Two Criteria for Model Selection in Multiclass Support Vector Machines

Lei Wang, *Member, IEEE*, Ping Xue, *Senior Member, IEEE*, and Kap Luk Chan, *Member, IEEE*

***Abstract*—Practical applications call for efficient model selection criteria for multiclass support vector machine (SVM) classification. To solve this problem, this paper develops two model selection criteria by *combining* or *redefining* the radius–margin bound used in binary SVMs. The combination is justified by linking the test error rate of a multiclass SVM with that of a set of binary SVMs. The redefinition, which is relatively heuristic, is inspired by the conceptual relationship between the radius–margin bound and the class separability measure. Hence, the two criteria are developed from the perspective of model selection rather than a generalization of the radius–margin bound for multiclass SVMs. As demonstrated by extensive experimental study, the minimization of these two criteria achieves good model selection on most data sets. Compared with the k-fold cross validation which is often regarded as a benchmark, these two criteria give rise to comparable performance with much less computational overhead, particularly when a large number of model parameters are to be optimized.**

***Index Terms*—Class separability measure, model selection, multiclass classification, multiclass support vector machines (SVMs), radius–margin bound.**

I. Introduction

In RECENT years, multiclass support vector machines (SVMs) have attracted much attention due to the demands for multicategory classification in many practical applications and the success of SVMs in binary classification. The methods realizing the multiclass SVMs roughly fall into three categories, namely, the methods using the strategies of one-versus-all [1] or one-versus-one [2], [3], the methods based on the *error-correcting output codes* (ECOC) approach [4], [5], and those using the *single-machine* approach [6]–[8]. Comparative studies of these methods can be found in [1] and [9]. The one-versus-one- and one-versus-all-based methods are often recommended for practical use because of lower computational cost or conceptual simplicity.

Manuscript received February 1, 2007; revised October 31, 2007. First published September 16, 2008; current version published November 20, 2008. This work was supported in part by Nanyang Technological University under Grant LIT 2002-4 of A-STAR and Grant RGM 14/02 and in part by Australian Research Council Discovery Project under Grant DP0773761. The early work of this paper was carried out at Nanyang Technological University, Singapore, Singapore, and the further work of this paper was carried out at The Australian National University, Canberra, A.C.T., Australia. This paper was recommended by Associate Editor S. Singh.

L. Wang is with the Research School of Information Sciences and Engineering, The Australian National University, Canberra, A.C.T. 0200, Australia (e-mail: Lei.Wang@mail.rsise.anu.edu.au).

P. Xue and K. L. Chan are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: epxue@ntu.edu.sg; eklchan@ntu.edu.sg).

Digital Object Identifier 10.1109/TSMCB.2008.927272

Similar to binary SVMs, multiclass SVMs also require model selection to achieve good classification performance. Overcomplex models will overfit training data, whereas oversimple models cannot effectively represent the intrinsic data structure. Both will result in poor classification performance when the classifiers are put into use. Just as its binary counterpart, the model selection of multiclass SVMs is used to select the parameters of a kernel function and the regularization parameter that balances training error and machine complexity. Very often, a single model parameter set is uniformly used across all the involved classifiers (for example, the binary SVM classifiers in the one-versus-one- or one-versus-all-based methods), rather than using different parameter sets in different binary classifiers. This is favored because of the following: 1) Much less model parameters need to be determined, particularly when the kernel function has multiple parameters; 2) past studies show little difference on classification performance [10], [11]; and 3) the risk of overfitting is reduced by using a simpler model. Hence, the focus of this paper is on the model selection for multiclass SVMs by finding the best single model parameter set.

In most of the existing work, the model selection for multiclass SVMs uses an exhaustive grid-based search method. The criterion is the k-fold or leave-one-out cross-validation error rate. Although straightforward, the model selection process in this way can become unbearably time consuming because for multiclass SVMs, we are often required to solve larger scale optimization problems. A few methods have been proposed to speed up this process. In [12], generalized approximate cross validation, which is an estimator of the leave-one-out test error rate, is extended to the multiclass setting to tune model parameters. In [13], an error bound for a multiclass SVM using the ECOC approach is developed and applied to the model selection. The grid search is still needed to find the best parameter set. These methods soon become intractable when three or more model parameters are to be tuned. A genetic algorithm has been used to search the model parameter space for model selection [14], [15]. Again, the selection process becomes very slow when the number of model parameters is large.

Practical applications of multiclass SVMs call for efficient model selection criteria, which should be able to handle more model parameters without leading to unacceptable computation cost. In recent years, model selection for binary SVMs has been well studied, and many selection criteria and methods have been developed [16]–[18]. Our proposed approach in this paper is to develop new criteria based on the principles of the successful criteria in binary SVMs for the multiclass setting. In the model selection for binary SVMs, a class of methods use nonlinear optimization techniques to maximize or minimize a certain

criterion to obtain an optimal model parameter set [19], [20]. They can achieve much more efficient model selection than the straightforward grid search. A significant progress along this direction is the method of minimizing the radius–margin bound of a binary SVM classifier [16], [21]. Chapelle *et al.* determine the derivatives of this bound with respect to model parameters, making iterative gradient-based optimization techniques applicable. The optimal model parameter set can be efficiently found after a number of iterations. This method not only significantly shortens the model selection process but can also optimize multiple model parameters simultaneously. It is much desired if such a criterion could also be extended to the multiclass setting. However, such a theoretical generalization of this bound is not that straightforward because this bound is rooted in the theoretical basis of binary SVMs. In [22], a theoretical generalization of this bound was reported but without further experimental investigation.

Although an error bound can certainly be used as a model selection criterion, it is unnecessary for a model selection criterion to be a valid error bound. As pointed out in [16], when model selection is of concern, whether the minimum (or maximum) of a criterion aligns well with lower test error rates is more important. Hence, instead of aiming to derive an error bound for a multiclass SVM, our paper focuses on developing practical and efficient model selection criteria by observing the principle of such criteria in a binary setting. In detail, the radius–margin bound for binary SVMs is exploited in the following two ways: 1) by linking the test error rates from binary and multiclass SVM classifiers, the first criterion is developed based on the pairwise combination of the radius–margin bounds of a set of binary SVMs for model selection; and 2) inspired by the relationship between the radius–margin bound and the class separability measure, the second criterion defines a new radius and margin to accommodate multiple classes. As shown later, both criteria inherit the elegant properties of the original radius–margin bound. Their derivatives with respect to model parameters can also be analytically computed, and thus, gradient-based optimization techniques are still applicable. The two criteria allow for efficient optimization for several hundreds of model parameters simultaneously. As before, the optimized kernel parameters can be used to identify more discriminative features, which can be used to perform feature selection in a multiclass scenario. To evaluate the model selection performance of the two criteria, extensive experiments were conducted on a variety of benchmark data sets with different numbers of model parameters. Although the two criteria are developed for a multiclass SVM classifier using the one-versus-one classification strategy, the model parameters selected by them are also tested on the classifiers using other classification strategies, including the one-versus-all, ECOC, and the single-machine approach. The experimental results demonstrate the simplicity, effectiveness, and efficiency of the two criteria for model selection in multiclass SVMs.

The rest of this paper is organized as follows. In Section II, the radius–margin bound is briefly introduced. To stay in focus, the details of binary and multiclass SVMs are omitted, and readers are referred to the papers cited earlier. Sections III and IV present the two model selection criteria in detail. In Section V, computational issue is discussed. Section VI presents experimental results, and the concluding remarks are drawn in Section VII.

II. Radius-Margin Bound for Binary SVMs

Let $\mathcal{D}$ denote a set of l training samples and $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\} \in (\mathbb{R}^d \times \mathcal{Y})^l$, where $\mathbb{R}^d$ denotes a d-dimensional input space, $\mathcal{Y}$ denotes the label set of $\mathbf{x}$, and y is $\{\pm 1\}$ in binary classification. A kernel is defined as $k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$, where $\phi(\cdot)$ is a possibly nonlinear mapping from $\mathbb{R}^d$ to a feature space $\mathcal{F}$, and $\boldsymbol{\theta}$ denotes the kernel parameter set. For nonseparable data, a regularization parameter C will be used, and the model parameter set becomes $\{\boldsymbol{\theta}, C\}$.

Let $\mathcal{L}(\mathcal{D})$ be the number of errors in a leave-one-out procedure performed on $\mathcal{D}$. The radius–margin bound is an upper bound of $\mathcal{L}(\mathcal{D})$. For a hard margin binary SVM, it is shown in [16] that

$$\mathcal{L}(\mathcal{D}) \leq \frac{4R^2}{\gamma^2} = 4R^2 \|\mathbf{w}\|^2 \tag{1}$$

where R is the radius of the smallest sphere enclosing the l training samples in $\mathcal{F}$, γ is the margin, $\mathbf{w}$ is the normal vector of the optimal separating hyperplane, and $\gamma = 1/\|\mathbf{w}\|$. For nonseparable data, an L2-norm soft margin SVM will be used, and the aforementioned result still holds. This is because an L2-norm soft margin can be shown as a hard margin with a slightly modified kernel function $\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j)$ [16], [23]. The relationship between $\widetilde{k}$ and k is $\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + (1/C)$ if $i = j$ and $\widetilde{k}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$ otherwise, where C is the regularization parameter mentioned earlier. This is also adopted in this paper. The squared radius R^2 is expressed as $R^2 = \min_{\|\phi(\mathbf{x}_i) - \hat{\mathbf{c}}\|^2 \leq \hat{R}^2} (\hat{R}^2)$, where $\phi(\mathbf{x}_i)$ $(i = 1, \ldots, l)$ is the image of $\mathbf{x}_i$ in $\mathcal{F}$, $\hat{R}$ is the radius of a sphere enclosing all the $\phi(\mathbf{x}_i)$, and $\hat{\mathbf{c}}$ is the center of this sphere. This leads to a quadratic optimization problem, and it can be obtained that

$$R^2 = \max_{\boldsymbol{\beta} \in \mathbb{R}^l} \left[\sum_{i=1}^{l} \beta_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j=1}^{l} \beta_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) \right]$$
$$\text{subject to} : \sum_{i=1}^{l} \beta_i = 1; \quad \beta_i \geq 0 \ (i = 1, 2, \ldots, l) \tag{2}$$

where β_i is the ith Lagrange multiplier and the center of the sphere is represented as $\hat{\mathbf{c}} = \sum_{i=1}^{l} \beta_i \phi(\mathbf{x}_i)$. As for $\|\mathbf{w}\|^2$, it can be obtained once the SVM optimization problem is solved. In detail

$$\frac{1}{2}\|\mathbf{w}\|^2 = \max_{\boldsymbol{\alpha} \in \mathbb{R}^l} \left[\sum\nolimits_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right]$$
$$\text{subject to} : \sum_{i=1}^{l} \alpha_i y_i = 0; \quad \alpha_i \geq 0 \ (i = 1, 2, \ldots, l) \tag{3}$$

where α_i is the ith Lagrange multiplier. The derivatives of R^2 and $\|\mathbf{w}\|^2$ with respect to the model parameters are given in

[16]. Let $\theta_t (\theta_t \in \boldsymbol{\theta})$ be the tth model parameter

$$\frac{\partial R^2}{\partial \theta_t} = \sum_{i=1}^{l} \beta_i^\star \frac{\partial k(\mathbf{x}_i, \mathbf{x}_i)}{\partial \theta_t} - \sum_{i,j=1}^{l} \beta_i^\star \beta_j^\star \frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial \theta_t} \tag{4}$$

where $\beta_i^\star (i = 1, 2, \ldots, l)$ is the solution of (2). The derivative of $\|\mathbf{w}\|^2$ with respect to θ_t is given as

$$\frac{\partial \|\mathbf{w}\|^2}{\partial \theta_t} = (-1) \cdot \sum_{i,j=1}^{l} \alpha_i^\star \alpha_j^\star y_i y_j \frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial \theta_t} \tag{5}$$

where $\alpha_i^\star (i = 1, 2, \ldots, l)$ is the solution of (3). This way, the derivative of the radius–margin bound with respect to θ_t is

$$\frac{\partial \left(R^2 \|\mathbf{w}\|^2\right)}{\partial \theta_t} = \|\mathbf{w}\|^2 \frac{\partial R^2}{\partial \theta_t} + R^2 \frac{\partial \|\mathbf{w}\|^2}{\partial \theta_t}. \tag{6}$$

The model selection with the radius–margin bound is briefly described as follows.

1) Set $\boldsymbol{\theta}_r$ to an initial value $\boldsymbol{\theta}_0$.
2) Based on the current $\boldsymbol{\theta}_r$, optimize for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ based on (3) and (2), respectively, and denote the optimal solutions by $\boldsymbol{\alpha}_r^\star$ and $\boldsymbol{\beta}_r^\star$.
3) Once $\boldsymbol{\alpha}_r^\star$ and $\boldsymbol{\beta}_r^\star$ are obtained, the derivative in (6) can be explicitly computed for a given $\boldsymbol{\theta}_r$. Thus, a gradient-based search method can be used to minimize $R^2\|\mathbf{w}\|^2$ with respect to $\boldsymbol{\theta}_r$. The minimizer is denoted by $\boldsymbol{\theta}_{r+1}$.
4) Stop if a given stopping criterion is satisfied and $\boldsymbol{\theta}_{r+1}$ is the selected model. Otherwise, let $\boldsymbol{\theta}_r \longleftarrow \boldsymbol{\theta}_{r+1}$ and go to Step 2).

As demonstrated, the radius–margin bound is rooted in the theoretical basis of binary SVMs, and it cannot be directly used in model selection for multiclass SVMs. In the rest of this paper, two criteria are developed based on this bound to deal with model selection in multiclass SVMs.

III. Model Selection Criterion I

Let $\mathcal{D}$ and $\mathcal{D}_t$ denote the training and test data sets, respectively. $E(\mathcal{D}_t)$ denotes the number of misclassified samples obtained by applying a multiclass SVM classifier to $\mathcal{D}_t$. The classifier and the test set are assumed to be fixed but unknown. For a c-class problem

$$E(\mathcal{D}_t) = \sum_{i=1}^{c} \sum_{j=1, j\neq i}^{c} E_{ij}(\mathcal{D}_t). \tag{7}$$

$E_{ij}(\mathcal{D}_t)$ denotes the number of samples misclassified from class i to class j,[1] and it is expressed as

$$E_{ij}(\mathcal{D}_t) = \left|\left\{\mathbf{x} | \mathbf{x} \in \mathcal{D}_t, y^0(\mathbf{x}) = i, y^m(\mathbf{x}) = j\right\}\right| \tag{8}$$

where $|\cdot|$ denotes the size of a set.[2] A sample $\mathbf{x}$ in $\mathcal{D}_t$ will be counted into $E_{ij}(\mathcal{D}_t)$ if and only if its true label y^0 is i, whereas the label y^m predicted by a multiclass SVM classifier is j. Considering that both true and predicted labels are unique for each sample,[3] a misclassified sample will not be counted into two different E_{ij}'s. Hence, there is no overlapping among these E_{ij}'s.

Let us focus on the one-versus-one strategy with the *max wins* classification rule [9]. It is commonly used to solve multiclass SVM problems. With this strategy, a set of $c(c-1)/2$ pairwise binary SVM classifiers are constructed. Let SVM_{ij} denote the binary SVM classifier trained with the samples from classes i and j. The $E'_{ij}(D_t)$ is the number of test samples which belong to class i but are misclassified to class j when SVM_{ij} is applied to classes i and j. For the convenience of notation, the label predicted by SVM_{ij} is written as i or j although it is $+1$ or -1 in general. The $E'_{ij}(\mathcal{D}_t)$ is formally expressed as

$$E'_{ij}(\mathcal{D}_t) = \left|\left\{\mathbf{x} | \mathbf{x} \in \mathcal{D}_t, y^0(\mathbf{x}) = i, y^b_{ij}(\mathbf{x}) = j\right\}\right| \tag{9}$$

where $y^b_{ij}(\mathbf{x})$ stands for the label predicted by the binary SVM classifier, SVM_{ij}. The total number of errors made by the $c(c-1)/2$ binary SVM classifiers is

$$E'(\mathcal{D}_t) = \sum_{1\leq i,j\leq c, i\neq j} E'_{ij}(\mathcal{D}_t). \tag{10}$$

The following proves that $E(\mathcal{D}_t)$ *is upper bounded by* $E'(\mathcal{D}_t)$. Under the rule of max wins [9], the label of a test sample $\mathbf{x}$ is decided by

$$y^m(\mathbf{x}) = \arg \max_{i=1,\ldots,c} S_i(\mathbf{x}) = \arg \max_{i=1,\ldots,c} \left(\sum_{j=1, j\neq i}^{c} \text{sign}\left[\langle \mathbf{w}_{ij}, \phi(\mathbf{x})\rangle + b_{ij}\right] \right) \tag{11}$$

where $\langle \mathbf{w}_{ij}, \phi(\mathbf{x})\rangle + b_{ij}$ is positive if $\mathbf{x}$ is classified to class i. The $\text{sign}(a)$ denotes the sign function, and it is $+1$ for $a > 0, 0$ for $a = 0$, and -1 otherwise. The summation over the $(c-1)$ sign functions is a score, and it is denoted by $S_i(\mathbf{x})$ for class i. The sample $\mathbf{x}$ is assigned to the class having the highest score. This rule immediately leads to the following three results.

1) $\forall \mathbf{x} \in \mathcal{D}_t$, there must be $S_i(\mathbf{x}) \leq (c-1)(i = 1, \ldots, c)$, and the equality is achieved if and only if all the $(c-1)$ binary SVM classifiers SVM_{ij} $(j = 1, \ldots, c, j \neq i)$ classify $\mathbf{x}$ to class i.
2) If $S_i(\mathbf{x}) < S_j(\mathbf{x})$, there must be $S_i(\mathbf{x}) < (c-1)$. Referring to result 1), this indicates that at least one of the $(c-1)$ binary SVM classifiers does not classify $\mathbf{x}$ to class i.

[1] Without loss of generality, the cost of misclassification is considered as identical among the classes in (7). The case having different misclassification costs will be discussed at the end of Section IV.

[2] Please note that according to the definition of $\mathcal{D}_t$, "$\mathbf{x} \in \mathcal{D}_t$" in (8) should be written as "$(\mathbf{x}, y^0(\mathbf{x})) \in \mathcal{D}_t$." However, the former is used in this paper for the convenience of notation.

[3] In multilabel classification, the true and predicted labels may not be unique for a sample. This paper confines itself to multiclass problems.

3) If $S_i(\mathbf{x}) = S_j(\mathbf{x})$, then both of them must be smaller than $(c-1)$. This is because the binary SVM$_{ij}$ cannot classify $\mathbf{x}$ to both classes i and j simultaneously.

Assume that a multiclass SVM misclassifies a test sample $\mathbf{x}_t (\mathbf{x}_t \in \mathcal{D}_t)$. That is, the true label $y^0(\mathbf{x}_t)$ is i, whereas the predicted label $y^m(\mathbf{x}_t)$ is j. This contributes one count to $E_{ij}(\mathcal{D}_t)$ based on (8). By referring to (11), this means that $S_j(\mathbf{x}_t)$ is the highest score, and hence, $S_i(\mathbf{x}_t) \leq S_j(\mathbf{x}_t)$. By applying results 2) and 3), it is obtained that $S_i(\mathbf{x}_t) < (c-1)$, indicating that at least one of the $(c-1)$ binary SVMs has misclassified the sample $\mathbf{x}_t$. This contributes one count to $E'_{ik}(\mathcal{D}_t)$; however, please note that k *is not necessary to be exactly the* j *in* $E_{ij}(\mathcal{D}_t)$. Therefore, for any test sample misclassified by a multiclass SVM, it must have been misclassified by at least one binary SVM classifier. Summing E_{ij} and E'_{ik} over i and j (or k) gives rise to

$$\sum_{1\leq i,j\leq c, i\neq j} E_{ij} \leq \sum_{1\leq i,k\leq c, i\neq k} E'_{ik} \Longleftrightarrow E(\mathcal{D}_t) \leq E'(\mathcal{D}_t). \quad (12)$$

This proves that $E(\mathcal{D}_t)$ is upper bounded by $E'(\mathcal{D}_t)$. Meanwhile, it is worth mentioning that $E_{ij}(\mathcal{D}_t) \leq E'_{ij}(\mathcal{D}_t)$ is not necessary to be true.

The aforementioned result suggests that to reduce the value of $E(\mathcal{D}_t)$, we could seek to minimize its upper bound $E'(\mathcal{D}_t)$. This leads to one model selection criterion as follows. As known from (1) in Section II, the test error $(E'_{ij} + E'_{ji})$ can be estimated through the leave-one-out error of SVM$_{ij}$, which is denoted by $\mathcal{L}_{ij}$, that satisfies

$$\mathcal{L}_{ij} \leq 4R_{ij}^2 \|\mathbf{w}_{ij}\|^2. \quad (13)$$

Thus, the $E'(\mathcal{D}_t)$ can be estimated by $\sum_{1\leq i<j\leq c} \mathcal{L}_{ij}$, and it satisfies

$$\sum_{1\leq i<j\leq c} \mathcal{L}_{ij} \leq \sum_{1\leq i<j\leq c} 4R_{ij}^2 \|\mathbf{w}_{ij}\|^2. \quad (14)$$

To minimize $E'(\mathcal{D}_t)$ (or more precisely, to minimize its estimate), the right side has to be minimized.

Based on the aforementioned analysis, the $\sum_{1\leq i<j\leq c} R_{ij}^2 \|\mathbf{w}_{ij}\|^2$ is defined as a model selection criterion for multiclass SVMs. It is a pairwise combination of the radius–margin bounds of the binary SVM classifiers. The optimal model parameter set is obtained by

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\sum_{1\leq i<j\leq c} R_{ij}^2 \|\mathbf{w}_{ij}\|^2 \right). \quad (15)$$

The derivative of this criterion with respect to the tth model parameter θ_t is

$$\frac{\partial}{\partial\theta_t} \left(\sum_{1\leq i<j\leq c} R_{ij}^2 \|\mathbf{w}_{ij}\|^2 \right) = \sum_{1\leq i<j\leq c} \left(\|\mathbf{w}_{ij}\|^2 \frac{\partial R_{ij}^2}{\partial\theta_t} + R_{ij}^2 \frac{\partial \|\mathbf{w}_{ij}\|^2}{\partial\theta_t} \right). \quad (16)$$

The calculation of $\partial R_{ij}^2/\partial\theta_t$ and $\partial\|\mathbf{w}_{ij}\|^2/\partial\theta_t$ follows (4) and (5). As in a binary classification, the optimal model parameter set $\boldsymbol{\theta}^*$ can be found by using gradient-based optimization techniques.

Before ending this section, it is interesting to look into the relationship between the proposed model selection criterion and the radius–margin bound generalized for multiclass SVMs in [22]. In that work, the multiclass SVM is solved by the *single-machine* approach. With the notations in this paper, the generalized bound in [22] can be expressed as

$$\mathcal{L}(\mathcal{D}) \leq (4K/c) \left(R^2 \sum_{1\leq i<j\leq c} \|\mathbf{w}_i - \mathbf{w}_j\|^2 \right) \triangleq (4K/c) \left(R^2 \sum_{1\leq i<j\leq c} \|\widetilde{\mathbf{w}}_{ij}\|^2 \right) \quad (17)$$

where K is a constant and c is the number of classes. In [22], R denotes the radius of the smallest sphere enclosing the support vectors only. In this paper, R is changed to enclose all the training samples. Note that such a change will not affect the "$\leq$" in (17) because the new R is an upper bound of the original one. The work in [22] adopts the multiclass SVMs proposed by [6]. There, the $(\mathbf{w}_i - \mathbf{w}_j)$ can be understood as a $\widetilde{\mathbf{w}}_{ij}$, which is a normal vector of an SVM hyperplane separating classes i and j. For the proposed Criterion I in (15), R_{ij} is the radius of the smallest sphere enclosing the training samples from classes i and j, and therefore, $R_{ij}^2 \leq R^2$. Replacing all R_{ij}^2 in (15) with R^2 and moving R^2 out of the summation sign turn Criterion I to $(R^2 \sum_{1\leq i<j\leq c} \|\mathbf{w}_{ij}\|^2)$. If the constant $(4K/c)$ is ignored, the proposed Criterion I and the generalized bound in [22] will share similar structures. Surely, from the perspective of generalizing a bound in a strict theoretical sense, the approach in [22] is more suitable.

IV. Model Selection Criterion II

Class separability is a concept widely used in pattern recognition [24]–[26]. The scatter-matrix-based measure is often favored, thanks to its simplicity and applicability to both binary and multiclass problems. They are defined as

$$\mathbf{S}_W = \sum_{i=1}^{c} \left[\sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^\top \right]$$

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^\top$$

$$\mathbf{S}_T = \sum_{i=1}^{c} \left[\sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^\top \right] = \mathbf{S}_W + \mathbf{S}_B. \quad (18)$$

c is the number of classes, $\mathcal{D}_i$ is the set of training samples from class i, and n_i is the size of $\mathcal{D}_i$. $\mathbf{m}_i$ and $\mathbf{m}$ are the class and total means, respectively. Many combinations of two of $\mathbf{S}_W$, $\mathbf{S}_B$, and $\mathbf{S}_T$ can be used as a class separability measure. The commonly used ones include $\mathrm{tr}(\mathbf{S}_B)/\mathrm{tr}(\mathbf{S}_W)$ and $|\mathbf{S}_B|/|\mathbf{S}_W|$,

where $\mathrm{tr}(\mathbf{A})$ and $|\mathbf{A}|$ denote the trace and determinant of a square matrix $\mathbf{A}$, respectively. Other combinations can be found in [26].

In our previous work [27], we restrict to binary classification and preliminarily discuss the relationship between the scatter-matrix-based class separability measure and the radius–margin bound. Now, this discussion is extended to a multiclass case and is used to develop the second model selection criterion. To do so, the following first extends the class separability to a kernel-induced feature space $\mathcal{F}$. Considering that the high dimensionality of $\mathcal{F}$ can easily make the scatter matrices singular and their determinants zero, the trace-based measure is used instead. In the following, the superscript ϕ is used to distinguish the variables in $\mathcal{F}$ from those in $\mathbb{R}^d$. Recall that $\mathcal{D}_i$ denotes the training samples from the ith class. $\mathcal{D}$ is defined as the union of $\mathcal{D}_i$ $(i = 1, 2, \ldots, c)$, which is expressed as $\mathcal{D} = \cup_{i=1}^{c} \mathcal{D}_i$. $\mathbf{K}_{\mathcal{A},\mathcal{B}}$ is a kernel matrix where $\{\mathbf{K}_{\mathcal{A},\mathcal{B}}\}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, with the constraints of $\mathbf{x}_i \in \mathcal{A}$ and $\mathbf{x}_j \in \mathcal{B}$. $\mathrm{Sum}(\cdot)$ denotes the summation of all the elements in a matrix. The traces are obtained as

$$\mathrm{tr}\left(\mathbf{S}_B^{\phi}\right) = \sum_{i=1}^{c} \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_i,\mathcal{D}_i})}{n_i} - \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n} \tag{19}$$

$$\mathrm{tr}\left(\mathbf{S}_W^{\phi}\right) = \mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) - \sum_{i=1}^{c} \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_i,\mathcal{D}_i})}{n_i} \tag{20}$$

$$\mathrm{tr}\left(\mathbf{S}_T^{\phi}\right) = \mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) - \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n}. \tag{21}$$

To facilitate analysis, the class separability measure in $\mathcal{F}$ is defined as $\mathrm{tr}(\mathbf{S}_B^{\phi})/\mathrm{tr}(\mathbf{S}_T^{\phi})$ instead of $\mathrm{tr}(\mathbf{S}_B^{\phi})/\mathrm{tr}(\mathbf{S}_W^{\phi})$. Note that they are essentially identical because $\mathrm{tr}(\mathbf{S}_T^{\phi}) = \mathrm{tr}(\mathbf{S}_B^{\phi}) + \mathrm{tr}(\mathbf{S}_W^{\phi})$.

Recall that n_1 and n_2 are the sizes of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. The relationship between $\mathrm{tr}(\mathbf{S}_B^{\phi})$ and the squared margin γ^2 can be proven as (the proof is omitted)

$$\gamma^2 \leq \frac{1}{4 - \left(\frac{n_1+n_2}{n_1 n_2}\right)\mathrm{tr}\left(\mathbf{S}_B^{\phi}\right)} = \frac{1}{4 - \left\|\mathbf{m}_1^{\phi} - \mathbf{m}_2^{\phi}\right\|^2}. \tag{22}$$

This result indicates that $1/(4 - \|\mathbf{m}_1^{\phi} - \mathbf{m}_2^{\phi}\|^2)$ is an upper bound of γ^2. The equality in "$\leq$" is achieved if and only if the solution of the problem in (3), denoted by $\alpha_i^{\star}$, is $1/n_1$ for $\mathbf{x}_i \in \mathcal{D}_1$ and $1/n_2$ for $\mathbf{x}_i \in \mathcal{D}_2$. Considering that such a solution seldom occurs in practice, $1/(4 - \|\mathbf{m}_1^{\phi} - \mathbf{m}_2^{\phi}\|^2)$ is a strict upper bound in general. Recall that when minimizing the radius–margin bound for the model selection, γ^2 is to be maximized. Based on (22), to allow γ^2 to be maximized, its upper bound needs to be adequately large, and it will prevent γ^2 from being increased otherwise. This, in turn, requires $\|\mathbf{m}_1^{\phi} - \mathbf{m}_2^{\phi}\|^2$ to be adequately large. Meanwhile, decreasing the value of $\|\mathbf{m}_1^{\phi} - \mathbf{m}_2^{\phi}\|^2$ will reduce the upper bound value, forcing γ^2 to be kept small. Please note that although a larger (or smaller) $\|\mathbf{m}_1^{\phi} - \mathbf{m}_2^{\phi}\|^2$ does not necessarily lead to a larger (or smaller) γ^2, their values are often strongly positively correlated to each other in practice, which can be seen from the results comparing the values of $-\mathrm{tr}(\mathbf{S}_B^{\phi})$ and $\|\mathbf{w}\|^2$ in our previous work [27].

A similar result can be proven for the squared radius R^2 as

$$R^2 \geq \frac{1}{(n_1+n_2)}\mathrm{tr}\left(\mathbf{S}_T^{\phi}\right). \tag{23}$$

It shows that $\mathrm{tr}(\mathbf{S}_T^{\phi})/(n_1+n_2)$ is a lower bound of R^2. The equality in "$\geq$" is achieved if and only if the solution of the problem in (2), denoted by $\beta_i^{\star}$, is $1/(n_1 + n_2)$ for all the training samples. Again, such a solution is rare in practice, and this is a strict lower bound in general. When minimizing the radius–margin bound for the model selection, R^2 is to be minimized. Based on (23), this needs $\mathrm{tr}(\mathbf{S}_T^{\phi})$ to be adequately small to avoid hindering the decrease of R^2. In addition, it can be seen from [27] that the values of $\mathrm{tr}(\mathbf{S}_T^{\phi})$ and R^2 are often strongly positively correlated.

Conceptually speaking, $\|\mathbf{m}_1^{\phi} - \mathbf{m}_2^{\phi}\|^2$ and γ^2 reflect the similar property of data separation, whereas $\mathrm{tr}(\mathbf{S}_T^{\phi})$ and R^2 measure the similar property of data scattering. Inspired by the aforementioned results, this paper transplants the radius–margin bound to a multiclass scenario by mimicking the class separability measure. At the same time, please note that this new model selection criterion will still be based on R and $\|\mathbf{w}\|$ rather than the traces of the scatter matrices.

In a multiclass case, $\mathrm{tr}(\mathbf{S}_T^{\phi})/(n_1+n_2)$ measures the average of the squared scattering radius of the training samples in $\mathcal{F}$. Considering the analogy between $\mathrm{tr}(\mathbf{S}_T^{\phi})/(n_1+n_2)$ and R^2 in a binary classification, the new criterion redefines R^2 as the radius of the smallest sphere enclosing *all* the training samples from the c classes

$$R_c^2 = \min_{\|\phi(\mathbf{x}) - \hat{\mathbf{c}}\|^2 \leq \hat{R}^2} (\hat{R}^2) \qquad \forall \mathbf{x} \in \mathcal{D}. \tag{24}$$

For $\mathrm{tr}(\mathbf{S}_B^{\phi})$, it can be shown that in the case of c classes

$$\mathrm{tr}\left(\mathbf{S}_B^{\phi}\right) = \frac{\sum_{1 \leq i < j \leq c} n_i n_j \left\|\mathbf{m}_i^{\phi} - \mathbf{m}_j^{\phi}\right\|^2}{n^2}. \tag{25}$$

By noting the analogy between $\|\mathbf{m}_1^{\phi} - \mathbf{m}_2^{\phi}\|^2$ and γ^2 in a binary classification, the margin in the new criterion is redefined as

$$\overline{\gamma}^2 = \frac{\sum_{1 \leq i < j \leq c} n_i n_j \gamma_{ij}^2}{n^2} = \sum_{1 \leq i < j \leq c} P_i P_j \|\mathbf{w}_{ij}\|^{-2} \tag{26}$$

where γ_{ij} is the margin of the binary SVM classifier trained with the training samples of classes i and j, and $P_i = n_i/n$, which is the prior probability of class i estimated from the training samples. The redefined margin is a weighted average of those from the pairwise binary SVM classifiers, and the weight is the product of the prior probabilities of the two involved classes. This implies that the margins between the classes dominating the training and test sets need to be emphasized. Otherwise, the number of misclassified samples will be high. This agrees with the intuition. In this way, the second model

TABLE I
COMPARISON OF COMPUTATIONAL LOAD

Method	Computational load
CV/LOO (One-vs-all) [1]	$s^{\lvert\boldsymbol{\theta}\rvert} k \cdot c \cdot \mathrm{QP}(n)$
CV/LOO (One-vs-one) [9]	$s^{\lvert\boldsymbol{\theta}\rvert} k \cdot \sum_{1\le i<j\le c} \mathrm{QP}(n_i+n_j)$
LOO-bound (ECOC dense encoding) [13]	$s^{\lvert\boldsymbol{\theta}\rvert} \cdot [10 \log_2 c] \cdot \mathrm{QP}(n)$
LOO-bound (ECOC sparse encoding) [13]	$s^{\lvert\boldsymbol{\theta}\rvert} \cdot [15 \log_2 c] \cdot \mathrm{QP}(0.5n)$
GACV (Single-machine) [12]	$s^{\lvert\boldsymbol{\theta}\rvert} \cdot \mathrm{QP}(cn)$
Criterion I	$e \cdot \left[2 \sum_{1\le i<j\le c} \mathrm{QP}(n_i+n_j)\right]$
Criterion II	$e \cdot \left[\mathrm{QP}(n) + \sum_{1\le i<j\le c} \mathrm{QP}(n_i+n_j)\right]$

* CV : Cross-validation, LOO : Leave-one-out.

selection criterion is obtained, and the optimal model parameter set is given by

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\frac{R_c^2}{\overline{\gamma}^2}\right). \tag{27}$$

The derivative of this criterion with respect to the tth model parameter θ_t is

$$\frac{\partial}{\partial\theta_t}\left(\frac{R_c^2}{\overline{\gamma}^2}\right) = \frac{1}{\overline{\gamma}^4}\left(\overline{\gamma}^2 \frac{\partial R_c^2}{\partial\theta_t} - R_c^2 \frac{\partial\overline{\gamma}^2}{\partial\theta_t}\right) \tag{28}$$

where

$$\frac{\partial\overline{\gamma}^2}{\partial\theta_t} = -\left(\sum_{1\le i<j\le c} P_i P_j \|\mathbf{w}_{ij}\|^{-4} \frac{\partial\|\mathbf{w}_{ij}\|^2}{\partial\theta_t}\right). \tag{29}$$

Again, the minimization of this bound can be achieved by using the gradient-based optimization techniques. Compared with Criterion I, this criterion is more heuristic and is farther from being interpreted as a bound of generalization error.

Finally, please note that these two criteria can be conveniently extended to handle the case where the misclassification costs between different classes are different. The first criterion is currently a pairwise combination of the radius–margin bounds with equal weights. When different misclassification costs are defined, a weighted combination can be applied instead. A larger weight will be assigned if the misclassification cost among a certain pair of classes is higher. For the second criterion, the weighting can be applied to the pairwise combination of margins. In the terminology of class separability, this means that two classes with higher misclassification cost will be pushed farther away from each other to reduce the potential misclassification chances.

V. COMPUTATIONAL ISSUE

Both criteria have continuous first- and second-order derivatives with respect to the model parameters as long as the employed kernel function has. The minimization of them can be efficiently solved by using gradient-based optimization techniques. For example, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton method [28] is often favored because it commonly takes a smaller number of iterations before convergence. At each iteration, the computational load is largely due to the evaluation of the objective function. Let $\mathrm{QP}(n)$ denote a quadratic programming problem with n training samples. Let s be the number of values tried for each model parameter in an exhaustive grid-based search method. $|\boldsymbol{\theta}|$ is the number of model parameters to be optimized, and k is the number of folds of cross validation. Besides these, e stands for the number of function evaluations in an optimization process. The computational loads of the model selection methods reviewed in Section I are listed in Table I. Following [16], the measure in terms of *the total number of QP problems to be solved* is used.[4] As shown in Table I, for a multiclass SVM classifier using the one-versus-all strategy, training this classifier results in the computational cost of $c \cdot \mathrm{QP}(n)$. For that using the one-versus-one strategy, this result becomes $\sum_{1\le i<j\le c} \mathrm{QP}(n_i+n_j)$. Calculating R_c^2 and R_{ij}^2 (or $\|\mathbf{w}_{ij}\|^2$) in the proposed criteria involves one $\mathrm{QP}(n)$ and one $\mathrm{QP}(n_i+n_j)$, respectively. From this table, it is found that the computational load of grid-based search methods increases rapidly with the increasing value of s, $|\boldsymbol{\theta}|$, or k. They quickly become intractable when $|\boldsymbol{\theta}|$ is larger than three. In contrast, the proposed criteria have a much lower computational load, thanks to the applicability of gradient-based optimization techniques. As shown in the experimental study, the minimization of them can be accomplished in a few iterations with a number of function evaluations, even if $|\boldsymbol{\theta}|$ is as large as several hundreds. Compared with the existing methods, the proposed criteria can save considerable computational cost, and the more the model parameters, the more the savings will be.

VI. EXPERIMENTAL RESULT

This experiment evaluates the effectiveness of the proposed criteria for the model selection of multiclass SVMs. Two forms of the Gaussian radial basis function (GRBF) kernel are used. One is the *spherical* GRBF kernel $k(\mathbf{x},\mathbf{y}) = \exp(-(\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2))$, where σ is the kernel parameter known as the Gaussian width. In this case, the model parameter set is $\boldsymbol{\theta} = \{C, \sigma\}$, and C is the regularization parameter. The other is the *ellipsoidal* GRBF kernel that assigns different σ's to each feature dimension. It is expressed as $k(\mathbf{x},\mathbf{y}) = \exp(-\sum_{i=1}^{d}((x_i-y_i)^2/2\sigma_i^2))$, where σ_i is the Gaussian width for the ith dimension. At this time, $\boldsymbol{\theta}$ expands to $\{C, \sigma_1, \sigma_2, \ldots, \sigma_d\}$. In this experiment, the two kernels are used to evaluate the performance of the proposed criteria in

[4]The computational load is also affected by the dimensionality of data and the computation of a kernel function. They are considered as constants for a given multiclass SVM classification problem.

TABLE II
ATTRIBUTES OF THE MULTICLASS BENCHMARK DATA SETS

Data	d	#Class	n	n_{test}	max/min/avg number of samples per class
Iris	4	3	150	–	50/50/50
Wine	13	3	178	–	71/48/59.3
Glass	13	6	214	–	76/0/35.7
Car	6	4	1728	–	1210/65/432
Dermatology	34	6	366	–	112/20/61
E.coli	7	8	336	–	143/2/42
Yeast	8	10	1484	–	463/5/148.4
Zoo	16	7	101	–	41/4/14.4
Vowel	10	11	528	–	48/48/48
Vehicle	18	4	846	–	218/199/211.5
Segment	19	7	2310	–	330/330/330
DNA	180	3	2000	1186	1654/765/1062
Satimage	36	6	4435	2000	1533/626/1072.5

dealing with small- and large-sized model parameter sets, respectively.

The BFGS quasi-Newton method is employed to minimize the two criteria to find the optimal model parameter set. To avoid the constraints of $C > 0$ and $\sigma > 0$, the transforms of $\mu = -\ln(C)$ and $\nu = \ln(g) = \ln(1/2\sigma^2)$ are applied, where $\ln(\cdot)$ denotes the natural logarithm. μ and ν (or $\nu_1, \ldots, \nu_d$ when an ellipsoidal GRBF kernel is used) are optimized instead. Thus, the minimization of the criteria becomes an unconstrained optimization problem. The initial values of μ and ν are set as $\mu_0 = 0$ (or, equally, $C = 1$) and $\nu_0 = -\ln(2d)$ (or, equally, $\sigma = \sqrt{d}$), where d is the dimension of the feature vector. Note that for a data set, the feature components along each dimension have been linearly scaled to $[-1, +1]$ by using the training samples. The feature components of the test samples will be scaled with the same scaling parameters when doing classification.

Following the work in [21], two stop criteria are used, and the optimization will terminate when either of them is satisfied. Let $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$ denote the model parameters obtained in the tth and $(t+1)$th iterations, respectively. $f(\boldsymbol{\theta}_t)$ and $f(\boldsymbol{\theta}_{t+1})$ are the corresponding values of the objective function. The first stop criterion is $|f(\boldsymbol{\theta}_{t+1}) - f(\boldsymbol{\theta}_t)| \leq \epsilon f(\boldsymbol{\theta}_t)$, where ϵ is a small positive number which is set as 10^{-5} in this experiment. With this stop criterion, the optimization will terminate if the difference of the function values in two consecutive iterations is less than a predefined tolerance. The second stop criterion is specifically designed for the minimization of the radius–margin bound in [21]. At each iteration of the BFGS quasi-Newton algorithm, a line search is carried out to find the starting point and direction for the next iteration. However, as pointed out in [21], too many line searches may be conducted at a single iteration when the minimum of the bound is being approached. The reason is that in practice, the derivatives computed based on (4) and (5) may slightly deviate from their theoretical values. When the minimum of the bound is being approached, the derivatives will be small, and the impact of this deviation will become significant. Due to the inaccurate derivative information, an iteration may take many line searches to find a solution. To deal with this, the second stop criterion terminates the optimization if the number of line searches at an iteration exceeds a predefined value, for example, ten.

Benchmark data sets from **UCI** Machine Learning Repository and **Statlog** Project are used, and they are listed in Table II. d denotes the dimensionality, "#Class" is the number of classes, and n is the size of a data set. For "DNA" and "Satimage," n and n_{test} are the sizes of the training and test sets, respectively. The last column in Table II lists the maximum, minimum, and average numbers of samples in each class. Following the work in [16] and [29], for a data set without predefined training/test sets, the whole data set is randomly split as 100 pairs of training/test subsets (50%:50%), and the first five training subsets are used for model selection. For "DNA" and "Satimage," the predefined training sets are randomly split as five pairs of training/test subsets, and the five training subsets are used for the model selection. This experiment uses a mixture of codes in **C** and **Matlab**. The **C** version of LIBSVM [30] is used to optimize R^2 and $\|\mathbf{w}\|^2$, as well as in training and testing an SVM classifier. The BFGS quasi-Newton algorithm is realized by using the Matlab function $\mathbf{fminunc}()$.

There are three parts in this experiment. First, the properties of the proposed two criteria are demonstrated. Second, their effectiveness for multiclass SVM model selection is evaluated on the benchmark data sets. Finally, its application to feature selection is briefly demonstrated through a toy problem and an optical digit recognition task.

A. Properties of the Two Model Selection Criteria

1) Relation Between $\sum E_{ij}$, $\sum E'_{ij}$, and Criterion I: As defined earlier, E_{ij} is the number of test samples misclassified from class i to j when a multiclass SVM classifier is applied, whereas E'_{ij} is such a number when a binary SVM classifier SVM_{ij} is used. In this paper, the minimization of $\sum E_{ij}$ is sought by minimizing its upper bound $\sum E'_{ij}$, and this gives rise to Criterion I. In this experiment, with different pairs of C and σ, the values of $\sum E_{ij}$, $\sum E'_{ij}$, and Criterion I are calculated and compared. The results on the "Wine" and "Vowel" data sets are shown in Fig. 1. The axes are $\ln(C)$, $\ln(\sigma)$, and test error, respectively. The curved surfaces of $\sum E_{ij}$ and $\sum E'_{ij}$ are labeled by arrows. It can be seen from both subfigures that the two surfaces show similar profiles with respect to $\ln(C)$ and $\ln(\sigma)$, although their magnitudes are different when the test error is large. To quantitatively measure the correlation between them, $\sum E_{ij}$ and $\sum E'_{ij}$ are treated as two random variables. The correlation coefficient ρ is calculated and listed under each subfigure. The two variables are found to be strongly positively correlated, indicating that a smaller $\sum E'_{ij}$ generally corresponds to a smaller $\sum E_{ij}$. Similar results are also observed from other data sets. These results preliminarily show that it is sensible to use $\sum E'_{ij}$ to estimate $\sum E_{ij}$ for the model selection.

Now, the correspondence between the test error $\sum E_{ij}$ and Criterion I is further checked. The $\sum E_{ij}$ and the criterion values often do not have such a strong correlation as $\sum E_{ij}$ and $\sum E'_{ij}$ do. This is because the radius–margin bound is not a tight bound of test error in a binary classification [16]. Fig. 2 shows their correspondence on the data set of "Vowel." Fig. 2(a) and (b) show the values of $\sum E_{ij}$ and Criterion I, respectively. This is a top view, and the color bar on the right side indicates

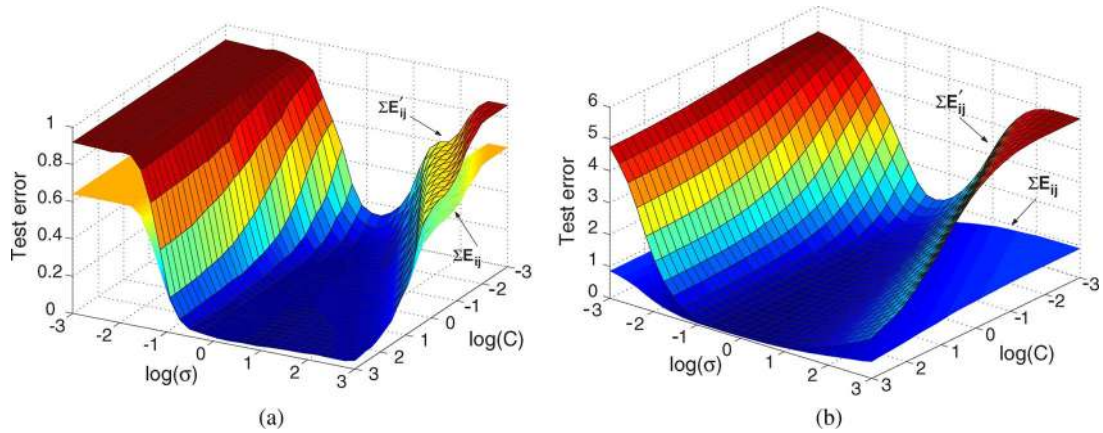

Fig. 1. Correspondence between $\sum E'_{ij}$ and $\sum E_{ij}$ (ρ is the correlation coefficient between them). (a) Wine, $\rho = 0.998$. (b) Vowel, $\rho = 0.942$.

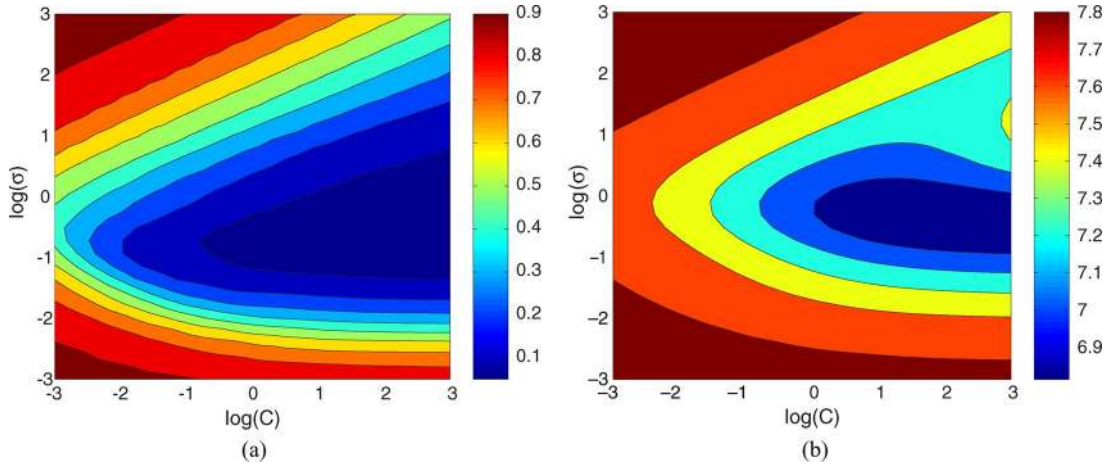

Fig. 2. Correspondence between $\sum E_{ij}$ and Criterion I. (a) Vowel, $\sum E_{ij}$. (b) Vowel, Criterion I.

the magnitude. $\sum E_{ij}$ and Criterion I show similar contours, and the region of smaller criterion values aligns well with that of lower test error rates. Similar results are also seen from other data sets. These observations suggest that it is possible to locate the region of lower test error rates by minimizing Criterion I. This will be further verified later by more experimental study. On the other hand, one exceptional case is also found on the data set of "Vehicle." There, the region of lower criterion values does not align with the area of lower test error rates. Further investigation finds that the radius–margin bound on classes 1 and 2 is too loose to reasonably reflect the value of $(E'_{12} + E'_{21})$.

2) Relationship Between the Class Separability Measure and Criterion II: As mentioned before, there is some relationship between the scatter-matrix-based class separability measure and the radius–margin bound. Inspired by this, Criterion II is developed. Although this relationship is not directly related to the efficiency of this criterion, it is still demonstrated in this experiment for the sake of integrity. After this, the correspondence between $\sum E_{ij}$ and Criterion II will be shown. Fig. 3 shows the class separability and the radius–margin bound calculated by using the first two classes of the "Wine" data set. The results are shown in Fig. 3(a) and (b), respectively. Along the axis showing the natural logarithm of σ, the two surfaces reach lower values within their nearby locality. Finally, the correspondence between $\sum E_{ij}$ and Criterion II is shown in Fig. 4.

B. Experimental Result on the Benchmark Data Sets

The benchmark data sets in Table II are used in this experiment. They have different dimensionalities, unknown real distributions, and different sample sizes. Some of them have unbalanced classes, such as "Car," "E.coli," and "Yeast." These data sets form a good test bed for evaluating the two model selection criteria. In this experiment, the model selection result from the proposed criteria is compared with that using a five-fold cross-validation approach [16], [29], which is regarded as a benchmark here. In this approach, different pairs of $\{\sigma, C\}$ are evaluated via a 30×30 grid search on the top five training subsets. For each training subset, the pair giving rise to the minimal cross-validation error is selected, and five pairs of $\{\sigma, C\}$ are obtained in total. The median of the five σ values is selected as the optimal σ, and the median of the five C values is selected as the optimal C.

For Criteria I and II, the optimal σ and C are found by using the BFGS quasi-Newton optimization method. To reflect its computational load, the number of iterations and that of

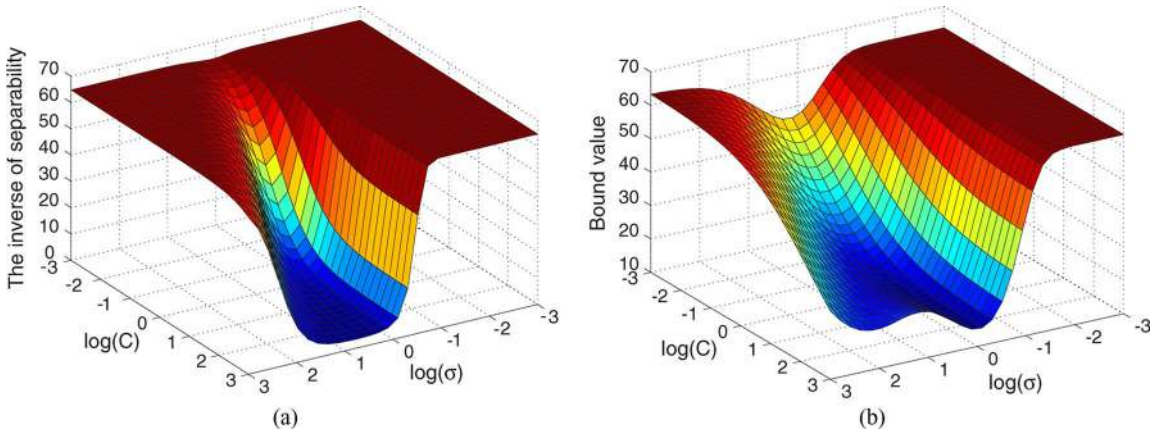

Fig. 3. Correlation between the class separability and the radius–margin bound. (a) Wine, inverse of separability. (b) Wine, bound value.

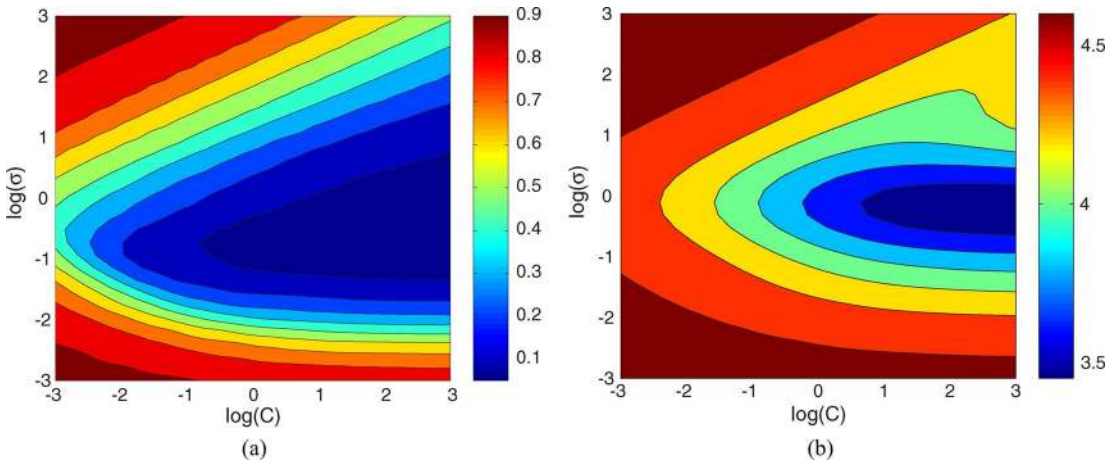

Fig. 4. Correspondence between $\sum E_{ij}$ and Criterion II. (a) Vowel, $\sum E_{ij}$. (b) Vowel, Criterion II.

TABLE III
SIX TESTED MULTICLASS SVM METHODS

Name in short	Classification Strategy	Final classification rule
1vsall	One-vs-all	Max wins
1vs1	One-vs-one	Max wins
dag	One-vs-one	Directed acyclic graph
dense	ECOC (dense encoding)	Exp. loss-based decoding
sparse	ECOC (sparse encoding)	Exp. loss-based decoding
single	One single optimization	Max wins

function evaluations are recorded. Afterward, a multiclass SVM classifier with the optimized model parameters is created and evaluated by using each of the 100 pairs of training and test subsets, and the average test error rate is obtained. Six different multiclass SVM methods listed in Table III are investigated. This is to evaluate the efficiency of the proposed model selection criteria for multiclass SVM methods using different strategies and classification rules.

Table IV lists the results for the methods using the one-versus-one and one-versus-all strategies. The spherical GRBF kernel is employed. The columns are separated into three groups, showing the results from the two criteria and the five-fold cross validation, respectively. In the first two groups, t and e denote the number of iterations and the number of function evaluations, respectively. The "1vs1" and "1vsall" stand for the average test error rates from the methods of one-versus-one and one-versus-all, respectively. The numbers in the brackets are the standard deviations, and the minimal test error rates are highlighted in bold.

By comparing the test error rates, it is seen that for both multiclass SVM methods, the proposed criteria give rise to the classification performance comparable to that obtained by using the five-fold cross validation. For Criterion I, it achieves the minimal test error rate on "Wine," "Zoo," and "DNA." On "Iris," "Car," "Dermatology," "E.coli," and "Yeast," the test error rates are similar to those from the cross validation. Criterion I produces slightly higher error rates on "Glass," "Vowel," "Segment," and "Satimage." As for Criterion II, it is comparable to the cross validation on "Iris," "Wine," "Glass," "Car," "Zoo," and "DNA." On "Dermatology," "E.coli," "Yeast," "Segment," and "Satimage," the test error rates from Criterion II are a bit higher. Comparison between the two criteria shows that Criterion I leads to slightly better overall classification performance. This can be seen from the data sets of "Dermatology," "E.coli," "Yeast," and "DNA." For both criteria, the test errors of the methods of "1vs1" and "1vsall" are comparable on most data sets; however, on "Vowel" and

TABLE IV
TEST ERRORS (SPHERICAL GRBF KERNEL, ONE-VERSUS-ONE, AND ONE-VERSUS-ALL)

Data set	Criterion I			Criterion II			5-fold CV	
	(t, e)	1vs1	1vsall	(t, e)	1vs1	1vsall	1vs1	1vsall
Iris	(6.4, 25.6)	4.09 (±1.71)	4.17 (±1.72)	(8.8, 29.6)	4.15 (±1.93)	4.17 (±1.83)	**4.33** (±**1.99**)	5.03 (±2.04)
Wine	(7.0, 31.6)	2.22 (±1.40)	**2.15** (±**1.44**)	(6.2, 25.6)	2.43 (±1.51)	2.33 (±1.51)	2.42 (±1.44)	2.40 (±1.49)
Glass	(7.8, 34.4)	35.82 (±4.89)	35.18 (±4.97)	(7.0, 30.0)	33.99 (±4.04)	33.94 (±3.89)	**33.23** (±**3.75**)	34.20 (±3.66)
Car	(6.2, 23.0)	12.28 (±0.67)	12.22 (±0.66)	(6.6, 29.8)	12.27 (±0.67)	12.29 (±0.67)	**11.93** (±**0.62**)	12.03 (±0.66)
Dermat.	(7.2, 27.6)	2.72 (±1.04)	2.48 (±1.10)	(10.0, 41.2)	3.23 (±1.23)	3.86 (±1.19)	2.55 (±1.02)	**2.42** (±**0.97**)
E.coli	(7.4, 34.8)	13.99 (±2.03)	13.55 (±1.84)	(6.8, 30.0)	14.44 (±1.92)	14.08 (±1.86)	13.99 (±1.93)	**13.18** (±**1.84**)
Zoo	(8.8, 33.8)	**6.61** (±**3.91**)	6.80 (±3.88)	(7.0, 31.2)	6.65 (±3.93)	6.80 (±3.94)	7.69 (±4.33)	7.75 (±4.42)
Yeast	(5.8, 24.2)	41.06 (±1.35)	40.44 (±1.43)	(7.6, 31.2)	43.11 (±1.34)	41.88 (±1.58)	40.23 (±1.33)	**40.14** (±**1.39**)
Vowel	(8.2, 34.8)	4.92 (±2.03)	6.23 (±2.20)	(7.8, 37.4)	**4.78** (±**2.01**)	6.34 (±2.21)	4.84 (±2.08)	5.77 (±2.18)
Vehicle$^{\triangle}$	(7.8, 35.2)	28.89 (±1.92)	27.32 (±1.98)	(6.6, 27.8)	32.08 (±2.20)	29.27 (±1.93)	16.62 (±1.88)	**16.44** (±**1.92**)
Segment	(5.8, 24.4)	4.84 (±0.57)	5.29 (±0.60)	(4.2, 17.4)	3.84 (±0.45)	4.69 (±0.56)	3.39 (±0.47)	**3.34** (±**0.49**)
DNA	(9.4, 39.8)	4.55	**4.22**	(5.4, 23.4)	4.81	4.72	4.55	4.30
Satimage	(4.6, 24.2)	10.15	10.10	(5.4, 32.0)	9.20	9.50	**7.90**	7.95

* The error rates of "DNA" and "Satimage" reported in Statlog are 4.1% and 9.4%, respectively.

"Segment," the test error rates for the "1vsall" method are a bit higher. On the other hand, a failure of the model selection is also observed on "Vehicle," where the obtained test error rates are significantly higher than that from the five-fold cross validation. This result can be expected from the discussion given at the end of Section VI-A1.

From the values of t and e, it is known that the minimization process is often accomplished in a few iterations with a small number of function evaluations. For example, on "Iris," when Criterion I is used, the model selection on a training subset completes in about seven iterations, including 26 function evaluations in total. Referring to Table I, this means that 156 $(26 \times 2 \times 3)$ QP problems are solved. However, for the five-fold cross-validation method, it has to solve 2700 $(30 \times 30 \times 3)$ QP problems when the one-versus-one strategy is used. Even if the grid number reduces to, for example, 10×10, by taking larger steps or by using a "coarse-to-fine" search, this number is still as high as 300.

The results for the classification methods of "dag," "dense," "sparse," and "single" are presented in Table V. The proposed criteria still provide good classification performance on most data sets except for "Vehicle." In detail, Criterion I obtains minimal test error rates on "Wine" and "Zoo," whereas the test error rates on "Glass," "Segment," and "Satimage" are a little higher. For Criterion II, the test error rates are generally comparable to those of the five-fold cross validation except that some increases are seen on "Dermatology," "E.coli," "Yeast," "Segment," and "Satimage." With respect to the test error rate obtained by the five-fold cross validation, the maximum increase caused by using Criterion I is 4.35% on "Satimage" and that caused by using Criterion II is 3.89% on "Yeast." Criterion I still shows marginally better classification performance than Criterion II. In addition, the test error rates listed in Tables IV and V are compared across the six classification methods. No strong evidence shows that Criteria I and II consistently perform better or worse on a particular method.

In short, it is observed from Tables IV and V that both criteria work well for the six classification methods. Although differences are observed between the test error rates obtained by the proposed criteria and those of the five-fold cross validation, they are not significant, particularly when the standard deviations are taken into account. Finally, to illustrate the details in a practical optimization process, the evolution of the values of Criterion I and the corresponding test error rate are shown in Fig. 5. As shown, the test error rate quickly drops with the decreasing criterion value.

The following part presents the experimental results when the ellipsoidal GRBF kernel is employed. Considering that different σ values are assigned to each dimension, the number of model parameters increases to the feature dimensionality

TABLE V
TEST ERRORS (SPHERICAL GRBF KERNEL, DAG, DENSE, SPARSE, AND SINGLE)

Data set	Criterion I				Criterion II				5-fold CV			
	dag	dense	sparse	single	dag	dense	sparse	single	dag	dense	sparse	single
Iris	4.31 (±1.87)	4.29 (±1.83)	4.33 (±1.78)	4.15 (±1.76)	4.03 (±1.92)	**4.09** (±**1.88**)	4.23 (±1.94)	4.49 (±2.10)	4.33 (±1.99)	5.03 (±2.04)	4.81 (±2.55)	4.55 (±1.91)
Wine	2.27 (±1.44)	2.18 (±1.43)	**2.13** (±**1.29**)	2.19 (±1.43)	2.43 (±1.50)	2.40 (±1.45)	2.27 (±1.33)	2.45 (±1.52)	2.42 (±1.44)	2.40 (±1.49)	2.29 (±1.39)	2.34 (±1.46)
Glass	35.24 (±4.57)	34.51 (±4.50)	34.83 (±4.52)	34.57 (±4.42)	34.10 (±3.94)	34.11 (±3.60)	33.60 (±3.56)	33.94 (±3.79)	**32.79** (±**4.38**)	33.22 (±3.76)	33.55 (±3.80)	33.43 (±3.98)
Car	12.27 (±0.67)	12.38 (±0.93)	12.38 (±0.93)	12.30 (±0.67)	12.27 (±0.67)	12.01 (±0.87)	12.01 (±0.87)	12.37 (±0.66)	11.91 (±0.71)	**11.68** (±**0.81**)	11.77 (±0.85)	11.99 (±0.68)
Dermat.	2.72 (±1.04)	2.52 (±1.07)	2.90 (±1.38)	2.50 (±1.00)	3.24 (±1.21)	3.94 (±1.20)	3.58 (±1.17)	3.30 (±1.03)	2.55 (±1.02)	**2.39** (±**0.98**)	2.66 (±1.17)	2.42 (±0.98)
E.coli	13.98 (±2.04)	13.57 (±1.80)	13.95 (±2.15)	13.29 (±1.93)	14.44 (±1.92)	14.07 (±1.90)	14.38 (±1.87)	13.61 (±1.91)	14.00 (±1.94)	**13.30** (±**1.99**)	13.53 (±2.01)	13.27 (±1.98)
Zoo	**6.61** (±**3.91**)	6.96 (±3.97)	6.78 (±4.05)	6.69 (±3.93)	6.67 (±3.92)	6.98 (±3.97)	6.92 (±4.17)	6.69 (±3.99)	7.65 (±4.30)	7.45 (±4.24)	6.94 (±3.94)	6.88 (±4.02)
Yeast	41.06 (±1.34)	40.53 (±1.42)	43.35 (±1.90)	40.45 (±1.36)	43.09 (±1.31)	42.51 (±1.45)	44.12 (±1.78)	40.27 (±1.32)	**40.23** (±**1.33**)	40.54 (±1.37)	42.56 (±1.86)	40.35 (±1.42)
Vowel	5.52 (±2.08)	5.70 (±2.15)	6.00 (±2.16)	6.31 (±2.23)	**4.78** (±**2.02**)	6.01 (±2.29)	5.98 (±2.34)	6.22 (±2.25)	5.02 (±1.85)	5.94 (±2.34)	5.94 (±2.35)	5.67 (±2.10)
Vehicle$^{\triangle}$	28.89 (±1.92)	28.84 (±2.08)	25.54 (±2.05)	28.84 (±2.08)	32.06 (±1.27)	32.82 (±2.08)	32.82 (±2.08)	24.27 (±1.95)	**16.72** (±**1.82**)	17.76 (±1.76)	17.38 (±1.66)	16.77 (±1.85)
Segment	4.66 (±0.55)	4.94 (±0.52)	5.33 (±0.70)	5.40 (±0.59)	4.70 (±0.62)	7.08 (±0.67)	6.36 (±0.51)	4.15 (±0.45)	3.41 (±0.49)	3.71 (±0.48)	**3.40** (±**0.52**)	3.69 (±0.52)
DNA	4.89	4.64	4.64	4.38	4.81	4.72	4.55	4.38	4.72	**4.22**	4.30	5.56
Satimage	10.75	11.00	12.15	10.00	9.95	10.50	11.25	9.40	8.10	7.95	**7.80**	8.10

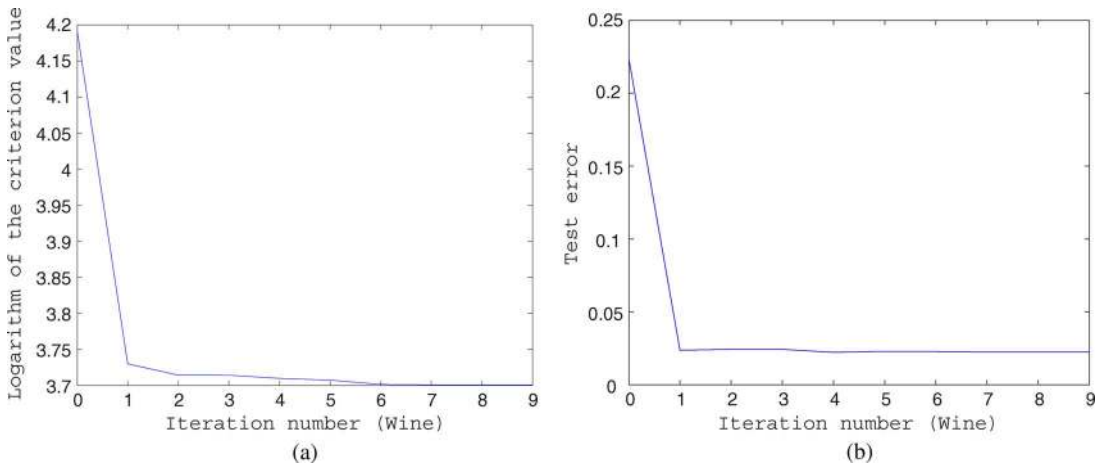

Fig. 5. Evolution of Criterion I and test error rate (Wine, spherical GRBF kernel). (a) Criterion value. (b) Test error rate.

plus one. In this case, even for the data represented by low-dimensional feature vectors, model selection with the exhaustive grid-based search methods becomes intractable.[5] Both proposed criteria can still work well. As shown in Table VI, the minimization process still finishes in a number of iterations, although the number becomes larger due to more model parameters to be optimized. Compared with the case using a spherical kernel, lower test error rates are observed on "Iris," "Wine," "Car," "Zoo," and "Segment." This may be the benefit of using an ellipsoidal kernel where feature components are

[5]Considering that the five-fold cross-validation method is intractable for the case of multiple model parameters, the test errors obtained by using the spherical kernel have to be used in Table VI.

TABLE VI
TEST ERRORS (ELLIPSOIDAL GRBF KERNEL, ONE-VERSUS-ONE, AND ONE-VERSUS-ALL)

Data set	Criterion I			Criterion II			5-fold CV	
	(t, e)	1vs1	1vsall	(t, e)	1vs1	1vsall	1vs1	1vsall
Iris	(11.0, 50.6)	**3.99** (±**1.71**)	4.01 (±1.70)	(11.68, 50.0)	4.07 (±2.10)	4.03 (±2.06)	4.33 (±1.99)	5.03 (±2.04)
Wine	(33.0, 147.2)	**1.53** (±**1.68**)	1.55 (±1.76)	(15.0, 68.8)	2.03 (±1.97)	1.90 (±1.95)	2.42 (±1.44)	2.40 (±1.49)
Glass$^{\triangle}$	(13.0, 61.6)	40.02 (±4.99)	39.45 (±4.89)	(11.40, 46.4)	40.97 (±5.54)	40.65 (±5.27)	**33.23** (±**3.75**)	34.20 (±3.66)
Car	(15.0, 66.2)	12.10 (±0.71)	12.17 (±0.70)	(14.6, 62.4)	12.06 (±0.71)	12.14 (±0.70)	**11.93** (±**0.62**)	12.03 (±0.66)
Dermat.	(19.0, 86.4)	2.44 (±0.98)	2.48 (±0.99)	(17.2, 79.2)	3.49 (±1.14)	4.15 (±1.17)	2.55 (±1.02)	**2.42** (±**0.97**)
E.coli	(15.0, 69.4)	14.10 (±2.10)	13.45 (±1.88)	(10.6, 47.8)	14.76 (±2.30)	14.30 (±2.23)	13.99 (±1.93)	**13.18** (±**1.84**)
Zoo	(5.2, 23.0)	**5.57** (±**3.78**)	5.57 (±3.78)	(3.0, 13.8)	5.69 (±3.98)	5.84 (±3.95)	7.69 (±4.33)	7.75 (±4.42)
Yeast	(15.0, 62.8)	40.87 (±1.31)	40.34 (±1.36)	(13.0, 59.2)	42.87 (±1.62)	41.88 (±1.58)	40.23 (±1.33)	**40.14** (±**1.39**)
Vowel	(32.0, 146.4)	5.16 (±1.88)	6.16 (±2.04)	(17.8, 83.2)	5.01 (±1.87)	5.74 (±2.15)	**4.84** (±**2.08**)	5.77 (±2.18)
Vehicle$^{\triangle}$	(17.0, 79.2)	27.54 (±1.82)	27.57 (±1.89)	(10.0, 45.6)	28.78 (±1.79)	28.91 (±1.97)	16.62 (±1.88)	**16.44** (±**1.92**)
Segment	(20.0, 88.2)	4.02 (±0.48)	4.06 (±0.49)	(10.2, 48.8)	3.43 (±0.42)	3.92 (±0.49)	3.39 (±0.47)	**3.34** (±**0.49**)
Dna	(11.0, 48.6)	4.72	4.47	(5.4, 23.4)	4.72	4.55	4.55	**4.30**
Satimage	(15.2, 70.8)	11.50	12.55	(5.4, 32.0)	10.95	11.40	**7.90**	7.95

combined in a weighted fashion. Similar test error rates are obtained on "E.coli," "Vowel," and "DNA," whereas higher error rates are seen on "Satimage." The model selection performance of Criterion I is still better than that of Criterion II in general.

On the other hand, a degraded performance is seen on "Glass," which is not as good as that obtained when a spherical kernel is used. Through analysis, we believe that this is because the number of training samples in some classes of "Glass" is so small (e.g., there are 2–6 training samples only) that the model selection process suffers from *overfitting*, i.e., when training samples are scarce, the minimization of the criteria may fit sample noise and fail to capture the real pattern there. In this case, although the minimum has been achieved, the selected model may not be good, and an SVM classifier using this model will not attain satisfactory classification performance on the test data. In addition, the model selection performance in this case often becomes sensitive to the optimization setting.[6] How to effectively avoid overfitting is also an active research area, and the regularization technique [31] seems to be a promising solution.

[6]With another optimization setting, we obtain a lower test error rate (33.96% ± 3.48%) on the "Glass" data set, which is comparable to that from the five-fold cross validation. However, the original result is reported for the sake of consistency of optimization settings for all the data sets.

Finally, Table VII presents the results from the "dag," "dense," "sparse," and "single" methods. Similarly, on the data sets such as "Iris," "Wine," "Car," and "Segment," some decreases on test error rates are observed, whereas on the other data sets such as "Dermatology," "Vowel," and "Satimage," the test error rates increase a bit. Generally speaking, for the case of using the ellipsoidal GRBF kernel, the proposed two criteria still demonstrate good performance for model selection in multiclass SVMs. There is no considerable scale performance degradation when the number of model parameters significantly increases, for example, to as high as 180 on "DNA."

Before the end of this part, the model selection time taken by the proposed criteria is compared with that taken by the five-fold cross-validation approach. The comparison is carried out on a Linux server with 3.0-GHz CPU and 1.0-GB memory. In this experiment, the proposed criteria are computed and optimized by using the code written in **Matlab**. It calls the Linux binaries in LIBSVM to calculate R^2 and $\|\mathbf{w}\|^2$, and then loads the results from the output files. This is not the most efficient implementation in terms of computational time (for example, less efficient than realizing all the steps with a single **C** program). However, as shown in Table VIII, the two criteria have been able to achieve faster model selection than the five-fold cross-validation approach realized by using LIBSVM in **C**. The model selection time for the ellipsoidal GRBF kernel is longer because more kernel parameters have to be optimized.

TABLE VII
TEST ERRORS (ELLIPSOIDAL GRBF KERNEL, DAG, DENSE, SPARSE, AND SINGLE)

Data set	Criterion I				Criterion II				5-fold CV			
	dag	dense	sparse	single	dag	dense	sparse	single	dag	dense	sparse	single
Iris	3.99 (±1.71)	4.01 (±1.70)	4.00 (±1.71)	4.01 (±1.70)	4.00 (±2.05)	4.03 (±2.06)	**3.95 (±2.08)**	4.27 (±2.19)	4.33 (±1.99)	5.03 (±2.04)	4.81 (±2.55)	4.55 (±1.91)
Wine	1.53 (±1.68)	1.55 (±1.76)	**1.36 (±1.49)**	1.56 (±1.56)	2.03 (±1.97)	1.90 (±1.95)	1.65 (±1.81)	1.56 (±1.56)	2.42 (±1.44)	2.40 (±1.49)	2.29 (±1.39)	2.34 (±1.46)
Glass$^{\triangle}$	40.03 (±4.99)	39.44 (±4.89)	41.18 (±5.36)	34.90 (±3.84)	40.97 (±5.57)	40.69 (±5.31)	44.03 (±6.06)	34.70 (±3.45)	**32.79 (±4.38)**	33.22 (±3.76)	33.55 (±3.80)	33.43 (±3.98)
Car	12.10 (±0.71)	11.70 (±0.72)	11.70 (±0.72)	12.17 (±0.71)	12.06 (±0.71)	**11.67 (±0.74)**	11.67 (±0.74)	12.14 (±0.70)	11.91 (±0.71)	11.68 (±0.81)	11.77 (±0.85)	11.99 (±0.68)
Dermat.	2.44 (±0.98)	2.53 (±0.99)	2.73 (±1.12)	2.89 (±1.05)	3.53 (±1.13)	4.32 (±1.22)	3.84 (±1.21)	4.27 (±1.25)	2.55 (±1.02)	**2.39 (±0.98)**	2.66 (±1.17)	2.42 (±0.98)
E.coli	14.09 (±2.11)	13.31 (±1.87)	13.80 (±1.99)	**13.05 (±1.93)**	14.74 (±2.31)	14.25 (±2.31)	14.97 (±2.59)	13.57 (±2.00)	14.00 (±1.94)	13.30 (±1.99)	13.53 (±2.01)	13.27 (±1.98)
Zoo	**5.57 (±3.78)**	5.57 (±3.78)	5.59 (±3.75)	8.10 (±3.83)	5.69 (±3.98)	5.78 (±3.92)	5.88 (±3.93)	8.16 (±4.58)	7.65 (±4.30)	7.45 (±4.24)	6.94 (±3.94)	6.88 (±4.02)
Yeast	40.85 (±1.31)	40.50 (±1.29)	44.06 (±2.24)	**39.73 (±1.38)**	42.87 (±1.63)	42.15 (±1.66)	46.49 (±2.30)	39.82 (±1.33)	40.23 (±1.33)	40.54 (±1.37)	42.56 (±1.86)	40.35 (±1.42)
Vowel	5.20 (±1.89)	6.74 (±2.10)	6.46 (±2.36)	6.08 (±2.05)	5.10 (±1.89)	6.47 (±2.19)	6.11 (±2.35)	5.92 (±2.21)	**5.02 (±1.85)**	5.94 (±2.34)	5.94 (±2.35)	5.67 (±2.10)
Vehicle$^{\triangle}$	27.52 (±1.82)	27.86 (±2.03)	27.86 (±2.03)	24.49 (±1.72)	28.77 (±1.80)	28.30 (±1.86)	28.30 (±1.86)	24.11 (±1.71)	**16.72 (±1.82)**	17.76 (±1.76)	17.38 (±1.66)	16.77 (±1.85)
Segment	4.00 (±0.48)	4.10 (±0.50)	4.26 (±0.53)	3.63 (±0.45)	3.44 (±0.42)	3.88 (±0.46)	3.86 (±0.55)	4.33 (±0.44)	3.41 (±0.45)	3.71 (±0.48)	**3.40 (±0.52)**	3.69 (±0.52)
DNA	4.72	4.47	4.38	4.30	4.72	4.55	4.47	4.64	4.72	**4.22**	4.30	5.56
Satimage	11.45	12.60	13.35	11.25	10.90	11.50	11.95	9.85	8.10	7.95	**7.80**	8.10

TABLE VIII
COMPARISON OF THE MODEL SELECTION TIME (UNIT: SECONDS)

Data set	Criterion I		Criterion II		5-fold CV
	Spatial GRBF	Ellipsoidal GRBF	Spatial GRBF	Spatial GRBF	Spatial GRBF
Iris	4.30	7.46	2.77	5.55	14.40
Wine	7.79	11.58	3.07	7.88	18.00
Glass	19.37	56.00	5.62	12.74	39.60
Car	37.83	46.35	32.07	20.79	367.20
Dermatology	19.45	64.35	6.48	44.56	82.80
E.coli	30.56	32.08	8.66	7.94	57.60
Yeast	170.31	225.35	97.40	74.67	918.00
Zoo	25.44	36.77	3.59	8.82	28.80
Vowel	66.75	101.37	14.96	17.11	151.20
Vehicle	34.89	107.47	26.41	95.71	248.40
Segment	88.11	198.72	44.74	66.49	1386.00
Dna	322.12	1219.5	300.62	1853.25	5346.00
Satimage	403.14	1020.60	232.29	962.87	4932.00

* In this experiment, the proposed criteria are computed and optimized by using the mixture codes of Matlab and C. Faster model selection can be expected when entirely implemented in C.

This is particularly true for the data sets with high-dimensional input spaces. In addition, Criterion II generally costs less model selection time than Criterion I. They use different ways to evaluate the radius R (solving a single larger quadratic problem versus solving multiple smaller ones). However, it is believed that in practical applications, determining which criterion is faster will depend on the code design. The evolution of Criterion I and the corresponding test error rate is also shown in

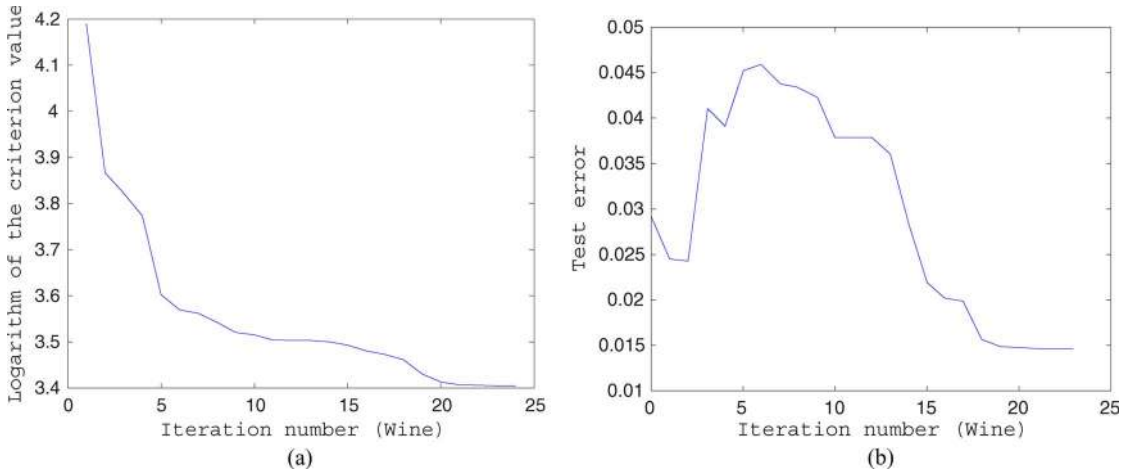

Fig. 6. Evolution of Criterion I and test error rate (Wine, ellipsoidal GRBF kernel). (a) Criterion value. (b) Test error rate.

Fig. 6. At the initial stage, the criterion value decreases, but the test error rate increases. We believe that this is because the criterion is not a tight estimation of the test error rate. When the criterion value is relatively large, its decrease may not lead to an immediate reduction on the test error rate. However, as shown in Fig. 6, when the criterion value converges to its minimum, a lower enough test error rate will be achieved.

C. Application to Feature Selection

The optimized model parameters can be used to identify features important for classification.[7] For example, when the ellipsoidal GRBF kernel is used, the value of $g_i (g_i = 1/(2\sigma_i^2), i = 1, 2, \ldots, d)$ can reflect the importance of the ith feature, and the larger the g_i value, the more important this feature is. This has been observed from the binary classification where the radius–margin bound is applicable [16]. This experiment will demonstrate that the proposed criteria well preserve this property in multiclass classification. Two data sets are used. One is a toy problem, and the other is the U.S. Postal Service (USPS) data set on optical digit recognition.

1) Toy Data Set: This data set is created by following [16]. However, in this experiment it consists of multiple classes. There are 52 features in total, and only the first two of them are useful. The two features are shown in Fig. 7(a). There are three concentric circles, forming a three-class classification problem. The remaining 50 features are randomly sampled from a Gaussian distribution of $\mathcal{N}(0, 20)$. Three hundred samples are generated in total. This experiment is to check whether the first two features can be identified by using the proposed criteria, i.e., being assigned larger g values. The toy data set is randomly split into 100 pairs of training/test subsets, and Criterion II is applied to each of the training subsets. Considering that the three classes are completely nonlinearly separable, the initial value of the regularization parameter C is set as a bit higher value, e.g., 10.0. A promising result is obtained. The first two features are correctly assigned higher g values on all the 100 trials. The g values averaged over the 100 trials are shown in Fig. 7(b). As shown, the first two features can be easily identified by sorting the g values. Once they are correctly selected, the three circles can be classified without error. It is assumed here that "only two features are really useful" is known beforehand. In a general case, the top k features will be selected, and some noisy features will be brought in if k is larger than the number of useful features, for example, two for this data set.

2) Optical Digit Recognition: The USPS data set contains 7291 training samples and 2007 test samples. They form ten classes corresponding to digits from "0" to "9." Each sample is characterized by a 256-D feature vector. It is obtained by reshaping a 16×16 gray-level thumbnail image. Some examples are shown in Fig. 8.

With the training data, the proposed criteria are minimized to find the optimal model parameters. Afterward, the multiclass SVMs with the optimized model parameters are created and evaluated. The test error rates are listed in Table IX. The value in the column of "Reported best" is the lowest test error rate given in [32] when a spherical GRBF kernel is used (that for an ellipsoidal GRBF kernel is left blank because the five-fold cross-validation approach is intractable in this case). As seen from this table, the multiclass SVMs with the optimized model parameters give rise to the classification performance comparable to the reported best. When the ellipsoidal kernel is used, slightly increased test error rates are observed, but the difference between our results and the reported best is still less than 1%.

By reshaping the optimal value of $g_1, g_2, \ldots, g_{256}$ back into a 16×16 matrix, the map of g values is shown in Fig. 9, where each block corresponds to one of the 256 g's and its magnitude is reflected by the gray value. The blocks having larger g values distribute at the central part of this map, whereas those having lower values are mostly at the borders and corners. This result implies that the pixels at the central part are more important for classification. This result well matches the case of the

[7]Please note that the feature selection here is different from feature extraction that considers the feature dependence and seeks the optimal combination of features, for example, in the way of principal component analysis or linear discriminant analysis. Here, features are treated individually, and those important for classification are identified. As for the feature dependence, it is left to the SVM classifier that can handle it automatically.

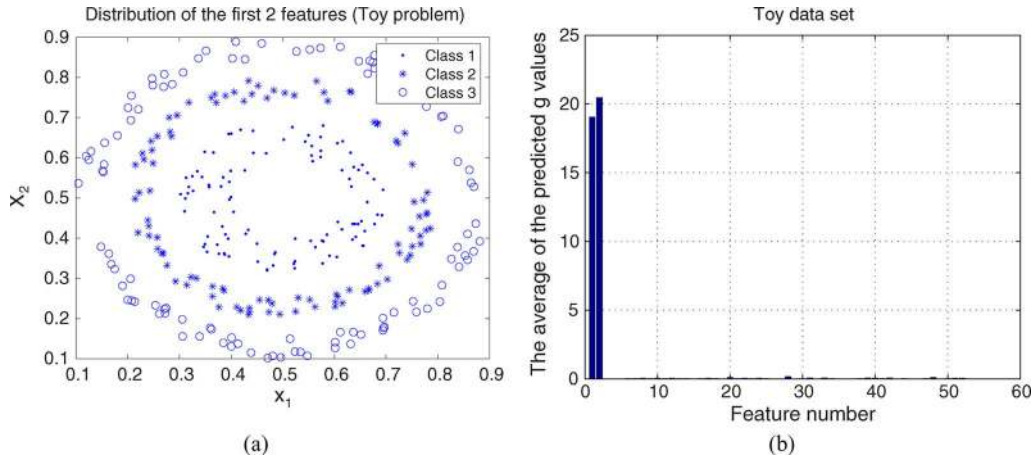

Fig. 7. Results for the toy data set. (a) Distribution of the first two features. (b) Optimized value of $g = 1/(2\sigma^2)$ averaged on 100 trails.

Fig. 8. Thumbnail images of digits (USPS data set).

TABLE IX
TEST ERRORS OBTAINED WITH THE SELECTED MODEL PARAMETERS (THE USPS DATA SET)

Kernel type	Criterion I			Criterion II			Reported best [32]
	1vs1	1vsall	dag	1vs1	1vsall	dag	
Spherical GRBF	4.73	4.58	4.63	4.68	4.48	4.63	**4.30**
Ellipsoidal GRBF	4.93	4.68	4.88	5.08	5.28	4.93	—

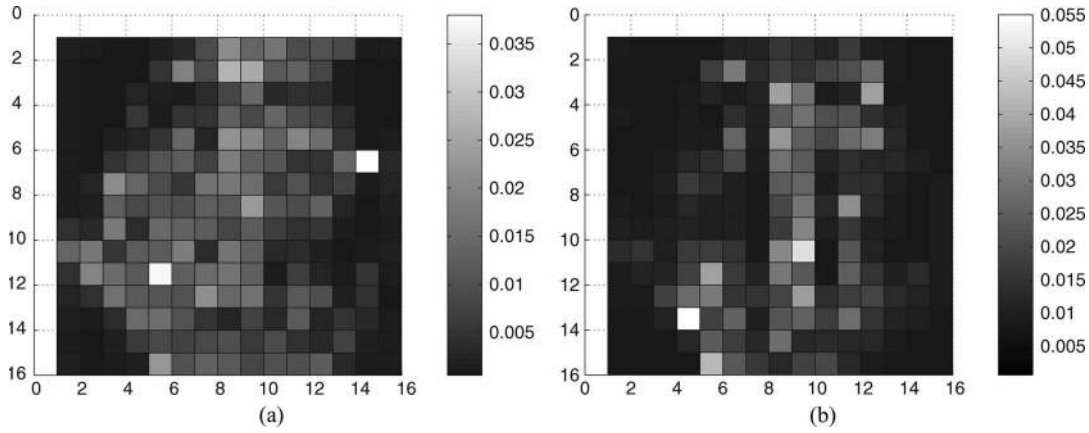

Fig. 9. Map of optimized values of g (USPS data set). (a) Criterion I. (b) Criterion II.

thumbnail images in Fig. 8 that a digit is commonly displayed at the central part of an image.[8] To investigate the performance of feature selection using these g values, the following experiments are conducted. The 256 features are sorted according to a descending order of the corresponding g values, and only the top k features are used to train a multiclass SVM to perform classification. For comparison, another three feature selection methods named "Fisher criterion score," "Pearson correlation coefficient," and "Kolmogorov–Smirnov test" in [16] are also used and adapted to the multiclass case. The following two points will be checked: 1) whether the test error rate rapidly decreases with the increasing value of k, and 2) whether the proposed criteria can produce feature selection performance comparable to the other three methods. The result is shown in Fig. 10, where the horizontal axis is the number of selected features and the vertical one is the test error rate. It is seen that the test error rate drops quickly with the increasing number of selected features. Compared with the other three methods,

[8]This experiment was first presented in [16], where a binary classification problem of discriminating two groups of digits (group I, "0"–"4"; group II: "5"–"9") is considered and the radius–margin bound is used. This paper develops two model selection criteria and makes such a feature selection applicable to the multiclass classification that discriminates each digit from each other.

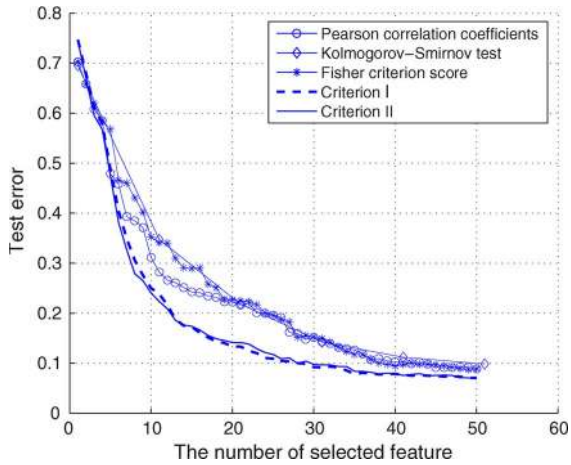

Fig. 10. Feature selection result. Test error rate versus number of selected features (USPS data set).

both criteria give better feature selection performance. By using only the top 50 selected features, the test error rate of 0.07 has been achieved (the lowest error rate is about 0.05 when all the 256 features are used). Many redundant features are recognized by using the proposed criteria. The curves in Fig. 10 show the change of test error rate with respect to the number of selected features. In practical applications, a point on the curve can be selected to balance between classification accuracy and the number of features used in an SVM classifier.

D. Summary

Based on the aforementioned experimental results, the following summary can be made.

1) Both criteria demonstrate good model selection performance on most of the data sets, giving rise to classification accuracy comparable to that obtained by using the five-fold cross-validation approach.
2) The two criteria work well with both spherical and ellipsoidal GRBF kernels, and the latter verifies their ability in handling a large number of model parameters. At the same time, it is also observed that satisfactory model selection may not be attained if training samples are scarce. How to solve this problem is still an ongoing research, and it is worth exploring in future work.
3) Six different kinds of multiclass SVM methods are investigated. Although the proposed criteria are developed on the multiclass SVM using the one-versus-one strategy, they help all the six multiclass SVM classifiers achieve good enough classification performance. It seems that the two criteria are promising to be generally used for model selection of multiclass SVM methods.
4) Model parameters optimized by the two criteria are used to do feature selection. Compared with the existing selection criteria, they achieve comparable or even better selection results. We believe that this property has wide applications in real-world problems and that it is worth further investigation.
5) The comparison of the two criteria finds that Criterion I shows marginally better performance. Meanwhile, there is a difference between their computational loads. To compute the radius R^2, Criterion I solves multiple smaller scale QP problems, whereas Criterion II solves one larger scale QP problem. When the number of classes is large, model selection with the second criterion might be faster.

VII. Conclusion

This paper has proposed two criteria to perform model selection in multiclass SVMs. They are realized by combining or redefining the radius–margin bound of binary classification to accommodate multiple classes. Both criteria are not the radius–margin bound generalized for multiclass SVMs. Nevertheless, they are simple and practical, and most importantly, they demonstrate satisfactory performance in the task of model selection for which they are proposed. These two criteria well preserve the elegant properties of the radius–margin bound in the model selection of binary SVMs. Their derivatives with respect to model parameters are analytically calculated, and thus, the gradient-based optimization technique is used to find the best model efficiently. The two criteria handle hundreds of model parameters well and save much computational cost than the grid-based search methods. In addition, they do not need to put a part of training samples aside for validation and make full use of all the training samples available. In the application to feature selection, the optimized model parameters successfully identify features important for classification. This is very helpful for the reduction of system complexity and feature discovery. Extensive experimental study on multiple benchmark data sets and different multiclass SVM methods verify the effectiveness of the proposed criteria and their applicability for multiclass SVM model selection.

References

[1] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004.
[2] J. H. Friedman, "Another approach to polychotomous classification," Dept. Stat. Stanford Linear Accelerator Center, Stanford Univ., Stanford, CA, 1996. Tech. Rep.
[3] U. Kressel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999.
[4] T. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. Artif. Intell. Res.*, vol. 2, pp. 263–286, 1995.
[5] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, 2000.
[6] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *Proc. 7th Eur. Symp. Artif. Neural Netw.*, 1999, pp. 219–224.
[7] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," in *Proc. 13th Annu. Conf. Comput. Learn. Theory*, 2000, pp. 35–46.
[8] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines," in *Proc. 33rd Symp. Interface Comput. Sci. Statist.*, 2001, pp. 498–512.
[9] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
[10] W.-C. Kao, K.-M. Chung, C.-L. Sun, and C.-J. Lin, "Decomposition methods for linear support vector machines," *Neural Comput.*, vol. 16, no. 8, pp. 1689–1704, Aug. 2004.
[11] P.-H. Chen, C.-J. Lin, and B. Schölkopf, "A tutorial on ν-support vector machines," *Appl. Stoch. Models Bus. Ind.*, vol. 21, no. 2, pp. 111–136, 2005.

[12] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data," *J. Amer. Stat. Assoc.*, vol. 99, no. 465, pp. 67–81, Mar. 2004.
[13] A. Passerini, M. Pontil, and P. Frasconi, "New results on error correcting output codes of kernel machines," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 45–54, Jan. 2004.
[14] P. Xu and A. K. Chan, "Support vector machines for multi-class signal classification with unbalanced samples," in *Proc. IJCNN*, 2003, pp. 1116–1119.
[15] M. Liepert, "Topological fields chunking for German with SVM's: Optimizing SVM-parameters with GA's," in *Proc. Int. Conf. RANLP*, Borovets, Bulgaria, 2003.
[16] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 131–159, 2002.
[17] K. Duan, S. S. Keerthi, and A. N. Poo, "Evaluation of simple performance measures for tuning SVM hyperparameters," Nat. Univ. Singapore, Singapore, Tech. Rep. CD-01-11, 2001. [Online]. Available: http://guppy.mpe.nus.edu.sg/~mpessk/svm.shtml
[18] M. Seeger, "Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers," in *Proc. NIPS*, 2000, vol. 12, pp. 603–609.
[19] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, Jul. 2003.
[20] N. Cristianini, C. Campbell, and J. Shawe-Taylor, "Dynamically adapting kernels in support vector machines," in *Proc. Advances Neural Inf. Process. Syst.*, 1999, vol. 11, pp. 204–210.
[21] S. Keerthi, "Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1225–1229, Sep. 2002.
[22] Y. Darcy and Y. Guermeur, "Radius–margin bound on the leave-one-out error of multi-class SVMs," INRIA, Rocquencourt, France, Tech. Rep., No. 5780, Dec. 2005.
[23] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Base Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[24] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. New York: Academic, 1999.
[25] R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern Classification*, 2nd ed. Hoboken, NJ: Wiley, 2001.
[26] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.
[27] L. Wang and K. L. Chan, "Learning kernel parameters by using class separability measure," in *Proc. 6th Kernel Mach. Workshop Learning Kernels, NIPS*, 2002.
[28] D. F. Shanno and K. H. Phua, "Remark on "Algorithm 500: Minimization of unconstrained multivariate functions [E4]"," *ACM Trans. Math. Softw.*, vol. 6, no. 4, pp. 618–622, Dec. 1980.
[29] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for AdaBoost," *Mach. Learn.*, vol. 42, no. 3, pp. 287–320, Mar. 2001.
[30] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[31] Z. Chen and S. Haykin, "On different facets of regularization theory," *Neural Comput.*, vol. 14, no. 12, pp. 2791–2846, Dec. 2002.
[32] B. Schölkopf and A. Smola, *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.

Lei Wang (M'07) received the B.Eng. and M.Eng. degrees from Southeast University, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore, in 2004.

He was a Research Associate and Research Fellow with Nanyang Technological University, from 2003 to 2005. He is currently a Research Fellow with the Department of Information Engineering, Research School of Information Sciences and Engineering, The Australian National University, Canberra, A.C.T., Australia. His research interests include computer vision, pattern recognition, information retrieval, and machine learning.

Dr. Wang was awarded the Australian Postdoctoral Fellowship by the Australian Research Council in 2007.

Ping Xue (M'91–SM'03) received the B.S. degree in electronic engineering from the Harbin Institute of Technology, Harbin, China, in 1968, and the M.S.E., M.A., and Ph.D. degrees in electrical engineering and computer science from Princeton University, Princeton, NJ, in 1981, 1982, and 1985, respectively.

He was a Member of the Technical Staff in the David Sarnoff Research Center, Princeton University, from 1984 to 1986 and of the faculty of Shanghai Jiao Tong University, Shanghai, China, from 1986 to 1990. He was with Chartered Semiconductor, Singapore, from 1991 to 1994 and the Institute of Microelectronics from 1994 to 2001. Since 2001, he has been with Nanyang Technological University, Singapore, Singapore, where he was an Associate Professor and where he is currently with the School of Electrical and Electronic Engineering. His research interests include multimedia signal processing, content/perceptual based analysis for video indexing and retrieval, and applications in communication networks.

Kap Luk Chan (S'88–M'90) received the Ph.D. degree in robot vision from the Imperial College of Science, Technology, and Medicine, University of London, London, U.K., in 1991.

He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests are in image analysis and computer vision, particularly in texture analysis, statistical image analysis, perceptual grouping, image and video retrieval, application of machine learning in computer vision, computer vision for human–computer interaction, and biomedical signal and image analysis.

Dr. Chan is a member of The Institution of Engineering and Technology.