Feature Selection with Kernel Class Separability

Lei Wang, *Member*, *IEEE*

Abstract—Classification can often benefit from efficient feature selection. However, the presence of linearly nonseparable data, quick response requirement, small sample problem, and noisy features makes the feature selection quite challenging. In this work, a class separability criterion is developed in a high-dimensional kernel space, and feature selection is performed by the maximization of this criterion. To make this feature selection approach work, the issues of automatic kernel parameter tuning, numerical stability, and regularization for multiparameter optimization are addressed. Theoretical analysis uncovers the relationship of this criterion to the radius-margin bound of the Support Vector Machines (SVMs), the Kernel Fisher Discriminant Analysis (KFDA), and the kernel alignment criterion, thus providing more insight into feature selection with this criterion. This criterion is applied to a variety of selection modes using different search strategies. Extensive experimental study demonstrates its efficiency in delivering fast and robust feature selection.

Index Terms—Kernel class separability, feature selection, Support Vector Machines, Kernel Fisher Discriminant Analysis, pattern classification.

1 Introduction

In many classification tasks, numerous features can be extracted, but only a small number of them are really discriminative. In this case, feature selection becomes critical and leads to many benefits. It reduces the dimensions of a feature space, thus giving rise to more reliable parameter estimation, lower system complexity, and less storage requirement. Removing noisy and irrelevant features can improve the performance of the classifier. In addition, time, labor, and expense can be well saved by only extracting useful features. Nevertheless, the following problems make the feature selection quite challenging:

- *Linearly nonseparable classes.*[1] This case generally exists in real-world applications. In image classification, the semantic gap between the high-level concepts (used by humans to interpret visual content) and the low-level features (used by computers for classification) often results in a nonlinear mapping between them.
- *Quick response.* This is always desirable, especially when real-time processing is needed. A computationally efficient criterion will be preferred. Furthermore, such a criterion will lend itself to being combined with a more sophisticated (and often more computationally expensive) selection strategy to further improve the selection performance.
- *Selection with a small number of samples.* When extracting a large number of features from every sample is expensive or time consuming (for example, intensive laboratory tests are needed), it will be good to first identify the useful ones via a small-sized sample set. By doing so, future feature extraction can focus on only the useful features. In this case, a selection criterion that is less sensitive to sampling process is needed.
- *Noisy features.* These are features that are statistically irrelevant to class labels or are heavily corrupted by noise during data generation. To deal with noisy features, a more robust selection criterion is required.

In this paper, a kernel-based feature selection criterion is proposed. Compared with the existing criteria, it can achieve overall better performance in the presence of the above problems.

Introduced with the Support Vector Machines (SVMs), the *kernel* trick [1], [2] has attracted much attention because of its efficient and elegant way of modeling nonlinear patterns. Via a kernel function, data can be nonlinearly mapped to a high-dimensional kernel space.[2] As stated by the Cover theorem, the data will be more likely linearly separable when they are nonlinearly mapped to a higher dimensional space [3]. Moreover, this mapping is implicitly and efficiently realized through a kernel function. Aside from the SVMs, the Kernel Principal Component Analysis (KPCA) [4] and the Kernel Fisher Discriminant Analysis (KFDA) [5] are also commonly used kernel-based algorithms.

1. More precisely, the "linearly nonseparable classes" means that the boundary that optimally separates the classes from each other is not a hyperplane.

2. This space is often called the "feature space." For convenience of notation, it is called "kernel space" in this work.

• *The author is with the Research School of Information Sciences and Engineering, The Australian National University, RSISE, Building 115, Canberra, ACT, Australia, 0200.*
E-mail: Lei.Wang@mail.rsise.anu.edu.au.

Manuscript received 4 Oct. 2006; revised 13 June 2007; accepted 25 Sept. 2007; published online 15 July 2008.
Recommended for acceptance by Z. Ghahramani.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0701-1006.
Digital Object Identifier no. 10.1109/TPAMI.2007.70799.

In this paper, the kernel trick is incorporated into a class separability measure. Class separability is a classic concept in pattern recognition [6], [7], [8]. A widely used separability measure is based on the scatter matrices of data. In this paper, the scatter-matrix-based class separability measure is extended to a kernel space and developed as a feature selection criterion. This is based on the idea that *the features that lead to larger class separability are more important*. In a high-dimensional kernel space, the scatter matrices are often singular, and their determinants become zero. To measure the class separability in such a case, a trace-based kernel class separability criterion is derived in this paper. However, straightforwardly applying this criterion to feature selection will be problematic, because the value of a kernel-based criterion depends on the parameters of the kernel function. A poor setting of these parameters can easily remove the difference between good and bad features, which makes the feature selection fail. In this paper, the kernel parameters are treated as variables and are automatically tuned by the maximization of the kernel class separability criterion. This not only avoids the adverse affect of manual parameter setting but also improves the feature selection efficiency. To ensure the numerical stability in the process of kernel parameter optimization, a lower bound of the proposed criterion is further derived by assuming that a stationary or normalized kernel is used. In addition, in the case of small sample set but a large number of features, simply maximizing the proposed criterion may result in overfitting and degrade the feature selection performance. To address this problem, a regularization strategy is proposed in this paper.

This kernel class separability criterion is successfully applied to a variety of selection modes, including the simplest Best individual N (BIN),[3] the sequential forward selection (SEQ), and the state-of-the-art kernel parameter optimization (KPO) approach. The radius-margin bound, which is an upper bound of the leave-one-out cross-validation error of the SVMs, has demonstrated excellent performance as a feature selection criterion [10], [11]. In this paper, our theoretical analysis uncovers the intrinsic relationship between the kernel class separability criterion and the radius-margin bound. Compared with the radius-margin bound, the proposed criterion has the following advantages. Feature selection with this criterion is faster, because it is conceptually and computationally simpler. The proposed criterion is more robust in the case of small sample set and is less vulnerable to noisy features. This is because it is based on the information averaged over all of the data. Aside from these, the relationship between the proposed criterion and the KFDA algorithm is clarified. This criterion is proven to be a lower bound of the maximum value of the KFDA's objective function. It is not a reinvention of the KFDA algorithm.

Our extensive experimental study is conducted on synthetic and real benchmark data sets to compare the proposed criterion with the existing methods, particularly the radius-margin bound. The result demonstrates the efficiency of the proposed criterion in achieving fast and robust feature selection. The rest of this paper is organized as follows: In Section 2, existing feature selection criteria are reviewed. Section 3 develops the class separability measure in a kernel space. It is then tailored for feature selection and applied to three different feature selection modes. Section 4 discusses its relationship to the radius-margin bound, the KFDA algorithm, and the kernel alignment (KA) criterion, as well as its advantages. Finally, experimental results and concluding remarks are given in Sections 5 and 6.

3. This term is taken from [9]. It means that a selection criterion is individually applied to each of the features to evaluate its goodness. The features that give the larger criterion values will be selected.

2 Feature Selection

Feature selection, more precisely *feature subset selection*, aims at finding p features out of the original d ones according to a selection criterion. Note that it is different from *feature extraction* (*or feature combination*), where a d-dimensional feature vector is projected to a p-dimensional subspace, for example, the case in the PCA. For a classification-oriented feature selection, the p selected features are expected to produce low classification errors when they are used by a classifier.[4] Feature selection often consists of a selection criterion and a search strategy (or the selection mode in this paper). An efficient search strategy is critical, since feature selection is essentially a combinatorial optimization problem. Many search algorithms have been developed in the literature, for example, the branch-and-bound procedure, the sequential forward/backward selection, and the floating search methods [6], [9]. In this work, the selection criterion is focused. Some widely used selection criteria are reviewed as follows, and they will be compared with the criterion proposed by this work:

- *Pearson correlation coefficient.* By treating a feature and the class label as two random variables, this method evaluates the strength of relevance between them. High relevance is used to identify good features. This method is simple and efficient when the two variables are linearly correlated. Nevertheless, it cannot handle linearly nonseparable data.
- *Kolmogorov-Smirnov test.* For a given feature, this test evaluates whether the samples in two classes are actually generated by the same underlying distribution. The less the possibility (or the higher the test value), the better this feature for discrimination. This test is applicable to linearly nonseparable data, but it needs a sufficient number of samples to estimate the distribution.
- *Class separability (nonkernel).* For a given feature subset, the scatter-matrix-based class separability measure evaluates the ratio of the between-class scattering to the within-class scattering of data. A subset that gives rise to high class separability is regarded as a good one [9]. This criterion is simple, robust, and unified for both binary and multiclass classification. Nevertheless, it cannot handle linearly nonseparable data. In this paper, we will extend this measure to a kernel space and make it an efficient feature selection criterion.
- *Radius-margin bound.* This is an upper bound of the leave-one-out cross-validation error of the SVMs. In

4. Note that feature selection cannot completely be separated from the classifier. For a linearly nonseparable problem, a wise user will employ a nonlinear classifier, rather than a linear one, to obtain the result most agreeable to human perception. It is believed that this assumption can be generally satisfied.

[10] and [11], it is minimized to optimize the kernel parameter assigned to each feature. The larger the parameter value, the more important the corresponding feature. This bound is theoretically elegant and well handles linearly nonseparable data. However, it is not computationally efficient. In addition, it is sensitive to small sample sets and may fail when too many noisy features exist.

3 Proposed Feature Selection Criterion

3.1 Class Separability

Class separability includes divergence, the Bhattacharyya distance, and scatter-matrix-based measures [6], [8], [9]. The scatter-matrix-based measure is often favored because of its simplicity. It includes the *Within-class scatter matrix* $\mathbf{S}_W$, *Between-class scatter matrix* $\mathbf{S}_B$, and *Total scatter matrix* $\mathbf{S}_T$. Let $(\mathbf{x}, y) \in (\mathbb{R}^d \times \mathcal{Y})$ denote a sample, where $\mathbb{R}^d$ stands for a d-dimensional feature space, $\mathcal{Y}$ is the set of class labels, and the size of $\mathcal{Y}$ is the number of classes c. The number of samples in the ith class is denoted by n_i. Let $\mathbf{m}_i$ be the mean vector for the ith class and $\mathbf{m}$ be the mean vector for all classes. The scatter matrices are defined as

$$\begin{aligned}\mathbf{S}_W &= \sum_{i=1}^{c}\Big[\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\mathbf{m}_i)(\mathbf{x}_{ij}-\mathbf{m}_i)^\top\Big],\\ \mathbf{S}_B &= \sum_{i=1}^{c} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^\top,\\ \mathbf{S}_T &= \sum_{i=1}^{c}\Big[\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\mathbf{m})(\mathbf{x}_{ij}-\mathbf{m})^\top\Big]\\ &= \mathbf{S}_W+\mathbf{S}_B.\end{aligned} \tag{1}$$

A large class separability means small within-class scattering but large between-class scattering. The commonly used measures 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 measures can also be found in [6]. In these measures, the scattering of data is evaluated via the mean and variance. This implicitly assumes a Gaussian distribution for each class. The resulting drawback is that these measures cannot correctly evaluate the class separability when the data presents a non-Gaussian structure such as two classes distributed along two concentric circles. This will be remedied by incorporating the kernel trick.

3.2 Kernel-Based Class Separability

A kernel function is an inner product in a kernel space. It is written 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 the feature space $\mathbb{R}^d$ to a kernel space $\mathcal{K}$. $\boldsymbol{\theta}$ denotes the kernel parameter set. A kernel function plays a central role in a kernel-based algorithm. Geometrically, it defines a distance metric in $\mathcal{K}$, because it can be shown that $\|\phi(\mathbf{x}_i)-\phi(\mathbf{x}_j)\|^2 = k_{ii}-2k_{ij}+k_{jj}$, where k_{ij} denotes $k(\mathbf{x}_i,\mathbf{x}_j)$. In other words, it implicitly determines the scattering of data in the kernel space $\mathcal{K}$.

Let us develop the class separability measure in $\mathcal{K}$. Since only $\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle$ is accessible via a kernel function, none of the scatter matrices in (1) can be explicitly computed in $\mathcal{K}$. Moreover, the high dimensionality of $\mathcal{K}$ often makes the scatter matrices singular and their determinants zero, leaving the determinant-based measure invalid. Hence, the trace-based measure is used in this work. The superscript ϕ distinguishes the variables in $\mathcal{K}$ from those in $\mathbb{R}^d$. Let $\mathcal{D}_i$ denote the set of samples from the ith class. In addition, it is defined that $\mathcal{D}=\cup_{i=1}^{c}\mathcal{D}_i$. The sizes of $\mathcal{D}_i$ and $\mathcal{D}$ are n_i and n, respectively. $\mathbf{K}$ denotes a kernel matrix with $\{\mathbf{K}\}_{ij}=k_{\boldsymbol{\theta}}(\mathbf{x}_i,\mathbf{x}_j)$. $\mathbf{K}_{\mathcal{A},\mathcal{B}}$ is a kernel matrix with the constraints of $\mathbf{x}_i\in\mathcal{A}$ and $\mathbf{x}_j\in\mathcal{B}$. The operator $\mathrm{Sum}(\cdot)$ denotes the summation of all elements in a matrix. The traces are derived as

$$\begin{aligned}&\mathrm{tr}(\mathbf{S}_B^\phi)\\ &=\mathrm{tr}\Big[\sum_{i=1}^{c} n_i(\mathbf{m}_i^\phi-\mathbf{m}^\phi)(\mathbf{m}_i^\phi-\mathbf{m}^\phi)^\top\Big]\\ &=\sum_{i=1}^{c} n_i\Big[(\mathbf{m}_i^\phi-\mathbf{m}^\phi)^\top(\mathbf{m}_i^\phi-\mathbf{m}^\phi)\Big]\\ &=\sum_{i=1}^{c}\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_i,\mathcal{D}_i})}{n_i}-\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n},\end{aligned} \tag{2}$$

$$\begin{aligned}&\mathrm{tr}(\mathbf{S}_W^\phi)\\ &=\mathrm{tr}\Big[\sum_{i=1}^{c}\sum_{j=1}^{n_i}(\phi(\mathbf{x}_{ij})-\mathbf{m}_i^\phi)(\phi(\mathbf{x}_{ij})-\mathbf{m}_i^\phi)^\top\Big]\\ &=\sum_{i=1}^{c}\sum_{j=1}^{n_i}\Big[(\phi(\mathbf{x}_{ij})-\mathbf{m}_i^\phi)^\top(\phi(\mathbf{x}_{ij})-\mathbf{m}_i^\phi)\Big]\\ &=\mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}})-\sum_{i=1}^{c}\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_i,\mathcal{D}_i})}{n_i},\end{aligned} \tag{3}$$

and

$$\begin{aligned}\mathrm{tr}(\mathbf{S}_T^\phi)&=\mathrm{tr}(\mathbf{S}_B^\phi)+\mathrm{tr}(\mathbf{S}_W^\phi)\\ &=\mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}})-\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n}.\end{aligned} \tag{4}$$

The class separability in the kernel space can be measured as

$$\mathcal{J}^\phi=\frac{\mathrm{tr}(\mathbf{S}_B^\phi)}{\mathrm{tr}(\mathbf{S}_W^\phi)}. \tag{5}$$

This criterion is still conceptually simple and computationally light, although the kernel trick has been incorporated. The main computation is to calculate $\mathbf{K}$ only. To maintain the numerical stability in the maximization of $\mathcal{J}^\phi$, the denominator $\mathrm{tr}(\mathbf{S}_W^\phi)$ has to be prevented from approaching zero. This may be realized by employing a modified kernel matrix like $\mathbf{K}'=\mathbf{K}+\mu\mathbf{I}$, where μ is a regularization parameter, and $\mathbf{I}$ is an identity matrix. In this paper, the regularization is bypassed by deriving a lower bound of $\mathcal{J}^\phi$ as follows.

From (4), it is known that maximizing $\mathcal{J}^\phi=\mathrm{tr}(\mathbf{S}_B^\phi)/\mathrm{tr}(\mathbf{S}_W^\phi)$ is equivalent to maximizing $\mathcal{J}_1^\phi=\mathrm{tr}(\mathbf{S}_B^\phi)/\mathrm{tr}(\mathbf{S}_T^\phi)$. Let k_s denote a *stationary* kernel or a *normalized* kernel. The value of a stationary kernel only depends on the difference of two input samples, that is, $k_s(\mathbf{x}_i,\mathbf{x}_j)=k_s(\mathbf{x}_i-\mathbf{x}_j)$. A normalized kernel is defined as

$$k_s(\mathbf{x}_i,\mathbf{x}_j)=\left\langle\frac{\phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|},\frac{\phi(\mathbf{x}_j)}{\|\phi(\mathbf{x}_j)\|}\right\rangle=\frac{k(\mathbf{x}_i,\mathbf{x}_j)}{\sqrt{k(\mathbf{x}_i,\mathbf{x}_i)k(\mathbf{x}_j,\mathbf{x}_j)}}.$$

Geometrically, both kernels map the data onto a hypersphere in $\mathcal{K}$ with the radius of $\sqrt{k_s(\mathbf{x},\mathbf{x})}$. This is because $\|\phi(\mathbf{x})\|^2$, which is equal to $k_s(\mathbf{x},\mathbf{x})$, is constant. In addition, it is assumed that $k_s(\mathbf{x}_i,\mathbf{x}_j)\geq 0,\ \forall\mathbf{x}_i,\mathbf{x}_j\in\mathcal{D}$. A good example of the stationary kernel is the commonly used Gaussian RBF kernel $k(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2})$, where the

Gaussian width σ is the kernel parameter. Based on the above conditions, it can be shown that

$$\mathrm{tr}(\mathbf{S}_T^{\phi}) \leq \mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) - \frac{\mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n} = (n-1)k_s(\mathbf{x},\mathbf{x}). \quad (6)$$

Thus, a lower bound of $\mathcal{J}_1^{\phi}$ can be obtained as

$$\mathcal{J}_1^{\phi} \geq \frac{\mathrm{tr}(\mathbf{S}_B^{\phi})}{(n-1)k_s(\mathbf{x},\mathbf{x})} = \mathcal{J}_l^{\phi}, \quad \text{and} \quad \mathcal{J}_l^{\phi} \triangleq \mathrm{tr}(\mathbf{S}_B^{\phi}). \quad (7)$$

The last step removes $(n-1)k_s(\mathbf{x},\mathbf{x})$, since it is a constant in a classification task. By doing so, maximizing $\mathcal{J}_1^{\phi}$ can be approximated by maximizing its lower bound $\mathcal{J}_l^{\phi}$. Compared with $\mathcal{J}_1^{\phi}$, the criterion $\mathcal{J}_l^{\phi}$ is simpler and avoids manually setting a regularization parameter. *In this work, assuming that a stationary or normalized kernel is used, $\mathcal{J}_l^{\phi}$ is proposed as a criterion for feature selection.* In the rest of this paper, $\mathcal{J}^{\phi}$ is used to denote $\mathcal{J}_l^{\phi}$ for the convenience of notation.

3.3 Feature Selection by Maximizing Class Separability

Let $\boldsymbol{\alpha}(\boldsymbol{\alpha} \in \{0,1\}^d)$ be a d-dimensional indicator vector consisting of "0"(unselected) and "1"(selected). A set of selected features is written as $\mathbf{x}(\boldsymbol{\alpha}) = \boldsymbol{\alpha} \otimes \mathbf{x}$, where $\otimes$ denotes the componentwise multiplication. Given a criterion $\mathcal{C}$, the feature selection can be expressed as the following maximization (or minimization) problem:

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}\in\Lambda}[\mathcal{C}(\boldsymbol{\alpha} \otimes \mathbf{x})], \quad (8)$$

where Λ is the parameter space of $\boldsymbol{\alpha}$. Finding $\boldsymbol{\alpha}^*$ is a combinatorial optimization problem and is computationally intractable in general. A variety of suboptimal search methods have been developed to maintain a balance between computational feasibility and selection performance. This paper demonstrates the applicability of the proposed kernel class separability criterion to the following feature selection modes.

3.3.1 Best Individual N and Sequential Forward Selection

BIN and SEQ are two suboptimal search methods that are widely used in practical applications. In BIN, a selection criterion is individually applied to each of the features. Those giving larger criterion values are selected. It may be the simplest search method. In SEQ, one feature is selected and transferred to the set of selected features at each iteration. By adding this particular feature, the selection criterion computed with the selected feature set can be maximized.

Straightforwardly taking $\mathcal{J}^{\phi}$ as a selection criterion for these two search methods will be problematic, because the criterion value with a given feature set is subject to the value of kernel parameters. Fig. 1 shows an example from the synthetic data set used in the experiment section (Section 5.1). As shown in the figure, the value of $\mathcal{J}^{\phi}$ significantly varies with the value of σ. A poor setting of σ can easily blur the difference between good and bad features, making the feature selection fail. To handle this problem, the *maximal class separability* over the kernel parameter set $\boldsymbol{\theta}$ is used as a selection criterion, which is independent of the value of $\boldsymbol{\theta}$. It is expressed as

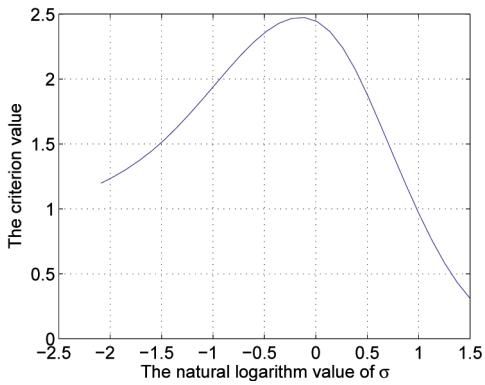

Fig. 1. $\mathcal{J}^{\phi}$ versus σ in a Gaussian RBF kernel (for a given feature).

$$\mathcal{J}^{\phi}(\boldsymbol{\alpha},\boldsymbol{\theta}^*) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\big[\mathcal{J}^{\phi}(\boldsymbol{\alpha},\boldsymbol{\theta})\big], \quad (9)$$

where $\boldsymbol{\Theta}$ denotes the parameter space of $\boldsymbol{\theta}$. The criterion $\mathcal{J}^{\phi}$ has continuous first-order and second-order derivatives with respect to $\boldsymbol{\theta}$, as long as the kernel function has. Hence, the maximization of $\mathcal{J}^{\phi}$ over $\boldsymbol{\theta}$ can be efficiently solved by gradient-based optimization techniques. This not only eliminates the affect of poor parameter setting but also frees users from manual parameter tuning, giving rise to a friendly selection criterion.

3.3.2 Selection via Kernel Parameter Optimization

Recently, in kernel-based feature selection [10], [11], there has been a trend of relaxing $\boldsymbol{\alpha}$ to a weighting vector in $\mathbb{R}^d$ with the constraint of $\alpha_i > 0$ $(i = 1, 2, \cdots, d)$. By doing so, the gradient-based optimization techniques can be employed to efficiently search for the optimal $\boldsymbol{\alpha}$, even if there are a large number of features. Feature selection is essentially converted to a KPO problem. The proposed criterion $\mathcal{J}^{\phi}$ is also applicable to this selection mode. For example, the Gaussian RBF kernel in this case becomes

$$\begin{aligned} k(\boldsymbol{\alpha}\otimes\mathbf{x},\boldsymbol{\alpha}\otimes\mathbf{y}) &= \exp\left[-\sum\nolimits_{i=1}^{d}\frac{\alpha_i^2(x_i-y_i)^2}{2\sigma^2}\right] \\ &= \exp\Big[-\sum\nolimits_{i=1}^{d}\eta_i(x_i-y_i)^2\Big], \end{aligned} \quad (10)$$

where $\eta_i = \alpha_i^2/(2\sigma^2)$. Finding the optimal $\boldsymbol{\alpha}$ is equivalent to finding the optimal kernel parameter set $\boldsymbol{\eta}$. That is

$$\boldsymbol{\eta}^* = \arg\max_{\boldsymbol{\eta}\in\mathbb{R}^d,\boldsymbol{\eta}>0}\big[\mathcal{J}^{\phi}(\boldsymbol{\eta})\big]. \quad (11)$$

A larger η_i^* indicates more important features when the features have been normalized into the similar scales.

When there are a large number of features (and, thus, a high-dimensional $\boldsymbol{\eta}$) but a small number of training samples, simply maximizing $\mathcal{J}^{\phi}(\boldsymbol{\eta})$ may fit the noise in the training samples and fail to correctly reflect the importance of features. This is often known as *overfitting*. In this case, a regularization term has to be added. In this paper, a regularized $\mathcal{J}^{\phi}(\boldsymbol{\eta})$ is proposed as

$$\mathcal{J}_{reg}^{\phi}(\boldsymbol{\eta}) = (1-\lambda)\mathcal{J}^{\phi}(\boldsymbol{\eta}) + \lambda\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2, \quad (12)$$

where λ $(0 \leq \lambda < 1)$ is the regularization parameter that penalizes the deviation of $\boldsymbol{\eta}$ from a preset $\boldsymbol{\eta}_0$. Mathematically, this imposes a Gaussian prior over $\boldsymbol{\eta}$. In this work, $\boldsymbol{\eta}_0$ is automatically found by optimizing (11) with the constraint of $\eta_1 = \eta_2 = \cdots = \eta_d$. Since this constraint reduces the number of free parameters to one, the overfitting will less likely happen. The regularization parameter λ needs to be set beforehand. Empirically, the larger the number of features, or the smaller the number of training samples, the larger the λ value. In addition, the k-fold cross validation can be used to tune λ if the computational load is not a critical issue and there are a sufficient number of training samples. In this paper, the λ value is mainly empirically set, with the k-fold cross validation employed when there are adequate training samples. Efficiently seeking the optimal λ will be a direction of the future work.

The function $\mathcal{J}^{\phi}_{reg}(\boldsymbol{\eta})$ is not convex, and a gradient-based search technique will find a local optimum. Meanwhile, it is found that $\mathcal{J}^{\phi}_{reg}(\boldsymbol{\eta})$ can be written as a difference of two convex functions when the Guassian RBF kernel is used.[5] Thus, the global optimum may be sought via the Difference of Convex functions (DC) Programming, which is an active research area in global optimization [12]. In this paper, $\mathcal{J}^{\phi}_{reg}(\boldsymbol{\eta})$ is optimized by the commonly used BFGS Quasi-Newton method. Reasonable experimental results are still obtained.

Before ending this section, it is worth noting that besides the above KPO approach, more complicated (and thus possibly more accurate) variants on finding $\boldsymbol{\alpha}^*$ are also available in [10] and [11]. The proposed criterion is still applicable. However, they are out of the focus of this work.

4 Analysis and Discussion

4.1 Relationship with the Radius-Margin Bound

Applying the proposed criterion to feature selection can be theoretically justified by its relationship to the radius-margin bound. This also helps us in understanding why the proposed criterion is faster and more robust than this bound.

Let $\mathcal{D}$ be a set of n training samples from two classes, and $\mathcal{L}(\mathcal{D})$ is the number of errors in the leave-one-out cross-validation procedure performed over $\mathcal{D}$. The value of $\mathcal{L}(\mathcal{D})$ is an estimate of the generalization error of an SVM classifier trained with $\mathcal{D}$. It is upper bounded as

$$\mathcal{L}(\mathcal{D}) \leq \frac{4R^2}{\gamma^2} = 4R^2\|\mathbf{w}\|^2, \tag{13}$$

where R is the radius of the smallest hypersphere enclosing the n training samples in the kernel space, γ is the margin, $\mathbf{w}$ is the normal vector of the optimal separating hyperplane of the SVM classifier, and $\gamma^{-1} = \|\mathbf{w}\|$. This result is based on the SVM with a hard margin that assumes separable data. For nonseparable data, the SVM with L2-norm soft margin will be used, because it can be interpreted as the SVM with a hard margin that employs a slightly modified kernel. This is followed in this paper. The R^2 and $\|\mathbf{w}\|^2$ are obtained by solving two quadratic optimization problems:

$$\begin{aligned} R^2 = \max_{\boldsymbol{\beta} \in \mathbb{R}^n} & \left[\sum_{i=1}^{n} \beta_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j=1}^{n} \beta_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) \right] \\ \textit{subject to}: & \quad \sum_{i=1}^{n} \beta_i = 1;\ \beta_i \geq 0, \end{aligned} \tag{14}$$

and

$$\begin{aligned} \frac{1}{2}\|\mathbf{w}\|^2 = \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} & \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right] \\ \textit{subject to}: & \quad \sum_{i=1}^{n} \alpha_i y_i = 0;\ \alpha_i \geq 0. \end{aligned} \tag{15}$$

The relationship between $\mathrm{tr}(\mathbf{S}^{\phi}_B)$ and γ^2 (or $1/\|\mathbf{w}\|^2$) and that between $\mathrm{tr}(\mathbf{S}^{\phi}_T)$ and R^2 can be proven as

$$\gamma^2 \leq \frac{1}{4 - \left(\frac{n}{n_1 n_2}\right)\mathrm{tr}(\mathbf{S}^{\phi}_B)}, \tag{16}$$

and

$$R^2 \geq \frac{1}{n}\mathrm{tr}(\mathbf{S}^{\phi}_T), \tag{17}$$

where n_1 and n_2 are the numbers of training samples from the two classes, respectively.

When minimizing the radius-margin bound, γ^2 is maximized, whereas R^2 is minimized. According to (16), γ^2 is upper bounded by $1/[4 - (\frac{n}{n_1 n_2})\mathrm{tr}(\mathbf{S}^{\phi}_B)]$. When the latter is small, the maximization of γ^2 will adversely be affected. Hence, a larger $1/[4 - (\frac{n}{n_1 n_2})\mathrm{tr}(\mathbf{S}^{\phi}_B)]$ will be preferred (although it does not necessarily lead to a larger γ^2). Similarly, according to (17), decreasing the value of $\mathrm{tr}(\mathbf{S}^{\phi}_T)$ will facilitate the minimization of R^2. The maximization of the kernel class separability criterion $\mathrm{tr}(\mathbf{S}^{\phi}_B)/\mathrm{tr}(\mathbf{S}^{\phi}_T)$ well reflects the above idea (note that larger $1/[4 - (\frac{n}{n_1 n_2})\mathrm{tr}(\mathbf{S}^{\phi}_B)]$ just means larger $\mathrm{tr}(\mathbf{S}^{\phi}_B)$). To some extent, this justifies the application of this criterion to feature selection. That is, *maximizing the kernel class separability selects the features by using which an SVM classifier is prone to achieving lower generalization error*. In addition, the kernel class separability criterion can also be interpreted as a special case of the radius-margin bound where all training samples are considered to be equally important, rather than being distinguished as *support* or non*support* vectors.

4.2 Relationship with the Kernel Alignment and the Kernel Fisher Discriminant Analysis

Although KA and KFDA are not designed for feature selection, it is still helpful to clarify their relationship and difference to the proposed criterion.

In [13], KA is developed as a measure of the alignment of a kernel function to a given classification task. Recall that $\mathbf{K}$ denotes a kernel matrix. An ideal kernel matrix is defined as $\mathbf{K}' = \mathbf{y}\mathbf{y}^\top$, where $\mathbf{y}$ $(\mathbf{y} \in \{+1, -1\}^n)$ is the vector consisting of the labels of n training samples. The measure is defined as

$$A(\mathbf{K}, \mathbf{y}\mathbf{y}^\top) = \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle}{\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle \langle \mathbf{y}\mathbf{y}^\top, \mathbf{y}\mathbf{y}^\top \rangle}} = \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle}{n\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle}}. \tag{18}$$

5. The theoretical analysis and proofs of some results in this paper can be found in the author's homepage.

KA can be viewed as a special case of the kernel class separability criterion. It is proven that

$$\begin{aligned}\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle &= (n_1+n_2)\mathrm{tr}(\mathbf{S}_B^\phi)\Big|_{n_1=n_2},\\ \langle \mathbf{K}, \mathbf{K} \rangle &\le (n_1+n_2)\Big[(n_1+n_2) - \mathrm{tr}(\mathbf{S}_T^\phi)\Big].\end{aligned} \tag{19}$$

Note that the second result uses the condition of $k(\mathbf{x}_i, \mathbf{x}_j) \in (0, 1]$, which can be satisfied by a Gaussian RBF kernel and a part of the normalized kernels. As observed, $\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle$ is a special case of $\mathrm{tr}(\mathbf{S}_B^\phi)$ when the numbers of training samples from the two classes n_1 and n_2 are the same. In [13], by constraining $\langle \mathbf{K}, \mathbf{K} \rangle$ as a constant C_0, the value of $\langle \mathbf{K}, \mathbf{y}\mathbf{y}^\top \rangle$ is maximized to optimize the combination coefficients of a set of kernels. From the viewpoint of kernel class separability, this imposes a constraint of $\mathrm{tr}(\mathbf{S}_T^\phi) \le \left[(n_1+n_2) - \frac{C_0}{(n_1+n_2)}\right]$ and maximizes $\mathrm{tr}(\mathbf{S}_B^\phi)\Big|_{n_1=n_2}$.

KFDA finds an optimal projection from the kernel space $\mathcal{K}$ to a lower dimensional subspace in the sense that two classes are maximally separated in that subspace. The optimal projection $\mathbf{w}$ ($\mathbf{w} \in \mathcal{K}, \mathbf{w} \neq \mathbf{0}$) is sought by maximizing

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B^\phi \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_T^\phi \mathbf{w}}. \tag{20}$$

The projection of a sample $\mathbf{x}$ into the subspace is obtained as $\mathbf{y} = \mathbf{w}^\top \phi(\mathbf{x})$, which is a linear combination of all features in $\phi(\mathbf{x})$.

Feature selection with the proposed criterion and the KFDA algorithm are essentially different, although both of them aim at maximizing the class separability. In KFDA, $\frac{\mathbf{w}^\top \mathbf{S}_B^\phi \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_T^\phi \mathbf{w}}$ measures the class separability *in a lower dimensional subspace*. The variable is $\mathbf{w}$, which is a projection from a high-dimensional kernel space to a lower dimensional subspace. Its optimal value is found by the maximization of the class separability, which is solved via an eigendecomposition. On the other hand, feature selection with the proposed criterion aims at evaluating the importance of each feature and identifying more important ones. To achieve this, a class separability criterion *in a high-dimensional kernel space* $\mathrm{tr}(\mathbf{S}_B^\phi)/\mathrm{tr}(\mathbf{S}_T^\phi)$ is maximized. The variable is the parameter set of a kernel function. Its optimal value is sought via a gradient-based search method.

In the theoretical aspect, it is proven that the kernel class separability criterion is actually a lower bound of the maximum value of the KFDA's objective function. It is expressed as

$$\frac{\mathrm{tr}(\mathbf{S}_B^\phi)}{\mathrm{tr}(\mathbf{S}_T^\phi)} \le \mathcal{J}(\mathbf{w}^\star) = \max_{\mathbf{w}\in\mathcal{K}, \mathbf{w}\neq\mathbf{0}} [\mathcal{J}(\mathbf{w})]. \tag{21}$$

This lower bound is independent of $\mathbf{w}$ and is a function of the kernel parameters only. This result can be understood as follows: In KFDA, the maximum class separability that could be achieved in a lower dimensional subspace is subject to the class separability in the high-dimensional kernel space. Improving the value of $\frac{\mathrm{tr}(\mathbf{S}_B^\phi)}{\mathrm{tr}(\mathbf{S}_T^\phi)}$ may help boost the value of $\mathcal{J}(\mathbf{w}^\star)$. Through the above analysis, it can be clearly found that the proposed kernel class separability criterion is *not* a simple reinvention of the KFDA algorithm.

4.3 Benefit of Using the Kernel Class Separability Criterion

Generally speaking, when there are a sufficient number of training samples, little noise in data, and adequate selection time, minimizing the radius-margin bound can give a better feature selection result, since the proposed criterion is only an approximation of this bound. However, it is observed that the proposed criterion often provides more benefits in practice.

Computational efficiency. For a given kernel parameter set, each evaluation of the radius-margin bound needs to solve two quadratic optimization problems, which can considerably prolong the feature selection process. Comparatively, each evaluation of the proposed criterion has much less computational load, since it does not involve any optimization. It can significantly reduce the time cost, leading to faster feature selection.

Robustness. The radius-margin bound considers the *worst* case (for example, it maximizes the minimum margin between two classes) by solving (14) and (15). When training samples are scarce, the estimated margin will less likely reflect the true margin due to its high functional complexity (overfitting happens). In addition, the estimation of R^2 is prone to being affected by noisy samples. In these cases, the radius-margin bound can no longer accurately predict the generalization error, and this, in turn, affects its feature selection performance. Improving the robustness of the SVMs has attracted much attention and has been an active research topic [14]. Comparatively, the proposed kernel class separability criterion is less sensitive to the scarcity of training samples and the presence of data noise, because it evaluates the *average* case (for example, it maximizes the distance between two class means) of data separability and has much lower functional complexity.

Stability. Feature selection with the radius-margin bound needs two loops of optimization. The outer loop minimizes the radius-margin bound with respect to the kernel parameters, whereas the inner loop computes the radius and margin by solving (14) and (15). In the outer loop, when a search direction is determined, a line search will be performed to find the minimum of the radius-margin bound along this direction. This involves evaluating the radius-margin bound with a series of kernel parameters. However, these parameter values are suggested by the linear search mechanism and are not necessarily reasonable for the training data. For example, a large Gaussian width or regularization parameter C may be suggested when the training samples are not separable. This will result in a very long or even endless optimization process in solving (14) and (15). The proposed class separability criterion is completely free of this problem.

5 Experimental Result

This experiment evaluates the performance of the proposed feature selection criterion in dealing with linearly nonseparable classes, fast feature selection, small sample set, and noisy features. This criterion is compared with the Pearson

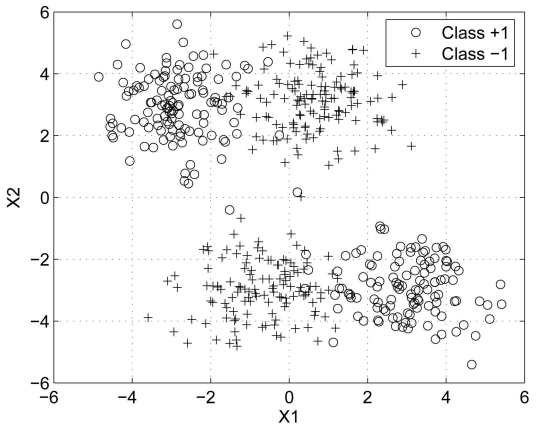

Fig. 2. Illustration of the synthetic data set.

correlation coefficient, the Kolmogorov-Smirnov test, class separability (nonkernel), Boosting feature selection, and, particularly, the radius-margin bound. The feature selection performance is measured by 1) the percentage of correct selection, 2) the test error of an SVM classifier using the selected features, and 3) the time cost by feature selection. Higher percentage of correct selection, lower test error, and less selection time indicate better performance. Three feature selection modes of "BIN," "SEQ," and "KPO" are investigated. The Gaussian RBF kernel is employed. The LIBSVM [15] is used for training, test, and timing. The BFGS Quasi-Newton method is employed to maximize the proposed criterion or minimize the radius-margin bound. The code provided in [16] is used.

5.1 Result on the Synthetic Data Set

A synthetic data set is created by following [11]. It is a nonlinear binary classification problem and has been used as a benchmark to test feature selection criteria. In this data set, only two out of the 52 features are statistically relevant to the class label, whereas all the others are random noise. For each sample $\mathbf{x}$ ($\mathbf{x} \in \mathbb{R}^{52}$), its label y has the equal probability of being $+1$ or -1. The features x_1 and x_2 are drawn from a normal distribution of $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I})$ or $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{I})$ with equal probability, where $\mathbf{I}$ is an identity matrix. When $y = -1$, $\boldsymbol{\mu}_1 = [-0.75, -3]^\top$, and $\boldsymbol{\mu}_2 = [0.75, 3]^\top$. When $y = +1$, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ become $[3, -3]^\top$ and $[-3, 3]^\top$, respectively. The remaining 50 features $x_3, \cdots, x_{52}$ are randomly sampled from $\mathcal{N}(0, 20)$. The distribution of x_1 and x_2 is illustrated in Fig. 2. x_1 is the most discriminative feature. The two classes can be best separated by using both x_1 and x_2. Assuming that we have known that only two features are useful, the following experiments investigate if they can be correctly selected by the aforementioned criteria.

5.1.1 Result on the Mode of Best Individual N

Fig. 3 compares the percentage of correct selection and the test error of an SVM classifier with the two features selected by different criteria. All of the results are averaged over 30 groups. Fig. 3a shows the percentage of correctly selecting x_1 within the top two. As observed, both the Pearson correlation coefficient and class separability (nonkernel) fail to identify x_1. This is not surprising, because they cannot effectively handle the linearly nonseparable data. Free of this problem, the Kolmogorov-Smirnov test produces better performance. However, it needs enough samples to estimate the underlying distribution. This affects its performance when the number of training samples is small. Higher selection percentage is obtained by the proposed criterion and the radius-margin bound. The percentage of exactly selecting (x_1, x_2) as the top two features is plotted in Fig. 3b. All selection criteria give a poor result. This can be expected for the BIN selection mode. With the two selected features, the test error of an SVM classifier is further compared across these selection criteria. The same 100 training samples and 500 test samples are applied. For a fair comparison, the hyperparameters in each SVM classifier are equally optimized via a 5-fold cross validation. The mean and standard deviation of the test errors are plotted in Figs. 3c and 3d, respectively. Consistent with the selection percentage, the SVM classifier with the features selected by the proposed criterion or the radius-margin bound produces the lowest test error. The significance test is conducted, as shown in Table 1, where the **McNemar** test [17] with the significance level of 0.05 is used. According to the test result, each of the 30 groups is categorized as "KCSM (significantly) better," "KCSM (significantly) worse," or "No statistical difference."[6] The number of groups in each category is listed. The proposed criterion achieves a performance comparable to that of the radius-margin bound. The time spent by different feature selection criteria is compared in the first part (the BIN mode) of Table 5. For the kernel class separability and the radius-margin bound, the time mainly includes the portion of kernel evaluation and quadratic optimization. The portion for reading and writing data and preprocessing is excluded, because it varies with programming. Hence, the sign "$>$" is put before these numbers. As observed, the proposed criterion is faster than the radius-margin bound. In terms of the ratio of selection performance to selection time, the proposed criterion is the best one.

5.1.2 Result on the Mode of Sequential Forward Selection

In this selection mode, the proposed criterion is compared with the radius-margin bound and the Boosting feature selection [18]. Other criteria are omitted due to their relatively poor selection performance. Fig. 4 shows the percentage of correct selection and the SVM test error. The proposed criterion obtains comparable or slightly better performance than the radius-margin bound. This is confirmed by the significance test result in Table 2. The Boosting feature selection correctly selects the most useful feature x_1, but it fails to identify the best combination of (x_1, x_2). In this selection mode, the proposed criterion still achieves excellent selection performance. Meanwhile, it maintains the faster feature selection than the radius-margin bound, as shown in the second part (the SEQ mode) of Table 5.

5.1.3 Result on the Mode of Kernel Parameter Optimization

This selection mode is newly proposed in [10], where feature selection is performed by optimizing a criterion with respect to the kernel parameter assigned to each

6. "KCSM significantly better" means that a statistically significant difference is detected between the test errors and that the mean of the test errors with respect to the proposed criterion is lower. "KCSM (significantly) worse" is defined in a similar way.

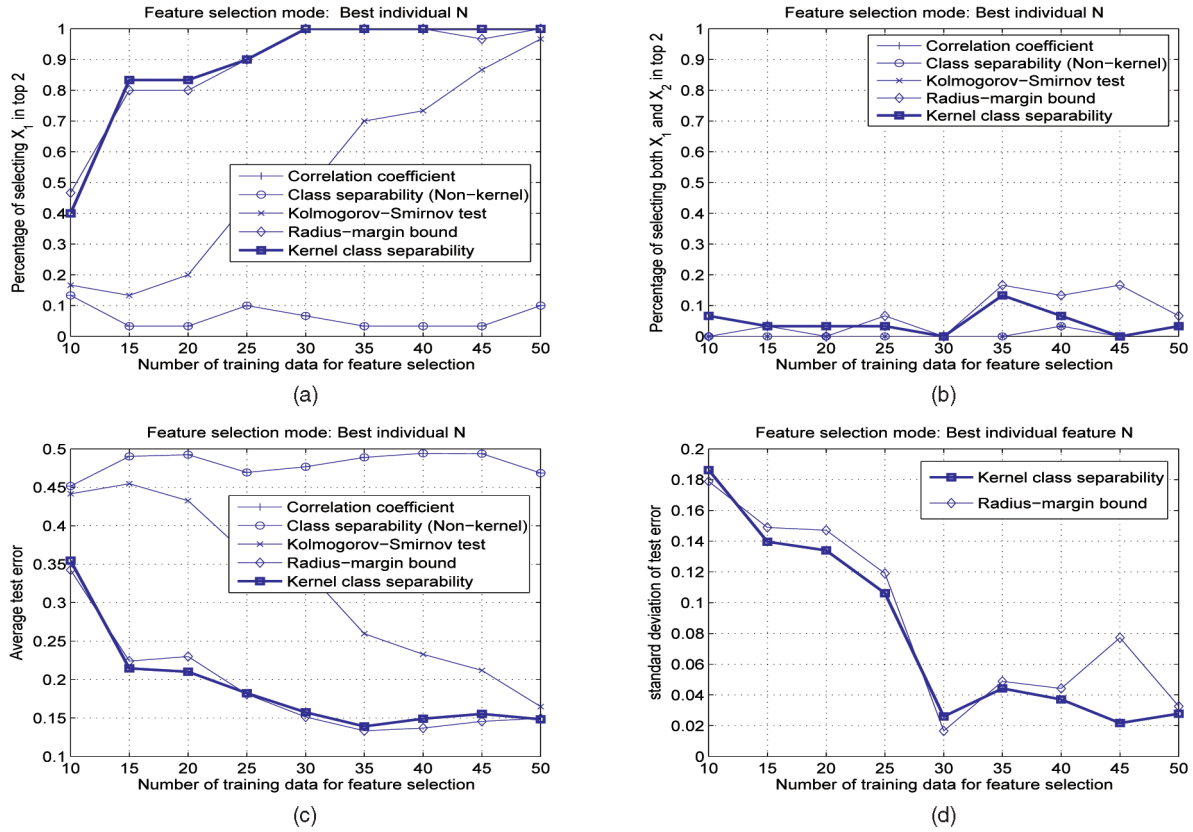

Fig. 3. Feature selection result and the SVM test error with the two selected features (the BIN mode and synthetic data). The Significance Test result is reported in Table 1. (a) Correctly selecting x_1 within the top two. (b) Correctly selecting both x_1 and x_2 as the top two. (c) Mean of the test errors. (d) Standard deviation of the test errors.

TABLE 1
Significance Test of the SVM Test Errors with respect to KCSM and RMB in Fig. 3 (the BIN Mode)

	Number of training data for feature selection									
McNemar Test ($\alpha = 0.05$)	10	15	20	25	30	35	40	45	50	Average
No. of groups KCSM better	2	7	7	7	0	4	1	4	6	4.22
No. of groups KCSM worse	2	5	6	6	4	5	6	8	1	4.78
No. of groups No difference	26	18	17	17	26	21	23	18	23	21.00

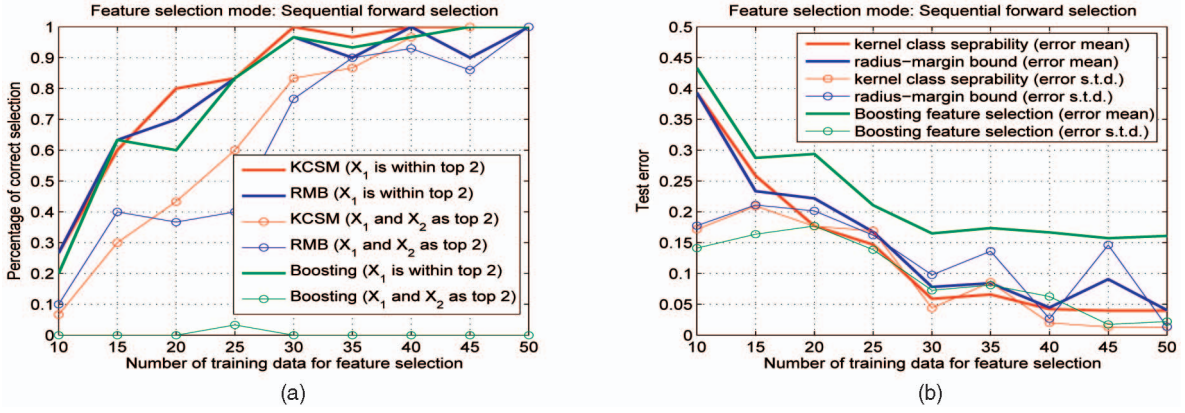

Fig. 4. Feature selection result and the SVM test error with the two selected features (the SEQ mode and synthetic data). The Significance Test result is reported in Table 2. (a) Percentage of correct selection. (b) SVM test error with the two selected features.

feature. This mode is attractive, because the optimization problem is efficiently solved by using the gradient-based search methods, rather than the exhaustive search, in the modes of BIN and SEQ. This selection mode is the focus of the rest of this experimental study. As before, the proposed criterion and the radius-margin bound are compared, and the results are reported in Fig. 5 and Table 3. Comparatively, the proposed criterion achieves a higher percentage

TABLE 2
Significance Test of the SVM Test Errors with respect to KCSM and RMB in Fig. 4 (the SEQ Mode)

	Number of training data for feature selection									
McNemar Test ($\alpha = 0.05$)	10	15	20	25	30	35	40	45	50	Average
No. of groups KCSM better	2	4	7	7	6	3	3	5	1	4.22
No. of groups KCSM worse	5	6	4	3	3	3	1	0	0	2.78
No. of groups No difference	23	20	19	20	21	24	26	25	29	23.00

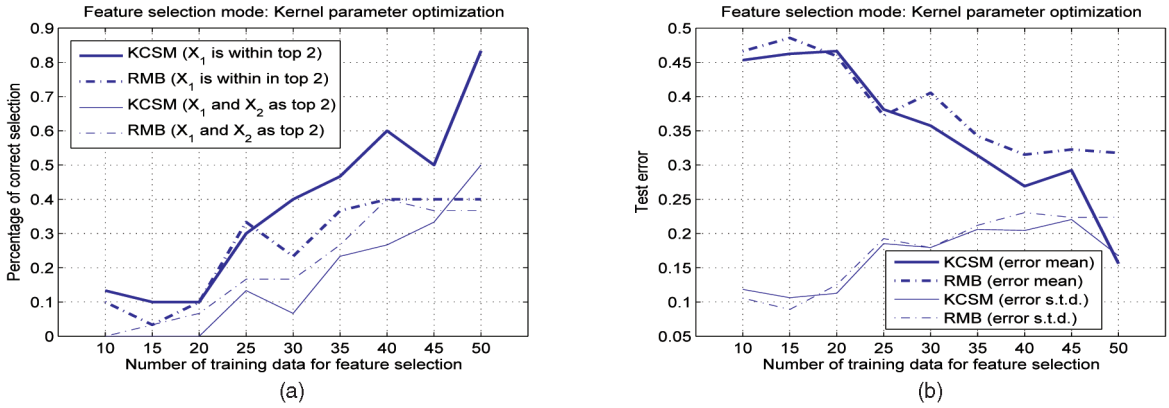

Fig. 5. Feature selection result and the SVM test error with the two selected features (the KPO mode and synthetic data). The Significance Test result is reported in Table 3. (a) Percentage of correct selection. (b) SVM test error with the two selected features.

TABLE 3
Significance Test of the SVM Test Errors with respect to KCSM and RMB in Fig. 5 (the KPO Mode)

	Number of training data for feature selection									
McNemar Test ($\alpha = 0.05$)	10	15	20	25	30	35	40	45	50	Average
No. of groups KCSM better	3	3	3	7	8	10	9	11	17	7.89
No. of groups KCSM worse	2	1	3	6	5	9	9	9	4	5.33
No. of groups No difference	25	26	24	17	17	11	12	10	9	16.78

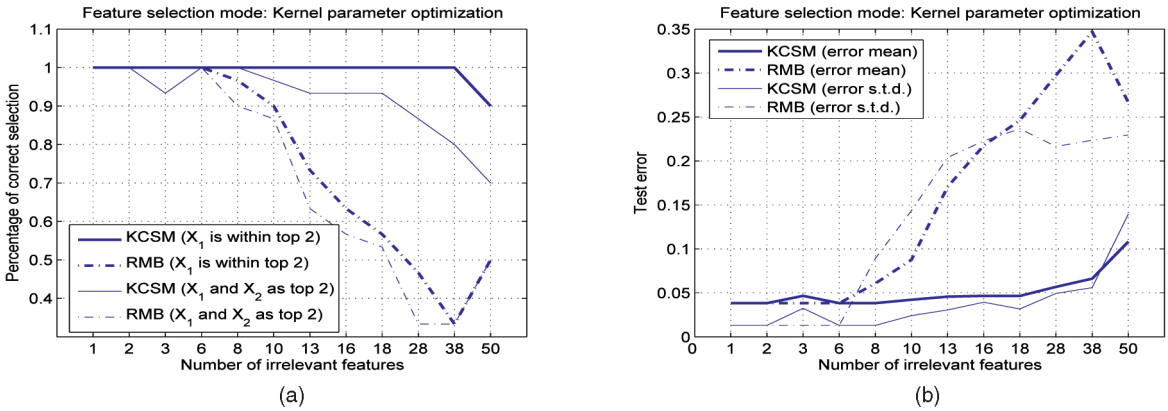

Fig. 6. Feature selection result and the SVM test error with the two selected features (the KPO mode and synthetic data). The Significance Test result is reported in Table 4. (a) Percentage of correct selection. (b) SVM test error with the two selected features.

in selecting the feature x_1, whereas the percentage of selecting the combination of (x_1, x_2) is a bit lower. However, as demonstrated by the SVM test error, the significance test, and the timing result in the third part (the KPO mode) of Table 5, the proposed criterion still produces overall better performance.

Since this selection mode simultaneously considers all features by a one-shot optimization, it is important to investigate the selection performance of a criterion with respect to the number of irrelevant features. A better criterion should be able to maintain a high percentage of correct selection with the increasing number of irrelevant features. Fig. 6 compares the proposed criterion and the radius-margin bound in this case. To remove the affect of small sample set, a sufficient number (100) of training samples are used in feature selection. Again, the result is averaged over 30 groups. Fig. 6a shows that the selection percentage of the radius-margin bound quickly drops when

TABLE 4
Significance Test of the SVM Test Errors with respect to KCSM and RMB in Fig. 6 (the KPO Mode)

	Number of irrelevant features												
McNemar Test ($\alpha = 0.05$)	1	2	3	6	8	10	13	16	18	28	38	50	Average
No. of groups KCSM better	0	0	0	0	3	4	11	13	14	19	20	14	8.17
No. of groups KCSM worse	0	0	2	0	0	1	1	2	1	2	2	8	1.58
No. of groups No difference	30	30	28	30	27	25	18	15	5	9	8	8	20.25

TABLE 5
Time Cost versus the Number of Training Samples

Method (with BIN mode)	10	15	20	25	30	35	40	45	50	Unit
Correlation coefficient	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	sec.
Class separability (No-kernel)	0.16	0.09	0.10	0.10	0.10	0.11	0.12	0.13	0.07	sec.
Kolmogorov-Smirnov test	0.07	0.15	0.09	0.06	0.06	0.06	0.07	0.06	0.06	sec.
Kernel class separability	> 0.08	0.22	0.36	0.57	0.83	1.07	1.50	1.81	2.29	sec.
Radius-margin bound	> 0.54	1.22	1.96	2.50	2.97	4.11	4.04	6.78	8.42	sec.
Method (with SEQ mode)	10	15	20	25	30	35	40	45	50	Unit
Kernel class separability	> 0.16	0.43	0.72	1.13	1.65	2.12	2.97	3.59	4.53	sec.
Radius-margin bound	> 1.06	2.42	3.89	4.95	5.88	8.15	8.02	13.44	16.67	sec.
Method (with KPO mode)	10	15	20	25	30	35	40	45	50	Unit
Kernel class separability	> 2.90	7.3	11.5	19.3	22.8	39.4	49.7	89.9	97.9	millisec.
Radius-margin bound	> 5.40	15.1	17.2	23.5	40.9	84.5	70.6	108.4	75.0	millisec.

TABLE 6
Percentage of Selecting (x_1, x_2) versus the Number of Irrelevant Features (Different λ Values)

Method (KPO mode)	1	3	6	8	10	13	16	18	28	38	50
Kernel CS ($\lambda = 0$)	100	93.3	100	100	96.7	93.3	93.3	93.3	86.7	80	70
Kernel CS ($\lambda = 0.10$)	100	100	100	100	100	100	100	100	100	100	80
Kernel CS ($\lambda = 0.25$)	100	100	100	100	100	100	100	100	100	96.7	76.7
Kernel CS ($\lambda = 0.50$)	100	100	100	100	100	100	100	100	100	96.7	76.7
Kernel CS ($\lambda = 0.75$)	100	100	100	100	100	100	100	100	100	93.3	76.7
Kernel CS ($\lambda = 0.99$)	100	100	100	100	100	100	100	100	100	93.3	76.7
RM bound ($\lambda = 0$)	100	100	100	96.7	90	73.3	73.3	60	40	40	50
RM bound ($\lambda = 0.10$)	100	100	100	96.7	90	73.3	63.3	56.7	46.7	33.3	50
RM bound ($\lambda = 0.25$)	100	100	100	96.7	93.3	73.3	63.3	56.7	33.3	33.3	50
RM bound ($\lambda = 0.5$)	100	100	100	96.7	93.3	73.3	60	56.7	36.7	43.3	53.3
RM bound ($\lambda = 0.75$)	100	100	100	100	90	73.3	66.7	63.3	33.3	36.7	50
RM bound ($\lambda = 0.99$)	100	100	100	96.7	93.3	63.3	66.7	56.7	36.7	30	56.7

more irrelevant features are included. In other words, it cannot deliver a good-enough selection result by simply applying a one-shot KPO (in [10], a sequential backward elimination of the worst features obtains better selection performance, but it involves multiple times of KPO). Comparatively, the proposed criterion demonstrates much better selection performance with the increasing number of irrelevant features. It is believed that the poor performance of the radius-margin bound is due to its sensitivity to data noise. Recall that R is the radius of the smallest hypersphere enclosing all training samples. The value of R heavily depends on the sample that most deviates from the center of data. When this deviation is caused by noise rather than the underlying data distribution, R will become noisy. Applying the kernel class separability criterion can considerably mitigate this problem, because it measures the average radius of data scattering (via $\mathrm{tr}(\mathbf{S}_T^{\phi})$). Moreover, this work takes the lower bound in (7) as a feature selection criterion, which further reduces the impact of the radius estimation. The SVM test error in Fig. 6b and the significance test in Table 4 confirm the advantage of the proposed criterion. Aside from the above comparison, it is also investigated whether the poor performance of the radius-margin bound in this case could be improved by applying regularization. By setting the regularization parameter λ in (12) to different values ranging from 0 to 0.99, a regularized radius-margin bound is used. In Table 6, it is seen that for this feature selection problem, the radius-margin bound cannot

Fig. 7. Example of the digit images in the USPS data set.

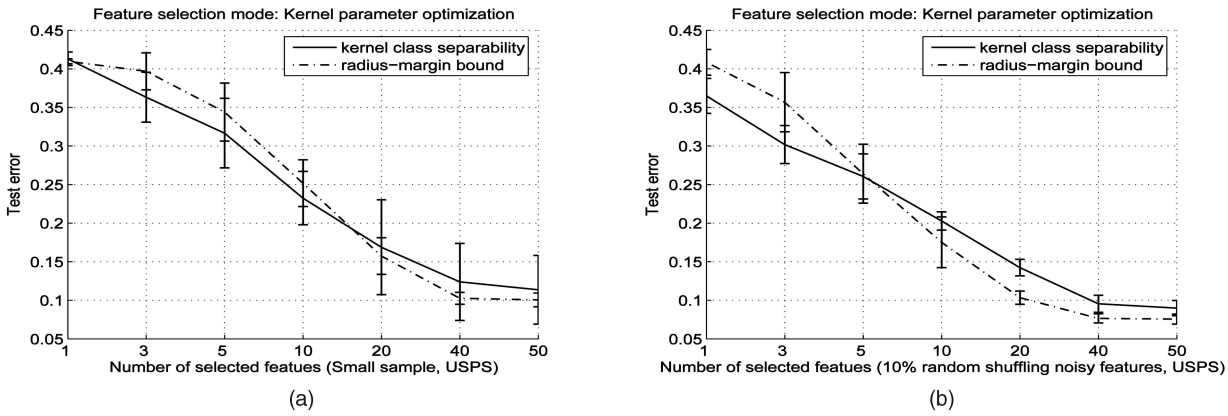

Fig. 8. (a) Test error of SVM with selected features (small sample). (b) Test error of SVM with selected features (10 percent noisy features obtained by random shuffling are used). Note that the radius-margin bound in this figure uses the regularization proposed in this work. The original radius-margin bound gives poorer performance in this case.

produce better performance, even if the regularization is applied. As for the proposed criterion, the regularization slightly improves its performance.

5.2 Result on the Data Set of the US Postal Service

The USPS data set has been widely used as a benchmark for evaluating learning algorithms [2]. It contains 7,291 training samples and 2,007 test samples, which can be downloaded from [15]. They are from 10 classes of digits from 0 to 9, as shown in Fig. 7. Each sample is characterized by 256 features obtained by reshaping a 16×16 gray-level thumbnail image to a long vector. Following the experimental setting in [10], a binary classification problem is created to separate the digits of $0\sim 4$ from $5\sim 9$. Feature selection in this case is to identify more discriminative ones from the 256 features. The performance of the proposed criterion in the presence of small sample set and noisy features is investigated.

To simulate the case of small sample set, m $(m = 4, 5, 6, \cdots, 10)$ training samples are randomly chosen from each class, forming a small-sized training set. Seven training sets are obtained in total. After performing feature selection on each of them, an SVM classifier with the top-k $(k = 1, 3, 5, 10, 20, 40, 50)$ selected features is trained with 1,000 training samples and tested on the predefined 2,007 test samples. Finally, the seven test errors are averaged. Different regularization parameters $(\lambda = 0, 0.1, 0.25, 0.5, 0.75, 0.99)$ are applied, and the lowest average test error from each criterion is compared to each other. As shown in Fig. 8a, the proposed criterion is better in identifying the most important features, which is reflected by its lower test error in the early stage. The radius-margin bound catches up when more features are selected.

In the experiment on noisy features, 10 percent of the 256 features are replaced by the noisy features generated by randomly reshuffling each of them. The noisy features generated this way maintain their original unconditional distributions. Thirty groups of noisy training sets are created, each including 1,000 samples randomly selected from the predefined training set. Again, feature selection is performed on each of them. This time, the regularization parameter λ is selected by applying the 5-fold cross validation. Each training set is randomly partitioned into five subsets. Each subset is used as a validation subset *once*, and the remaining four subsets are correspondingly used as a training subset. Based on this training subset, feature selection with a given λ value is performed, and an SVM classifier with the top-k selected features is trained. The test error of the SVM classifier is then computed via the validation subset. This process is repeated five times, and the five validation errors are averaged, forming a criterion of the goodness of this λ value. From a given selection pool, the λ value that gives rise to the minimum average validation error is chosen. In this experiment, the selection pool of λ is (0, 0.1, 0.25, 0.5, 0.75, 0.99). With the 5-fold cross validation, the value of 0.99 is consistently selected for all the 30 training sets when the proposed criterion is used. For the radius-margin bound, the value of 0.75 is selected for 27 training sets, the value of 0.5 is selected for 2, and the value of 0.25 is selected for 1. With the selected λ value, feature selection is then carried out on each of the 30 training sets. The SVM classifier with the top-k selected features is trained and then tested on the predefined 2,007 test samples (10 percent of the 256 features of these samples have also been replaced by the noisy features). As plotted in Fig. 8b, the proposed criterion still shows better performance in selecting a small number of features. The significance test is listed in Table 7, from which a similar conclusion can be drawn. In this experiment, the selection time of the radius-margin bound is much longer than that of the proposed criterion. This is because in each evaluation of the radius-margin bound, two quadratic programming problems with 1,000 training samples have to be solved. In addition, the minimization of the radius-margin bound needs more function evaluations.

5.3 Result on the Data Set of Deoxyribonucleic Acid

This data set is taken from the Statlog Project database. Its original task is to decide whether there is a *splice junction* in a given DNA sequence and infer the type of this junction if there is. In this experiment, this data set is used as a binary

TABLE 7
Significance Test of the SVM Test Errors with respect to KCSM and RMB in Fig. 8b (the KPO Mode)

	Number of selected features							
McNemar Test ($\alpha = 0.05$)	1	3	5	10	20	40	50	Average
No. of groups KCSM better	23	21	10	4	0	0	0	8.29
No. of groups KCSM worse	0	1	6	16	27	20	13	11.86
No. of groups No difference	7	8	14	10	3	10	17	9.86

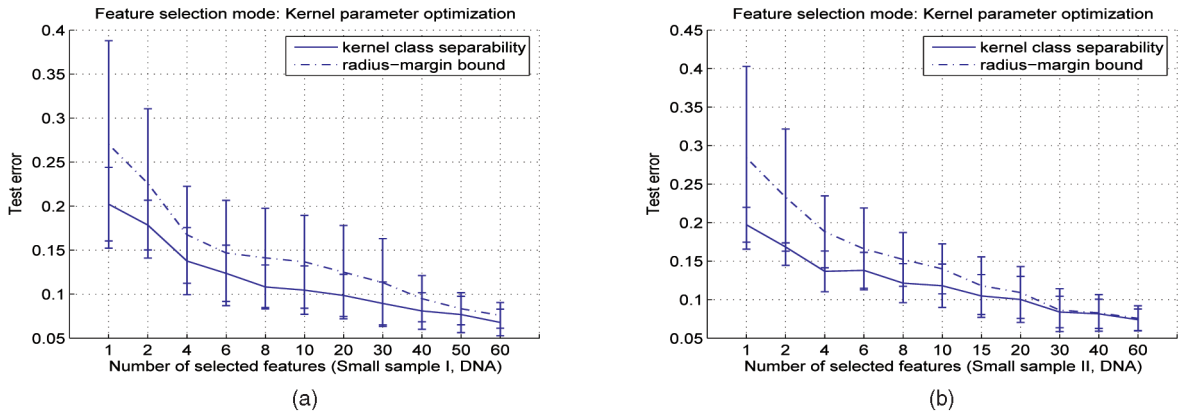

Fig. 9. (a) Test error of the SVM with the selected features (50 samples). (b) Test error of the SVM with the selected features (40 samples). Note that the radius-margin bound in this figure uses the regularization proposed in this work. The original radius-margin bound gives poorer performance in this case.

TABLE 8
Significance Test of the SVM Test Errors with respect to KCSM and RMB in Fig. 9 (the KPO Mode)

	Number of selected features (Small sample, DNA)											
McNemar Test ($\alpha = 0.05$)	1	2	4	6	8	10	15	20	30	40	60	Average
For the sub-figure (a)												
No. of groups KCSM better	9	9	11	8	11	10	9	9	9	8	8	9.18
No. of groups KCSM worse	0	0	3	3	2	1	1	2	3	4	2	1.91
No. of groups No difference	11	11	6	9	7	9	10	9	8	8	10	8.91
For the sub-figure (b)												
No. of groups KCSM better	10	9	9	6	10	8	8	6	4	4	4	7.09
No. of groups KCSM worse	0	0	0	1	2	2	4	4	4	4	3	2.18
No. of groups No difference	10	11	11	13	8	10	8	10	12	12	13	10.73

classification problem by only deciding the "presence" or "absence" of a splice junction. This data set includes 2,000 training samples and 1,186 test samples. Each sample is a DNA sequence consisting of 60 nucleotides. Each nucleotide is described by three binary features. This way, a DNA sequence is represented by $60 \times 3 = 180$ features in total. This data set is used, because it has a "ground truth" to some extent: "*Much better performance is generally observed if attributes closest to the junction are used. This means using attributes A61 to A120 only.*"[7] It provides a good way of evaluating the feature selection performance. Twenty small-sized training subsets are randomly sampled from the predefined training set, each of which contains only 50 samples. As before, with the proposed criterion or the radius-margin bound, feature selection is performed on each training subset via KPO. After that, an SVM classifier is trained with the top-k ($k = 1, 2, 4, 6, 8, 10, 15, 20, 30, 40, 60$) selected features and evaluated on the predefined test set. By applying different regularization parameter values, the lowest test error from each criterion is picked and compared, as shown in Fig. 9a. The proposed criterion produces lower test errors than the radius-margin bound, showing better feature selection performance. This experiment is repeated by further reducing the number of training samples in each training subset to 40. The SVM test errors are compared in Fig. 9b, from which a similar result is observed. The significance test result is reported in Table 8. The number of groups on which the proposed criterion wins is clearly higher. This verifies the better performance achieved by the proposed criterion.

Finally, since this data set has a "ground truth" about the features that should be selected, the following will check whether they are really picked by the proposed criterion. The optimized values of η_i ($i = 1, 2, \cdots, 180$) are averaged over the 20 trials and are plotted in Fig. 10. The features from the 80th to the 100th are assigned higher η values, showing that they are identified as more discriminative features. This result well matches the aforementioned "ground truth."

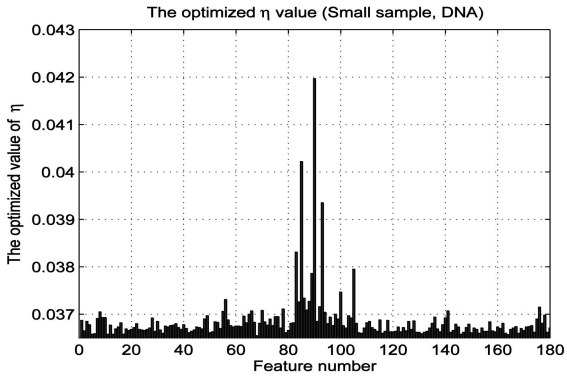

Fig. 10. The average of the optimized value of η.

7. http://www.liacc.up.pt/ML/old/statlog/datasets/dna/dna.descri.html.

6 Conclusion and Future Work

In this paper, a kernel-based class separability measure is developed as a feature selection criterion. A feature subset that gives rise to higher class separability is considered to be more important. With this criterion, different modes of feature selection are studied. Via our theoretical analysis, the relationship of the proposed criterion to the radius-margin bound, the KFDA, and the KA is exposed. This helps us in understanding the advantages and disadvantages of the proposed criterion for feature selection. As experimentally demonstrated, this criterion gives the overall best feature selection performance among the compared ones. It delivers faster feature selection and well handles linearly nonseparable data. In addition, it is robust to both the scarcity of training samples and the presence of noisy features.

Much future work is worth exploring. For example, the proposed kernel class separability criterion is unified for both binary and multiclass classification. It can be readily applied to the feature selection of a multiclass problem that is more common in practice. In addition, it would be appealing if the optimal number of features to be selected could automatically be determined. It is believed that accurately and efficiently identifying this number still remains an open issue. A possible way is to treat this number as an extra variable and maximize the criterion with respect to it. Finally, since this criterion is proven as a lower bound of the maximum value of the KFDA's objective function, its maximization may be used to tune the kernel parameters in the KFDA. This approach is expected to be more computationally efficient than those computing the leave-one-out cross-validation bound [19], [20].

Acknowledgments

The author thanks the anonymous reviewers for their constructive comments and suggestions, Chunhua Shen for discussing the concept of DC programming, and Richard Hartley and Luping Zhou for reading the paper draft and giving valuable comments. This work is supported by the ARC Discovery Project under Grant DP0773761.

References

[1] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer Verlag, 1995.
[2] B. Schölkopf and A. Smola, *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002.
[3] S. Haykin, *Neural Networks: A Comprehensive Foundation,* second ed. Prentice Hall, 1999.
[4] B. Schölkopf, A.J. Smola, and K.-R. Müller, "Kernel Principal Component Analysis," *Advances in Kernel Methods—Support Vector Learning,* pp. 327-352, 1999.
[5] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher Discriminant Analysis with Kernels," *Neural Networks for Signal Processing IX,* Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, eds., pp. 41-48, IEEE, 1999.
[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.
[7] S. Theodoridis and K. Koutroumbas, *Pattern Recognition.* Academic Press, 1999.
[8] R.O. Duda, D.G. Stork, and P.E. Hart, *Pattern Classification,* second ed. John Wiley & Sons, 2001.
[9] A.R. Webb, *Statistical Pattern Recognition,* second ed. John Wiley & Sons, 2002.
[10] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing Multiple Parameters for Support Vector Machines," *Machine Learning,* vol. 46, nos. 1-3, pp. 131-159, 2002.
[11] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature Selection for SVMs," *Advances in Neural Information Processing Systems 13—Proc. Ann. Conf. Neural Information Processing Systems (NIPS '00),* T.K. Leen, T.G. Dietterich, and V. Tresp, eds., pp. 668-674, MIT Press, 2000.
[12] L.T.H. An and P.D. Tao, "DC Programming. Theory, Algorithms and Applications: The State of the Art," *Proc. First Int'l Workshop Global Constrained Optimization and Constraint Satisfaction (COCOS '02),* pp. 131-159, Oct. 2002.
[13] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J.S. Kandola, "On Kernel-Target Alignment," *Advances in Neural Information Processing Systems 14—Proc. Ann. Conf. Neural Information Processing Systems (NIPS '01),* T.G. Dietterich, S. Becker, and Z. Ghahramani, eds., pp. 367-373, MIT Press, 2001.
[14] L. Xu, K. Crammer, and D. Schuurmans, "Robust Support Vector Machine Training via Convex Outlier Ablation," *Proc. 21st Nat'l Conf. Artificial Intelligence (AAAI '06),* 2006.
[15] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
[16] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge Univ. Press, 1988.
[17] T.G. Dietterich, "Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms," *Neural Computation,* vol. 10, no. 7, pp. 1895-1923, 1998.
[18] K. Tieu and P. Viola, "Boosting Image Retrieval," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '00),* pp. 228-235, 2000.
[19] L. Bo, L. Wang, and L. Jiao, "Feature Scaling for Kernel Fisher Discriminant Analysis Using Leave-One-Out Cross Validation," *Neural Computation,* vol. 18, no. 4, pp. 961-978, 2006.
[20] G. Cawley and N.L.C. Talbot, "Efficient Leave-One-Out Cross-Validation of Kernel Fisher Discriminant Classifiers," *Pattern Recognition,* vol. 36, no. 11, pp. 2585-2592, Nov. 2003.

Lei Wang received the BEng and MEng degrees from Southeast University, China, in 1996 and 1999, respectively, and the PhD degree from Nanyang Technological University, Singapore, in 2004. From 2003 to 2005, he was a research associate and research fellow in Nanyang Technological University. He joined the Department of Information Engineering, Research School of Information Sciences and Engineering (RSISE), Australian National University, in 2005 as a research fellow. His research interests include computer vision, information retrieval, and machine learning. He received an Australian Postdoctoral Fellowship from the Australian Research Council in 2007. He is a member of the IEEE and the IEEE Computer Society.