Neurocomputing 159 (2015) 219–226

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# A novel wrapper method for feature selection and its applications

Gang Chen *, Jin Chen

*State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, NO. 800 Dongchuan Road, Minhang District, Shanghai, China*

ARTICLE INFO

*Article history:*
Received 2 August 2014
Received in revised form
19 January 2015
Accepted 20 January 2015
Communicated by "Jiayu Zhou"
Available online 16 February 2015

*Keywords:*
Cosine similarity measure
Feature selection
Support vector machines
Bayesian interpretation

ABSTRACT

This paper introduces a wrapper method, namely cosine similarity measure support vector machines (CSMSVM), to eliminate irrelevant or redundant features during classifier construction by introducing the cosine distance into support vector machines (SVM). Traditionally, feature selection approaches typically extract features and learn SVM parameters independently or in the attribute space, which might result in a loss of information related to classification process or lead to the increase of classification error when introduce the kernel SVM. The proposed CSMSVM framework, however, jointly performs feature selection, SVM parameter learning and remove low relevance features by optimizing the shape of an anisotropic RBF kernel in feature space. Moreover, the Bayesian interpretation of the novel methodology reveals its Bayesian character, which builds the proposed method on solid theory foundation, and the iteration algorithm, which is proposed to optimize the feature weight, has achieved to maximize the maximum a posterior (MAP). Comparing the novel method with well-known feature selection techniques with experiments, CSMSVM outperformed the other methodologies in improving the pattern recognition accuracy with fewer features.



## 1. Introduction

Feature selection addresses the dimensionality reduction issue by determining a subset of the available features from large dimensionality domains via predetermined evaluation criteria. Thus feature selection is significance for reduction of the computational complexity and improvement of classifier's generalization ability. The reason for this is quite evident, since features irrelevant and redundant usually occur in application fields, especially in high-dimensional feature vectors or large dataset. Additionally, a low-dimensional representation will reduce the risk of over fitting [1,2]. Therefore, many researchers have spent their energy in the studying of feature selection process and have developed dozens of feature selection algorithms [3–7].

Based on the evaluation procedure, the existing feature selection methods can be sorted into three categories, namely the wrapper method, filter methods and hybrid methods. Filter methods perform the feature selection process independently without involving any learning algorithm [8], while wrapper methods [9] utilize a predetermined learning algorithm for feature subset evaluation which makes the final selected subset features be correlated with the chosen relevance measure and the hybrid methods combine the filter and wrapper methods, respectively. Intuitively, the hybrid methods are based on the other two methods and proposed to overcome the drawbacks of the filter and wrapper methods. Since the filter methods have low computational cost with the selected feature subset shows insufficient reliability for classification. In the other side, the wrapper approaches achieve superior classification accuracy, but need much more computational power. The drawbacks and complementarity of the two methods lead to the development of the hybrid method, such as the SAGA [5] and the normalized mutual information feature selection method which used a genetic algorithm to form a hybrid method called GAMIFS [10].

These existing methods, including the filter, the wrapper and the hybrid method, have improved the features' discrimination for classification. When they come to the classification algorithm itself, however, they have not overcome the classification algorithm's drawbacks. In the other words, the feature selection process has not enhanced the classification algorithm but enhanced the features. Moreover, the wrapper and hybrid method, even though they have achieved a high classification accuracy, they fail to address the computational efficiency. To deal with this issue, this paper proposes a cosine similarity measure support vector machines (CSMSVM) that selects relevant features during classifier construction by introducing the cosine distance into SVM. The CSMSVM not only optimizes the margin in SVM, but also decreases the intra-class distance during the feature selection process which will reduce the classification error rate. In terms of the computational effort, the CSMSVM performs feature selection process and classification simultaneously, which makes it evade a

* Corresponding author. Tel.: +86 158 0075 0723.
*E-mail addresses:* megangchen@gmail.com (G. Chen), Jinchen@sjtu.edu.cn (J. Chen).

http://dx.doi.org/10.1016/j.neucom.2015.01.070
0925-2312/

further validation step to find the adequate number of ranked feature. Thus the proposed CSMSVM has improved the computational efficiency to a large extent.

The rest of the paper is organized as follows: Section 2 reviews previous work on SVMs and some relative work. Section 3 introduces the proposed CSMSVM and the Bayesian interpretation of the CSMSVM is presented in Section 4; The learning of feature weights in kernel space is presented in Section 5; Section 6 demonstrates the proposed methodology with two experiments; Then the conclusions are drawn in Section 7.

## 2. Previous work

### 2.1. Support vector machine

Support vector machine (SVM) is a state-of-the-art learning machine which is an effective classification method with significant advantages. Since it is an algorithm that absence of local minima, a representation that depends on few parameters and an adequate generalization to new objects [11,12], SVM has seen its prosperity and has been widely applied to many fields for the past 20 years. It has exerted an indispensable role in pattern recognition [13], disease diagnosis [14], forecasting [15], etc.

For a typical binary classification problem with dataset S in the form of $\{x_i, y_i\}_{i=1}^{m}$, where the training vectors $x_i \in R^n$, a vector of labels $y \in R^m$, $y_i \in \{-1, 1\}$, and $x^{(j)}$ denotes the $j$th feature of vector $x$. Hence $x_i^{(j)}$ is the $j$th feature of the $i$th instance. Linear SVM aims to separate the training patterns by find the optimal hyper plane $f(x) = w^T x + b$ though machine learning technique. This hyperplane can be obtained by solving the following convex optimization problem

$$\min_{w,b} \frac{1}{2}||\mathbf{w}||^2$$

subject to

$$y_i(\mathbf{w}^T \bullet \mathbf{x}_i + \mathbf{b}) \geq 1, \quad i = 1, 2, \ldots, m \tag{1}$$

To solve the optimization problem, we look at the dual formulation of the problem, introducing the Lagrangian multipliers $\alpha_i (i = 1, 2, \ldots, m)$ for the constraint and the Lagrangian is as follows:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(w^T \bullet x_i + b) - 1] \tag{2}$$

Then the primal problem can be expressed as finding the saddle point of Lagrange. Hence, the dual Lagrangian is transformed into:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j$$

$$\text{subject to} : \sum_{i=1}^{m} \alpha_i y_i = 0, 0 \leq \alpha_i \tag{3}$$

Obviously, it is a quadratic optimization problem (QP) with linear constraints and can be solved by many methods and the linear discriminant function $f(x)$ given by SVM can be defined by

$$f(x) = sgn\left(\sum_{i=1}^{m} \alpha_i y_i x^T x + b\right) \tag{4}$$

In many cases, the features in attribute space cannot be linearly separated. However, Kernel representations offer an alternative solution by projecting the data into a high dimensional feature space, namely the feature space, which has increased the computational power of the linear learning machines greatly. Fortunately, the use of linear machines in the dual representation makes it possible to perform this process implicitly, as in this representation the number of tunable parameters does not depend on the number of attributes being used. In the nonlinear SVM, the kernel functions are used to perform a nonlinear mapping to a high dimensional or infinite dimensional feature space without increasing the number of tunable parameters and the computation of the kernel function takes the place of computing the inner product of the feature vectors. For the applications where linear SVM does not produce satisfactory performance, nonlinear SVM is a good choice. The nonlinear SVM is to map the feature matrix by a kernel function. When map the feature vectors into a higher dimensional space, the dual formulation of the maximal margin of SVM's Lagrange multipliers can be found from [16]

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathrm{x}_i, \mathbf{x}_j)$$

$$\text{subject to} : \sum_{i=1}^{m} \alpha_i y_i = 0, \ 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, m \tag{5}$$

For the question whether a function can be used as a kernel function, Mercer's theorem pictures the characteristic of a kernel function $K(x,y)$. It indicates that many symmetric functions that satisfy the Mercer conditions can be a kernel function [17]. Among a variety of existing kernel function, the polynomial and the radial basis function (RBF) are the most popular functions and have a widely applications [18].

- Polynomial function: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \bullet \mathbf{x}_j + 1)^d$
- Radial basis function: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-(\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\rho^2))$

where $d \in N$ is the degree of the polynomial and $\rho > 0$ is the parameter controlling the width of the kernel.

### 2.2. Cosine similarity measure

Cosine similarity measure is a classical criterion for evaluating the distance between two vectors or points. The cosine similarity measure used in this paper is the within-class scatter matrix whitened cosine similarity measure which is among the most used similarity measures in the pattern recognition field. The cosine similarity measure can be described as follows:

$$\mathrm{CSM}(u, v) = \frac{\langle (W^t u)^t, (W^t v) \rangle}{\|W^t u\| \|W^t v\|} \tag{6}$$

where $\mathrm{CSM}(u, v)$ represents the whitened cosine similarity measure between feature vector $u$ and vector $v$. The two pattern vectors are with a dimension of d, namely $u, v \in R^d$ and $W$ is the whitening transformation matrix, which can be expressed by means of the covariance matrix. The covariance matrix of all instances can be described as the within-class scatter matrix $\Sigma_i = S_w = \sum_{i=1}^{L} P(w_i) \varepsilon\{(\chi - M_i)(\chi - M_i)^t \mid w_i\}$. When in the view of PCA, the covariance matrix can be described as $\Sigma_i = \Phi \Lambda \Phi^t$. Where $\varepsilon\{\bullet\}$ is the expectation operator and $M_i$ is the mean vector of class $i$, $\Phi$ is a matrix constructed by the eigenvector thus it is an orthogonal matrix, and $\Lambda$ is a diagonal eigenvalue matrix. Then the whitening transformation can be described by $W = \Phi \Lambda^{-1/2}$, and Eq. (5) can be represented by

$$\mathrm{CSM}(u, v) = \frac{u^t \Sigma^{-1} v}{\|W^t u\| \|W^t v\|} \tag{7}$$

However, owing to its inadequacy in addressing both the distance and the angular measure, the cosine similarity measure fails to indicate the actual distance between two pattern vectors in Euclidean space. Moreover, when the angle between the two vectors is greater than $\pi/2$, ambiguity will arise and lead to misunderstanding. To overcome these problems, the normalized

correlation measure [19] which uses the absolute value of the cosine similarity, have been proposed. These methods, however, could not fully address all the problems and usually lead to new drawbacks. To solve the inadequacy problems of cosine similarity measure, this paper introduces the SVM's maximum margin property to evaluate its actual distance between two pattern vectors.

## 3. The novel methodology

It is universally acknowledged that when the support vector machine finds a hyperplane, which separates the feature with maximum margin in feature space, it fails to address the scatter of the features in feature space. On the one hand, when the features in the training set are closed to each other, the probability that a smaller distance between features in the same class will be higher. Thereby the classifier will have a lower classification error. Taking the fact that the cosine similarity measure cannot address the real distance of two pattern vectors into consideration and to overcome the SVM's ignorance of feature scatter and the cosine similarity measure's intrinsic problems, this paper combines them together and construct a novel criterion for feature selection and classification, which selects the features and classifies the patterns simultaneously. We name the novel feature selection method as cosine similarity measure support vector machine and CSMSVM for short. The CSMSVM can be defined as follows:

$$\max_{\overline{\boldsymbol{w}},\overline{\boldsymbol{b}},\boldsymbol{v},\boldsymbol{M}} \frac{M}{\sqrt{\sum_{i,j}\frac{1+y_iy_j}{2}\left(1-CSM(\phi(\mathbf{x}_i,\boldsymbol{v}),\phi(\mathbf{x}_j,\boldsymbol{v})\right)}}$$

$$\text{subject to}: y_i(\overline{\mathbf{w}}\bullet\phi(x_i,v)+\overline{b})\geq M, \|\overline{\mathbf{w}}\|=1, \forall i \tag{8}$$

where $\boldsymbol{v}$ is a parameter vector that adds weight values to the features and maps from the input space to the feature space, $M$ is the margin of the standard SVM. According to the definition of the CSMSVM, the proposed method takes both the margin of the separation hyperplane and the within-class distance into consideration simultaneously, and the function for optimization is the margin to within-class distance ratio.

Let $w=\overline{w}/M, b=\overline{b}/M$ and $\psi(\boldsymbol{v})=\sum_{i,j}(1+y_iy_j/2)(1-\mathrm{CSM}(\phi(\mathbf{x}_i,\boldsymbol{v}),\phi(\mathbf{x}_j,\boldsymbol{v}))$ then substitutes them in Eq. (7), we have

$$\max_{\mathbf{w},\mathbf{b},M,\boldsymbol{v}} \frac{1}{\|\mathbf{w}\|_2\sqrt{\psi(\boldsymbol{v})}}$$

$$\text{subject to}: y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i,\boldsymbol{v})+b\right)\geq 1 \quad \forall i \tag{9}$$

Transform the above equation to a minimize optimal problem, Eq. (8) is equivalent to

$$\min_{\mathbf{w},\mathbf{b},\boldsymbol{v}} \frac{1}{2}\psi(\boldsymbol{v})\|\mathbf{w}\|_2^2$$

$$\text{subject to}: y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i,\boldsymbol{v})+b\right)\geq 1 \quad \forall i \tag{10}$$

Using the soft-margin instead of hard-margin, we have

$$\min_{\mathbf{w},\mathbf{b},\boldsymbol{v},\xi} \frac{1}{2}\psi(\boldsymbol{v})\|\mathbf{w}\|_2^2+C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to}: y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i,\boldsymbol{v})+b\right)\geq 1-\xi_i, \xi_i\geq 0, \forall i \tag{11}$$

Let $W(x)=\max_{w,b,M,v}\|w\|_2/\sqrt{\psi(v)}$, then the criterion for feature selection or the score for a feature p can be defined as

$$C_{\text{score}}^{(p)}=\frac{[W(x^{-p})-W(x)]}{W(x)} \tag{12}$$

where $x$–$p$ means the training data with feature $p$ removed.

Obviously, the feature selection criterion evaluates the influence of the removed feature $p$ on the margin to cosine similarity measure ratio, and the lower of a $C_{\text{score}}^{(p)}$ for a feature, the more important of the feature will be. After feature rank algorithm has been obtained the feature rank, the feature selection process can be realized by the kick-one-out strategy. Namely, we remove the feature with maximum score, then goes to the next ranking cycle to kick out the next feature until the classification accuracy reaches its maximum value.

## 4. Bayesian interpretation of the novel methodology

To build the proposed method on a theoretical foundation, this section discusses the Bayesian interpretation of the novel method. Through the analysis of the novel criterion, the connection of the novel method with the Bayes decision rule for minimum error will be revealed.

First, we reconsider the cosine similarity measure. As in a d-dimensional feature space, the feature vector $\alpha\in R^d$ and $\alpha$ belongs to one of the predefined $L$ classes which can be defined as $\omega_1,\omega_2,\ldots,\omega_L$, the conditional probability density functions and the prior probabilities can be described as $p(\alpha|\omega_1), p(\alpha|\omega_2), \ldots, p(\alpha|\omega_L)$ and $P(\omega_1), P(\omega_2),\ldots,P(\omega_L)$, respectively. Then the multiclass Bayes decision rule for minimum error could be written as follows [20]:

$$\ln[p(\alpha|w_k)P(w_k)]=\max_i^L \ln[p(\alpha|w_i)P(w_i)]\rightarrow\alpha\in w_k \tag{13}$$

Eq. (13) denotes that the conditional density function of $\alpha$ given $\omega_k$ and if its prior probability is the largest among the L classes, the feature vector $\alpha$ can be classified to $\omega_k$. When it comes to a multivariate normal distribution with mean vector $M_i\in R^d$, and the covariance matrix, $\Sigma_i\in R^{d\times d}$, we will have

$$p(\alpha|w_i)=\frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(\alpha-M_i)^t\Sigma_i^{-1}(\alpha-M_i)\right\} \tag{14}$$

Then the multiclass Bayes decision rule for minimum error described by Eq. (14) becomes

$$\delta_i(\alpha)=-\frac{1}{2}\{(\alpha-M_i)^t\Sigma_i^{-1}(\alpha-M_i)+\ln(2\pi)+\ln(\Sigma_i)\}+ln[P(w_i)] \tag{15}$$

To make Eq. (15) simple, we assume that the prior probabilities are equal with each other, which is usually used, then the covariance matrix of the L classes, $\Sigma_i$, are identical to the covariance matrix of all instances, in other words $\Sigma_i=\Sigma$. Then Eq. (14) can be simplified as follows:

$$\delta_i(\alpha)=-\frac{1}{2}(\alpha-M_i)^t\Sigma_i^{-1}(\alpha-M_i) \tag{16}$$

In fact, Eq. (16) is valid under two other assumptions, namely, the conditional probability density functions, $p(\alpha|\omega_i)$, are multivariate normal and the prior probability, $p(\omega_i)$, are all equal. After we introduce the within-class scatter whitened matrix into Bayes rule, the connection between the whitened cosine similarity measure and the Bayes rule can be revealed by:

$$\delta_i(\alpha)=-\frac{1}{2}\{\|W^t\alpha\|^2+\|W^tM_i\|^2-2\|W^t\alpha\|\|W^tM_i\|CSM(\alpha,M_i)\} \tag{17}$$

For the probabilistic interpretation of SVM classification [21], one can regard the optimization function (5) as defining a negative log-posterior probability for the parameters **w** and **b** for SVM. Then the traditional SVM classifier can be interpreted as the maximum a posterior (MAP) solution of the corresponding probabilistic inference problem. As the novel criterion only relates to the first term of function (1), here we give the prior $Q(w,b)\propto\exp\left(-(1/2)\|w\|^2-(1/2)b^2B^2\right)$. Obviously, this is a

Gaussian prior on **w** with the components of **w** that are uncorrelated with each other. Moreover, the components have unit variance. The Gaussian prior on **b** with variance $B^2$ is usually used as only the 'latent function' values $\theta(x)=\mathbf{w}\phi(x)+\mathbf{b}$ rather than **w** and **b** individually occur in the second term of (1). The $\theta(x)$ also has a joint Gaussian distribution with covariance can be defined as follows:

$$\langle\theta(x)\theta(x')\rangle=\langle(\phi(x)\bullet w)(\phi(x')\bullet w)\rangle+B^2=\phi(x)\phi(x')+B^2 \tag{18}$$

Then the SVM prior can be represented as a Gaussian process (GP) over the function $\theta$ with zeros mean whose covariance function can be described as

$$K(x,x')=\phi(x)\phi(x')+B^2 \tag{19}$$

Only take the first term of (5) into consideration, the MAP solution for a data set S is $\theta^*=\arg\max P(\theta|S)$. Where log-posterior of the model is

$$\ln P(\theta|S)=-\frac{1}{2}\sum_{x,x'}\theta(x)K^{-1}(x,x')\theta(x') \tag{20}$$

Combining the cosine similarity measure and the SVM, one can regard the first tern of (11) as defining a negative log-posterior probability for the parameters θ which can be defined as follows:

$$lnP(\theta|S)=-\frac{1}{2}\left(\sum_{ij}\frac{1+y_iy_j}{2}(1-CSM[\phi(x_i,v),\phi(x_j,v)]\right)\bullet \sum_{x,x'}\theta(x))K^{-1}(x,x')\theta(x') \tag{21}$$

If we assume that the whitened pattern vectors, $W^T\alpha$ and $W^tM_i$ are normalized to unit norm, then Eq. (21) can be represent as

$$lnP(\theta|S)=\left(\sum_{i=1}^{L}\sum_{j=1}^{l_i}\mathrm{CSM}(x_j,M_i)\right)\sum_{x,x'}\theta(x)K^{-1}(x,x')\theta(x') \tag{22}$$

where $L$ is the number of classes and $l_i$ is the number of instances in class $i$.

According to Eq. (22), the Bayesian property of the novel criterion is revealed, since the feature selection process selects the feature subset that can achieve the maximum MAP. In other words, the feature selection result of the novel criterion has achieved a optimal feature subset that improve the performance of SVM as it has optimized the MAP. Taking the angle and margin into consideration, the introduction of cosine similarity measurement into SVM have enhanced the performance of SVM in classification and equips the standard SVM with minimum error character.

## 5. Learning feature weights

The feature ranking method defined by Eq. (12) has the ability to select the feature already. However, if directly rank the feature with Eq. (12), the features in feature space have not been reconstructed. Fortunately, Eq. (12) also provides an access to improve the feature space as the features' weight $\boldsymbol{v}$ in Eq. (12) are set to 1. To achieve a better future space, this section proposes a method to learn the feature weights and reconstruct the features in feature space.

As $\phi(\mathbf{x}_i,\boldsymbol{v})$ assigns different weights to different features, hence $\phi(\mathbf{x}_i,\boldsymbol{v})=\phi(\boldsymbol{diag}(\boldsymbol{v})\mathbf{x}_i)$, where $v=[v_1,v_2,v_3,\ldots,v_l]^T$ are the feature weights. Then we simplify the cosine similarity measure and have

$$\psi(\boldsymbol{v})=\sum_{ij}\frac{1+y_iy_j}{2}\left(1-\frac{\langle\phi(\mathbf{x}_i,\boldsymbol{v}),\phi(\mathbf{x}_j,\boldsymbol{v})\rangle}{\sqrt{\phi(\mathbf{x}_i,\boldsymbol{v})^T\bullet\phi(\mathbf{x}_i,\boldsymbol{v})}\bullet\sqrt{\phi(\mathbf{x}_j,\boldsymbol{v})^T\bullet\phi(\mathbf{x}_j,\boldsymbol{v})}}\right) \tag{23}$$

In the CSMSVM, the RBF kernel is utilized as the kernel function, and define $K(\phi(\mathbf{x}_i,\boldsymbol{v})\bullet\phi(\mathbf{x}_j,\boldsymbol{v}))=\boldsymbol{K}(\mathbf{x}_i,\mathbf{x}_j,\boldsymbol{v})=\exp\left(-\|\mathbf{x}_i^*\boldsymbol{v}-\mathbf{x}_j^*\boldsymbol{v}\|^2\right)$.Then Eq. (23) is equivalent to

$$\psi(\boldsymbol{v})=\sum_{i,j}\frac{1+y_iy_j}{2}(1-\boldsymbol{K}(\mathbf{x}_i,\mathbf{x}_j,\boldsymbol{v})) \tag{24}$$

Therefore, the optimization problem can be described as follows:

$$\min_{\mathbf{w},\mathbf{b},\boldsymbol{v},\boldsymbol{\xi}}\frac{1}{2}\left(\sum_{i,j}\frac{1+y_iy_j}{2}(1-\boldsymbol{K}(\mathbf{x}_i,\mathbf{x}_j,\boldsymbol{v}))\right)\|\mathbf{w}\|_2^2+C\sum_{i=1}^{m}\xi_i$$
$$\text{subject to}: y_i(\mathbf{w}^T\phi(\mathbf{x}_i,\boldsymbol{v})+b)\geq 1-\xi_i,\quad \xi_i\geq 0,\forall i \tag{25}$$

To solve the optimization problem defined by Eq. (25), the Lagrangian function was introduced and was defined as

$$\boldsymbol{L}(\mathbf{w},\boldsymbol{b},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{v})=\frac{1}{2}\psi(\boldsymbol{v})\langle\mathbf{w}\bullet\mathbf{w}\rangle+\boldsymbol{C}\sum_{i=1}^{m}\boldsymbol{\xi_i}-\sum_{i=1}^{m}\boldsymbol{\alpha_i}[\boldsymbol{y_i}(\langle\mathbf{w}\bullet\phi(\mathbf{x_i},\boldsymbol{v})\rangle+\boldsymbol{b})-1+\boldsymbol{\xi_i}] \tag{26}$$

where $\boldsymbol{\alpha_i}$ are the Lagrange multipliers.

The corresponding dual is found by differentiating with the variances as follows

$$\frac{\partial\boldsymbol{L}(\mathbf{w},\boldsymbol{b},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{v})}{\partial\mathbf{w}}=\boldsymbol{\Psi}(\boldsymbol{v})\mathbf{w}-\sum_{i=1}^{m}\boldsymbol{\alpha_i}\boldsymbol{y_i}\mathbf{x_i}=0$$
$$\frac{\partial\boldsymbol{L}(\mathbf{w},\boldsymbol{b},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{v})}{\partial\mathbf{b}}=\sum_{i=1}^{m}\boldsymbol{\alpha_i}\boldsymbol{y_i}=0$$
$$\frac{\partial\boldsymbol{L}(\mathbf{w},\boldsymbol{b},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{v})}{\partial\boldsymbol{v}}=\frac{1}{2}\frac{d\boldsymbol{\Psi}(\boldsymbol{v})}{d\boldsymbol{v}}\langle\mathbf{w},\mathbf{w}\rangle-\sum_{i=1}^{m}\boldsymbol{\alpha_i}\boldsymbol{y_i}\left\langle\mathbf{w},\frac{d\phi(\mathbf{x_i},\boldsymbol{v})}{d\boldsymbol{v}}\right\rangle=0 \tag{27}$$

Since the above formulation is difficult to solve, while it shows some relationships between the variances. To address this issue, an iterative algorithm has been developed as an approximation for this optimization problem. According to [22], the two-step methodology can be utilized to solve this problem.

- First the tradition dual formulation of SVM for a fixed weight value or fixed kernel width $\boldsymbol{v}$ is solved, and the corresponding Lagrangian for the 1-norm soft margin optimization problem is

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2\Psi(\boldsymbol{v})}\sum_{i,s=1}^{m}\alpha_i\alpha_sy_iy_sK(\mathbf{x}_i,\mathbf{x}_s,\boldsymbol{v})$$
$$\text{subject to}:\sum_{i=1}^{m}\alpha_iy_i=0;\quad 0\leq\alpha_i\leq\frac{C}{\Psi(\boldsymbol{v})},\quad i=1,2,\ldots,m \tag{28}$$

- In the second step the algorithm solves the above non-linear formulation for a given solution $\alpha$. To obtain the optimal feature weights, we introduce a penalization function $f(\boldsymbol{v})$ based on the 0-norm in [23], and the penalization function is as follows

$$f(v)=\sum_{j=1}^{l}[1-\exp(-\beta v_j)] \tag{29}$$

And according to Eq. (26), the non-linear optimization problem can be described as follows

$$\min_{\boldsymbol{v}}\ H(\boldsymbol{v})=\frac{1}{2\Psi(\boldsymbol{v})}\sum_{i,s=1}^{m}\alpha_i\alpha_sy_iy_sK(\mathbf{x}_i,\mathbf{x}_s,\boldsymbol{v})+C_2f(\boldsymbol{v})$$
$$\text{subject to}: v_j\geq 0,\quad j=1,2,\ldots,l \tag{30}$$

To appropriate the optimal value of $\boldsymbol{v}$, a numerical iterative algorithm using the gradient of the objective function that updates the weight value $\boldsymbol{v}$ is used. The mechanism of the feature section

process is eliminating the features that their weight values are below a given threshold. The algorithm that learns the weight value and feature selection process can be described as follows.

Algorithm 1
Weight value learning and feature selection process

**Initialization:**
$\boldsymbol{v}=\boldsymbol{v}_0$; $\epsilon=\epsilon_0$; $\eta=\eta_0$; $\gamma=\gamma_0$; $\rho=\rho_0$;flag=true; $loop$=0;
Start:
**While** ($\|\mathbf{w}^{cycle+1}-\mathbf{w}^{cycle}\|\le\rho$)
  **While** (flag==true)
    Train SVM (step 1) for given $\boldsymbol{v}$;
    calculate $\mathbf{w}^{cycle+1}$
    $v^{loop+1}=v^{loop}-\gamma\triangle H(v^{loop})$
    **for** ($v_j^{loop+1}\le\epsilon$)
      $v_j^{loop+1}=0$;
    **end for;**
    **if** ($|v^{loop}-v^{loop+1}|\leqslant\eta$) **then**
    flag=false;
    **end if**
    loop=loop+1;
  **end while;**
cycle=cycle+1;
**end while**
**end**

After the optimal $\boldsymbol{v}$ has been obtained, the iteration algorithm goes back to step one to get the updated $\mathbf{w}$. The cycle of updating $\boldsymbol{v}$ and $\mathbf{w}$ will be continued until the change of $\boldsymbol{v}$ and $\mathbf{w}$ are below a pre-set threshold.

Algorithm 1 indicates that the main point of the iterative process is computing the gradient of the objective function described by Eq. (30) for a given solution from the standard SVM $\alpha$. After introducing the features' weight value, the feature selection process can be conducted by a more efficient way. Algorithm 1 shows that when the weight value is below a given threshold, the weight value will be set to 0, then the corresponding features will be eliminated from the feature subset. In other words, the feature selection process is achieved with the optimization of weight value. Look in detail and for a given feature $j$, the gradient of function $H(v)$ can be given by

$$\Delta_j H(v)=\frac{\left(\sum_{i,s=1}^{m} v_j\left(x_i^{(j)}-x_s^{(j)}\right)^2\alpha_i\alpha_s y_i y_s K(x_i,x_s,v)\right)\left(\sum_{i,s=1}^{m}\frac{1+y_i y_s}{2}(1-K(x_i,x_s,v))\right)+\left(\sum_{i,s=1}^{m} v_j\left(x_i^{(j)}-x_s^{(j)}\right)^2(1+y_i y_s)K(x_i,x_s,v)\right)\left(\sum_{i,s=1}^{m}\alpha_i\alpha_s y_i y_s K(x_i,x_s,v)\right)}{\left[\sum_{i,s=1}^{m}\frac{(1+y_i y_s)}{2}(1-K(x_i,x_s,v))\right]^2}$$

$$+C_2\beta\exp(-\beta v_j) \tag{31}$$

As discussed in [22], the main objective of Eq. (30 )is to find the sparse solutions which makes zero as many components of $v$ as possible. For this reason, this paper takes the 0-norm penalization into consideration instead of 1-norm penalty (LASSO penalty) or 2-norm, even though they might bring us a good feature selection and classification result.

## 6. Applications

### 6.1. Rolling element bearings fault diagnosis

To validate the efficiency of the proposed CSMSVM, an experiment based on the rolling element bearings was conducted. For rolling-element bearing, many features are useful for fault diagnosis, including the time-domain features and frequency-domain features. Intuitively, however, we know many features are sensitive to the classification algorithm. In other words, some features are efficient for classification only for some certain algorithms and inefficient for some other algorithms. To improve the algorithm's generalization ability and decrease the computation, feature selection process is a perfect choice to handle this situation. In this paper, 40 features are extracted as candidate features for selection, namely the wavelet packet energy (8 features), singular value spectrum (former 15 order singular value), time domain statistics (12 features, they are average, variance, peak, average amplitude, RMS, skewness, kurtosis, 2-order central moment etc.) and intrinsic mode functions (IMFs) energy (former 5 order IMFs). The wavelet packet energy comes from the wavelet package transform and each instance was decomposed into eight bins of wavelet package coefficients via WPT at level 3(in the wavelet package transform, the db4 wavelet is used); the singular value spectrum comes from the singular value decomposition of a Hankel matrix composed by the time domain signal and the IMFs come from the empirical mode decomposition (EMD). Before these features are fed as input to the machine learning algorithm, three kinds of features are normalized.

In order to evaluate the classification performance of the CSMSVM, we compared the results for a given number of features (determined by the threshold ε set in Algorithm 1) with different features selection algorithms for SVMs. The features for standard SVMs come from the Fisher score based feature selection result and the genetic algorithm based feature selection method. Fisher based method, which is a filter method, was chosen because the fisher score is relative to the within classes distance to between classes distance ratio, and the proposed criterion also addresses the margin to within classes distance. The genetic algorithm was chosen because it is hybrid method that is close to the wrapper method.

In the experiments, the time series data collected from the rolling element bearing test rig was split into 200 parts and 2048 samples for each part. Then we got 800 pieces of data as we have four kinds of running state for the rolling element bearing. To construct the dataset, 410 instances are chosen for each running state and the number of variables, number of examples, and proportion of examples in the predominant class are 40,410 and 0.64, respectively. The detail of the instances is shown in Table 1.

### 6.2. Mild cognitive impairment diagnosis

Effective and accurate diagnosis of mild cognitive impairment (MCI), which is the early stage of Alzheimer's disease (AD), has played a pivotal role in dealing with AD. The electroencephalogram (EEG) signal has been proved to be a good biomarker for the diagnosis of

**Table 1**
Number of variables, number of examples, and proportion of examples in the predominant class for all six data sets.

| | Variables | Examples | Predominant class proportion |
|---|---|---|---|
| Inner fault | 40 | 410 | 0.64 |
| Outer fault | 40 | 410 | 0.64 |
| Rolling element | 40 | 410 | 0.64 |
| Normal bearing | 40 | 410 | 0.64 |
| MCI | 128 | 560 | 0.643 |
| Normal cognitive | 128 | 560 | 0.786 |

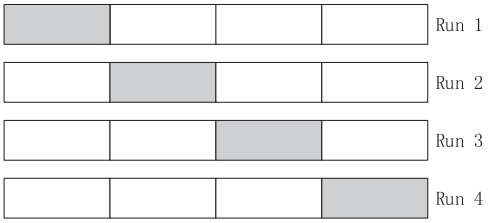



**Fig. 1.** The technique of S-fold cross-validation, illustrated here for the case of $S=4$, involves taking the available data and partitioning it into 4 groups. The 3 of the groups are used to train the models that are then evaluated on the remaining group. The procedure is then repeated for all S possible choices for the held-out group, and the performance scores from the S runs are the averaged.

**Table 2**
Number of selected features n, effectiveness (error rate), change of within classes distance to between classes distance ratio (only for CSMSVM) using three different feature selection methods.

| | $n$ | Fisher+SVM | CSMSVM | GA+SVM | Change of distance ratio |
|---|---|---|---|---|---|
| MCI | 6 | 0.1429 | 0.0286 | 0.10 | 0.91 |
| Normal cognitive | 5 | 0.1286 | 0.0143 | 0.0857 | 0.90 |
| Normal | 6 | 0.0625 | 0.0063 | 0.0625 | 0.91 |
| Inner fault | 5 | 0.1125 | 0.0375 | 0.0437 | 0.93 |
| Outer fault | 7 | 0.0438 | 0.0125 | 0.0563 | 0.94 |
| Race fault | 9 | 0.0125 | 0 | 0.0063 | 0.97 |

MCI. The EEG signal used in this paper came from the Shanghai Sixth People's Hospital. In the test, the data were digitized at a sampling rate of 500 Hz, 50 MCI patients, 30 Normal cognitive (NC) people and 60 AD patients 's data were used in this experiment. 16 channels of EEG activity were recorded continuously from 21 sites by using electrodes set in an elastic cap and positioned according to the 10–20 international system (Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, CZ, C4, T4, T5, P3, Pz, P4, T6, O1, O2, A1 and A2).

As the frequency band is an important feature for the diagnosis of MCI, this paper decomposed the EEG signals with wavelet package transform. In this paper, the Daubechies wavelet with order 5 was chosen in the wavelet package transform and the signals were decomposed by 3 levels, thus 8 frequency bands were obtained.

In this case, the EEG signal of each person was divided into 4 parts, and wavelet package transform was applied to each part. The features were the normalized energy of the wavelet coefficient at level 3. Thus we have 4 features for every piece of signal and 128 features for 16 channels. For every person, we have extracted 4 feature vectors and every feature vector we have 64 features. Therefore, in the feature subset, there are 560 feature vector, which is a $560\times 64$ data matrix. The number of variables, number of examples, and proportion of examples in the predominant class for the diagnosis of NC and MCI are shown in Table 1. In this experiment, the feature dimensions are above 100, which can indicate the proposed method's ability in feature selection when the feature dimensions are high. The number of samples, however, is restricted by the number of patients in our projects and is still relatively small. To address this problem and the relatively noisy estimate of predictive performance, the cross-validation, which is illustrated in Fig. 1, is used in this experiment. The technique of $S$-fold cross-validation, illustrated here for the case of $S=4$, allows a proportion $(S-1)/S$ of the available data to be used for training while making use of all the data to assess performance. The 3 of the groups are used to train the models that are then evaluated on the remaining group. The procedure is then repeated for all $S$ possible choices for the held-out group, and the performance scores from the $S$ runs are the averaged.

### 6.3. Results and discussions

For the diagnosis of rolling element bearing fault, MCI and Normal cognitive (NC), the extracted features were fed to the CSMSVM, and the result is shown in Table 2. To make the result comparable, the number of features selected by Fisher score and genetic algorithm are the same with the number of the CSMSVM. Table 2 shows that the proposed method outperforms all other approaches in terms of classification error for a given number of features, especially for the diagnosis of MCI and rolling element's normal state. For the CSMSVM, as the weight values are vary with the features, after the weight values have been added to the features, the feature space has been changed. To investigate the change of the feature space, the within classes distance to between classes distance ratio is calculated. Table 2 also shows the change of the ratio between the weight values has been added to feature and without adding the weight values to features, and the change shows that the CSMSVM has gotten a smaller within classes distance to between classes distance ratio. Even though this change is very small, but its influence on diagnosis result is significance. Table 2 also shows that a larger variables will lead to more significant performance for the proposed method.

The proposed CSMSVM method can achieve an optimal feature subset and maximize the MAP to reduce the classification error. However, through the application to the rolling element's fault diagnosis and the diagnosis of MCI, we found that some parameters should be carefully chosen. As shown in Algorithm 1, there are many parameters should be predefined, such as $\boldsymbol{v}_0, \epsilon_0, \boldsymbol{\gamma}_0, \ etc.$ To find the optimal value of these parameters, some experiments should be conducted before applying the CSMSVM to the real fields. Fig. 2 shows the results of some experiments. Fig. 2(a) shows the searching of optimal parameter $\epsilon_0$ for the diagnosis of rolling element's fault and Fig. 2(b) shows the searching of optimal parameter $\epsilon_0$ for the diagnosis of outer race fault. Fig. 2 (a) and (b) also show the change of the number of selected features against the varying of the parameter $\epsilon_0$. In Fig. 2(a) and (b), a larger $\epsilon_0$ will lead to fewer features in the optimal feature subset, while a larger $\epsilon_0$ do not mean a higher performance of the CSMSVM. Only a suitable $\epsilon_0$ is chosen, can the higher performance of the CSMSVM be achieved. In our application, the suitable $\epsilon_0$ is chosen according to the experiment which varies the value of $\epsilon_0$ and then chooses the $\epsilon_0$ that corresponding to the best performance. For the parameter $\boldsymbol{v}_0$, to make it simple, all the experiments' $\boldsymbol{v}_0$ are set to 1. Fig. 2(c) and (d) shows the optimization of parameter $\beta$ and $\gamma$ for the diagnosis of rolling element's fault. Fig. 2 (c) indicates that $\beta$ is not very sensitive to the performance of the CSMSVM, thus the a wide range of $\beta$ can be chosen. Fig. 2(d) shows that even though a smaller value of $\gamma$ can reach a better performance, but smaller $\gamma$ also means the larger number of iteration. However, in the other side, a larger $\gamma$ will lead to the oscillation of the performance. In the searching of optimal parameter, the number of features in the feature subset is the same and by varying the parameter and observe the classification performance, some conclusions can be drawn as follows: $C_2$ has a great influence to the iteration algorithm, a smaller $C_2$ will increase the iteration cycle or increase the computational time to obtain an optimal weight value. However, a larger $C_2$ will lead to the oscillation iteration process and cannot approach the optimal value in a pre-set limit cycle; $\gamma$ is also related to the necessary iterations for optimization process: a larger $\gamma$ leads to a smaller number of iterations while if too large, oscillation phenomenon will arise; $\beta$ have a weak influence to the classification and have less influence to the optimization process, while $\beta$ and $C_2$ should match each other if an optimal optimization process is needed.

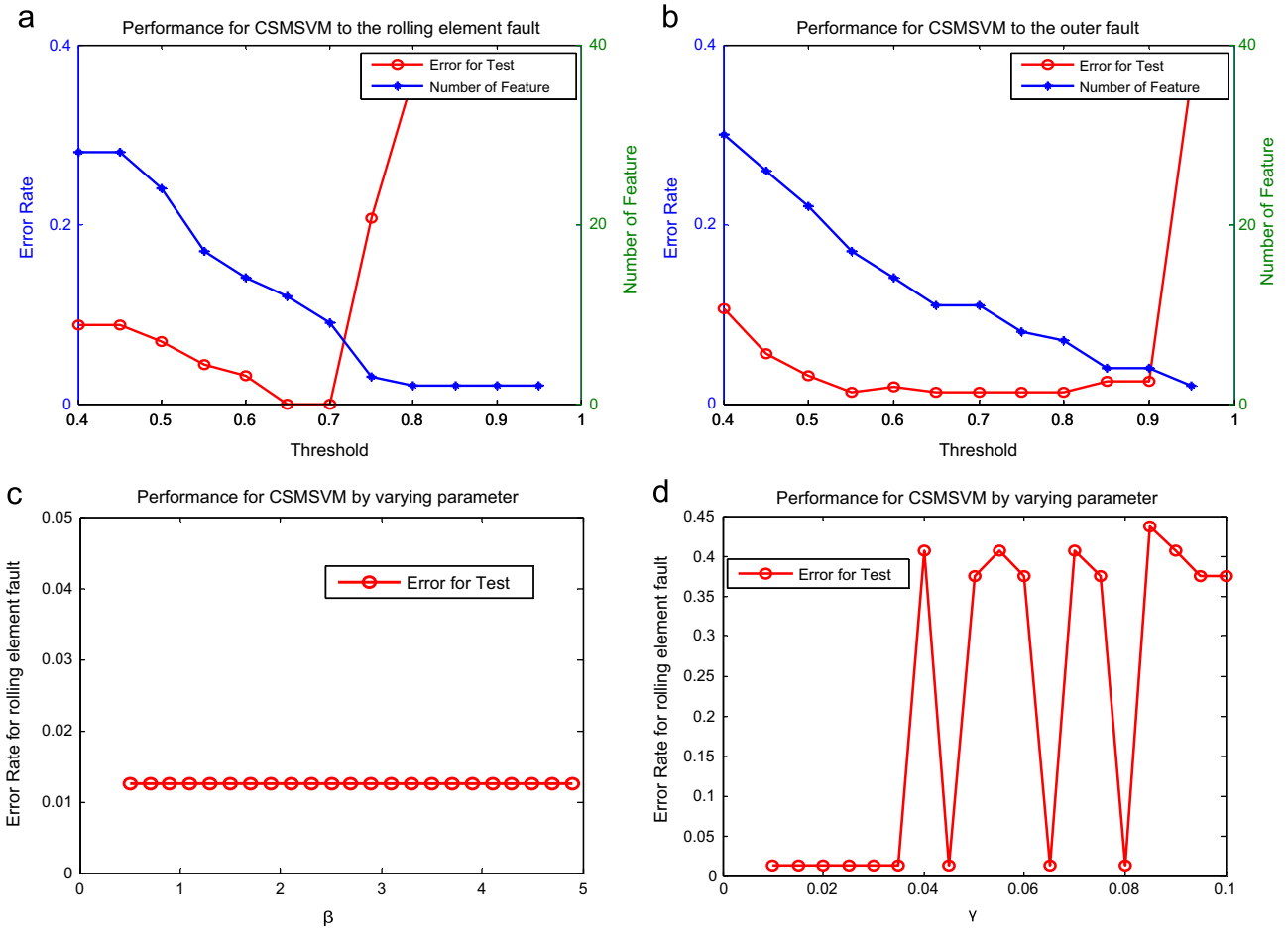



**Fig. 2.** Parameter optimization process (a) the optimization of parameter $\varepsilon$ for the diagnosis of rolling element's fault; (b) the optimization of parameter $\varepsilon$ for the diagnosis of outer race's fault; (c) the optimization of parameter $\beta$ for the diagnosis of rolling element's fault; (d) the optimization of parameter $\gamma$ for the diagnosis of rolling element's fault.

## 7. Conclusions

In this paper, a cosine similarity measure support vector machine has been present. The proposed feature selection method is a kind of wrapper method as it is an integration of feature selection and pattern classification. The CSMSVM has utilized a novel feature selection criteria, namely the margin verses cosine distance ratio, which adds a weight value to the features to maximize the margin verses cosine distance ratio. Compared to other feature selection proposals, the CSMSVM has decreased classification error by increasing the degree of polymerization of data. Additionally, the cosine distance has the advantage over the Euler distance or other distance on increasing the probability of correct classification, which can be obtained from the Bayesian interpretation of the novel methodology.

The optimization process of the proposed feature selection method has shown that the iterative approximation algorithm can achieve the optimal weight value. The application of the CSMSVM to rolling element bearing's fault diagnosis and MCI diagnosis show that the CSMSVM has great capacity in feature selection and pattern recognition. As the proposed methodology is a general method, and it has been built on a solid theory foundation. Hence, it can be used in an active area in pattern recognition, machine learning, data mining and statistic. However, as the novel method is based on the SVM, it application fields could not overpass the application fields of standard SVM, when the data set is too larger, or the features is too larger, the efficiency of the proposed method is open to doubt. Although much progress has been carried out by applying the CSMSVM in industrial field, there are still many aspects of the CSMSVM can be enhanced and many characteristics of the CSMSVM need excavating. In the application of the proposed methodology, we find that the iteration process is very sensitive to the parameters which is bad for the application of the novel method to industry fields. Therefore, for future work, we will improve the computational efficiency, especially enhance the optimization process and extend the application fields of the cosine similarity measure support vector machine. Other directions, such as combining the proposed method with multi-class SVM, replacing the RBF kernel with other kernel functions, investigating the undesirable effects caused by unbalance data sets, etc.

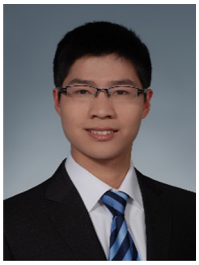
## Acknowledgement

The authors are grateful to the support by the National Natural Science Foundation under Grants nos. 51105243 and 51035007.

## References

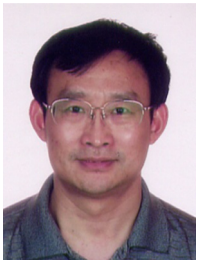
[1] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[2] Y. Liu, Y.F. Zheng, FS_SFS: a novel feature selection method for support vector machines, Pattern Recognit. 39 (2006) 1333–1345.
[3] S. Maldonado, R. Weber, A wrapper method for feature selection using support vector machines, Inf. Sci. 179 (2009) 2208–2217.
[4] M. Pal, G.M. Foody, Feature selection for classification of hyperspectral data by SVM, IEEE Trans. Geosci. Remote Sens. 48 (2010) 2297–2307.
[5] I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recognit. 43 (2010) 5–13.

[6] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (2010) 392–398.
[7] M. Yousef, W. Khalifa, A zero-norm feature selection method for improving the performance of the one-class machine learning for microRNA target detection," in: 2010 5th International Symposium on Health Informatics and Bioinformatics (HIBIT), 2010, pp. 45–50.
[8] H. Liu, E.R. Dougherty, J.G. Dy, K. Torkkola, E. Tuv, H. Peng, et al., Evolving feature selection, IEEE Intell. Syst. 20 (2005) 64–76.
[9] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[10] P.A. Estévez, M. Tesmer, C.A. Perez, J.M. Zurada, Normalized mutual information feature selection, IEEE Trans. Neural Networks 20 (2009) 189–201.
[11] V.N. Vapnik, Statistical Learning Theory (1998).
[12] N. Christianini, S.J. Taylor, An Introduction to Support Vector Machines (And othre Kernel-based Learning Methods) (2000).
[13] A. Mueller, G. Candrian, V.A. Grane, J.D. Kropotov, V.A. Ponomarev, G.-M. Baschera, Discriminating between ADHD adults and controls using independent ERP components and a support vector machine: a validation study, Nonlinear Biomed. Phys. 5 (2011) 5.
[14] B. Magnin, L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, O. Colliot, M. Sarazin, et al., Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI, Neuroradiology 51 (2009) 73–83.
[15] O. Kisi, M. Cimen, A wavelet-support vector machine conjunction model for monthly streamflow forecasting, J. Hydrol. 399 (2011) 132–140.
[16] P. Coloma, J. Guajardo, J. Miranda, R. Weber, Modelos analíticos para el manejo del riesgo de crédito, Trend Manage. 8 (2006) 44–51.
[17] R. Courant, D. Hilbert, Methods of Mathematical Physics vol. 1: CUP Archive, 1966.
[18] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.
[19] V. Struc, N. Pavesic, The corrected normalized correlation coefficient: a novel way of matching score calculation for lda-based face verification, in: FSKD'08. Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 110–115.
[20] K. Fukunaga, Introduction to Statistical Pattern Recognition: Access Online via Elsevier, 1990.
[21] C. Gold, A. Holub, P. Sollich, Bayesian approach to feature selection and parameter tuning for support vector machine classifiers, Neural Networks 18 (2005) 693–701.
[22] S. Maldonado, R. Weber, J. Basak, Simultaneous feature selection and classification using kernel-penalized support vector machines, Inf. Sci. 181 (2011) 115–128.
[23] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.



Gang Chen received the B.S. in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China in 2012, where he is currently working toward the M.S. degree in the Mechanical Engineering Department.

G. Chen is currently with the State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University. His research interests include signal processing, machine fault diagnosis, prognostics and machine learning.



Jin Chen received the B.S. and M.S degrees in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1982 and 1984, and received his Ph. D. degrees from Tokyo Institute of Technology in 2003. He was a senior visiting scholar of the University of Iowa, USA from 1995 to 1996.

He is currently a Professor with the State Key Laboratory of Mechanical System and Vibration, a visiting professor of the Tokyo Institute of Technology and the director of the library of Shanghai Jiao Tong University. His research interests include Mechanical Systems and Signal Processing, mechanical fault diagnosis system.