Kernel discriminant analysis based feature selection

Ishii, Tsuneyoshi

Ashihara, Masamichi

Abe, Shigeo

(Citation)
Neurocomputing,71(13-15):2544-2552

(Issue Date)
2008-08

(Resource Type)
journal article

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/90000924

Kernel Discriminant Analysis Based Feature Selection

Tsuneyoshi Ishii, Masamichi Ashihara, and Shigeo Abe
Graduate School of Engineering
Kobe University
Rokkodai, Nada, Kobe, Japan

May 21, 2008

Abstract

For two-class problems we propose two feature selection criteria based on kernel discriminant analysis (KDA). The first one is the objective function of kernel discriminant analysis called the KDA criterion. We show that the KDA criterion is monotonic for the deletion of features, which ensures stable feature selection. The second one is the recognition rate obtained by a KDA classifier, called the KDA-based recognition rate, which is defined in the one-dimensional space obtained by KDA. Namely, a conditional probability of a datum for a given class is calculated and the datum is classified into the class with the maximum conditional probability. To ensure stable feature selection, we evaluate the KDA-based recognition rate by cross-validation. By computer experiments we compare the two criteria for two-class problems and the recognition rate of the support vector Machine (SVM) evaluated by cross-validation, called the SVM-based recognition rate. The selection performance of the KDA criterion and the KDA-based recognition rate is comparable and is better than that by the SVM-based recognition rate.

Keywords: Feature Selection, Kernel Discriminant Analysis, Support Vector Machines.

1 Introduction

Feature selection is to select from the original set of features, namely, input variables, the minimum subset of features that realizes the maximum generalization ability [1, 2]. To realize this, during the process of feature selection, the generalization ability of a subset of features needs to be estimated. This type of feature selection is called a wrapper method [3]. But since it is time-consuming to directly estimate the generalization ability, some selection criterion, which is considered to well reflect the generalization ability, is used. This method is called a filter method and various selection criteria have been developed [4, 5].

The forward or backward selection method using a selection criterion is widely used. In backward selection, we start from all the features and delete one feature at a time, which deteriorates the selection criterion the least. We delete features until the selection criterion reaches a specified value. In forward

selection, we start from an empty set of features and add one feature at a time, which improves the selection criterion the most. We iterate this procedure until the selection criterion reaches a specified value. Because forward or backward selection is slow, we may add or delete more than one feature at a time based on feature ranking, or we may combine backward and forward selection.

Because these selection methods are local optimization techniques, global optimality of feature selection is not guaranteed. Usually, backward selection is slower but is more stable in selecting optimal features than forward selection [6]. If a selection criterion is monotonic for deletion or addition of a feature, we can terminate feature selection when the selection criterion violates a predefined value [7] or we can use optimization techniques such as the branch-and-bound technique. An exception ratio defined based on the overlap of class regions approximated by hyperboxes [7] is proved to be monotonic for the deletion of features. But the exception ratio defined in the feature space is not monotonic [8].

By the introduction of support vector machines (SVMs), various selection methods suitable for support vector machines have been developed. The selection criterion for filter methods used in the literature is, except for some cases [8, 9, 10], the margin [2, 11, 12, 13, 14]. In addition, in most cases, a linear support vector machine is used. The idea of feature selection is as follows: If some elements of the coefficient vector of the hyperplane are zero, the deletion of the associated input variables does not change the optimal hyperplane for the remaining variables. But if we delete variables associated with nonzero elements, the optimal solution changes. Thus the magnitude of the margin decreases. In [15], selection of features in support vector machines with polynomial kernels is discussed, but this is for deletion of feature space variables, not input variables. In [8], the objective function of kernel discriminant analysis called the KDA criterion, namely the ratio of the between-class scatter and within-class scatter, is proved to be monotonic for the deletion of features, and feature selection based on the KDA criterion was shown to be robust for benchmark data sets.

As a wrapper method, in [16, 17], block deletion of features in backward feature selection is proposed using the generalization ability by cross-validation as the selection criterion.

In addition to filter and wrapper methods, the embedded methods combine training and feature selection; because training of support vector machines results in solving a quadratic optimization problem, feature selection can be done by modifying the objective function [18, 19, 20, 21].

In this paper, for two-class problems, in addition to the KDA criterion in [8], which is one of the filter methods, we propose using the recognition rate of the KDA classifier evaluated by cross-validation, called the KDA-based recognition rate. In the KDA classifier, a conditional probability of a datum for a given class is calculated and the datum is classified into the class with the maximum conditional probability. Here, the conditional probability is calculated assuming that the data of each class obey a normal distribution.

The feature selection is done by backward selection. We start from all the features. We temporarily delete one feature, calculate the selection criterion, and delete the feature that improves the selection criterion the most. This process is iterated until the stopping condition is satisfied.

By computer experiments we evaluate the two feature selection criteria as well as the SVM-based recognition rate evaluated by cross-validation, from the

standpoints of the number of deleted features that do not deteriorate the generalization ability and the validity of the stopping conditions of the selection criteria.

In Sections 2 and 3, we summarize SVMs and KDA and in Section 4, we discuss two selection criteria and their monotonicity. In Section 5, we explain backward feature selection used and in Section 6 we demonstrate the validity of the proposed methods by computer experiments.

2 Support Vector Machines

In this section we summarize two-class support vector machines [1, 22, 23].

Let m-dimensional inputs $\mathbf{x}_i\,(i = 1, \ldots, M)$ belong to Class 1 or 2 and the associated labels be $y_i = 1$ for Class 1 and -1 for Class 2. To enhance separability, the input space is mapped into the high-dimensional dot-product space called the feature space. Let the l-dimensional mapping function be $\mathbf{g}(\mathbf{x})$. If the dot product in the feature space is expressed by $H(\mathbf{x}, \mathbf{x}') = \mathbf{g}^T(\mathbf{x})\mathbf{g}(\mathbf{x})$, $H(\mathbf{x}, \mathbf{x}')$ is called the kernel function, and we do not need to explicitly treat the feature space. In the following we use polynomial kernels with degree d:

$$H(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + 1)^d, \tag{1}$$

and RBF kernels with positive parameter γ:

$$H(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2). \tag{2}$$

Let the decision function in the feature space be

$$D(\mathbf{x}) = \mathbf{w}^T\,\mathbf{g}(\mathbf{x}) + b, \tag{3}$$

where $\mathbf{w}$ is an l-dimensional vector in the feature space, b is a scalar, and

$$y_i\,D(\mathbf{x}_i) \geq 1 - \xi_i \quad \text{for} \quad i = 1, \ldots, M. \tag{4}$$

Here ξ_i are nonnegative slack variables.

The distance between the separating hyperplane $D(\mathbf{x}) = 0$ and the training datum, with $\xi_i = 0$, nearest to the hyperplane is called the margin. The hyperplane $D(\mathbf{x}) = 0$ with the maximum margin is called the optimal separating hyperplane.

To determine the optimal separating hyperplane, we minimize

$$\frac{1}{2}\,\|\mathbf{w}\|^2 + C\sum_{i=1}^{M}\xi_i \tag{5}$$

subject to the constraints

$$y_i\,(\mathbf{w}^T\,\mathbf{g}(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad \text{for} \quad i = 1, \ldots, M, \tag{6}$$

where C is the margin parameter that determines the tradeoff between the maximization of the margin and minimization of the classification error.

If the dimension of the feature space is very large, we convert the original problem into the dual problem. Introducing the nonnegative Lagrange multipliers, we obtain the following dual problem. Maximize

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \, \alpha_j \, y_i \, y_j \, H(\mathbf{x}_i, \, \mathbf{x}_j) \tag{7}$$

subject to the constraints

$$\sum_{i=1}^{M} y_i \, \alpha_i = 0, \quad C \geq \alpha_i \geq 0 \quad \text{for} \quad i = 1, \ldots, M, \tag{8}$$

where $\alpha_i (i = 1, \ldots, M)$ are dual variables associated with $\mathbf{x}_i$.

3 Kernel Discriminant Analysis

In this section we summarize kernel discriminant analysis, which finds the component that maximally separates two classes in the feature space [1, 24, 25, 26].

Here we redefine the training data to make the definition of KDA simpler. Let the sets of m-dimensional data belong to Class i $(i = 1, 2)$ be $\{\mathbf{x}_1^i, \ldots, \mathbf{x}_{M_i}^i\}$, where M_i is the number of data belonging to Class i, and data $\mathbf{x}$ be mapped into the l-dimensional feature space by the mapping function $\mathbf{g}(\mathbf{x})$. Now we find the l-dimensional vector $\mathbf{w}$, in which the two classes are separated maximally in the direction of $\mathbf{w}$ in the feature space.

The projection of $\mathbf{g}(\mathbf{x})$ on $\mathbf{w}$ is $\mathbf{w}^T \, \mathbf{g}(\mathbf{x}) / \|\mathbf{w}\|$. In the following we assume that $\|\mathbf{w}\| = 1$. We find such $\mathbf{w}$ that maximizes the difference of the centers, and minimizes the variances, of the projected data.

The square difference of the centers of the projected data, d^2, is

$$d^2 = (\mathbf{w}^T (\mathbf{c}_1 - \mathbf{c}_2))^2 = \mathbf{w}^T (\mathbf{c}_1 - \mathbf{c}_2) \, (\mathbf{c}_1 - \mathbf{c}_2)^T \mathbf{w}, \tag{9}$$

where $\mathbf{c}_i$ are the centers of class i data:

$$\mathbf{c}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} \mathbf{g}(\mathbf{x}_j^i) \quad \text{for} \quad i = 1, 2. \tag{10}$$

We define

$$Q_B = (\mathbf{c}_1 - \mathbf{c}_2) \, (\mathbf{c}_1 - \mathbf{c}_2)^T \tag{11}$$

and call Q_B the between-class scatter matrix.

The variances of the projected data, s_i^2, are

$$s_i^2 = \mathbf{w}^T \, Q_i \, \mathbf{w} \qquad \text{for} \quad i = 1, 2, \tag{12}$$

where

$$Q_i = \frac{1}{M_i} \, (\mathbf{g}(\mathbf{x}_1^i), \ldots, \mathbf{g}(\mathbf{x}_{M_i}^i)) \, (I_{M_i} - \mathbf{1}_{M_i}) \begin{pmatrix} \mathbf{g}^T(\mathbf{x}_1^i) \\ \vdots \\ \mathbf{g}^T(\mathbf{x}_{M_i}^i) \end{pmatrix} \quad \text{for} \quad i = 1, 2. \tag{13}$$

Here, I_{M_i} is the $M_i \times M_i$ unit matrix and $\mathbf{1}_{M_i}$ is the $M_i \times M_i$ matrix with all elements being $1/M_i$. We define

$$Q_W = Q_1 + Q_2 \tag{14}$$

and call Q_W the within-class scatter matrix.

Now, we want to maximize

$$J(\mathbf{w}) = \frac{d^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T Q_B \mathbf{w}}{\mathbf{w}^T Q_W \mathbf{w}}, \tag{15}$$

but since $\mathbf{w}$, Q_B, and Q_W are defined in the feature space, we need to use kernel tricks. Assume that a set of M' vectors $\{\mathbf{g}(\mathbf{y}_1), \ldots, \mathbf{g}(\mathbf{y}_{M'})\}$ spans the space generated by $\{\mathbf{g}(\mathbf{x}_1^1), \ldots, \mathbf{g}(\mathbf{x}_{M_1}^1), \mathbf{g}(\mathbf{x}_1^2), \ldots, \mathbf{g}(\mathbf{x}_{M_2}^2)\}$, where $\{\mathbf{y}_1, \ldots, \mathbf{y}_{M'}\} \subset \{\mathbf{x}_1^1, \ldots, \mathbf{x}_{M_1}^1, \mathbf{x}_1^2, \ldots, \mathbf{x}_{M_2}^2\}$ and $M' \leq M_1 + M_2$.

We use the Cholesky factorization of the kernel matrix H in selecting independent vectors [1]. Let H be positive definite. Then H is decomposed by the Cholesky factorization into

$$H = L\, L^T, \tag{16}$$

where L is the regular lower triangular matrix and each element L_{ij} is given by

$$L_{op} = \frac{H_{op} - \sum_{n=1}^{p-1} L_{pn}\, L_{on}}{L_{pp}} \quad \text{for} \quad o = 1, \ldots, M, \quad p = 1, \ldots, o-1, \tag{17}$$

$$L_{aa} = \sqrt{H_{aa} - \sum_{n=1}^{a-1} L_{an}^2} \quad \text{for} \quad a = 1, \ldots, M. \tag{18}$$

Here, $H_{ij} = H(\mathbf{x}_i, \mathbf{x}_j)$.

Then during the Cholesky factorization, if the diagonal element is smaller than the prescribed value $\eta\,(> 0)$:

$$H_{aa} - \sum_{n=1}^{a-1} L_{an}^2 \leq \eta, \tag{19}$$

we delete the associated row and column and continue decomposing the matrix. The training data that are not deleted in the Cholesky factorization are independent.

Using the independent vectors, $\mathbf{w}$ is expressed as

$$\mathbf{w} = (\mathbf{g}(\mathbf{y}_1), \ldots, \mathbf{g}(\mathbf{y}_{M'}))\, \boldsymbol{\alpha}, \tag{20}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{M'})^T$ and $\alpha_1, \ldots, \alpha_{M'}$ are scalars. Substituting (20) into (15), we obtain

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T K_B\, \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T K_W\, \boldsymbol{\alpha}}, \tag{21}$$

where

$$K_B = (\mathbf{k}_{B_1} - \mathbf{k}_{B_2})\, (\mathbf{k}_{B_1} - \mathbf{k}_{B_2})^T, \tag{22}$$

$$\mathbf{k}_{Bi} = \begin{pmatrix} \frac{1}{M_i}\sum_{j=1}^{M_i} H(\mathbf{y}_1, \mathbf{x}_j^i) \\ \cdots \\ \frac{1}{M_i}\sum_{j=1}^{M_i} H(\mathbf{y}_{M'}, \mathbf{x}_j^i) \end{pmatrix} \quad \text{for} \quad i = 1, 2, \tag{23}$$

$$K_W = K_{W_1} + K_{W_2}, \tag{24}$$

$$K_{W_i} = \frac{1}{M_i} \begin{pmatrix} H(\mathbf{y}_1, \mathbf{x}_1^i) \cdots H(\mathbf{y}_1, \mathbf{x}_{M_i}^i) \\ \cdots \\ H(\mathbf{y}_{M'}, \mathbf{x}_1^i) \cdots H(\mathbf{y}_{M'}, \mathbf{x}_{M_i}^i) \end{pmatrix} (I_{M_i} - \mathbf{1}_{M_i}) \times \begin{pmatrix} H(\mathbf{y}_1, \mathbf{x}_1^i) \cdots H(\mathbf{y}_1, \mathbf{x}_{M_i}^i) \\ \cdots \\ H(\mathbf{y}_{M'}, \mathbf{x}_1^i) \cdots H(\mathbf{y}_{M'}, \mathbf{x}_{M_i}^i) \end{pmatrix}^T \quad \text{for} \quad i = 1, 2. \tag{25}$$

Taking the partial derivative of (21) with respect to $\mathbf{w}$ and equating the resulting equation to zero, we obtain the following generalized eigenvalue problem:

$$K_B\, \boldsymbol{\alpha} = \lambda\, K_W\, \boldsymbol{\alpha}, \tag{26}$$

where λ is a generalized eigenvalue.

Substituting

$$K_W\, \boldsymbol{\alpha} = \mathbf{k}_{B_1} - \mathbf{k}_{B_2} \tag{27}$$

into the left-hand side of (26), we obtain

$$(\boldsymbol{\alpha}^T\, K_W\, \boldsymbol{\alpha})\, K_W\, \boldsymbol{\alpha}. \tag{28}$$

Thus, by letting $\lambda = \boldsymbol{\alpha}^T\, K_W\, \boldsymbol{\alpha}$, (27) is a solution of (26).

Since K_{W_1} and K_{W_2} are positive semi-definite, K_W is positive semi-definite. If K_W is positive definite, $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha} = K_W^{-1}\, (\mathbf{k}_{B_1} - \mathbf{k}_{B_2}). \tag{29}$$

Even if we choose independent vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{M'}$, for non-linear kernels, K_W may be positive semi-definite, i.e., singular. One way to overcome singularity is to add positive values to the diagonal elements [24]:

$$\boldsymbol{\alpha} = (K_W + \varepsilon I)^{-1}\, (\mathbf{k}_{B_1} - \mathbf{k}_{B_2}), \tag{30}$$

where ε is a small positive parameter.

Now, from (20) the projection of $\mathbf{g}(\mathbf{x})$ on $\mathbf{w}$, $\mathbf{w}^T\mathbf{g}(\mathbf{x})$, is calculated as follows:

$$\begin{aligned} \mathbf{w}^T\mathbf{g}(\mathbf{x}) &= \mathbf{g}^T(\mathbf{x})\mathbf{w} \\ &= \mathbf{g}^T(\mathbf{x})(\mathbf{g}(\mathbf{y}_1), \ldots, \mathbf{g}(\mathbf{y}_{M'}))\, \boldsymbol{\alpha} \\ &= (H(\mathbf{y}_1, \mathbf{x}), \ldots, H(\mathbf{y}_{M'}, \mathbf{x}))\, \boldsymbol{\alpha}. \end{aligned} \tag{31}$$

We define $p_r(\mathbf{x}) = \mathbf{w}^T\mathbf{g}(\mathbf{x})$ and call the one-dimensional space KDA space.

4 Selection Criteria and Their Monotonicity

4.1 KDA Criterion

The first selection criterion is the value of (15) for optimum $\mathbf{w}$. We call this KDA criterion. The KDA criterion with linear kernels, i.e., the LDA criterion is often used for a feature selection criterion but its monotonicity for deletion of features is not known.

We can easily prove that the KDA criterion is monotonic for the deletion of features [8]. Let $\mathbf{x}^i$ be the m-dimensional vector, in which the ith element of $\mathbf{x}$ is replaced with 0 and other elements are the same with those of $\mathbf{x}$. Then the resulting feature space $S^i = \{\mathbf{g}(\mathbf{x}^i) \,|\, \mathbf{x}^i \in R^m\}$ is the subspace of $S = \{\mathbf{g}(\mathbf{x}) \,|\, \mathbf{x} \in R^m\}$, where the feature space variables in S^i that include the ith element of $\mathbf{x}^i$ are zero for polynomial and RBF kernels.

Let the coefficient vectors obtained by KDA in S and S^i be $\mathbf{w}_{\rm opt}$ and $\mathbf{w}^i_{\rm opt}$, respectively. Then

$$J(\mathbf{w}_{\rm opt}) \geq J(\mathbf{w}^i_{\rm opt}) \tag{32}$$

is satisfied. This is proved as follows. Assume that the above relation does not hold. Namely, $J(\mathbf{w}_{\rm opt}) < J(\mathbf{w}^i_{\rm opt})$ is satisfied. Then $\mathbf{w}_{\rm opt}$ is not optimal in S since $\mathbf{w}^i_{\rm opt} \in S$.

Monotonicity of the selection criterion is very important because we can terminate the selection procedure by setting a threshold, or we can use optimization techniques such as branch and bound for feature selection.

4.2 KDA-Based Recognition Rates

The second selection criterion is the recognition rate calculated by the KDA classifier. In the KDA classifier, we calculate conditional probabilities in the KDA space. Let ω be a random variable that takes ω_1 or ω_2 and ω_i denote class i $(i = 1, 2)$. The probability that $\mathbf{x}$ belongs to class i, $p(\mathbf{x}|\omega_i)$, is calculated assuming that the class i data obey the normal distribution:

$$p(\mathbf{x}|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-(p_r(\mathbf{x}) - \mathbf{w}^T\mathbf{c}_i)^2/\sigma_i^2) \quad \text{for} \quad i = 1, 2, \tag{33}$$

where $\mathbf{c}_i$ is the center of class i in the feature space and using (31) $\mathbf{w}^T\mathbf{c}_i$ and σ_i are given, respectively, by

$$\mathbf{w}^T\mathbf{c}_i = \frac{1}{M_i}\sum_{j=1}^{M_i}(H(\mathbf{y}_1, \mathbf{x}_j), \ldots, H(\mathbf{y}_{M'}, \mathbf{x}_j))\,\boldsymbol{\alpha}, \tag{34}$$

$$\sigma_i = \frac{1}{M_i}\sum_{j=1}^{M_i}(p_r(\mathbf{x}_j) - \mathbf{w}^T\mathbf{c}_i)^2. \tag{35}$$

Then $\mathbf{x}$ is classified into class

$$\arg_i \max_{i=1,2} p(\mathbf{x} \,|\, \omega_i). \tag{36}$$

Using (36) we calculate the recognition rate of the training data set or the validation data set in cross-validation. Unfortunately, monotonicity of the KDA-based

recognition rate for the deletion of features is not guaranteed. Thus, to increase stability of the selection criterion we evaluate the KDA-based recognition rate by cross-validation.

5 Backward Feature Selection

We select features using backward feature selection. In the backward feature selection, first we calculate the value of the selection criterion using all the features. Then starting from the initial set of features we temporarily delete each feature, calculate the value of the selection criterion, and delete the feature with the largest value of the selection criterion from the set. We iterate feature deletion so long as the value of the selection criterion is larger than the prescribed threshold.

To determine the threshold we normalize the selection criterion by that evaluated using all the features. Since the KDA criterion is nonincreasing for the deletion of features, we set the threshold smaller than 1. It is difficult to set a proper value but in the following study based on some preliminary experiment we set 0.5.

For the KDA-based recognition rate, we set the threshold to be 1 and stop deletion if the normalized selection criterion is smaller than 1. This means that if the recognition rate evaluated by cross-validation is smaller than that using all the features, we stop deleting features. This is to avoid deteriorating the generalization ability by deleting too many features.

Let the initial number of features be m and F^k and F^k_j denote the set of k features and the set of k features with the j element temporarily deleted from the set, respectively. And let the selection criterion for F^k_j be $T(F^k_j)$. Then the normalized selection criterion $c(F^k_j)$ is

$$c(F^k_j) = \frac{T(F^k_j)}{T(F^m)}. \tag{37}$$

The procedure of backward feature selection is as follows:

1. Using the m features, evaluate the selection criterion T^m. Set the initial set of features as $F^m = \{1, \ldots, m\}$. Set $k = m$ and go to Step 2.

2. Delete the ith $(i = 1, \ldots, k)$ feature temporarily from F^k and calculate the normalized selection criterion $c(F^k_i)$. For the KDA criterion, if

$$c(F^k_j) > \delta_{\mathrm{KDA}} \quad \text{for} \quad j = \arg\max_{i \in F^k} c(F^k_i), \tag{38}$$

   where δ_{KDA} is the threshold for the KDA criterion and for the KDA-based recognition rate, if

$$c(F^k_j) > \delta_{\mathrm{Rec}} \quad \text{for} \quad j = \arg\max_{i \in F^k} c(F^k_i), \tag{39}$$

   where δ_{Rec} is the threshold for the KDA-based recognition rate, go to Step 3. Otherwise stop feature selection.

Table 1: Two-class benchmark data sets.

Data	Inputs	Train.	Test	Sets
B. cancer	9	200	77	100
Diabetes	8	468	300	100
German	20	700	300	100
Heart	13	170	100	100
Image	18	1300	1010	20
Ringnorm	20	400	7000	100
F. solar	9	666	400	100
Thyroid	5	140	75	100
Titanic	3	150	2051	100
Twonorm	20	400	7000	100
Waveform	21	400	4600	100

3. Permanently delete j from F^k:

$$F^{k-1} = F^k - \{j\} \tag{40}$$

and go to Step 2.

After feature selection, set F^k includes the features selected by backward feature selection. Instead of deleting one feature at a time, to speed up feature selection we may delete more than one feature as discussed in [16]. But here we use the conventional backward feature selection strategy since our main purpose is to demonstrate the usefulness of the selection criteria.

6 Performance Evaluation

6.1 Data Sets and Evaluation Conditions

We evaluated performance of the selection criteria using the two-class problems [24][1] listed in 1. The table shows the numbers of input variables, training data, test data, and data sets for the problems. Except for the image problem, each problem has 100 training data sets and their associated test data sets. We selected these problems because each problem consists of multiple training data sets and their associated test data sets and thus we could analyze the statistical difference of the proposed selection methods.

As a classifier to evaluate the performance of the proposed feature selection methods, we used the SVM. Therefore to determine parameters for feature selection, we determined the kernel and its parameter value by fivefold cross-validation for the SVM using the first five training data sets. Namely, in cross-validation for each training data set, we divided the data set randomly into five subsets, trained the SVM using the four subsets and evaluate the recognition rate for the remaining subset, i.e., for the validation subset, repeated training and evaluation of the SVM changing the subsets, and calculated the

[1] http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm

total recognition rate for the validation subsets. By cross-validation we selected the kernel, its parameter, and the margin parameter from among polynomial kernels with $d = [2, 3, 4]$, RBF kernels with $\gamma = [0.1, 1, 10]$, and from $C = [1, 10, 50, 100, 500, 1000, 2000, 3000, 5000, 8000, 10000, 50000, 100000]$.

Then we selected the value of ε, which is used to avoid matrix singularity in KDA and the threshold value of Cholesky factorization, η, from among $\varepsilon = [10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$ and $\eta = [10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$ so that the KDA criterion [27] or the KDA-based recognition rate is maximized as follows:

1. Calculate the KDA criterion or the KDA-based recognition rate for all the combinations of the values of ε and η for the first five training data sets.

2. Select the values of ε and η that correspond to the maximum value of the KDA criterion or the KDA-based recognition rate.

According to our preliminary experiment, since the KDA-based recognition rate evaluated using the training data was not robust for feature selection, we calculated the recognition rate by fivefold cross-validation. The values of ε and η were fixed during feature selection.

Since each problem consists of 100 or 20 data sets, we combined the first five training data sets into one and selected features by backward feature selection for the two selection criteria with $\delta_{\mathrm{KDA}} = 0.5$ and $\delta_{\mathrm{Rec}} = 1.0$. Since the KDA criterion is monotonic for the deletion of variables, δ_{KDA} needs to be smaller than 1. The selection threshold of $\delta_{\mathrm{Rec}} = 1.0$ means that if the recognition rate is smaller than that using all the features we stop feature selection.

As a reference selection criterion we used the SVM-based recognition rate evaluated by cross-validation. The feature selection procedure is the same with that of the KDA-based recognition rate. The only difference is that the recognition rate is evaluated training the SVM instead of training the KDA-based classifier.

After feature selection we evaluated performance of the selected features by the recognition rate of the test data using the SVM. To evaluate the recognition rate of the test data for the selected features, we fixed the kernel and the kernel parameter with those determined using all the features and determined the margin parameter C by cross-validation using the first five training data sets. According to the order of deleted features, we deleted one feature at a time from the set of features, determined the margin parameter C for the selected feature set by cross-validation, and calculated the mean and the standard deviation for the test data sets. Then we statistically analyzed the means and standard deviations for the initial feature set and the selected feature set with the significance level of 0.05.

Table 2 lists the parameter values obtained by the above procedure. In the table, for example, $\gamma 0.1$ means the RBF kernels with $\gamma = 0.1$ and $d3$ means the polynomial kernels with degree 3. For the KDA criterion, $\varepsilon = 10^{-8}$ was selected for all the problems. Thus, we do not list the values of ε in the table.

6.2 Experimental Results

Figure 1 shows the recognition rates of the thyroid data set when features were deleted using the KDA criterion. The horizontal axis shows the deleted features

Table 2: Parameter setting.

Data	Kernel	KDA Criterion	KDA Classifier	
		η	ε	η
B. cancer	$\gamma 0.1$	10^{-8}	10^{-4}	10^{-2}
Diabetes	$d3$	10^{-8}	10^{-8}	10^{-7}
German	$\gamma 0.1$	10^{-8}	10^{-7}	10^{-6}
Heart	$\gamma 0.1$	10^{-8}	10^{-8}	10^{-2}
Image	$\gamma 10$	10^{-8}	10^{-8}	10^{-5}
Ringnorm	$\gamma 10$	10^{-4}	10^{-3}	10^{-8}
F. solar	$d2$	10^{-3}	10^{-2}	10^{-3}
Thyroid	$\gamma 10$	10^{-8}	10^{-6}	10^{-2}
Titanic	$d3$	10^{-3}	10^{-8}	10^{-8}
Twonorm	$d3$	10^{-8}	10^{-5}	10^{-2}
Waveform	$\gamma 10$	10^{-3}	10^{-2}	10^{-8}

at each selection step and the vertical axis shows the recognition rates of the training data set in the right and test data sets in the left for each selection step. The vertical axis also shows the value of the selection criterion with the initial value normalized to 1.

As seen from the figure, the selection criterion is monotonic for the deletion of features. Since $\delta_{\mathrm{KDA}} = 0.5$, three features: 4th, 3rd, and 1st features were deleted and 2nd and 5th features were left.

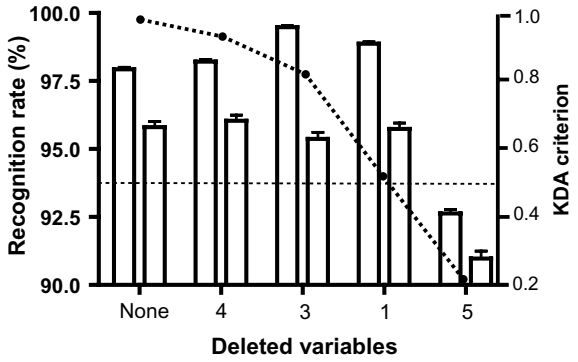

Figure 1: Feature deletion for the thyroid data set by KDA criterion.

Table 3 shows the feature selection results using the KDA criterion. In the table, the "Deleted" column lists the features deleted. The "C" column lists the value of C selected by fivefold cross-validation. Using the determined value of C the SVM was trained and the recognition rates were evaluated. The "Train." and "Test" columns list the average recognition rates with the standard deviations for the training and test data sets, respectively. The "KDA" column lists the values of the selection criterion.

In the table, the results for each classification problem consist of two or three

lines. The average recognition rate of the test data in the first line is in Roman. If the average recognition rate of the test data for the second or third line is statistically the same with, better than, or inferior to that of the first line, i.e., that without deleting features, the average recognition rate is shown in Roman, boldface, or in parentheses, respectively. If the recognition rate is in Roman or boldface, it also means that the average recognition rates during the feature deletion process are all statistically equal to or better than that with all the features.

First we explain the three-line results using the waveform problem. According to the KDA criterion, features, 3, 16, 6, 15, 19, and 8 were deleted and the average recognition rate was 88.41 ± 0.39, which was statistically inferior to 90.00 ± 0.44 using all the features. The second line shows the deleted features, where the average recognition rates were statistically the same with or better than that using all the features until all the features listed in the second line were deleted. For the waveform problem, even if feature 3 was deleted, the average recognition rate was statistically the same with that without deleting feature 3. But if feature 16, in addition to feature 3 was deleted, the average recognition rate became inferior.

There are two cases for two-line results: 1) the average recognition rates during feature deletion are better than that with all the features or 2) deletion of a feature from all the features results in statistical deterioration of the average recognition rate. For instance for the breast cancer problem, deletion of features 4, 1, 8, 7, 5, 3, and 2 did not degrade the generalization ability. Whereas for the ringnorm problem, the deletion of feature 18 resulted in deterioration of the generalization ability.

From Table 3, except for the image problem, the KDA criterion is monotonic for the deletion of features. Namely, except for the image problem, the KDA criterion decreased as the features were deleted, but for the image data, it increased during feature deletion. Except for the ringnorm and twonorm problems, at least one feature was deleted without deteriorating the recognition rate of the test data.

Table 4 shows the feature selection results using the KDA-based recognition rate evaluated by cross-validation. In the table, "Rate" denotes the KDA-based recognition rate for the validation data set in cross-validation. Comparing Tables 3 and 4 the performance of the two methods are similar but the KDA-based recognition rate did not delete features for the titanic problem. For the waveform problem, the KDA-based recognition rate performed better.

Table 5 shows the feature selection results using the SVM-based recognition rate. Compared to Tables 3 and 4, performance is very poor. The reason is that the stopping condition was too conservative. Namely, we could delete features even though the recognition rate by cross-validation was lower than that using all the features. For example, for the breast cancer problem, if we ignored the stopping condition, we could delete eight features without deteriorating the generalization ability. This is better than the KDA criterion and the KDA-based recognition rate. If we select features for each of the first five training data sets, six features were commonly deleted without deteriorating the generalization ability. Thus the early stopping of feature selection was caused by combining the first five training data sets. Although the stopping condition of the KDA-based recognition rate evaluated by cross-validation worked well, that of the SVM-based criterion did not work. From Table 5 the discrepancy between the

Table 3: Recognition performance for feature selection using the KDA criterion. Bold face numerals and numerals in parentheses mean that they are statistically superior and inferior to the associated numerals using all the features, respectively.

Data	Deleted	C	Train.	Test	KDA
B. cancer	None	500	77.57±1.87	72.36±4.67	0.94
	4,1,8,7,5,3,2	2000	74.41±2.59	72.57±5.01	0.46
Diabetes	None	100	78.95±1.27	76.42±1.79	1.35
	4,1,5,3,7	2000	78.12±1.16	76.29±1.92	1.03
	4,1,5,3,7,6	3000	76.56±1.15	(75.45)±**1.18**	0.92
German	None	50	77.80±1.03	76.19±2.27	0.94
	4,16,17,20,13,5,11,9,12,19,8,18	1000	78.26±1.09	76.62±2.16	0.83
	4,16,17,20,13,5,11,9,12,19,8,18,6,15,1,10,14	5000	74.93±1.03	(74.60)±2.26	0.53
Heart	None	50	85.96±1.91	83.69±3.41	2.95
	4,5,1,10,6,8,9,7,2,11	10000	83.72±2.88	82.72±3.81	1.93
Image	None	1000	98.60±0.17	97.13±0.47	18.9
	8,6,12,9,10,3,14,16,4,5,7,13	500	98.81±0.19	**97.64**±0.41	20.9
	8,6,12,9,10,3,14,16,4,5,7,13,1,18	50000	97.96±0.26	(96.06)±0.52	9.68
Ringnorm	None	10	99.51±0.33	97.67±0.33	27.6
	18,20,15,11,5,17,14	10	98.33±0.54	(95.50)±0.39	13.9
F. solar	None	10	67.50±1.04	67.61±1.72	0.57
	9,7,8,3,6,2,1	100000	67.46±1.09	67.67±1.81	0.44
Thyroid	None	10	97.93±0.78	95.80±2.09	26.1
	4,3,1	8000	98.87±0.64	95.75±2.16	14.2
Titanic	None	100	79.49±3.66	77.47±1.43	0.74
	2,1	100000	78.09±3.60	77.57±**0.26**	0.54
Twonorm	None	10	98.09±0.59	97.59±0.12	8.31
	12,18,19,7,4,10,2,6,5,16,9	10	91.53±1.39	(90.70)±(0.20)	4.20
Waveform	None	1	93.53±1.36	90.00±0.44	22.8
	3	1	93.45±1.25	89.99±0.45	20.9
	3,16,6,15,19,8	1	91.63±1.43	(88.41)±0.39	12.5

recognition rate for the validation data set and that of the test data set is large although that for the KDA-based recognition rate is not prominent as seen from Table 4. Because we evaluated the SVM-based recognition rate combining the first five training data sets and the SVM is more powerful than the KDA-based classifier, the SVM-based recognition rate for the validation data set improved more than that for the test date set. Thus, the validity of the stopping condition was deteriorated.

Table 6 lists the number of features selected without deteriorating the recognition rate of the test data and the number of selected features in parenthesis for three methods. In the table, "KDA" denotes the KDA criterion, "Rec." denotes the KDA-based recognition rate, and " SVM" denotes the SVM-based recognition rate. In each row, the largest number of deleted features without deteriorating the generalization ability is shown in boldface. The "Best" row shows the numbers that the associated selection criterion performed best among the selection criteria and the "Total" row shows the total number of features

Table 4: Recognition performance for feature selection using the KDA-based recognition rate with cross-validation.

Data	Deleted	C	Train.	Test	Rate
B. cancer	None	500	77.57±1.87	72.36±4.67	75.90
	2,8,4,9,7,1,3	500	76.27±1.97	**75.79**±4.72	77.40
Diabetes	None	100	78.95±1.27	76.42±1.79	79.19
	1,4,3	500	79.17±1.14	76.64±1.62	78.76
German	None	50	77.80±1.03	76.19±2.27	76.89
	16,5,4,12,17,15,18,13,20,10,7,11	100	77.05±1.00	76.74±2.44	77.06
Heart	None	50	85.96±1.91	83.69±3.41	84.71
	2,6,11,8,5,4,1,10,7	100	84.21±2.13	83.48±3.74	84.82
Image	None	1000	98.60±0.17	97.13±0.47	97.75
	8,9,6,3,16,10,14,12,5,4,7,11,18	2000	99.10±0.26	**97.50**±0.48	97.96
Ringnorm	None	10	99.51±0.33	97.67±0.33	98.50
	14,20	10	99.27±0.39	(97.27)±0.37	98.50
F. solar	None	10	67.50±1.04	67.61±1.72	66.42
	6,2,7,3,1,8	10	67.46±1.09	67.67±1.81	67.15
	6,2,7,3,1,8,4,9	10	57.23±1.16	(57.22)±1.93	100
Thyroid	None	10	97.93±0.78	95.80±2.09	97.14
	4,3	100	98.79±0.68	95.67±2.12	97.43
Titanic	None	100	79.49±3.66	77.47±1.43	79.73
Twonorm	None	10	98.09±0.59	97.59±0.12	97.25
	20	50	98.08±0.74	(97.16)±(0.19)	97.45
Waveform	None	1	93.53±1.36	90.00±0.44	91.20
	6,2,1,4,20,21,13	1	92.65±1.29	89.96±0.44	91.45

selected by the associated selection criterion without deteriorating the generalization ability. The "Coincidences" row shows the number of coincidences in which the number of features deleted by the stopping condition of the selection criterion and the maximum number of deleted features that do not deteriorate the generalization ability are the same.

From the standpoint of the Best criterion, the KDA criterion is the best and the KDA-based recognition rate is the second best but their difference is not so large. And from the total number of deleted features and the number of coincidences they are the best. Thus, KDA criterion and the KDA-based recognition rate are the comparably best selection criterion. It is surprising that the KDA criterion, which does not use the information of the validation data performed comparably with the KDA-based recognition rate with cross-validation.

The SVM-based recognition rate with cross-validation performed poorly. The main reason is that the stopping condition was too conservative.

We measured feature selection time using an Athlon 64×2 4800+ personal computer running on Linux. Table 7 lists the feature selection time. To make comparison clear, we measured the deletion time of three variables by KDA criterion, the KDA-based recognition rate, and the SVM-based recognition rate with cross-validation for $\gamma = 10$ and $d = 2$. In training the SVM we used the

Table 5: Recognition performance for feature selection using the SVM-based recognition rate with cross-validation.

Data	Deleted	C	Train.	Test	Rate
B. cancer	None	500	77.57±1.87	72.36±4.67	91.70
Diabetes	None	100	78.95±1.27	76.42±1.79	85.51
	4	100	78.97±1.18	76.64±1.65	85.74
German	None	50	77.80±1.03	76.19±2.27	96.34
	20,4,16,5	50	78.02±0.91	76.32±2.21	96.46
Heart	None	50	85.96±1.91	83.69±3.41	98.24
	10	100	86.37±1.92	83.37±3.08	98.47
Image	None	1000	98.60±0.17	97.13±0.47	99.69
	5,7,6,18,3,8,9,14,11,4,10,12,16	5000	98.62±0.16	97.38±0.32	99.78
	5,7,6,18,3,8,9,14,11,4,10,12,16,15	1000	97.23±0.28	(95.54)±0.48	99.72
Ringnorm	None	10	99.51±0.33	97.67±0.33	98.60
F. solar	None	10	67.50±1.04	67.61±1.72	67.18
	8,9,1	10	67.48±1.07	67.63±1.78	67.24
Thyroid	None	10	97.93±0.78	95.80±2.09	100
	4,3	100	98.79±0.68	95.67±2.12	100
Titanic	None	100	79.49±3.66	77.47±1.43	79.87
Twonorm	None	10	98.09±0.59	97.59±0.12	97.25
	2	10	97.71±0.66	(97.25)±(0.13)	97.25
Waveform	None	1	93.53±1.36	90.00±0.44	92.30
	10,4,2	1	92.79±1.37	(89.56)±0.45	92.55

primal-dual interior-point method combined with the decomposition techniques [28]. From the table, the feature selection time for the KDA criterion is two to four times shorter than that of the KDA-based recognition rate. This is because the KDA-based recognition rate was evaluated by fivefold cross-validation. Usually, feature selection by the KDA criterion and the KDA-based recognition rate was much faster than that by the SVM-based recognition rate. But in some cases, feature selection by the SVM-based recognition rate was faster. The reason may be that the KDA criterion and the KDA-based recognition rate used the Cholesky factorization and for a large training data set, the decomposition became slower. For example for the twonorm problem, the matrix size is 2000 combining the first five training data sets.

7 Conclusions

In this paper, we proposed two measures for feature selection: the KDA criterion which is the objective function of KDA and the KDA-based recognition rate, which calculate the conditional probability of a datum belonging to a class assuming that each class data obey the normal distribution.

We show that the KDA criterion is monotonic for the deletion of features. Thus we can evaluate the KDA criterion using the training data. Backward feature selection is terminated when the KDA criterion is below the predetermined

Table 6: Comparison of feature selection methods.

Data	KDA	Rec.	SVM
B. cancer	**7** (7)	**7** (7)	0 (0)
Diabetes	**5** (6)	3 (3)	1 (1)
German	**12** (17)	**12** (12)	4 (4)
Heart	**10** (10)	9 (9)	1 (1)
Image	12 (14)	**13** (13)	**13** (14)
Ringnorm	0 (7)	0 (2)	0 (0)
F. solar	**7** (7)	6 (8)	3 (3)
Thyroid	**3** (3)	2 (2)	2 (2)
Titanic	**2** (2)	0 (0)	0 (0)
Twonorm	0 (11)	0 (1)	0 (1)
Waveform	1 (6)	**7** (7)	0 (3)
Best	7	4	1
Total	59	59	17
Coincidences	5	7	3

Table 7: Comparison of feature selection time in seconds.

Data	$\gamma 10$			$d2$		
	KDA	Rec.	SVM	KDA	Rec.	SVM
B. cancer	145	608	7672	20	94	21229
Diabetes	2183	7115	200494	88	289	254509
German	19505	69374	635227	2716	8734	571241
Heart	168	1095	1753	34	94	2769
Image	94594	261559	199106	4266	14539	3766942
Ringnorm	18040	46989	5254	820	2617	186991
F. solar	795	1909	1589827	210	742	1508135
Thyroid	16	52	340	3	10	1192
Titanic	1	5	1336	1	5	1593
Twonorm	18301	49004	3364	1738	2694	9270
Waveform	18402	49112	10564	1027	3269	37829

threshold.

Since the KDA-based recognition rate is not monotonic for the deletion of features, we evaluate the KDA-based recognition rate by cross-validation. The backward feature selection process is terminated when the KDA-based recognition rate is below that using all the features.

Using the two-class benchmark data sets we compared the two criteria. As a reference feature selection method, we used the SVM-based recognition rate evaluated by cross-validation. The KDA criterion deleted most for the seven benchmark data sets out of 11, but the KDA criterion and the KDA-based recognition rate showed comparable performance in the number of total deleted features that did not deteriorate the generalization ability and in the validity of the stopping criteria. But the SVM based recognition rate performed poorly in the number of deleted features and the feature selection time.

As future work, for a large number of training data we need to accelerate

feature selection for the KDA criterion; a fast matrix inversion method other than the Cholesky factorization needs to be used.

References

[1] S. Abe. *Support Vector Machines for Pattern Classification*. Springer-Verlag, London, 2005.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002.

[3] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–324, 1997.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition, second edition*. Academic Press, San Diego, 1990.

[5] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, Third Edition*. Academic Press, London, UK, 2006.

[6] S. Abe. *Pattern Classification: Neuro-Fuzzy Methods and Their Comparison*. Springer-Verlag, London, 2001.

[7] R. Thawonmas and S. Abe. A novel approach to feature selection based on analysis of class regions. *IEEE Transactions on Systems, Man, and Cybernetics—Part B*, 27(2):196–207, 1997.

[8] M. Ashihara and S. Abe. Feature selection based on kernel discriminant analysis. In S. Kollias, A. Stafylopatis, W. Duch, and E. Oja, editors, *Artificial Neural Networks (ICANN 2006)—Proceedings of the Sixteenth International Conference, Part II, Athens, Greece*, pages 282—291. Springer-Verlag, Berlin, Germany, 2006.

[9] T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio. Image representations for object detection using kernel classifiers. In *Proceedings of Asian Conference on Computer Vision (ACCV 2000)*, pages 687–692, Taipei, Taiwan, 2000.

[10] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. P. Mesirov, and T. Poggio. Support vector machine classification of microarray data. Technical Report AI Memo 1677, Massachusetts Institute of Technology, 1999.

[11] J. Bi, K. P. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–1243, 2003.

[12] S. Perkins, K. Lacker, and J. Theiler. Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research*, 3:1333–1356, 2003.

[13] A. Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.

[14] Y. Liu and Y. F. Zheng. FS_SFS: A novel feature selection method for support vector machines. *Pattern Recognition*, 39(7):1333–1345, 2006.

[15] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003.

[16] S. Abe. Modified backward feature selection by cross validation. In *Proceedings of the Thirteenth European Symposium on Artificial Neural Networks (ESANN 2005)*, pages 163–168, Bruges, Belgium, 2005.

[17] T. Nagatani and S. Abe. Backward variable selection of support vector regressors by block deletion. In *Proceedings of the 2007 International Joint Conference on Neural Networks (IJCNN 2007)*, pages 1540–1545, Orlando, FL, 2007.

[18] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, pages 82–90, Madison, 1998.

[19] M. Brown. Exploring the set of sparse, optimal classifiers. In *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR 2003)*, pages 178–184, Florence, Italy, 2003.

[20] M. Brown, N. P. Costen, and S. Akamatsu. Efficient calculation of the complete optimal classification set. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, volume 2, pages 307–310, Cambridge, UK, 2004.

[21] C. Gold, A. Holub, and P. Sollich. Bayesian approach to feature selection and parameter tuning for support vector machine classifiers. *Neural Networks*, 18(5–6):693–701, 1999.

[22] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[23] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK, 2000.

[24] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX—Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48, 1999.

[25] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

[26] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

[27] S. Kita, S. Maekawa, S. Ozawa, and S. Abe. Boosting kernel discriminant analysis with adaptive kernel selection. In *Proceedings of Seventh International Conference on Adaptive and Natural Computing Algorithm*, CD-ROM, 2005.

[28] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing VII—Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 276–285, 1997.