MULTI-PCA MODELS FOR PROCESS MONITORING AND FAULT DIAGNOSIS

Liling Ma, Yunbo Jiang, Fuli Wang

P.O.Box 131
The School of Information Science and Engineering,
Northeastern University, Shenyang, 110004, P.R.China
E-mail: maliling1974@yahoo.com.cn

and

Furong Gao

Department of Chemical Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong

Abstract: Multivariate statistical approaches have been proved effective for reducing the dimension of highly correlated process variables and subsequently simplifying the tasks of process monitoring and fault diagnosis. However, for the process with distinctive stages, a single statistical model is not sufficient or even incapable to map the substantive process information. In this paper, multi-PCA models are proposed for promptly detecting faults and improving the exactness of the diagnosis as well. The effectiveness of the approach is demonstrated on a complicated fermentation process.

Keywords: process monitoring; multi-PCA models; clustering technology; fault diagnosis

1. INTRODUCTION

With the ever-increasing demand of control precision, modern industrial plants become more and more complicated. As a result, the tasks of prompt detection of any abnormal process behaviour, which is caused by breakdowns or malfunctions of plant instruments or grievous working conditions, is more challenging nowadays. Traditional model-based approaches based on the assumption that the occurrence of any unexpected faults will change the physical parameters or states, are no longer applicable in most cases because of the difficulty to get the theoretical models from the control theory to setup any precise parameter estimators or state estimators (J.Zhang, et al, 1996). The knowledge-based approach known as expert system demands a deep and comprehensive understanding of the whole process (J.Zhang, et al, 1996). To setup a reasonable rule set is rather difficult and time consuming. Fortunately, with the application of modern process computers, thousands of variables can be collected and processed within a few seconds. The distributions of and correlations among these variables encapsulate precious knowledge of the plant (Theodora Kourti, et al, 1996). Thus, by analyzing the variance of the historical operating data, the characters of the plants can be learnt through multivariate statistical techniques. In recent years, data based multivariate statistical techniques, such as principal component analysis (PCA) and projection to latent structure (PLS), have received much attentions for the simplicity and practicality. Their excellent abilities in extracting the chief information of the process and casting away the noises have been fully demonstrated on many applications. By projecting the highly correlated process data onto a lower dimensional variable space without discarding any useful process information, these methods can greatly simplify the task of process monitoring and make it easier for fault diagnosis as well (P.R.Goulding, et al ,2000). The main advantage of multivariate statistical approaches is that they are largely dependent on the historical operating data and need not have a comprehensive knowledge of the complicated process.

However, when the plant works through several different phases during a batch process, the relationships of the variables will be quite different (Svante Wold, et al, 1996). In other words, the plant will exhibit different collinear behavior in each phase. From this point, a single PCA or PLS model is not sufficient to map the whole process information. When taking different stages into consideration, the multivariate statistic confidence bounds will be inappropriately set and are always larger than needed. Consequently, the probability of failure to report the abnormal sample will be greatly increased.

The aim of this work is to overcome these annoying problems and thus improve the precision of the PCA models for prompt fault detection and diagnosis. A practical approach based on the sub-PCA models is proposed. Hyper-ellipsoid based clustering procedure is designed to categorize data. Then, supervised training approach of SOFM network is described for clustering faults features. This approach is fully demonstrated by the experiments on the fermentation process. The results show the feasibility and effectiveness of the proposed method.

2. PROCESS MONITORING AND FAULT DIAGNOSIS SCHEME

Principal components analysis was first proposed by Hotelling to analyze the correlated structures of the multi-variables. It has become one of the most popular multivariate statistical techniques and has received wide application in industrial processes. By projecting the original information onto a lower dimensional space, the principal components can summarize the chief information about the variance in the original data set (Parthasarathy Kesavan, et al, 2000). Suppose X is the original data set which is composed by m variables and k principal components are enough for summarizing the main information, X can be decomposed as the following equation:

$$X = \sum_{i=1}^{k} t_i \times p_i^{\ T} + E \qquad (1)$$

The number of proper principal components can be determined by the accumulated contributions of the principal components or cross validation. Process monitoring is based on the two statistics called T^2 and SPE (E.B.Martin, et al, 1996), which conform to F-distribution and normal distribution respectively.

$$T^2 = T_k \Lambda^{-1} T_k \qquad (2)$$

where Λ is the diagonal matrix composed of the first k eigenvalues of $X^T X$.

$$SPE = trace(R \times R^T) \qquad (3)$$

R is the residual matrix;

$$\begin{aligned} R &= X - \hat{X} = X - X \times P \times P^T . \\ &= X \times (I - P \times P^T) \end{aligned} \qquad (4)$$

As has been discussed in Section 1， multi-PCA models are necessary for the process with distinctive stages to improve the promptness of fault detection and to ease the following fault diagnosis as well. To perform the task of the process monitoring using multi-PCA models, the data sampled from which phase should be identified first, that is, the fitness of the data to each cluster should be determined. Then the data are projected onto the related single PCA model or the combination of several PCA models and corresponding control limits are set to monitor the performance of the process. The diagram of the whole procedure is illustrated in Figure 1. This scheme is composed of three steps. At first, sampled data is classified based on hyper ellipsoid clustering technique. Then, analysing the assorted result, process monitoring is realized. If the fault is detected in this phase, the last fault diagnosis will be accomplished by SOFM network with these samples.

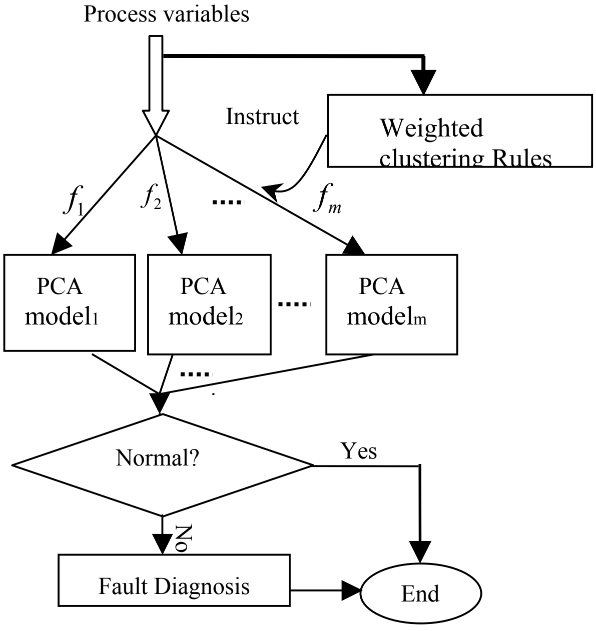


Fig1. The sketch map of the fault diagnosis

3. DESIGN OF MULTI-PCA MODELS

The core of building multi-PCA models is how to classify sampling data currently. A proper clustering technique is fundamental for reasonable decision-making. Former researches have investigated various clustering techniques ranging from simple identical sphere windows with fixed centres to intelligent approaches using neural networks, such as RBFN and SOFMN. In practical, the distribution of each group is not necessary of the same size, so the identical sphere windows will not work in most cases. Though clustering techniques using NN are powerful at processing nonlinear information and some have excellent self-learning ability in determining the numbers of the clusters, to assign the proper neurons and to train the weights of the NN are rather tough work. For example, when training the RBFN, to select the centres of the radial basis functions of hidden neurons and to determine their widths are indeed demanding jobs (Gao Daqi, et al, 2001). Further more, the more the input variables, the more complicated of the NN structure; clustering the newly sampled data would be time-consuming because of the over-burden computing procedure.

K-means clustering approach is a well-developed technique. Currently the trial-and-error method is adopted to determine the number of the clusters. However, it is based on the Euclid distance and the data of the same cluster are confined within a hyper-sphere. Since the variance of each variable is not necessary of the same size, the bounds of the clusters should be hyper-ellipsoid rather than hyper-sphere. The traditional K-means clustering approach based on the hyper-sphere bound, improper classification of the data often happens (Johnston, et al, 1994).

To reduce the probability of the misclassification, a set of clustering rules are suggested in this paper. First, suppose the number of the clusters is known according to the knowledge of the character of the process, use K-means clustering algorithm to grossly divide the data set into several clusters. (If the number of the clusters is not known, adopt the trial-and-error method to determine the number of the clusters.) Then analyze the variance of each cluster and adjust its bound. The following procedures are detailed as follows:

Find out the direction, along which the variance is the largest, and then the next. Those directions are orthogonal to each other. Project the original data onto each direction and find the centre of the projection. In fact, the first direction contains the largest amount of the information of the process and the main information of the process can be expressed by the first few projections. The information contained in the last few projections can often be explained as noise. When there are many variables and the variances along the last few directions are small, those projections can be neglected. The whole process is similar to the procedure of subtracting principal components. The bounds of the clusters are hyper-ellipsoids whose axes are overlapped with the principal variance directions. The size of the hyper-ellipsoid can be determined according to statistical confidence level.

After finding out the directions, the fitness μ of the data X to each cluster is measured by following equation:

$$\mu = \frac{1}{S_\alpha} \times T^2 \qquad (5)$$

where T^2 is Hotelling's statistic:

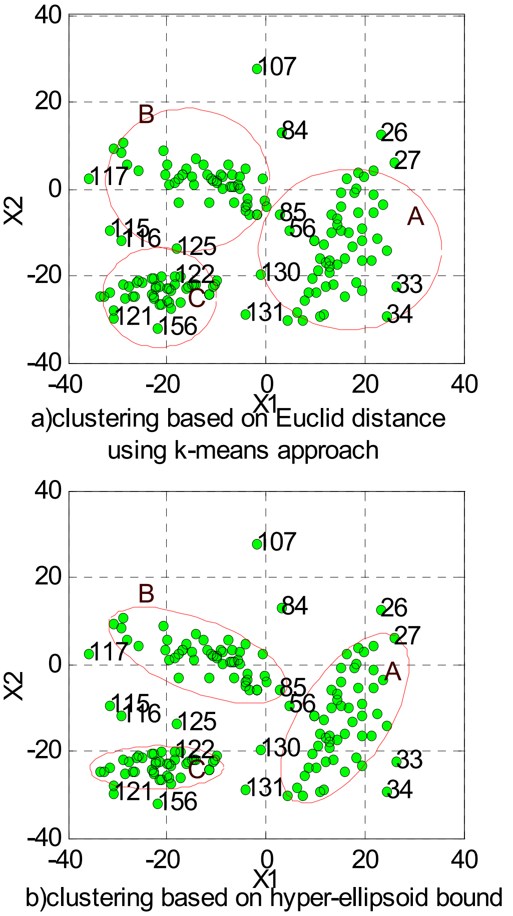
$$T^2: \quad T^2 = T_k \Lambda^{-1} T_k \qquad (6)$$

where Λ is the diagonal matrix composed of the first k eigenvalues of $X^T X$. T_k is the first k principal components, $T^2 \in R^{1\times k}$. T^2 obeys F distribution. Define S_α based on F_α as follows:

$$S_\alpha = \frac{k(n-k+1)}{n(n-k)} \times F_\alpha(k, n-k) \qquad (7)$$

where n is the size of the cluster, k is the dimension of the original data or the number of the principal components and α is the confidential level , here, $\alpha = 0.95$. $\mu = 1$ represents the hyper-elliptic bound. If $\mu < 1$, it means the data is in the inner of the bound. Any samples falling into the clustering bound can be regarded as the same type. Then reclassify the data set and adjust the chief variance directions and centres, repeat the former steps until the classification of each data will not change.

The advantages of the clustering technique based on hyper-elliptic bound are illustrated by a simple two-dimensional clustering problem in Figure1. Figure1a) demonstrates the clustering results by pure K-means and Figure1b) shows the clustering results based on elliptic bound. From the distribution of the samples, sample 125 is far away from the other samples in cluster B and it is more reasonably be classified as singularity as sample 117 and 115 etc. Sample 85 should be classified to cluster B though its Euclid distance from the centre of B is farther than that from the centre of A. From the above, the clustering based on the hyper-elliptic bound can overcome the shortcoming of the traditional K-means clustering approach.



a)clustering based on Euclid distance using k-means approach

b)clustering based on hyper-ellipsoid bound

Fig 2 Comparison of the two clustering techniques

Thus, using the clustering method proposed above, multi-PCA models for process monitoring and fault diagnosis are built as follows:

$$X^{(k)} = T^{(k)} P^{(k)T} = \sum_{i=1}^{m} t^{(k)}{}_i \; p^{(k)}{}_i{}^T + E^{(k)}{}_m \qquad (8)$$

Where k presents the kth sub-PCA model. These sub models make up the multi-PCA models.

4. PROCESS MONITORING USING MULTI-PCA MODEL

When a new sample comes during process monitoring, the data samples from which phase should be identified first, that is, the fitness of the data to each sub-PCA model should be determined. Then the data are projected onto the related single PCA model or the combination of several PCA models and corresponding control limits are set to monitor the performance of the process.

The smaller the T^2 to the cluster, the better of the fitness to that cluster. Certainly, most of the data sampled during each phase can be clearly classified. However, since the process is continuous, the transitory data are likely to contain both characters of the neighbor clusters. When the cluster bounds are rigidly set, some transitory data will be likely classified as singularities. On the other hand, the probability of misclassification will be increased. To solve this problem, fuzzy-clustering rules are proposed (Yang Yinghua, et al 2002). Two bounds of a cluster are suggested and their sizes are determined by two radius, namely, kernel radius and class radius. Here, we can also set two hyper-elliptic clustering bounds based on different confidence levels:

$$\text{Class bound: } \frac{1}{S_{0.99}} \times T^2 = 1 . \qquad (9)$$

$$\text{Kernel bound: } \frac{1}{S_{0.90}} \times T^2 = 1 \qquad (10)$$

The fitness of the samples to each cluster can be computed by the following rules:

1) If T^2 statistic of the new sample falls into one of the kernel bounds, that is, $\frac{1}{S_{0.90}} \times T^2 \le 1$, the fitness of the sample to the cluster can be assigned to 1;

2) If $1 < \frac{1}{S_{0.90}} \times T^2$ and $\frac{1}{S_{0.99}} \times T^2 \le 1$, and T^2 is beyond any other class bound, the fitness of the sample to the cluster can also be assigned to 1.

3) If T^2 falls into the overlapped area of several class bounds, suppose the new T^2 statistic falls into m clusters, define

$$L_1 = \frac{1}{S_{0.99}} \times T_1^2 , \ldots, L_m = \frac{1}{S_{0.99}} \times T_m{}^2 \qquad (11)$$

obviously $L_1,\ldots,L_m < 1$, the smaller the L_k, the closer of the new sample to the kernel of the cluster. The fitness of the new sample to each related cluster can be defined as follows:

$$f_i = \frac{1/L_i}{\sum_{k=1}^{m} 1/L_k} \qquad (12)$$

4) If T^2 statistic falls into neither class bounds, it can be regarded as a singularity.

If the new sample is regarded as totally subjected to one classification, the procedure of the monitoring is the same as that based on a single PCA models. When the new comer falls into the common region of several regions, the fitness to each cluster is computed according to equation (12) first. Then adjust the directions of the principal components based on its fitness to each clusters (Yang Yinghua, et al 2002):

$$P = \sum_{i=1}^{m} f_i \times p_i \qquad (13)$$

here $\sum_{i=1}^{m} f_i = 1$ and p_i is the principal components directions of each sub-PCA models. The principal components can be achieved by projecting the original data on subspace explained by P. The SPE control limit is computed as follows:

$$U_{SPE} = \sum_{i=1}^{m} f_i \times U_{SPEi} \qquad (14)$$

5. FAULT DIADNOSIS USING SOFM NETWORK

When a fault is detected by previous step, SOFM is used to diagnose the fault, dealing with the current sample data as inputs. SOFM neural network was originally developed by a Finland scientist Kohonen. It is similar to the memory mode of the human beings. Different to other kinds of neural network, the information of one pattern is not memorized by one cell in SOFM neural network, but by a set of neurons in certain region. The excited region in the network is like a Mexican Hat, with the central neuron cell being most excited when stimulated by the corresponding pattern. The excitement of the neuron nearby reduces and the neurons outside this region are restrained. Further more, the distribution of the weight vectors reflect the statistical characters of the input mode. When reminiscing, pattern classification is mainly based on the most excited neuron.

The chief advantage of the SOFM is its self-learning ability. It can automatically categorize the input mode without supervision. When the former knowledge of the clusters is not sufficient, SOFMNN is adept at extracting the character of each cluster through self-organized learning. The structure of the network is composed of two layers, input layer and output layer. The output layer is a competing layer in the form of two-dimensional array. The structure of the network is shown in Figure 3.

The training algorithm can be found in many literatures. The adjustment of the connection weights is based on the following equation:

$$W_{k+1} = \begin{cases} W_k + \eta(k)[U_k - W_k] & j \in N_k \\ W_k & j \notin N_k \end{cases} \qquad (15)$$

where N_k is the neighbor field, $\eta(k)$ is the learning factor. N_k begins with a large area and contains all the neurons from the origin, and then shrinks to only contain one to two neurons from the centre (C.W.Chan, et al, 2001):

$$N_k = int\left[N_0 \times \left(1 - \frac{k}{K}\right)\right] \tag{16}$$

The learning speed also reduces with the increase of k. It can be adjusted according to the following equation:

$$\eta(t) = \eta_0 \times exp\left(-\frac{k}{K}\right) \tag{17}$$

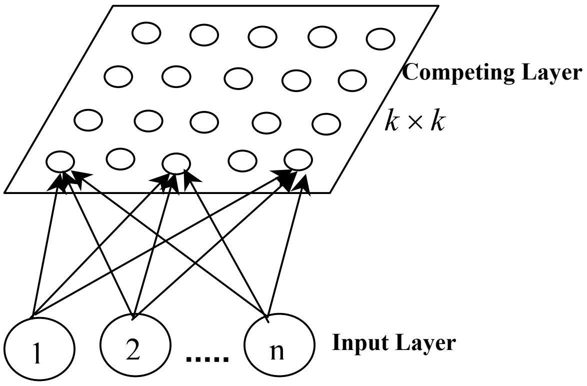


Fig 3. The structure of SOFM network

6. CASE STUDY ON FERMENTATION PROCESS

The fermentation plant for producing glutamic-acid is introduced to evaluate the approach proposed in this paper. It experiences three distinct phases, namely, the growing phase, fermenting phases and perishing phase. The acidity and the amount of dissolved oxygen have different characters in three phases. In the growing phase, the acidity increase slowly with the production of glutamic-acid. The demand for dissolved oxygen increase too. In the following fermenting phase, with a large number of glutamic-acid being produced, the PH value decrease quickly and the demand for dissolved oxygen increase markedly. When the production peak passes away, the acidity falls slightly. So in this experience, the PH value and the amount of the dissolved oxygen as well as their tendencies are used for pattern classifications(Xu Ling, et al, 1999). Three hyper-ellipsoids are defined for classification on the historical normal operating data. In this experiment, the class hyper-elliptic bound is set based on 0.99 confidence level and the kernel bound based on 0.90 confidence level. The distribution of the historical normal operating data and the clustering bounds for each clusters are illustrated in Figure4.

Based on the classification, three PCA models are developed for monitoring. Seven variables are used while analyzing the fermenting process. They are PH value, dissolved oxygen density (DO), the changing rate of DO, temperature of the fermenting environment, the inflow of the atmosphere, the position of the outlet valve and the pressure of the fermenting environment. Figure 5 and Figure 6 show the performance of the PCA models when monitoring a normal process and detecting the occurrence of the fault1 and fault2 using single PCA model and multi-PCA models respectively. Fault1 represents the failure of outlet valve. The solid line in the figures represents the control limits based on 0.99 confidence level and the dash based on 0.90 confidence level. The plus signs represent the abnormal samples identified during clustering. The diamond signs represent the samples falling into the overlapped regions of the clusters' class bounds.

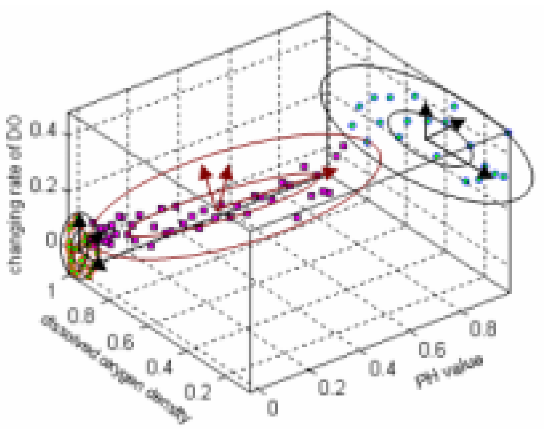


Fig. 4 Clustering illustration of fermentation plant

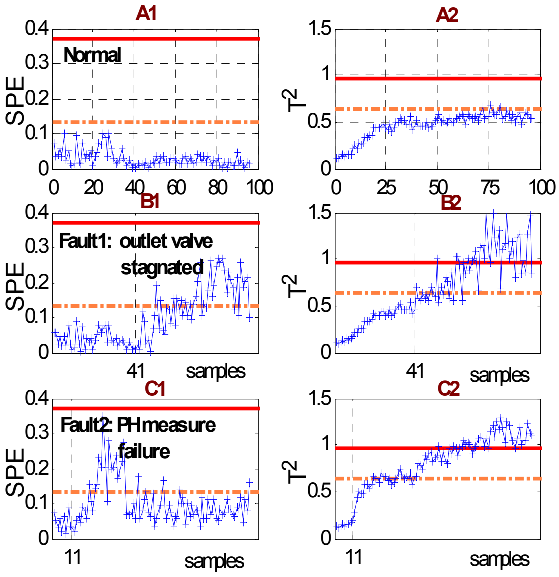


Fig.5 Process monitoring using single PCA model

When the pressure inside the fermentation is out of control, there is a contamination. As a result, the PH value will be affected too. Fault2 simulated a sensor failure, that is, the PH instrument doesn't work. When the faults occur, the correlated structure of the data will be changed, the two T^2 and SPE statistics will be out of control theoretically. However, since the fermentation plant contains three distinct phases, whose correlated variable structures are quite different, the bound of the control is difficult to be adjusted. It is obviously that the SPE control limit is lager than needed during the fermenting phases and perishing phases, which leads to the failure to report the Fault2 illustrated in the C1 sub-chart. The

T^2 control limit is also inappropriately set for the growing phase. From Figure6, the precision of the monitoring models is greatly improved when using multi-PCA models and consequently the promptness of detecting faults is improved too.

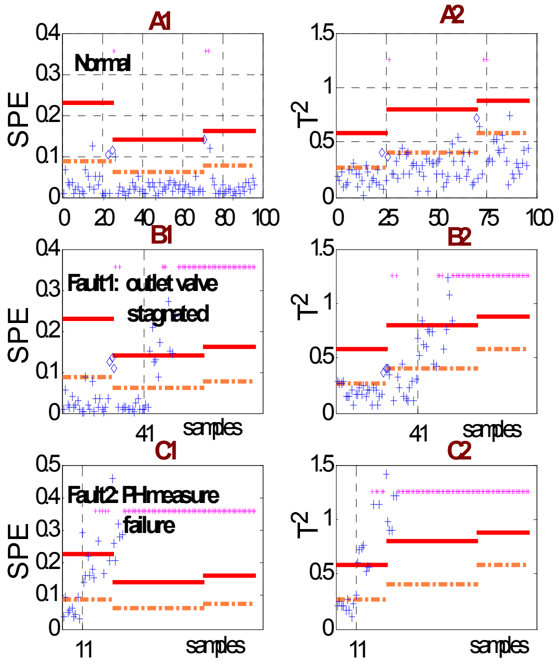


Fig.6 Process monitoring using multi-PCA models

Figure 7 shows the results of fault classification. Choose a 8 × 8 array of neurons to compose a competing layer and select the tendency of the PH value and DO, temperature, pressure etc. as the input of the SOFM network. After training, three regions of neurons are stimulated corresponding to three pattern inputs. When the Fault1 is detected and the current sampled data is input to the SOFM network, the 18th neuron or the neurons nearby will be the most excited according to the reminiscence. The Fault2 data will stimulate the neurons with the centre of 13th neuron. The results of Figure 7 show that the corresponding faults can be diagnosed exactly.

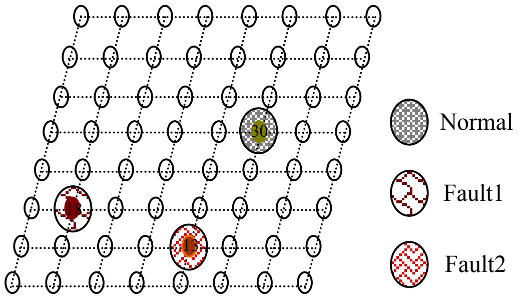


Fig.7 Illustration of faults classification on SOFMN's competing layer

6. CONCLUSIONS

Multivariate statistical approaches have received widely application for the processes rich in measurement data. However, for those the data structures are quite different in different stages, setting proper control limits is difficult. In this paper, multi-PCA models are suggested. Process monitoring is based on the combinations of related sub-PCA models and the weigh of each sub-PCA model is assigned according to the weighted clustering technique. Fault classification is realized by SOFM network. The feature of the fault can be stored in the weights of the network through self-organize learning. The effectiveness of the proposed approach is demonstrated by the experiment on fermentation process.

REFERENCES

C.W.Chan, Hong Jin, K.C.Cheung (2001) Fault Detection of Systems with Redundant Sensors using Constrained Kohonen Networks. *Automatica*. 37, 1671-1676.

E.B.Martin, A,J,Morris (1996). Process Performance Monitoring Using Multivariate Statistical Process Control. *IEE Proc-Control Theory Appl.* 143 (2), 132-144.

Gao Daqi, Yang Genxing, (2001). Basic Principles of Pattern Classification Methods Based on Improved RBF Neural Networks. *Journal of East China University of Science and Technology*. 27(6), 667-683.

Johnston, L.P.M.; Kramer, M.A. (1994). Probability Density Estimation Using Elliptical Basis Functions. *AJCHE J*. 40, 1639.

J.Zhang, E.B.Martin, A.J.Morris(1996).Fault Detection and Diagnosis Using Multivariate Statistical Techniques. *Trans IchemE*. 74, Part A, January, 89-96.

Parthasarathy Kesavan, Jay H.Lee (2000). Partial Least Squares(PLS) Based Monitoring and Control of Batch Digests. *Journal of Process Control*, 10, 229-236.

P.R.Goulding, B.Lennox, (2000). Fault Detection in Continuous Processes Using Multivariate Statistical Methods. *International Journal of Systems Science*,.31 (11), 1459-1471.

Svante Wold, Nouna Kettaneh (1996). Hierarchical Multi-block PLS and PC Models for Easier Model Interpretation and as an Alternative to Variable Selection. *Journal of Chemo-metrics*. 10, 463-482.

Theodora Kourti, Jennifer Lee (1996). Experiences With Industrial Applications of Projection Methods for Multivariate Statistical *Process Control. Computers chem. Engng*. 20, S745-S750.

Xu Ling, Xu Wenbo, (1999) The Fuzzy PID Control for Dissolved Oxygen in Fermentation Process. *Process Automation Instrumentation*.20, 3-7.

Yang Yinghua, Lu Ningyun, Wang Fuli (2002) Statistical Process Monitoring Using Multiple PCA Models. *American Control Conference (ACC02)*. 5072-5073.