A Support Vector Clustering-Based Probabilistic Method for Unsupervised Fault Detection and Classification of Complex Chemical Processes Using Unlabeled Data

Jie Yu
Dept. of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L7

DOI 10.1002/aic.13816
Published online April 23, 2012 in Wiley Online Library (wileyonlinelibrary.com).

A new support vector clustering (SVC)-based probabilistic approach is developed for unsupervised chemical process monitoring and fault classification in this article. The spherical centers and radii of different clusters corresponding to normal and various kinds of faulty operations are estimated in the kernel feature space. Then the geometric distance of the monitored samples to different cluster centers and boundary support vectors are computed so that the distance–ratio–based probabilistic-like index can be further defined. Thus, the most probable clusters can be assigned to the monitored samples for fault detection and classification. The proposed SVC monitoring approach is applied to two test scenarios in the Tennessee Eastman Chemical process and its results are compared to those of the conventional K-nearest neighbor Fisher discriminant analysis (KNN-FDA) and K-nearest neighbor support vector machine (KNN-SVM) methods. The result comparison demonstrates the superiority of the SVC-based probabilistic approach over the traditional KNN-FDA and KNN-SVM methods in terms of fault detection and classification accuracies. © 2012 American Institute of Chemical Engineers *AIChE J,* 59: 407–419, 2013

Keywords: unsupervised process monitoring, fault detection, fault classification, support vector clustering, probabilistic-like index, Tennessee Eastman Chemical process

Introduction

Effective monitoring of chemical processes is critically important to ensure safe plant operation, reliable product quality, consistent environment compliance and maximized production profit. Any abnormal process operation should be detected in the early stage to avoid the occurrence of serious incidents in a plant. Then the root causes of process faults can be diagnosed so that the corrective actions may be further taken to move the plant back to normal status.[1–3] Traditional methods of process monitoring are mainly based on mechanistic models, which require significant effort and in-depth knowledge to develop.[4] For complex processes, it is not a trivial task to build fundamental models so that this type of approaches are difficult to implement in practice.

With the development of advanced measurement and data storage techniques, data driven multivariate statistical techniques have been widely applied to chemical process monitoring over the past decades. The most popular multivariate process monitoring methods are principal component analysis (PCA) and partial least squares (PLS).[5–10] These techniques can project the measurement data from the original high-dimensional space into low-dimensional linear subspace with covariance or cross-correlation information retained. Then the fault detection and diagnosis can be performed within the latent variable subspace using Hotelling's T^2 and squared prediction error indexes. The conventional PCA or PLS monitoring methods are targeted at linear systems, and, thus cannot handle nonlinearity in the processes. To deal with nonlinear processes, kernel function-based PCA and PLS approaches have been developed and applied to chemical process monitoring.[11,12] Basically, kernel PCA or PLS converts the input space into high-dimensional feature space through nonlinear kernel mapping and then the fault detection statistics can be derived in the kernel feature space. In PCA- or PLS-based monitoring methods, the objective is to de correlate latent variables and, therefore, only second-order statistics are taken into account. However, industrial processes are often of non-Gaussianity and, thus, higher-order statistics should not be ignored. More recently, independent component analysis (ICA)-based monitoring approach has been proposed to tackle non-Gaussian processes.[13,14] The statistically independent latent variables are extracted to track the abnormal operation events in complex processes with significant non-Gaussianity. Alternately, Gaussian mixture model (GMM) has been integrated with Bayesian inference for multimode non-Gaussian process monitoring and fault diagnosis. The finite mixture model can well characterize the multimodality due to shifting operation conditions in chemical processes.[15,16]

The aforementioned PCA, PLS, ICA, and GMM methods essentially require fault-free data set to train the normal operation model. In practice, however, the collected training data often include both normal and faulty samples. Though

Correspondence concerning this article should be addressed to J. Yu at jieyu@mcmaster.ca.

the supervised monitoring techniques such as FDA[17,18] and support vector machine (SVM)[19–21] can deal with such modeling set, the prior assumption is that the faulty samples have been isolated from the normal ones so that the class labels are available on the training data. Typically, a preliminary clustering step is required to identify normal and faulty data labels in order for the above FDA or SVM methods to be applicable. The traditional clustering techniques such as k-nearest neighbor (KNN) and k-means methods are not the best choices in handling process nonlinearity and non-Gaussianity. In this work, a support vector clustering-(SVC)-based fault detection and classification approach is proposed and applied to the chemical processes with unlabeled data set. The original multivariate measurements are mapped into a high-dimensional feature space with Gaussian kernel function. Then the spherical boundaries enclosing different data clusters are identified within the feature space. Further the spheres are mapped back to the original measurement space as a series of contours to delineate the underlying probability distributions of various clusters, which correspond to normal or different types of faulty samples. With the identified cluster boundaries, a geometric distance-based probability index is established to determine whether the operation data are normal or faulty. Moreover, different types of process faults can be classified on the abnormal samples.

The article is laid out as follows. The preliminaries on SVM technique are briefly reviewed as shown in the Preliminaries. The new SVC-based fault detection and classification approach is described as shown in the SVC-Based Probabilistic Method for Fault Detection and Classification. The performance of the proposed monitoring approach through the application example of the Tennessee Eastman Chemical Process is illustrated as shown in the Case Study. This work is finally concluded in the Conclusion.

Preliminaries

SVM is a powerful technique for pattern classification, function regression and probability density estimation.[22] The formulation of SVM is based on the structural risk minimization principle, which minimizes the upper bound on the expected risk of models. For the classification problem, the goal of SVM is to search for the optimal separating hyperplane with the maximized separation margin. Consider a set of training samples $\{(x_1,\ y_1),\ (x_2,\ y_2),\ldots,(x_m,\ y_m)\}$ with $x_i \in R^n$ and $y_i \in \{-1, 1\}$. The two-class separation hyperplane can be expressed as[23,24]

$$\langle \omega, x \rangle + b = 0 \tag{1}$$

where the parameters ω and b satisfy the following constraints

$$y_i[\langle \omega, x_i \rangle + b] \geq 1, \quad i = 1, 2, \ldots, n \tag{2}$$

Then the optimal separating hyperplane can be solved from the minimization problem below[25]

$$\min \frac{1}{2}||\omega||^2 \tag{3}$$

which is subject to the constraints in Eq. 2. The solution to the above optimization problem can be further obtained from the saddle point of the following Lagrange function

$$\phi(\omega, b, \alpha) = \frac{1}{2}||\omega||^2 - \sum_{i=1}^{n} \alpha_i (y_i[\langle \omega, x_i \rangle + b] - 1) \tag{4}$$

where $\alpha_i \geq 0$ is the Lagrange multiplier.

The corresponding dual problem is given by

$$\max_{\alpha} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{k=1}^{n} \alpha_k \right\} \tag{5}$$

whose solution is expressed as

$$\arg\min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_{k=1}^{n} \alpha_k \right\} \tag{6}$$

with

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{7}$$

The parameters of optimal separating plane are estimated as

$$\hat{\omega} = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{8}$$

and

$$\hat{b} = -\frac{1}{2} \langle \hat{\omega}, x_{\text{sv}}^A + x_{\text{sv}}^B \rangle \tag{9}$$

where x_{sv}^A and x_{sv}^B are the support vectors belonging to two different classes A and B, respectively. The SVM-based classifier can be then defined as the following sign function

$$f(x) = \text{sgn}(\langle \hat{\omega}, x \rangle + b) \tag{10}$$

With the nonlinear kernel function K introduced, the optimization problem is updated to

$$\max_{\alpha} \left\{ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum_{k=1}^{n} \alpha_k \right\} \tag{11}$$

with the constraints

$$0 \leq \alpha_i \leq C \quad i = 1, 2, \ldots, n \tag{12}$$

where C is the regularization parameter corresponding to the penalty level of errors. The nonlinear classifier is expressed as

$$f(x) = \text{sgn}\left(\sum_{i \in \text{SVs}} \alpha_i K(x_i, x) \right) \tag{13}$$

The class labels of data samples are needed to train the SVM model.

SVC-Based Probabilistic Method for Fault Detection and Classification

Different from the SVM-based classification, the SVC method has the capability of unsupervised learning with

unlabeled training samples. A series of support vectors are identified to characterize the probability density distribution of training data and then the minimum spheres surrounding different classes are estimated within the high-dimensional feature space.[26,27] Let $X = \{x_1,x_2,\ldots,x_m\}$ be a set of n-dimensional input measurements in the training data set while the corresponding class labels are unavailable. It is noted that a preliminary data normalization step is conducted to scale the training samples to zero means and unit variances along all the measurement variables. Meanwhile, a nonlinear mapping function Φ is defined to project the observations from the low-dimensional measurement space into the high-dimensional feature space. Thus, the mapped samples in the feature space are denoted as

$$\{\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_m)\} \tag{14}$$

Assume that there are total $L + 1$ classes

$$\{S_1, S_2, \ldots, S_{L+1}\} \tag{15}$$

which correspond to the normal operation and L different types of process faults as follows

$$\begin{aligned} &S_1 \quad : \text{normal operation} \\ &S_2 \quad : \text{fault type 1} \\ &S_3 \quad : \text{fault type 2} \\ &\vdots \\ &S_{L+1} \quad : \text{fault type L} \end{aligned}$$

For an arbitrary class S_k, the corresponding sphere in the feature space can be characterized as

$$||\Phi(x_i) - \gamma_k||^2 \le r_k^2 \tag{16}$$

where $||\cdot||$ is the $L2$ norm, γ_k represents the spherical center of the kth class and r_k denotes the radius of the sphere. If a slack variable $\xi_i \ge 0$ is introduced to slightly relax the boundary of the sphere, the above formulation becomes

$$||\Phi(x_i) - \gamma_k||^2 \le r_k^2 + \xi_i \tag{17}$$

The Lagrange function of the optimization problem is written as

$$F = r_k^2 - \sum_{i=1}^{m} \alpha_i (r_k^2 + \xi_i - ||\Phi(x_i) - \gamma_k||^2) - \sum_{i=1}^{m} \beta_i \xi_i + C \sum_{i=1}^{m} \xi_i \tag{18}$$

where $\alpha_i \ge 0$ and $\beta_i \ge 0$ are the Lagrange multipliers, and C is the regularization constant to determine the penalty on slack variables.[26] The optimal solutions on normal or faulty cluster spheres are obtained by minimizing the above Lagrange function and the sphere centers are estimated as

$$\hat{\gamma}_k = \sum_{i=1}^{m} \alpha_i \Phi(x_i) \tag{19}$$

which is derived by setting to zeros the first-order derivatives of F with respect to r_k, γ_k and ξ_i. Here, the Lagrange multipliers also satisfy the conditions of $\sum_{i=1}^{m} \beta_i = 1$ and $\alpha_i + \beta_i = C$ ($i = 1, 2,\ldots,m$). Meanwhile, the Karush–Kuhn–Tucker conditions further lead to

$$\beta_i \xi_i = 0 \tag{20}$$

and

$$\alpha_i (r_k^2 + \xi_i - ||\Phi(x_i) - \gamma_k||^2) = 0 \tag{21}$$

The dual problem can then be expressed as

$$W = \sum_{i=1}^{m} \alpha_i K(x_i, x_i) - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \tag{22}$$

where $K(x_i, x_j)$ is the Gaussian kernel function defined as

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle = e^{-||x_i - x_j||^2 / 2\sigma^2} \tag{23}$$

with σ being the kernel width.

The mapped points in the kernel feature space that are outside the cluster spheres are termed as bounded support vectors and the corresponding Lagrange multipliers satisfy the condition of $\alpha_i = C$. On the other hand, the boundary points on the spheres within the kernel feature space form support vectors and the relevant multipliers meet the inequality of $0 \le \alpha_i \le C$. The radius of the support vectors is equivalent to the cluster sphere radius as

$$\hat{r}_k^2 = K(x_{\mathrm{SV}}, x_{\mathrm{SV}}) - 2 \sum_{i=1}^{m} \alpha_i K(x_{\mathrm{SV}}, x_i) + \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \tag{24}$$

where x_{SV} is an arbitrary support vector with respect to the kth cluster S_k. The boundary of the cluster is characterized by a set of points that satisfy the following criterion

$$\{x | r(x) = \hat{r}_k^2\} \tag{25}$$

where $r(x)$ is the distance of the boundary point x to the cluster center γ_k. It should be noted that there are only two user-specified parameters, the Gaussian kernel width σ and the regularization constant C, to implement SVC. In this study, the values of σ and C are selected through cross validation procedure.

For any monitored sample x_t of the process, its geometric distance to the kth cluster sphere center is given by

$$\hat{r}^2(x_t, S_k) = K(x_{\mathrm{SV}}, x_t) - 2 \sum_{i=1}^{m} \alpha_i K(x_t, x_i) + \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \tag{26}$$

If $\hat{r}^2(x_t, S_k)$ is equal or less than the cluster radius $\hat{r}_k^2$, then the sample point is determined to be within the kth cluster sphere. Otherwise, it is outside the cluster sphere. For the monitored sample, a distance–ratio–based probabilistic-like index with respect to all different clusters is defined as follows

$$P(S_k | x_t) = \frac{\min\{\hat{r}_k^2, \hat{r}^2(x_t, S_k)\}}{\hat{r}^2(x_t, S_k)} \tag{27}$$

where $1 \le k \le L + 1$ and $0 \le P(S_k|x_t) \le 1$. The probability index value satisfies $P(S_k|x_t) = 1$ as long as $\hat{r}^2(x_t, S_k) \le \hat{r}_k^2$,

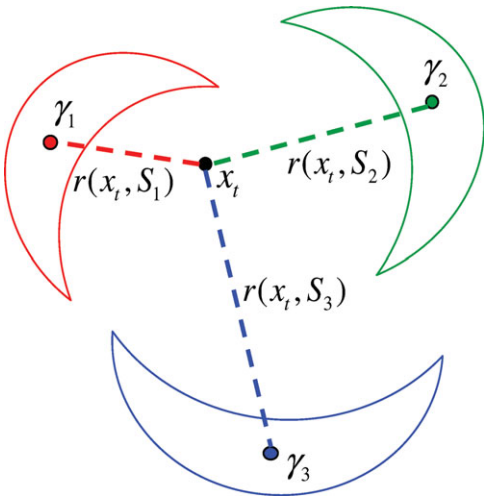

Figure 1. Geometric illustration of the distance from the monitored sample to different cluster centers.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

which corresponds to the sample points on or within the identified cluster spheres in the kernel feature space. The geometric illustration is shown in Figure 1.

The probabilistic rules of cluster assignment for the sample point x_t can then be established as

$$S(x_t) = \underset{S_k \in \{S_1, S_2, \cdots, S_{L+1}\}}{\arg\max} P(S_k | x_t) \tag{28}$$

where $S(x_t)$ is the most probable cluster that the monitored sample x_t is categorized into. Hence, the fault detection can be conducted according to the assigned clusters as follows

- Normal operation if $S(x_t) = S_1$
- Faulty event if $S(x_t) = \{S_2, S_3, \ldots, S_{L+1}\}$

For the detected abnormal samples, the corresponding fault types may be further determined as

- Fault type 1 if $S(x_t) = S_2$
- Fault type 2 if $S(x_t) = S_3$
- …
- Fault type L if $S(x_t) = S_{L+1}$

The procedure of the SVC-based process fault detection and classification approach is summarized below and the flow diagram of the proposed method is shown in Figure 2.

1. Gather the multivariate measurement data with unknown class labels from the monitored process to form the training set;
2. Conduct preliminary normalization on the training data to scale the samples to zero means and unit variances along all measurement variables;
3. Estimate the sphere centers and radii of different clusters based on Gaussian kernel function and constrained optimization;
4. Identify the support vectors on the spheres of all different clusters within kernel feature space;
5. For any monitored sample from the process, normalize the multivariate measurements based on the means and variances of training data set;
6. Compute the geometric distance of the monitored sample from the centers of all different clusters based on the corresponding support vectors;
7. Estimate the distance–ratio–based probabilistic-like index values of the monitored sample with respect to all the identified clusters;
8. Assign the most probable cluster of the monitored sample by maximizing the probabilistic-like index with respect to different clusters;
9. Determine normal or faulty operation based on the above cluster assignment;
10. Further perform fault classification through the identified cluster from Step (8) to isolate different types of process faults.

Case Study

Tennessee Eastman chemical process

In this work, the Tennessee Eastman Chemical process data are used to evaluate the performance of the SVC-based fault detection and classification approach. Meanwhile, the results of SVC-based probabilistic monitoring method are compared to those of KNN-FDA and KNN-SVM techniques.

The Tennessee Eastman process includes five major unit operations, which are a reactor, a product condenser, a vapor–liquid separator, a recycle compressor, and a stripper. Among the process input streams, four chemical reactants A, C, D, and E are fed into the reactor to form two products of G and H along with a byproduct of F.[28] The process flow diagram is shown in Figure 3. The whole process is operated continuously with overall 12 manipulated variables and 41 measurement variables, among which 22 variables provide continuous measurements while the remain 19 variable

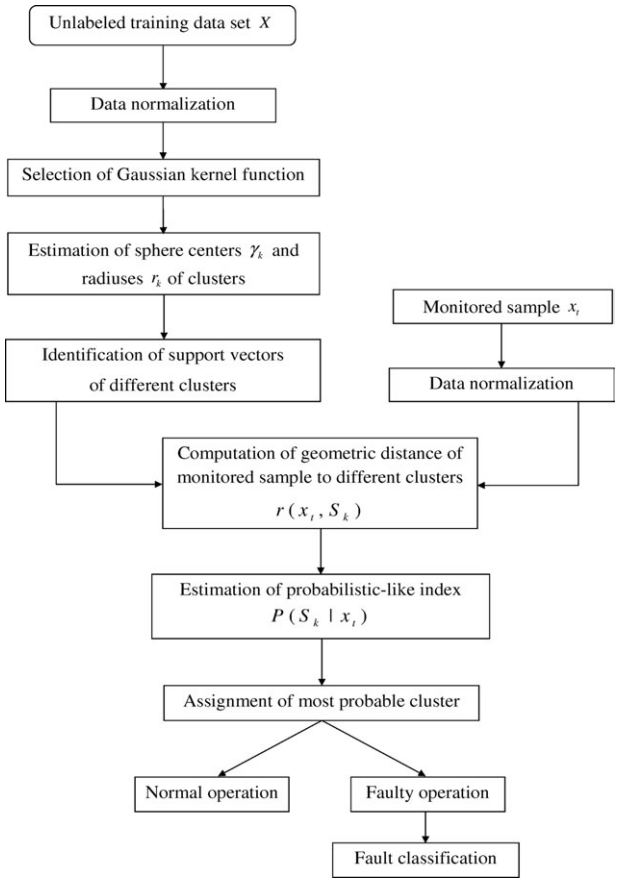

Figure 2. Flow diagram of the SVC-based fault detection and classification approach.

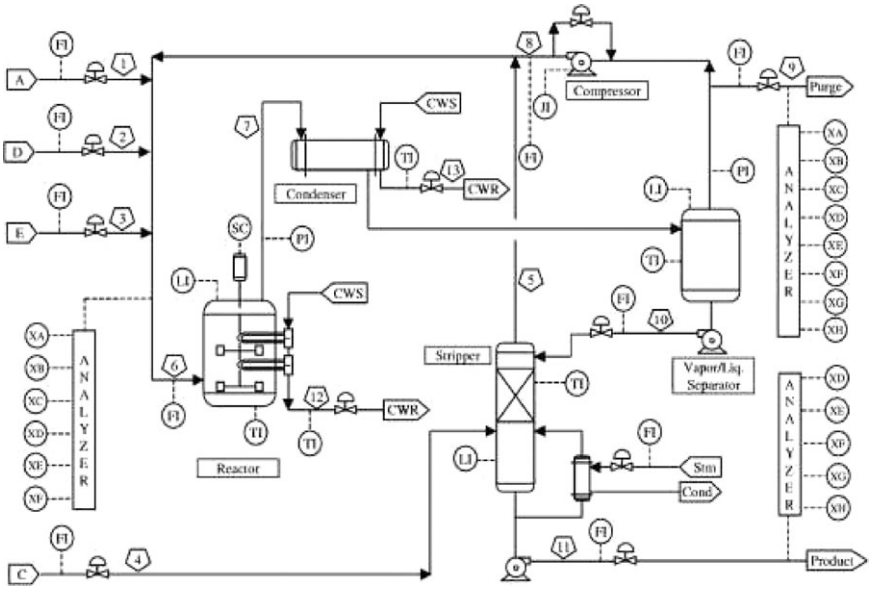

Figure 3. Flow sheet of the Tennessee Eastman Chemical process.

record chemical compositions through either off-line lab analysis or on-line analyzers. The 22 continuous measurement variables and the 12 manipulated variables are listed in Table 1. The process involves a plant-wide decentralized control implementation with multiple feedback or loops.[29] In this study, the 22 continuous measurement variables are adopted for process monitoring purpose and the sampling time of those measurements is 3 min.

As listed in Tables 2 and 3, two different scenarios are designed to assess and compare the accuracy and reliability of various monitoring methods. In the first case, the training set starts with 500 normal samples from the steady-state operation and then it is followed by 200 faulty points with the occurrence of step error in D feed temperature. After that, the increased random variation in condenser cooling water inlet temperature is taking place in the process with the duration of additional 200 samples. The test set is initiated with normal operation data and then the increased random variation in condenser coolant temperature occurs from the 201st sample. After the random variation error lasts 100 samples, the process operation returns to normal state with a duration of 200 samples. From the 501st sample, the step error in D feed temperature takes place in the plant and remains for 100 samples. The second scenario is more complicated and composed of four types of faults, which are step error in reactor cooling water inlet temperature, increased random variation error in A, B, C feed composition, slow drift in reaction kinetics, and condenser cooling water valve stiction. The training data set involves 600 normal samples that are followed by 600 faulty ones from the above four kinds of faults, respectively. Meanwhile, the normal and faulty segments alternate in the test set and each of the segments includes 150 or 200 samples as described in Table 3.

Table 1. Continuous Measurement and Manipulated Variables in the Tennessee Eastman Chemical Process

No.	Measured Variable	No.	Manipulated Variable
1	A Feed rate	1	D Feed flow valve
2	D Feed rate	2	E Feed flow valve
3	E Feed rate	3	A Feed flow valve
4	A+C Feed rate	4	A+C Feed flow valve
5	Recycle flow rate	5	Recycle valve
6	Reactor feed rate	6	Purge valve
7	Reactor pressure	7	Separator valve
8	Reactor level	8	Stripper valve
9	Reactor temperature	9	Steam valve
10	Purge rate	10	Reactor coolant flow
11	Separator temperature	11	Condenser coolant flow
12	Separator level	12	Agitator speed
13	Separator pressure		
14	Separator underflow		
15	Stripper level		
16	Stripper pressure		
17	Stripper underflow		
18	Stripper temperature		
19	Stem flow rate		
20	Compressor work		
21	Reactor coolant temperature		
22	Condenser coolant temperature		

Comparison of fault detection and classification results of KNN-FDA-, KNN-SVM-, and SVC-based probabilistic methods

In both test scenarios, the training data set is first used to build the KNN-FDA, KNN-SVM, and SVC models. Then the test sets are fed into the models for fault detection and classification. The class labels indicating normal operation

Table 2. Training Sets of Four Simulated Scenarios in the Tennessee Eastman Chemical Process

Case No.	Training Set
1	500 normal samples 200 faulty samples with step error in D feed temperature 200 faulty samples with random variation in condenser coolant temperature
2	600 normal samples 150 faulty samples with step error in reactor coolant inlet temperature 150 faulty samples with random variation in A, B, C feed composition 150 faulty samples with drift error in reaction kinetics 150 faulty samples with condenser coolant flow valve stiction

Table 3. Test Sets of Four Simulated Scenarios in the Tennessee Eastman Chemical Process

Case No.	Test Set
1	1st–200th samples: normal operation 201st–300th samples: random variation in condenser coolant temperature 301st–500th samples: normal operation 501st–600th samples: step error in D feed temperature
2	1st–150th samples: normal operation 151st–300th samples: step error in reactor coolant inlet temperature 301st–500th samples: normal operation 501st–650th samples: drift error in reaction kinetics 651st–800th samples: normal operation 801st–1000th samples: random variation in A, B, C feed composition 1001st–1200th samples: normal operation 1201st–1350th samples: condenser coolant flow valve stiction

and faulty events are assumed to be unknown for model learning so that the KNN- or SVC-based clustering procedure is used to identify various categories before the abnormality detection and fault type classification.

For the first case, the fault detection results of KNN-FDA, KNN-SVM, and SVC-based probabilistic methods are shown in Figures 4, 5, and 6, respectively. The quantitative performance index values including fault detection rate, false alarm rate, and fault classification rate of the three methods are summarized in Table 4. It can be observed from Figures 4 and 5 that there are significant numbers of normal samples misidentified as faulty ones as well as abnormal points undetected by both KNN-FDA and KNN-SVM approaches, though SVM appears to be a little better than FDA in fault detection capability. In contrast, the SVC-based probabilistic method results in much fewer misidentified normal samples and undetected faulty ones, as illustrated in Figure 6. The fault detection rate and false alarm rate of SVC-based monitoring method are 97.5 and 4.0%, both of which are more desirable than those of KNN-FDA and KNN-SVM methods. The FDA algorithm essentially searches for the linear latent variables or subspace that optimize the separation index in terms of the ratio of the between-class over the within-class distances. Hence, the underlying nonlinearity and non-Gaussianity in the process data may not be effectively handled by FDA method. Though the SVM-based classification can deal with process nonlinearity, it is a kind of supervised monitoring technique and the preliminary KNN clustering step is necessary to isolate the various normal and faulty clusters of the process data. However, the identified cluster separation in KNN may not be the optimal so that the accuracy of normal and faulty sample isolation tends to be degraded. As a comparison, SVC method identifies the nonlinear cluster boundaries with the maximized separation margins. Further, the probabilistic inference strategy can assign test samples to different clusters with the highest likelihood. Consequently, the best fault detection performance can be achieved by SVC approach in this test scenario.

Within the detected faulty segments, the classification of various process fault types can be further conducted and the results of the three methodsare shown in Figures 7, 8, and 9, respectively. As seen from Figure 7a, there are 11 samples misclassified as step error by KNN-FDA method when the

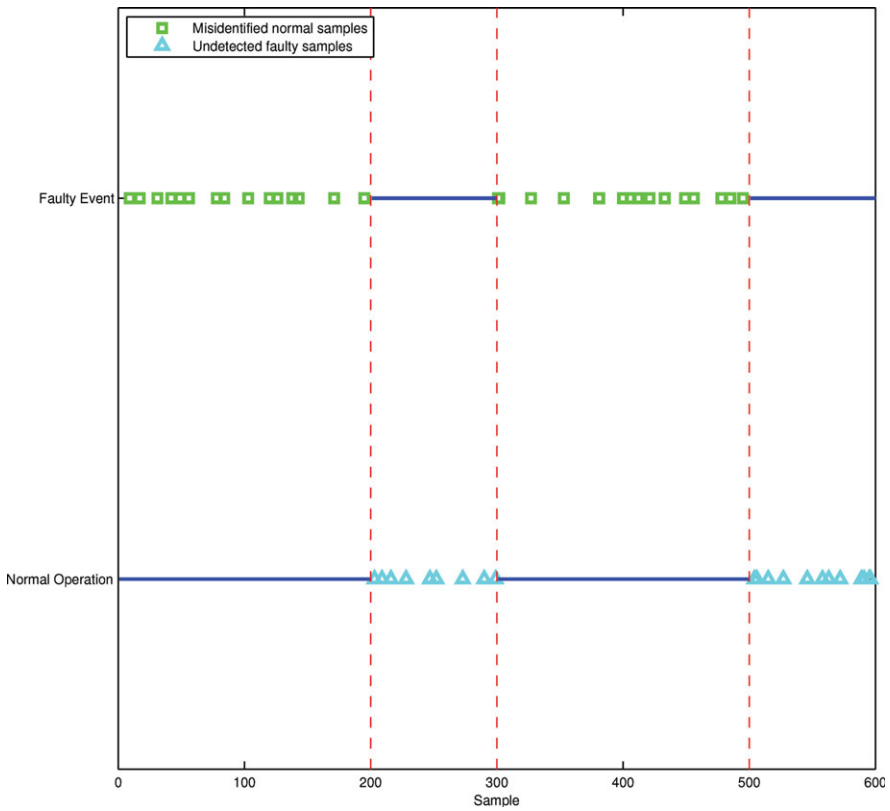

Figure 4. First test case of the Tennessee Eastman Chemical process: fault detection results of KNN-FDA method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

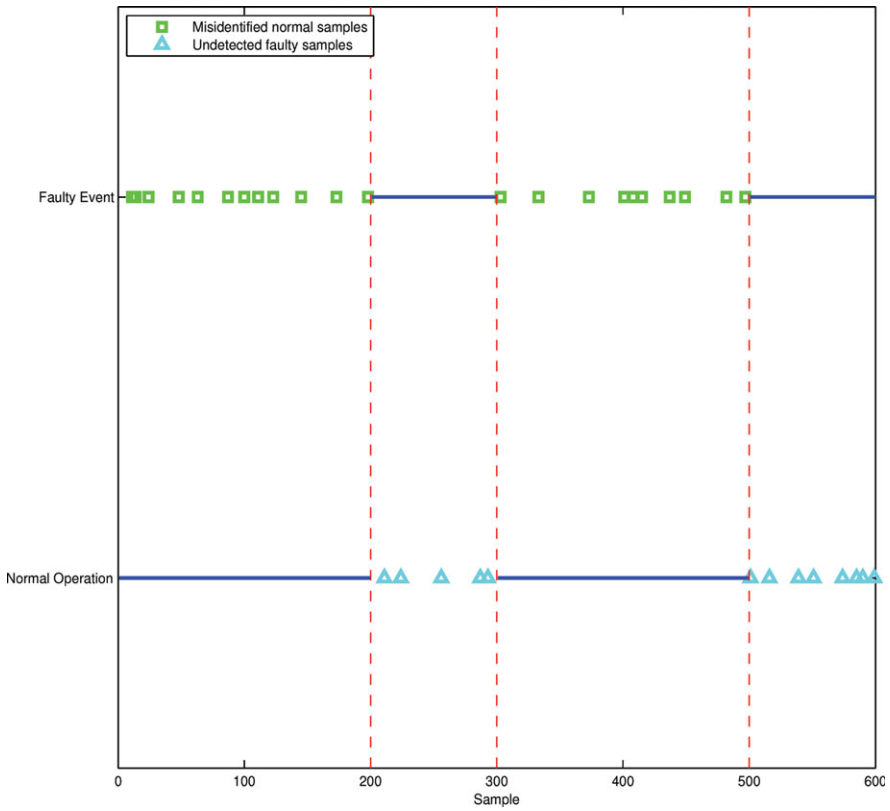

Figure 5. First test case of the Tennessee Eastman Chemical process: fault detection results of KNN-SVM method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

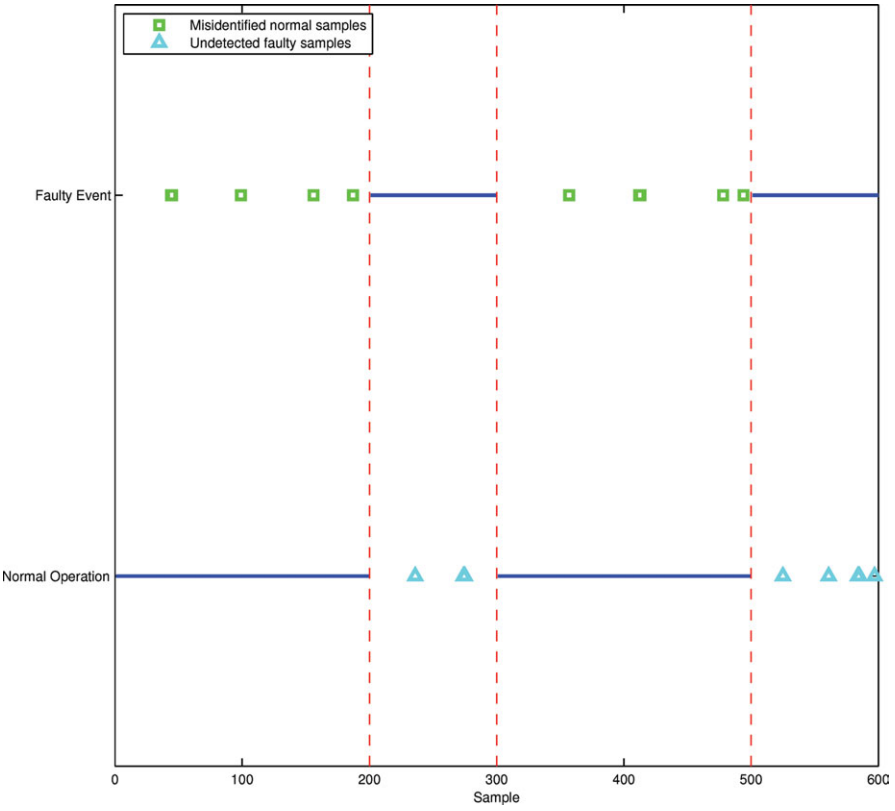

Figure 6. First test case of the Tennessee Eastman Chemical process: fault detection results of SVC based probabilistic method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 4. Comparison of Fault Detection and Classification Results among KNN-FDA, KNN-SVM and SVC-Based Probabilistic Methods

	Test Case 1			Test Case 2		
Performance Index	KNN-FDA (%)	KNN-SVM (%)	SVC (%)	KNN-FDA (%)	KNN-SVM (%)	SVC (%)
Fault detection rate	92.5	94.5	97.5	88.3	90.9	96.4
False alarm rate	10.5	6.5	4.0	12.3	8.0	4.8
Fault classification rate	89.0	93.5	95.5	86.5	90.5	95.2

fault of increased random variation occurs in the process during that operation period. Meanwhile, another 11 points with random variation type of fault are wrongly categorized as step error. The KNN-SVM method, as shown in Figure 8a,b, leads to slightly improved fault classification results with six samples misidentified as step error while 7 points as random variation. Nevertheless, the SVC-based probabilistic method provides the best fault classification rate with only four samples misclassified as step error and 5 as random variation. The above comparison demonstrates the strong fault classification ability of the proposed SVC monitoring method. Because of process uncertainty and noises, the clustering on the unlabeled training data may lead to kind of biased cluster boundaries. Then the test samples that are affected by the process uncertainty and noises may go across the distorted normal or faulty cluster boundaries and, thus, be misclassified. That is why certain number of individual samples are misidentified by different unsupervised monitoring methods. The better performance of SVC monitoring approach in this aspect indicates that the identified cluster boundaries from training data are more accurate and have improved generalization capability.

To further examine the fault detection and classification performance of different methods, the second test scenario is designed with more complicated faulty events occurring in the process. The fault detection results are depicted in Figures 10, 11 and 12, respectively. The fault detection accuracy of SVC-based probabilistic method is still better than that of both KNN-FDA and KNN-SVM approaches. The fault detection rate and false alarm rate of SVC method are 96.4% and 4.8%. In contrast, the KNN-FDA and KNN-SVM methods have lower fault detection rates of 88.3% and 90.9% while higher false alarm rates of 12.3% and 8.0%, respectively. Such comparison indicates that the SVC-based probabilistic method is very robust in monitoring complex processes with multiple types of faulty events. After the faulty operations are detected, the effort can be focused on the abnormal periods to classify fault types. The fault classification results of three different methods in the second case are compared in Figures 13, 14, and 15. It is obvious that the KNN-FDA method has the worst fault classification accuracies in all the four faulty periods with the largest numbers of misclassified samples. Similar to the first case, the proposed SVC monitoring approach consistently leads to the

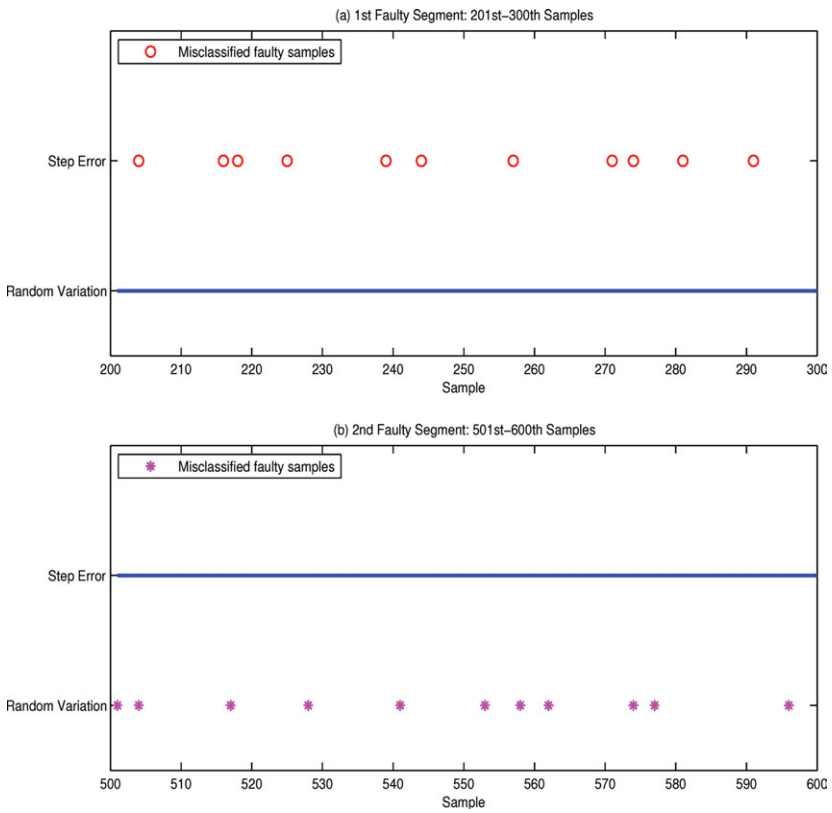

Figure 7. First test case of the Tennessee Eastman Chemical process: fault classification results of KNN-FDA method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

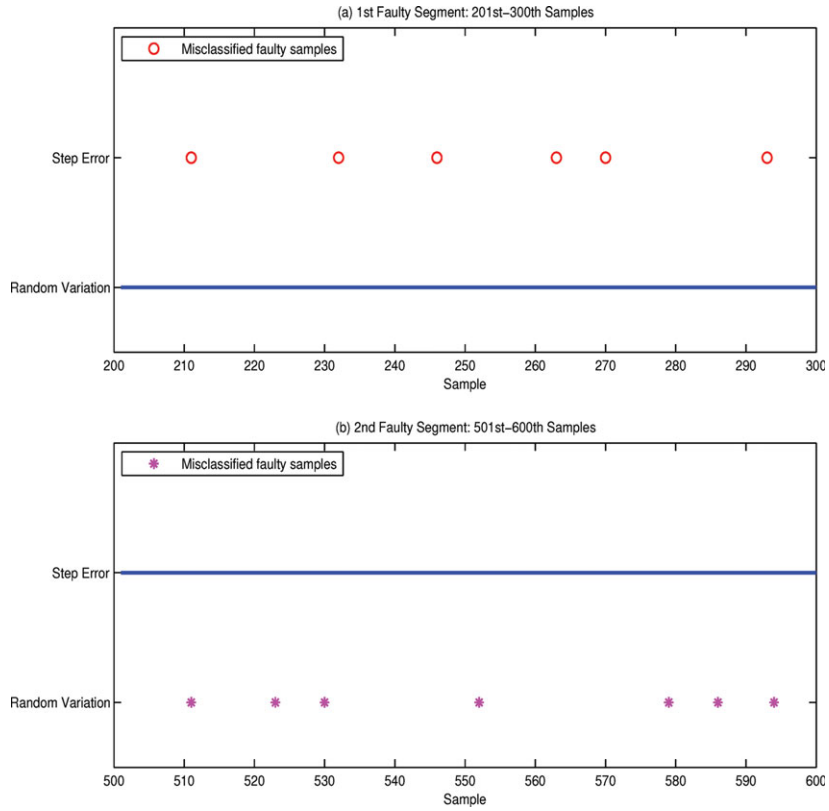

Figure 8. First test case of the Tennessee Eastman Chemical process: fault classification results of KNN-SVM method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

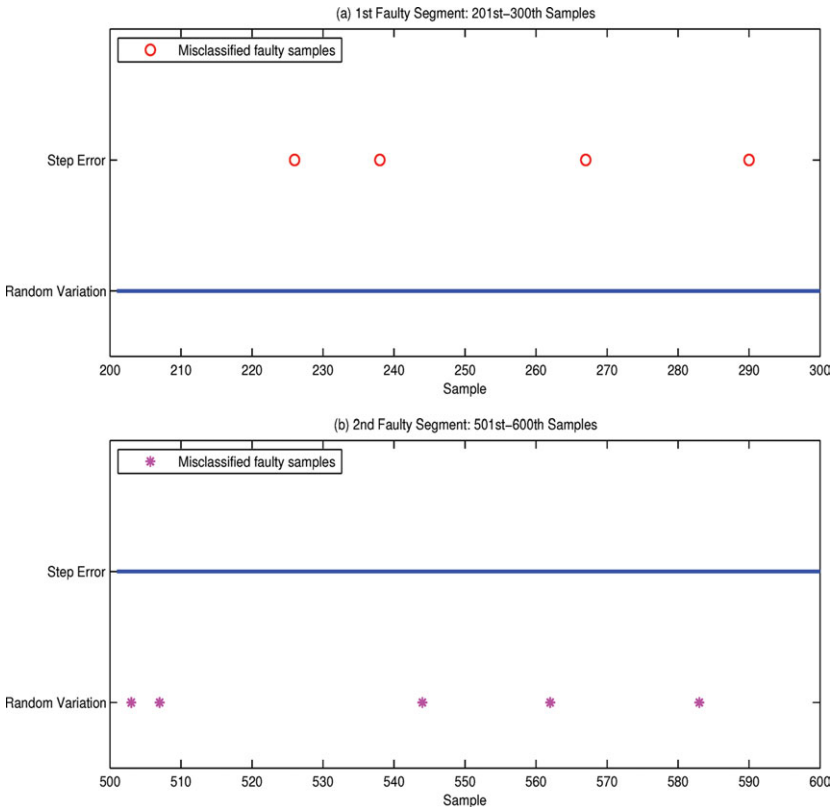

Figure 9. First test case of the Tennessee Eastman Chemical process: fault classification results of SVC-based probabilistic method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

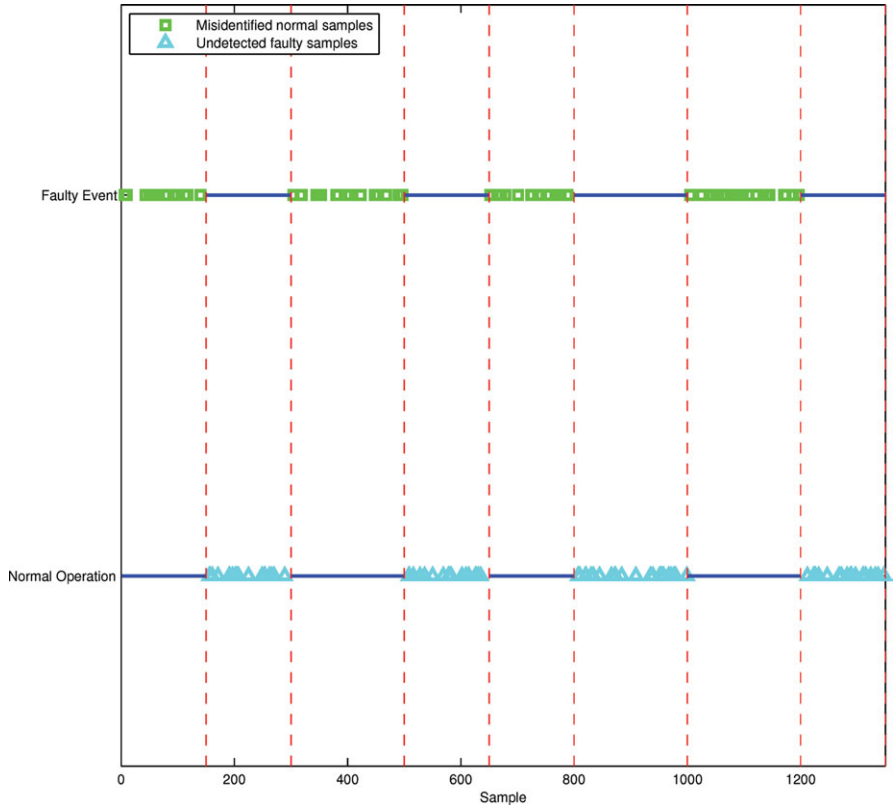

Figure 10. Second test case of the Tennessee Eastman Chemical process: fault detection results of KNN-FDA method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

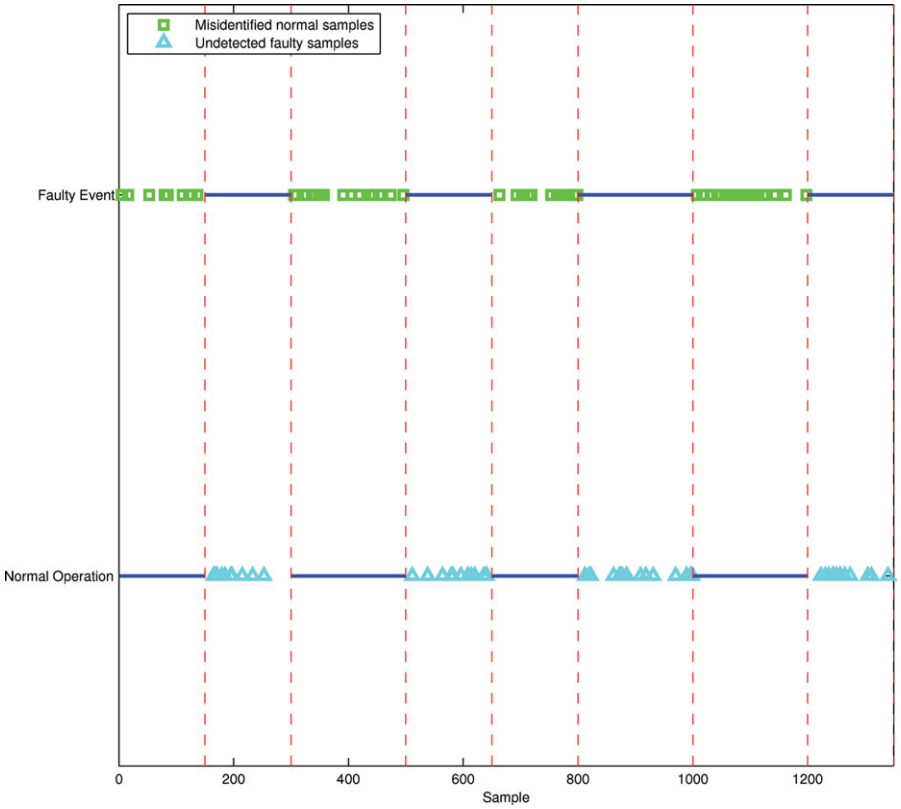

Figure 11. Second test case of the Tennessee Eastman Chemical process: fault detection results of KNN-FDA method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

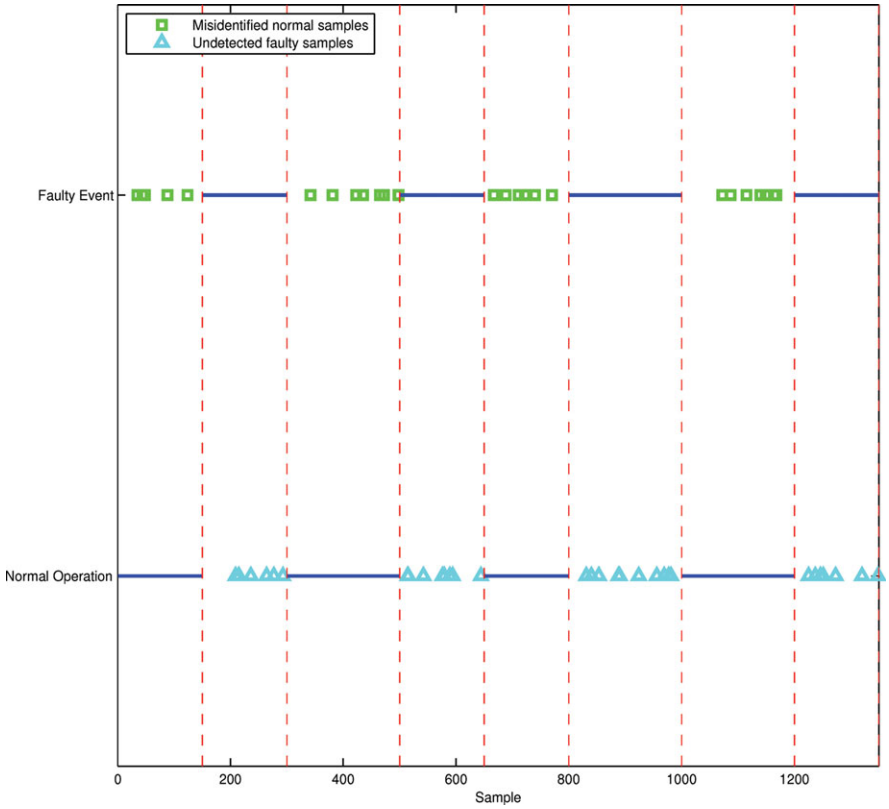

Figure 12. Second test case of the Tennessee Eastman Chemical process: fault detection results of SVC-based probabilistic method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

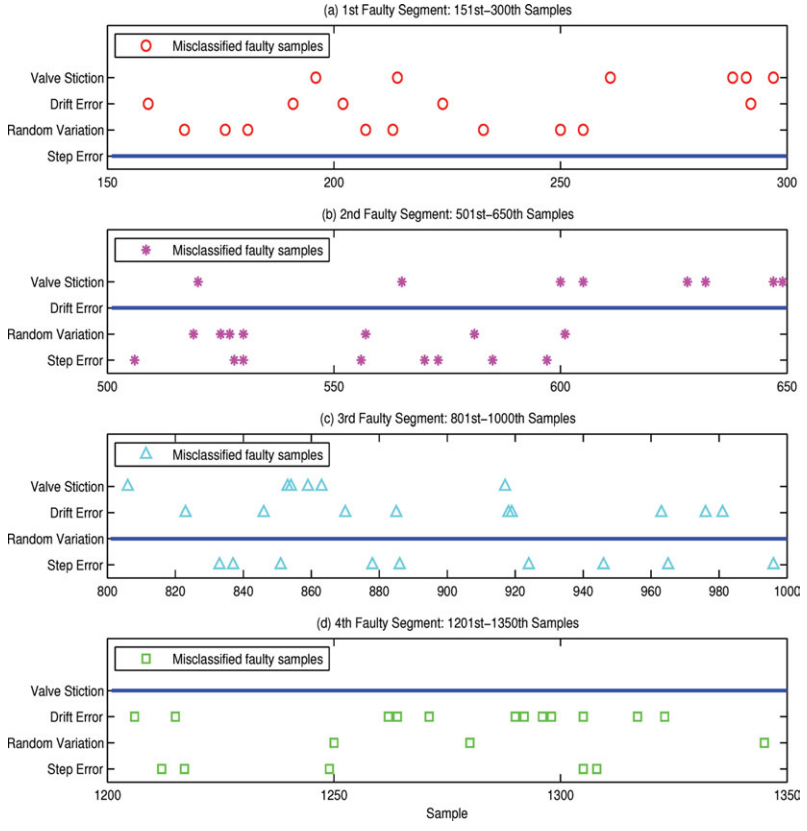

Figure 13. Second test case of the Tennessee Eastman Chemical process: fault classification results of KNN-FDA method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

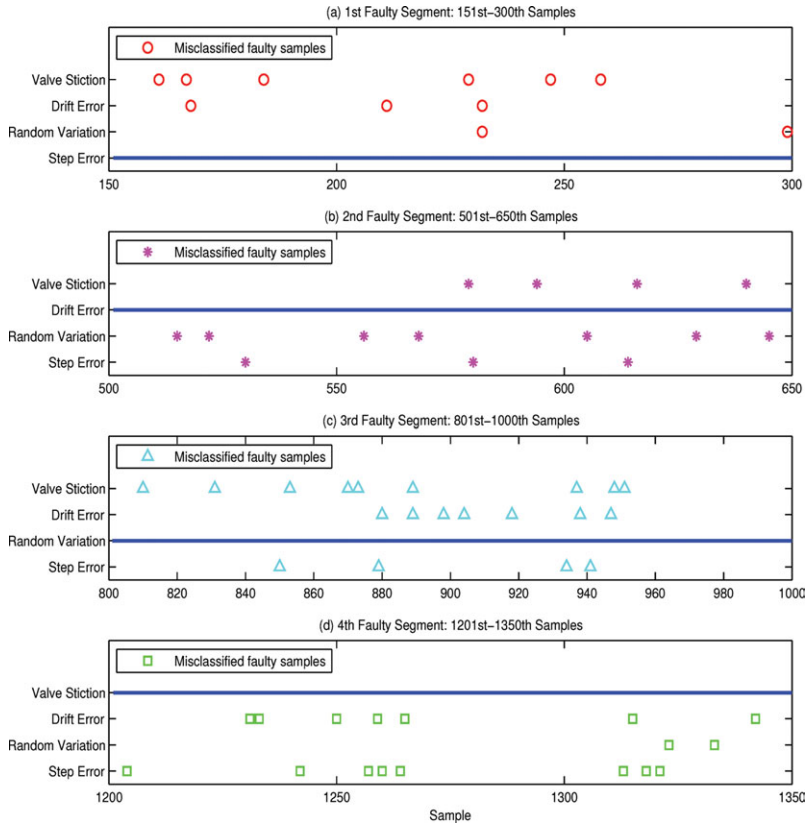

Figure 14. Second test case of the Tennessee Eastman Chemical process: fault classification results of KNN-SVM method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

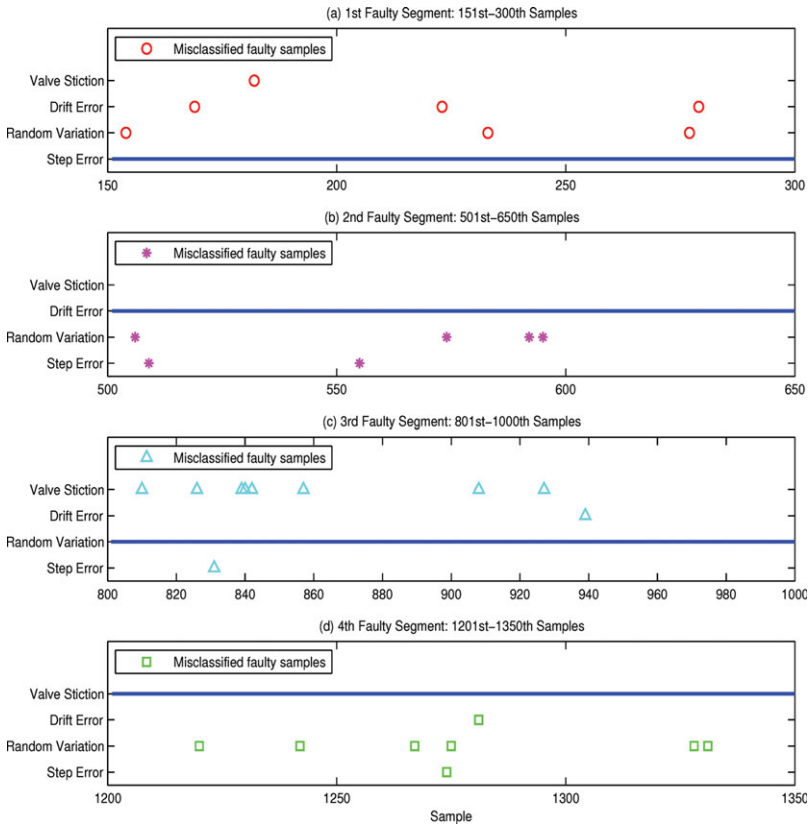

Figure 15. Second test case of the Tennessee Eastman Chemical process: fault classification results of SVC-based probabilistic method.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

best fault classifications with the overall accuracy of 95.2%. Both test cases of the Tennessee Eastman Chemical process demonstrate that the SVC-based probabilistic method has reliable and satisfactory performance in detecting and classifying multiple kinds of process faults. In industrial practice, if the fault detection and identification are preferred on a set of data rather than individual samples, the additional heuristic rules can be designed in the monitoring systems so that the set of operation data are treated as faulty only if more than certain threshold percentage of samples are detected as faulty.

Conclusions

A novel SVC-based probabilistic method is proposed to detect and classify multiple types of operation faults in complex chemical processes. The spherical boundaries of normal and different faulty classes can be identified in the kernel feature space by SVC algorithm. Thus, a geometric distance–ratio–based probabilistic-like index is defined to assign test samples into the most likely clusters, which can be used to detect faulty operation and further classify different fault types.

The SVC-based probabilistic approach is able to deal with unlabeled process data and may serve as an unsupervised monitoring tool. The proposed method is applied to the Tennessee Eastman Chemical process and its fault detection and classification results are compared to those of the KNN-FDA and KNN-SVM approaches. Two test cases are designed and the computation results show that the SVC-based probabilistic approach has significantly improved fault detection and classification accuracies than both KNN-FDA and KNN-SVM methods. The SVC monitoring method not only has reliable performance in handling complex faulty scenarios, but also integrates the clustering and classification functions so that it can deal with the unlabeled training data including both normal and faulty events. Therefore, the SVC-based probabilistic approach may serve as a powerful tool for complex process monitoring and fault classification.

It should be noted that the harmless disturbances that do not cause significant operation upsets or abnormal behaviors on process variables will be classified into normal operation by the proposed SVC monitoring method, because the measurement data under this kind of harmless disturbances are closer to the normal rather than faulty cluster center within the feature space. However, if the disturbances result in substantial process upsets and abnormal measurement responses, then they will be more likely categorized as faults. In industrial applications, the operator feedback would be very useful for validating the monitoring decisions with the aid of operation knowledge. Future research can be focused on enhancing the adaptive feature of the SVC monitoring method to handle the potential faults that do not exist in the training data. The new types of faults in the test data should be automatically detected and further used to update the clustering model for improved fault type classification.

Literature Cited

1. Chiang L, Russell E, Braatz R. *Fault Detection and Diagnosis in Industrial Systems. Advanced Textbooks in Control and Signal Processing*. London, Great Britain: Springer-Verlag. 2001.
2. Venkatasubramanian V. Prognostic and diagnostic monitoring of complex systems for product lifecycle management: challenges and opportunities. *Comput Chem Eng*. 2005;29:1253–1263.
3. Kano M, Fujioka T, Tonomura O, Hasebe S, Noda M. Data-based and model-based blockage diagnosis for stacked microchemical processes. *Chem Eng Sci*. 2007;62:1073–1080.
4. Venkatasubramaniana V, Rengaswamy R, Yin K, Kavuri S. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Comput Chem Eng*. 2003;27:293–311.
5. Nomikos P, MacGregor JF. Monitoring of batch processes using multi-way principal component analysis. *AIChE J*. 1994;40:1361–1375.
6. MacGregor JF, Jaeckle C, Kiparissides C, Koutoudi M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE J*. 1994;40:826–838.
7. Kosanovich K, Dahl K, Piovoso M. Improved process understanding using multiway principal component analysis. *Ind Eng Chem Res*. 1996;35:138–146.
8. Bakshi BR. Multiscale PCA with application to multivariate statistical process monitoring. *AIChE J*. 1998;44:1596–1610.
9. Qin SJ. Recursive PLS algorithms for adaptive data modeling. *Comput Chem Eng*. 1998;22:503–514.
10. Wang X, Kruger U, Lennox B. Recursive partial least squares algorithms for monitoring complex industrial processess. *Control Eng Practice*. 2003;11:613–632.
11. Lee JM, Yoo CK, Lee IB. Nonlinear process monitoring using kernel principal component analysis. *Chem Eng Sci*. 2004;59:223–234.
12. Cho JH, Lee JM, Choi S, Lee D, Lee IB. Fault identification for process monitoring using kernel principal component analysis. *Chem Eng Sci*. 2005;60:279–288.
13. Kano M, Tanaka S, Hasebe S, Hashimoto I, Ohno H. Monitoring independent components for fault detection. *AIChE J*. 2003;49:969–976.
14. Lee JM, Yoo CK, Lee IB. Statistical monitoringnext term of dynamic previous termprocessesnext term based on dynamic independent component analysis. *Chem Eng Sci*. 2004;59:2995–3006.
15. Yu J, Qin SJ. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE J*. 2008;54:1811–1829.
16. Yu J. A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. *Chem Eng Sci*. 2012;68:506–519.
17. Chiang L, Kotanchek M, Kordon A. Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Comput Chem Eng*. 2004;28:1389–1401.
18. Yu J. Localized Fisher discriminant analysis based complex chemical process monitoring. *AIChE J*. 2011;57:1817–1828.
19. Kulkarni A, Jayaraman V, Kulkarni B. Knowledge incorporated support vector machines to detect faults in Tennessee Eastman Process. *Comput Chem Eng*. 2005;29:2128–2133.
20. Yélamos I, Graells M, Puigjaner L, Escudero G. Simultaneous fault diagnosis in chemical plants using a multilabel approach. *AIChE J*. 2007;53:2871–2884.
21. Zhang Y. Enhanced statistical analysis of nonlinear processes using KPCA, KICA and SVM. *Chem Eng Sci*. 2009;64:801–811.
22. Vapnik V. *Statistical Learning Theory*. Wiley: New York, NY, 1998.
23. Vapnik V. *The Nature of Statistical Learning Theory*. Springer: New York, NY. 1995.
24. Smola A, Schölkopf B. A tutorial on support vector regression. *Stat Comput*. 2004;14:199–222.
25. Gunn S. Support vector machines for classification and regression. 1998. ISIS technical report.
26. Ben-Hur A, Horn D, Siegelmann H, Vapnik V. Support vector clustering. *J Mach Learn Res*. 2002;2:125–137.
27. Lee J, Lee D. An improved cluster labeling method for support vector clustering. *IEEE Trans Pattern Anal Mach Intell*. 2005;27:461–464.
28. Downs JJ, Vogel EF. Plant-wide industrial process control problem. *Comput Chem Eng*. 1993;17:245–255.
29. Ricker NL. Decentralized control of the Tennessee Eastman Challenge Process. *J Proc Cont*. 1996;6:205–221.

Manuscript received Jan. 25, 2012, and revision received Mar. 19, 2012.