Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

On-line Fault Detection and Classification for a Compressor Process in the Oxygen Plant

Jialin Liu* and Ding-Sou Chen**

** Department of Information Management, Fortune Institute of Technology, 1-10, Nwongchang Rd., Neighborhood 28, Lyouciyou Village, Daliao Township, Kaohsiung Country, Taiwan, Republic of China (Tel: 886-7-7889888 ext. 6122; e-mail: jialin@center.fotech.edu.tw).*
***Chemical Process & Water Treatment Section, New Materials Research & Development Department, China Steel Corporation, 1, Chung Kang Rd., Hsiao Kang, Kaohsiung, Taiwan, Republic of China (e-mail: t6410@mail.csc.com.tw)*

Abstract: In this paper, a data-driven model is proposed for on-line monitoring a process with high-dimensional variables, outliers, and time-varying characteristics. In this research, principal component analysis (PCA) is used to eliminate collinearity between process variables. After that, fuzzy rules are generated by using the compressed data and an outlier rejection clustering algorithm, named distanced-based fuzzy c-means (DFCM), from which a feasible solution can be obtained to reflect the actual data gatherings. When new event emerge, the data are collected for next model update. An adaptive PCA algorithm is utilized to accommodate the new event data without recalculating the trained data. The known event rules can be transferred to the new PCA subspace by rotating and shifting coordinates of the subspace. Therefore, only new event data need to be clustered on the new subspace. The proposed approach has been applied to monitor a compressor process of the steel plant. Results show the challenges of process monitoring can be effectively dealt with.

1. INTRODUCTION

China Steel Corporation (CSC) is an integrated steel maker, which produces steel form iron ore. In the production steps, blast furnace and oxygen converter are the most important processes, in which the former extracts iron from iron ore by reacting iron oxide with carbon monoxide, and the latter reduce impurities of iron by blowing oxygen through the metal in the converter. The both ones are oxygen-intensive processes.

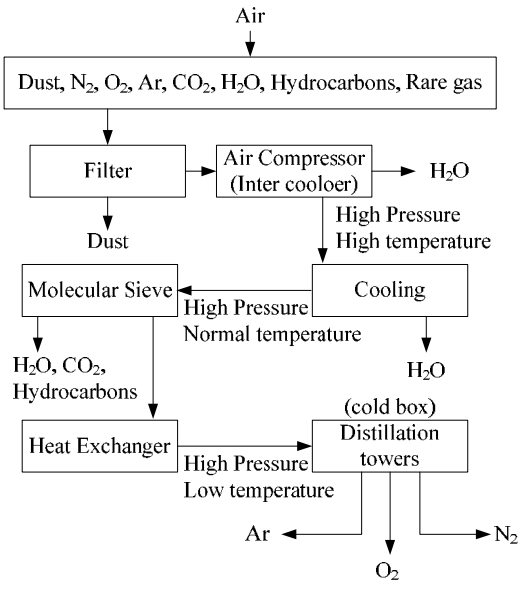

Fig. 1. Brief flow diagram of the oxygen plant

In order to supply high purity of oxygen, CSC runs several stand-alone oxygen plants. Eventually, it is an air separation process in the oxygen plant. The process input is air, coming from atmosphere, and the products are oxygen, nitrogen and argon. Fig. 1 shows a brief flow diagram of an oxygen plant, in which the process can be roughly split into two steps. Before entering distillation towers (cold box), air purification has to be proceeded in order to remove ingredients, such as, dust, moisture, CO_2, hydrocarbons, and so on. During the purification procedures, air compressor is one of the most important and energy consuming units. If the compressor discharged pressure cannot meet the specification, the air is hardly liquefied. Furthermore, it results in the air cannot be separated effectively in the cold box. In addition, production quantity of the plant directly depends on the air quantity that the compressor can handle with. Therefore, monitoring the compressor operation is a crucial task for the oxygen plant.

2. PROPOSED APPROACH

The proposed approach consists of three phases, including (1) off-line modelling phase, (2) on-line monitoring and updating phase, and (3) off-line updating phase, to monitor a process with collinearity, outliers, and time-varying characteristics. In the off-line modelling phase, a distance-based FCM (DFCM) algorithm has been applied to the score vectors instead of raw data. This approach is not only capable of reducing variable dimensions effectively, but also gradually discarding outliers to reach a feasible clustering result. In the on-line monitoring and updating phase, the statistic Q and T^2 of on-line data are examined with their control limits. If any one of the statistics is out of its control limits, alarms should be triggered and the data are stored for next off-line model

update. On the other hand, the Mahalanobis distances of score vectors to each cluster centers are evaluated with the corresponding cluster boundary distance when the on-line data belong to the PCA subspace. If any one of the Mahalanobis distances is less than the corresponding cluster boundary, the data point is part of known event groups, and then it can be classified into known groups according to its membership values whereas the cluster parameters also are updated. Otherwise, the data point is an outlier, even though it is within the PCA subspace, alarms should be triggered and the outlier is collected for next model update. In the off-line updating phase, the PCA subspace is adapted by using new event data and a recursive PCA algorithm in order to account for known and new events without recalculating the trained data. After that, the cluster parameters on the previous subspace are transferred to the new one. A concept of rotating and shifting coordinates of subspace has been applied to transfer the cluster parameters. Therefore, only new event data need to be clustered on the new subspace.

2.1 Off-line Modelling

Consider the data matrix $\mathbf{W} \in R^{m \times n}$ with m rows of observations and n columns of variables. Each column is normalized to zero means and unit variances: $\mathbf{X} = \left(\mathbf{W} - \mathbf{1}\overline{\mathbf{W}}\right)\mathbf{S}^{-1}$ where $\overline{\mathbf{W}}$ is a mean vector, $\mathbf{1}$ is a column vector in which all elements are one, and $\mathbf{S}$ is a diagonal matrix of standard deviations. The eigenvectors ($\mathbf{P}$) of the covariance matrix can be obtained from the normalized dataset. The score vectors are the projection of the data matrix $\mathbf{X}$ to each eigenvector. The data matrix $\mathbf{X}$ can be decomposed as:

$$\mathbf{X} = \sum_{i=1}^{k} \mathbf{t}_i \mathbf{p}_i^{\mathrm{T}} + \sum_{i=k+1}^{n} \mathbf{t}_i \mathbf{p}_i^{\mathrm{T}} \tag{1}$$

The first term of right side of above equation is the systematic part, which is described by the PCA subspace with the first k terms of eigenvectors, and the second term is remainder of $\mathbf{X}$ that is orthogonal to the PCA subspace. The statistics Q and T^2 can be measured to identify whether the data point belongs to the PCA subspace; the details can be found in Jackson (1991).

The objective of fuzzy clustering is to partition the dataset $\mathbf{T}$, which is composed of the score vectors, into c clusters with vague boundaries, which are determined by the fuzzifier q. The fuzzy c-means (FCM) algorithm is as follows:

1. Randomly initialize the degrees of membership following the constraints:
$$u_{ij} \in [0,1], \quad 1 \le i \le c, \quad 1 \le j \le m \tag{2}$$
$$\sum_{i=1}^{c} u_{ij} = 1, \quad 1 \le j \le m, \qquad 0 < \sum_{j=1}^{m} u_{ij} < m, \quad 1 \le i \le c$$
2. Compute the cluster centers and covariances:
$$\boldsymbol{\mu}_i^{(k)} = \sum_{j=1}^{m} \left[u_{ij}^{(k-1)}\right]^q \mathbf{t}_j \Big/ \sum_{j=1}^{m} \left[u_{ij}^{(k-1)}\right]^q, \quad i = 1...c \tag{3}$$
$$\boldsymbol{\Sigma}_i^{(k)} = \sum_{j=1}^{m} \left[u_{ij}^{(k-1)}\right]^q \left(\mathbf{t}_j - \boldsymbol{\mu}_i^{(k)}\right)^{\mathrm{T}} \left(\mathbf{t}_j - \boldsymbol{\mu}_i^{(k)}\right) \Big/ \sum_{j=1}^{m} \left[u_{ij}^{(k-1)}\right]^q \tag{4}$$
3. Compute the distances and update the degrees of membership:
$$D_{ij,\boldsymbol{\Sigma}_i}^{2(k)} = \left\|\boldsymbol{\Sigma}_i^{(k)}\right\|^{1/n} \left(\mathbf{t}_j - \boldsymbol{\mu}_i^{(k)}\right) \boldsymbol{\Sigma}_i^{-1(k)} \left(\mathbf{t}_j - \boldsymbol{\mu}_i^{(k)}\right)^{\mathrm{T}} \tag{5}$$
$$u_{ij}^{(k)} = 1 \Big/ \sum_{l=1}^{c} \left(D_{ij,\boldsymbol{\Sigma}_i}^{2(k)} \Big/ D_{lj,\boldsymbol{\Sigma}_l}^{2(k)}\right)^{2/(q-1)}, \quad i = 1...c, \quad j = 1...m \tag{6}$$
4. If the norm of membership changes is larger than the predefined tolerance (ε), i.e., $\sum_{j=1}^{m}\sum_{i=1}^{c} \left\|u_{ij}^{(k)} - u_{ij}^{(k-1)}\right\| \ge \varepsilon$, k=k+1 go back to step 2.

The FCM used the constraints that the memberships of an observation in all clusters must sum to 1, so that the outliers significantly affect the clustering results. In this paper, a boundary distance for the cluster has been used in order to identify outliers from the reference dataset. The boundary distance of the i^{th} cluster is defined as:

$$D_{b,i} \equiv D_{ave,i} + 3D_{std,i} \tag{7}$$
$$D_{ave,i} = \sum_{j=1}^{m} u_{ij}^q D_{ij,\boldsymbol{\Sigma}_i}^2 \Big/ \sum_{j=1}^{m} u_{ij}^q$$
$$D_{std,i} = \left(\sum_{j=1}^{m} u_{ij}^q \left(D_{ij,\boldsymbol{\Sigma}_i}^2 - D_{ave,i}\right)^2 \Big/ \left(\sum_{j=1}^{m} u_{ij}^q - 1\right)\right)^{0.5}$$

where $D_{ave,i}$ and $D_{std,i}$ respectively are the average and standard deviation of the Mahalanobis distances that the data points belong to the i^{th} cluster. Assuming that the data points belonging to the i^{th} cluster are randomly distributed and the distances of data points to the cluster center will follow a Gaussian distribution. Therefore, the boundary distance covers 99.87% data points. The distances of an outlier to each cluster center should be larger than the respective boundary distances. Therefore, before performing FCM iterations, the outliers should be discarded according to their distances. The proposed algorithm is as follows:

1. The initial clustering results of FCM are used to calculate the boundary distance for each cluster $D_{b,i}^{(k)}, \quad i = 1...c$. Set the number of retained observations equaling to the number of the reference data, i.e. $m_{ret}^{(k)} = m$, and $k = 0$.
2. When the observations are inliers, at least one of the Mahalanobis distances should be less than the corresponding boundary distances. Otherwise, the outliers are discarded.
3. Perform FCM iterations by using the retained observations, in which the numbers of the retained data points are $m_{ret}^{(k+1)}$. After converging, the boundary distances, $D_{b,i}^{(k+1)}, \quad i = 1...c$, can be found.
4. If the changes of the boundary distances are larger than the predefined tolerance (ε), $\sum_{i=1}^{c} \left\|D_{b,i}^{(k+1)} - D_{b,i}^{(k)}\right\| > \varepsilon$, k=k+1 go back to step 2.

2.2 On-line Model Updating

When on-line data belong to known event groups, the respective cluster parameters, including centers, covariances, and boundary distances, are updated to accommodate additional information. Marsili-Libelli and Müller (1996) enhanced the FCM with an adaptive capability. The centers and covariance matrices of clusters are updated as follows:

$$\boldsymbol{\mu}_i\Big|_{m+1} = \left(\sum_{j=1}^{m} u_{ij}^q \mathbf{t}_j + u_{i,m+1}^q \mathbf{t}_{m+1}\right) \Big/ \left(\sum_{j=1}^{m} u_{ij}^q + u_{i,m+1}^q\right) \tag{8}$$

$$\boldsymbol{\Sigma}_i\Big|_{m+1} = \frac{\sum_{j=1}^{m} u_{ij}^q \left(\mathbf{t}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \left(\mathbf{t}_j - \boldsymbol{\mu}_i\right) + u_{i,m+1}^q \left(\mathbf{t}_{m+1} - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \left(\mathbf{t}_{m+1} - \boldsymbol{\mu}_i\right)}{\sum_{j=1}^{m} u_{ij}^q + u_{i,m+1}^q}$$

where $u_{i,m+1}$ is the membership of the new data, which was computed from the unchanged prototypes. Since the $\sum_{j=1}^{m} u_{ij}^q \mathbf{t}_j$, $\sum_{j=1}^{m} u_{ij}^q \left(\mathbf{t}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \left(\mathbf{t}_j - \boldsymbol{\mu}_i\right)$, and $\sum_{j=1}^{m} u_{ij}^q$ can be recursively obtained, the computational requirement of the adaptation is moderate. In this paper, the boundary distances of the clusters also have been adapted:

$$D_{b,i}\Big|_{m+1} \equiv D_{ave,i}\Big|_{m+1} + 3 D_{std,i}\Big|_{m+1} \tag{9}$$

$$D_{ave,i}\Big|_{m+1} = \left(\sum_{j=1}^{m} u_{ij}^q D_{ave,i} + u_{i,m+1}^q D_{i,m+1}^2\right) \Big/ \left(\sum_{j=1}^{m} u_{ij}^q + u_{i,m+1}^q\right)$$

$$D_{std,i}\Big|_{m+1} = \left\{ \left[\begin{array}{l} \left(\sum_{j=1}^{m} u_{ij}^q - 1\right) D_{std,i}^2 \\ + \sum_{j=1}^{m} u_{ij}^q \left(D_{ave,i} - D_{ave,i}\Big|_{m+1}\right)^2 \\ + u_{i,m+1}^q \left(D_{i,m+1}^2 - D_{ave,i}\Big|_{m+1}\right)^2 \end{array} \right] \Big/ \left(\sum_{j=1}^{m} u_{ij}^q + u_{i,m+1}^q\right) \right\}^{0.5}$$

where $D_{i,m+1}^2$ is the Mahalanobis distance, which is the new data point to the i^{th} prototype; and the $D_{ave,i}\Big|_{m+1}$ and $D_{std,i}\Big|_{m+1}$ respectively are the updated average and standard deviation of the distances, which the data points belong to the i^{th} cluster.

2.3 Off-line Adapting Model

Assuming new dataset with m' rows of observations, denoted as $\mathbf{W}' \in R^{m' \times n}$. The mean vector ($\overline{\mathbf{W}}'$) and the diagonal matrix of standard deviations ($\mathbf{S}'$) have to be prepared for normalizing the dataset $\mathbf{X}' = \left(\mathbf{W}' - \mathbf{1}\overline{\mathbf{W}}'\right)\mathbf{S}'^{-1}$ with zero means and unit variances. The covariance matrix of the new dataset can be obtained from the normalized data matrix: $\boldsymbol{\Sigma}'_{m'} = \mathbf{X}'^{\mathrm{T}} \mathbf{X}' \big/ \left(m' - 1\right)$. The mean vector and the standard deviations of combining datasets can be derived as:

$$\overline{\mathbf{W}}^* = \frac{m}{m^*}\overline{\mathbf{W}} + \frac{m'}{m^*}\overline{\mathbf{W}}', \quad \mathbf{S}^* = diag\left[\sigma_1^* \quad \sigma_2^* \quad \cdots \quad \sigma_n^*\right] \tag{10}$$

$$\sigma_i^* = \sqrt{\frac{\left(m-1\right)\sigma_i^2 + m\,\overline{w}_i^2 + \left(m'-1\right)\sigma_i'^2 + m'\,\overline{w}_i'^2 - m^*\,\overline{w}_i^{*2}}{m^* - 1}} \tag{11}$$

where the m^* are the total numbers of observations in the combined dataset, i.e. $m^* = m + m'$. Based on the updated means and standard deviations, the covariance matrix of the combined dataset is written as:

$$\boldsymbol{\Sigma}_{m^*}^* = \mathbf{X}_{m^*}^{*\mathrm{T}} \mathbf{X}_{m^*}^* \big/ \left(m^* - 1\right) = \left(\mathbf{X}_m^{*\mathrm{T}} \mathbf{X}_m^* + \mathbf{X}_{m'}^{*\mathrm{T}} \mathbf{X}_{m'}^*\right) \big/ \left(m^* - 1\right) \tag{12}$$

where $\mathbf{X}_{m^*}^*$ is the combined data matrix $\mathbf{X}_{m^*}^* = \left[\mathbf{X}_m^{*\mathrm{T}} \quad \mathbf{X}_{m'}^{*\mathrm{T}}\right]^{\mathrm{T}}$. The $\mathbf{X}_m^{*\mathrm{T}} \mathbf{X}_m^*$ of the above equation can be obtained from the covariance matrix of the reference dataset.

$$\mathbf{X}_m^{*\mathrm{T}} \mathbf{X}_m^* = \left(m-1\right)\mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_m \mathbf{A} + m\mathbf{Z} \tag{13}$$

$$\mathbf{A} \equiv \mathbf{S}\mathbf{S}^{*-1}, \quad \mathbf{Z} \equiv \left(\mathbf{S}^{*-1}\right)^{\mathrm{T}} \Delta\overline{\mathbf{W}}^{\mathrm{T}} \Delta\overline{\mathbf{W}} \mathbf{S}^{*-1}, \quad \Delta\overline{\mathbf{W}} \equiv \overline{\mathbf{W}} - \overline{\mathbf{W}}^*$$

where $\boldsymbol{\Sigma}_m$ is the covariance matrix of the trained dataset. In the same way, the covariance matrix of the new dataset can be written based on the means and standard deviations of the combined dataset. The covariance of the combined dataset can be obtained from the previous covariance matrices.

$$\boldsymbol{\Sigma}_{m^*}^* = \left(m-1\right) \big/ \left(m^* - 1\right) \boldsymbol{\Sigma}_m^* + \left(m'-1\right) \big/ \left(m^* - 1\right) \boldsymbol{\Sigma}_{m'}^* \tag{14}$$

$$\boldsymbol{\Sigma}_m^* \equiv \mathbf{X}_m^{*\mathrm{T}} \mathbf{X}_m^* \big/ \left(m-1\right), \quad \boldsymbol{\Sigma}_{m'}^* \equiv \mathbf{X}_{m'}^{*\mathrm{T}} \mathbf{X}_{m'}^* \big/ \left(m'-1\right)$$

The singular value decomposition (SVD) is applied to the updated covariance matrix. The eigenvectors ($\mathbf{P}^*$) can be obtained to span the new PCA subspace.

$$\boldsymbol{\Sigma}_{m^*}^* = \mathbf{P}^{*\mathrm{T}} \boldsymbol{\Lambda}^* \mathbf{P}^* \tag{15}$$

where $\boldsymbol{\Lambda}^*$ is the diagonal matrix of the eigenvalues.

When the subspace has been adapted, the known cluster parameters can be transferred from the previous subspace, the details can be found in Liu (2004).

$$\boldsymbol{\mu}_i^* = \boldsymbol{\mu}_i \mathbf{C}_{k,k^*} + \mathbf{1}\Delta\overline{\mathbf{W}}\mathbf{S}^{*-1}\mathbf{P}_{k^*}^* \tag{16}$$

$$\boldsymbol{\Sigma}_i^* = \mathbf{C}_{k,k^*}^{\mathrm{T}} \boldsymbol{\Sigma}_i \mathbf{C}_{k,k^*} + \mathbf{C}_{n-k,k^*}^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_i \mathbf{C}_{n-k,k^*} \tag{17}$$

$$\tilde{\boldsymbol{\Sigma}}_i^* = \mathbf{C}_{k,n-k^*}^{\mathrm{T}} \boldsymbol{\Sigma}_i \mathbf{C}_{k,n-k^*} + \mathbf{C}_{n-k,n-k^*}^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_i \mathbf{C}_{n-k,n-k^*} \tag{18}$$

$$\mathbf{C}_{k,k^*} \equiv \mathbf{P}_k^{\mathrm{T}} \mathbf{S}\mathbf{S}^{*-1} \mathbf{P}_{k^*}^*, \quad \mathbf{C}_{n-k,k^*} \equiv \mathbf{P}_{n-k}^{\mathrm{T}} \mathbf{S}\mathbf{S}^{*-1} \mathbf{P}_{k^*}^*, \quad \mathbf{C}_{n-k,n-k^*} \equiv \mathbf{P}_{n-k}^{\mathrm{T}} \mathbf{S}\mathbf{S}^{*-1} \mathbf{P}_{n-k^*}^*$$

where $\mathbf{P}_{k^*}^*$ are the first k^* terms of eigenvectors, which span the new subspace, and $\Delta\overline{\mathbf{W}} \equiv \overline{\mathbf{W}} - \overline{\mathbf{W}}^*$. The covariance matrix of the $(k+1)^{\text{th}}$ to n^{th} term score vectors for the i^{th} cluster is defined as $\tilde{\boldsymbol{\Sigma}}_i \equiv \mathbf{T}_{i,n-k}^{\mathrm{T}} \mathbf{T}_{i,n-k} \big/ \left(m_i - 1\right)$.

The boundary distance of the i^{th} cluster is given by

$$D_{b,i} = \left\|\boldsymbol{\Sigma}_i\right\|^{1/k} \left(\mathbf{t}_b - \boldsymbol{\mu}_i\right) \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{t}_b - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \tag{19}$$

where $\mathbf{t}_b$ is an arbitrary data point on the previous subspace. The updated boundary distance of the i^{th} cluster can be derived as follows:

$$D_{b,i}^* = \left\|\boldsymbol{\Sigma}_i^*\right\|^{1/k^*} \left\|\boldsymbol{\Sigma}_i\right\|^{-1/k} D_{b,i} \tag{20}$$

where the $\boldsymbol{\Sigma}_i^{*-1} \approx \left(\mathbf{C}_{k,k^*}^{\mathrm{T}} \boldsymbol{\Sigma}_i \mathbf{C}_{k,k^*}\right)^{-1}$ are used.

3. ILLUSTRATIVE EXAMPLE

The compressor process is a 4-stage centrifugal compressor, equipped with an intercooler between stages to cool down the compressed air, as Fig. 2 shows. In order to reduce the shaft work, each stage should be maintained with a close compression ratio and as lowest inlet temperature as possible. In addition, the production quantity varies with the demand of the downstream processes, such as, blast furnaces and oxygen converters. The compressor's inlet flowrate needs to be adjusted as well. However, the relation among inlet flowrate, motor speed, and discharged pressure of the compressor should be confined in the form of curves, called performance map. The feasible operating conditions of the compressor should be maintained within two operating lines, surge line and choke line. The former represents the minimum volumetric flowrate and the latter represents the maximum volumetric flowrate on the performance map. Compressor surge is denoted as an uncontrollable process, causing strong fluctuations in pressure and cyclic abrupt reversal of the flow direction when the operating condition is below the surge line. It results in violent vibration, large noise, and excessive thermal stresses. Finally, an unrepairable damage may be occurred to bearings, seals, even to impellers and rotors.

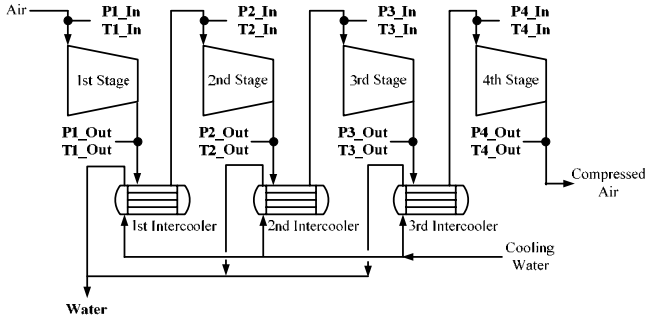

Fig. 2. Air compressor process flow diagram

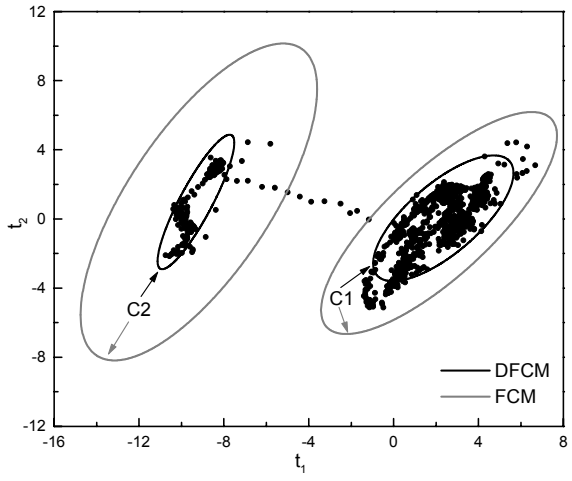

Fig. 3. Cluster results from FCM and DFCM.

In this paper, 29 measured variables were chosen according to the plant expert's advice. The training data were collected every 5 minutes for 4 days. There were 1152 observations in the training dataset. There were 1147 observations that could be explained by the PCA subspace, in which the captured variances were 88% with 2 PCs, within 99% confidence limits. The clustering results of the projection of the first two score vectors are shown in Fig. 3, in which the gray and black lines represent the boundary distances of clusters by using FCM and DFCM. It is obvious that the shapes of the FCM clusters were stretched due to the outliers. The situation would be significantly improved by using DFCM.

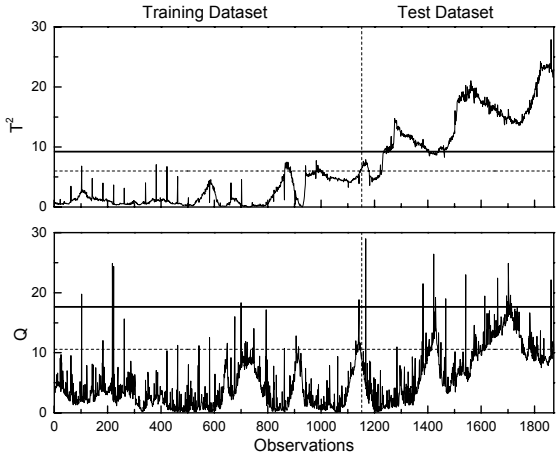

Fig. 4. Process monitoring charts for the first test dataset.

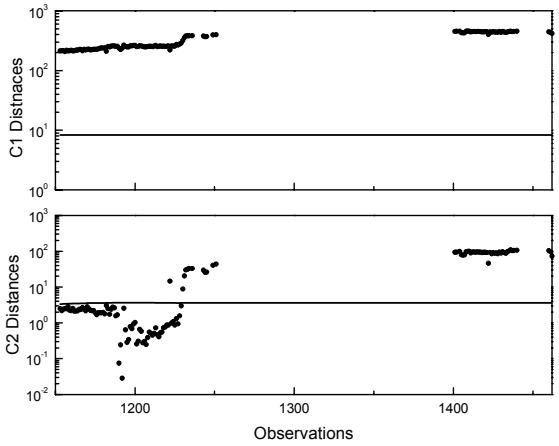

Fig. 5. The Mahalanobis distances of data points.

The data after the training dataset were collected every five minutes for 2.5 days. Fig. 4 shows the statistic Q and T^2 of the first test dataset, in which the solid and dashed lines represent the 99% and 95% confidence limits respectively, and the control charts suggested the data belonged to the subspace before the 1235th observation and the sample number between 1400 and 1440. Since a data point may be an outlier, even it belongs to the PCA subspace. It is necessary to validate whether the data point belongs to the known groups, in order to avoid misleading operator's decisions. So, before classifying the data points, which belonged to the PCA subspace, into the known groups, the Mahalanobis distances of a data point to each cluster center needed to be compared with the corresponding cluster boundary distances respectively. Fig. 5 shows the data points neither belonged to C1 nor C2 cluster after the 1230th observation, so these data should be stored for next off-line model update. In the figure, it should be noted that the Mahalanobis distances of the data points, which were beyond the subspace, were not calculated in the period of the

sampling number between 1251 and 1400. On the other hand, the data points, in which at least one of the distances was less than the corresponding boundary distance, were classified into the C2 cluster according to their membership values. The center, covariance, and boundary distance of C2 cluster were finely adapted by using the additional data, which belonged to the cluster, as Fig. 6 shows. In the figure, the outliers, representing with white points, were collected for next model update. After that, the PCA subspace was accommodated by using the new event data and the adaptive PCA method. The adaptive PCA subspace, which captured 90% variances with 2 PCs, was applied to the first test dataset. Fig. 7 shows the Q and T^2 statistics, in which almost all of the statistics of the test data were under their control limits. It shows the proposed method is capable of adapting the subspace to accommodate the known and new events without recalculating the trained data.

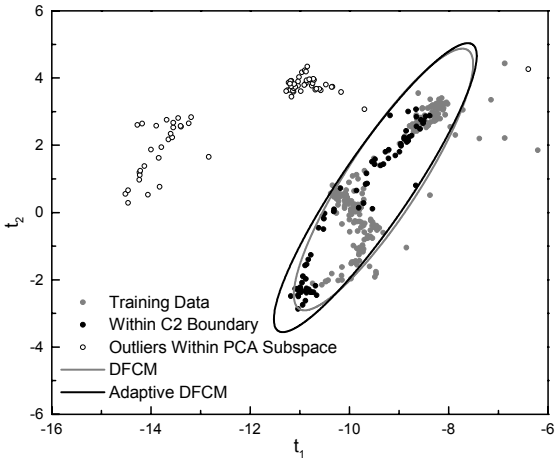

Fig. 6. Adapting C2 cluster with additional information.

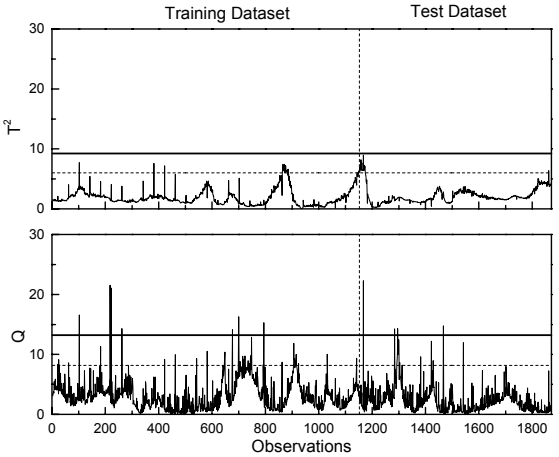

Fig. 7. Statistic Q and T^2 for the training and the first test datasets with the adaptive PCA subspace.

Only the new data need to be clustered on the new PCA subspace, the known cluster parameters, including cluster centers, covariances, and boundaries were transferred to the new subspace through rotating and shifting coordinates of the subspace. The trained data projected on the new subspace were plotted in Fig. 8 with gray points. In the figure, the cluster C1 and C2 were transferred from the previous subspace. It demonstrates the proposed method is capable of transferring the known clusters to the new subspace. The scores of the new event data and the outliers on the previous subspace were clustered by using DFCM. The cluster boundaries were labelled as C3 and C4 respectively in the figure. Eventually, the outliers on the previous subspace, plotted with black points in Fig. 8, were parts of the C3 group that was an unknown event in the previous subspace. It is obvious that the operator's decisions would be easily misled, if the outliers were not identified in the on-line classification phase.

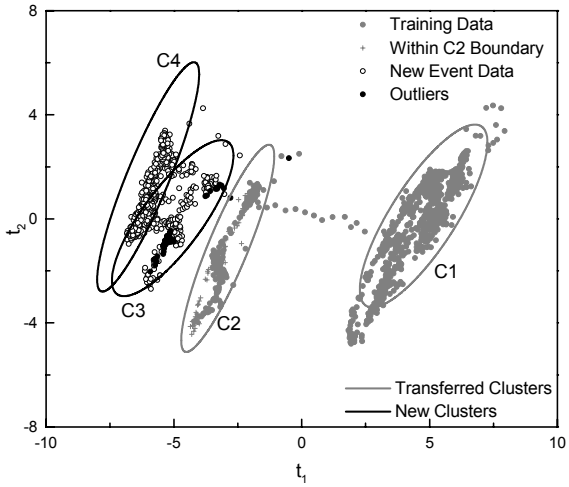

Fig. 8. Clustering results on the new subspace.

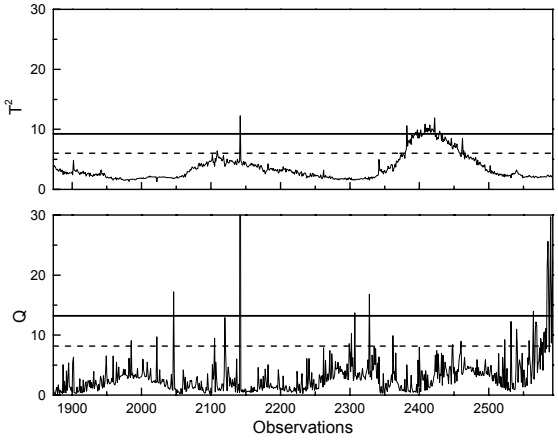

Fig. 9. Process monitoring charts by using the adaptive PCA.

The second test dataset, which was collected every five minutes after the first test dataset for 2.5 days, was examined by utilizing the adaptive PCA subspace. Fig. 9 shows most of the statistic Q and T^2 of the second test dataset are within their control limits except that the 2142th observation and the observation number between 2407 and 2430, and after the 2584th observation, which were stored for next model update. Fig. 10 compares the distances of the data points belonging to the adaptive PCA subspace with each cluster boundary distance. It shows all data points were far from C1 and C2 clusters, most of them were within the C4 boundary. The first two scores of the data points also have been plotted in Fig. 11. In the figure, the cluster parameters of C3 and C4 were adapted by using additional information. It shows C3 cluster slightly moved toward C4 group and the C4 cluster

was compactly shrunken in order to accommodate the new information. Therefore, the on-line adaptive method is not only capable of rejecting outliers to mislead the model updating direction, but also is effective to tackle the time-varying process.

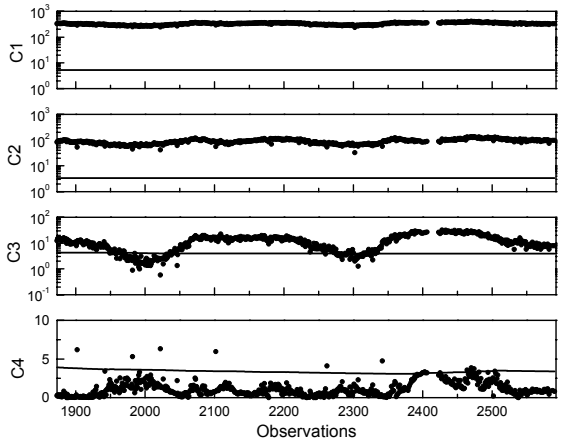

Fig. 10. Data points to each cluster distances.

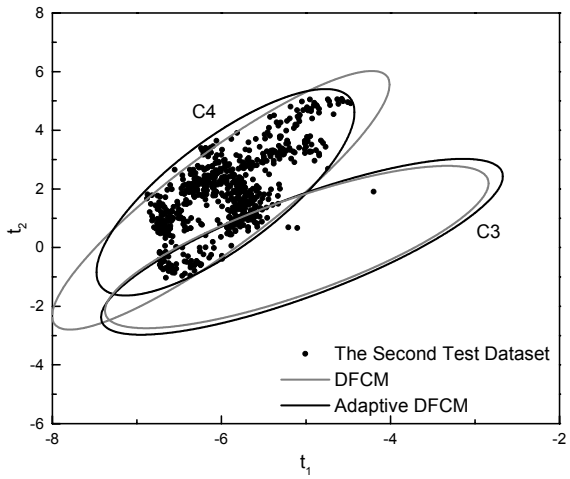

Fig. 11. Adapting clusters by using the second test data.

To look up the operator logs, the inlet flowrate of the compressor was decreased around the 900th observation due to the downstream processes reducing their demand. Fig. 12 shows air inlet flowrate was cut to 80% of normal operating flowrate at that time point. After that, it further reduced to 70% around the 1250th sampling interval. In order to avoid the compressor surge, the output pressure had to be decreased simultaneously. Fig. 12 shows the output pressure gradually decreased to 82% of normal operating pressure. It should be noted that the set point of output pressure was manually adjusted according to the anti-surge control line by field operator. For the purpose of smoothing transitions, operator carefully adjusted the set point in the low flowrate region. It can be found the set point changes of the discharged pressure were more frequent than in the normal operating region, as Fig. 12 shows. Compare with clustering results of training dataset, the C1 and C2 clusters represented the normal operation and 80% of inlet flowrate respectively, as Fig. 12 shows. The first test dataset was grouped into C3 and C4 clusters, which were partly overlapped as Fig. 8 shows. It is reasonable that the operator finely manipulated set points of the controlled variables to bring the process to the target. Therefore, the operating regions of the two clusters were close, but not identical. It resulted in the part of the second test data were shared by C3 and C4 clusters, as Fig. 12 shows.

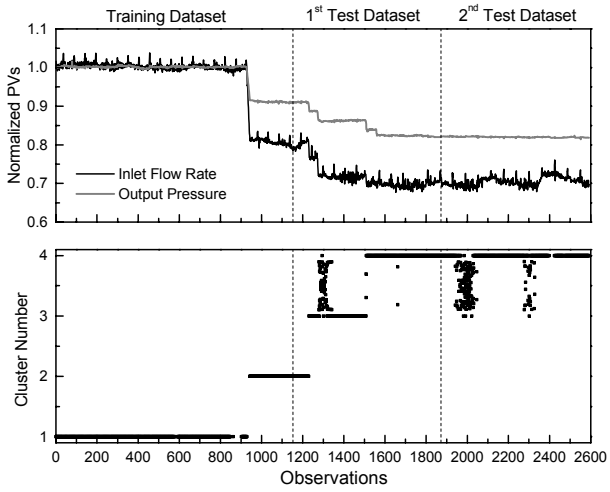

Fig. 12. Compare the operating conditions with the clustering results.

4. CONCLUSIONS

In this paper, an outlier rejection clustering algorithm (DFCM) has been applied to a real plant dataset in order to trim transition data and reach more feasible clustering results. In addition, blockwise recursive PCA algorithm has been utilized to accommodate new event data. When the subspace has been adapted, the known event clusters have to be transferred to the new subspace by rotating and shifting coordinates of the subspace. The real plant data from a compressor process in the oxygen plant have been demonstrated by using the proposed approach. In this case, new events emerged while on-line monitoring. Before on-line classification, it is necessary to identify the new observations whether they belong to the known event groups; even they are within the PCA subspace. Results show the challenges of process monitoring, such as collinearity, outliers, and time-varying characteristics, can be effectively dealt with.

ACKNOWLEDGMENT

This work was supported by the National Science Council, Republic of China, under Grant NSC-96-2221-E-268-004, and by the China Steel Corporation.

REFERENCES

Jackson, J.E. (1991) *A User's Guide to Principal Components*, Wiley, New York.

Liu, J. (2004) Process Monitoring Using Bayesian Classification on PCA Subspace. *Ind. Eng. Chem. Res.* **43**, 7815-7825.

Marsili-Libelli, S. and Műller, A. (1996) Adaptive fuzzy pattern recognition in the anaerobic digestion process. *Pattern Recognition Letter*, **17**, 651-659.