



Journal of Applied Sciences

ISSN 1812-5654

science
alert

ANSI*net*
an open access publisher
<http://ansinet.com>

Improved Evolving Kernel of Fisher's Discriminant Analysis for Classification Problem

¹Hanaa E. Sayed, ²Hossam A. Gabbar and ¹Shigeji Miyazaki

¹Department of Intelligent Systems Engineering,

Okayama University, 1-1 Tsushima Naka 3-chome, 700-8530 Okayama, Japan

²Faculty of Energy Systems and Nuclear Science, UOIT, ON L1H 7K4 Ontario, Canada

Abstract: This study presents a kernel approach called Improved Evolving Kernel Fisher's Discriminant Analysis (IE-KFDA), which chooses kernel function parameters using proposed Gaussian mutation operator. Evolutionary Programming (EP) known by enhancing the search performance without increasing the computational time. An integration of KFDA with evolutionary optimization algorithm is presented. Classification example shows improving classification results using IE-KFDA in classifying different groups of data over KFDA, KPCA and PLS.

Key words: Genetic algorithms, optimization, search engine, gaussian operator, mutation operation, multivariate analysis, nonlinear analysis, faults diagnosis

INTRODUCTION

Classification problem as per definition is the problem of separating objects into smaller classes and giving criteria for determining whether a particle object in the domain is in a particular class or not. One of the most famous applications of classification is in biology, which is Carolus Linnaeus famous classification of living things by class, order, genus and species. Another application in physics is the classification of the physical world into matter and energy, or more precisely, the classification of elementary particles into fermions and bosons and the further classification of these particles using quantum numbers. In industrial applications there is a famous example, fault diagnosis problem. Fault diagnosis is the problem of classifying new fault into one of predefined faults groups and hence we can know causes, consequences and the fault mode. Pattern recognition, features extractions, signal processing and image classification also need classification methods to solve them.

Many multivariate analysis techniques have been used in classification problems and features extraction. Linear Discriminant Analysis (LDA) is a traditional statistical method that has proved successful on classification problems (Fukunaga, 1990). The procedure is based on an eigenvalue resolution and gives an exact solution of the maximum of the inertia. Most known methods Principal Component Analysis (PCA), Fisher's Discriminant Analysis (FDA), Partial Least Squares (PLS) and Discriminant Partial Least Squares (DPLS)

(Kemsley, 1996; Kourti and MacGregor, 1996). PCA has played a prominent role in features extraction and classification problems. Although that, FDA has been shown recently that it gives better results than PCA (Chiang *et al.*, 2000). Also, Hudlet and Johnson (1977) and Duda and Hart (1973) showed that FDA provides an optimal lower dimensional representation in terms of discriminating among classes of data. Due to the linearity nature of FDA, a lot of classification mistakes occur. Most real-world modeling problems are not linear. Baudat and Anouar (2000) has proposed kernel version for FDA for two-class and multi-class problems. Kernel idea solved the problem of nonlinearity but still choosing the kernel function in order to obtain the optimal nonlinear feature transformation is an open research problem. For an introduction and review on existing methods of classification and pattern recognition (Devijver and Kittler, 1982; Fukunaga, 1990; McLachlan, 2004).

This study tackles this issue using an optimization method, evolutionary computation. Evolutionary computation, such as Genetic Programming (GP), Genetic Algorithms (GA), Evolutionary Programming (EP) and Evolutionary Strategies (ES), has been known as a powerful intelligent optimization methodology. It has been extensively utilized to optimize more complicated real-world problems in engineering and industry, such as scheduling and planning. In this research, we propose an improved evolutionary algorithm to solve the optimal selection problem of kernel when applying to KFDA. Based on earlier research work which proofed that FDA is outperforming PCA, KPCA is the nonlinear solution for

PCA (Scholkopf *et al.*, 1999, 1998) and KFDA is the nonlinear solution of FDA.

Kernel fisher’s discriminant analysis: Let us consider a set of m observations in an n -dimensional space. And the column vector $x_i \in R^n$, where, $i = 1, \dots, m$, is the transpose of the i th row of $X \in R^{n \times m}$, where, X_h is the subset containing m_h samples that belong to the fault h . Thus,

$$X = \bigcup_{h=1}^c X_h \text{ and } m = \sum_{h=1}^c m_h$$

where, c is the number of fault classes. For a given nonlinear mapping Φ , the input space R^n can be mapped into feature space F , $\Phi: R^n \rightarrow F, x \rightarrow \Phi(x)$. F can have a much higher dimensionality could reach to infinity.

The objective of KFDA is to find certain directions in the original variable space, along with latent groups or clusters in R^n are discriminant as clearly as possible. Kernel FDA performs FDA in the feature space F , which is nonlinearly related to the input space R^n . As a result, kernel FDA produces a set of nonlinear discriminant vectors in the input space. The discriminant weight vector is determined by maximizing between-class matrix S_b^Φ while minimizing total scatter matrix S_t^Φ , which are defined in F as follows:

$$S_b^\Phi = \frac{1}{m} \sum_{i=1}^c c_i (m_i^\Phi - m^\Phi)(m_i^\Phi - m^\Phi)^T \quad (1)$$

$$S_t^\Phi = \frac{1}{m} \sum_{i=1}^m c_i (\Phi(x_i) - m^\Phi)(\Phi(x_i) - m^\Phi)^T \quad (2)$$

Where:

- m_i^Φ : Represents the mean vector of the mapped observations of class i
- m^Φ : The mean vector of the mapped m observations
- c_i : The number of observations of class i
- C : The total number of class of $x_h, h = 1, \dots, m$

The idea of KFDA is to solve the problem of FDA in the feature space F . This can be achieved by maximizing the following Fisher criterion (Baudat and Anouar, 2000):

$$J^\Phi(\varphi) = \frac{\varphi^T S_b^\Phi \varphi}{\varphi^T S_t^\Phi \varphi}, \varphi \neq 0 \quad (3)$$

The intuition behind maximizing $J(\varphi)$ is to find a direction which maximizes the projected class means (the numerator) while minimizing the classes variance in this direction (the denominator).

The optimal discriminant vectors can be expressed as a linear combination of the observations in the feature space F , i.e.

$$\varphi = \sum_{i=1}^m \alpha_i (x_i) = Q\alpha \quad (4)$$

where, $Q = [\Phi(x_1), \dots, \Phi(x_m)]$ and $\alpha = (\alpha_1, \dots, \alpha_m)^T$. For a given kernel matrix K ($k_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle, j = 1, \dots, m$). Substituting Eq. 4 into Eq. 3, we have:

$$J^\Phi(\alpha) = \frac{\alpha^T (KWK)\alpha}{\alpha^T (KK)\alpha} \quad (5)$$

where, $W = \text{diag}(W_1, \dots, W_c)$ and W_1 is an $n_1 \times n_1$ matrix with terms all equaling to $1/n_1$. Then, the optimal discriminant vectors in feature space are given by:

$$\varphi_i = Q \alpha_i = QP\Lambda^{-1/2} \beta_i, i = 1, \dots, d \quad (6)$$

where, $P = (p_1, \dots, p_r)$ and $\Lambda = (\lambda_1, \dots, \lambda_r)$. p_1, \dots, p_r are orthogonal eigenvectors of the matrix K corresponding to r nonzero eigenvalues $\lambda_1 \geq \dots \geq \lambda_r > 0$. β_1, \dots, β_d are actually the eigenvectors of $\Lambda^{1/2} P^T W P \Lambda^{1/2} - \Lambda$ corresponding to $d(d \leq c-1)$ largest eigenvalues.

Consequently, given a new sample x_{new} and its mapped observation $\Phi(x_{new})$, its KFDA discriminant score vector can be obtained as follows:

$$z = \varphi^T \Phi(x_{new}) = (\beta_1, \dots, \beta_d)^T \left(\frac{p_1}{\sqrt{\lambda_1}}, \dots, \frac{p_r}{\sqrt{\lambda_r}} \right)^T k_{new} \quad (7)$$

Where:

$$k_{new} = Q^T \Phi(x_{new}) = [k(x_1, x_{new}), \dots, k(x_n, x_{new})]$$

PROPOSED IMPROVED EVOLUTIONARY ALGORITHM

During the last 30 years there has been a growing interest in problem solving systems based on principles of evolution: such systems maintain a population of potential solutions, they have some selection process based on fitness of individuals and some recombination operators. One type of such systems is a class of evolution strategy i.e., algorithms which imitate the principles of natural evolution for parameters optimization problems. Fogel’s evolutionary programming (Fogel, 1999) is a technique for searching through a space of small finite-state machines. It maintains a population of reference points and generates offspring by weighted linear combinations. Another type of evolution based systems is Holland’s genetic algorithm

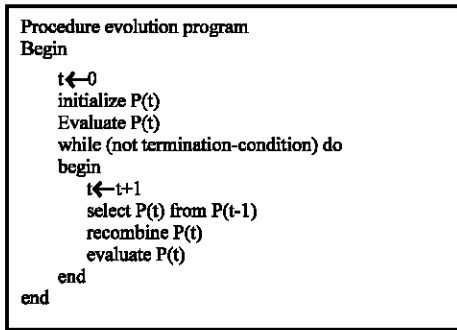


Fig. 1: Evolution program structure

(Goldberg, 1989; Holland, 1975). We use a common term, Evolutionary Programming (EP), for all evolution-based systems (including systems described above). The structure of an evolution program is shown in Fig. 1.

The evolution program is a probabilistic algorithm which maintains a population of individuals, $P(t) = \{x_1^t, \dots, x_n^t\}$ for iteration t . Each individual represents a potential solution to the problem at hand and in any evolution program, is implemented as some (possibly complex) data structure S . Each solution x_i^t is evaluated to give some measure of its fitness. Then, a new population (iteration $t+1$) is formed by selecting the more fit individuals (select step).

Some members of the new population undergo transformations (recombine step) by means of genetic operators to form new solutions. There are unary transformations m_i (mutation step), which create new individuals by a small change in a single individual ($m_i: S \rightarrow S$) and higher order transformations c_j (two or more) individuals ($c_j: S \times \dots \times S \rightarrow S$). After some number of generations the program converges - the best individual hopefully represents the optimum solution.

Genetic algorithms are famous of their global search capability. But usually they can't ensure that the choosing the best optimal from population in every generation will cover the global optimum. Miller *et al.* (1993) used three types of local search operators to decrease the issue of premature convergence. A new evolutionary algorithm of mutation in evolutionary strategies and evolutionary programming is presented in this study. And it will be shown how it is enhancing the performance of the global and local search capabilities. It uses a new mutation operator, new Gaussian mutation. And it is defined as follows:

$$\sigma_j = k \cdot \text{Min} \{x_j^{t-1} - a_j, b_j - x_j^{t-1}\} \left(1 - \frac{t}{M_g}\right)^{s-1} \left(1 + \frac{t}{M_g}\right), \quad j = 1, 2, \dots, M \quad (8)$$

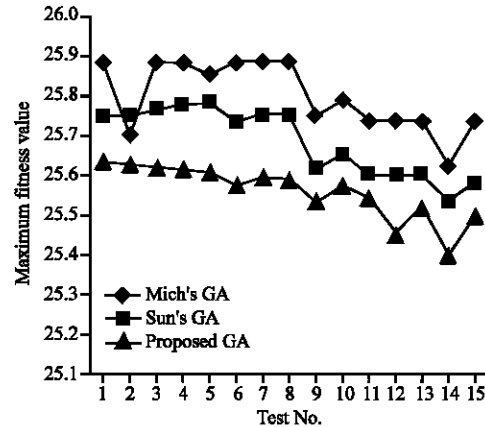


Fig. 2: Comparison of performance for Mich's GA, Sun's GA and proposed GA

where, k is a constant within the closed interval $[0,1]$; t is the generation; x_j^{t-1} is the j th variable to be optimized in the $(t-1)$ th generation; $[a_j, b_j]$ is the j th variable's scope; M_g is the maximum generation; s is a shape parameter and M is the number of variables.

The mutation of the j th variable x_j is given by:

$$x_j' = x_j + \epsilon_j, \quad j = 1, 2, \dots, M \quad (9)$$

$$\epsilon_j \sim N(0, \sigma_j) \quad (10)$$

where, ϵ_j is distributed as a Gaussian random number variable with zero mean and σ_j standard deviation.

We compare here the commonly used genetic operators designed by Michalewicz (1996), we call it Mich's GA and evolutionary algorithm of Sun *et al.* (2007) and we call it Sun's GA, with present proposed evolutionary algorithm Proposed GA when solving the following optimization problem:

$$\begin{aligned} \text{Max } f(x_1, x_2) &= 21.5 + x_1 \cdot \sin(4\pi x_1) + x_2 \cdot \sin(20\pi x_2) \quad (11) \\ -3.0 &\leq x_1 \leq 12.1 \\ \text{S.T.} \\ 4.1 &\leq x_2 \leq 5.8 \end{aligned}$$

A comparison of the maximum fitness value for the three methods is shown in Fig. 2. Outperformance for the proposed evolutionary algorithm is shown.

Results are obtained by making 15 runs and run the experiment using Mich's GA, Sun's GA and present improved Genetic Algorithm (Proposed GA) with fixed parameters and recording the results. And by making several experiments, the values of the control parameters of the improved evolutionary algorithm are shown in Table 1.

Table 1: Improved evolutionary algorithm's control parameters

Parameters	Values
Population size	15.00
Maximum generation	150.00
Crossover rate	0.70
Mutation rate	0.15
Gauss mutation rate	0.32
k in Eq. 8	1/6

IE-KFDA method for classification: The objective of KFDA is to find certain directions in the original variable space, along with latent groups or clusters in R^n are discriminated as clearly as possible. Kernel FDA performs LFDA in the feature space F , which is nonlinearly related to the input space R^n . Then, kernel FDA produces a set of nonlinear discriminant vectors in the input space. As a result, it is important to select an appropriate kernel function according to the initial data set before performing KFDA. There are many kernel functions exist and the most widely used are given by:

$$\text{Gaussian kernel: } k(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma}\right) \quad (12)$$

$$\text{Polynomial kernel: } k(x,y) = (x,y)^{\delta} \quad (13)$$

$$\text{Sigmoid kernel: } k(x,y) = \tanh(\tau_0(x,y)^{\delta} + \tau_1) \quad (14)$$

where, the parameters δ , τ_0 , τ_1 and σ have to be predetermined. We can use the proposed evolutionary algorithm to find the optimal parameter from the viewpoint of optimization. Thus, the fitness criterion is first required for evaluating individual parameters.

In the classification problem, we can measure the separability of the classes by the scatter matrix. Therefore, we use the following fitness function:

$$F_j = \ln \left| \frac{S_w}{S_t} \right| \quad (15)$$

where, S_w , S_t are the between-class scatter matrix and total scatter matrix, respectively. And they are defined in Eq. 1 and 2.

So, after choosing the best kernel function using the proposed evolutionary algorithm, we can apply KFDA on the data. And hence, we can extract features and obtain the classification results. Figure 3 show a flowchart for the described steps.

CLASSIFICATION EXAMPLE AND RESULTS

Using benchmark data sets from UCI and Statlog can be found freely on (<http://archive.ics.uci.edu/ml/datasets.html?format=andtask=claandatt=andarea=andnumAtt=a>

Table 2: Benchmark methods error comparison

Data sets	KPCA	KFDA	DPLS	IE-KFDA
Cancer	9.0±0.9	11.0±0.4	10.0±0.1	9.0±0.5
B. cancer	27.2±1.1	27.2±1.2	27.2±0.2	27.2±0.3
Diabetes	24.6±0.9	24.6±0.1	24.6±1.5	24.6±0.3
Image	3.5±1.1	3.5±1.4	3.5±0.3	3.5±0.1
Waveform	10.5±0.8	10.5±0.4	10.5±1.2	10.5±0.6
Twonorm	2.8±0.8	2.8±0.9	2.8±1.1	2.8±0.1
Titanic	23.0±0.7	23.0±1.4	23.0±0.2	23.0±0.5
Heart	17.4±0.8	17.4±1.1	17.4±0.8	17.4±0.3
Adult	2.2±1.5	2.2±0.9	2.2±2.1	2.2±0.2
Zoo	30.2±0.9	30.0±1.1	30.0±0.1	30.0±0.5
Gisette	15.1±1.4	15.1±0.3	15.1±0.9	15.1±0.1
Nursery	3.6±1.2	3.5±0.7	3.5±0.2	3.5±0.5
Coverttype	24.0±0.8	24.0±2.2	24.0±1.5	24.0±0.4
F. sonar	34.1±1.1	34.0±2.4	34.0±1.5	34.0±0.5
Thyroid	4.2±0.2	4.2±0.3	4.2±0.5	4.2±0.1

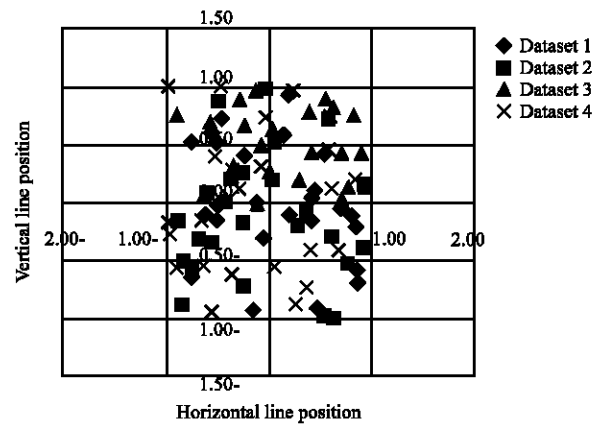


Fig. 3: Initial groups of data sets

ndnumIns=andtype=andsort=taskUpandview=table). We used 15 multivariate datasets. Datasets are from different categorical, real and integer data types with different distribution sets. We experienced different classification methods: KPCA, DPLS, KFDA and IE-KFDA, Table 2 shows the results of the experiments. It shows average test errors for different datasets and it is shown that IE-KFDA is showing better results for most of the datasets, next to it KFDA, then DPLS. We estimated the parameters using 7 fold cross validation on the first seven realizations of the training sets and took the model parameters to be the median over the seven estimates.

Then we choose another real example for faults database of experimental chemical process. We are going to show by graphs how the improved evolutionary algorithm is enhancing KFDA when it is solving a classification problem.

Let us have four classes of data inputs, which are normally distributed. In Fig. 3, a point plot of the initial data set is shown. The number of data points in each class is 25. Standard deviation is same for the four groups of data 0.22. The data is normalized to the area [0, 1] as

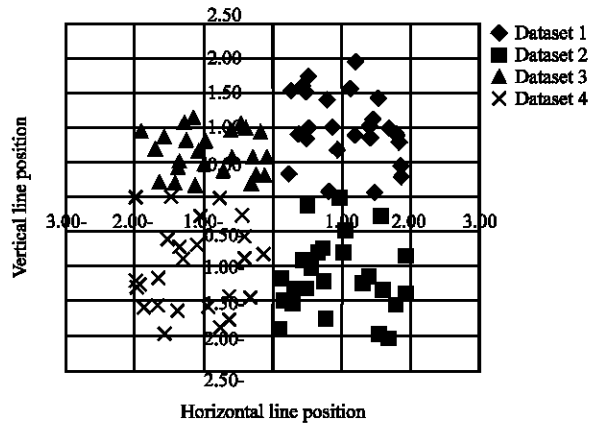


Fig. 4: KFDA classification

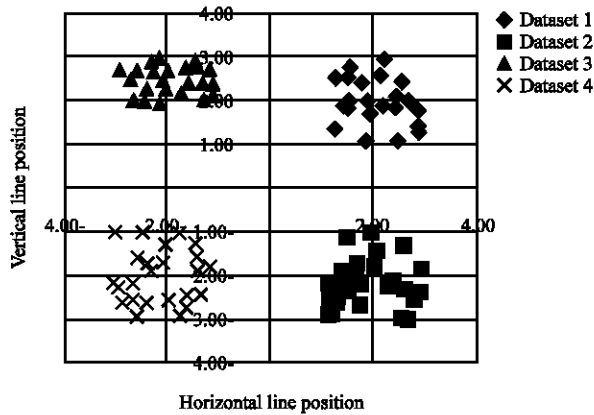


Fig. 5: IE-KFDA classification

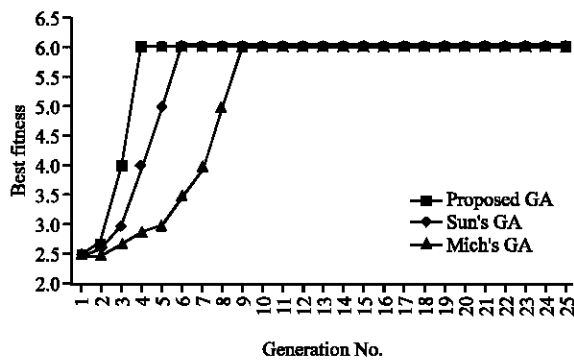


Fig. 6: Performance comparison between Mich's, Sun's and proposed GA

shown in Fig. 4. FDA, KFDA is performed on the same normalized groups of data. The fitness functions, F_j is calculated and the results of the classification is shown in Fig. 5. The fitness results comparison between Mich's GA, Sun's GA and present proposed Genetic Algorithm

is shown in Fig. 6, it shows that it is reaching the maximum fitness function at the least generation number, then follows it Sun's GA and finally Mich's GA.

CONCLUSION

Classification problem is considered one of the focus research areas nowadays. In this research we have improved the KFDA classification capability by adding improved genetic algorithm to it. This algorithm solves the issue of choosing kernel function before applying KFDA. Using the optimization searching capability of evolution algorithm, it searches for the best kernel function and chooses it. The classification case study shows how the proposed algorithm improved the classification results. Also comparing the number of generations in order to reach the best fitness function, we found that improved genetic algorithm needs less number of generations than using Sun's Genetic Algorithm and Michalewicz's Genetic Algorithm. And also it is shown how classification problem solution is enhanced using IE-KFDA for different benchmark and real data sets.

REFERENCES

Baudat, G. and F. Anouar, 2000. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12: 2385-2404.

Chiang, L.H., E.L. Russell and R.D. Braatz, 2000. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares and principal component analysis. *Chemomet. Intell. Lab. Syst.*, 50: 243-252.

Devijver, P. and J. Kittler, 1982. *Pattern Recognition: A Statistical Approach*. 1st Edn., Prentice Hall, London.

Duda, R.O. and P.E. Hart, 1973. *Pattern Classification and Scene Analysis*. 2nd Edn., John Wiley and Sons, New York, ISBN: 0-471-22361-1, pp: 482.

Fogel, L.J., 1999. *Artificial Intelligence through Simulated Evolution*. 3rd Edn., John Wiley and Sons, New York, ISBN: 0-471-33250-X, pp: 271-301.

Fukunaga, K., 1990. *Introduction to Statistical Pattern Recognition*. 2nd Edn., Academic Press, Orlando, FL, ISBN: 0-12-269851-7.

Goldberg, D.E., 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. 1st Edn., Addison-Wesley, New York, USA., ISBN: 0201157675.

Holland, J.H., 1975. *Adaptation in Natural and Artificial Systems*. 1st Edn., University of Michigan Press, Michigan, Ann, Arbor, ISBN: 0472084607.

- Hudlet, R. and R. Johnson, 1977. Classification and Clustering. Academic Press, New York, ISBN: 9780127142500, pp: 371-394.
- Kemsley, E.K., 1996. Discriminant analysis of high-dimensional data: A comparison of principal components analysis and partial least squares data reduction methods. *Chemomet. Intell. Lab. Syst.*, 33: 47-61.
- Kourti, T. and J.F. MacGregor, 1996. Multivariate SPC methods for process and product monitoring. *J. Quality Technol.*, 28: 409-428.
- McLachlan, G.J., 2004. Discriminant Analysis and Statistical Pattern Recognition. 1st Edn., John Wiley and Sons, UK., ISBN: 978-0-471-69115-0.
- Michalewicz, Z., 1996. Genetic Algorithms + Data Structures = Evolution Programs. 1st Edn., Springer, Verlag, Berlin.
- Miller, J.A., W.D. Potter, R.V. Gandham and C.N. Lapena, 1993. An evaluation of local improvement operators for genetic algorithms. *IEEE Trans. Syst., Man Cybernetics*, 23: 1340-1351.
- Scholkopf, B., A.J. Smola and K.R. Muller, 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10: 1299-1319.
- Scholkopf, B., S. Mika and C.J.C. Burges, 1999. Input space versus feature space in kernel-based methods. *IEEE. Trans. Neural Networks*, 10: 1000-1017.
- Sun, R., F. Tsung and Q.U. Liangsheng, 2007. Evolving Kernel principal component analysis for fault diagnosis. *Comput. Ind. Eng.*, 53: 361-371.