Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2015, Article ID 251945, 9 pages
http://dx.doi.org/10.1155/2015/251945

Research Article

Multiclass Posterior Probability Twin SVM for Motor Imagery EEG Classification

Qingshan She, Yuliang Ma, Ming Meng, and Zhizeng Luo

Institute of Intelligent Control and Robotics, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China

Correspondence should be addressed to Qingshan She; qsshe@hdu.edu.cn

Received 27 May 2015; Accepted 10 September 2015

Academic Editor: Ladislav Hluchy

Motor imagery electroencephalography is widely used in the brain-computer interface systems. Due to inherent characteristics of electroencephalography signals, accurate and real-time multiclass classification is always challenging. In order to solve this problem, a multiclass posterior probability solution for twin SVM is proposed by the ranking continuous output and pairwise coupling in this paper. First, two-class posterior probability model is constructed to approximate the posterior probability by the ranking continuous output techniques and Platt's estimating method. Secondly, a solution of multiclass probabilistic outputs for twin SVM is provided by combining every pair of class probabilities according to the method of pairwise coupling. Finally, the proposed method is compared with multiclass SVM and twin SVM via voting, and multiclass posterior probability SVM using different coupling approaches. The efficacy on the classification accuracy and time complexity of the proposed method has been demonstrated by both the UCI benchmark datasets and real world EEG data from BCI Competition IV Dataset 2a, respectively.

1. Introduction

Among different brain imaging techniques, electroencephalography (EEG) is commonly used due to its noninvasive acquisition, high temporal resolution, ease of use, and low cost [1]. The motor imagery based noninvasive BCI systems have provided users the ability to control movements of a computer cursor and interactive robotic wheelchairs and explore virtual environments [2]. The EEG measurement is a high dimensional scalp measurement, and thus it reflects the global cerebral electrophysiological activity. At the same time, it results in two weaknesses [3]. On the one hand, EEG is inherently multivariate and exhibits a high correlation [4] between the measured potentials at different electrodes (channels). On the other hand, the observed data are of low amplitude and sensitive to noise of biological, environmental, and instrumental origin. Due to these characteristics of EEG, adequate signal processing techniques are required to handle the problems of noise cancelation, feature extraction, and classification of multichannel EEG.

One core part of motor imagery based BCI systems is a pattern-classification process, where effective classification methods are crucial to achieving high performance. Conventional learning methods have been applied to binary classification of EEG, such as Fisher linear discriminant, k-nearest neighbor, artificial neural network, support vector machine (SVM) [5], Bayesian classifier [6], and hidden Markov models [7]. Owing to good generalization performance, absence of local minima, and sparse representation of solution, SVMs are combined with different feature extraction methods to improve classification performance of EEG patterns, but they do not provide posterior probability. They are actually useful for constructing a classifier producing a posterior probability instead of hard class labels, due to the fact they can be suitable for postprocessing and correcting the error introduced by wrongly labeled/noisy instances [8, 9]. Platt proposed probability output methods for SVM and validated its performance on three data-mining-style datasets [8]. Recently, Shao proposed the two-class probability estimates for twin SVM (TSVM) by making use of its superiority over SVM [10]. Different from SVMs in one fundamental way, TSVM aims to generate two nonparallel planes by solving two small size quadratic programming problems such that it is approximately four times faster than SVMs [11].

Consider the task of learning to recognize several motor imagery tasks performed by a given subject. This is a multiclass recognition problem. As for the posterior probability SVM based multiclassification, it is usually solved by using their multiclass extensions via the single-machine approach [9] and the multimachine approach [12]. The former involves directly solving a single optimization problem, while the latter involves decomposing a multiclass problem to a set of two-class problems via one-against-rest (OAR) and one-against-one (OAO). For the OAR approach, the number of data samples for a given task is a small fraction of the whole data, and thus the sizes of the two classes in each binary classification problem are unbalanced [9, 13]. By use of the OAO approach, the probability estimates for multiclass SVM have been proposed by combining all pairwise comparisons [12] and demonstrated that it is more stable than existing methods, that is, the voting scheme and the method in [14].

To the best of our knowledge, no previous work has been done for twin SVM with multiclass posterior probability outputs, or more specially multiclass posterior TSVM modeling by the ranking continuous output and pairwise coupling. The main advantage of our work is that multiclass posterior TSVM model can provide the posterior probability with high classification accuracy and low time complexity as compared with conventional SVM methods, and it is suitable for BCI applications. In order to handle multiclass classification of motor imagery EEG, a multiclass posterior probability TSVM (PPTSVM) solution is proposed by combining two-class PPTSVM with pairwise coupling in this paper. First, two-class PPTSVM model is constructed to approximate the posterior probability by the ranking continuous output techniques [10] and Platt's estimating method [8]. Secondly, a solution of multiclass probabilistic outputs for TSVM is provided by combining every pair of class probabilities according to the method of pairwise coupling [12]. Finally, the proposed method is compared with multiclass SVM via voting, multiclass TSVM by voting, and multiclass posterior probability SVM via the coupling approach. The classification accuracy and time complexity of the proposed method have been manifested by both the UCI benchmark datasets and real world EEG data from BCI Competition IV Dataset 2a.

This paper is organized as follows: Section 2 states the multiclass posterior probability TSVM model including a brief introduction to standard TSVM, two-class PPTSVM formulation and its multiclass extension via pairwise coupling. Section 3 presents some experimental results on both benchmark datasets and motor imagery EEG datasets to demonstrate the validity of the proposed method. Finally, concluding remarks are given in Section 4.

2. Foundations for the Proposed Model Description

2.1. Formulation of Twin Support Vector Machine. In a two-class problem, the patterns to be classified are denoted by a set of m row vectors $A_i \in R^n$, where $A_i = (A_{i1}, A_{i2}, \ldots, A_{in})$, and let $\mathbf{y} \in \{-1, 1\}$ represent the class to which the pattern belongs. As compared with the classical soft margin SVM classifier finding one hyperplane $w^T x + b = 0$ that separates patterns from the two classes, TSVM seeks two nonparallel hyperplanes [11]. Let the number of patterns in classes +1 and −1 be, respectively, given by m_1 and m_2, where $m_1 + m_2 = m$. For simplicity of expression, the patterns of classes +1 and −1 are denoted by matrices $A \in R^{m_1 \times n}$ and $B \in R^{m_2 \times n}$, respectively. In nonlinear discrimination, the TSVM with nonlinear kernel finds the following kernel-generated surfaces:

$$\begin{aligned} f_1(x) &= K\left(x^T, C^T\right) u_1 + b_1 = 0, \\ f_2(x) &= K\left(x^T, C^T\right) u_2 + b_2 = 0 \end{aligned} \tag{1}$$

Here, $C^T = [A \;\; B]^T$, and K is an appropriately chosen kernel, where $K(x^T, C^T) = \phi(x^T) \cdot \phi(C^T)$ and $\phi(\cdot)$ is a nonlinear mapping function.

The parameters of the two hypersurfaces are obtained by solving the following pair of quadratic programming problems:

$$\begin{aligned} &\underset{u_1, b_1, q}{\text{Min}} \quad \frac{1}{2} \left\| K\left(A, C^T\right) u_1 + e_1 b_1 \right\|^2 + c_1 e_2^T q \\ &\text{s.t.} \quad -\left(K\left(B, C^T\right) u_1 + e_2 b_1\right) + q \geq e_2, \quad q \geq 0, \end{aligned} \tag{2}$$

$$\begin{aligned} &\underset{u_2, b_2, q}{\text{Min}} \quad \frac{1}{2} \left\| K\left(B, C^T\right) u_2 + e_2 b_2 \right\|^2 + c_2 e_1^T q \\ &\text{s.t.} \quad \left(K\left(A, C^T\right) u_2 + e_1 b_2\right) + q \geq e_1, \quad q \geq 0, \end{aligned} \tag{3}$$

where the positive constants c_1 and c_2 are penalty parameters and e_1 and e_2 are the vectors of parameters of appropriate dimensions and $\|\cdot\|$ denotes the L_2 norm. The objective functions have clear geometric meaning; for example, function (2) shows that the first term tends to keep the hypersurface close to points of class +1, while the constraints require the hypersurface to be at a distance of at least 1 from points of class −1.

By introducing Lagrange multipliers and exploiting Karush-Kuhn-Tucker (KKT) conditions, the Wolfe dual formulation of the above optimization problem can be obtained as follows [11]:

$$\begin{aligned} &\underset{\alpha}{\text{Max}} \quad e_2^T \alpha - \frac{1}{2} \alpha^T G \left(H^T H\right)^{-1} G^T \alpha \\ &\text{s.t.} \quad 0 \leq \alpha \leq c_1, \\ &\underset{\beta}{\text{Max}} \quad e_1^T \beta - \frac{1}{2} \beta^T H \left(G^T G\right)^{-1} H^T \beta \\ &\text{s.t.} \quad 0 \leq \beta \leq c_2, \end{aligned} \tag{4}$$

where $G = [K(B, C^T) \;\; e_2]$, $H = [K(A, C^T) \;\; e_1]$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{m_2})^T$, and $\beta = (\beta_1, \beta_2, \ldots, \beta_{m_1})^T$ are the vectors of Lagrangian multipliers. Once the above dual formations are solved to obtain the surfaces (1), a new sample $x \in R^n$ is assigned to class i ($i = +1, -1$), depending on which of the two surfaces it lies closest to, given by

$$\text{class } i = \arg \min_{k=1,2} \frac{\left| K\left(x^T, C^T\right) u_k + b_k \right|}{\sqrt{u_k^T K\left(C, C^T\right) u_k}}, \tag{5}$$

where $|\cdot|$ is the perpendicular distance of the point x from the surface $K(x^T, C^T)u_k + b_k = 0, k = 1, 2$.

2.2. Multiclass Posterior Probability TSVM. The multiclass extension to TSVM can also be modified to include posterior probability estimates instead of hard labels. In this section, multiclass posterior probability TSVM model is established. Firstly, similar to [10], a continuous output value is defined by calculating the distances from a sample to two decision hypersurfaces of TSVM with nonlinear kernels. Secondly, two-class posterior probability TSVM model is constructed to approximate the posterior probability using Platt's estimating method [8]. Here it is noted that the current SVM probability models cannot be used directly due to the different mechanism of TSVM and SVM. Finally, a solution of multiclass probabilistic outputs for TSVM is provided by combining every pair of class probabilities according to the pairwise coupling method [12].

2.2.1. Two-Class PPTSVM Model. In TSVM binary classification, once the surfaces (1) are obtained, the label of any sample can be predicted. However, many applications require a posterior class probability instead of predicting the hard label. Since traditional TSVMs do not provide such probabilities, a modeling method of two-class PPTSVM is constructed by the following [8, 10].

To reflect the degree of new sample belonging to a certain class, let a function $f(x)$ be the continuous output of a two-class TSVM [10]

$$f(x) = \begin{cases} d_{\min}(x)\left(\dfrac{d_{\min}(x)}{d_{\max}(x)}\right)^{\gamma}, & \dfrac{\left|K\left(x^T, C^T\right)u_1 + b_1\right|}{\left|K\left(x^T, C^T\right)u_2 + b_2\right|} \cdot \dfrac{\sqrt{u_2^T K\left(C, C^T\right)u_2}}{\sqrt{u_1^T K\left(C, C^T\right)u_1}} < 1, \\ 0, & \dfrac{\left|K\left(x^T, C^T\right)u_1 + b_1\right|}{\left|K\left(x^T, C^T\right)u_2 + b_2\right|} \cdot \dfrac{\sqrt{u_2^T K\left(C, C^T\right)u_2}}{\sqrt{u_1^T K\left(C, C^T\right)u_1}} = 1, \\ -d_{\min}(x)\left(\dfrac{d_{\min}(x)}{d_{\max}(x)}\right)^{\gamma}, & \dfrac{\left|K\left(x^T, C^T\right)u_1 + b_1\right|}{\left|K\left(x^T, C^T\right)u_2 + b_2\right|} \cdot \dfrac{\sqrt{u_2^T K\left(C, C^T\right)u_2}}{\sqrt{u_1^T K\left(C, C^T\right)u_1}} > 1, \end{cases} \tag{6}$$

where $\gamma > 0$ is a weight parameter and two relevant variables $d_{\min}(x)$ and $d_{\max}(x)$ denote the minimum and maximum value in the distances $d_+(x)$ and $d_-(x)$ from the sample x to two separating hypersurfaces, respectively. Here, the distances $d_+(x)$ and $d_-(x)$ are given by

$$\begin{aligned} d_+(x) &= \frac{\left|\left(K\left(x^T, C^T\right)u_1 + b_1\right)/\sqrt{u_1^T K\left(C, C^T\right)u_1} + \left(K\left(x^T, C^T\right)u_2 + b_2\right)/\sqrt{u_2^T K\left(C, C^T\right)u_2}\right|}{\sqrt{2 + 2u_1^T K\left(C, C^T\right)u_2 / u_1^T K\left(C, C^T\right)u_1 u_2^T K\left(C, C^T\right)u_2}}, \\ d_-(x) &= \frac{\left|\left(K\left(x^T, C^T\right)u_1 + b_1\right)/\sqrt{u_1^T K\left(C, C^T\right)u_1} - \left(K\left(x^T, C^T\right)u_2 + b_2\right)/\sqrt{u_2^T K\left(C, C^T\right)u_2}\right|}{\sqrt{2 - 2u_1^T K\left(C, C^T\right)u_2 / u_1^T K\left(C, C^T\right)u_1 u_2^T K\left(C, C^T\right)u_2}}. \end{aligned} \tag{7}$$

Here, the probability of the sample x belonging to class i ($i = +1, -1$) depends on the value of the output function $f(x) \in (-\infty, +\infty)$ whose value range is similar to the continuous output in SVM. The larger the value is, the bigger the probability of the sample x belongs to class +1, and vice versa.

According to the posterior probability estimating method proposed by Platt [8], the posterior output of the two-class TSVM can be approximated using a parametric form of a sigmoid function which plays an important role in the classification settings,

$$\begin{aligned} p\left(y = +1 \mid f(x)\right) &= \frac{1}{1 + e^{af(x)+b}}, \\ p\left(y = -1 \mid f(x)\right) &= \frac{e^{af(x)+b}}{1 + e^{af(x)+b}}, \end{aligned} \tag{8}$$

where a and b are the fitting parameters which can be determined using the maximum likelihood estimation. In this way, the two-class PPTSVM model is established to produce the posterior probability. Due to the fact that the posterior probability at a point is the combined effect of a number of neighboring samples, the advantage is that it can give a chance to correct the error introduced by wrongly labeled or noisy points.

2.2.2. Proposed Multiclass PPTSVM Model. In multiclass pattern recognition, there are two conventional approaches to extend the binary classification problem to the multiclass one. The first method couples the constraints of having multiple classes in a single formulation. The other one aims to convert the multiclass problem to a set of independent two-class problems by different decomposition methods. Pairwise coupling is popular multiclass classification method that combines together all pairwise comparisons for each pair of classes [12]. In this section, a multiclass extension to TSVM with probabilistic outputs is proposed based on pairwise coupling due to its stability and effectiveness.

For the multiclass classification, a sample x needs to be discriminated to belong to one of $M > 2$ classes and the class label is denoted as $y = \{1, 2, \ldots, M\}$. The goal is to estimate the class probabilities $P(x) = \{p_i(x)\}_{i=1}^{M}$, and the decision rule is $\arg\max_i p_i$, where $p_i = p(y = i \mid x)$, $i = 1, 2, \ldots, M$. Let r_{ij} $(i \neq j)$ be the estimated pairwise class probabilities, and let μ_{ij} be expectation of r_{ij}; that is, $\mu_{ij} = E(r_{ij}) = p_i/(p_i + p_j)$ [12, 14]. Hastie and Tibshirani proposed the pairwise coupling approach to estimate the class probabilities by minimizing the Kullback-Leibler (KL) distance between r_{ij} and μ_{ij} [14]. Although the method outperforms the traditional voting rule wherein it only predicts a class label rather than a class probability [15], the assumptions that the sampling mode is binomial and r_{ij} are independent do not hold in the classification problem as pointed out in [14]. To solve the problem, a more stable method is proposed in [12]. The class probability vector $P(x)$ is derived by the following optimization formulations:

$$\begin{aligned} &\underset{P}{\mathrm{Min}} \quad \frac{1}{2}\sum_{i=1}^{M}\sum_{j:j\neq i}^{M}\left(r_{ij}p_j - r_{ji}p_i\right)^2 \\ &\text{s.t.} \quad \sum_{i=1}^{M} p_i = 1, \quad p_i \geq 0. \end{aligned} \tag{9}$$

The objective function of (9) can be replaced by

$$\underset{P}{\mathrm{Min}} \frac{1}{2} P^T Q P, \quad \text{where } Q_{ij} = \begin{cases} \sum_{s:s\neq i} r_{si}^2, & i = j, \\ r_{ji} r_{ij}, & i \neq j, \end{cases} \tag{10}$$

where the matrix Q is positive semidefinite. Its solution can be obtained by solving a linear-constrained convex quadratic programming problem [12]. The practical iterative algorithm is described as follows.

Input. It is the matrix Q.

Output. It is the class probability vector P.

Process

Step 1. Set the iteration number $t = 0$, and initialize the threshold of the iterative error eps $= 0.005/M$, the maximum number of iterations itmax $= 100$, and $P_t = 1.0/M$.

Step 2. Compute the matrix QP_t and $P_t{}^T QP_t$.

Step 3. Calculate the maximum deviation errmax $= \max|QP_t - P_t{}^T QP_t|$. If errmax $<$ eps, then the iterative process stops, and output the class probability vector P; otherwise go to Step 4.

Step 4. Iteratively compute P_t according to the following pseudocode:

```
for $t = 1 : M$
{
    $\Delta = (-QP_t + P_t{}^T QP_t)/Q_{tt}$
    $P_t = P_t + \Delta$
    $P_t{}^T QP_t = (P_t{}^T QP_t + \Delta(\Delta Q_{tt} + 2QP_t))/(1 + \Delta)^2$
    for $i = 1 : M$
    {
        $QP_i = (QP_i + \Delta Q_{ti})/(1 + \Delta)$
        $P_i = P_i/(1 + \Delta)$
    }
}.
```

Step 5. If the iteration number is smaller than itmax, go to Step 2; otherwise stop iteration and output the class probability vector P.

In the implementation of multiclass PPSVM formulation, the detailed steps are given as follows:

(1) There are $M(M - 1)/2$ binary PPTSVM models for each possible pair of classes, and the corresponding pairwise class probabilities r_{ij} are estimated at each sample x.

(2) Adopt the iterative method to solve the optimization problem (10), obtaining the class probability vector $\widehat{P}(x) = \{\widehat{p}_i(x)\}_{i=1}^{M}$.

(3) The label of the test sample is discriminated according to the decision rule as $\arg\max_i \widehat{p}_i$, $i = 1, 2, \ldots, M$.

3. Results and Discussion

In this section, some experiments are conducted on UCI benchmark datasets and BCI competition datasets to study the performance of multiclassification approaches and verify the effectiveness of our proposed method. Six algorithms are used for multiclassification and separated into two groups. The first group produces hard class labels, consisting of conventional SVM and TSVM by voting [15]. The second group estimate class probabilities, including posterior probability SVM via minimizing the KL distance [14] (PPSVM_HT), posterior probability SVM by pairwise coupling [12] (PPSVM), and the proposed posterior probability TSVM via minimizing the KL distance [14] (PPTSVM_HT), as well as pairwise coupling [12] (PPTSVM).

Multiclass posterior probability TSVM is constructed based on the binary TSVM classification code (http://www.optimal-group.org/Resource/TWSVM.html) provided by Shao et al. [10] and multiclass posterior probability SVM code can refer to the mcpIncSVM package (http://www-ti.informatik.uni-tuebingen.de/~spueler/mcpIncSVM/). For all the six methods, Gaussian kernel function is chosen due to its validity and stability in experiments. All the methods are implemented in MATLAB 2013a environment on a PC with a 2.5 GHz processor and 4.0 GB RAM.

Table 1: Descriptions of benchmark datasets used in experiments.

Datasets	Total class number	Number of each class	Feature dimension	Training samples/ total samples	Testing samples/ total samples
Iris	3	[50, 50, 50]	4	96/150	54/150
Lineblobs	3	[118, 75, 73]	2	173/266	93/266
Square 1	4	[250, 250, 250, 250]	2	650/1000	350/1000
Zoo	7	[41, 20, 5, 13, 4, 8, 10]	16	66/101	35/101
Vowel_gy	11	[48, 48, 48, 48, 48, 48, 48, 48, 48, 48, 48]	10	343/528	185/528

Table 2: Best model parameters and sigmoid parameters for five UCI datasets.

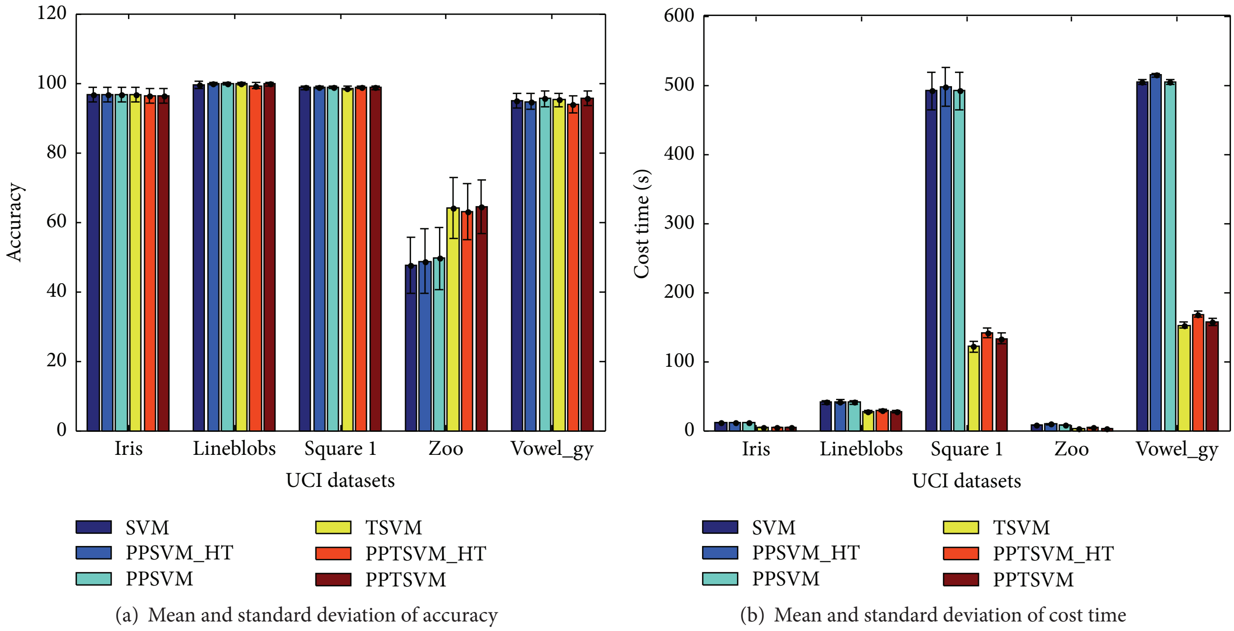
Datasets	TSVM (PPTSVM)			SVM (PPSVM)		
	c_1/c_2	g	b/a	c	g	b/a
Iris	0.25/0.25	2.25	0.4518/−8.1787	0.25	0.25	−0.0202/−4.0336
Lineblobs	0.25/0.25	0.25	0.0451/−79.2594	2.25	0.25	−0.1062/−4.8808
Square 1	0.25/0.25	3.25	−0.0967/−14.6328	1.25	1.25	0.0660/−4.9734
Zoo	0.25/0.25	0.25	0.7365/−15.2263	0.25	0.25	0.0884/−3.0649
Vowel_gy	0.25/3.25	1.25	0.2548/−93.0990	2.25	0.25	−0.2565/−4.0580

3.1. Experiments on UCI Datasets. First, our proposed method is evaluated on several multiclass datasets from the UCI Machine Learning Repository [16]. In this experiment, five datasets with multiple labels are chosen and the details are listed in Table 1. It is shown that the number of classes range from three to eleven, and the feature dimension is also different from two to sixteen.

In classification process, each dataset is first divided into training and testing sets with the ratio 65% : 35%. Secondly, k-fold cross-validation and grid search techniques are used to select the optimal kernel parameter g and penalty parameters c, c_1, c_2 for different methods, and then the sigmoid parameters a, b are obtained. Multiple rounds of cross-validation are performed using different partitions of training data in order to reduce variability, and the partitions are randomly generated by the "crossvalind" function in the MATLAB Bioinformatics toolbox. To ensure a fair comparison between different approaches, the same data partitions are used in cross-validation. Next, the trained models from training sets are applied to predict the testing sets. The classification process is repeated 100 times, and the average of these outcomes is the final classification rate. Table 2 lists the best parameters for each benchmark dataset. For each dataset, there are $M(M-1)/2$ binary classification models. For Zoo, a twofold cross-validation is conducted on the grid points $\{2^{-2}, \ldots, 2^{2}\}$ in both TSVM and SVM models because its minimum number of classes is 4. For all other four datasets, the best model parameters are searching by tenfold cross-validation from $\{2^{-2}, \ldots, 2^{2}\}$ in both TSVM and SVM models. The parameters g, c_1, c_2 of PPTSVM were set to be the same with TSVM, and the penalty parameters g, c of PPSVM are also set to be the same with SVM, as shown in Table 2.

The performance is evaluated in terms of average value and standard deviation of the accuracy and cost time. Figure 1 shows the classification performances for each benchmark dataset using PPTSVM, PSVM, TSVM, and SVM. It is shown that the TSVM-type classification methods perform similarly in all the five datasets, and the same is true for the SVM-type approaches. At the same time, TSVM-type competitors are consistently faster than SVM-type contenders due to the fact that TSVM-type methods solve two smaller sized quadratic programming problems instead of a single one with relatively larger size. The figure also indicates that TSVM-type classification methods yield higher accuracy in 3 of the 5 datasets (Lineblobs, Zoo, and Vowel_gy) and perform best on Zoo, while SVM-type approaches perform slightly better on two datasets (Iris and Square 1). Furthermore, PPTSVM performs better on two datasets (Zoo and Vowel_gy) than the other algorithms, especially gaining about 15% improvement over all the SVM-type contenders on the Zoo dataset. It demonstrates the potential advantage of TSVM-type model in dealing with smaller samples. However, all the classification methods obtain worse performance on the Zoo datasets because the number of its training datasets is comparatively small, and additionally it belongs to the imbalanced data in that it has seven classes with different class numbers. As proposed in [17] recently, it is interesting to construct corresponding TSVM-type models for the imbalanced classification problem although it is not the scope of this paper.

3.2. Applications in Classification of Motor Imagery EEG. The availability of BCI data from past competitions is an important contribution to stimulate the interdisciplinary engagement of researchers. This paper selects BCI Competition IV Dataset 2a provided by the BCI research center in Berlin, Germany [18]. Nine participants (e.g., A1~A9) were invited as experimental subjects. For each subject, four different motor imagery tasks were executed, that is, left hand, right hand, feet, and tongue. Two sessions on different days were recorded for each participant and each session was comprised of six runs separated by short breaks where one run consists of 48 trails and thus there are a total of



(a) Mean and standard deviation of accuracy

(b) Mean and standard deviation of cost time

Figure 1: Comparisons of average and standard deviation of the classification accuracy and cost time on UCI datasets using different methods.

288 trials per session. The data were recorded using 22 Ag/AgCl electrodes. The measured signals were sampled with 250 Hz and band-pass filtered between 0.5 and 100 Hz.

Concerning the classification of motor imagery tasks, it is involved with feature extraction of EEG. The common spatial patterns (CSP) method is used to extract features since it can construct new time series whose variances are optimal for the discrimination of two populations of data [19], and thus it has been applied successfully to the classification of raw movement-related EEG. Standard CSP algorithm is proposed for binary classification, and its multiclass extensions contain using CSP within the classifier, one-versus-the-rest CSP (OVR), and simultaneous diagonalization of covariance matrices from the multiclass data [20]. The OVR approach computes the CSP features that discriminate each class from the rest of the classes [20]. CSP is extended for the multiclass data using the OVR scheme in this paper. The detailed steps of EEG processing are outlined in the following;

(1) Preprocess the multichannel EEG data using a 5-order Butterworth filter, obtaining a band-pass filtered signal with the frequency band 7–35 Hz.

(2) Perform the CSP algorithm on the filtered EEG data using the OVR scheme, getting the corresponding feature vector:

$$f_p = \log\left(\frac{\operatorname{var}(Z_p)}{\sum_{j=1}^{2J}\operatorname{var}(Z_j)}\right), \quad p = 1,\ldots,2J, \tag{11}$$

where Z_p denotes the signals processed by CSP, var(⋅) is the operation of computing vector variance, and J represents the number of spatial filters. For four-class data, there are 4 projection matrices, and the optimal $2J$ directions are taken in each projection matrix. When $J = 2$, the feature vector of 16 dimensions is obtained in the experiment.

(3) Recognize the classes of motor imagery tasks using the different classifiers and compare their performances evaluated in terms of mean kappa value and cost time. Used as a performance measure in BCI competition, the kappa coefficients consider the distribution of wrong classifications and the frequency of occurrence is normalized [21]. The kappa coefficient is denoted in the following equality:

$$\text{kappa} = \frac{p_0 - p_e}{1 - p_e}, \tag{12}$$

where p_0, the overall agreement, is equal to the classification accuracy and the chance agreement p_e equals the accuracy of a trivial classifier. If the actual number of samples is equally distributed across classes, the chance expected agreement is $p_e = 1/M$, and the kappa coefficient is given by

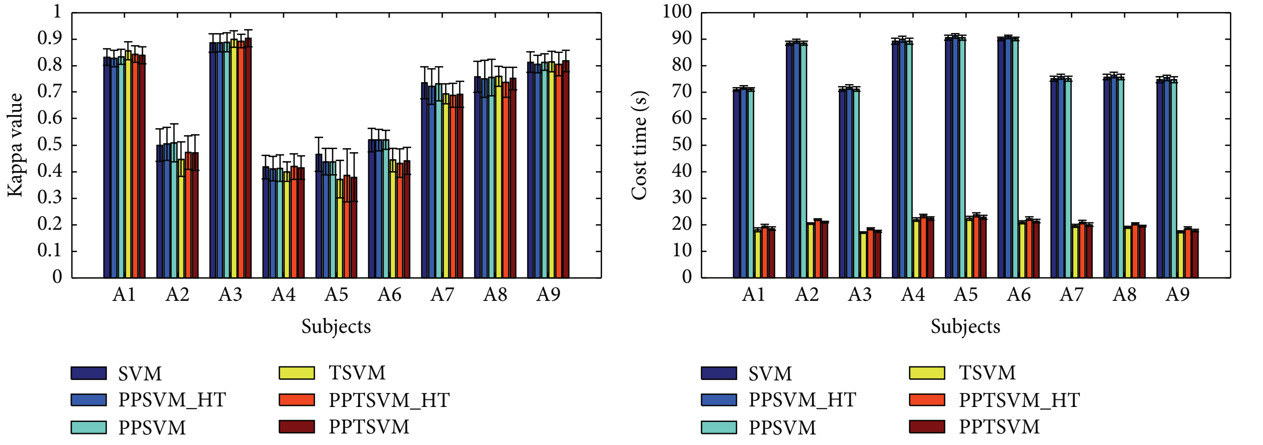
$$\text{kappa} = \frac{Mp_0 - 1}{M - 1}. \tag{13}$$

For example, an accuracy of 50% in a two-class problem is equivalent to an accuracy of 25% in a four-class problem, but the kappa value is zero in both cases, and thus it can be used to do a fair comparison of multiclass problems.

The experiment consists of two parts. In the first part, each piece of training data from BCI Competition IV Dataset 2a is randomly divided into training and testing sets with the ratio 65% : 35%, and tenfold cross-validations are performed on the cross-validation training sets, and then the trained models are applied to predict the cross-validation testing sets (part of the training data). Similar to the parameter selection process described in UCI datasets, the parameters g, c_1, c_2 of PPTSVM were set to be the same with TSVM, and the penalty parameters g, c of PPSVM are also set to be the same with SVM where the best model parameters are searching

Table 3: Best model parameters and sigmoid parameters for each subject in BCI Competition IV Dataset 2a.

Subjects	TSVM (PPTSVM)			SVM (PPSVM)		
	c_1/c_2	g	b/a	c	g	b/a
A1	0.25/0.25	2.25	0.0021/−10.8360	1.25	3.25	0.5137/−3.1760
A2	2.25/2.25	2.25	0.1625/−143.8536	2.25	2.25	0.5102/−2.9896
A3	0.25/0.25	2.25	0.0550/−13.2102	0.25	1.25	−0.2401/−3.8991
A4	0.25/2.25	2.25	0.1680/−126.5687	1.25	1.25	−0.3945/−3.1796
A5	0.25/0.25	0.25	0.0059/−186.2766	3.25	1.25	0.6197/−3.5109
A6	0.25/2.25	1.25	0.1853/−85.6740	2.25	2.25	−0.4072/−2.6546
A7	0.25/0.25	2.25	0.0806/−21.3598	3.25	2.25	−0.1001/−2.9976
A8	0.25/1.25	1.25	0.5234/−229.3538	1.25	1.25	−0.5236/−3.7116
A9	0.25/0.25	2.25	−0.1761/−19.3798	2.25	2.25	0.3187/−3.3938



(a) Mean and standard deviation of kappa value

(b) Mean and standard deviation of cost time

Figure 2: Experimental results on the mean kappa value, cost time, and their standard deviations of 10 × 10-fold cross-validation on the training data in BCI Competition IV Dataset 2a.

from $\{2^{-2}, \ldots, 2^2\}$ in both TSVM and SVM models. Table 3 summarizes the best parameters for each subject in BCI Competition IV Dataset 2a.

Figure 2 provides the classification results using kappa coefficient and cost time with mean values and standard deviations of 10 × 10-fold cross-validation. The results show that the posterior probability TSVM-type classification methods via different pairwise coupling techniques perform similarly in terms of kappa value on BCI Competition IV Dataset 2a. The same is true for different posterior probability SVM-type approaches, and the results coincide with the demonstrations in [12] when the number of classes is small; for example, it is 4 in the BCI Competition IV Dataset 2a. All the TSVM-type competitors have obvious superiority in the cost time and achieve slightly higher kappa values for 5 subjects (1, 3, 4, 8, and 9) while the SVM-type contenders perform better in 4 subjects (2, 5, 6, and 7). Over all the nine subjects, the PPTSVM method yields higher averaged mean kappa value (0.634) than TSVM (0.632) and PPTSVM_HT (0.630), but slightly lower than the best competitor SVM (0.658). Compared with the results in [22] using multiclass extensions of the CSP algorithm via the OVR approach and the naïve Bayesian Parzen window (NBPW) classifier, our method yields higher averaged mean kappa value over all the subjects than the NBPW algorithm with CSP (0.538), but lower than the NBPW algorithm with filter bank CSP approach (0.663) that employs additional feature selection based on mutual information with respect to the classical CSP algorithm. On the other hand, it is almost four times faster than the conventional SVM methods since the averaged cost time of our method is 20.040 s while that of SVM is 80.593 s, and the advantage in time complexity is helpful to BCI applications.

In the second part of the experiment, the trained models from cross-validation training sets are applied to the whole evaluation set. The classification results using average kappa value for all the algorithms are plotted in Figure 3. It is shown that the PPTSVM method achieves slightly higher mean kappa values in 3 subjects (1, 3, and 9). By contrast, SVM via voting yields the best averaged mean kappa value on the evaluation data (0.480) over all the nine subjects, and the PPTSVM method obtains higher averaged mean kappa value (0.454) than TSVM (0.452) and PPTSVM_HT (0.449).

Comparing the results of Figures 2 and 3, the results on the evaluation data are consistently lower than the cross-validation results for all six methods. Specially, PPTSVM yields lower mean kappa value averaged over all the subjects

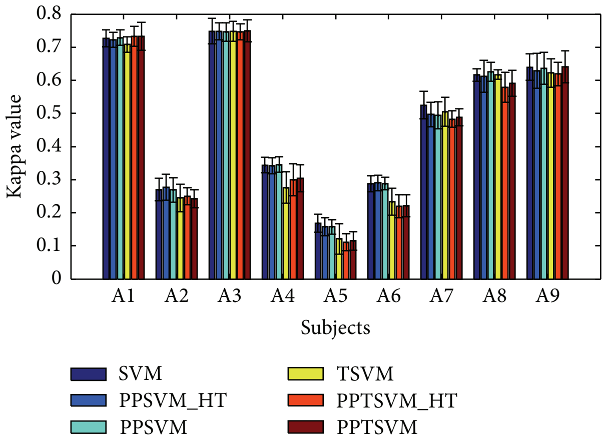


Figure 3: Experimental results on the mean kappa value and standard deviations of the evaluation data in BCI Competition IV Dataset 2a.

on the evaluation data (0.454) than the cross-validation results (0.634) in all the nine subjects.

4. Conclusions

In this paper, the posterior probability twin SVM is extended to the multiclass case by the ranking continuous and pairwise coupling. The performance of the proposed method has been tested in terms of classification accuracy and cost time on both the UCI benchmark datasets and real world EEG data from BCI Competition IV Datasets and compared with multiclass SVM/TSVM by the voting rule [15], multiclass posterior probability SVM/TSVM via minimizing the KL distance [14], and pairwise coupling in [12], respectively. The experimental results have demonstrated that the proposed method yields slightly higher averaged mean kappa value than TSVM by voting and PPTSVM_HT via the minimization of the KL distance on the BCI Competition IV Dataset 2a. In addition, it can achieve comparatively close performance to SVM competitors with lower time complexity on the UCI datasets and BCI Competition IV Dataset 2a. The decrease in time complexity is valuable in BCI applications and also its posterior probability output is useful for multimodal information fusion.

Although our method can yield satisfactory classification performance with low time cost, it is still designed in off-line settings, and thus it is worthwhile to research a framework for exact incremental learning and adaptation of multiclass posterior probability TSVM in future work. It is quite possible to have multiclass machine learning problem where one or more classes are rare compared with others, and constructing the TSVM-type models for the imbalanced classification problem is also part of the future work.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant (no. 61201302, 61172134, and 61372023), the State Scholarship Fund of China (no. 201308330297), and Zhejiang Province Natural Science Foundation of China (LY15F010009). The authors would like to thank the providers of the UCI benchmark datasets and BCI Competition IV Dataset 2a. Comments from reviewers are very much appreciated.

References

[1] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, 2006.

[2] A. J. Doud, J. P. Lucas, M. T. Pisansky, and B. He, "Continuous three-dimensional control of a virtual helicopter using a motor imagery based brain-computer interface," *PLoS ONE*, vol. 6, no. 10, Article ID e26322, 2011.

[3] N. Jrad, M. Congedo, R. Phlypo et al., "sw-SVM: sensor weighting support vector machines for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 8, no. 5, Article ID 056004, 2011.

[4] M. Hu and H. Liang, "Noise-assisted instantaneous coherence analysis of brain connectivity," *Computational Intelligence and Neuroscience*, vol. 2012, Article ID 275073, 12 pages, 2012.

[5] H. Zhang, Z. Y. Chin, K. K. Ang, C. Guan, and C. Wang, "Optimum spatio-spectral filtering network for brain-computer interface," *IEEE Transactions on Neural Networks*, vol. 22, no. 1, pp. 52–63, 2011.

[6] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs," *Pattern Recognition*, vol. 45, no. 6, pp. 2137–2144, 2012.

[7] T. Wissel, T. Pfeiffer, R. Frysch et al., "Hidden Markov model and support vector machine based decoding of finger movements using electrocorticography," *Journal of Neural Engineering*, vol. 10, no. 5, Article ID 056020, 14 pages, 2013.

[8] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, MIT Press, Cambridge, Mass, USA, 2000.

[9] M. Gönen, A. G. Tanuğur, and E. Alpaydin, "Multiclass posterior probability support vector machines," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 130–139, 2008.

[10] Y.-H. Shao, N.-Y. Deng, Z.-M. Yang, W.-J. Chen, and Z. Wang, "Probabilistic outputs for twin support vector machines," *Knowledge-Based Systems*, vol. 33, pp. 145–151, 2012.

[11] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, 2007.

[12] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *The Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[13] S. Soman and Jayadeva, "High performance EEG signal classification using classifiability and the twin SVM," *Applied Soft Computing*, vol. 30, pp. 305–318, 2015.

[14] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *The Annals of Statistics*, vol. 26, no. 2, pp. 451–471, 1998.

[15] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: a stepwise procedure for building and training a neural network," in *Neurocomputing: Algorithms, Architectures and Applications*, vol. 68 of *NATO ASI Series*, pp. 41–50, Springer, Berlin, Germany, 1990.

[16] M. Lichman, *UCI Machine Learning Repository*, School of Information and Computer Science, University of California, Irvine, Calif, USA, 2013, http://archive.ics.uci.edu/ml.

[17] Y.-H. Shao, W.-J. Chen, J.-J. Zhang, Z. Wang, and N.-Y. Deng, "An efficient weighted Lagrangian twin support vector machine for imbalanced data classification," *Pattern Recognition*, vol. 47, no. 9, pp. 3158–3167, 2014.

[18] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain-computer interface: fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539–550, 2007.

[19] G. Pfurtscheller and F. H. L. da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[20] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 993–1002, 2004.

[21] A. Schlögl, J. Kronegg, J. E. Huggins, and S. G. Mason, "Evaluation criteria for BCI research," in *Toward Brain-Computer Interfacing*, chapter 19, pp. 327–342, MIT Press, Cambridge, Mass, USA, 2007.

[22] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers in Neuroscience*, vol. 6, article 39, 9 pages, 2012.