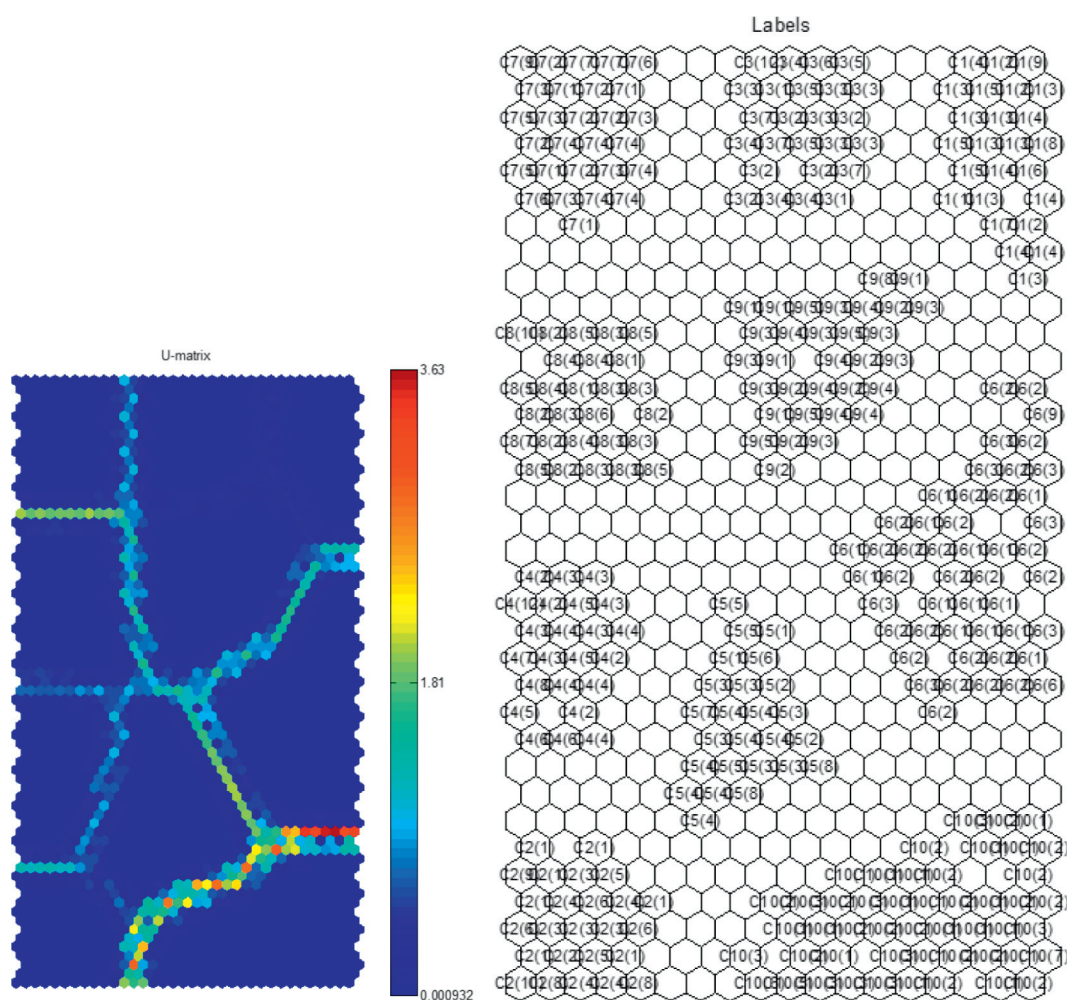
3.3.1. Data training

In the process of data training, the KPCA algorithm is used to extract the features of different performance data. Suppose there are N performance statuses, $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$, where $\mathbf{X}_i \in R^{n \times m}$, using the KPCA algorithm to reduce the dimensionality of the data and the data will be turned into $\mathbf{X}'_1, \mathbf{X}'_2, \dots, \mathbf{X}'_N$, where $\mathbf{X}'_i \in R^{n \times k}$. The processed data $\mathbf{X}'_1, \mathbf{X}'_2, \dots, \mathbf{X}'_N$ are used as the input data of SOM algorithm. Using Eq. (17) to obtain the number of neurons in output layer and adjusting the weight vector $\mathbf{w}$ by Eqs. (16), (18) and (19). Then mapping the input data into neurons and labeling them according to the categories. A two-dimensional grid chart is acquired to divide training data.

3.3.2. Data testing

After obtaining the well-trained mapping diagram of different data, we need to extract the information of the training data to recognize the current operation status. In the training process, *NMAP* derived from Eq. (14) have been achieved through KPCA algorithm for feature extraction. First, we need to determine what kind do the test data belong to, that is, the data classification, and then multiply the corresponding matrix.

In the classification process, the base case is introduced as a reference value. The data of base case $\mathbf{X}_{\text{base}}$ is imported, and its mean



(a) U-matrix diagram

(b) Mapping diagram

Fig. 7. Training diagrams of LTSA-SOM (PCs 4).

value $\bar{X}_{\text{base}}$ is computed. Similarly, for the training samples, the mean value $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_N$ is computed. The test data X_{test} is deducted from $\bar{X}_i$, $i = 1, 2, \dots, N$, the difference is divided by $\bar{X}_{\text{base}}$. The absolute value X_i^* is then obtained.

$$X_i^* = \text{abs}((X_{\text{test}} - \bar{X}_i) / \bar{X}_{\text{base}}), \quad i = 1, 2, \dots, N \quad (21)$$

where $X_i^* \in R^{1 \times m}$. In this way, the difference of data based on X_{base} can be calculated and, to some extent, the impact of different measurement units of data in the chemical process is reduced. The data classification index S is defined as follows:

$$S_i = \sum_{j=1}^m (x_{ij}^*)^2, \quad i = 1, 2, \dots, N \quad (22)$$

where x_{ij}^* is an element of X_i^* , $j = 1, 2, \dots, m$. If the test data belong to a certain category of train data, then the S index it corresponds to will be relatively small compared with other categories. Assuming that $S_{\min}$ is the minimum value of the S index and $S_{2\min}$ is the second small S index, if

$$\frac{S_{2\min}}{S_{\min}} > 10 \quad (23)$$

then from Eq. (15), we can get $X_{\text{test}}' = X_{\text{test}} \times \text{MAP}_{\min}$ and map the X_{test} into SOM diagram like data training process. $\text{MAP}_{\min}$ is the mapping matrix corresponding to $S_{\min}$. If

$$\frac{S_{2\min}}{S_{\min}} < 10 \quad (24)$$

then the test data do not belong to a certain category of train data, and other performance indexes provided in Section 2 need to be introduced to judge the performance of the chemical process. Hence, the entire flow of the training process is given below:

- If the test data come from the training data, namely, $\frac{S_{2\min}}{S_{\min}} > 10$, the test data belong to the training data with $S_{\min}$, else
- If the test data are beyond the range of constraints, it belongs to **case 10**, else
- If the flowrate change of test data is beyond $\pm 5\%$ in the range $8-16 \text{ h}^{-1}$, it belongs to **case 9**, else
- If composition variability of G is beyond $\pm 5\text{mol}\%$ in the range $8-16 \text{ h}^{-1}$, it belongs to **case 8**, else
- Judge the profit range of the test data, end.

The results of Simulink are shown in Section 4.

4. Experimental Results and Analysis

In this section, the proposed performance assessment method based on KPCA-SOM will be applied to the TE process. To demonstrate the advantages of the KPCA-SOM method, SOM, PCA-SOM, and LTSA-SOM methods are used as comparisons.

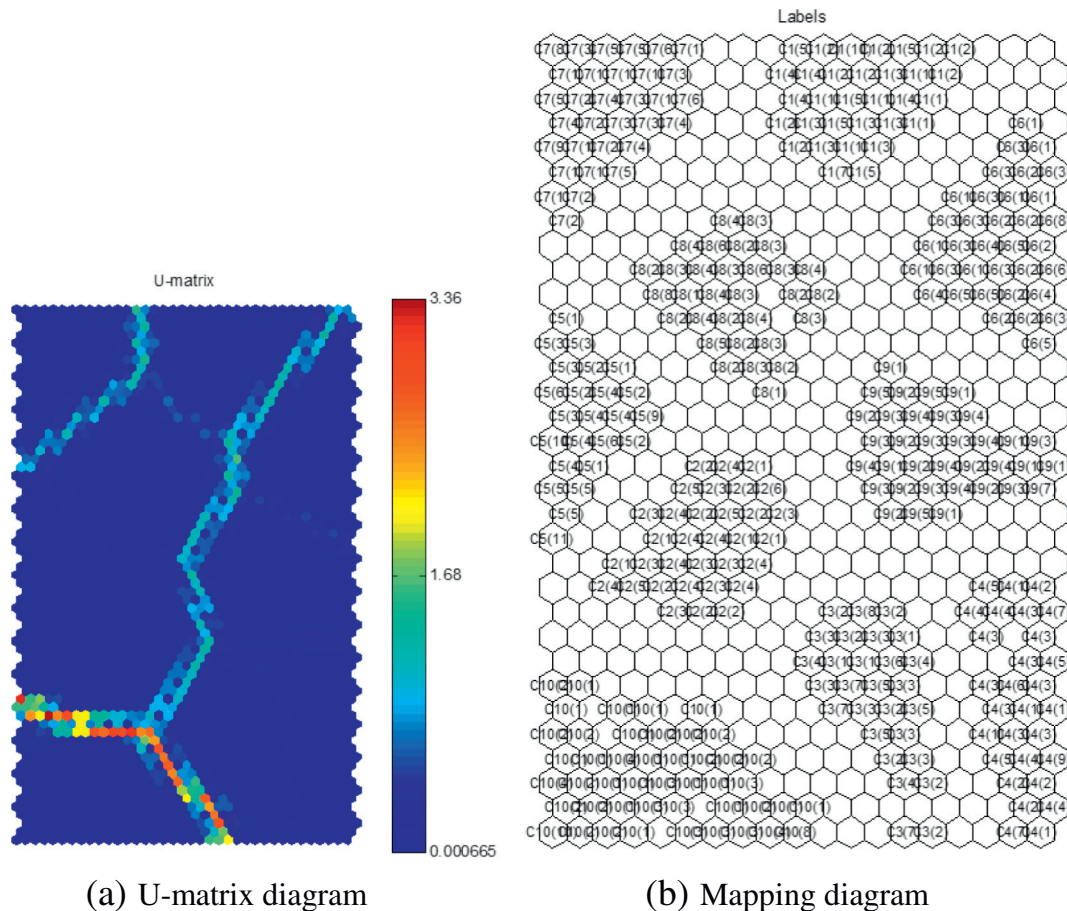


Fig. 8. Training diagrams of LTSA-SOM (PCs 4).

4.1. Defect of simple SOM

To examine the capability of the proposed method, the test data are selected in two ways. First, we choose the data from the same sets of the training data as the test data (way 1). In addition, given that several different statuses may be encountered in the process, it is necessary to test the performance of some unknown statuses (way 2). Hence, some data with different profits are chosen as the test data for case 1 to case 7. The out of regulated range data are constraints, and the variability of flowrate and G exceeding 5% in set time are of concern.

The SOM results are displayed. Fig. 2 shows the training result of the SOM. The U-matrix represents the average distance between the weight vector of a neuron and the weight vector of its neighboring neurons. SOMs are visualized through different colors. Typically, dark colors represent small distances and light colors represent larger distances. Areas with low values in the U-matrix form clusters, whereas high values of the U-matrix represent cluster boundaries. Thus, a dark region on the U-matrix can be regarded as a cluster with an optical boundary. The mapping diagram corresponds with the U-matrix, and C1–C10 represent cases 1 to case 10. Numbers in parentheses indicate the amount of training data mapped to the neuron. As shown in Fig. 2, the boundaries of 10 sets of training data in the U-matrix are distinctly differentiated. We can obtain a well-trained diagram of 10 sets of data based on SOM algorithm. Fig. 3 shows the test diagram of case 1. Fig. 3 (a) shows the test data selected in way 1. Given that the test data of way 1 comes from the same data set of training data, we can obtain a well-tested diagram for these test data, such as Fig. 3(a). However, the

test data of way 2 cannot obtain a good test diagram using a simple SOM method. In the training section, the profit of the training data for case 1 is 5235 USD·h⁻¹, and in the test section, the profit of test data is 5217 USD·h⁻¹, which belongs to the range of profit ≥ 5200 . However, in Fig. 3(b), the test data of case 1 surround the training region of case 2 instead of mapping it into the region of case 1. From the mapping results of simple SOM, it can be concluded that the SOM method is capable of obtaining good training diagram and monitoring known statuses but cannot achieve good results of unknown statuses.

4.2. Some existing methods combined with SOM

Considering the disadvantages of SOM, the PCA-SOM is first used to improve the performance of simple SOM. The number of principal components is four. However, as shown in Fig. 4, the training results of PCA-SOM appear to be not good enough. The change of the number of principal components to 10 and the training results are shown in Fig. 5. On the basis of the training diagrams, the test data of case 1 selected in way 2 are tested, as shown in Fig. 6. It can be assumed that it is correctly mapped because the test data of case 1 are mapped into the region of case 1. On the contrary, when the rate of contribution is 99% of PCA algorithm, the number of principal components is 4, whereas there is no reason to select number 10.

Tian [34] used LTSA-SOM to diagnose the faults of p-xylene oxidation process and proved it is superior to PCA-SOM. Local tangent space alignment (LTSA) was proposed by Zhang and Zha [35]. LTSA is

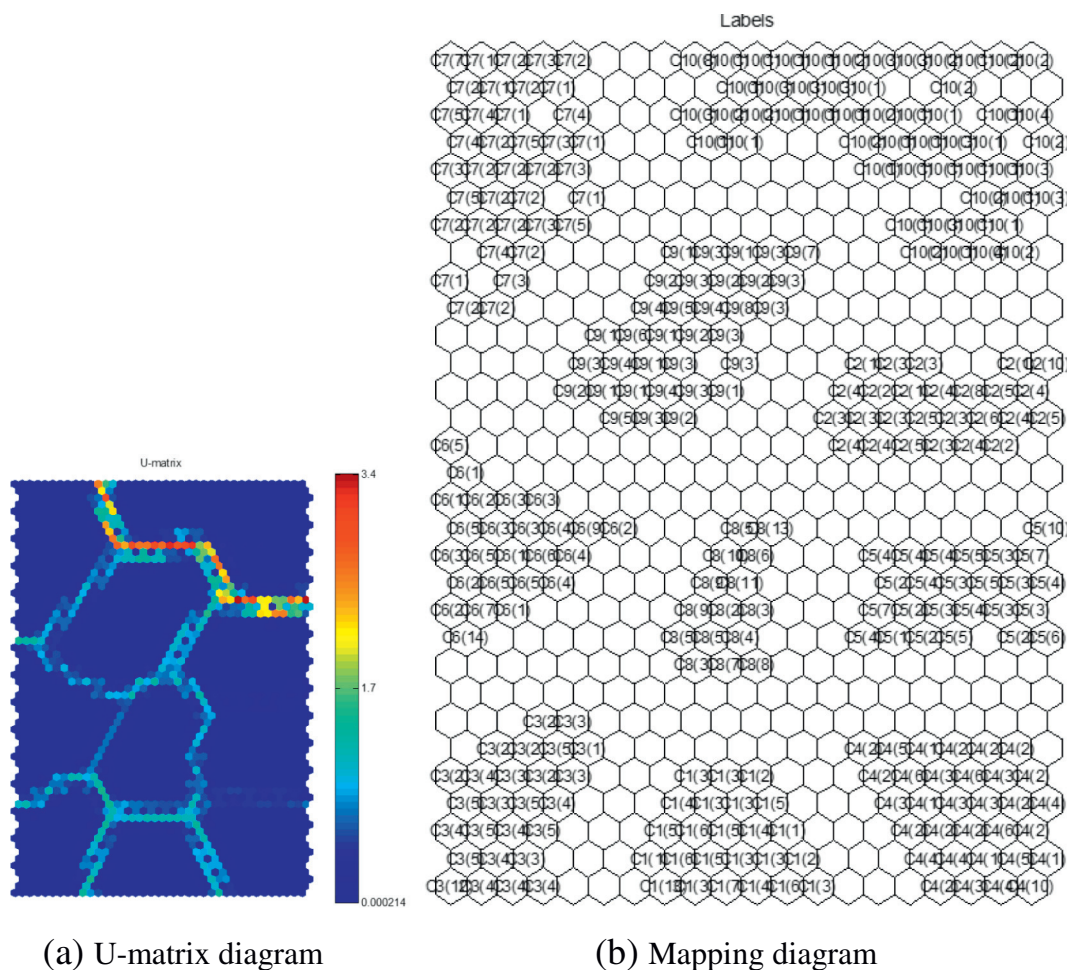
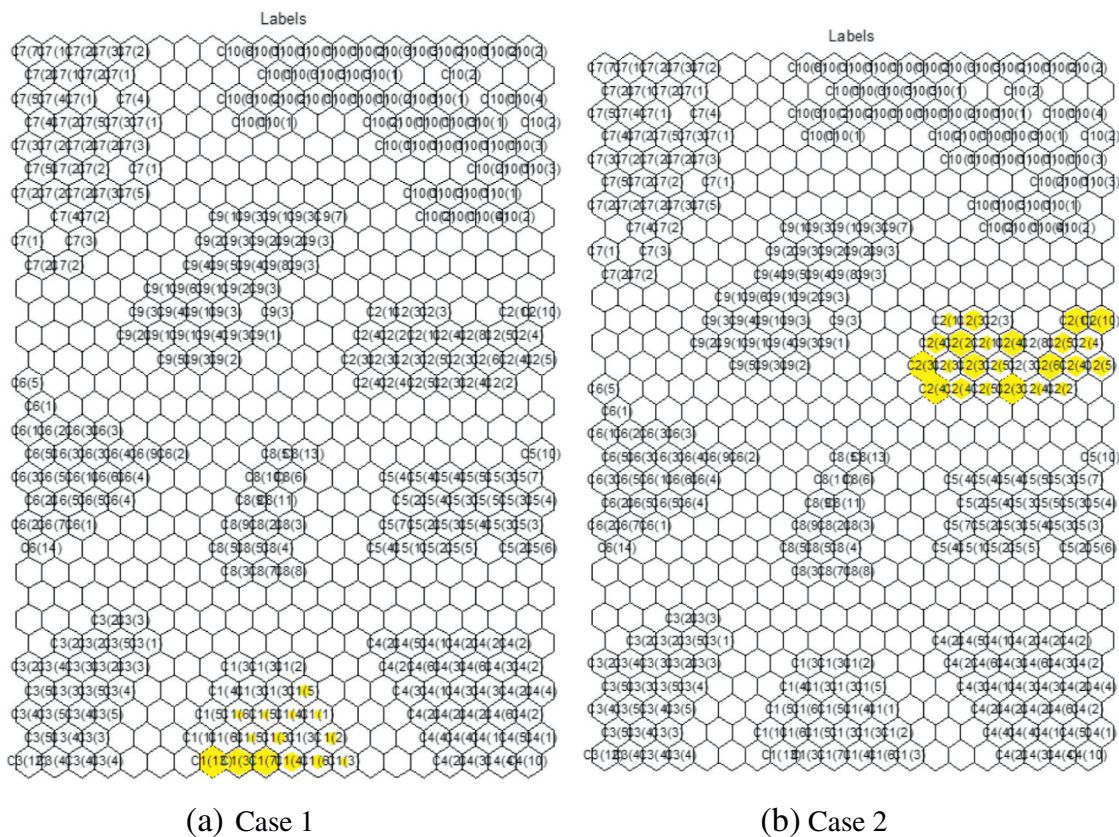
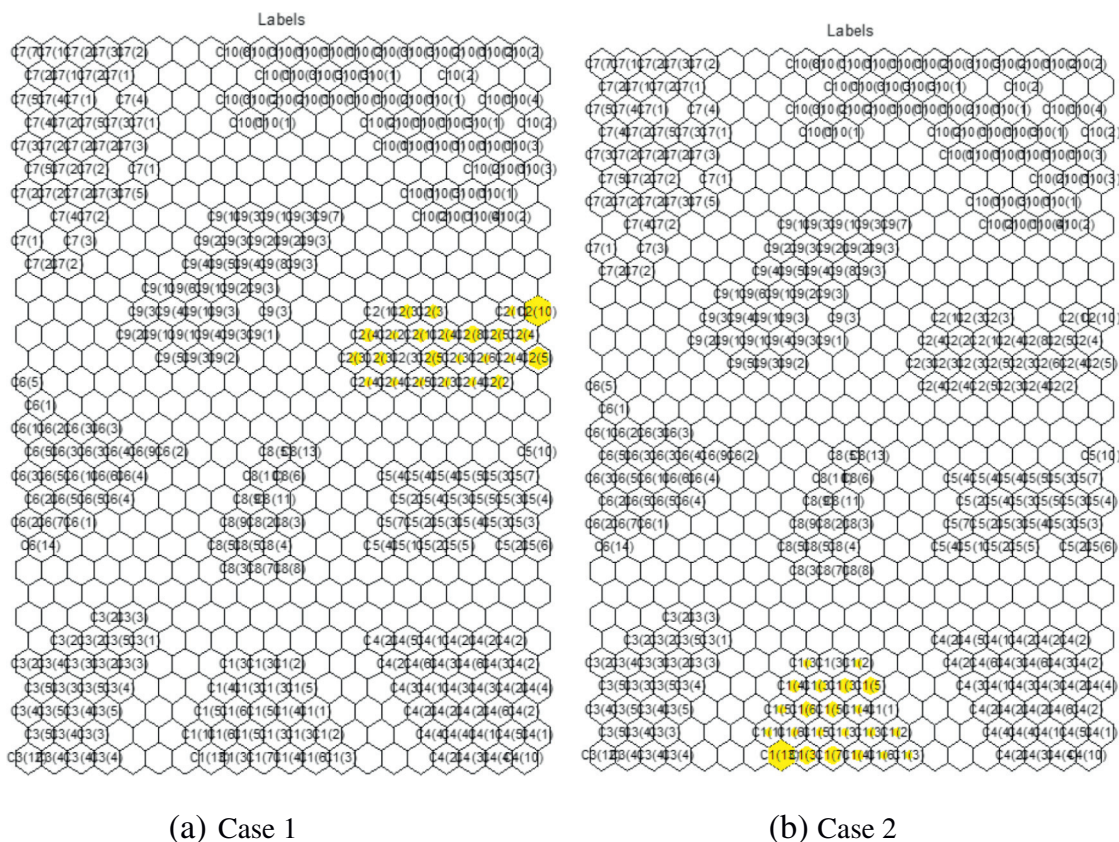


Fig. 9. Training diagrams of KPCA-SOM.



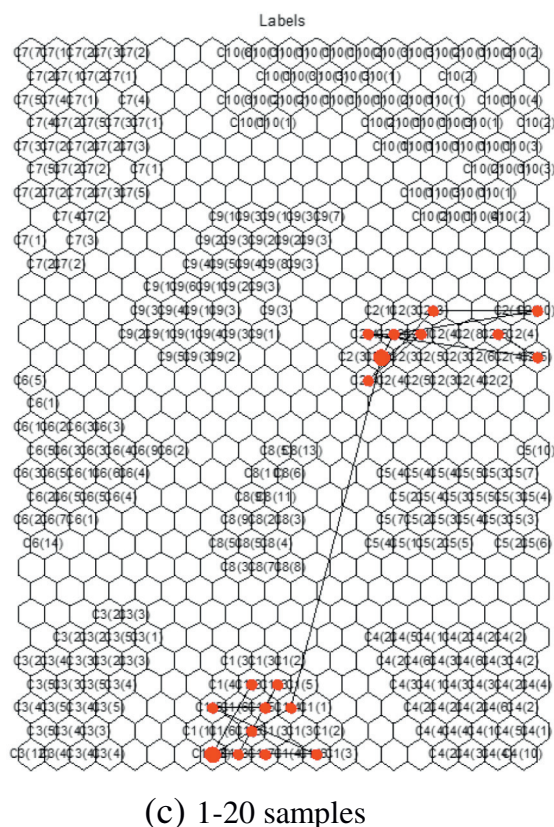
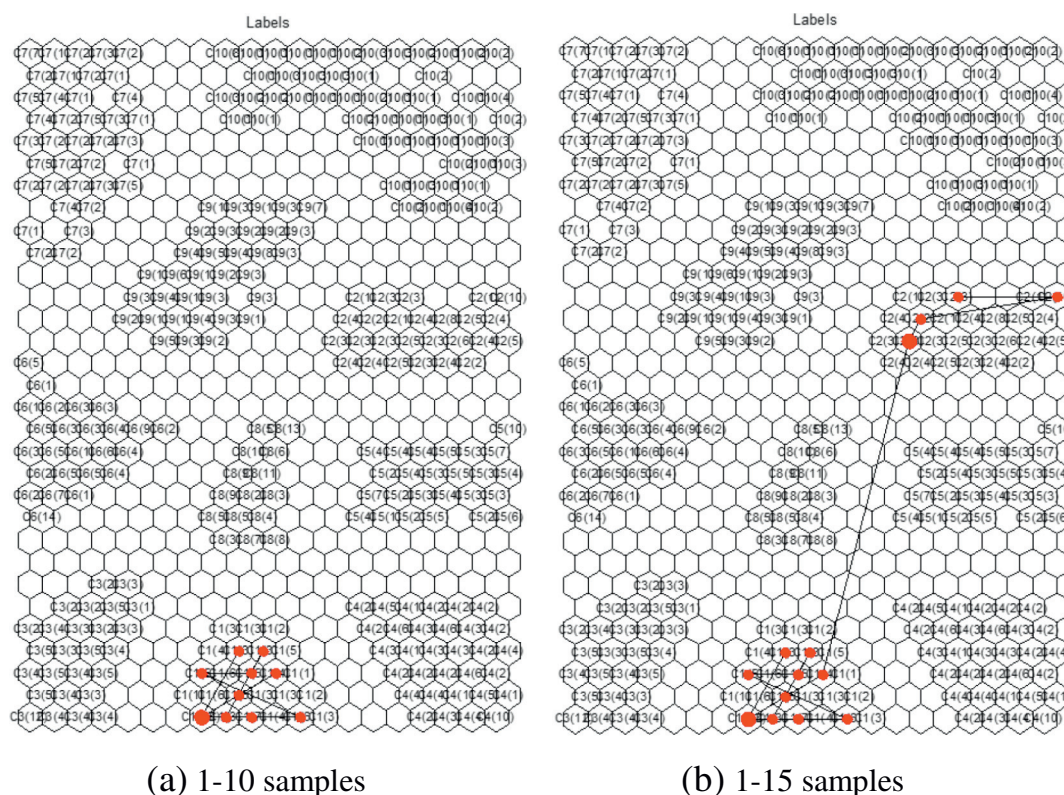


Fig. 12. The trajectory of performance variation.

a neighborhood method based on feature extraction and dimension reduction. Hence, the LTSA-SOM method is also considered. The number of principal components is also 4. Fig. 7 shows that the U-matrix diagram can clearly distinguish the different cases. However, the LTSA

method has a fatal shortcoming, *i.e.*, its mapping results are not invariant, which means periodically it will yield good and bad results. Furthermore, the defect is determined by its inherent properties that it will find k local nearest neighbors during the training process, and the selections

of neighbors are random. Fig. 8 also displays the results of LTSA-SOM, and the numbers of neighbors and principal components are the same as those of Fig. 7. However, the results obviously are not good enough, and case 3 is even divided into two regions. There are no standards in selecting the number of principal components, although increasing of the principal components will improve the performance of LTSA-SOM method. Therefore, it is faced with the same problem of PCA-SOM. In addition, Chen and Yan also combined FDA [36] and CCA [37] with SOM for fault diagnosis in chemical process, but it seems that these methods are even inferior to LTSA-SOM.

Therefore, we can recognize that these existing methods combined SOM with feature extraction methods cannot obtain a good result in performance assessment based on data.

4.3. The off-line and on-line monitoring of KPCA-SOM

The simple SOM can obtain good training results and monitor known statuses but cannot acquire good results of unknown statuses. The PCA-SOM and LTSA-SOM can attain good test results of unknown statuses; however, the training results are affected by the number of principal components. The KPCA-SOM algorithm is then proposed to acquire good training and test results at the same time. To measure the capability of KPCA-SOM, it is used to monitor the off-line and on-line data of the TE process.

First, the off-line data are tested. When the contribution rate is 99% of KPCA, the number of principal components is 4 as well. As shown in Fig. 9, 10 sets of data are distinctly distinguished. Fig. 10 shows the test data of cases 1 and 2 selected in way 1. The test data are found to be uniformly distributed in the regions of cases 1 and 2. The test data selected in way 2 are shown in Fig. 11. The test data of case 1 is introduced above. The profit of the training data for case 2 is 5186 USD·h⁻¹ in the training and test sections. The profit of test data is 5183 USD·h⁻¹, which belongs to the range of 5150–5200 USD·h⁻¹. Similarly, they are correctly mapped into the regions of cases 1 and 2. Therefore, the KPCA-SOM algorithm cannot only gain well-mapped training diagrams but also obtain the well-mapped test diagrams in both way 1 and way 2.

The on-line monitoring capability is then measured. Fig. 12 shows the trajectory of performance variation from case 1 to case 2 to realize the on-line monitoring of performance changes. Case 1 represents the optimal performance state of the process which gains maximum benefits, and case 2 represents the suboptimal status. The process is in an optimal status at the beginning, and as time goes on, the performance of the process decreases to suboptimal status. Therefore, the KPCA-SOM method can be used to monitor the real-time performance change trajectory of the process.

5. Conclusions

This paper proposed a hierarchy framework to assess the off-line and on-line performance of the process industry. Hard safety indexes and soft safety indexes are used to ensure the safety of the entire process. Profit is also a criterion to evaluate the performance when the process is in normal condition. The KPCA-SOM algorithm is used to visualize the trajectory of performance variation to realize the real-time monitoring of the process. KPCA can extract the features and reduce the dimensions of data, and SOM maps the training data and test data into a two-dimensional diagram. The results of simulation show that the combination of hierarchy framework and KPCA-SOM algorithm is sufficient to obtain good off-line monitoring effects and to monitor the real-time performance change trajectory of the process.

This paper does not provide the explanations for the variations of the performance; however, we can use a multiple mapping method to find the cause of the performance variations. In the TE process, there are a total of five constraints, and they are all mapped into case 10. Therefore, we can rebuild a training map consisting of five

training data of constraints, such as the training diagram of case 1 to case 10 if the test data is mapped into case 10. At that point, the test data on the training map is retested to find which constraint exceeds the boundary. In this way, we can find the reason leading to the performance variation and take some actions to maintain the optimal status of the process.

Appendix A

A1. the calculation of $c_2(t)$

The TE process has four feed streams. The component cost, molecular weight and are respectively given in Table 9, Table 2 and Table 9 of the paper [16]. Suppose that the mole fractions of A, D, and E in streams 1, 2, and 3 are 100% and the mole fractions of A and C in stream 4 are 48.5% and 51% respectively. The cost of B is ignored.

a. The cost of stream 1, namely A feed

The mole flow of A feed:

$$y_1(t) = 10^3 y_1(t) = 10^3 y_1(t) = \frac{10^3 y_1(t)}{22.4} \text{ kmol} \cdot \text{h}^{-1} \quad (\text{A1})$$

The cost of A feed:

$$\text{cost}_A = 2.206 \text{ USD} \times \frac{10^3 y_1(t)}{22.4} \approx 98.48 y_1(t) \text{ USD} \cdot \text{h}^{-1} \quad (\text{A2})$$

b. The cost of stream 2, namely D feed

The mass flow of D feed: $y_2(t) \text{ kg} \cdot \text{h}^{-1}$.

The cost of D feed:

$$\text{cost}_D = 22.06 \text{ USD} \times y_2(t) = \frac{22.06 \text{ USD}}{32} \times y_2(t) \approx 0.689 y_2(t) \text{ USD} \cdot \text{h}^{-1} \quad (\text{A3})$$

c. The cost of stream 3 namely E feed

The mass flow of E feed: $y_3(t) \text{ kg} \cdot \text{h}^{-1}$.

The cost of E feed:

$$\text{cost}_E = 14.56 \text{ USD} \times y_3(t) = \frac{14.56 \text{ USD}}{46} \times y_3(t) \approx 0.3165 y_3(t) \text{ USD} \cdot \text{h}^{-1} \quad (\text{A4})$$

d. The cost of stream 4, namely A and C feed

The mole flow of A and E feed:

$$y_4(t) = 10^6 y_4(t) = \frac{10^6 y_4(t)}{22.4} \text{ mol} \cdot \text{h}^{-1} \quad (\text{A5})$$

the mole flow of A: $0.485 \times \frac{10^6 y_4(t)}{22.4} \text{ mol} \cdot \text{h}^{-1}$.

the mole flow of C: $0.51 \times \frac{10^6 y_4(t)}{22.4} \text{ mol} \cdot \text{h}^{-1}$.

The cost of E feed:

$$\text{cost}_A = 2 \times 0.485 \times \frac{y_4 10^3(t)}{22.4} \times \frac{2.206 \text{ USD}}{2} \approx 47.76 y_4(t) \text{ USD} \cdot \text{h}^{-1} \quad (\text{A6})$$

$$\text{cost}_C = 28 \times 0.51 \times \frac{10^3 y_4(t)}{22.4} \times \frac{6.177 \text{ USD}}{28} \approx 140.64 y_4(t) \text{ USD} \cdot \text{h}^{-1} \quad (\text{A7})$$

$$cost_4 = cost_A + cost_C = 188.4y_4(t) \text{ USD} \cdot \text{h}^{-1} \quad (\text{A8})$$

From the calculation above

$$c_2(t) = 98.48y_1(t) + 0.689y_2(t) + 0.3165y_3(t) + 188.4y_4(t) \text{ USD} \cdot \text{h}^{-1} \quad (\text{A9})$$

A2. the calculation of $profit(t)$

Table 9 [16] gives the cost of components and the cost of G is the cost of the sum of A, C, and D, the cost of H is the sum of A, C, and E. Since the paper does not provide the profit of G and H, it is assumed that the profits of G and H are the twice of the costs of G and H. That is, the profit of G is $60.88 \text{ USD} \cdot \text{kg}^{-1} \cdot \text{mol}^{-1}$ and the profit of H is $45.88 \text{ USD} \cdot \text{kg}^{-1} \cdot \text{mol}^{-1}$.

In 1 mol production:

the quality of G: $\frac{y_{40}(t)}{100} \times 62 \text{ g}$

the quality of H: $\frac{y_{41}(t)}{100} \times 76 \text{ g}$

the quality of D: $\frac{y_{37}(t)}{100} \times 32 \text{ g}$

the quality of E: $\frac{y_{38}(t)}{100} \times 46 \text{ g}$

the quality of F: $\frac{y_{39}(t)}{100} \times 48 \text{ g}$

Assuming that when the components G, H, D, E, and F are mixed, the volume of production will not change. According to the liquid density in Table 2 [16], the volume of 1 mol production is

$$V = \frac{m}{\rho} = \frac{y_{40}(t)}{100} \times 62/\rho_G + \frac{y_{41}(t)}{100} \times 76/\rho_H + \frac{y_{37}(t)}{100} \times 32/\rho_D + \frac{y_{38}(t)}{100} \times 46/\rho_E + \frac{y_{39}(t)}{100} \times 48/\rho_F \quad (\text{A10})$$

Then the volume of 1 mol production is V and the production volume in time t is $y_{17}(t)\text{m}^3 \cdot \text{h}^{-1}$. Then the profits of G and H are

$$profit_G = \frac{y_{17}(t)}{V} \times \frac{y_{40}(t)}{100} \text{ mol} \times 60.88 \text{ USD} \cdot (\text{kg} \cdot \text{mol})^{-1} \quad (\text{A11})$$

$$profit_H = \frac{y_{17}(t)}{V} \times \frac{y_{41}(t)}{100} \text{ mol} \times 45.88 \text{ USD} \cdot (\text{kg} \cdot \text{mol})^{-1} \quad (\text{A12})$$

References

- [1] H.A. Alhashimi, C.B. Aktas, Life cycle environmental and economic performance of biochar compared with activated carbon: A meta-analysis, *Resour. Conserv. Recycl.* 118 (2017) 13–26.
- [2] Y. Arushanyan, E. Ekener, Åsa Moberg, Sustainability assessment framework for scenarios – SAFS, *Environ. Impact Assess. Rev.* 63 (2017) 23–34.
- [3] J. Shao, M. Taisch, M.O. Mier, Influencing factors to facilitate sustainable consumption: From the experts' viewpoints, *J. Clean. Prod.* 142 (6) (2017) 203–216.
- [4] M.P. Dočekalová, A. Kocmanová, Composite indicator for measuring corporate sustainability, *Ecol. Indic.* 61 (2016) 612–623.
- [5] T.L. Saaty, Group decision making and the AHP, *Analytic Hierarchy Process Application & Studies* 1989, pp. 59–67.
- [6] R.K. Singh, H.R. Murty, S.K. Gupta, A.K. Dikshit, Development of composite sustainability performance index for steel industry, *Ecol. Indic.* 7 (3) (2007) 565–588.
- [7] A.G. Frank, N.D. Molle, W. Gerstlberger, J.A.B. Bernardi, D.C. Pedrini, An integrative environmental performance index for benchmarking in oil and gas industry, *J. Clean. Prod.* 133 (2016) 1190–1203.
- [8] Y. Long, J. Pan, S. Farooq, H. Boer, A sustainability assessment system for Chinese iron and steel firms, *J. Clean. Prod.* 125 (2016) 133–144.
- [9] N.D. Pour, B. Huang, S.L. Shah, Performance assessment of advanced supervisory-regulatory control systems with subspace LQG benchmark, *Automatica* 46 (8) (2010) 1363–1368.
- [10] C. Zhao, Y. Zhao, H. Su, B. Huang, Economic performance assessment of advanced process control with LQG benchmarking, *J. Process Control* 19 (4) (2009) 557–569.
- [11] S. Wei, J. Cheng, Y. Wang, Data-driven two-dimensional LQG benchmark based performance assessment for batch processes under ILC, *IFAC PapersOnline* 48 (8) (2015) 291–296.
- [12] X. Tian, G. Chen, S. Chen, A data-based approach for multivariate model predictive control performance monitoring, *Neurocomputing* 74 (4) (2011) 588–597.
- [13] F. Xu, B. Huang, S. Akande, Performance assessment of model predictive control for variability and constraint tuning, *Ind. Eng. Chem. Res.* 46 (4) (2007) 1208–1219.
- [14] N.A. And, B. Huang, E.C. Tamayo, Assessing model prediction control (MPC) performance. 1. Probabilistic approach for constraint analysis, *Ind. Eng. Chem. Res.* 46 (24) (2007) 8101–8111.
- [15] N.A. And, B. Huang, E.C. Tamayo, Assessing model prediction control (MPC) performance. 2. Bayesian approach for constraint tuning, *Ind. Eng. Chem. Res.* 46 (24) (2007) 8112–8119.
- [16] J.J. Downs, E.F. Vogel, A plant-wide industrial process control problem, *Comput. Chem. Eng.* 17 (3) (1993) 245–255.
- [17] X. Gao, J. Hou, An improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastman process, *Neurocomputing* 174 (2016) 906–911.
- [18] H. Chen, P. Tiño, X. Yao, Cognitive fault diagnosis in Tennessee Eastman Process using learning in the model space, *Comput. Chem. Eng.* 67 (3) (2014) 33–42.
- [19] M. Golshan, M.R. Pishvaie, R. Bozorgmehry Boozarjomehry, Stochastic and global real time optimization of Tennessee Eastman challenge problem, *Eng. Appl. Artif. Intell.* 21 (2) (2008) 215–228.
- [20] M. Golshan, R.B. Boozarjomehry, M.R. Pishvaie, A new approach to real time optimization of the Tennessee Eastman challenge problem, *Chem. Eng. J.* 112 (1–3) (2005) 33–44.
- [21] N.L. Ricker, Optimal steady-state operation of the Tennessee Eastman challenge process, *Comput. Chem. Eng.* 19 (9) (1995) 949–959.
- [22] N.L. Ricker, Decentralized control of the Tennessee Eastman Challenge Process, *J. Process Control* 6 (4) (1996) 205–221.
- [23] I.T. Jolliffe, *Principal Component Analysis*, Springer, Berlin, 1986.
- [24] K.I. Diamantaras, S.Y. Kung, *Principal Component Neural Networks*, Wiley, New York, 1996.
- [25] Y. Xu, D. Zhang, F. Song, J.Y. Yang, Z. Jing, A method for speeding up feature extraction based on KPCA, *Neurocomputing* 70 (4–6) (2007) 1056–1061.
- [26] C.Y. Cheng, C.C. Hsu, M.C. Chen, Adaptive kernel principal component analysis (KPCA) for monitoring small disturbances of nonlinear processes, *Ind. Eng. Chem. Res.* 49 (5) (2010) 2254–2262.
- [27] I. Elaissi, I. Jaffel, O. Taouali, H. Messaoud, Online prediction model based on the SVD-KPCA method, *ISA Trans.* 52 (1) (2013) 96–104.
- [28] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, Z. Jin, KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (2) (2005) 230–244.
- [29] J. Ni, C. Zhang, S.X. Yang, An adaptive approach based on KPCA and SVM for real-time fault diagnosis of HVCBs, *IEEE Trans. Power Delivery* 26 (3) (2011) 1960–1971.
- [30] T. Kohonen, The self-organizing map, *Proc. IEEE* 78 (9) (1990) 1464–1480.
- [31] Y.S. Ng, R. Srinivasan, Multivariate temporal data analysis using self-organizing maps. 2. Monitoring and diagnosis of multistate operations, *Ind. Eng. Chem. Res.* 47 (20) (2008) 7758–7771.
- [32] Y.S. Ng, R. Srinivasan, Multivariate temporal data analysis using self-organizing maps. 1. Training methodology for effective visualization of multistate operations, *Ind. Eng. Chem. Res.* 47 (20) (2008) 7744–7757.
- [33] B. Lamrini, E.K. Lakhal, M.V.L. Lann, L. Wehenkel, Data validation and missing data reconstruction using self-organizing map for water treatment, *Neural Comput. & Applic.* 20 (4) (2011) 575–588.
- [34] Y. Tian, W. Du, F. Qian, High dimension feature extraction based visualized SOM fault diagnosis method and its application in p-xylene oxidation process, *Chin. J. Chem. Eng.* 23 (9) (2015) 1509–1517.
- [35] Z.Y. Zhang, H.Y. Zha, Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, *J. Shanghai Univ.* 8 (4) (2004) 406–424.
- [36] X. Chen, X. Yan, Using improved self-organizing map for fault diagnosis in chemical industry process, *Chem. Eng. Res. Des. Trans. Inst.* 90 (12) (2012) 2262–2277.
- [37] X. Chen, X. Yan, Fault diagnosis in chemical process based on self-organizing map integrated with fisher discriminant analysis, *Chin. J. Chem. Eng.* 20 (4) (2013) 382–387.