Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 13 (20): 4195-4200, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.4195.4200

Multi-Kernel PCA with Discriminant Manifold for Hoist Monitoring

[1,2]Shijin Ren, [1]Ping Ling, [1]Maoyun Yang, [1]Yinlong Ni and [2]Zhihuan Song
[1]School of Computer Science and Technology, Jiangsu Xuzhou Normal University, Xuzhou, Jiangsu, 221116, China
[2]National Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China

Abstract: Dimensional reduction is crucial to machine condition monitoring and diagnosis since the extracted features are often redundant and heterogeneous as well as high dimensional data often are embedded in lower-dimensional manifold. Inspired by manifold learning and multiple-kernel learning theory, multi-kernel principal component analysis with discriminant manifold (DMMKPCA) is proposed for host fault monitoring and diagnosis. The method not only preserves the local and global structures of data set but also handles heterogeneous characteristic sets, which inherits the excellences of LPMVP and multiple kernel learning. A two-stage iterative optimization algorithm is proposed to obtain the optimal combining weights of multiple kernel functions and parameters of each kernel functions. A case study of hoist illustrates the efficiency of the proposed algorithm on the information extraction and fault detection.

Key words: Manifold learning, multiple kernel learning, multi-kernel principal component analysis with discriminant manifold (DMMKPCA), Hoist, fault detection

INTRODUCTION

The research on fault monitoring and diagnosis is the hot topics in the mechanical community due to the wide applications of large scale machine equipments. In the past decade, many successful applications of the machinery monitoring have been reported (Lei *et al*., 2011; He, 2013; Wang *et al*., 2011). Most machinery monitoring researches mainly focus on the removal of undesired uncertainties, extracting features that capture the underlying operation patterns of the machine and online fault detection methods like pattern classifiers in a supervised way and fault monitoring statistics like Multivariate Statistical Process Control (MSPC) in an unsupervised way.

In machinery monitoring, features are extracted from time and frequency domains of vibration signals which are are heterogeneous and mutually complementary. The underlying signal can be adequately exploited by a linear combination of kernels at different scales can effectively enhance performance of kernel machines (Lin *et al*., 2011). Although these features extracted from. Although these features extracted from mechanical signals are high dimensionality, those often lie in a nonlinear manifold with low dimensionality (Wang, 2010; Jiang *et al*., 2009). Manifold learning can explore the low-dimensional latent representation corresponding to high-dimensional data, which is an efficient dimensionality reduction approach, successfully applied into machinery monitoring and process monitoring (Wang, 2010; Jiang *et al*., 2009; Yu, 2012).

Inspired from manifold learning and Multi-Kernel Learning (MKL), a novel multiple-kernel principal component analysis with discriminant manifold algorithm (DMMKPCA) is proposed. Firstly, through multiple explicit empirical feature mapping functions, the original data with heterogeneous characteristic are mapped to multiple high- dimensional feature spaces which can be regarded as various views of input data. Secondly, features are extracted by manifold learning and the latent data in the low-dimensional space approximately follow Gaussian distribution. Meanwhile, the global and local geometrics of the data are preserved. In order to remove the redundant kernel functions, an optimization algorithm is proposed. Finally, the feasibility and performance of the proposed method for hoist fault detection are illustrated.

DMMKPCA ALGORITHM

As a traditional multivariate projection method, although PCA can preserve the global geometric structure of data and discover the manifold which lies in data set, it does not describe the manifold. Locality Preserving Projection (LPP) can describe the low dimensional manifold, which takes the local information of data set into account. However, the global or non-local information in

Corresponding Author: Shijin Ren, School of Computer Science and Technology, Jiangsu Normal University, 221116, Xuzhou, China

data set may be lost (Yu, 2012; Zhang and Song, 2009). DMMKPCA is proposed to reduce the dimensionality of high-dimensional data with heterogeneous feature sets by integrating manifold learning with multiple kernel learning.

Given a matrix $X = [x_1, x_2, ..., x_n]$ with $x_i \in R^d$, a high-dimensional feature mapping function, $\phi(x)$, corresponds to a positive definite kernel function, k(.,.) which is computed by:

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

where, $\phi(x_i)$ is assumed to be centered. DMMKPCA intends to seek a projection matrix $W^\phi \in R^{R^{N_{FHM}}}$ in feature space such that the local geometry of the reduction data:

$$Y = \left[W^{\phi T}\phi(x_i)\right]_{i=1}^{n} \in R^M$$

is similar to that of original data. Thus the global information of data set can be preserved as well simultaneously.

Laplacian adjacent matrix is constructed to approximate the local geometry of points, which is defined by []:

$$S_{ij}^w = \begin{cases} \exp\left(\dfrac{\left\|\phi(x_j)-\phi(x_i)\right\|^2}{-t}\right), x_j \in kN(x_i) \text{ or } x_i \in kN(x_j) \\ 0 \quad\quad \text{otherwise} \end{cases}$$

where kN (x_i) is the neighborhood of x_i, S^w_{ij} represents the similarity between the two neighbor samples and t is a constant associated with samples. The optimization problem of DMMKPCA is formulated as follows:

$$J_{Local}\left(w_k^\phi\right) = \min_{w_k^\phi} \sum_{i,j=1}^{N} \left\|y_i^k - y_j^k\right\|^2 S_{ij}^w \tag{1}$$

$$J_{Global+NonLocal}\left(w_k^\phi\right) = \max_{w_k^\phi} \sum_{i=1}^{N} \left(y_i^k - \mu_y^k\right)^2 \tag{2}$$
$$\lambda \in [0,1]$$

$$\text{s.t.} \begin{cases} \left\|w_k^\phi\right\|^2 = 1 \\ \left\langle w_k^\phi, w_j^\phi \right\rangle = 0, j = 1, 2, \cdots, k-1 \end{cases} \tag{3}$$

Where:

$$\mu_y^k = \sum_{i=1}^{N} w_k^\phi \phi(x_i) \Big/ N$$

The value of S^w_{ij} increases with their similarity. Therefore, the local information of data set is preserved by exploiting the Laplacian matrix.

The above problems is rewritten as:

$$\sum_{i=1}^{n}\left(y_i^k - \mu_y^k\right)^2 = \sum_{i=1}^{n}\left(w_k^\phi \phi(x_i) - \mu_y^k\right)^2 = w_k^{\phi T} C^\phi w_k^{\phi T} \tag{4}$$

$$\sum_{i,j=1}^{N}\left\|y_i^k - y_j^k\right\|^2 S_{ij}^w = w_k^{\phi T} F\left(D^w - S^w\right) F^T w_k^\phi \tag{5}$$

Where:

$$D^w = \text{diag}\left[D_{11}^w, \cdots, D_{NN}^w\right],\ D_{ii}^w = \sum_{j=1}^{N} S_{ij}^w$$

and $L^W = D^W - S^W$ The two above optimization problems can be integrated into an overall optimization problem with constraints in Eq. 3, which is formulated as:

$$\max_{w_k^\phi} J\left(w_k^\phi\right) = \frac{w_k^{\phi T} C^\phi w_k^\phi}{w_k^{\phi T} F\left(D^w - S^w\right) F^T w_k^\phi} \tag{6}$$

The advantage of doing this is that the projection vectors are uncorrelated and the geometrical structure of low-dimensional representations is consistent with that of original input data. The derivation can be briefly deduced as follows. Denote the solution of optimization problem by W, the distance between two lower-dimensional representations, y_i and y_j, is calculated as follows:

$$\text{Dist}\left(y_i, y_j\right) = \left\|y_i - y_j\right\| = \sqrt{\left(x_i - x_j\right)^T WW^T \left(x_i - x_j\right)}$$

Note that WW^T is an identity matrix, which leading to Dist $(y_i, y_j) = \|x_i - x_j\|_2$. Generally, the formulation of $\phi(\cdot)$ is implicit and the dimensionality of w^ϕ_k is very high, even infinite. As a result, Eq. 6 is not intractable. To address this issue an explicit Empirical Kernel Map (EKM) (Zhu *et al.*, 2009) is introduced. Let $K = Q\Lambda Q^T$ be the rank-r SVD of a kernel matrix K, i.e. $Q \in R^{N\times r}$, $B = KQ\Lambda^{-1/2}$, the explicit kernel map $\phi^e(\cdot)$ is defined as:

$$\phi^e(x) = \Lambda^{-1/2} Q^T \left[\ker(x, x_1), \cdots, \ker(x, x_N)\right]^T$$

The dot product on EKM matrix is calculated by $\langle \Phi^e, \Phi^e \rangle = K$, indicating that span$\{\phi^e(x_i)\}$. Whilst, the empirical kernel space is easier to interpret and optimize.

For k = 1, the Lagrangian equation corresponding to the above optimization problem in the absence of the constraints can be written as:

$$J\left(w_k^\phi\right) = w_k^{\phi T} C^\phi w_k^\phi - \gamma\left(w_k^{\phi T} F\left(D^w - S^w\right) F^T w_k^\phi\right) \tag{7}$$

Setting derivatives of the Lagrange with respect to W^ϕ_k zero, we have:

$$\delta J\left(w_k^\phi\right)/\delta w_k^\phi = 2C^\phi w_k^\phi + -2\gamma\left(F\left(D^w - S^w\right)F^T w_k^\phi\right) = 0 \quad (8)$$

The optimum w_1^ϕ can be achieved by solving the following generalized eigenvalue problem:

$$\left(F\left(D^w - S^w\right)F^T\right)^{-1} C^\phi w_k^\phi = \gamma w_k^\phi \quad (9)$$

where w_1^ϕ is the eigenvector corresponding to the largest eigenvalue of matrix.

$$\left(F\left(D^w - S^w\right)F^T\right)^{-1} C^\phi$$

For $k \geq 2$, the optimal w^ϕ_k is obtained in a similar way. For the sake of simplicity, we define $Le = C^\phi$ and $Ri = \Phi(D^w\text{-}S^w)\Phi^T$. The Lagrangian function associated with (6) and (3) is written as:

$$J\left(w_k^\phi\right) = w_k^{\phi T} Le w_k^\phi - \gamma w_k^{\phi T} R_i w_k^\phi - \sum_{i=1}^{k-1} \alpha_i \left(w_i^{\phi T} w_k^\phi\right) \quad (10)$$

The Lagrangian derivatives with respect to w^ϕ_k and α_i are set to zero, then we get:

$$\alpha_i = w_i^{\phi T} L_e w_k^\phi - \gamma w_i^{\phi T} R_i w_k^\phi \quad (11)$$

Let:

$$W_{(k-1)} = \left[w_1^\phi, w_2^\phi, \cdots, w_{k-1}^\phi\right]$$

and $G^{(k-1)} = W_{(k-1)}^T R_i^{-1} W_{(k-1)}$. Substitute (11) into (10) and note that $w_j^{\phi T} w_i^\phi = 0$, $j \neq i, i, j = 1, 2, \cdots, k-1$, we get:

$$\left\{I - R_i^{-1} W_{(k-1)} G^{(k-1)T} W_{(k-1)}^T\right\} R_i^{-1} L_e w_k^\phi = \gamma w_k^\phi \quad (12)$$

where, I is an identity matrix. Given a new sample x, the projected data is calculated by

$$y = W^T \phi^e(x) = W^T \Lambda^{-1/2} Q^T \left[\ker(x, x_1), \cdots, \ker(x, x_N)\right]^T$$

SELECTION OF DMMKPCA MODEL PARAMETERS

MKL can be deemed as a relaxed version of kernel method by replacing a kernel function with a linear combination of multiple kernels, dealing with the problem of the choice of kernel function in kernel methods. Specifically, given L base kernel functions $k_l(x_i, x_j)$ satisfying mercer condition, a linear combination of these base kernel functions is defined as:

$$k\left(x_i, x_j\right) = \sum_{l=1}^{L} \beta_l k_l\left(x_i, x_j\right)$$

with $\beta_l \geq 0$ and:

$$\sum_{l=1}^{L} \beta_l = 1$$

Geometrically, the linear combination of kernels can be interpreted as taking the Cartesian product of the associated feature spaces. From view of multi-view learning (Wang *et al.*, 2008), MKL can capture the different ‘views’ of the problem or features from different sources (Fig. 1, 2). Compared to wavelet transform, MKL can also decompose data set into different scales (Fig. 2). In this article, we make use of these two properties of multi-kernel learning to enhance the capability to deal with nonlinear heterogeneous features.

In order to remove redundant kernel functions, CCA is utilized to extract the common information from two views with different and specific sets of features by maximizing all the correlations between the projections of pairwise feature sets, which can be equivalently transform into a distance minimization problem as follows (Wang *et al.*, 2008):

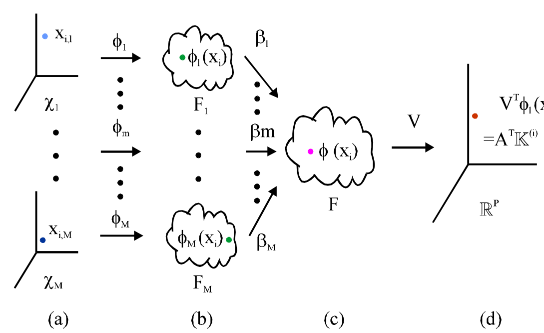

Fig. 1: MKL for different representation

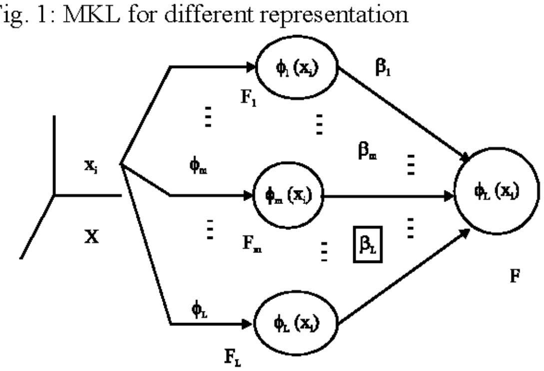

Fig. 2: Multi-resolution analysis of MKL

$$\min_{\beta_l} \sum_{k,l=1,k\neq l}^{M} \left\|\beta_k K^{(k)} - \beta_l K^{(l)}\right\|_F^2, \text{ s.t. } \sum_{k=1}^{M} \beta_k = 1$$

where, M is the number of specific feature sets.

In simulation, the initial values of kernel parameters are as follows: Calculate the standard variance σ_0 of data set, then parameters of base RBF kernels are set as: $\sigma_1 = \sigma_0 * i$, i = 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5.

DMMKPCA APPROACH TO HOIST MONITORING

In multivariate statistical control, conventional process monitoring methods, such as PCA, LPMVP, etc., often use fault statistics T^2 and SPE under the assumption that the latent vectors are normally distributed with zeros mean. Given a sample x_i, its low-dimensional representation is $y_i = W^{\Phi T} \phi(x_i)$ through DMMKPCA. Original variable $\phi(x_i)$ can be decomposed into principal space and residual space, respectively. Following conventional multivariate statistic process monitoring methods, we develop Hotelling T^2 and SPE monitoring statistics on Y and $\tilde{F}$ respectively. T^2 statistic on Y is calculated by:

$$T^2 = y^T \Lambda^{-1} y \leq F_{d,N-d,\alpha}\, d(N-1)/(N-d)$$

where $\Lambda = YY^T/(N-1)$ is covariance matrix, $F_{d,N\text{-}d,\alpha}$ is F-distribution with d and N-d degrees of freedom at the significance level α. Under the assumption that a residual vector is multivariate normal, Q-statistic SPE on $\tilde{F}$ is calculated by:

$$SPE = e^T e \sim g\chi_h^2 \leq SPE_\alpha = g\chi_{h,\alpha}^2$$

where, $g\chi^2_{h,a}$ is χ^2-distribution with scaling factor g and h degrees of freedom at a significance level α (Xie and Shi, 2012). For Q-statistic, $g = S/2\mu$ and $h = 2\mu^2/S$ are estimated based on the matching moments between a$g\chi^2_h$ distribution and the reference distribution of $\tilde{\Phi}$, where μ, S are mean and variance of $\tilde{\Phi}$. If the statistics of a new sample fall into these limits, the process is considered to be in control statistically.

SIMULATIONS

Hoist plays a very important role in coal production. The task of hoist is to pull coarse coal and workers from ground to coal, or vice verse. Hoist system is compose of motors, control unit etc (Liu *et al*., 2012a). In this study, only speed reducer is studied to illustrate the efficiencies of DMMKPCA for hoist monitoring.

The vibration signals with 3 working loads are collected from three sensors with sampling rate 10 Hz in normal operations, the sampling time lasts 200 sec for each mode. Some of measurements are illustrated in Fig. 3. Two types of fault data associated with crack of outer and inner races are obtained from history database. The number of data associated with each type of fault is 10000x200=2000000, which are illustrated in Fig. 4.

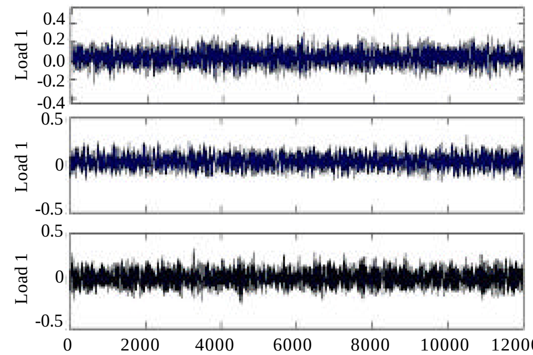

Fig. 3: Vibration signals with three working loads under normal conditions

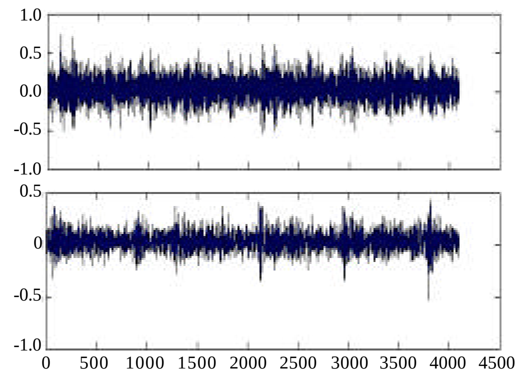

Fig. 4: Two types of fault (cracks of outer and inner races)

Before the application of the proposed algorithm, the vibration signals must be preprocessed, such as removing noises from signals, normalization etc. In this simulation, we generate the following different representations (Liu *et al*., 2012b; Ghoraani and Krishnan, 2011; Zhu *et al*., 2009):

- **Time representation:** Standard variation, correlation coefficient, Kurtosis, Skewness and fractal coefficient are generated from time domain of signals
- **Frequency presentation:** HHT is used to decompose signals into multiple scales and find "true" intrinsic mode functions (IMFs) (Liu *et al*., 2012b). Then, features such as amplitude, phase, transient frequency, weighted frequency, average amplitude,

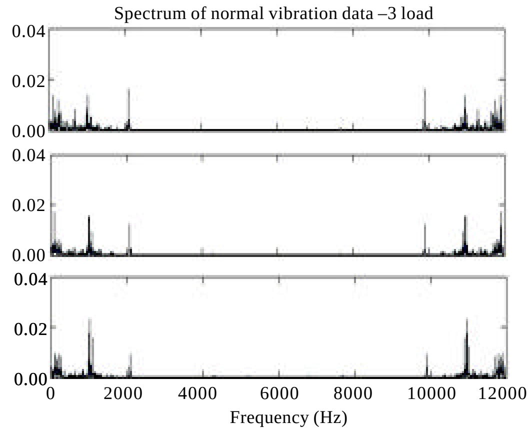

Fig. 5: Spectrum of normal signals with three working loads

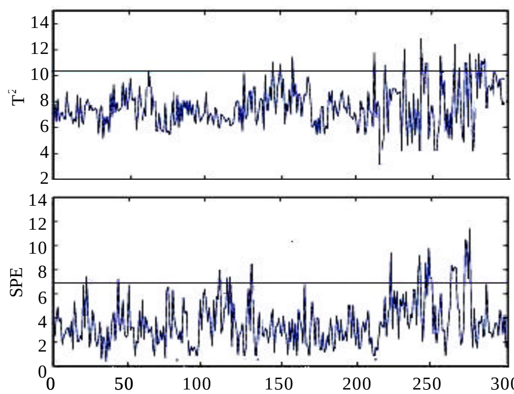

Fig. 6: Monitoring results of DMMKPCA

Table 1: Data representations

Representation	No. of variables
Time	6
Frequency	11
Time series	(AR) 5

Kurtosis, moment energy and residual energy are generated from those selected IMFs

- **Time series model representation:** Robust time series AR is used to model vibration signals, the associated AR coefficients and fitting residual energy are taken as features

To study the characters of vibration signals, frequency spectrum of the signals are obtained by FFT (Fig. 5). It is obvious that most of spectrum concentrates in 1000 Hz and double-frequency 2000 Hz. For the sake of simplicity, 2048 points are regarded as one unit and features are generated from one unit. Table 1 illustrates the number of dimensions associated with above data representations. We establish 440 samples for 3 normal working loads and 40 samples for each type of fault.

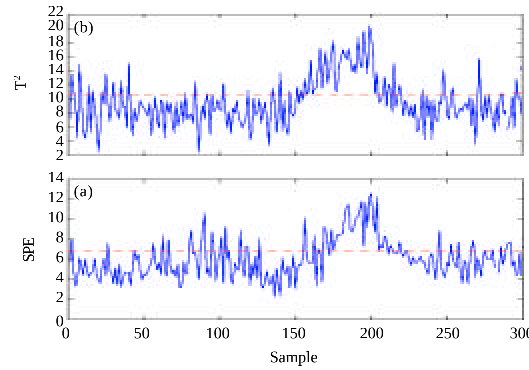

Fig. 7: Monitoring results of KPCA

In this simulation, Gaussian kernel functions are used as base kernel functions and half of training samples are randomly sampled from normal samples, the remains are used as test samples. Five base kernel functions are used for each data representation. The number of samples in neighborhood set is 4. In DMMKPCA, the weights of kernel functions are initialized to 1/12, the maximum iteration is 100. After the optimal weights are obtained, we remove the redundant kernel functions whose weight is less than 0.001, 11 kernel functions are retained to perform DMMKPCA. The number of reduced dimensionality is 7 which is selected by cross validation.

Monitoring results of DMMKPCA on 220 normal samples and 80 fault samples are shown in figure 6. Figure 6 shows T^2 on normal samples falls into the control limit while T^2 on fault samples is outside of control limit. To measure the performance of fault detection, detection rate is defined as the percentage of the samples outside control limit at a significance level 99%. The detection rate of T^2 on fault samples is 52% while detection rate on normal test is 4%. The detection rate of T^2 on fault samples is 43% while detection rate on normal test is 5%. Furthermore, To evaluate the fault detection capability of the proposed method, we compare our algorithm with KPCA. The monitoring result of KPCA is demonstrated in Fig. 7. the detection rates of T^2 or SPE on normal samples (i.e. alarm rate)is over 25% while that on fault data is less than 20%. Irrespective of the alarm rate and detection rate on fault samples, the detection ability of the proposed method outperforms that of KPCA.

CONCLUSION

As a novel nonlinear dimension reduction algorithm, DMMKPCA can exploit different complementary representations and explore local and global geometrics of data set and work well on complex data from multiple sources. The projected data in latent space approximately follow Gaussian distribution, which is very convenient to develop fault statistics and perform online fault monitoring. The model complexity is controlled by removing redundant kernel functions as well. The present work highlights the promise of DMMKPCA approach to machinery monitoring. However, it should be further investigated as to how to determine the number of reduced dimensionality and the proper parameters of kernel functions.

ACKNOWLEDGMENT

This study was financially supported by National Science funder of China with Grant No. 60974056.

REFERENCE

Ghoraani, B. and S. Krishnan, 2011. Time-frequency matrix feature extraction and classification of environmental audio signals. IEEE Trans. Audio Speech Language Process., 19: 2197-2210.

He, Q., 2013. Time-frequency manifold for nonlinear feature extraction in machinery fault diagnosis. Mech. Syst. Signal Process., 35: 200-218.

Jiang, Q., M. Jia, J. Hu and F. Xu, 2009. Machinery fault diagnosis using supervised manifold learning. Mech. Syst. Signal Process., 23: 2301-2311.

Lei, Y., Z. He and Y. Zi, 2011. EEMD method and WNN for fault diagnosis of locomotive roller bearings. Expert Syst. Appl., 38: 7334-7341.

Lin, Y.Y., T.L. Liu and C.S. Fuh, 2011. Multiple kernel learning for dimensionality reduction. IEEE Trans. Pattern Anal. Machine Intell., 33: 1147-1160.

Liu, X.P., G.Y. Xu and S.J. Ren, 2012a. Research on hoist machinery fault monitoring based on HHT-DDKICA and SVDD method. Chin. J. Machinery Eng., 39: 21-32.

Liu, L., W. Zuo, D. Zhang, N. Li and H. Zhang, 2012b. Combination of heterogeneous features for wrist pulse blood flow signal diagnosis via multiple kernel learning. IEEE Trans. Inform. Technol. Biomedicine, 16: 598-606.

Wang, Z., S.C. Chen and T. Sun, 2008. MultiK-MHKS: A novel multiple kernel learning algorithm. IEEE Trans. Pattern Anal. Mach. Intell., 30: 348-353.

Wang, G.B., 2010. Manifold learning based fault diagnosis for rotating machine. Ph.D. Thesis, Central South University.

Wang, S., W. Huang and Z.K. Zhu, 2011. Transient modeling and parameter identification based on wavelet and correlation filtering for rotating machine fault diagnosis. Mech. Syst. Signal Process., 25: 1299-1320.

Xie, X. and H. Shi, 2012. Dynamic multimode process modeling and monitoring using adaptive gaussian mixture models. Ind. Eng. Chem. Res., 51: 5497-5505.

Yu, J., 2012. Semiconductor manufacturing process monitoring using gaussian mixture model and bayesian method with local and nonlocal information. IEEE Trans. Semiconduct. Manuf., 25: 480-493.

Zhang, M.G. and Z.H. Song, 2009. LPMVP algorithm and its application to fault monitoring. ACTA Automatic Sinica, 35: 766-771.

Zhu, W., X.F. Hu and Z. Zhen, 2009. MPCA-AR-based Fault predictionalgorithm. Chin. J. Device Instrument, 3: 1778-1782.