Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Multiscale classification and its application to process monitoring*

Yu-ming LIU, Lu-bin YE, Ping-you ZHENG, Xiang-rong SHI, Bin HU, Jun LIANG†‡

(*Institute of Industrial Control, State Key Lab of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China*)

†E-mail: jliang@iipc.zju.edu.cn

Received July 15, 2009; Revision accepted Oct. 26, 2009; Crosschecked May 4, 2010

Abstract: Multiscale classification has potential advantages for monitoring industrial processes generally driven by events in different time and frequency domains. In this study, we adopt stationary wavelet transform for multiscale analysis and propose an applicable scale selection method to obtain the most discriminative scale features. Then using the multiscale features, we construct two classifiers: (1) a supported vector machine (SVM) classifier based on classification distance, and (2) a Bayes classifier based on probability estimation. For the SVM classifier, we use 4-fold cross-validation and grid-search to obtain the optimal parameters. For the Bayes classifier, we introduce dimension reduction techniques including kernel Fisher discriminant analysis (KFDA) and principal component analysis (PCA) to investigate their influence on classification accuracy. We tested the classifiers with two simulated benchmark processes: the continuous stirred tank reactor (CSTR) process and the Tennessee Eastman (TE) process. We also tested them on a real polypropylene production process. The performance comparison among the classifiers in different scales and scale combinations showed that when datasets present typical scale features, the multiscale classifier had higher classification accuracy than conventional single scale classifiers. We also found that dimension reduction can generally contribute to a better classification in our tests.

Key words: Multiscale analysis, Stationary wavelet transform, Multi-class classifier, Feature extraction, Process monitoring
doi:10.1631/jzus.C0910430 **Document code:** A **CLC number:** TP277

1 Introduction

Process monitoring is increasingly in demand to guarantee a safe and productive production for modern industrial processes. To apply a specific analytical model or to do further fault diagnosis in practical process monitoring, a large amount of acquired process data usually need to be classified into known operating conditions or fault types. Consequently, classification methods for industrial processes have gained much attention in recent years.

The current investigations into the classification problem were mainly focused on single scale classification, related to the following methods: Fisher discriminant analysis (FDA), support vector machine (SVM), discriminant partial least square (DPLS), kernel Fisher discriminant analysis (KFDA), distance-based fuzzy C-means (FCM), the probabilistic clustering approach, the limited finite Gaussian mixture model (GMM), etc. (Chiang *et al*., 2000; 2004; He *et al*., 2005; 2008; Detroja *et al*., 2006; Wang *et al*., 2006; Yu and Qin, 2008). However, in fact, data from real industrial processes are multiscale in nature since the processes are generally driven by events located in different time and frequency domains. Hence, a multiscale analysis on the acquired data should have the advantage over single scale analysis for providing useful modeling information. To this end, Bakshi (1998) proposed a wavelet-based multiscale principal component analysis (PCA), which takes the coefficients at each scale into account and has been regarded as the framework of later multiscale analysis based process monitoring (Misra *et al*., 2002; Aradhye *et al*., 2003; 2004; Li *et al*., 2004; Yoon and

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 60574047), the National High-Tech R & D Program (863) of China (Nos. 2007AA04Z168 and 2009AA04Z154), and the Research Fund for the Doctoral Program of Higher Education in China (No. 20050335018)

MacGregor, 2004; Reis and Saraiva, 2006; Reis *et al.*, 2008). However, most of the multiscale methods have been employed for fault detection while little work has been undertaken for fault diagnosis or classification problems. Zhou *et al.* (2005) combined multiscale PCA with the hidden Markov model (HMM) for fault identification, but they focused on HMM modeling without detailed investigation on scale feature extraction or scale selection. Moreover, HMM has a limited application range. Woody and Brown (2007) presented three methods in detail to select the wavelet transform scales for multi-class classifier design. Unfortunately, they used only relatively simple classifiers, i.e., a linear discriminate classifier and a *K*-nearest neighbor classifier, and the application objects are mostly unrelated to typical chemical processes.

In this study, considering the multiscale nature of industrial processes and being motivated by previous works on applying multiscale methods to fault detection and classification, we aim to investigate the effects of multiscale feature extraction on classifier performance and to provide methods to design applicable multiscale multi-class classifiers for a better fault or operating condition classification.

2 Methodology

Our methodology is illustrated in Fig. 1. The essentials of this methodology has two aspects. One is how to extract and select discriminative multiscale features, and the other is how to use these features to design a multiscale classifier. We will describe them in detail below.

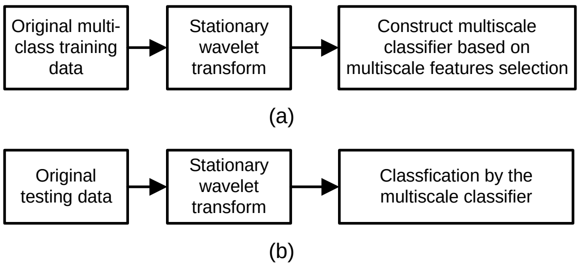
Original multi-class training data
Stationary wavelet transform
Construct multiscale classifier based on multiscale features selection
(a)
Original testing data
Stationary wavelet transform
Classfication by the multiscale classifier
(b)

Fig. 1 Schematic diagram of our multiscale classification method
(a) Training the classifier; (b) Testing the classifier

2.1 Multiscale feature extraction

2.1.1 Stationary wavelet transform

For the purpose of multiscale classification, we choose stationary wavelet transform (SWT, also referred to as translation-invariant or maximum overlapping wavelet transform) for multiscale analysis as suggested in Aradhye *et al.* (2003) and Woody and Brown (2007). Compared with the traditional discrete orthogonal wavelet transform (DOWT), SWT can guarantee both a translation-invariant performance for a circularly shifted signal and a good alignment between the original signal and the transformed coefficients (including SWT wavelet coefficients, details, and approximations) (Percival and Walden, 2000). In particular, the transform coefficients at each level in SWT are not down-sampled as those in DOWT, but keep the same length as the original signal. Assume that a signal *s* with a length of *m* is analyzed by *L*-level ($L\le\log_2 m$, $L\in\mathbb{Z}$) SWT. The corresponding wavelet function can be expressed as

$$\psi_{a,b}(t)=\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \tag{1}$$

where $\psi(t)$ is the mother wavelet, *a* is the scaling (or dilation) parameter, and *b* is the translation parameter. In DOWT, *a* and *b* are discretized into $a=2^j$ and $b=2^jk$ ($1\le j\le L$; $j, k\in\mathbb{Z}$), respectively, and the corresponding discretization wavelet function can be represented by

$$\psi_{j,k}(t)=2^{-j/2}\psi(2^{-j}t-k). \tag{2}$$

This means *b* is down-sampled when *j* changes to *j*+1. However, in SWT, *b* is discretized into *k* at each scale. Thus, the down-sampling does not occur, guaranteeing that the number of the transformed coefficients at each level is the same as that of the original signal.

We select SWT details and approximation as the features for classification. We refer to D_j and A_j as the *j*th level wavelet details and approximation, respectively. Consequently, the original signal *s* can be represented by

$$s(t)=\sum_{j=1}^{L}D_j(t)+A_L(t). \tag{3}$$

Daubechies (DB) and symlets (SYM) wavelets can be used for the multiscale feature extraction as suggested in Percival and Walden (2000) and Aradhye *et al*. (2003). To select the suitable associated wavelet filter, we consider mainly the wavelet filter with the shortest width, which can give reasonable results. This consideration is due to the fact that wavelet filters with very short widths can sometimes introduce undesirable effects into the resulting analysis, whilst the filters with a large width can provide a better match to the characteristic features but they may result in, for example, more coefficients being influenced by boundary conditions (Percival and Walden, 2000). In our applications, we test DB and SYM wavelet filters with different filter widths, and finally choose the symlets-4 (SYM4) wavelet whose associated filter is a near-linear-phase filter with length 8.

2.1.2 Scale selection

The aim of scale selection is to find the scale or scales combination containing the most discriminative features. As for chemical process data, the approximation at the coarsest scale generally reflects the dynamic trend of the original variable, while the details reflect, for example, sensor and process oscillations in different scales. Thus, all of them should be involved in scale selection.

We consider two schemes for such scale selection:

1. Voting scheme, similar to the one-vs.-one strategy for a training multi-class classifier. After adding a classifier of one scale, its 4-fold cross-validation accuracy is calculated, based on which a corresponding confidence weight is given to the classifier. For example, we can calculate a weight for the *i*th classifier by the following equation (Woody and Brown, 2007):

$$e_i = \ln C_{\mathrm{v}i}^2, \tag{4}$$

where $C_{\mathrm{v}i}$ is the cross-validation accuracy for the *i*th classifier. The weights can be applied to the testing observation to make the final classification.

2. Reconstruction scheme, where all of the possible reconstruction combinations of the transformed wavelet and scaling coefficients, for example, $2^{L+1}-1$ reconstructed signal for *L*-level SWT, are used as a training dataset, and the reconstruction combination with the highest 4-fold cross-validation accuracy is regarded as the best scale combination. However, the voting scheme may fail when the classifiers at different levels have significantly different performances. This is common for process data, causing the classifier on the most discriminative scale to dominate the final classification performance, while the reconstruction scheme is essentially an exhaustive and time-consuming algorithm.

Therefore, we adopt a more practical reconstruction method: (1) Construct *L*+1 single-scale classifiers including *L* detail classifications at all levels and one approximation classifier at the coarsest level. (2) Select the classifier with the best accuracy according to the results of both cross-validation and testing data validation. (3) Add other scales one by one to the present scale(s) to check whether the overall classification accuracy is improved or not. If yes, we keep this scale and go on adding other scales one by one as mentioned. If no, it indicates that the optimal scale(s) is/are obtained, and this selection process is complete.

2.2 Multi-class classifier design

In this study, we construct two multi-class classifiers: the SVM classifier and the Bayes classifier. The SVM classifier is preferable for its small size and linearly inseparable datasets, and is particularly useful when in some cases only a few fault samples are available. The Bayes classifier design, combined with the dimension reduction technique (either PCA or KFDA) can be used for large size datasets, such as those under different operating conditions, based on which probability estimation is possible. We can expect that both PCA- and KFDA-based dimension reduction, used for linear and nonlinear feature extraction respectively, may transform the data to the reduced features along the relative discriminative directions and simplify the classifier design to some extent. Further, since SVM is a non-parametric data-driven method, it is also applicable for large size datasets.

2.2.1 SVM classifier design

SVM for two-class classification is the essential part and its goal is to find the representative training observations referred to as a ‘support vector’ to define two boundary hyperplanes with a maximum margin between them. With regard to linearly separable two-class classification, the training data can be represented by

$$(\boldsymbol{x}_i, y_i),\ \ i=1,2,\cdots,m,\ \ \boldsymbol{x}_i\in\mathbb{R}^n,\ \ y_i\in\{+1,-1\}, \quad (5)$$

whereby the SVM classifier can be expressed by a constrained optimization problem (Bian and Zhang, 2000):

$$\begin{aligned}&\min\varphi(\boldsymbol{w})=\|\boldsymbol{w}\|^2/2\\ &\text{s.t.}\ \ y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1\geq 0,\ \ i=1,2,\cdots,n,\end{aligned} \quad (6)$$

where $\boldsymbol{w}\cdot\boldsymbol{x}_i+b$ is the linear discriminant function, $\boldsymbol{w}$ is the weight vector, b is the threshold, and $\varphi(\boldsymbol{w})$ is the reciprocal of the margin. This constrained optimization problem can be converted into a relatively simple dual problem using the Lagrange multiplier method based on dual theory. The final decision function can be expressed as

$$f(\boldsymbol{x})=\operatorname{sgn}(\boldsymbol{w}^*\cdot\boldsymbol{x}+b^*)=\operatorname{sgn}\left(\sum_{i=1}^{s}\alpha_i^* y_i\boldsymbol{x}_i\cdot\boldsymbol{x}+b^*\right), \quad (7)$$

where $\boldsymbol{w}^*$ is the weight vector of the optimal classification plane, α_i^* is the optimal solution, and b^* is the classification threshold. If there are some inseparable observations, the generalized classification plane can be used to control the classification accuracy, which can be represented by

$$\sum_{i=1}^{n}y_i\alpha_i=0,\ \ 0\leq\alpha_i\leq C,\ \ i=1,2,\cdots,n. \quad (8)$$

For linearly inseparable two-class classification, a nonlinear transform can be used to convert it to a linearly separable problem in a higher dimensional space. The nonlinear transform from lower dimensional space to higher dimensional space can be expressed as $\varphi(\cdot)$: $\mathbb{R}^d\rightarrow\mathbb{R}^k$, implemented by the so-called kernel trick, which constructs a kernel function K as $K(\boldsymbol{x}_i, \boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)$ so that the inner product operation in higher dimensional space can be directly obtained using the values of original variables without needing an explicit expression of the nonlinear transform $\varphi(\cdot)$. The corresponding decision function can be expressed as

$$f(\boldsymbol{x})=\operatorname{sgn}\left(\sum_{i=1}^{s}\alpha_i^* y_i K(\boldsymbol{x}_i,\boldsymbol{x})+b^*\right). \quad (9)$$

In our applications, we prefer the RBF kernel function

$$k(\boldsymbol{x}_i,\boldsymbol{x}'_j)=\exp\left(-\gamma\left\|\boldsymbol{x}_i-\boldsymbol{x}'_j\right\|^2\right),\ \ \gamma>0, \quad (10)$$

where γ is a kernel parameter. Hence, the SVM classifier needs to determine two parameters, C and γ. In order to find the best two parameters, we use a 'grid-search' method according to the cross-validation accuracy as recommended in Hsu *et al*. (2008).

For multi-class classification, we adopt a one-vs.-one strategy for training the SVM classifier (Hsu and Lin, 2002), in which all possible two-class classifiers (there are $c(c-1)/2$ classifiers for c classes) are firstly trained, and then integrated together to classify a certain observation by means of voting. The voting procedure is carried out as follows. First the observation is tested in each two-class classifier. The class to which the observation is assigned will get one vote. Finally the observation is assigned to the class with most votes.

2.2.2 Bayes classifier design with dimension reduction

Assume a training dataset follows multivariate Gaussian distribution. Using a linear Bayes classifier the decision value of observation $\boldsymbol{x}$ can be calculated to determine the possibility as to whether it belongs to the ith class according to the following decision function (Woody and Brown, 2007):

$$g_i(\boldsymbol{x})=\bar{\boldsymbol{x}}_i\boldsymbol{\Sigma}_p^{-1}\boldsymbol{x}^{\mathrm{T}}, \quad (11)$$

where $\bar{\boldsymbol{x}}_i$ is the mean vector of observations of the ith class, and $\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}$ is the covariance matrix of the pooling dataset. The observation $\boldsymbol{x}$ will be assigned to the class with the maximum decision value $g_i(\boldsymbol{x})$, i=1, 2, …, c.

In case of a large number of variables, we can choose dimension reduction technologies including PCA and KFDA to the SWT details and approximations. PCA is a commonly used dimension reduction method, which can extract a linear relationship among variables (Russell *et al*., 2000). However, PCA cannot extract a nonlinear relationship and is unable to use known classification information. KFDA can extract a nonlinear relationship among variables; meanwhile it can take the classification information into account. Given the proper parameters, KFDA may project the variables along some discriminative directions, and thus KFDA has the potential advantage for better feature extraction.

KFDA is a more generalized FDA, which uses the kernel trick in the same manner as SVM in order to solve a linearly separable problem in a nonlinearly-transformed higher dimensional space (Baudat and Anouar, 2000). Conventional FDA reduces the dimensionality whilst it guarantees that the between-class distance is maximized and the within-class distance is minimized, which can be represented by an optimization problem (Russell *et al*., 2000):

$$\max\ J_{\mathrm{LF}}(\boldsymbol{p}) = \boldsymbol{p}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{p}\big/(\boldsymbol{p}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{p}), \tag{12}$$

where $\boldsymbol{S}_{\mathrm{b}}$ is the sum of the sampling covariance matrix of each class, i.e., the between-class covariance matrix, and $\boldsymbol{S}_{\mathrm{w}}$ is the within-class covariance matrix; $\boldsymbol{p}$ denotes the projection direction. Similarly, we can denote in the KFDA the nonlinear transform function as φ, which is not necessarily an explicit formula. Thus, the optimization problem can be expressed as (Baudat and Anouar, 2000)

$$\max\ J_{\mathrm{KF}}(\boldsymbol{p}) = \boldsymbol{p}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{b}}^{\varphi}\boldsymbol{p}\big/(\boldsymbol{p}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}}^{\varphi}\boldsymbol{p}), \tag{13}$$

where $\boldsymbol{S}_{\mathrm{b}}^{\varphi}$ and $\boldsymbol{S}_{\mathrm{w}}^{\varphi}$ represent the between-class and within-class covariance matrices in the higher dimension, respectively. This problem is equivalent to a generalized singular value decomposition (GSVD) problem (Baudat and Anouar, 2000):

$$\lambda\boldsymbol{S}_{\mathrm{w}}^{\varphi}\boldsymbol{p} = \boldsymbol{S}_{\mathrm{b}}^{\varphi}\boldsymbol{p}, \tag{14}$$

where λ is the singular value. To solve this GSVD, we need only to choose a kernel function to perform the nonlinear transform. In our applications, we prefer the Gauss kernel function with kernel width σ determined by cross-validation, and choose the number of projection directions as $c-1$; i.e., the data dimensions of the classification problem are reduced to $c-1$. We also assume that the data are kept centralized after nonlinear transform.

In addition, we notice that the Bayes-KFDA classifier is essentially a linear classifier. We need to undertake more investigation to determine how much the KFDA-based nonlinear dimension reduction can affect a nonlinear classification problem.

3 Applications

The proposed multiscale methods as well as the conventional single-scale methods were tested and compared with one another in three typical chemical processes, including two simulated benchmark problems—the continuous stirred tank reactor (CSTR) process and the Tennessee Eastman (TE) process, and one real industrial polypropylene production process. The classifiers were implemented under MATLAB. The general procedure of these applications is as follows:

1. Select three classes named C1, C2, and C3 in the three applications. Data from each application have unique characteristics; that is, the data are characterized by a specific frequency band, simulated random and step disturbance, and real industrial noise, respectively.

2. Use three-level SWT with the ‘SYM4’ wavelet to extract scale based features, where the training and testing datasets both contain 256 continuous observations.

3. Apply a multiscale SVM classifier for the features obtained in Step 2 to determine the optimal scale(s) and scale-combination with higher testing accuracy.

4. Apply the Bayes classifier to the typical scale selections obtained in Step 3.

5. Make a performance comparison amongst multiscale classifiers with or without dimension reduction and single scale classifiers. Also make a performance comparison with the multiscale *K*-nearest neighbor (KNN) classifier presented in Woody and Brown (2007).

3.1 Continuous stirred tank reactor process

The nonisothermal CSTR process simulated in our study has a first order reaction and its reaction temperature is under feedback control. This simulator has been extensively used in fault detection and diagnosis and a detailed description can be observed in

Yoon and MacGregor (2001). To illustrate the multi-scale classification performance for datasets with typical frequency features, we considered the following three classes: (1) C1, normal data; (2) C2, sensor oscillation fault at reactor inlet temperature (denoted as T_i) with a period of 150 s and a magnitude of 0.4, expressed as $d_1(t)=0.4\sin(2\pi t/150)$; (3) C3, sensor oscillation fault at cooling water temperature (denoted as T_c) with a period of 300 s and a magnitude of 1, expressed as $d_2(t)=\sin(2\pi t/300)$.

We selected nine variables including the flow rate of reactant, cooling water, and solvent, the inlet concentration of reactant and solvent, the outlet concentration of the product, and the inlet temperature, reaction temperature, and cooling water temperature. The sampling time is one minute. Fig. 2 shows the inlet temperature and cooling water temperature for normal and oscillation datasets. Table 1 shows the parameters and the performance of the SVM classifier. Table 2 shows the classification accuracy of both the Bayes classifiers and the KNN classifiers.

As shown in Table 1, both the D_1 SVM classifier and the D_2 SVM classifier have a higher classification accuracy (67.06% and 59.90% respectively) than the original signal (i.e., $A_3+D_3+D_2+D_1$) SVM classifier (57.16%). Moreover, when D_1 and D_2 are combined, the classification accuracy is further increased to 75.52%. Thus, we can assert that levels 1 and 2 are the discriminative scales. Table 2 shows that the D_1+D_2 Bayes-KFDA classifier has the highest classification accuracy of 72.93% amongst all of the Bayes classifiers and KNN classifiers, which is comparable to that of the previous SVM classifier. The Bayes classifiers with PCA-dimension reduction and without dimension reduction have a far lower accuracy than the previous two classifiers. The reason may be that they do not take the class information into account.

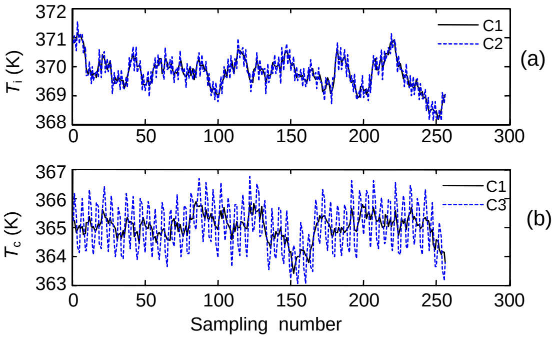

Fig. 2 Comparison between normal and oscillation data for (a) inlet temperature and (b) cooling water temperature

C1: normal data; C2: $d_1(t)=0.4\sin(2\pi t/150)$; C3: $d_2(t)=\sin(2\pi t/300)$

Table 1 SVM three-class classifier for the CSTR process

Scale selection	C	γ	Accuracy of 4-fold CV (%)	Testing accuracy (%)
A_3	2^3	2^{-10}	27.47	33.20
D_1	2^{10}	2^{-3}	62.63	67.06
D_2	2^3	2^{-3}	55.49	59.90
D_3	2^{10}	2^{-12}	28.39	33.33
D_1+D_2	2^9	2^{-5}	74.87	75.52
Original signal	2^{10}	2^{-7}	44.27	57.16

CV: cross-validation. D_j and A_j are the jth level wavelet detail and approximation, respectively

3.2 Tennessee Eastman process

The TE process problem, proposed by Downs and Vogel (1993), simulates a real industrial chemical process and is more complicated than the previous CSTR process. The TE simulator has 52 process variables and 21 fault types. Recently, many researchers have used the TE process as a crucial bench- mark problem to illustrate their methods for fault or operating condition classification. The faults include step change, slow drift, random fluctuation, valve stick, and unknown disturbance. Since some of the faults are overlapped, making an accurate classification for them is a challenge. With reference to Chiang *et al.* (2004) and to easily do a comparison among different classifiers, we selected three overlapped faults, i.e., 4th, 9th, and 11th faults, labeled C1, C2, and C3 respectively (Table 3). Both the training

Table 2 Classification accuracy of Bayes and KNN classifiers on typical scale selections for the CSTR process

Scale selection	Classification accuracy (%)				
	Bayes-No DR	Bayes-PCA*	Bayes-KFDA**	KNN-No DR	KNN-PCA*
D_1	33.07	33.20(4)	66.02(1.1)	57.55	49.74(4)
D_1+D_2	33.33	33.46(4)	72.93(1.4)	64.71	59.11(4)
Original signal	33.46	32.55(6)	56.12(48.5)	46.74	35.67(6)

* Within the brackets is the number of remaining principal elements; ** Within the brackets is the kernel width. DR: dimension reduction. D_j: the jth level wavelet detail

and testing data were chosen from the standard datasets, containing 256 continuous (i.e., 201–456) observations representing the stable condition of each fault. Fig. 3 shows that the 9th variable (V9, reactor temperature) for fault 4 and both V9 and the 51st variable (V51, cooling water valve position) for fault 11 have significant changes as disturbances are introduced after the 160th sampling time point. Table 4 shows the parameters and the performance of the SVM classifier. Table 5 shows the classification accuracy of both the Bayes classifiers and the KNN classifier.

Table 4 shows that, among the SVM classifiers using single scale information, the A_3 classifier has the highest testing accuracy of 53.39%, and amongst all of the SVM classifiers combining A_3 with other details, the A_3+D_3 classifier has the highest testing accuracy of 56.38%, which is higher than that of the original signal classifier, 50.65%. This indicates that discriminative features lie in both A_3 (the profile of the signal) and D_3 (the most disturbances within the relatively low frequency band), while details of other scales contain much higher frequency noises, which decreases the classification accuracy. Table 5 shows that both the Bayes-PCA and Bayes-KFDA classifiers for A_3 and A_3+D_3 have a higher accuracy than KNN classifiers, and the Bayes-PCA classifier has a higher accuracy than the Bayes-KFDA classifier. The possible reason may be that the Bayes classifier assumes features following the Gaussian distribution, but this assumption might not be satisfied after a KFDA-based dimension reduction. Since PCA is a linear transform which does not have influences on the data

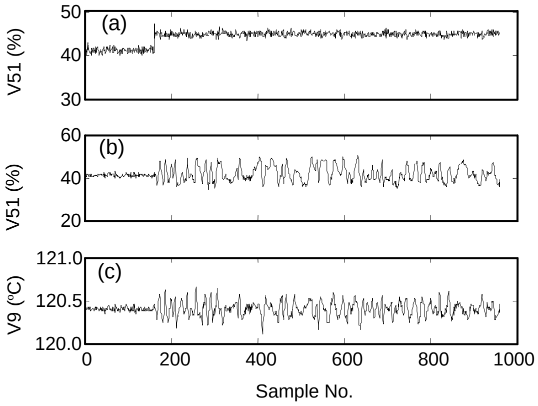

Fig. 3 (a) Cooling water valve position (V51) for C1; (b) Cooling water valve position (V51) for C2; (c) Reactor temperature (V9) for C3

C1, C2, and C3 are the 4th, 9th, and 11th overlapped faults, respectively

Table 3 The selected three classes in the TE process

Label	Fault	Change location	Type	Training size	Testing size
C1	IDV 4	Temperature of the reactor cooling water	Step	256	256
C2	IDV 9	Temperature of the feed D	Random	256	256
C3	IDV 11	Inlet temperature of the reactor cooling water	Random	256	256

C1, C2, and C3 are the 4th, 9th, and 11th overlapped faults, respectively

Table 4 SVM three-class classifiers on typical scale selections for the TE process

Scale selection	C	γ	Accuracy of 4-fold CV (%)	Testing accuracy (%)	Scale selection	C	γ	Accuracy of 4-fold CV (%)	Testing accuracy (%)
A_3	2^2	2^{-8}	93.10	53.39	A_3+D_2	2^2	2^{-7}	93.36	53.91
D_1	2^{10}	2^{-7}	50.65	33.72	A_3+D_1	2^0	2^{-6}	91.67	50.26
D_2	2^{10}	2^{-6}	81.12	35.16	A_3+D_3+D_2	2^{-1}	2^{-5}	92.84	48.83
D_3	2^{10}	2^{-4}	99.74	49.48	No selection	2^{10}	2^{-6}	91.15	50.65
A_3+D_3	2^1	2^{-7}	92.19	56.38					

CV: cross-validation. D_j and A_j are the jth level wavelet detail and approximation, respectively

Table 5 Classification accuracy of Bayes and KNN classifiers on typical scale selections for the TE process

Scale selection	Classification accuracy (%)				
	Bayes-No DR	Bayes-PCA*	Bayes-KFDA**	KNN-No DR	KNN-PCA*
A_3	36.98	49.48(19)	42.32(26.5)	40.36(19)	42.19(19)
A_3+D_3	41.80	47.79(19)	43.88(38.2)	43.10(19)	43.22(19)
Original signal	46.74	40.86(28)	48.05(32.1)	37.11(28)	47.00(28)

* Within the brackets is the number of remaining principal elements; ** Within the brackets is the kernel width. DR: dimension reduction. D_3 and A_3 are the 3rd level wavelet detail and approximation, respectively

distribution, the Gauss assumption will not show poor deviation because of dimension reduction. Further, the Bayes classifier without dimension reduction has substantially lower classification than the classifier with either KFDA- or PCA-based dimension reduction. In addition, we noticed that the SVM classifier has the best classification accuracy amongst all of the classifiers.

We also noticed that the classification accuracy of the single scale SVM here is not equal to that presented in Chiang *et al*. (2004), although they are comparable. The main reason could be that we have used different training and testing datasets, rather than the datasets in Chiang *et al*. (2004), which include the in-transition stage data, whilst we intended to select as much stationary data as possible in our application.

3.3 Polypropylene process

The polypropylene production device in this study is operating in a large scale industrial process (Fig. 4). This process includes four reactors, labeled R200, R201, R202, and R401, and one attachment unit, labeled D201. In this application, we selected reactor R202 as the monitored process, in which most polymers are produced. Based on the process principle, after canceling the variables related to production rate changes such as the propylene feeding rate, hydrogen feeding rate, and jacket flow rate, which may otherwise impair the significances of the discriminative variable, we selected 10 key variables in R202 as the modeling variables (Table 6). In addition, we selected three of the typical grades that can be produced through this polypropylene process, i.e., T36F, EPS30R, and EPS30RA, labeled as C1, C2, and C3 respectively. The sampling time was 5 min. The bivariate plots of four grade-sensitive variables in Fig. 5 show that there are some overlapping observations among C1, C2, and C3.

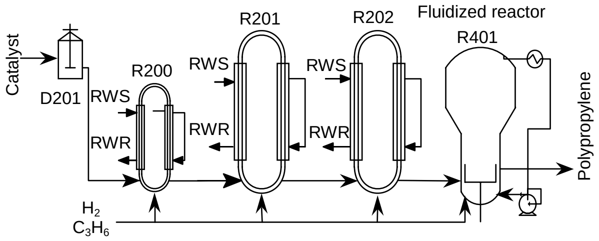

Fig. 4 Illustrative flow sheet of a polypropylene production process
RWS: recycled water supply; RWR: recycled water return

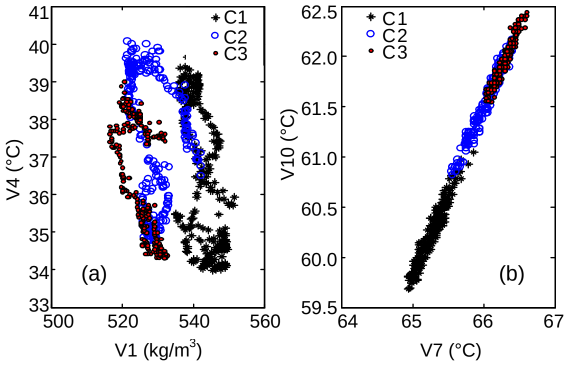

Fig. 5 Bivariate plots for (a) density in R202 (V1) vs. propylene feeding temperature (V4) and (b) cooling water temperature (V7) vs. jacket temperature of R202 (V10)
C1: T36F; C2: EPS30R; C3: EPS30RA

Table 6 The selected 10 variables in R202 of the polypropylene process

No.	Name (unit)	Description
V1	DRC251PV, density in R202 (kg/m^3)	Measured variable, whose mean varies among different grades and has a specific change trend
V2	FRCA251PV, propylene for rinsing (kg/hr)	Controlled variable, which fluctuates around the set point
V3	JRA251PV, power of P202 (kW)	Measured variable, whose mean varies among different grades and has a specific change trend
V4	TIA231PV, propylene feeding temperature (°C)	Measured variable, whose mean varies among different grades and has a specific change trend
V5	TIA256, wall surface temperature (°C)	Measured variable, which fluctuates within the predetermined range
V6	TIA 258, wall surface temperature (°C)	Measured variable, which fluctuates within the predetermined range
V7	TR253PV, cooling water temperature (°C)	Measured variable, whose mean varies among different grades and has a specific change trend
V8	TRA273PV, overall cooling water temperature (°C)	Measured variable, whose mean varies among different grades and has a specific change trend
V9	TRCA251PV, temperature of R202 (°C)	Controlled variable, which fluctuates around a set point
V10	TRCA 252PV, jacket temperature of R202 (°C)	Measured variable, whose mean varies among different grades and has a specific change trend

Table 7 shows the parameters and the performance of the SVM classifier. Table 8 shows the classification accuracy of both the Bayes classifiers and the KNN classifier.

Table 7 SVM three-class classifier for the polypropylene process

Scale selection	C	γ	Accuracy of 4-fold CV (%)	Testing accuracy (%)
D_1	2^9	2^{-2}	41.67	41.28
D_2	2^9	2^1	47.53	35.81
D_3	2^{10}	2^2	57.94	34.64
A_3	2^1	2^{-10}	95.18	80.08
A_3+D_3	2^1	2^{-10}	95.44	80.34
A_3+D_2	2^1	2^{-10}	95.31	80.21
A_3+D_1	2^1	2^{-10}	95.05	79.42
A_3+D_3+D_2	2^1	2^{-10}	93.88	79.56
Original signal	2^1	2^{-10}	94.79	78.65

CV: cross-validation. D_j and A_j are the jth level wavelet detail and approximation, respectively

Table 7 shows that amongst the SVM classifiers using single scale information, the A_3 classifier has the highest testing accuracy of 80.08%, and amongst all of the SVM classifiers combining A_3 with other details, the A_3+D_3 classifier has the highest testing accuracy of 80.34%, which is higher than that of the original signal classifier, 78.65%. These results are similar to those in the previous TE process which indicate that the discriminative features lie in A_3 and D_3, whilst details at other scales contain more high-frequency noises, which therefore decreases the classification accuracy. Table 8 shows that the Bayes-PCA for A_3 and A_3+D_3 has a higher accuracy than other Bayes classifiers and the KNN classifiers. The reason may be that PCA effectively reduced the noises in addition to dimension reduction by means of keeping the proper number of principal components whilst the KFDA, in this case, did not keep the most discriminative reduced features. The best Bayes-KFDA classifier (on selection A_3+D_3) is comparable to the corresponding SVM classifier. This may imply that, given the Gaussian distribution assumption, the Bayes-KFDA classifier may be equivalent to the SVM classifier to some degree. We can expect that the SVM classifier with PCA dimension reduction in this case would also improve its performance. In addition, we found that the Bayes classifier without dimension reduction had substantially lower classification accuracy than the Bayes classifier with either KFDA- or PCA-based dimension reduction.

4 Conclusions and discussion

Based on SWT, we combined scale feature extraction with multi-class classification design and tested the multiscale SVM classifier without dimension reduction and the linear Bayes classifier with KFDA- or PCA-based dimension reduction in three cases. In our applications, we found that:

1. Compared with single-scale classification, the multiscale classification can utilize extra scale information, thus providing a way to construct a potentially better classification for process monitoring. The results verified their advantages over single scale methods, particularly when the process data have typical scale or frequency features or significant frequency noises.

2. The SVM classifier is more general than the linear Bayes classifier, but has more parameters to be determined and therefore is time consuming, while the linear Bayes classifier can simplify the training process using Gauss distribution, and when combined with suitable dimension reduction, the Bayes classifier may have a comparative performance to the SVM classifier. Moreover, multiscale SVM and/or Bayes classifiers generally have higher classification accuracy than the KNN classifiers.

3. Dimension reduction using KFDA or PCA can contribute to a better classification, in that it can reduce not only the number of training features but also the data noises to some degree. Moreover, the KFDA can further take both the nonlinear relationship among variables and the class information into

Table 8 Classification accuracy of Bayes and KNN classifiers on typical scale selections for polypropylene

Scale selection	Classification accuracy (%)				
	Bayes-No DR	Bayes-PCA*	Bayes-KFDA**	KNN-No DR	KNN-PCA*
A_3	76.95	88.15(3)	80.34(1.7)	76.04	80.34(3)
A_3+D_3	77.34	88.15(3)	80.86(1.2)	76.30	80.34(3)
Original signal	76.82	88.02(4)	76.43(5.6)	76.30	83.33(4)

* Within the brackets is the number of remaining principal components; ** Within the brackets is the kernel width. DR: dimension reduction. D_3 and A_3 are the 3rd level wavelet detail and approximation, respectively

account, but the number of reduced features needs to be properly determined.

Obviously, the multiscale classification methods presented in this paper can be applied to relatively large multi-class classification problems. However, although we can obtain some benefits from multiscale classification, the corresponding computation time would increase rapidly if more levels (such as five or more) are to be considered. This would inevitably result in an exponential increase in the number of possible scale combinations. The current work is looking for other more efficient optimization algorithms to select the desirable scales or scale combination and to choose optimal parameters for the classifiers with or without dimension reduction. Another concerning issue about the multiscale classification is its generalization ability. The current method has limitations for systems with time-dependent characteristics. Further investigation may lie in modeling wavelet coefficients at each level, for example, constructing discriminative scale features (Reis and Bauer, 2009), for a better classification.

References

Aradhye, H.B., Bakshi, B.R., Strauss, R.A., Davis, J.F., 2003. Multiscale SPC using wavelets: theoretical analysis and properties. *AIChE J.*, **49**(4):939-958. [doi:10.1002/aic.690490412]

Aradhye, H.B., Bakshi, B.R., Davis, J.F., Ahalt, S.C., 2004. Clustering in wavelet domain: a multiresolution ART network for anomaly detection. *AIChE J.*, **50**(10):2455-2466. [doi:10.1002/aic.10245]

Bakshi, B.R., 1998. Multiscale PCA with application to multivariate statistical process monitoring. *AIChE J.*, **44**(7):1596-1610. [doi:10.1002/aic.690440712]

Baudat, G., Anouar, F.E., 2000. Generalized discriminant analysis using a kernel approach. *Neur. Comput.*, **12**(10):2385-2404. [doi:10.1162/089976600300014980]

Bian, Z., Zhang, X., 2000. Pattern Recognition (2nd Ed.). Tsinghua University Press, Beijing, p.298-299 (in Chinese).

Chiang, L.H., Russell, E.L., Braatz, R.D., 2000. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemometr. Intell. Lab. Syst.*, **50**(2):243-252. [doi:10.1016/S0169-7439(99)00061-1]

Chiang, L.H., Kotanchek, M.E., Kordon, A.K., 2004. Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Comput. Chem. Eng.*, **28**(8):1389-1401.

Detroja, K.P., Gudi, R.D., Patwardhan, S.C., 2006. A possibilistic clustering approach to novel fault detection and isolation. *J. Process Control*, **16**(10):1055-1073. [doi:10.1016/j.jprocont.2006.07.001]

Downs, J.J., Vogel, E.F., 1993. A plant-wide industrial process control problem. *Comput. Chem. Eng.*, **17**(3):245-255. [doi:10.1016/0098-1354(93)80018-I]

He, Q.P., Qin, S.J., Wang, J., 2005. A new fault diagnosis method using fault directions in Fisher discriminant analysis. *AIChE J.*, **51**(2):555-571. [doi:10.1002/aic.10325]

He, X.B., Yang, Y.P., Yang, Y.H., 2008. Fault diagnosis based on variable-weighted kernel Fisher discriminant analysis. *Chemometr. Intell. Lab. Syst.*, **93**(1):27-33. [doi:10.1016/j.chemolab.2008.03.006]

Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. *IEEE Trans. Neur. Networks*, **13**(2):415-425. [doi:10.1109/72.991427]

Hsu, C.W., Chang, C.C., Lin, C.J., 2008. A Practical Guide to Support Vector Classification. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf [Accessed on May 21, 2010].

Li, J., Zhang, J., Ge, W., Liu, X., 2004. Multi-scale methodology for complex systems. *Chem. Eng. Sci.*, **59**(8-9):1687-1700. [doi:10.1016/j.ces.2004.01.025]

Misra, M., Yue, H.H., Qin, S.J., Ling, C., 2002. Multivariate process monitoring and fault diagnosis by multi-scale PCA. *Comput. Chem. Eng.*, **26**(9):1281-1293. [doi:10.1016/S0098-1354(02)00093-5]

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, Cambridge, p.160, 195-200.

Reis, M.S., Bauer, A., 2009. Wavelet texture analysis of on-line acquired images for paper formation assessment and monitoring. *Chemometr. Intell. Lab. Syst.*, **95**(2):129-137. [doi:10.1016/j.chemolab.2008.09.007]

Reis, M.S., Saraiva, P.M., 2006. Generalized multiresolution decomposition frameworks for the analysis of industrial data with uncertainty and missing values. *Ind. Eng. Chem. Res.*, **45**(18):6330-6338. [doi:10.1021/ie051313b]

Reis, M.S., Saraiva, P.M., Bakshi, B.R., 2008. Multiscale statistical process control using wavelet packets. *AIChE J.*, **54**(9):2366-2378. [doi:10.1002/aic.11523]

Russell, E.L., Chiang, L.H., Braatz, R.D., 2000. Data-Driven Methods for Fault Detection and Diagnosis in Chemical Process. Springer, London, p.64, 103-107.

Wang, H., Li, P., Gao, F., Song, Z., Ding, S.X., 2006. Kernel classifier with adaptive structure and fixed memory for process diagnosis. *AIChE J.*, **52**(10):3515-3531. [doi:10.1002/aic.10982]

Woody, A.A., Brown, S.D., 2007. Selecting wavelet transform scales for multivariate classification. *J. Chemometr.*, **21**(7-9):357-363. [doi:10.1002/cem.1060]

Yoon, S., MacGregor, J.F., 2001. Fault diagnosis with multivariate statistical models: part I. using steady state fault signatures. *J. Process Control*, **11**(4):387-400. [doi:10.1016/S0959-1524(00)00008-1]

Yoon, S., MacGregor, J.F., 2004. Principal-component analysis of multiscale data for process monitoring and fault diagnosis. *AIChE J.*, **50**(11):2891-2903. [doi:10.1002/aic.10260]

Yu, J., Qin, S.J., 2008. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE J.*, **54**(7):1811-1829. [doi:10.1002/aic.11515]

Zhou, S., Xie, L., Wang, S., 2005. On-line fault diagnosis in industrial process using variable moving window and hidden Markov model. *Chin. J. Chem. Eng.*, **13**(3):388-395.