Feature selection based on kernel discriminant analysis

Ashihara, Masamichi

Abe, Shigeo

(Citation)
Lecture Notes in Computer Science : Artificial Neural Networks - ICANN 2006,4132:282-291

(Issue Date)
2006

(Resource Type)
journal article

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/90000153

Feature Selection Based on Kernel Discriminant Analysis

Masamichi Ashihara and Shigeo Abe

Graduate School of Science and Technology
Kobe University
Rokkodai, Nada, Kobe, Japan
abe@eedept.kobe-u.ac.jp
http://www2.eedept.kobe-u.ac.jp/~abe

Abstract. For two-class problems we propose two feature selection criteria based on kernel discriminant analysis. The first one is the objective function of kernel discriminant analysis (KDA) and the second one is the KDA-based exception ratio. We show that the objective function of KDA is monotonic for the deletion of features, which ensures stable feature selection. The KDA-based exception ratio defines the overlap between classes in the one-dimensional space obtained by KDA. The computer experiments show that the both criteria work well to select features but the former is more stable.

1 Introduction

Feature selection, i.e., deletion of irrelevant or redundant input variables from the given input variables, is one of the important steps in constructing a pattern classification system with high generalization ability [1, 2]. And many selection methods for kernel-based methods have been proposed [2–7]. The margin [5, 8, 9] is often used for feature selection for support vector machines. Instead of the margin, in [7], block deletion of features in backward feature selection is proposed using the generalization ability by cross-validation as the selection criterion.

Feature selection has a long history of research and many methods have been developed. In [10], an exception ratio is defined based on the overlap of class regions approximated by hyperboxes. This exception ratio is monotonic for the deletion of input variables. By this monotonicity, we can terminate feature selection when the exception ratio exceeds a predefined value.

In this paper we propose two feature-selection criteria based on kernel discriminant analysis (KDA) for two-class problems. The first criterion uses the objective function of KDA. Namely the ratio of the between-class scatter and within-class scatter. We prove that this criterion is monotonic for the deletion of input variables. The second criterion is the exception ratio defined on the one-dimensional space generated by KDA according to [10].

The feature selection is done by backward selection. We start from all the input variables. We temporally delete one input variable, calculate the selection

criterion, and delete the input variable that improves the selection criterion the most. This process is iterated until the stopping condition is satisfied.

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

2 Kernel Discriminant Analysis

In this section we summarize kernel discriminant analysis, which finds the component that maximally separates two classes in the feature space [11, 12], [13, pp. 457–468].

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 $\mathbf{w}^T\,\mathbf{g}(\mathbf{x})/\|\mathbf{w}\|$. 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

$$d^2 = (\mathbf{w}^T(\mathbf{c}_1 - \mathbf{c}_2))^2 = \mathbf{w}^T(\mathbf{c}_1 - \mathbf{c}_2)\,(\mathbf{c}_1 - \mathbf{c}_2)^T\mathbf{w}, \tag{1}$$

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

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

We define

$$Q_B = (\mathbf{c}_1 - \mathbf{c}_2)\,(\mathbf{c}_1 - \mathbf{c}_2)^T \tag{3}$$

and call Q_B the *between-class scatter matrix*.

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

$$s_i^2 = \mathbf{w}^T\,Q_i\,\mathbf{w} \qquad \text{for} \quad i = 1, 2, \tag{4}$$

where

$$Q_i = \frac{1}{M_i}\,(\mathbf{g}(\mathbf{x}_1^i), \ldots, \mathbf{g}(\mathbf{x}_{M_i}^i))\,(I_{M_i} - \mathbf{1}_{M_i})\begin{pmatrix}\mathbf{g}^T(\mathbf{x}_1^i)\\ \vdots \\ \mathbf{g}^T(\mathbf{x}_{M_i}^i)\end{pmatrix} \quad \text{for} \quad i = 1, 2. \tag{5}$$

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

$$Q_W = Q_1 + Q_2 \tag{6}$$

and call Q_W the *within-class scatter matrix*.

Now, we want to maximize

$$J(\mathbf{w}) = \frac{d^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T\, Q_B\, \mathbf{w}}{\mathbf{w}^T\, Q_W\, \mathbf{w}}, \tag{7}$$

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 $M' \leq M_1 + M_2$. Then $\mathbf{w}$ is expressed as

$$\mathbf{w} = (\mathbf{g}(\mathbf{y}_1), \ldots, \mathbf{g}(\mathbf{y}_{M'}))\, \boldsymbol{\alpha}, \tag{8}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{M'})^T$ and $\alpha_1, \ldots, \alpha_{M'}$ are scalars. Substituting (8) into (7), we obtain

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T\, K_B\, \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\, K_W\, \boldsymbol{\alpha}}, \tag{9}$$

where

$$K_B = (\mathbf{k}_{B_1} - \mathbf{k}_{B_2})\, (\mathbf{k}_{B_1} - \mathbf{k}_{B_2})^T, \tag{10}$$

$$\mathbf{k}_{Bi} = \begin{pmatrix} \dfrac{1}{M_i} \displaystyle\sum_{j=1}^{M_i} H(\mathbf{y}_1, \mathbf{x}_j^i) \\ \cdots \\ \dfrac{1}{M_i} \displaystyle\sum_{j=1}^{M_i} H(\mathbf{y}_{M'}, \mathbf{x}_j^i) \end{pmatrix} \quad \text{for} \quad i = 1, 2, \tag{11}$$

$$K_W = K_{W_1} + K_{W_2}, \tag{12}$$

$$K_{W_i} = \frac{1}{M_i} \begin{pmatrix} H(\mathbf{y}_1, \mathbf{x}_1^i) \cdots H(\mathbf{y}_1, \mathbf{x}_{M_i}^i) \\ \cdots \\ H(\mathbf{y}_{M'}, \mathbf{x}_1^i) \cdots H(\mathbf{y}_{M'}, \mathbf{x}_{M_i}^i) \end{pmatrix} (I_{M_i} - \mathbf{1}_{M_i}) \times \begin{pmatrix} H(\mathbf{y}_1, \mathbf{x}_1^i) \cdots H(\mathbf{y}_1, \mathbf{x}_{M_i}^i) \\ \cdots \\ H(\mathbf{y}_{M'}, \mathbf{x}_1^i) \cdots H(\mathbf{y}_{M'}, \mathbf{x}_{M_i}^i) \end{pmatrix}^T \quad \text{for} \quad i = 1, 2. \tag{13}$$

Taking a partial derivative of (9) with respect to $\mathbf{w}$ and equating the resulting equation to zero, we obtain the following generalized eigenvalue problem:

$$K_B\, \boldsymbol{\alpha} = \lambda\, K_W\, \boldsymbol{\alpha}, \tag{14}$$

where λ is a generalized eigenvalue.

Substituting

$$K_W\, \boldsymbol{\alpha} = \mathbf{k}_{B_1} - \mathbf{k}_{B_2} \tag{15}$$

into the left-hand side of (14), we obtain

$$(\boldsymbol{\alpha}^T\, K_W\, \boldsymbol{\alpha})\, K_W\, \boldsymbol{\alpha}. \tag{16}$$

Thus, by letting $\lambda = \boldsymbol{\alpha}^T K_W \boldsymbol{\alpha}$, (15) is a solution of (14).

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

$$\boldsymbol{\alpha} = K_W^{-1} (\mathbf{k}_{B_1} - \mathbf{k}_{B_2}). \tag{17}$$

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 [11]:

$$\boldsymbol{\alpha} = (K_W + \varepsilon I)^{-1} (\mathbf{k}_{B_1} - \mathbf{k}_{B_2}), \tag{18}$$

where ε is a small positive parameter.

3 Selection Criteria and Their Monotonicity

3.1 KDA Criterion

The first selection criterion is the value of (7) 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 input variables. 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 $\mathbf{w}_{\mathrm{opt}}$ and $\mathbf{w}^i_{\mathrm{opt}}$, respectively. Then

$$J(\mathbf{w}_{\mathrm{opt}}) \geq J(\mathbf{w}^i_{\mathrm{opt}}) \tag{19}$$

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

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.

3.2 KDA-based Exception Ratio

In this section, we discuss the exception ratio defined in the one-dimensional space, $\mathbf{w}^T \mathbf{g}(\mathbf{x})/\|\mathbf{w}\|$, obtained by KDA, which is an extension of the exception ratio [10] defined in the input space. We call the space obtained by KDA *KDA space*. We define the class overlap by the overlap of class data in the KDA space. Namely, for class i ($i = 1, 2$), we define the activation regions with level 1, $A_{ii}(1)$, calculating the maximum $V_{ii}(1)$ and minimum $v_{ii}(1)$ of class i data in the KDA

space. If the activation regions $A_{11}(1)$ and $A_{22}(2)$ overlap we define the overlapping regions as the inhibition region $I_{12}(1)$ with the interval $[W_{12}(1), w_{12}(1)]$. If there are data in the inhibition region, we define the activation regions with level 2, $A_{12}(2)$ and $A_{21}(2)$. If there is an overlap between $A_{12}(2)$ and $A_{21}(2)$, we define the inhibition region $I_{12}(2)$. We repeat the above procedure until there are no data in the inhibition region.

The ratio of activation regions and inhibition regions indicates the difficulty of classification. Therefore, we define the exception ratio o_{ij} for classes i and j as the sum of the ratios of the activation and inhibition regions as follows:

$$o_{ij} = \sum_{l=1,\cdots,l_{ij}} p_{ij}(l) \frac{b_{I_{ij}}(l)}{b_{A_{ij'}}(l)}, \tag{20}$$

where $j' = i$ for $l = 1$, $j' = j$ for $l \geq 2$,

$$b_{I_{ij}} = \begin{cases} W_{ij}(l) - w_{ij}(l) & \text{for } W_{ij}(l) - w_{ij}(l) > \varepsilon, \\ \varepsilon & \text{otherwise,} \end{cases}$$

$$b_{A_{ij'}} = \begin{cases} V_{ij'}(l) - v_{ij'}(l) & \text{for } V_{ij'}(l) - v_{ij'}(l) > \varepsilon, \\ \varepsilon & \text{otherwise,} \end{cases}$$

$$p_{ij}(l) = \frac{\text{number of class } i \text{ training data in } I_{ij}(l)}{\text{total number of training data}}.$$

Here, ε is a small positive parameter. If there is no data in the inhibition region, the region does not affect separability of classes. Thus, in (20), we add $p_{ij}(l)$ to reflect this fact. We call the exception ratio given by (20) *KDA-based exception ratio*.

The exception ratio is zero if there is no overlap between classes. Thus, by this criterion, separability is considered to be the same even if the margins between classes are different. The exception ratio defined in the input space is monotonic for the deletion of input features [10], but unfortunately the KDA-based exception ratio is not monotonic as the computer experiments discussed later show.

4 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 temporally delete each feature, calculate the value of the selection criterion, and delete the feature with the highest value of the selection criterion from the set. We iterate feature deletion so long as class separability is higher than the prescribed level.

Let the initial set of selected features be F^m, where m is the number of input variables, and the value of the selection criterion be T^m. We delete the ith ($i = 1, \ldots, m$) feature temporally from F^m and calculate the selection criterion. Let the selection criterion be T_i^m. We iterate this procedure for all i

$(i = 1, \ldots, m)$. Then we delete the feature $\arg\max_{i \in F^m} T_i^m$ from F^m: $F^{m-1} = F^m - \{\arg\max_{i \in F^m} T_i^m\}$, if $T_i^m / T^m > \delta_{\mathrm{KDA}}$ or $T_i^m / T^m < \delta_{\mathrm{EXT}}$, where the first inequality is for the KDA criterion, the second inequality is for the KDA-based exception ratio, and δ_{KDA} and δ_{EXT} are thresholds for the KDA criterion and KDA-based exception ratio, respectively.

We iterate the above feature selection procedure so long as the above inequality is satisfied.

5 Performance Evaluation

We evaluated performance of the selection criteria using the two-class problems listed in Table 1 [11].[1] Each problem has 100 or 20 training and test data sets.

For the features selected by backward feature selection, we trained the L1 support vector machines, scaling the input range into $[0, 1]$, calculated the means and standard deviations of the recognition rates, and statistically analyzed the results with the significance level of 0.05. We used an AthlonMP2000+ personal computer running on Linux.

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

Table 2. Parameter setting

Data	Kernel	ε	η
B. cancer	$\gamma 10$	10^{-8}	10^{-8}
Diabetes	$\gamma 10$	10^{-8}	10^{-6}
German	$\gamma 10$	10^{-8}	10^{-8}
Heart	$\gamma 10$	10^{-8}	10^{-8}
Image	$\gamma 10$	10^{-8}	10^{-8}
Ringnorm	$\gamma 10$	10^{-8}	10^{-4}
F. solar	$\gamma 10$	10^{-8}	10^{-7}
Thyroid	$\gamma 10$	10^{-8}	10^{-8}
Titanic	$\gamma 10$	10^{-8}	10^{-4}
Twonorm	$\gamma 10$	10^{-8}	10^{-6}
Waveform	$\gamma 10$	10^{-8}	10^{-3}

We selected the kernel and its parameter, from among polynomial kernels with $d = [2, 3, 4]$ and RBF kernels with $\gamma = [0.1, 1, 10]$, so that the maximum value of the objective function of KDA [14] is realized. 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 $\varepsilon = [10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$, $\eta = [10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$ so that the KDA criterion is maximized as follows:

1. Calculate the KDA criterion, using all the features, for the first five training data sets. Thus we obtain 5 values of the objective function.
2. Select the values of ε and η that correspond to the maximum value of the KDA criterion.

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

Table 2 lists the parameter values obtained by the above procedure. For all the problems, RBF kernels with $\gamma = 10\,(\gamma = 10)$ were selected.

In evaluating the selected features by the support vector machine, we determined the kernel and parameter values by 5-fold cross-validation; for the original set of features, we used the same kernel types and parameter ranges as those for KDA and determined the value of the margin parameter C from $C = [1, 10, 50, 100, 500, 1000, 2000, 3000, 5000, 8000, 10000, 50000, 100000]$. For the selected feature set, we used the same kernel and kernel parameter as those for the initial set of features and determined the value of C by cross-validation.

Since each problem consists of 100 or 20 data sets, we combined the first 5 training data sets into one and selected features by backward feature selection for the two selection criteria with $\delta_{\mathrm{KDA}} = 0.5$ and $\delta_{\mathrm{EXT}} = 1.5$.

Figures 1 and 2 show the recognition rates of the thyroid data set when features were deleted using the KDA criterion and KDA-based exception ratio criterion, respectively. The horizontal axis shows the deleted features 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.

In Fig. 1, 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. In Fig. 2 the deletion sequence of features is the same with that by the KDA criterion. But the selected features are different. From the figure, for $\delta_{\mathrm{EXT}} = 1.5$ only the 4th feature was deleted compared with three features by the KDA criterion. Since the exception ratio decreased when the fourth feature was deleted, the exception ratio was not monotonic.

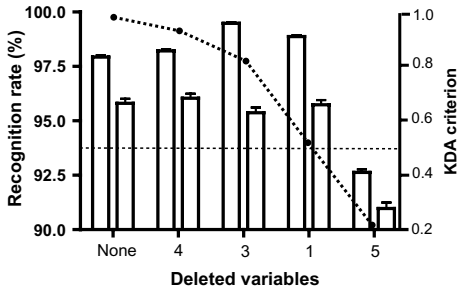

Fig. 1. Feature deletion for the thyroid data set by KDA criterion

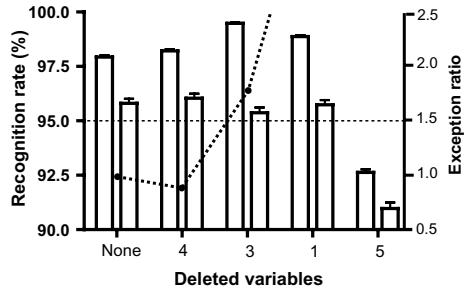

Fig. 2. Feature deletion for the thyroid data set by exception ratio

Tables 3 and 4 show the feature selection results using the KDA and KDA-based exception ratio criteria, respectively. In the tables, the "Deleted" column lists the features deleted. If for a classification problem two feature strings are shown, the numbers of features selected by the two criteria are different. The second feature string shows the features that are deleted after the first feature

string is deleted. And the asterisk shows that the number of features deleted is the same with that deleted by the other criterion not used in deleting the features. For example, in Table 3 for b. cancer the 5th and the 9th features are deleted using the KDA criterion and since four features are deleted by the KDA-based exception ratio criteria as shown in Table 4, we delete two more features: the 2nd and 7th. The best average recognition rate and standard deviation are shown in boldface and the second best italic. If there is no statistical difference they are shown in Roman.

In Table 3, "Parm" and "C" columns list the kernels and the values of C selected by 5-fold cross validation. For example, $\gamma 0.1$ means the RBF kernels with $\gamma = 0.1$ and $d3$ means the polynomial kernels with degree 3. (In Table 4, the "Parm" column is not included because it is the same with that in Table 3.) The "Train." and "Test" columns list the average recognition rates with the standard deviations. The "KDA (EXT)" column lists the values of the selection criterion. The "#F" column lists the number of features that are successively deleted without deteriorating the generalization ability in each deletion step. For example, in Table 3, according to the KDA criterion the four features are deleted for diabetes but by statistical analysis, additional one feature can be deleted.

From Table 3, except for the image data, the KDA criterion is monotonic for the deletion of features. For the image data, because of the memory overflow, we could not delete more than 6 features. Except for the ringnorm, twonorm, and waveform data sets, the selected features by the KDA criterion show comparable performance for the test data with the original features.

In Table 4, for german and heart data sets, since the exception ratio was 0, we deleted the features using KDA criterion until the exception ratio became non-zero. The exception ratio was monotonic for b. cancer, diabetes, ringnorm, twonorm, and waveform. For f. solar and titanic data sets, since the exception ratio monotonically decreased for the deletion of features, we could not stop the deletion procedure. The selected features by the exception ratio show comparable performance for the test data with the original features for b. cancer, diabetes, heart, image, and thyroid data sets. The "#F" for the KDA criterion is in most cases better than that for the exception ratio. And the feature selection is more stable.

6 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 exception ratio, which defines the overlap of classes in the one-dimensional space obtained by KDA. We show that the KDA criterion is monotonic for the deletion of features. According to the computer experiments for two-class problems, we showed that both criteria work well to selects features but the KDA criterion was more stable.

References

1. S. Abe. *Support Vector Machines for Pattern Classification*. Springer, 2005.

Table 3. Recognition performance for feature selection using the KDA criterion.

Data	Deleted	Parm	C	Train.	Test	KDA	#F
B. cancer	None	γ0.1	500	*77.57*±**1.87**	72.36±4.67	12.3	4
	5,9		2000	**78.68**±**1.83**	72.94±4.56	8.3	
	2,7*		100	74.56±4.04	72.77±5.39	3.7	
Diabetes	None	*d*3	100	**78.95**±1.27	76.42±1.79	3.31	5
	4,3*		50	78.44±1.05	76.67±1.76	2.50	
	5,1		100	78.35±1.11	**77.00**±1.67	1.78	
German	None	γ0.1	50	*77.80*±1.03	76.19±2.27	676	9
	4,20,16,5,18,15,10,17		500	**78.71**±0.90	75.82±2.14	359	
	19*		100	76.99±1.00	75.77±2.17	137	
Heart	None	γ0.1	50	**85.96**±1.91	83.69±3.41	1081	5
	6,11,9		100	**86.15**±1.93	83.76±3.52	694	
	4,1*		100	85.17±2.07	83.43±3.53	65	
Image	None	γ10	1000	98.60±0.17	97.13±0.47	18.9	6
	8,6,12,9,10,3		2000	**99.28**±**0.09**	97.37±0.37	22.2	
Ringnorm	None	γ10	10	**99.51**±**0.33**	**97.67**±0.33	27.6	0
	18*		10	**99.38**±**0.35**	*97.41*±0.37	25.8	
	20, 15, 11, 5, 17, 14		10	98.33±0.54	95.50±0.39	13.9	
F. solar	None	*d*2	10	67.50±1.05	67.61±1.72	0.730	1
	9,6,8,3,7,2,1		100000	67.46±1.09	67.67±1.81	0.436	
Thyroid	None	γ10	10	97.93±0.78	95.80±2.09	26.1	3
	4*		10	*98.21*±0.82	96.04±2.08	25.2	
	3,1		8000	**98.87**±**0.64**	95.75±2.16	14.2	
Titanic	None	*d*3	100	**79.49**±3.66	77.47±1.43	0.839	2
	2,1		100000	78.09±3.60	77.57±**0.26**	0.542	
Twonorm	None	*d*3	10	**98.09**±**0.59**	**97.59**±**0.12**	42.7	0
	18,7*		10	*97.62*±0.71	*96.95*±**0.14**	35.3	
	12,5,2		50	96.86±0.82	95.67±0.19	23.8	
Waveform	None	γ10	1	**93.53**±1.36	**90.00**±0.44	22.8	1
	3,16*		1	**93.18**±1.28	*89.77*±0.45	19.1	
	6,15,19,8		1	91.63±1.43	88.41±0.39	12.5	

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. P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. *Proc. ICML '98*, 82–90, 1998.
4. J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero-norm with linear models and kernel methods. *J. Machine Learning Research*, 3:1439–1461, 2003.
5. S. Perkins, K. Lacker, and J. Theiler. Grafting: Fast, incremental feature selection by gradient descent in function space. *J. Machine Learning Research*, 3:1333–1356, 2003.
6. Y. Liu and Y. F. Zheng. FS_SFS: A novel feature selection method for support vector machines. *Pattern Recognition* (to appear).
7. S. Abe. Modified backward feature selection by cross validation. In *Proc. ESANN 2005*, 163–168, 2005.

Table 4. Recognition performance for feature selection using the exception ratio

Data	Deleted	C	Train.	Test	EXT	#F
B. cancer	None	500	77.57±1.87	**72.36**±4.67	0.288	1
	1,9*	100000	**82.73**±1.92	70.55±4.73	0.170	
	4,2	100000	*78.45*±1.69	**72.72**±4.73	0.358	
Diabetes	None	100	**78.95**±1.27	76.42±1.79	14.6	5
	5,3	10	77.51±1.04	76.10±1.83	19.4	
	1,6*	100	77.59±1.26	75.90±1.82	27.2	
German	None	50	*77.80*±1.03	**76.19**±2.27	0	8
	4,20,16,5,18,15,10,17*	500	**78.71**±0.90	**75.82**±2.14	0	
	2	500	76.98±1.11	73.91±2.21	0	
Heart	None	50	**85.96**±1.91	83.69±3.41	0	6
	6,11,9*	100	**86.15**±1.93	83.76±3.52	0	
	4,1	100	85.17±2.07	83.43±3.53	0	
Image	None	1000	98.60±0.17	97.13±0.47	1.43	6
	3,10,6,8,9,14	2000	**99.23±0.13**	**97.40±0.37**	0.79	
Ringnorm	None	10	**99.51±0.33**	**97.67±0.33**	0.0993	0
	17	10	**99.40±0.34**	*97.52*±0.32	0.131	
	20,5,6,12,2,8*	50	99.25±0.42	94.80±0.38	1.26	
F. solar	None	10	**67.50±1.05**	**67.61±1.72**	2.97	0
	4,7,2,1,6,5,3*	50000	49.69±6.65	48.76±6.64	0.0158	
Thyroid	None	10	97.93±0.78	95.80±2.09	0.000257	3
	4	10	*98.21*± 0.82	96.04±2.08	0.000218	
	3,1*	8000	**98.87±0.64**	95.75±2.16	0.113	
Titanic	None	100	**79.49±3.66**	**77.47±1.43**	0.894	1
	1,3*	100000	46.6±30.0	45.92±29.5	0.0520	
Twonorm	None	10	**98.09±0.59**	**97.59±0.12**	0.00805	0
	8,3	50	**97.96±0.65**	*96.91*±0.17	0.0536	
	10,4,2*	50	96.65±0.89	95.46±0.19	0.154	
Waveform	None	1	**93.53**±1.36	**90.00**±0.44	0.264	0
	9,7	1	*92.88*±1.27	*89.41*±0.39	0.301	
	6,21,4,8*	1	91.72±1.29	88.54±0.41	1.28	

8. J. Bi, K. P. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *J. Machine Learning Research*, 3:1229–1243, 2003.
9. A. Rakotomamonjy. Variable selection using SVM-based criteria. *J. Machine Learning Research*, 3:1357–1370, 2003.
10. R. Thawonmas and S. Abe. A novel approach to feature selection based on analysis of class regions. *IEEE Trans. SMC–B*, 27(2):196–207, 1997.
11. S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. *NNSP 99*, 41–48, 1999.
12. G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.
13. B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.
14. S. Kita, S. Maekawa, S. Ozawa, and S. Abe. Boosting kernel discriminant analysis with adaptive kernel selection. In *Proc. ICANCA 05*, CD-ROM, 2005.