IEICE TRANS. INF. & SYST., VOL.E93–D, NO.12 DECEMBER 2010

PAPER

Optimal Gaussian Kernel Parameter Selection for SVM Classifier

Xu YANG[†a)], ***Student Member***, **HuiLin XIONG**[†], ***and*** **Xin YANG**[†], ***Nonmembers***

SUMMARY The performance of the kernel-based learning algorithms, such as SVM, depends heavily on the proper choice of the kernel parameter. It is desirable for the kernel machines to work on the optimal kernel parameter that adapts well to the input data and the learning tasks. In this paper, we present a novel method for selecting Gaussian kernel parameter by maximizing a class separability criterion, which measures the data distribution in the kernel-induced feature space, and is invariant under any non-singular linear transformation. The experimental results show that both the class separability of the data in the kernel-induced feature space and the classification performance of the SVM classifier are improved by using the optimal kernel parameter.

key words: *kernel optimization, model selection, kernel parameter selection, support vector machines, pattern recognition*

1. Introduction

The "kernel method" is well established as a way of nonlinear generalization of the linear machines, e.g. KPCA, KDA and SVM [1], by mapping the input data X into a high-dimensional feature space F, $\phi : X \rightarrow F$, where the linear machines perform. The map ϕ is implicitly presented by specifying a kernel function as the dot product between each pair of points in F. It is often assumed that the distribution of the mapped data in the feature space is more appropriate to the linear algorithms than that in the original space. However, this is not always the case, and a lot of researches [2]–[9] have shown that the performance of kernel based learning algorithms depends heavily on the kernel selection. It is desirable for the kernel machines to find a way that could automatically choose an optimal kernel.

The goal of kernel selection is to find a kernel which minimizes the generalization error of the corresponding classifier. Unfortunately, the error rate is not an explicit function of the kernel, so the problem of kernel selection is usually tackled by cross validation, which usually takes an exhaustive search on every possible values of a pre-defined discrete set. This may be time-consuming, and furthermore, the kernel selected in this way is usually sub-optimal. In literatures, the research concerning kernel selection focuses on approximating the error rate with an explicit continuous measure, so that the task could be considered within the framework of a tractable mathematic optimization problem. This is also known as the "kernel optimization" problem. In [2], the optimal kernel for KPCA is optimized by maximizing a quadratic cost function with respect to eigenvalues, which measures the variances deviation of the principal components. In [3], a measure, called kernel-alignment, is proposed to evaluate the degree of the agreement between a kernel matrix and the expected target kernel matrix, and the theoretical analysis proves that a high kernel alignment value corresponds to the low generalization error bound of a Parzen windows classifier. In [4], both the kernel alignment and the maximal margin measures are adopted and the kernel optimization is achieved by the Semi-Definite Programming technique. In [5], a more complicated measure, the radius-margin bound, is proposed and a gradient based method is used to select the optimal kernel parameters automatically for Support Vector Machine (SVM) classifier. Xiong et al. [6] firstly propose to optimize the kernel function by using a class separability measure. This measure is defined as the ratio between the trace of the between-class scatter matrix and that of the within-class scatter matrix, which corresponds to the J_4 criterion in [10]. This measure is also used latter by Yeung et al. [7] as the criterion to optimize the coefficients of a recombination kernel matrix, and further generalized by Chen et al. [8] to cope with the kernel optimization of the multimodally distributed data. Although convenient for mathematical handling the data in the high-dimensional feature space, the J_4 criterion has the disadvantage of being dependent on the coordinate system [10]. Recently, Sekiguchi et al. [9] select the optimal kernel parameter for KDA by maximizing the ratio of the within-class scatter matrix to the between-class scatter matrix. In their optimization scheme, they first project all the samples into a subspace of the feature space using the empirical kernel map [11], and then, evolve the optimization in this finite-dimensional sbuspace. They show that both the kernel optimization and the KDA learning can be expressed in an explicit form in this sub-space .

In this paper, we propose an alternative method to learn the Gaussian kernel parameter by maximizing the class separability of the mapped data in the kernel-induced feature space. This criterion is similar to the commonly used Fisher criterion [10], which is invariant under any non-singular linear transformation. We maximize the criterion by using a standard gradient descent approach, which is performed implicitly by specifying the inner product between each pair of data points rather than by giving their corrdinates explicitly. The Gaussian kernel learned by this method can implicitly

Manuscript received June 9, 2010.
Manuscript revised August 31, 2010.

[†]The authors are with the Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai 200240, China.
a) E-mail: feilang@foxmail.com

DOI: 10.1587/transinf.E93.D.3352

distribute the data in the feature space in a favorable way for the task of classification. Experimental results are encouraging in comparison with the cross-validation method.

The paper is organized as follows. In Sect. 2, the traditional scatter-matrix-based criterion is reviewed, and then we show how it is generalized in the kernel-induced feature space, and utilized efficiently as a measure to choose the Gaussian kernel parameter for SVM classifier. Section 3 presents the experimental results on the benchmark data sets. Finally, concluding remarks and future work are given in Sect. 4.

2. Proposed Kernel Optimization Method

SVM classifiers work by constructing a hyperplane in the kernel-induced feature space with the largest distance, called margin, between the nearest support vectors of different data classes. According to the Vapnik-Chervonenkis theory [1], the larger the margin is, the lower the generalization error of the classifier is. Since the kernel optimization using the Fisher criterion can make different data classes in the kernel-induced feature space well separated, that is with a relatively larger margin, the kernel parameter obtained by the Fisher criterion usually leads to the best performance of the SVM classifier.

2.1 Scatter-Matrix-Based Class Separability Criterion

Let X denote the input data set, which is a subset of R^d, and $Y = \{-1, +1\}$ the corresponding class labels. The input-output pairs (x_i, y_i), where $x_i \in \mathrm{X}$ and $y_i \in \mathrm{Y}$ $(i = 1, \ldots, n)$, constitute the whole sample set. The number of samples in class C_j (class label equals $l_1 = -1$ or $l_2 = +1$) is $n_j, (j = 1, 2)$, where $n = n_1 + n_2$. Let $m_j = \frac{1}{n_j} \sum_{y_i = l_j} x_i$ be the mean vector of class j and $m = \frac{1}{n} \sum_{i=1}^{n} x_i$ be the global mean vector. The between-class scatter matrix S_b and the within-class scatter matrix S_w are defined as

$$S_b = \frac{1}{n} \sum_{j=1}^{2} n_j \left(m_j - m\right)\left(m_j - m\right)^T \tag{1}$$

$$S_w = \frac{1}{n} \sum_{j=1}^{2} \sum\nolimits_{y_i = l_j} \left(x_i - m_j\right)\left(x_i - m_j\right)^T \tag{2}$$

We adopt the Fisher criterion $J_1 = Tr\left(S_w^{-1} S_b\right)$, where $Tr(\cdot)$ denotes the trace of a square matrix, to measure the class separability of the data. This measure is independent on the coordinate system, and invariant under any non-singular linear transformation.

2.2 Regularized Class Separability Criterion in the Kernel-Induced Feature Space

The kernel function $k_\gamma\left(x_i, x_j\right) = \left\langle \phi\left(x_i\right), \phi\left(x_j\right)\right\rangle$, where γ is the kernel parameter, determins the mapping $\phi(\cdot)$, and moreover, implicitly determins the distribution of the mapped data in the kernel-induced feature space. Therefore, we can derive the class separability criterion as the function of the kernel parameter. Replacing x by its image $\phi(x)$ in Eqs. (1) and (2), we define the between-class scatter matrix and within-class scatter matrix in the feature space as follows,

$$S_b^\phi = \frac{1}{n} \sum_{j=1}^{2} n_j \left(m_j^\phi - m^\phi\right)\left(m_j^\phi - m^\phi\right)^T \tag{3}$$

$$S_w^\phi = \frac{1}{n} \sum_{j=1}^{2} \sum\nolimits_{y_i = l_j} \left(\phi\left(x_i\right) - m_j^\phi\right)\left(\phi\left(x_i\right) - m_j^\phi\right)^T \tag{4}$$

where the superscript $^\phi$ is used to stress the variables are in the feature space F, rather than the input data space X.

Without loss of generality, let us assume that the first n_1 data belong to class C_1, that is, $y_i = -1, i \le n_1$, and the remaining n_2 data belong to class C_2, where $n = n_1 + n_2$. Let $\mathbf{1}_m$ be the $m-$dimensional vector whose all entries are equal to unity, $\mathbf{1}_m^{+k}\left(\mathbf{1}_m^{-k}\right)$ the $m-$dimensional vector whose first (last) k entries are equal to unity and the remaining entries are equal to 0, and the data matrix $\Phi = \left(\phi\left(x_1\right), \ldots, \phi\left(x_n\right)\right)$. Then, we have

$$m^\phi = \frac{1}{n} \sum_{i=1}^{n} \phi\left(x_i\right) = \frac{1}{n} \Phi \mathbf{1}_n \tag{5}$$

$$m_1^\phi = \frac{1}{n_1} \sum_{i=1}^{n_1} \phi\left(x_i\right) = \frac{1}{n_1} \Phi \mathbf{1}_n^{+n_1} \tag{6}$$

$$m_2^\phi = \frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \phi\left(x_i\right) = \frac{1}{n_2} \Phi \mathbf{1}_n^{-n_2} \tag{7}$$

The between-class scatter matrix can be formulated as follows:

$$\begin{aligned} S_b^\phi &= \frac{n_1}{n}\left(\frac{1}{n_1}\Phi\mathbf{1}_n^{+n_1} - \frac{1}{n}\Phi\mathbf{1}_n\right)\left(\frac{1}{n_1}\Phi\mathbf{1}_n^{+n_1} - \frac{1}{n}\Phi\mathbf{1}_n\right)^T \\ &\quad + \frac{n_2}{n}\left(\frac{1}{n_2}\Phi\mathbf{1}_n^{-n_2} - \frac{1}{n}\Phi\mathbf{1}_n\right)\left(\frac{1}{n_2}\Phi\mathbf{1}_n^{-n_2} - \frac{1}{n}\Phi\mathbf{1}_n\right)^T \\ &= \frac{n_1}{n}\Phi\left(\frac{1}{n_1}\mathbf{1}_n^{+n_1} - \frac{1}{n}\mathbf{1}_n\right)\left(\frac{1}{n_1}\mathbf{1}_n^{+n_1} - \frac{1}{n}\mathbf{1}_n\right)^T\Phi^T \\ &\quad + \frac{n_2}{n}\Phi\left(\frac{1}{n_2}\mathbf{1}_n^{-n_2} - \frac{1}{n}\mathbf{1}_n\right)\left(\frac{1}{n_2}\mathbf{1}_n^{-n_2} - \frac{1}{n}\mathbf{1}_n\right)^T\Phi^T \\ &= \Phi B \Phi^T \end{aligned} \tag{8}$$

where

$$\begin{aligned} B &= \frac{n_1}{n}\left(\frac{1}{n_1}\mathbf{1}_n^{+n_1} - \frac{1}{n}\mathbf{1}_n\right)\left(\frac{1}{n_1}\mathbf{1}_n^{+n_1} - \frac{1}{n}\mathbf{1}_n\right)^T \\ &\quad + \frac{n_2}{n}\left(\frac{1}{n_2}\mathbf{1}_n^{-n_2} - \frac{1}{n}\mathbf{1}_n\right)\left(\frac{1}{n_2}\mathbf{1}_n^{-n_2} - \frac{1}{n}\mathbf{1}_n\right)^T \end{aligned} \tag{9}$$

is a $n \times n$ constant symmetrical matrix. The within-class scatter matrix can be decomposed to two terms, and formulated as follows:

$$
\begin{aligned}
S_w^{\phi} &= \frac{1}{n}\left(\begin{array}{l}\sum_{i=1}^{n_1}\left(\phi(x_i)-m_1^{\phi}\right)\left(\phi(x_i)-m_1^{\phi}\right)^T \\ +\sum_{i=n_1+1}^{n}\left(\phi(x_i)-m_2^{\phi}\right)\left(\phi(x_i)-m_2^{\phi}\right)^T\end{array}\right) \\
&= \frac{1}{n}\left(\begin{array}{l}\sum_{i=1}^{n_1}\left(\phi(x_i)\phi(x_i)^T-m_1^{\phi}m_1^{\phi^T}\right) \\ +\sum_{i=n_1+1}^{n}\left(\phi(x_i)\phi(x_i)^T-m_2^{\phi}m_2^{\phi^T}\right)\end{array}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\phi(x_i)\phi(x_i)^T-n_1m_1^{\phi}m_1^{\phi^T}-n_2m_2^{\phi}m_2^{\phi^T}\right) \\
&= \frac{1}{n}\left(\Phi\Phi^T-\frac{1}{n_1}\Phi\mathbf{1}_n^{+n_1}\mathbf{1}_n^{+n_1T}\Phi^T-\frac{1}{n_2}\Phi\mathbf{1}_n^{-n_2}\mathbf{1}_n^{-n_2T}\Phi^T\right) \\
&= \frac{1}{n}\Phi\left(\mathrm{I}-\frac{1}{n_1}\mathbf{1}_n^{+n_1}\mathbf{1}_n^{+n_1T}-\frac{1}{n_2}\mathbf{1}_n^{-n_2}\mathbf{1}_n^{-n_2T}\right)\Phi^T \\
&= \Phi WW^T\Phi^T \qquad (10)
\end{aligned}
$$

where $W = \frac{1}{\sqrt{n}}\left(\mathrm{I}-\frac{1}{n_1}\mathbf{1}_n^{+n_1}\mathbf{1}_n^{+n_1T}-\frac{1}{n_2}\mathbf{1}_n^{-n_2}\mathbf{1}_n^{-n_2T}\right)$ is also a constant $n\times n$ symmetrical matrix.

The Fisher criterion $J_1 = Tr\left(S_w^{-1}S_b\right)$ requires in advance that the within-class scatter matrix S_w is nonsingular, which is usually not satisfied in practice, especially in the case when the sample size is smaller than the dimension of the sample data. The problem could become even worse as we use the criterion in the high-dimensional feature space. To address this problem, we adopt the regularization technique by adding a small term $\lambda\mathrm{I}$ to the within-class scatter matrix S_w^{ϕ}, where I is an identity matrix. Then, the regularized class separability criterion (RCSC) in the feature space is:

$$
\begin{aligned}
\tilde{J}_{reg}^{\phi} &= Tr\left(\left(\lambda\mathrm{I}+S_w^{\phi}\right)^{-1}S_b^{\phi}\right) \\
&= Tr\left(\left(\lambda\mathrm{I}+\Phi WW^T\Phi^T\right)^{-1}\Phi B\Phi^T\right) \qquad (11)
\end{aligned}
$$

Using the Woodbury formula,

$$(A+BC)^{-1} = A^{-1}-A^{-1}B(I+CA^{-1}B)^{-1}CA^{-1}$$

we obtain

$$\left(\lambda\mathrm{I}+\Phi WW^T\Phi^T\right)^{-1} = \frac{1}{\lambda}\left(\mathrm{I}-\Phi W\left(\lambda\mathrm{I}+W\Phi^T\Phi W\right)^{-1}W\Phi^T\right)$$

and the regularized class separability criterion can be reformulated as

$$
\begin{aligned}
\tilde{J}_{reg}^{\phi} &= Tr\left(\frac{1}{\lambda}\left(\Phi B\Phi^T-\Phi W\left(\lambda\mathrm{I}+W\Phi^T\Phi W\right)^{-1}W\Phi^T\Phi B\Phi^T\right)\right) \\
&= \frac{1}{\lambda}Tr\left(\Phi B\Phi^T-\Phi W(\lambda\mathrm{I}+WKW)^{-1}WKB\Phi^T\right) \\
&= \frac{1}{\lambda}Tr\left(\Phi B\Phi^T-\Phi\mathrm{A}\Phi^T\right) \\
&= \frac{1}{\lambda}Tr\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(b_{ij}-a_{ij}\right)\phi(x_i)\phi\left(x_j\right)^T\right) \\
&= \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(b_{ij}-a_{ij}\right)Tr\left(\phi(x_i)\phi\left(x_j\right)^T\right) \\
&= \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(b_{ij}-a_{ij}\right)k\left(x_i,x_j\right) \\
&= \frac{1}{\lambda}\mathbf{1}_n^T\left((B-\mathrm{A}).*K\right)\mathbf{1}_n \qquad (12)
\end{aligned}
$$

where ".*" denotes the element-by-element multiplication operation of two matrix, $K = \left[k_\gamma\left(x_i,x_j\right)\right]_{n\times n}$ is the kernel matrix, and $\mathrm{A} = W(\lambda\mathrm{I}+WKW)^{-1}WKB = \left[a_{ij}(\gamma)\right]_{n\times n}$ is a matrix, whose entries are also determined by the kernel parameter γ. As we can see that the $\tilde{J}_{reg}^{\phi}$ criterion is still conceptually simple and computationally efficient by incorporating the kernel trick.

2.3 Gaussian Kernel Parameter Selection by Maximizing the RCSC

We use the Gaussian function $k_{ij} = \exp\left(-\gamma\left\|x_i-x_j\right\|^2\right)$ as our kernel in this paper, where γ is the kernel parameter to be optimized. The optimal value γ_{opt} of the kernel parameter can be obtained through maximizing Eq. (12), i.e.

$$\gamma_{opt} = \arg\min_{\gamma}\tilde{J}_{reg}^{\phi}(\gamma) \qquad (13)$$

Unfortunately, this optimization problem can not be solved analytically from $\partial\tilde{J}_{reg}^{\phi}/\partial\gamma = 0$. In this paper, we employ the standard gradient algorithm to approximate the value of the optimal γ. The updating equation for maximizing the class separability criterion $\tilde{J}_{reg}^{\phi}$ is given by

$$\gamma^{(t+1)} = \gamma^{(t)}+\eta\left(\partial\tilde{J}_{reg}^{\phi}/\partial\gamma\right) \qquad (14)$$

where η is the learning rate. To ensure the convergence of the algorithm, a gradually decreasing learning rate is adopted

$$\eta(t) = \eta_0\left(1-\frac{t}{N}\right) \qquad (15)$$

where η_0 is the initial learning rate, N denotes a prespecified number of iterations, and t represents the current iteration number. Since B is a constant matrix, containing the label information of the samples, differentiating both sides of Eq. (12), we have

$$\frac{\partial\tilde{J}_{reg}^{\phi}(\gamma)}{\partial\gamma} = \frac{1}{\lambda}\mathbf{1}_n^T\left((B-\mathrm{A}).*K'-\mathrm{A}'.*K\right)\mathbf{1}_n \qquad (16)$$

where $K' = \left[\frac{\partial k_{ij}}{\partial\gamma}\right]_{n\times n} = \left[-k_{ij}\left\|x_i-x_j\right\|^2\right]_{n\times n}$, and $\mathrm{A}' = \left[\frac{\partial a_{ij}}{\partial\gamma}\right]_{n\times n}$. Let $C(\gamma)$ denote $\lambda\mathrm{I}+WKW$, then $\partial C/\partial\gamma = WK'W$ and the inverse of matrix C can be calculated using the Woodbury formula again. Since $\partial C^{-1}/\partial\gamma = -C^{-1}(\partial C/\partial\gamma)C^{-1}$, the differential of matrix A to γ can be expressed by

$$\mathrm{A}' = -WC^{-1}WK'WC^{-1}WKB+WC^{-1}WK'B$$

Now we can summarize our kernel optimization algorithm as follows.

1. Group the training samples according to their class labels. Calculate W and B.
2. Initialize λ, $\gamma^{(0)}$ and set the iteration number $t = 0$.
3. Calculate K and K' first, then A and A', and then $\partial \tilde{J}^{\phi}_{reg} \big/ \partial \gamma$.
4. Update the kernel parameter

$$\gamma^{(t+1)} = \gamma^{(t)} + \eta(t)\left(\partial \tilde{J}^{\phi}_{reg} \big/ \partial \gamma\right)$$

5. If t reaches a pre-specified number N, stop. Otherwise, set $t = t + 1$, and go back to step 3.

3. Experiments

In this section, we conduct three sets of experiments to investigate the effectiveness of using the proposed method to select the optimal Gaussian kernel parameter for the SVM classifier. In the first set of experiments, we study if the kernel parameters optimized by the regularized class separability criterion $\tilde{J}^{\phi}_{reg}$ match with the kernel parameters that lead to the best performances of SVM classifier. The second set of experiments examines the performance of the gradient-based iterative algorithm to maximize $\tilde{J}^{\phi}_{reg}$. The final set of experiments compares the performances of the proposed method and the cross validation method in kernel parameter selection. Seven real data sets, namely, the *Ionosphere*, *Wisconsin Breast Cancer*, *Sonar*, *Pima Indians diabetes*, *Liver disorder*, *Heart disease* and *Australian credit approval*, are collected from the UCI machine learning benchmark repository [12] to test our algorithm. These seven data sets are chosen, since they present different degrees of difficulty from the point of view of data classification. Table 1 presents some basic information about these data sets.

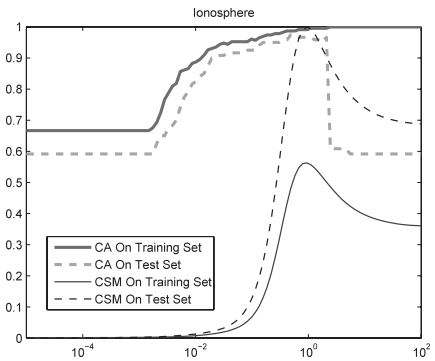
3.1 Experiment 1: The Effectiveness of $\tilde{J}^{\phi}_{reg}(\gamma)$

To show the effectiveness of the class separability criterion, $\tilde{J}^{\phi}_{reg}(\gamma)$, we illustrate the performance of SVM classifier and the values of $\tilde{J}^{\phi}_{reg}(\gamma)$, when different kernel parameters are used. Three real data sets, namely, *Ionosphere*, *Wisconsin Breast Cancer*, *Sonar*, are used. Each data set is first normalized to a distribution with zero mean and unit variance, and then randomly partitioned into two disjoint parts, consisting of $2/3$ and $1/3$ of the entire samples, respectively. The former is used as the training set, and the other as the test set. The regularization constant λ of $\tilde{J}^{\phi}_{reg}(\gamma)$ is set to 10^{-5}, and 100 different kernel parameters γ are uniformly sampled from the interval of $\left[10^{-5}, 10^{2}\right]$ on the logarithm scale. Figure 1 shows the values of the regularized class separability criterion (RCSC) and the classification accuracy (CA) of SVM, whose regularization constant is set to 1, over different kernel parameter γ values on the three data sets. For the convenience of comparisons, we normalize the values of $\tilde{J}^{\phi}_{reg}(\gamma)$ to $[0, 1]$.

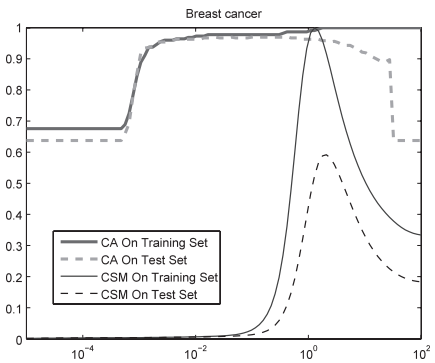
Table 1 Characteristics of datasets in the UCI repository.

Dataset	#Samples	#Dimensions
Ionosphere	351	34
Breast	683	10
Sonar	208	60
Pima	768	8
Liver	345	6
Heart	270	13
Australian	690	14

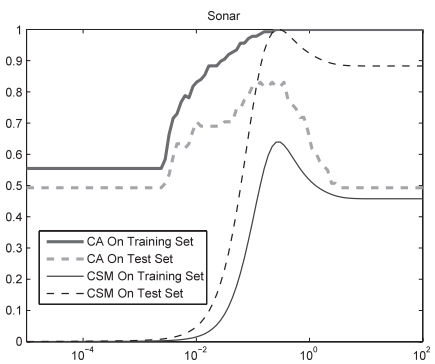
From Fig. 1, we see that the class separability of the test data varies in the same manner as that of the training data, and the maxima of $\tilde{J}^{\phi}_{reg}(\gamma)$ on both training set and test set are almost identical. Since both the training and test data are from the same probability distribution, optimizing the kernel parameter on the training data, or increasing the class sepa-



(a) Ionosphere



(b) Breast Cancer



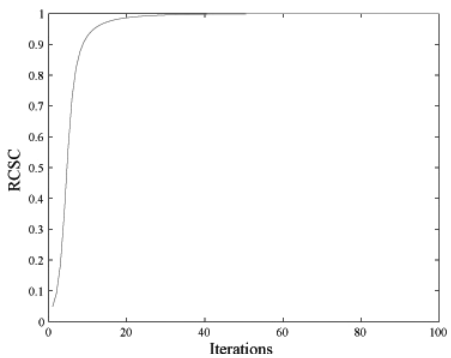
(c) Sonar

Fig. 1 Evaluation of the effectiveness of $\tilde{J}^{\phi}_{reg}(\gamma)$.

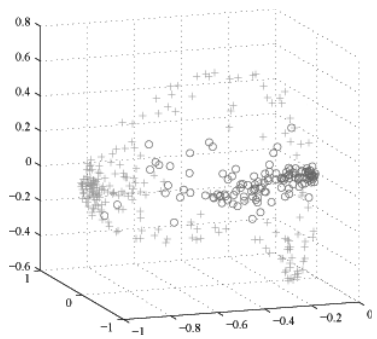
rability of the training data in the feature space, should lead to a similar effect on the test data. It can also be observed that the highest classification accuracies on the test set always coincide with the best class separability of the training data in the same position, where the classification accuracy on the training set just approach 100%. Therefore, using the optimal kernel parameter selected from the training data, the SVM achieves the highest classification accuracy on test set. This indicates that the maximization of $\tilde{J}_{reg}^{\phi}(\gamma)$ on the training set can result in a good kernel parameter selection for the SVM classifier.

3.2 Experiment 2: The Gradient-Based Kernel Optimization

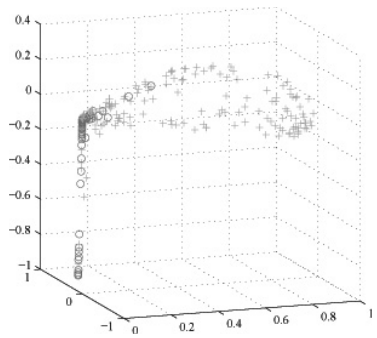
To show the performance of the gradient-based kernel optimization algorithm, we test it on the same 3 data sets as used in Sect. 3.1. The initial learning rate in Eq. (15) and the iteration number N are set to 10^{-4} and 100, respectively. The regularization term λ of $\tilde{J}_{reg}^{\phi}(\gamma)$ is always fixed at 10^{-6} and the initial value for the kernel parameter γ is set to $\gamma^{(0)} = n \Big/ \sum_{i=1}^{n} \|x_i - \bar{x}\|^2$, where $\bar{x}$ is the centroid of the whole data set. The values of the class separability on the training data of the three data sets are shown in Fig. 2 (a), 3 (a) and 4 (a), respectively. It is seen that the class separability of the data sets in the feature space can be improved substantially along with the iterations. Besides, by projecting the data onto its first three significant dimensions, Fig. 2 (b), 3 (b), 4 (b) and Fig. 2 (c), 3 (c), 4 (c) visualize the spatial distributions of the data before and after the kernel optimization, respectively. From these figures, the improve-



(a) $\tilde{J}_{reg}^{\phi}(\gamma)$ as a function of the number of iterations.

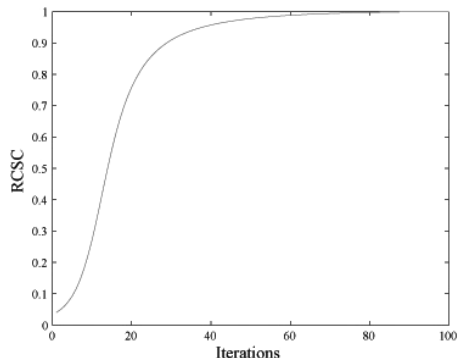


(b) 3-D embedding results with the initial kernel.

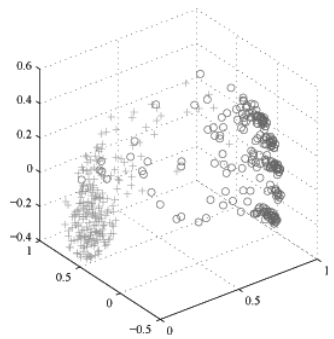


(c) 3-D embedding results with the optimized kernel.

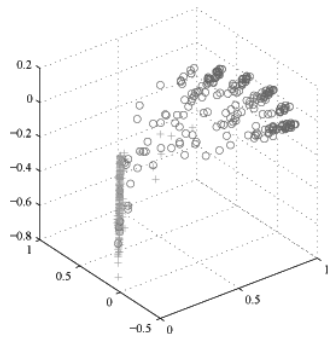
Fig. 2 Optimizing $\tilde{J}_{reg}^{\phi}(\gamma)$ on the *Ionosphere* data set.



(a) $\tilde{J}_{reg}^{\phi}(\gamma)$ as a function of the number of iterations.

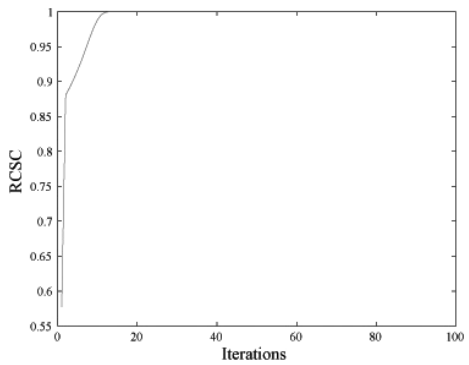


(b) 3-D embedding results with the initial kernel.

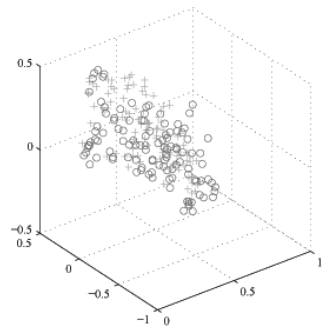


(c) 3-D embedding results with the optimized kernel.

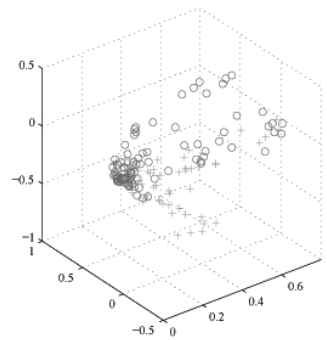
Fig. 3 Optimizing $\tilde{J}_{reg}^{\phi}(\gamma)$ on the *Breast Cancer* data set.



(a) $\tilde{J}^{\phi}_{reg}(\gamma)$ as a function of the number of iterations.



(b) 3-D projection by the initial kernel before optimization.



(c) 3-D projection by the optimized kernel.

Fig. 4 Optimizing $\tilde{J}^{\phi}_{reg}(\gamma)$ on the *Sonar* data set.

ment of the class separability can also be observed.

3.3 Experiment 3: Comparison with the Cross-Validation Method

In this experiment, we demonstrate the effectiveness of the proposed kernel parameter selection method by comparing with the commonly used cross-validation method. Eight real data sets in Table 1 are used in the experiment. Each data set is randomly partitioned into a training set and a test set as before. We use $2/3$ of the total samples as the training set to perform the kernel parameter selection. For the proposed method, the regularization constant λ is fixed to 10^{-5}, the initial learning rate is set to 10^{-4}, and the iteration number N is set to 100. The initial value for the kernel parameter γ is determined in the same way as that in Sect. 3.2. For the cross-validation method, the training data set is further divided into 10 equal subsets. Of the 10 subsets, one is retained as the validation set and the remaining subsets are used to train SVM with the kernel parameter choosing from $\{1\times10^{-5}, 2\times10^{-5}, \ldots, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. This process is repeated 10 times, so that each of the 10 subsets is used exactly once as the validation data. The 10 results of kernel parameter selection are averaged to produce a final estimation. The classification accuracies of the SVM classifier with kernel parameters selected by the two methods are compared to evaluate the performances. The average experimental results on the test data set over 20 trials are shown in Table 2. As what we can see, the optimal kernel parameters selected by maximizing the class separability criterion achieve better classification results in most cases than the traditional cross-validation method.

Table 2 Classification accuracies with kernels got by the proposed method K_{OPT} and the cross-validation method K_{CV}.

Dataset	K_{OPT}	K_{CV}
Ionosphere	95.81%	95.31%
Breast	97.69%	97.11%
Sonar	86.34%	84.28%
Pima	77.91%	76.86%
Liver	69.81%	67.68%
Heart	84.25%	83.79%
Australian	85.27%	85.35%

4. Conclusion

In this paper, a method for learning the optimal Gaussian kernel parameter is presented. We evaluate the kernel parameter by measuring the corresponding class separability of the mapped data in the kernel-induced feature space. A gradient-based optimization method is adopted to maximize the class separability criterion and to find the good parameter. Experimental results on real data sets show that the class separability of the data in the feature space is improved greatly by using the optimized kernel parameter, and the corresponding classification performance of the SVM classifier outperforms that of the cross-validation method in most cases. The method proposed in this paper can also be directly generalized to the other kinds of kernel parameter selection, provided that the kernel functions are differentiable to the parameter, and thus it provides a promising alternative for the traditional cross-validation method. Currently, the regularization constant λ is determined empirically, so we will explore the possibility of formulating another optimization procedure to find the optimal value of λ in our future research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.60775008), and the National High Technology Research and Development Program (863 Program) of China (Grant No.2007AA01Z196).

References

[1] B. Schölkopf and A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, The MIT Press, 2002.

[2] T. Nogayama, H. Takahashi, and M. Muramatsu, "Generalization of kernel PCA and automatic parameter tuning," 8th Australian and New Zealand Intelligent Information Systems Conference, Macquarie University, pp.173–178, Sydney, Australia, 2003.

[3] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel target alignment," in Advances in Neural Information Processing Systems 14, pp.367–373, Cambridge, MA, 2002.

[4] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan, "Learning the kernel matrix with semidefinite programming," J. Machine Learning Research, vol.5, pp.27–72, 2004.

[5] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," Mach. Learn., vol.46, no.1, pp.131–159, 2002.

[6] H. Xiong, M.N.S. Swamy, and M.O. Ahmad, "Optimizing the kernel in the empirical feature space," IEEE Trans. Neural Netw., vol.16, no.2, pp.460–474, 2005.

[7] D.Y. Yeung, H. Chang, and G. Dai, "Learning the kernel matrix by maximizing a KFD-based class separability criterion," Pattern Recognit., vol.40, no.7, pp.2021–2028, 2007.

[8] B. Chen, H. Liu, and Z. Bao, "A kernel optimization method based on the localized kernel fisher criterion," Pattern Recognit., vol.41, no.3, pp.1098–1109, 2008.

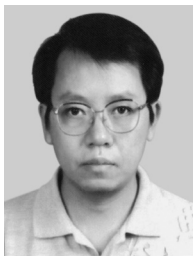
[9] R. Sekiguchi, H. Takahashi, and K. Hotta, "The automatic parameter tuning for multi-class learning with KDA," 2010 International Workshop on Nonlinear Circuits, Communication and Signal Processing, pp.190–193, Waikiki, Hawaii, 2010.

[10] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.

[11] B. Schölkopf, S. Mika, C.J.C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A.J. Smola, "Input space versus feature space in kernel-based methods," IEEE Trans. Neural Netw., vol.10, no.5, pp.1000–1017, 1999.

[12] A. Asuncion and D.J. Newman, "UCI Machine Learning Repository," Irvine, CA: University of California, School of Information and Computer Science, http://www.ics.uci.edu/ mlearn/ MLRepository.html, accessed 2007.

Xu Yang received the B.E. and M.E. degrees in Information and Control Engineering from Liaoning Shihua University, Fushun, China, in 2001 and 2004, respectively. He is currently a Ph.D. student in the Institute of Image Processing and Pattern Recognition at Shanghai Jiao Tong University Shanghai, China. His research interests include machine learning and medical image analysis.



HuiLin Xiong received the B.Sc. and M.Sc. degrees in Mathematics from Wuhan University, Wuhan, China, in 1985 and 1988, respectively. He received his Ph.D. degree in Pattern Recognition and Intelligent Control from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1999. He joined Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2007, and currently, he is a professor in the Department of Automation of SJTU. His research interests include kernel-based nonlinear pattern recognition, machine learning, and bioinfomatics.

Xin Yang received the M.Sc. degree in control engineering from Northwestern Polytechnic University, Xian, China, in 1982 and the Ph.D. degree of Applied Science degree in electronic engineering from the Free University of Brussels (ETRO/VUB) in 1995. Since 1997, he has been with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His current research activities are in the area of medical image analysis and partial differential equations in image processing. Dr. Yang is the author of more than 70 papers in refereed journals and conference proceedings.