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

Online Statistical Monitoring and Fault Classification of the Tennessee Eastman Challenge Process Based on Dynamic Independent Component Analysis and Support Vector Machine

Karim Salahshoor*, Fariborz Kiasi*

**Department of Automation and Instrumentation,*
Petroleum University of Technology, Tehran, Iran (email: salahshoor@ put.ac.ir, fkiasi@ put.ac.ir)

Abstract: This paper presents a new online statistical monitoring based on dynamic independent component analysis (DICA) to detect the Tennessee Eastman challenge process faults. The proposed method employs dynamic feature extraction approach to capture most of the inherent dynamic fault information. This leads to an efficient fault detection with superior performance compared to independent component analysis (ICA) approach in both detection rate and number of false alarms. A new statistic measure has been introduced to enhance the monitoring capabilities of ICA and DICA. An approach based on cumulative percent variance (CPV) has been incorporated to mechanize the selection of required number of independent components in both ICA and DICA online monitoring methods. To choose the best time-lag order for each fault dynamic model in the DICA augmented data matrix, a multivariate auto regressive exogenous (ARX) model structure has been adopted by validating the minimum Akaike's information criterion (AIC) index. An online procedure based on a multi-class support vector machine (SVM) with Gaussian kernel function, being set by sub-optimal width parameters, is employed to classify and isolate each fault. The SVM uses one against all (OAA) algorithm for fault classification and sequential minimization optimization (SMO) to solve the classification problem. Performances of the developed process monitoring methods (ICA-SVM, DICA-SVM) are evaluated on the Tennessee Eastman challenge process (TE).

Keywords: statistical monitoring, fault detection, Tennessee Eastman process, independent component analysis (ICA), dynamic independent component analysis (DICA), support vector machine (SVM)

1. INTRODUCTION

Occurrence of any abnormal situation or fault in modern chemical plants can lead to serious safety, environmental and economical implications. The exploitation of the enormous and highly correlated operational data by simple visual inspection to detect any possible fault is a difficult or impossible task. As a consequence, automatic online process monitoring approach is gaining great importance in the large complex chemical, oil and gas plants. Early and reliable fault detection and diagnosis is not only desirable but also crucial to minimize down-time, increase the operation safety, and to reduce the production costs. Research efforts for more than a decade have focused on models created with process history data based on statistical analysis tools. Principle component analysis (PCA) is the most widely data-driven technique for monitoring industrial processes. It is based on orthogonal decomposition of the covariance matrix of the process variables along directions that explain the maximum variation of the observed data. However, this method ignores the serial correlation among measurements at different times and assumes the monitored latent variables to be normally distributed. These characteristics limit the monitoring capabilities of the conventional PCA-based multivariate statistical approaches. Several extensions of the PCA have been developed in the literature (Ku et al., 1995, Nomikos et al., 1994; Wold et al., 1996; Bakhshi, 1998) to overcome these limitations. ICA is an emerging statistical technique which can extract basic underlying informative factors or independent components (ICs) from multivariate observed data. ICs can reveal more useful monitoring information from the observed process data than principle components (PCs) in the PCA-based monitoring approaches. Because, statistically speaking, PCA procedure can only impose independence up to second order statistics information (i.e. mean and variance) while ICA has no orthogonality constraint and hence accomplishes higher order statistics. Lee et al (2003) proposed a new statistical monitoring method that uses the ICA methodology. They proposed to use I^2, I_e^2 and SPE statistics as monitoring charts. Lee et al. (2004) introduced DICA as a more powerful monitoring approach for dynamic processes.

In this paper, DICA approach is used to present an integrated framework for online monitoring of Tennessee Eastman plant. To enhance the monitoring capability of ICA and

DICA methods, a new statistic measure SPE_e is proposed to take care of monitoring the excluded part of the independent vectors, which has not been captured by the main dominant part. An ARX reference model is used as a preliminary modelling study on the individual fault dynamics to determine the required number of time-lagged variables for developing the DICA models. The results are validated by Akaike information criterion (AIC) index. After the fault modelling step, the ICA procedure is performed on the lagged observed variables, stacked in the data matrix, using the FastICA software algorithm. The paper uses the statistics derived from the normal operating condition in the process data training data set to determine the 99% confidence limits for the four monitoring charts, I^2, I_e^2, SPE and SPE_e, by the kernel density estimation. Then, an automatic selection procedure is adopted to select the appropriate number of the ICs using a cumulative percent variance (CPV) measure. Finally, the SVM approach is utilized to classify and isolate the TE fault sources in an online manner. The paper is organized as follows. In Section 2, ICA and DICA monitoring approaches are developed. The ICA DICA monitoring approaches are then applied to TE process for fault detection. Following a brief review of SVM theory in Section 4, the proposed ICA and DICA combined with the SVM are utilized for TE fault diagnosis. Finally, some important conclusion remarks will be summarized in Section 5.

2. DEVELOPMENT OF ICA AND DICA

2.1 ICA approach

ICA is a generative model which can describe how the observed data are generated by the process of mixing the hidden ICs. In the ICA algorithm, it is assumed that the measured *d*-dimensional vector at time instant *k*, i.e., $x(k) = [x_1(k),...,x_d(k)]^T$ can be expressed as a linear combination of $m \le d$ hidden ICs, denoted by $s(k) = [s_1(k),...,s_m(k)]^T$, which can be represented by the following model:

$$x(k) = As(k) + e(k) \tag{1}$$

Where $A \in R^{d \times m}$ is an unknown full-rank matrix, called the mixing matrix and *e* is the residual or fitting error vector. The basic problem of ICA is to estimate the original component $s(k)$ or to estimate the unknown mixing matrix A from the measured data vector $x(k)$. Alternatively, the main objective of ICA is to estimate a demixing matrix where $W \in R^{m \times d}$ so that components of the reconstructed data vector $\hat{s}(k)$, given by $\hat{s}(k) = Wx(k)$, become as independent of each other as possible (i.e. $E(s(k)s^T(k)) = I$). The problem of estimating a full-rank matrix A can be reduced to the problem of estimating $s(k)$, as follows:

$$\hat{s}(k) = B^T Q x(k) \tag{2}$$

Where B is an orthogonal matrix (i.e. $BB^T = I$) and Q is the whitening matrix, given by $Q = D^{-1/2}V^T$, where V is the orthogonal matrix of eigenvectors of the data covariance matrix, $R_x = E(x(k)x^T(k))$, and D is the diagonal matrix of its corresponding eigenvalues. Comparing (2) and equations stated, leads to the following expression:

$$W = B^T Q \tag{3}$$

Hyvärinen (1999) presented a fast and robust fixed-point algorithm, called as FastICA, to perform ICA which entails maximizing the negentropy under the constraint of $\|b_i\| = 1$ (i.e., *i*th column of *B*). Finding *B*, the demixing matrix *W* is obtained from (3).

2.2 DICA approach

ICA approach assumes implicitly that the observations at one time instant are statistically independent of the observations at past time instances. This leads to neglecting of any useful serial correlation information in the observation. One simple approach to address this issue is to augment each observation vector with the previous *l* observations and stacking them in the data matrix as follows:

$$X(l) = \begin{bmatrix} x_k^T & x_{k-1}^T & \dots & x_{k-l}^T \\ x_{k+1}^T & x_k^T & \dots & x_{k+1-l}^T \\ x_{k+2}^T & x_{k+1}^T & \dots & x_{k+2-l}^T \\ \vdots & \vdots & \vdots & \vdots \\ x_{k+p}^T & x_{k+p-1}^T & \dots & x_{k+p-l}^T \end{bmatrix} \tag{4}$$

Where x_k^T denotes the *d*-dimensional observation vector in the training set at time interval *k*, *(p+1)* is the number of samples and *l* is the number of time-lagged measurements by performing the ICA algorithm on the augmented data matrix in (4). A DICA model is obtained which can extract ICs based upon both the cross-correlated and auto-correlated properties of the observed variables. The only major problem associated with the DICA approach is the appropriate selection of the number of the required time lags. Because, the number of time lags included in the DICA model may substantially affect its monitoring performance. In this paper, a procedure is presented which follows the idea of selecting the number of lags *l* to minimize the AIC criterion based on the following final prediction error (FPE) measure, using as ARX reference model:

$$FPE = V\left(\frac{1 + d/N}{1 - d/N}\right) \tag{5}$$

Where V is the loss function, d is the estimated parameters and N denotes the number of estimation data.

2.3 Process monitoring with ICA and DICA

In order to perform online process monitoring, the measured variables should be continuously analyzed to detect faults. To implement the process monitoring, the monitoring statistics of ICA or DICA should be estimated. Three types of statistics (I^2, I_e^2, SPE) have been already proposed by Lee et al. (2003) for process monitoring. The ICA or DICA model is determined based on the historical data collected during the normal operating condition (NOC) using the FastICA algorithm. Then, future process behaviour is compared against this normal or in-control model representation. To reduce the data dimensions, however, a few rows of W are only selected to make a reduced dominant part W_d . As a consequence, by collecting new data, $x_{new}(k)$ at every time instant k, the new decomposed independent data vectors can be obtained as follows in terms of the dominant and excluded parts:

$$\hat{s}_{newd}(k) = W_d x_{new}(k) \tag{6}$$

$$\hat{s}_{newe}(k) = W_e x_{new}(k) \tag{7}$$

Where W_e denotes the excluded part of W. Then, the three statistical monitoring measures are defined as follows:

$$I^2(k) = \hat{s}_{newd}(k)^T \hat{s}_{newd}(k) \tag{8}$$

$$I_e^2(k) = \hat{s}_{newe}(k)^T \hat{s}_{newe}(k) \tag{9}$$

$$SPE(k) = (x_{new}(k) - \hat{x}_{newd}(k))^T (x_{new}(k) - \hat{x}_{newd}(k)) \tag{10}$$

Where:

$$\hat{x}_{newd}(k) = Q^{-1} B_d \hat{s}_{newd}(k) = Q^{-1} B_d W_d x_{new}(k) \tag{11}$$

Noting that B_d is a reduced matrix of B whose indices correspond to the indices of W_d and can be computed directly by $B_d = (W_d Q^{-1})^T$. Similarly, the excluded or non-dominant part of B can be computed as $B_e = (W_e Q^{-1})^T$. Similar to the above reasoning for the *SPE* measure, suggested by Lee et al. (2003), another new statistical monitoring measure is proposed in this paper to take care of monitoring the excluded part of the independent vectors, defined by:

$$SPE_e(k) = (x_{new}(k) - \hat{x}_{newe}(k))^T (x_{new}(k) - \hat{x}_{newe}(k)) \tag{12}$$

Where:

$$\hat{x}_{newe}(k) = Q^{-1} B_e \hat{s}_{newe}(k) = Q^{-1} B_e W_e x_{new}(k) \tag{13}$$

Where $\hat{x}_{newd}(k)$ represents the main data captured by the dominant part of ICA or DICA model while $\hat{x}_{newe}(k)$ indicates the excluded part of the data sample time k. Thus, the new $SPE_e(k)$ is a useful measure to monitor the variation due to the excluded part which has not been captured by the main dominant part. Once the ICA or the DICA model has been developed in terms of the four statistics (I^2, I_e^2, SPE and SPE_e), any departure from the process NOC can be detected using the corresponding confidence limit values.

3. ONLINE FAULT DETECTION OF THE TE PROCESS USING ICA AND DICA

3.1 The TE process description

The Tennessee Eastman (TE) challenge process is a plant-wide process control problem which has been proposed by Downs and Vogel (1993) as a hypothetical challenge test problem for control and monitoring approaches. The flowsheet diagram of the TE process is depicted in Fig.1. The original process has 12 manipulated variables and 41 measurements (22 continuous process measurements and 19 composition measurements). The details on the process description are well explained in a book by Chiang et al (2001). In this research study, the same simulation data generated by Chiang et al. (2001) and used by Lee et al. (2004) has been employed for the sake of comparison. A total of 33 variables including manipulated variables and 22 measured variables as listed by Lee et al. (2004), have been selected to be used as monitoring variables. The study does not include the 19 composition measurements data to provide a more realistic monitoring problem. The set of used programmed faults, i.e. faults 1-21 has been introduced in Table 1. Each fault record consisted of 480 and 960 observations in the training and testing data sets, respectively. All faults in the testing data set were introduced from time sample instant of 151 (the instant of fault occurrence has been changed in this work).

3.2 Time-lag order determination in the DICA monitoring approach

To apply the DICA monitoring approach, the number of required time-lagged observation (l) in (4) should first be determined. For this purpose, the following multivariate ARX model structure is considered for each individual TE fault dynamic modelling:

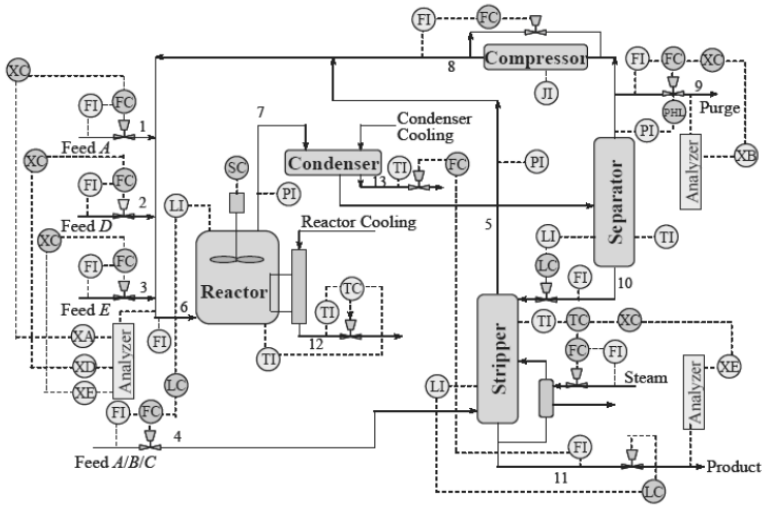

Fig. 1 Control system for the Tennessee Eastman process

Table 1. Process faults for Tennessee Eastman process

No	Fault	Type
1	A/C feed ratio, B composition constant (str.4)	Step
2	B composition, A/C feed ratio constant (str.4)	Step
3	D feed temp. (str.2)	Step
4	Reactor Cooling water inlet temp.	Step
5	Condenser cooling water inlet temp.	Step
6	A feed loss (str.1)	Step
7	C header press. Loss-reduced availability (str.4)	Step
8	A,B,C feed co position (str.4)	Random variation
9	D feed temp. (str.2)	Random variation
10	C feed temp. (str.4)	Random variation
11	Reactor Cooling water inlet temp.	Random variation
12	Condenser cooling water inlet temp.	Random variation
13	Reaction kinetics	Slow drift
14	Reactor cooling water valve	Sticking
15	Condenser cooling water valve	Sticking
16	Unknown	
17	Unknown	
18	Unknown	
19	Unknown	
20	Unknown	
21	Unknown	

$$y_i(t) = bias_i + \sum_{j=1}^{ny} \alpha_j y_i(t-j) + \sum_{j=1}^{n1} \beta_{j1} u_1(t-j) + \ldots$$
$$+ \sum_{j=1}^{n11} \beta_{j11} u_{11}(t-j) \qquad ;i = 1,\ldots,22 \tag{14}$$

In this way, each of the 22 measured output time responses $y_i(t)$ can be described by linear combination of their previous outputs $y_i(t-j)$ and the previous manipulated inputs $u_{jm}(t-j)$ where $1 \le m \le 11$. Assuming that $L_u = \max\{ny, n1,\ldots,n11\}$, (18) can be rewritten as follows:

$$y_i(t) = bias_i +$$
$$\sum_{j=1}^{L_u} \left\{ \alpha'_j y_j(t-j) + \beta'_{j1} u_1(t-j) + \ldots + \beta'_{j11} u_{11}(t-j) \right\} \tag{15}$$

Thus, without the loss of generality, the modelling problem now leads to estimation of a time-lag order (L_u) and two sets of model coefficients (α'_j, β'_{jn}) in (15). In this study, the search for obtaining the best number of lags (L_u) was conducted in order interval of 1 through 8 (i.e. $1 \le L_u \le 8$). Higher orders are practically undesirable in model identification application since it requires more computational time and might be in the favour of noise enhancement rather than the useful fault diagnostic data modelling. For any time-lag order, the ARX model parameters are estimated for each specific fault time-series pattern by the least squares solution of the over-determined system of equations, resulting from applying the relevant data samples of the output and input available in the TE training data set. In each test, the model mean squared error (MSE) and the Akaike's final prediction error (FPE) are computed. Finally, the best time lag order (L_u) for each fault model is determined on the basis of the minimum AIC

of all the performed time-lag tests. Table 2 summarizes the best obtained results for identification of all the fault models. However, a fixed number of time-lag order (L_u) should be selected for each output and manipulated variable to implement the DICA fault model in the proposed online process monitoring application. Table 3 shows the resulting average and selected time-lag orders.
It should be noted that the order of lags for the manipulated variables is set to the average of all lags in Table 3, resulting in selection of 4 lags for the all 11 manipulated variables.

3.3 Implementation of the proposed ICA and DICA monitoring

To obtain the ICA and DICA models for online TE process monitoring, the number of dominant ICs should be selected. Lee et al. (2003) suggested a graphical method to determine the number of dominant ICs by looking at the resulting ICs contribution bar graph, which is not appropriate for an online monitoring approach. Therefore, an automatic selection method is required for online monitoring purposes. In this paper, a measure based on the following cumulative percent variance (CPV) (Malinowski, 1991), captured by the first I_k independent components, is employed:

$$CPV(I_k) = \frac{\sum_{j=1}^{I_k} \lambda_j}{\sum_{j=1}^{m} \lambda_j} \tag{16}$$

Where $\lambda_j's$ are the eigenvalues of the covariance matrix R_x, sorted in the decreasing order:

$$R_x = \frac{\bar{x}^T . \bar{x}}{n-1} \tag{17}$$

$\bar{x}$ denotes the normalized version of the data matrix x corresponding to n samples of m measured variables. In this paper, the number of dominant ICs is chosen when the CPV measure reaches a predetermined limit of 97%, leading to 78 ICs for DICA monitoring. For ICA monitoring, however, 9 ICs where selected for the sake of comparison with the Lee et al. (2004). The 99% confidence limits for each monitoring statistics were also determined by the kernel density estimation method from the NOC in the training data set. This is due to the fact that there is no a priori knowledge of data in the testing set at this stage. However, Lee et al. (2004) adjusted the confidence limit of each statistic to its tenth highest value for the NOC of the testing data set. Table 4 shows the resulting false alarm rates due to the application of the proposed ICA and DICA monitoring procedure in the testing data set.
Comparing the corresponding results with those recorded in Lee et al. (2004), shows better false alarm rates achievements. Table 4 summarizes the obtained detection rates for all the 21 TE faults. In comparison with the corresponding Lee et al. (2004), results, the DICA method shows much better results with higher detection rates in almost all the TE faults especially for fault numbers 3, 9, 10, 11, 15, 16, 19 and 20 which demonstrates a considerable difference. This is mainly due to the fact that the proposed DICA uses the appropriate number of time-lagged measured variables information to extract the process dynamicity. The resulting DICA charts show that all the statistic measures can successfully detect all the TE faults except for the difficult faults 3, 9 and 15. The monitoring results for fault 5 have been illustrated in Fig. 2. Comparing the obtained monitoring charts with the corresponding charts presented by Lee et al. (2004) demonstrates that the effect of this fault is magnified more in our statistic measures. As shown, the new SPE_e statistic chart in Fig. 2 is an additional useful fault detection tool which can provide efficient information quite similar to the *SPE* statistic measure. Fig. 3 demonstrates the monitoring results for fault 10. Comparing the obtained DICA results with those recorded in Lee et al. (2004), shows a better fault detection rates for all the statistic measures. The detection rate for I^2 statistic is 95.5% while that of the Lee et al. (2004) is 87.84%. Similarly, the detection rate of I_e^2 and *SPE* statistics have increased from 94.36% and 74.19% to 96% and 89.88%, respectively. Inspecting the corresponding *SPE* monitoring chart of Lee et al. (2004) reveals that although the chart is able to detect the start of fault occurrence successfully, it can not show the persistent presence of the fault at many other samples, because the chart is below the confidence limit for those samples, giving the process operator an incorrect picture of the process status. While, all the obtained DICA monitoring charts in Fig. 3 successfully illustrate the detection effect up to the end of the processing time interval.

4. ONLINE TE PROCESS FAULT DIAGNOSIS USING SVM

Observing the multivariate monitoring charts (I^2, I_e^2, SPE and SPE_e) simply indicates the presence of a process fault. These measures can not give any information about the root-cause fault diagnostics. Thus, once a fault is detected by the statistical ICA or DICA monitoring approach, a fault classification is required to perform fault diagnostics. In this paper, a SVM technique is presented for this classification purpose.

4.1 A brief introduction to SVM

SVM is a relatively new computational learning method which is gaining more popularity in machine learning community due to its excellent generalization ability. SVM is based on the so-called structural risk minimization principle. The basic idea of applying SVM for solving classification problem can be stated briefly in two steps. First, SVM maps the input space to a higher dimensional feature space through the use of a nonlinear kernel function. Then, it seeks for an optimized linear division within the feature space. In this study, a Gaussian radial basis function (RBF), i.e. $\varphi(x_i) = \exp\{-\|x - x_i\|^2 / \sigma^2\}$, was used as the kernel function which maps the input data vectors (x_i) onto

Table 2. Number of lags determined by modelling
(Fault no. in rows and output number in columns)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
F1	6	3	6	3	1	2	6	4	3	5	8	1	8	1	1	6	1	7	8	6	4	8
F2	5	7	3	2	6	1	6	3	3	3	7	1	4	1	1	6	1	7	3	3	4	8
F3	5	3	3	2	1	1	4	1	6	4	5	1	6	1	1	6	1	4	2	3	2	4
F4	3	3	3	1	1	1	4	1	4	3	2	1	4	1	1	8	1	3	3	5	2	2
F 5	3	4	3	6	1	1	8	7	2	5	8	1	5	1	1	4	1	4	4	8	3	8
F 6	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
F 7	3	2	1	4	1	2	5	8	6	3	6	1	4	1	1	2	1	6	7	5	4	8
F 8	5	4	3	3	1	4	7	3	4	3	5	1	8	1	1	7	1	5	8	4	6	5
F 9	4	4	3	5	1	1	7	2	4	3	2	1	7	1	1	7	1	4	2	2	3	1
F10	3	3	3	2	1	1	4	2	5	5	4	1	5	1	1	4	1	2	3	4	3	1
F11	5	3	2	5	1	1	4	1	3	6	5	1	4	1	1	4	1	6	2	3	4	5
F12	3	2	8	4	1	3	3	4	5	5	3	1	5	1	2	5	2	6	7	5	3	2
F13	4	3	3	3	1	1	8	7	2	3	8	1	8	1	2	4	1	8	8	7	3	8
F14	3	8	3	1	2	2	8	6	4	5	2	1	8	1	2	4	1	5	2	8	6	1
F15	5	3	3	2	1	1	2	2	2	5	2	1	4	1	1	4	1	4	1	3	2	3
F16	5	3	3	3	1	1	4	3	4	3	3	1	4	1	1	4	1	3	3	8	3	3
F17	4	3	3	6	1	1	6	2	2	3	7	1	6	1	1	6	1	7	2	8	3	8
F18	7	6	7	8	6	8	3	8	3	8	7	1	4	7	6	8	2	5	7	8	5	5
F19	3	4	3	3	1	3	5	1	6	5	8	1	2	1	1	5	1	3	1	3	2	5
F20	7	2	3	7	1	1	4	3	4	7	6	2	2	1	1	7	2	5	3	3	3	4
F21	3	3	3	1	1	1	3	2	2	4	2	1	5	1	1	3	1	3	1	3	3	3

Table 3. Number of lags selected for the DICA monitoring

Variable No.	**Average Lag**	**Selected Lag**	**Variable No.**	**Average Lag**	**Selected Lag**
1	4.4762	5	12	1.381	2
2	3.8571	4	13	5.2857	6
3	3.6667	4	14	1.619	2
4	3.7619	4	15	1.7143	2
5	1.8571	2	16	5.3333	6
6	2.1429	3	17	1.4762	2
7	5.1905	6	18	5	5
8	3.7143	4	19	4.0476	4
9	3.9048	4	20	5.0952	5
10	4.5714	5	21	3.619	4
11	5.1429	6	22	4.7619	5

Table 4. Detection rates percentage of ICA and DICA statistics for the Tennessee Eastman process

Faults	ICA I^2	ICA I_e^2	ICA SPE	ICA SPE_e	DICA I^2	DICA I_e^2	DICA SPE	DICA SPE_e
1	99.75	100	99.5	99.75	99.88	99.75	99.88	99.75
2	98.00	98.62	98.12	98.00	98.88	98.75	97.75	98.25
3	0.63	8.13	8.50	0.63	2.25	7.50	7.12	3.38
4	61.88	100	90.75	57.63	100	100	100	97.12
5	100	100	100	99.88	100	100	100	100
6	100	100	100	100	100	100	100	100
7	91.13	100	100	88.12	100	100	100	100
8	97.00	98.38	98.38	95.37	98.25	98.38	97.88	98.00
9	0.63	6.25	5.75	0.37	2.62	7.25	3.88	5.00
10	80.25	87.38	76.75	68.63	95.50	96.00	89.88	89.88
11	48.75	78.13	71.00	41.50	93.13	95.13	76.75	75.25
12	99.75	99.88	99.75	99.12	99.88	99.88	99.88	99.88
13	94.63	95.25	94.87	94.63	96.00	96.13	95.75	95.50
14	99.88	100	100	99.88	99.88	100	100	99.88
15	1.87	16.75	12.13	0.50	25.75	17.13	9.88	7.00
16	74.00	91.75	69.75	58.63	97.38	98.25	91.50	88.12
17	84.38	96.25	91.25	84.00	97.88	97.88	97.12	96.63
18	89.88	90.13	89.50	89.62	90.87	91.00	90.38	91.13
19	46.50	93.25	43.75	29.25	99.88	99.88	88.75	86.00
20	88.12	86.88	72.75	71.13	91.13	91.50	80.37	81.87
21	45.37	63.62	50.50	25.25	53.37	62.50	46.25	42.38
False Alarm	0.13	4.5	3.62	0.37	1.25	5.12	2.75	2.38

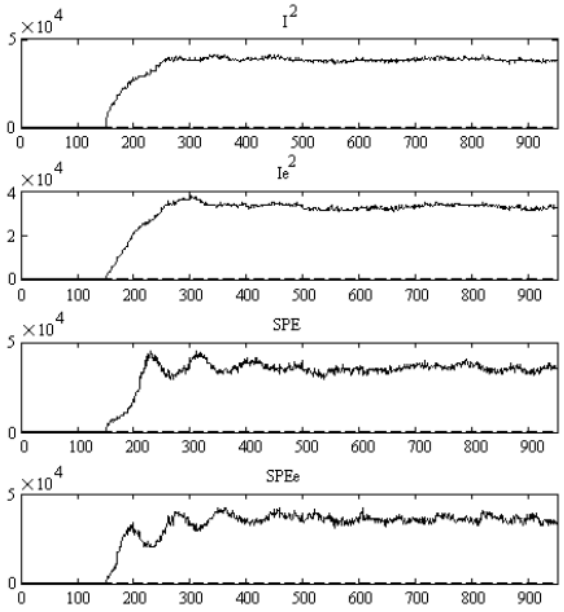

Fig. 2.Monitoring results of fault 5 by DICA

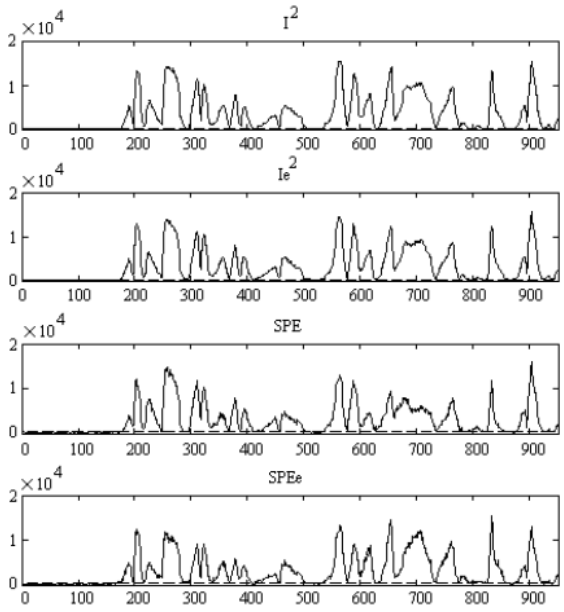

Fig. 3.Monitoring results of fault 10 by DICA

a high-dimensional feature space where x_i is called the support vector. The optimal hyperplane separating the ICA transformed data $\{(x_i, y_i)\}_{i=1}^{N}$ (where $x_i \in R^d$ is the dominant ICs and $y_t \in \{-1,+1\}$ is known as binary class or target) can be obtained by solving the following constrained quadratic problem (QP) (Cristianini et al., 2000):

$$\text{Min } \frac{1}{2}\|W\|^2 \tag{18}$$

Subject to: $y_i(W^T\varphi(x_i)+b) \geq 1$; $i=1,..,N$

Thus, the SVM problem has been formulated as an optimization problem to find the weighting vector (W) and the bias term (b) so that the margin between the two classes is maximized. In TE process, however, there are several classes of faults to be diagnosed. In this paper, the one against all (OAA) approach is used to implement the necessary multi-class fault classification objective.

4.2 Implementation of online TE fault diagnosis using multi-class SVM classifier

In this work, dominant independent data vectors $\hat{s}_i$ (i.e., rows of the matrix $\hat{s}_{newd}$) corresponding to the first 50 consecutive samples after the fault occurrence are used as extracted features for the classification purposes. Training the multi-class SVM classifier mainly includes initialization of parameters such as the width σ of the kernel function $\varphi(x_i)$. It is not known beforehand which values of σ are the best. In this paper, a direct searching algorithm was used to find the best σ values in a typical selected interval of [0,5].

The following performance function (PF^i) was employed as an assessment measure for the ith TE process fault:

$$PF^i = N^i{}_c - \sum_{j=1}^{18} N_m^j \quad ;i=1, 2,...,18\,,\ j \neq i \tag{19}$$

Where N_c^i refers to the number of correctly classified ith TE fault samples and N_m^j indicates its misclassified samples. In this work, the searching algorithm was repeated six times to find the sub-optimal σ values. Then, eighteen SVMs were trained with the corresponding training data set to generate the final classifier models based on the OAA approach. Then, all the 18 trained SVMs were tested to evaluate their online classification performance using the fault testing data set. A comparative result of this study is presented in Table 5. As shown, the best classification rates were found in the case of the proposed DICA-SVM approach. This characteristic has shown itself in the number of required support vectors for fault classification where DICA needs fewer number of SVs than ICA in majority of cases. The total average computation time for each sample in ICA-SVM approach, including the average fault detection time and the average fault diagnosis, is 0.0363 seconds, whereas this figure in DICA-SVM approach is equal to 0.0851 seconds. The calculations reported in this paper are performed in T2500@ 2.00 GHz Intel Centrino Duo processor with 2.00 GB RAM. The simulation results indicate that DICA-SVM is able to diagnose 11 out of 18 faults without any missed rates while that of ICA-SVM is 8 out of 18 faults. The DICA-SVM has no false alarm rates for 15 out of 18 faults whereas ICA-SVM approach has no false alarm rates for 11 out of 18 faults.

Table 5. SVM training and validation data set properties

	ICA-SVM					DICA-SVM				
	σ	Training CR %	Testing CR %	No. of SVs	CT In sec.	σ	Training CR %	Testing CR %	No. of SVs	CT In sec.
1	0.8	100	100	131	.0018	2	100	100	48	.0017
2	0.52	100	100	269	.0024	1.12	100	100	180	.0025
4	0.96	100	100	21	.0013	2	99.56	98	27	.0026
5	0.65	100	100	226	.0022	1.76	100	100	168	.0048
6	1.92	100	100	50	.0027	4.8	100	100	57	.0018
7	0.68	100	100	220	.0022	2.97	100	100	102	.0021
8	0.6	100	100	220	.0022	1.38	100	100	95	.002
10	0.64	98.89	94	175	.0020	2.42	99.78	98	128	.0023
11	0.9	98.67	94	44	.0014	2.15	99.56	96	75	.002
12	1.15	99.78	98	93	.0017	2.9	100	100	95	.0019
13	2.3	100	100	24	.0014	1.3	100	100	47	.0018
14	1.04	98.89	96	110	.0017	2.78	100	100	127	.0025
16	0.63	99.11	92	194	.0020	1.53	99.78	98	189	.0029
17	1.2	99.56	96	91	.0016	3.7	99.33	96	81	.0019
18	1.14	99.56	96	92	.0016	3.97	100	100	73	.0019
19	0.52	98.44	90	229	.0023	1.53	98.89	92	211	.0029
20	0.64	99.11	92	210	.0021	1.78	99.11	92	161	.0027
21	0.4	99.13	96	33	.0014	1.35	100	100	31	.0016

CT: Classification Time, CR: Classification Rate

5. CONCLUSIONS

A new online statistical approach has been proposed in this paper to detect and diagnose the TE process faults. In order to enhance the monitoring capabilities of the ICA and DICA, a new statistic measure SPE_e has been presented to take care of monitoring the excluded part of the independent vectors, which has not been captured by the main dominant part. A fault modelling procedure based on a multivariate ARX reference model and the AIC evaluation measure was developed to select the appropriate number of time-lagged measured variables in the DICA augmented data matrix. To implement the online process monitoring, a method based on the CPV measure was utilized to choose an appropriate number of ICs for DICA monitoring approach so as to capture a predetermined variance limit of 97% in the TE fault data. The developed ICA and DICA monitoring methods were evaluated on the TE process plant for fault detection purposes. In comparison with the corresponding Lee et al. (2004) results, the DICA method shows much better results with higher detection rates for almost all the TE faults especially the faults 3, 9, 10, 11, 15, 16, 19 and 20 which demonstrate considerable achievements. A multi-class SVM classifier was developed based on the OAA approach for online TE fault diagnosis. The SVM classifier uses the Gaussian RBF kernel functions with sub-optimal width parameters being found by direct searching algorithm to map the fault ICs patterns to a sufficiently higher dimensional feature space. The comparative studies done on 18 TE process faults, demonstrates the superiority of the DICA-SVM to the ICA-SVM in terms of classification rates, number of support vectors, misclassification rates and missed faults evaluating measures.

REFRENCES

Bakshi, B.R. (1998). Multiscale PCA with application to multivariate statistical process monitoring. *American Institute of Chemical Engineering Journal*, **44 (7)**, 1596–1610.

Chen, J. and C.-M. Liao (2002). Dynamic process fault monitoring based on neural network and PCA. *Journal of Process Control*, **12**, 277–289.

Chiang, L.H., E.L. Russell and R.D. Braatz, ,(2001). *Fault Detection and Diagnosis in Industrial Systems.* Springer, London.

Chiang, L.H., R.D. Braatz, (2003). Process monitoring using causal map and multivariate statistics: fault detection and identification. *Chemometrics and Intelligent Laboratory Systems*, **65**, 159–178.

Cristianini, N, J. Shawe-Taylor (2000). *An introduction to support vector machine and other kernel-based learning methods*, first edition (Cambridge: Cambridge university press)

Hyväarinen, A.(1999). Fast and robust Fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks,* **10**, 626–634.

Hyväarinen, A. and E. Oja (2000). Independent component analysis algorithms and applications. *Neural Networks*, **13 (4-5),** 411-430.

Hyväarinen, A., J. Karhunen, and E. Oja, (2001). *Independent Component Analysis*. Wiley, New York.

Lee, J.-M., C.K Yoo and I.-B Lee (2003). Statistical process monitoring with independent component analysis. *Journal of Process Contro,l* **14 (5)**, 467–485.

Lee, J.-M., C.K Yoo and I.-B Lee (2004). Statistical monitoring of dynamic Processes based on dynamic independent component analysis. *Chemical Engineering Science*, **59**, 2995 – 3006

Malinowski F.R. (1991). *Factor Analysis in Chemistry*, Wiley-Inter-science, New York.