Feature selection based on kernel discriminant analysis for multi-class problems

Ishii, Tsuneyoshi

Abe, Shigeo

(Citation)
Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on:2455-2460

(Issue Date)
2008-06

(Resource Type)
conference paper

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/90000926

Feature Selection Based on Kernel Discriminant Analysis for Multi-Class Problems

Tsuneyoshi Ishii and Shigeo Abe

***Abstract*— We propose a feature selection criterion based on kernel discriminant analysis (KDA) for an n-class problem, which finds $n-1$ eigenvectors on which the projected class data are locally maximally separated. The proposed criterion is the sum of the objective function values of KDA associated with the $n-1$ eigenvectors. The criterion results in calculating the sum of $n-1$ eigenvalues associated with the eigenvectors and is shown to be monotonic for the deletion or addition of features. Using the backward feature selection strategy, for several multi-class data sets, we evaluated the proposed criterion and the criterion based on the recognition rate of the support vector machine (SVM) evaluated by cross-validation. From the standpoint of generalization ability the proposed criterion is comparable with the SVM-based recognition rate, although the proposed method does not use cross-validation.**

I. Introduction

In pattern recognition, inputs variables, i.e., features, usually include unnecessary or redundant features, which may slow down classification speed or deteriorate the generalization ability. Feature selection is one of the approaches to avoid this problem, in which from the original set of features the minimum subset of features that realizes the maximum generalization ability [1], [2] is selected. To realize the maximum generalization ability, during the feature selection process, we need to estimate the generalization ability of feature subsets. This type of feature selection is called a wrapper method. But it is time-consuming to directly estimate the generalization ability. Therefore some selection criterion, which well reflects the generalization ability, is used. This method is called a filter method.

The filter or wrapper method usually uses the forward or backward selection method [3]. 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. In backward selection, we start from all the features and delete one feature at a time, which deteriorates the selection criterion the least. We iterate procedure until the selection criterion reaches a specified value. Usually, backward selection is slower but is more stable in selecting optimal features than forward selection.

With the advent of support vector machines (SVMs), feature selection methods and selection criteria suitable for SVMs are discussed. In addition to wrapper and filter methods, the embedded method, in which feature selection is combined with training the SVM is developed [4], [5], [6]. For the wrapper method a selection criterion such as the recognition rate of the SVM with cross-validation is used [1]. In [3], [7], to speedup feature selection, block deletion of features in backward feature selection is proposed using the generalization ability by cross-validation. As selection criteria of the filter method the margin [8] is widely used, and for linear kernels the absolute values of the coefficient vector elements of the decision function are used [9]. In [10], [11], kernel discriminant analysis (KDA) for two-class problems [12] is used. The objective function of KDA called KDA criterion is the ratio of the between-class scatter and within-class scatter and is proved to be monotonic for the deletion of features. If a selection criterion is monotonic for the deletion or addition of a feature, we can terminate feature selection when the selection criterion violates a predefined value. Feature selection based on the KDA criterion was shown to be robust for benchmark data sets.

In this paper, we extend the KDA criterion applicable to multi-class problems and demonstrate usefulness of the criterion as a filter method. The proposed KDA criterion is the sum of the locally maximum values of the objective function of KDA, which is the ratio of the between class scatter and the total scatter. For an n-class problem, KDA generates $n-1$ eigenvectors on which the projected class data are locally maximally separated. We show that the sum of the $n-1$ locally maximum values is equal to the sum of the $n-1$ eigenvalues of the generalized eigenvalue problem of KDA. The KDA criterion expressed by the sum of the $n-1$ eigenvalues is proved to be monotonic for the feature selection. This characteristic contributes in robust feature selection by backward selection without cross-validation. The stopping condition is specified according to the KDA criterion evaluated using all the features and the feature selection is terminated when the KDA criterion is lower than the threshold.

By computer experiments we compare the proposed feature selection criterion and the selection criterion based on the recognition rate of the SVM evaluated by cross-validation, called SVM-based recognition rate.

In Sections II and III, we summarize fuzzy pairwise SVMs used for evaluating multi-class problems and KDA for multi-class problems, respectively and in Section IV, we discuss the selection criterion, relationship between the locally maximum values of the objective function and the eigenvalues and monotonicity of the selection criterion. In Section V, we explain backward feature selection used and in Section V I, we demonstrate the validity of the proposed method by computer experiments.

Tsuneyoshi Ishii and Shigeo Abe are with Graduate School of Engineering, Kobe University, Kobe 657-8501, JAPAN (email: 070t219t@stu.kobe-u.ac.jp, abe@kobe-u.ac.jp).

978-1-4244-1821-3/08/$25.00©2008 IEEE

II. Fuzzy Pairwise Support Vector Machines

In this section we summarize fuzzy pairwise support vector machines [1], [13], which are used to evaluate the proposed selection criterion and are also used as the selection criterion, namely the SVM-based recognition rate.

In a fuzzy pairwise SVM, we determine the decision functions for all combinations of class pairs. Thus for an n class problem the number of decision functions is $n(n-1)/2$. To resolve unclassifiable regions occurred in pairwise SVMs, we introduce the membership functions.

Let a set of m-dimensional data belonging to class $i\,(i = 1,...,n)$ be $\mathbf{x}_{i1},...,\mathbf{x}_{in_i}$, and data $\mathbf{x}$ be mapped into the l-dimensional feature space by the mapping function $\mathbf{g}(\mathbf{x})$. 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.

Let the decision function for class i against class j be

$$D_{ij}(\mathbf{x}) = \mathbf{w}_{ij}^t \mathbf{g}(\mathbf{x}) + b_{ij}, \qquad (1)$$

where $\mathbf{w}_{ij}$ is the l-dimensional vector and b_{ij} is the bias term.

To avoid confusion of notations, let $\{\mathbf{x}_1,\ldots,\mathbf{x}_{n_i+n_j}\}$ be the training data for classes i and j, where n_i is the number of the data in class i. To determine the optimal separating hyperplane, we minimize

$$\frac{1}{2}||\mathbf{w}_{ij}||^2 + C\sum_{n=1}^{n_i+n_j}\xi_n \qquad (2)$$

subject to the constraints

$$y_n(\mathbf{w}_{ij}^t\mathbf{g}(\mathbf{x}_n) + b_{ij}) \le 1 - \xi_n \quad \text{for } n = 1,...,n_i+n_j, \qquad (3)$$

where C is the margin parameter that determines the tradeoff between the maximization of the margin and minimization of the classification error, y_n is the class label and 1 if $\mathbf{x}_n$ belongs to class i and -1 if it belongs to class j, and ξ_n is a nonnegative slack variable. Since the dimension of the feature space is usually very large, we convert the original problem into the dual problem. Namely we maximize

$$Q(\boldsymbol{\alpha}) = \sum_{k=1}^{n_i+n_j}\alpha_k - \frac{1}{2}\sum_{k,l=1}^{n_i+n_j}\alpha_k\alpha_l y_k y_l H(\mathbf{x}_k,\mathbf{x}_l) \qquad (4)$$

subject to the constraints

$$\sum_{k=1}^{n_i+n_j} y_k\alpha_k = 0, \quad C \le \alpha_k \le 0 \quad \text{for } k = 1,...,n_i+n_j, \qquad (5)$$

where α_k is a nonnegative Lagrange multiplier and is a dual variable associated with $\mathbf{x}_k$.

We define the membership function in the directions orthogonal to $D_{ij}(\mathbf{x}) = 0$ as follows:

$$m_{ij}(\mathbf{x}) = \begin{cases} 1 & \text{for } D_{ij}(\mathbf{x}) \le 1, \\ D_{ij}(\mathbf{x}) & \text{otherwise.} \end{cases} \qquad (6)$$

We define the class i membership function of $\mathbf{x}$ by the minimum operation for m_{ij}:

$$m_i(\mathbf{x}) = \min_{\substack{j \ne i \\ j = 1,...,n}} m_{ij}(\mathbf{x}). \qquad (7)$$

Because $m_i(\mathbf{x}) = 1$ holds for only one class, now unknown datum $\mathbf{x}$ is classified into the class

$$\arg\max_{i=1,...,n}\ \min_{\substack{j \ne i \\ j = 1,...,n}} D_{ij}(\mathbf{x}). \qquad (8)$$

III. Kernel Discriminant Analysis for Multi-Class Problems

In this section we summarize kernel discriminant analysis based on [14], which finds, for an n-class problem, the $n-1$ vectors onto which the projections of the data of one class is maximally separated from the remaining classes in the feature space.

Now we assume that the center of all data is zero in the feature space. Let S_T, S_B denote the total scatter matrix and the between-class scatter matrix in the feature space, respectively:

$$S_T = \frac{1}{M}\sum_{k=1}^{n}\sum_{l=1}^{n_k}\mathbf{g}(\mathbf{x}_{kl})\mathbf{g}^t(\mathbf{x}_{kl}), \qquad (9)$$

$$S_B = \frac{1}{M}\sum_{k=1}^{n} n_k\mathbf{m}_k\mathbf{m}_k^t, \qquad (10)$$

where M is the number of training data and $\mathbf{m}_k$ is the center of the kth class:

$$\mathbf{m}_k = \frac{1}{n_k}\sum_{l=1}^{n_k}\mathbf{g}(\mathbf{x}_{kl}). \qquad (11)$$

When samples in the feature space are projected onto vector $\mathbf{w}_i$, the total scatter matrix and the between scatter matrix on $\mathbf{w}_i$ are given, respectively, by

$$\frac{1}{M}\sum_{k=1}^{n}\sum_{l=1}^{n_k}(\mathbf{w}_i^t\mathbf{g}(\mathbf{x}_{kl}))^2 = \mathbf{w}_i^t S_T\mathbf{w}_i, \qquad (12)$$

$$\frac{1}{M}\sum_{k=1}^{n} n_l(\mathbf{w}_i^t\mathbf{m}_l)^2 = \mathbf{w}_i^t S_B\mathbf{w}_i. \qquad (13)$$

KDA seeks vector $\mathbf{w}_i$ $(i = 1,\ldots,n-1)$ that maximizes the ratio of total scatter and between-class scatter for maximum class separation. Namely, we want to maximize

$$J(\mathbf{w}_i) = \frac{\mathbf{w}_i^t S_B\mathbf{w}_i}{\mathbf{w}_i^t S_T\mathbf{w}_i}. \qquad (14)$$

But because $\mathbf{w}_i$, S_B, and S_T are defined in the feature space, we need to use kernel tricks. Then $\mathbf{w}_i$ is expressed as

$$\mathbf{w}_i = \sum_{k=1}^{n}\sum_{l=1}^{n_k} a_i^{kl}\mathbf{g}(\mathbf{x}_{kl}), \qquad (15)$$

where $\mathbf{a}_i = (a_i^{kl})\,(k \in \{1,...,n\}; l = 1,...,n_k)$. Substituting (15) into (14), we obtain

$$J(\mathbf{a}_i) = \frac{\mathbf{a}_i^t KWK\mathbf{a}_i}{\mathbf{a}_i^t KK\mathbf{a}_i}, \qquad (16)$$

where K is the kernel matrix and W is defined as

$$W_{ij} = \begin{cases} \dfrac{1}{n_k} & \text{if both } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to class } k, \\ 0 & \text{otherwise.} \end{cases} \quad (17)$$

Taking the partial derivative of (16) with respect to $\mathbf{w}_i$ and equating the resulting equation to zero, we obtain the following generalized eigenvalue problem

$$\lambda_i KK\mathbf{a}_i = KWK\mathbf{a}_i, \quad (18)$$

where λ_i are generalized eigenvalues.

Suppose the rank of K is r $(r \leq M)$ and the eigenvector decomposition of K is $K = P\Gamma P^t$, where P and Γ are $M \times M$ matrices. Now we remove zero eigenvalues to avoid singularity and redefine $M \times M$ P by $M \times r$ P and $M \times M$ Γ by $r \times r$ Γ. Thus Γ^{-1} exists and $P^tP = I$, where I is the identity matrix.

Let $\boldsymbol{\beta}_i = \Gamma P^t \mathbf{a}_i$ and substituting K into (16), we get

$$J(\boldsymbol{\beta}_i) = \frac{\boldsymbol{\beta}_i^t P^t W P \boldsymbol{\beta}_i}{\boldsymbol{\beta}_i^t P^t P \boldsymbol{\beta}_i} = \frac{\boldsymbol{\beta}_i^t P^t W P \boldsymbol{\beta}_i}{\boldsymbol{\beta}_i^t \boldsymbol{\beta}_i}. \quad (19)$$

Therefore we obtain the following eigenvalue problem:

$$\lambda_i \boldsymbol{\beta}_i = P^t W P \boldsymbol{\beta}_i. \quad (20)$$

Once $\boldsymbol{\beta}_i$ are calculated, $\mathbf{a}_i$ can be computed as $\mathbf{a}_i = P\Gamma^{-1}\boldsymbol{\beta}_i$.

To speed up the calculation of KDA, we can avoid the eigenvalue decomposition as discussed in [15].

IV. Selection Criterion and Its Monotonicity

A. Selection Criterion

The new feature selection criterion that we propose is the sum of the locally maximum values of the objective function of KDA. This is equivalent to the sum of the objective function values associated with the largest $n-1$ eigenvalues obtained by KDA:

$$\sum_{i=1}^{n-1} J(\mathbf{w}_i). \quad (21)$$

We can show that (21) is equivalent to

$$\sum_{i=1}^{n-1} \lambda_i. \quad (22)$$

Using (22) we need not calculate $\boldsymbol{\beta}$ and $\mathbf{a}$.

Now we prove that (21) is equal to (22). First, we show that the locally maximum value of (14) equals to the eigenvalue λ of (20). Taking the partial derivative of (14) with respect to $\mathbf{w}$ and equating the resulting equation to zero, we obtain

$$\frac{(\mathbf{w}^t S_T \mathbf{w})(2S_B\mathbf{w}) - (\mathbf{w}^t S_B \mathbf{w})(2S_T\mathbf{w})}{(\mathbf{w}^t S_T \mathbf{w})^2} = 0. \quad (23)$$

Thus

$$S_B\mathbf{w} = \left(\frac{\mathbf{w}^t S_B \mathbf{w}}{\mathbf{w}^t S_T \mathbf{w}}\right) S_T\mathbf{w}. \quad (24)$$

Letting

$$\frac{\mathbf{w}^t S_B \mathbf{w}}{\mathbf{w}^t S_T \mathbf{w}} = \lambda,$$

we obtain

$$S_B\mathbf{w} = \lambda S_T\mathbf{w}. \quad (25)$$

Therefore,

$$\text{local } \max\{J(\mathbf{w})\} = \lambda. \quad (26)$$

Thus, we obtain

$$\sum_{i}^{n-1} J(\mathbf{w}_i) = \sum_{i=1}^{n-1} \lambda_i. \quad (27)$$

B. Monotonicity

Monotonicity of the selection criterion is very important because we can terminate the selection procedure by setting a threshold. The proposed criterion is easily proved to be monotonic for the deletion of features. 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 resultant feature space $S^i = \{\mathbf{g}(\mathbf{x}^i) | \mathbf{x}^i \in R^m\}$ is the subspace of $S = \{\mathbf{g}(\mathbf{x}) | \mathbf{x} \in R^m\}$. Let the sums of eigenvalues given by (22) obtained in S and S^i be $\sum_{j=1}^{n-1} \lambda_j$ and $\sum_{j=1}^{n-1} \lambda_j^i$, respectively. Then

$$\sum_{j=1}^{n-1} \lambda_j \geq \sum_{j=1}^{n-1} \lambda_j^i \quad (28)$$

is satisfied. This is proved as follows. Assume that $\lambda_j < \lambda_j^i$ is satisfied. Then since λ_j^i is obtained by maximizing the objective function in S^i, which is a subspace of S, the above assumption cannot hold. Thus, (28) holds.

V. 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 highest value of selection criterion from the set. We iterate feature deletion so long as class separability is higher than the prescribed level. It is difficult to set a proper value but we set 0.95 in the following study.

Let the initial number of features be m and F^k and F_j^k denote the set of kth feature and the set of jth element temporarily deleted from the set. And let the selection criterion for F_j^k be T_j^k. Then we define the normalized selection criterion c_j^k:

$$c_j^k = \frac{T_j^k}{T^m}. \quad (29)$$

The procedure of backward feature selection is as follows:

1) Using all the features, evaluate the selection criterion T^m. And set $k = m$.

2) Delete the ith ($i = 1, ..., k$) feature temporarily from F^k and calculate the selection criterion T_i^k and get normalized selection criterion c_i^k. if

$$c_j^k > \delta \qquad \text{for } j = \arg\max_{i \in F^k} c_i^k,$$

where δ is the threshold, go to Step 3. Otherwise stop feature selection.

3) Delete j from F^k and go to Step 2.

Instead of deleting one feature at a time, to speed up feature selection we may use block deletion, namely we may delete more than one feature at a time. But here we use the conventional backward feature selection to demonstrate usefulness of the propose selection criterion.

VI. Experiments

A. Data Sets and Experimental Conditions

We evaluated performance of the selection criterion using benchmark data sets listed in Table I, which shows the numbers of input valuables, classes, training data, and test data. We scaled the input ranges into [0, 1].

In feature selection, we selected the kernel and its parameter from among polynomial kernels with $d = [1, 2, 3, 4]$ and RBF kernels with $\gamma = [0.1, 1, 10]$ so that the KDA criterion is maximized. As a result, we selected $\gamma = 10$ for all the problems. As a classifier to evaluate the performance of feature selection, we used the fuzzy pairwise SVM. Therefore, to evaluate the effectiveness of parameter selection by maximizing the KDA criterion, we also performed feature selection with the kernel and its parameter determined by the SVM using all the features.

Since the KDA criterion did not change much for the iris problem, we set threshold $\delta = 0.95$ for all the classification problems.

As the reference selection criterion we used the SVM-based recognition rate evaluated by fivefold cross-validation. We used the fuzzy pairwise L1-SVM. The feature selection procedure was the same. The only difference is the criterion. We determined the kernel, its parameters and margin parameter C by fivefold cross-validation. The kernels, their parameter ranges were the same for the KDA criterion and C was selected from $C = [1, 10, 50, 100, 500, 1000, 2000, 3000, 5000, 8000, 10000, 50000, 100000]$. During feature selection, we used the same kernel and kernel parameter as those for the initial feature set and determined the value of C by cross-validation.

After feature selection, we evaluated performance of the selected features by the recognition rate of the test data using the fuzzy pairwise L1-SVM. To evaluate the recognition rate of the test data for the selected features, we fixed the kernel, its parameter and the margin parameter C with those determined using all the features. Table II lists the selected kernels and their parameters, where $\gamma = 10$ means the RBF kernel with $\gamma = 10$ and $d = 3$ means the polynomial kernel with degree 3. Fixing the kernel and its parameter, we deleted one feature at a time from the set of features, determined the margin parameter C for selected feature set by cross-validation, and evaluated the recognition rates of the training and test data sets.

TABLE I
Benchmark data sets

Data	Inputs	Classes	Train. data	Test data
Iris	4	3	75	75
Numeral	12	10	810	820
Thyroid	21	3	3772	3428
Blood cell	13	12	3097	3100
Hiragana-13	13	38	8375	8375

TABLE II
SVM parameters

Data	kernel
Iris	$\gamma = 0.1$
Numeral	$d = 3$
Thyroid	$d = 1$
Blood cell	$\gamma = 10$
Hiragana-13	$\gamma = 10$

B. Experimental Results

Table III shows the feature selection results using the KDA criterion with the parameters determined by maximizing the KDA criterion. In the "Deleted" column, a list without parentheses shows the deleted features and that in parentheses shows the remaining features. The "C" column lists the value of margin parameter C selected by cross-validation using the SVM. Using the determined C, the SVM was trained and the recognition rates of the training and test data sets were evaluated. The "Train." and "Test" columns list the recognition rates for the training and test data sets, respectively. The "KDA" column lists the values of the normalized KDA criterion.

For each data set the results are shown in two or three lines. First we explain the three-line results. The first line shows the recognition rates of the training and test data sets when all the features are used. The second line shows the recognition rates when all the features listed in the "Deleted" column are deleted. In this case we mean that if the features are deleted in the listed order, the recognition rates of the test data are higher than that with all the features. The "Deleted" column in the third line shows the features deleted so long as the threshold δ is satisfied after the features in the second line are deleted. In this case, the recognition rates of the test data are lower than that of the initial set of features. For example, for numeral data, the second, fifth, and third features were deleted without deteriorating the recognition rate of the test data, but afterwards, the recognition rate was decreased and features were deleted until the eighth, first, ninth, and 10th features remained.

There are two cases for two-line results. In the first case, the recognition rate is better than that with all the features during feature selection. In the second case, deletion of a feature results in deterioration of the recognition rate. For the blood cell data, the recognition rate of the test data decreased when the first feature was deleted and the recognition rate was inferior to that using all the features.

Except for the thyroid data set, the KDA criterion was monotonic for the deletion of features. Figure 1 shows the value of the normalized selection criterion and the recognition rate of the numeral data set for each selection step. The horizontal axis shows the deleted features and the vertical axis shows the recognition rate of the training data set in the left and that of the test data set in the right. The vertical axis also shows the value of the normalized selection criterion in the dotted line. Figure 2 shows the value of the normalized selection criterion and the recognition rate of the blood cell data set for each selection step. From Figs. 1 and 2, the KDA criterion was monotonic for the deletion of features but it is difficult to set a proper threshold value. Figure 3 shows the value of the normalized selection criterion and the recognition rate of the thyroid data set for each selection step. The value of the normalized selection criterion was not monotonic and until 12 features were deleted it was higher than 1.

In Table IV, we show the feature selection results with kernel parameters determined by the SVM. Since the parameters were different for iris, numeral, and thyroid data sets, we show results only for these data sets. For the iris data set, the stopping threshold works better with the SVM parameter, but for the numeral data set, three features were deleted by the parameter determined by the KDA criterion, but two by the SVM parameter. For the thyroid data set, from the recognition rate of the test data, the stopping threshold works better for the parameter determined by the KDA criterion.

Table V shows the feature selection results using the SVM-based recognition rate with cross-validation. In the table, "Validation" denotes the recognition rate for the validation data sets in cross-validation. Comparing the results with Table III and Table IV, the selection performance of KDA criterion and that of SVM-based recognition rate are comparable. For the iris data set, the selected features were different but the recognition rates of the test data are the same for the KDA criterion with the parameter determined by the SVM and the SVM-based recognition rate. For the numeral data set, the KDA criterion deleted more features than SVM-based recognition rate with comparable performance. For the thyroid data set, more features were deleted by the KDA criterion but the recognition rate of test data was a little worse. For the blood cell and hiragana-13 data sets, there is not much difference in the generalization ability by the both methods.

VII. Conclusions

In this paper, we proposed the feature selection criterion for an n class problem: the KDA criterion, which is the sum of $n-1$ eigenvalues of KDA associated with the $n-1$ eigenvectors, on which the projected class data are locally maximally separated.

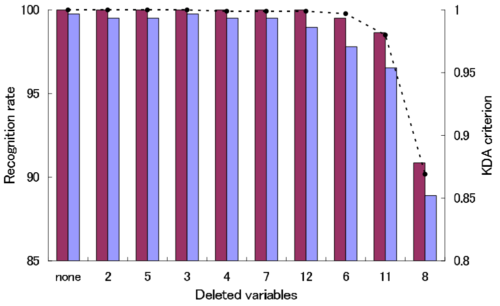

Fig. 1. Feature deletion for the numeral data set

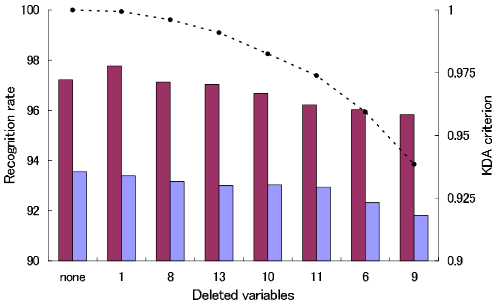

Fig. 2. Feature deletion for the blood cell data set

We show that the KDA criterion is monotonic for the deletion of features, which ensures termination of feature selection when the KDA criterion is below the predetermined threshold.

By computer experiments we compared the performance of the selection criterion that is the recognition rate of the SVM with cross-validation called SVM-based recognition rate. The performance of KDA criterion is comparable to that of the SVM-based recognition rate.

References

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

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

[3] S. Abe, "Modified Backward Feature Selection by Cross Validation," *Proc. European Symposium on Artificial Neural Networks (ESANN 2005)*, Bruges, Belgium, pp. 163–168, 2005.

[4] P. S. Bradley and O. L. Mangasarian, "Feature Selection via Concave Minimization and Support Vector Machines," *Proc. Fifteenth International Conference on Machine Learning (ICML '98)*, pp. 82–90, Madison, 1998.

[5] M. Brown, N. P. Costen, and S. Akamatsu, "Efficient Calculation of the Complete Optimal Classification Set," *Proc. Seventeenth International Conference on Pattern Recognition (ICPR 2004)*, Vol. 2, pp. 307–310, Cambridge, UK, 2004.

[6] C. Gold, A. Holub, and P. Sollich, "Bayesian Approach to Feature Selection and Parameter Tuning for Support Vector Machine Classifiers," *Neural Networks*, Vol. 18, Nos. 5-6, pp. 693–701, 1999.

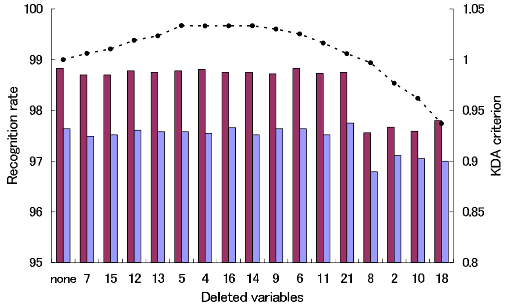

Fig. 3. Feature deletion for the thyroid data set

TABLE III

RECOGNITION PERFORMANCE FOR FEATURE SELECTION USING THE KDA CRITERION. A LIST OF FEATURES IN PARENTHESES IN "DELETED" COLUMN SHOWS THE REMAINING FEATURES

Data	Deleted	C	Train.	Test	KDA
Iris	None	100	100	97.33	1
	3, 1	10	97.33	96.00	0.999
Numeral	None	10^5	100	99.76	1
	2, 5, 3	10^5	100	99.76	1
	(8, 1, 9, 10)	5000	98.64	96.54	0.98
Thyroid	None	10^5	98.83	97.64	1
	(8,2,10,18,1,20,3,19,17)	10^5	98.75	97.75	1.006
	(18, 1, 20, 3, 19, 17)	10^5	97.56	97.05	0.962
Blood cell	None	50	97.22	93.55	1
	1, 8, 13, 10, 11, 6	500	96.03	92.32	0.96
Hiragana-13	None	500	100	99.76	1
	13	1000	100	99.76	0.997
	13, 11, 10	10^5	100	99.62	0.970

[7] T. Nagatani and S. Abe, "Backward Variable Selection of Support Vector Regressors by Block Deletion," *Proc. 2007 International Joint Conference on Neural Networks (IJCNN 2007)*, pp. 1540-1545, Orlando, FL, 2007.

[8] A. Rakotomamonjy, "Variable Selection Using SVM Based Criteria," *Journal of Machine Learning Research*, pp. 1357–1370, 2003.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, Vol. 46, No. 1-3, pp. 389–422, 2002.

[10] M. Ashihara and S. Abe, "Feature Selection Based on Kernel Discriminant Analysis," *Proc. International Conference on Artificial Neural Networks (ICANN 2006)*, Vol. 2, pp. 282–291, Athens, Greece, 2006.

[11] T. Ishii, M. Ashihara, and S. Abe, "Kernel Discriminant Analysis Based Feature Selection," *Neurocomputing* (in press).

[12] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher Discriminant Analysis with Kernels," Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IV—Proc. 1999 IEEE Signal Processing Society Workshop*, pp. 41–48, 1999.

[13] S. Abe and T. Inoue, "Fuzzy Support Vector Machines for Multiclass Problems," *Proc. European Symposium on Artificial Neural Networks (ESANN 2002)*, pp. 113-118, Bruges, Belgium, April 2002.

[14] G. Baudat and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach," *Neural Computation*, Vol. 12, pp. 2385-2404, 2000.

[15] D. Cai, X. He, and J. Han, "Efficient Kernel Discriminant Analysis via Spectral Regression," *Proc. 2007 International Conference on Data Mining (ICDM'07)*, Omaha, NE, pp. 427-432, 2007.

TABLE IV

RECOGNITION PERFORMANCE FOR FEATURE SELECTION USING THE KDA CRITERION WITH SVM PARAMETER. A LIST OF FEATURES IN PARENTHESES IN "DELETED" COLUMN SHOWS THE REMAINING FEATURES

Data	Deleted	C	Train.	Test	KDA
Iris	None	100	100	97.33	1
	2	3000	97.33	97.33	0.97
Numeral	None	10^5	100	99.76	1
	3, 12	10^5	100	99.76	1
	(1, 6, 8, 9, 11)	10^5	100	99.51	0.98
Thyroid	None	10^5	98.83	97.64	1
	(8, 2, 18, 3, 10, 17, 19, 20)	10^5	98.73	97.90	0.991
	(10, 17, 19, 20)	10^5	95.20	95.01	0.955

TABLE V

RECOGNITION PERFORMANCE FOR FEATURE SELECTION USING SVM-BASED RECOGNITION RATE WITH CROSS-VALIDATION.

Data	Deleted	C	Train.	Test	Validation
Iris	None	100	100	97.33	98.67
	1	50	100	97.33	100
Numeral	None	10^5	100	99.76	99.75
	3	10^5	100	99.76	99.88
	7, 12, 10	10^5	100	99.51	99.75
Thyroid	None	10^5	98.83	97.64	98.38
	(3, 8, 17, 19, 20)	10^5	98.59	97.81	98.56
Blood cell	None	50	97.22	93.55	94.61
	9, 8, 1, 6	50	96.96	92.41	94.64
Hiragana-13	None	500	100	99.76	99.77
	13	1000	100	99.72	99.78