Fault Diagnosis in Industrial Processes Using Principal Component Analysis and Hidden Markov Model

Shaoyuan Zhou, Jianming Zhang, and Shuqing Wang

Abstract—**An approach combining hidden Markov model (HMM) with principal component analysis (PCA) for on-line fault diagnosis is introduced. As a tool for feature extraction, PCA is used to reduce the large number of correlated variables to a small number of principal components in an optimal way. HMM is applied to classify various process operating conditions, which is based on pattern recognition principles and consists of two phases, training and testing. The moving window for tracking dynamic data is used. The impact of the window length is studied by simulation. The sampling rate used in training data and in test data is different for correct and quick fault diagnosis. Case studies from the Tennessee Eastman plant illustrate that the proposed method is effective.**

I. Introduction

With the increasing integration and complexity of chemical processes, it is essential for reliability and safety of the plant and for maintaining quality of the products to identify faults correctly and timely. With the widespread availability of distributed control systems (DCS), on-line fault diagnosis of chemical process is greatly facilitated, and has been studied intensively in recent years, which is recognized as a powerful support tool for operators. Much of the previous work on this topic is based on mathematical models and statistical models [8]. In the last two decades many contributions have been made using neural networks trained by steady-state data [2],[10], while some researchers trained the neural networks using dynamic data, a number of sets of time series data, and qualitative process dynamic trend [6],[10]. There are also some speech recognition approaches developed for fault diagnosis recently, such as dynamic time warping (DTW) for off-line diagnosis [1], and hidden Markov model (HMM) for detecting abnormal process operation [9]. Most approaches mentioned above contain two steps, feature extraction and pattern recognition. Selecting some important measurement variables via human experience is a critical step for fault diagnosis in these methods, which is difficult in many complex chemical processes, and the feature sequences extracted from training data and testing data have the same length and sampling rate.

In this work, all the measurement variables of plant are useful for fault diagnosis. Since all the measurement variables are highly correlated with each other, principal component analysis (PCA) will be used to reduce the large number of correlated variables to a small number of principal components in an optimal way without losing important information. These principal components can be used as the feature sequences (called observation sequences in this work), and indicate various kinds of process operating conditions. Then hidden Markov model is used for classification, which is based on pattern recognition principles and consists of two steps.

--First, a set of observation sequences $\boldsymbol{o}_{train}$ of training patterns (including normal and faults) is extracted via PCA and used to train corresponding HMMs.

--Second, when the pattern of an unknown fault is obtained, it is compared with all the reference patterns.

The moving windows for tracking dynamic data are used. For correct and quick fault diagnosis, the length of observation sequences $\boldsymbol{o}_t$ extracted from test patterns, which is determined by the moving window length, can be different from the one of observation sequences $\boldsymbol{o}_{train}$, and they also have the different sampling rate. The simulation of Tennessee Eastman (TE) plant with the decentralized Proportional-Integral-Differential (PID) control system is used to illustrate the proposed method.

This paper is organized as follows: In section Ⅱ and Ⅲ, PCA and HMM are briefly described. In section Ⅳ, PCA-CHMM based fault diagnosis method is developed in detail. In section Ⅴ, a simulation study using Tennessee Eastman plant is performed and the results are discussed.

Manuscript received September 19, 2003. This work was supported in part by the National High-Tech program of China under grant No. 2001AA413110.

Shaoyuan Zhou is with National Key Lab of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou, 310027, Zhejiang Province, China (phone: 86-571-8795-2441; fax: 86-571-8795-1445; e-mail: syzhou@iipc.zju.edu.cn).

Jianming Zhang is with National Key Lab of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou, 310027, China (e-mail: jmzhang@iipc.zju.edu.cn).

Shuqing Wang is with National Key Lab of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou, 310027, China (e-mail: sqwang@iipc.zju.edu.cn).

II. Principal Component Analysis

PCA is an optimal dimensionality reduction technique in terms of capturing the variance of the data. Given n observations of m measurement variables stacked into a training data matrix $\boldsymbol{X}$, which can be decomposed via singular value decomposition (SVD) as follows,

$$\boldsymbol{X}/\sqrt{n-1} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{1}$$

where $\boldsymbol{U} \in \Re^{n\times n}$ and $\boldsymbol{V} \in \Re^{m\times m}$ are unitary matrices and the matrix $\boldsymbol{\Sigma} \in \Re^{n\times m}$ contains the nonnegative real singular values of deceasing magnitude $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m)$. The loading vectors are the orthogonal column vectors in the matrix $\boldsymbol{V}$, and the variance of the i^{th} principal component (the projection of the training set along the i^{th} column of $\boldsymbol{V}$), is equal to σ_i^2. The first principal component is the direction in the physical variables along which the data exhibit the greatest variability. Subsequent principal components explain the remaining variability, while being orthogonal to the previous principal component. A small number of principal components, which still retain most of the information, can represent process operating condition. In this study, the raw data from training patterns and test patterns must be pre-processed via PCA, and a certain number of principal components are used as observation sequences for training and testing.

III. Hidden Markov Model

HMM is a double stochastic model, which not only can capture the serial correlations in the data, but also can take into account the random factors of process. The underlying backbone of HMM is a Markov process, the states of which only can be observed through another set of stochastic processes representing a sequence of observation. HMM usually has a chain structure (shown in Fig. 1), and can be characterized by five parameters.

1) N : the number of states in the model. The states are denoted as $\boldsymbol{S} = \{S_1, S_2, \cdots, S_N\}$.

2) M : the number of distinct observation symbols per state. The observation symbols correspond to the physical output of the system being modelled. The symbols are denoted as $\boldsymbol{V} = \{V_1, V_2, \cdots, V_M\}$.

3) $\boldsymbol{A} = \{a_{ij}\}$: the state transition probability distribution, where

$$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i), \quad 1 \leq i, j \leq N \tag{2}$$

a_{ij} is the probability of going to state S_j at time $t+1$, given that at time t, the state is S_i.

4) $\boldsymbol{B} = \{b_j(k)\}$: the observation symbol probability distribution in state S_j, where

$$b_j(k) = P(V_k(t) \mid q_t = S_j), \ 1 \leq j \leq N, 1 \leq k \leq M \tag{3}$$

$b_j(k)$ is the probability of the k^{th} observation symbol given that the state is state S_j and the time is time t.

5) $\boldsymbol{\pi} = \{\pi_i\}$: the initial state distribution, which is the probability of being in the i^{th} state at the initial time, $t = 1$. Where

$$\pi_i = P(q_1 = S_i), \ 1 \leq i \leq N \tag{4}$$

According to characteristics of the observation symbols, there are two kinds of HMMs, discrete hidden Markov model (DHMM) and continuous hidden Markov model (CHMM). The observation symbols of DHMM are mentioned above. The observation symbols of CHMM are continuous, Gaussian distribution of which is assumed in each hidden state. In this study, CHMM is used to classify various process operating conditions. There are three fundamental problems need to be solved in the HMM application.

1) Given the observation sequence $\boldsymbol{O} = \{o_1, o_2, \cdots, o_T\}$ and a model $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, how to calculate $P(\boldsymbol{O} \mid \boldsymbol{\lambda})$? The solution provides a score or measure of similarity between the observation sequence and the model.

2) Given the observation sequence $\boldsymbol{O} = \{o_1, o_2, \cdots, o_T\}$ and a model $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, how to determine the most likely state sequence that corresponds to the observation sequence $\boldsymbol{O}$.

3) How to refine model parameters $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ to maximize $P(\boldsymbol{O} \mid \boldsymbol{\lambda})$? The parameter re-estimation process is carried out using a set of observation sequences from training data.

HMM is formulated in two stages, training and testing. The first two problems are solved in the test phase, while the model re-estimation problem is solved during training phase. A well-known Bawm-Welch method can efficiently solve both the training and testing problems mentioned above [4].

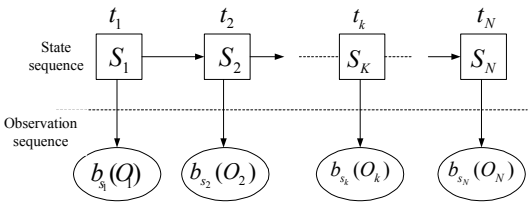

Fig. 1 Conventional HMM chain structure.

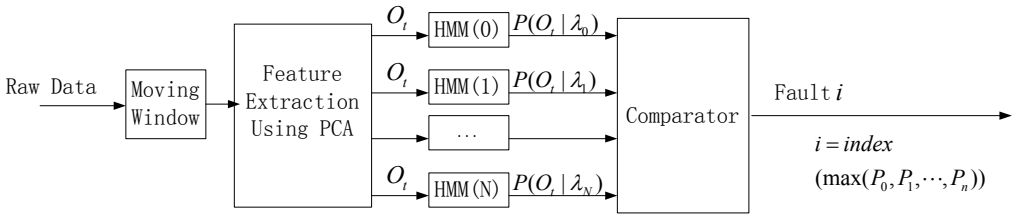

Fig. 2 Schematic diagram of fault diagnosis system.

IV. Structure of PCA-CHMM Based Fault Diagnosis Method

Same as other fault diagnosis methods, PCA-CHMM based fault diagnosis method also contains two mainly parts, feature extraction and pattern recognition. The particular attention of this study is to address,

--First, reduce many measurement variables to some principal components using PCA, and use the principal components as observation sequences.

--Second, use the different sampling rates and sampling number in the different kinds of observation sequences: training and testing, and classify various process operating conditions via CHMM.

The PCA-CHMM based fault diagnosis method developed in this study has a basic structure as shown in Fig. 2, where $\boldsymbol{\lambda}_0$ represents CHMM for normal operating condition, $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_N$ represent CHMMs for all faults, and $\boldsymbol{O}_t$ represents observation sequence for testing.

A. Feature Extraction Using PCA

It is almost impossible to use all the measurement variables directly for fault diagnosis in chemical engineering. In qualitative process trend analysis based on Neural Network, only some important variables are selected from all, and this method is well applied in a continuously well-stirred tank reactor (CSTR) [10] and a fluid catalytic cracking (FCC) process [6]. However many chemical processes like TE are complex, and it is almost impossible to select proper variables for fault diagnosis. There are a large number of variables and faults in these processes. And different faults may affect different variables. It is difficult to find a few common variables to explain all kinds of the information of process operating conditions. Since all the process variables are highly correlated, PCA will be used to reduce the large number of correlated variables to a small number of principal components in an optimal way without losing any important information. These principal components can be used as the feature sequences, which can indicate various kinds of process operating conditions. Note that PCA will only be a tool for feature extraction.

B. Moving Windows and CHMM Used for On-line Fault Diagnosis

A moving window is an indispensable technique to track dynamic data and widely used for on-line fault diagnosis. Sometimes it is important to select the proper time span of the moving window for fault diagnosis, which is the product of the sampling number (window length) in each moving window and time increment (sampling rate). If the window is chosen too small, one may capture process changes quickly, but the window may not contain enough information to sufficiently reflect the current process operating condition, thus leading to ambiguous classifications. Large window sizes can consider more information, but may lead to large time delays for the classification of various process operating conditions. The observation sequences $\boldsymbol{O}_{train}$ for training can have the different length from the observation sequences $\boldsymbol{O}_t$ for testing, if HMM is used for classification. The observation sequences $\boldsymbol{O}_{train}$ are longer for more information of process operating condition, whereas the observation sequences $\boldsymbol{O}_t$ are shorter for quicker fault diagnosis. The minimum length of the observation sequences $\boldsymbol{O}_t$, which is determined by the moving window length, is found by simulation.

The Shannon sampling theorem states that for a limited bandwidth (band-limited) signal with maximum frequency $f_{\max}$, the equally spaced sampling frequency f_s must be greater than twice of the maximum frequency $f_{\max}$ in order to have the signal be uniquely reconstructed without aliasing. Here the sampling rate in training data can also be different from the one in test data, if both of them are sufficiently high. The sampling rate in training data is relatively low for more information of process operating condition, whereas the sampling rate in test data is high for timely fault diagnosis.

Hidden Markov model method is mainly applied in the field of signal processing, and has been become a primary technique for speech recognition. In this study, CHMM is applied to classify various process operating conditions. The algorithm contains mainly five steps.

--First, a certain number of CHMMs are trained to

construct a database, including one CHMM corresponding to normal operating condition and other CHMMs corresponding to faults.

--Second, at time t, a limited number of data points are got from the raw data using moving window, and observation sequence $\boldsymbol{O}_t$ (principal components) is extracted via PCA.

--Third, the probabilities $P(\boldsymbol{O}_t \mid \boldsymbol{\lambda}_i)(i = 0,1\cdots,N)$ are calculated, which are the probabilities of the observation sequence $\boldsymbol{O}_t$, given all CHMMs $\boldsymbol{\lambda}_i$ in the database.

--Fourth, the maximum probability $P(\boldsymbol{O}_t \mid \boldsymbol{\lambda}_j)$ is found by comparing these probabilities, which indicates that the plant is running with fault $j(j = 0,1\cdots,N)$.

--Fifth, with time going on, the steps from second to fourth are repeated until we can make a correct fault classification.

V. Application

The Tennessee Eastman process simulator was created by the Eastman Chemical Company to provide a realistic industrial process for evaluating process control and monitoring methods. As a standard model, the Tennessee Eastman process simulator has been widely used by the process monitoring and diagnosis community as a source of data to estimate various fault detection and diagnosis methods [3]. The process consists of five major unit operations: a reactor, a product condenser, a vapor-liquid separator, a recycle compressor, and a product stripper. Two products are produced by two simultaneous gas-liquid exothermic reactions, and a byproduct is generated by two additional exothermic reactions. The control system used for dynamic simulations is the decentralized PID control system designed by McAvoy and Ye [7], which is shown in Fig. 3. The process has 12 manipulated variables, 22 continuous process measurements, and 19 composition measurements of reactor feed, purge gas and product.

In this study, the reference set consists of four patterns,

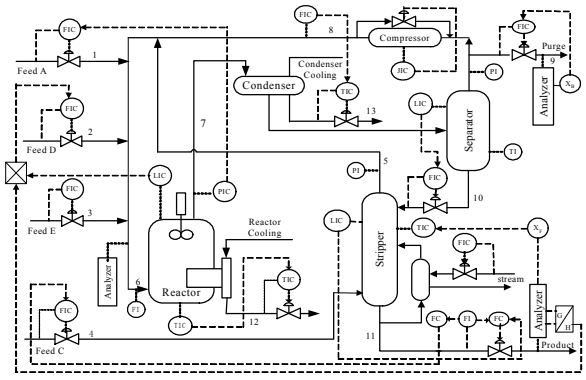

Fig. 3 A diagram of Tennessee Eastman simulator.

R_0, R_1, R_2, and R_3. The first pattern corresponds to the normal operating condition, whereas the other three patterns correspond to three major deterministic upsets, IDV(1), IDV(2), and IDV(7) [1]. A total of 41 variables is recorded every 6 minutes, and the details are shown in Table 1.

The patterns in the reference set in our case studies contain 41 measurement variables. However, not all 41 variables carry equally process information of variance for fault diagnosis purposes. Since the 41 variables are highly correlated with one another, PCA is used to reduce the dimension of the patterns in an optimal way.

TABLE I
Patterns in The Reference Set

Pattern	Fault	Type	Step size	Simulation time(minute)
R_0	Normal	——	——	480
R_1	IDV(1)	Step	+1.0	480
R_2	IDV(2)	Step	+1.0	480
R_3	IDV(7)	Step	+1.0	480

After PCA, the first four principal components, which can explain most information of process operating condition, are used as the observation sequence for training corresponding CHMM. Thus the observation sequence of each pattern in the reference set consists of four vectors with length of 80. Fig. 4a, 4b, 4c, and 4d show the first four principal components for R_0, R_1, R_2, and R_3 respectively.

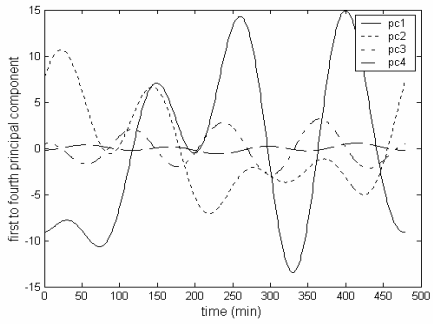

Fig. 4a Four principal components of R_0.

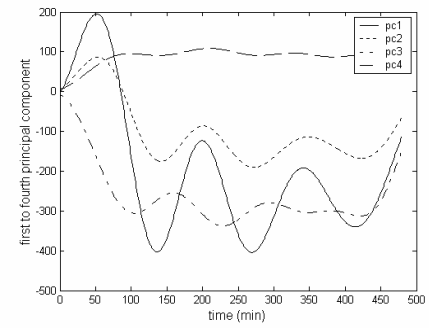

Fig. 4b Four principal components of R_1.

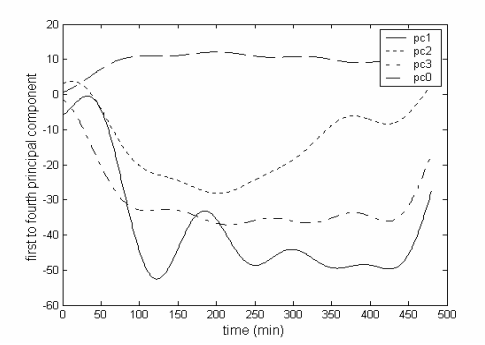

Fig. 4c Four principal components of R_2 .

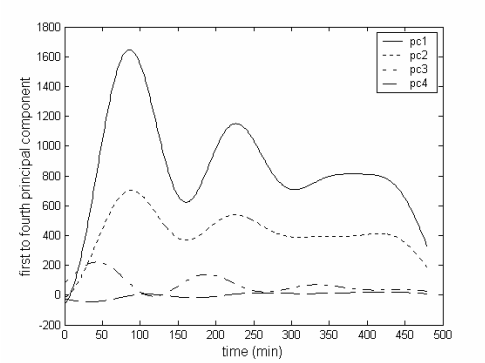

Fig. 4d Four principal components of R_3 .

There are four CHMMs need to be trained in this study, one for normal operation and other three for corresponding fault operations. For each reference pattern $R_i (i = 0,1,\cdots,3)$, the corresponding observation sequence is used to train the continuous hidden Markov model $\boldsymbol{\lambda}_i (i = 0,1,\cdots,3)$ via Baum-Welch method. Fig. 5 shows the results of training. From Fig. 5, we can see that the two training processes, R_0 and R_2 , are quite similar, which indicates that the dynamic response characteristic of pattern R_2 is close to normal, whereas the two other patterns, R_1 and R_3 , are quite different from normal operating condition.

There are also four patterns in the test set. All faults in the test set are identical to the faults of the reference set, and each test pattern begins with normal operation and introduces a fault after 480 minutes. A total of 41 variables is recorded every 1 minute (min). The details are shown in Table 2. The moving window is used for tracking dynamic data in the test set, and the window length should be as short as possible for quicker fault diagnosis. Here four windows with different length, 10, 20, 40 and 80, are used to test the effect of the window length on delay of diagnosis. Fig. 6a, 6b, 6c, and 6d show the results of four patterns in the test set with the window length of 10. The vertical coordinates represent $log(P(\boldsymbol{O}_t \mid \boldsymbol{\lambda}_i)(i = 0,1\cdots,3))$, the log probabilities of the observation sequence $\boldsymbol{O}_t$, given all four CHMMs in the database, and the horizontal coordinates represent time t . From these figures, we can see that various process operating conditions can be recalled correctly using the proposed fault diagnosis method. $log(P(\boldsymbol{O}_t \mid \boldsymbol{\lambda}_0)$ and $log(P(\boldsymbol{O}_t \mid \boldsymbol{\lambda}_2)$ are close in each figure, because the process operating conditions between R_0 and R_2 are quite similar, which mentioned above. For pattern T_4 , when fault IDV(7) is introduced at 480 min, the diagnostic performance is not so satisfying at the beginning (see Fig. 6d). Some random disturbance of process may worsen such performance. With time going on, the fault feature can be more discriminable, which is available of correct fault identification. Table 3 gives the diagnosis delay of different test patterns with the windows of different length, which shows that for the same test pattern with different length windows, the shorter of the window, the quicker of fault diagnosis. And for the same length window used for different patterns, the delay of diagnosis is shorter if the test pattern is closer to normal operation. In this study, if the window length is less than 10, the test patterns will not be recognized correctly.

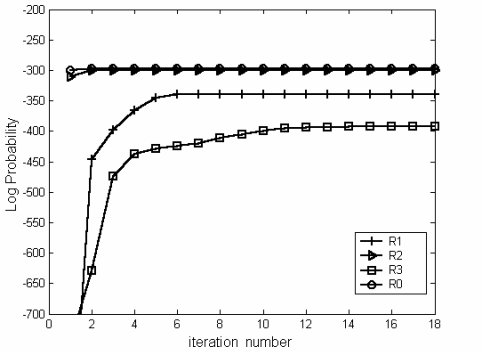

Fig. 5 Iteration process for training four CHMMs.

TABLE II
PATTERNS IN THE TEST SET

Pattern	Fault	Type	Step size	Fault occurs from/to (minute)
T_1	Normal	——	——	——
T_2	IDV(1)	Step	+1.0	480/960
T_3	IDV(2)	Step	+1.0	480/960
T_4	IDV(7)	Step	+1.0	480/960

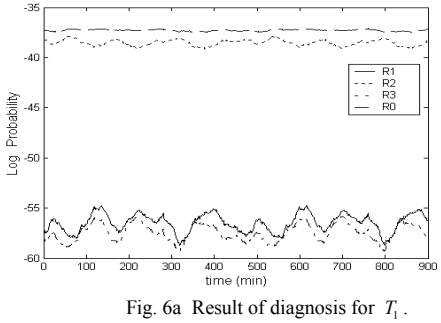

Fig. 6a Result of diagnosis for T_1 .

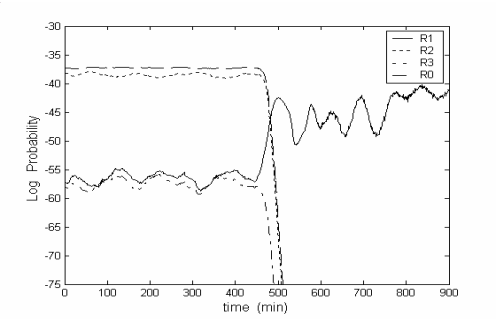

Fig. 6b Result of diagnosis for T_2 .

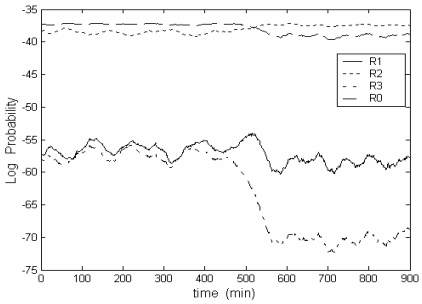

Fig. 6c Result of diagnosis for T_3 .

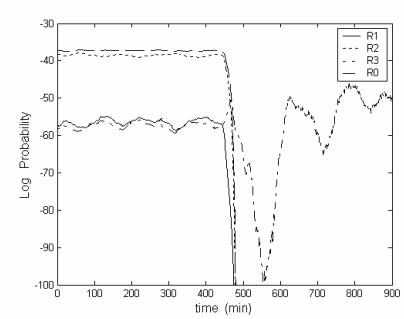

Fig. 6d Result of diagnosis for T_4 .

TABLE III
DIAGNOSIS DELAY ACCORDING TO WINDOW LENGTH

Window length	diagnosis time in T_2 (minute)	diagnosis time in T_3 (minute)	diagnosis time in T_4 (minute)
10	510	549	495
20	514	553	500
40	523	563	508
80	538	576	517

VI. CONCLUSION

An approach combing PCA with HMM for on-line fault diagnosis has been described. The use of PCA as a tool for feature extraction greatly reduces the dimensions of the patterns and results in large improvement in the discriminatory power of the classifier. The CHMM, which not only can capture the serial correlations in the feature sequences of the patterns, but also can take into account the process random factors, is used to classify various process operating conditions. It is very important to train a certain number of CHMMs accurately. These CHMMs construct a database, including one CHMM corresponding to normal operating condition and other CHMMs corresponding to all faults. The sampling rate and the sampling number of patterns in the reference set, are relatively low and more so that the observation sequences of patterns can contain sufficient process operating information and be used to train more accurate CHMMs. The moving window is used to tracking dynamic data for on-line fault diagnosis. The moving windows with shorter length and the high sampling rate of the test patterns are selected for quicker fault diagnosis. The proposed fault diagnosis method is demonstrated in Tennessee Eastman simulator, and results show that it can recall single faults correctly.

REFERENCES

[1] A. Kassidas, P. A. Taylor, and J. F. MacGregor, "Off-line diagnosis of deterministic faults in continuous dynamic multivariable processes using speech recognition methods," *J. Proc. Cont*, vol. 8, no. 5, pp. 381-393, 1998.

[2] A. T. Vemuri, and M. M. Polycarpou, "Neural-network-based robust fault diagnosis in robotic systems," *IEEE Transaction on Neural Networks*, vol. 8, no. 6, pp. 1410-1420, 1997.

[3] E. L. Russell, L. H. Chiang, and R. D. Braatz, "Fault detection in industrial processes using canonical variate analysis and dynamic principle component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 51, pp. 81-93, 2000.

[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[5] R. Li, J. H. Olson, and D. L. Chester, "Dynamic fault detection and diagnosis using neural networks," *Intelligent Control*, vol. 2, pp. 1169-1174, 1990.

[6] S. H. Yang, B. H. Chen, and X. Z. Wang, "Neural network based fault diagnosis using unmeasurable inputs," *Engineering Applications of Artificial Intelligence*, vol. 13, pp. 345-356, 2000.

[7] T. J. McAvoy, and N.Ye, "Base control for the Tennessee Eastman problem", *Computers and Chemical Engineering*, vol. 18, no. 5, pp. 383-413, 1994.

[8] V. Venkatasubramanian, R. Rengaswamy, and K. Yin, "A review of process fault detection and diagnosis (Part I)," *Computers and Chemical Engineering*, vol. 27, pp. 293-311, 2003.

[9] W. Sun, A. Palazoglu, and J. A. Romagnoli, "Detecting abnormal process trends by wavelet-domain hidden Markov models," *AIChE. J*, vol. 49, no. 1, pp. 140-150, 2003.

[10] Y. Maki, and K. A. Loparo, "A neural-network approach to fault detection and diagnosis in industrial process," *IEEE Trans. on control systems technology*, vol. 5, no. 6, pp. 529-541, 1997.