An Automatic Method for Selecting the Parameter of the Normalized Kernel Function to Support Vector Machines*

CHENG-HSUAN LI[1], HSIN-HUA HO[2], YU-LUNG LIU[3], CHIN-TENG LIN[1],
BOR-CHEN KUO+ AND JIN-SHIUH TAUR[2]
[1]*Institute of Electrical Control Engineering*
National Chiao Tung University
Hsinchu, 300 Taiwan
[2]*Department of Electrical Engineering*
National Chung Hsing University
Taichung, 402 Taiwan
[3]*Department of Computer Science and Information Engineering*
Asia University
Taichung, 413 Taiwan
+*Graduate Institute of Educational Measurement and Statistics*
National Taichung University of Education
Taichung, 403 Taiwan

Soft-margin support vector machine (SVM) is one of the most powerful techniques for supervised classification. However, the performances of SVMs are based on choosing the proper kernel functions or proper parameters of a kernel function. It is extremely time consuming by applying the *k*-fold cross-validation (CV) to choose the almost best parameter. Nevertheless, the searching range and fineness of the grid method should be determined in advance. In this paper, an automatic method for selecting the parameter of the normalized kernel function is proposed. In the experimental results, it costs very little time than *k*-fold cross-validation for selecting the parameter by our proposed method. Moreover, the corresponding soft-margin SVMs can obtain more accurate or at least equal performance than the soft-margin SVMs by applying *k*-fold cross-validation to determine the parameters.

Keywords: soft-margin support vector machine, SVM, kernel method, optimal kernel, normalized kernel, *k*-fold cross-validation

1. INTRODUCTION

In the recent years, soft-margin support vector machines (SVMs) are widely and successfully used in several remote sensing studies. In many studies, they performed more accurately than other classifiers or performed at least equally well [1-6], since soft-margin SVMs have three properties [3]:

1. They can handle large input spaces efficiently.
2. They are robust for dealing with noisy samples.
3. They can produce sparse solutions.

Received January 24, 2011; revised August 18, 2011; accepted August 30, 2011.
Communicated by Irwin King.
* The study was supported and funded by the National Science Council of Taiwan under Grants No. NSC 100-2628-E-142-001-MY3 and 99-2221-E-142-002.
+ Corresponding author.

However, the performances of soft-margin SVMs are based on choosing the proper kernel functions. For a given kernel function with some parameters, one can adjust the parameters to find different feature mappings which map samples from original space to a kernel induced feature spaces. Hence, the parameters play an important role by applying the soft-margin SVM or in the kernel method. How to determine the proper parameters of a kernel function is an import issue. Recently, some studies use a data-dependent kernel based on some basic kernel [16], *e.g.*, the Radial Basis Function (RBF) kernel, and use the optimization procedure to find the combination coefficients. However, it still needs to determine the basic kernel parameters, *e.g.*, the width in RBF kernel or the degree of the polynomial kernel. In generally, a "grid-search" on parameters, the parameters of kernel functions and the penalty parameter C which controls the magnitudes of the slack variables, of soft-margin SVMs with the k-fold cross-validation (CV) is used for choosing the parameters and prevents the over fitting problem [6, 7]. Nevertheless, it is time consuming. Furthermore, before doing a grid-search, a better region and fineness on the grid should be pre-determined.

In this paper, we will propose an automatic method for selecting the parameters of the normalized kernel function such as the RBF kernel function. The experimental results indicate that the searching efficiency is much improved and the corresponding performance is almost as good as the soft-margin SVM with the k-fold cross-validation.

The rest of the paper is organized as follows. The reviews of kernel method and SVM are introduced in section 2. The proposed search method will be introduced in section 3. The experiments on hyperspectral image data sets and UCI data sets are designed to evaluate the performances of the proposed method in section 4 and the experimental results are also reported in this section. Section 5 contains comments and conclusions.

2. REVIEWS OF KERNEL METHOD AND SOFT-MARGIN SUPPORT VECTOR MACHINES

2.1 Kernel Method

It is easier for classification if pixels are more sparsely distributed. Generally speaking, samples with high dimensionality, the number of spectral bands, potentially have better class separability. The strategy of kernel method is to embed the data from original space R^d into a feature space H, a Hilbert space with higher dimensionality, where more effective hyperplanes for classification are expected to exist in this space than in the original space. From this, we can compute the inner product of samples in the feature space directly from the original data items using a kernel function. This is based on the fact that any kernel function κ: $R^d \times R^d \to R$ satisfying the characterization of kernels [1], can be stated formally in the following theorem:

Theorem 1 Characterization of kernels. A function κ: $R^d \times R^d \to R$ which is either continuous or has a finite domain, can be decomposed

$\kappa(x, z) = \langle \phi(x), \phi(z) \rangle$, where $x, z \in R^d$,

into a feature map ϕ into a Hilbert space H applied to both its arguments followed by the evaluation of the inner product in H if and only if it is a symmetric function for which the matrices $K = [\kappa(x_i, x_j)]_{1 \le i,j \le N}$ formed by restriction to any finite subset $\{x_1, \ldots, x_N\}$ of the space R^d are positive semi-definite, *i.e.*, having no negative eigenvalues.

The following are some popular kernels.

- Linear kernel:

$$\kappa(x, z) = \langle x, z \rangle, \tag{1}$$

- Polynomial kernel:

$$\kappa(x, z, r) = (\langle x, z \rangle + 1)^r, \; r \in Z^+, \tag{2}$$

- Gaussian Radial Basis Function kernel (RBF kernel):

$$\kappa(x, z, \sigma) = \exp\left(- \frac{\|x - z\|^2}{2\sigma^2} \right), \; \sigma \in (0, \infty), \tag{3}$$

where x and z are the samples in R^d. Moreover, r and σ are the parameters which should be pre-determined of polynomial kernel and RBF kernel, respectively.

It is worth stressing here that the size of the kernel matrix is $N \times N$ and contains in each position K_{ij} the information of distance among all possible pixel pairs (x_i and x_j) measured with a suitable kernel function κ fulfilling the characterization of kernels and if we use the linear kernel, then the feature mapping ϕ is an identity map, that is, ϕ is linear. Otherwise, the feature mapping can be nonlinear. One important idea for using kernel method is without knowing the nonlinear mapping explicitly [15].

2.2 Normalized Kernel Function Kernel

In this paper, a normalized kernel function [1] is used. Given a based kernel function κ, *e.g.*, the RBF kernel or the polynomial kernel, with a parameter β, the corresponding normalized kernel is defined as

$$\bar{\kappa}(x, z, \beta) = \frac{\kappa(x, z, \beta)}{\sqrt{\kappa(x, x, \beta)}\sqrt{\kappa(z, z, \beta)}}. \tag{4}$$

One can observe that RBF kernel is a good example of a normalized kernel with a parameter β which is equal to σ, the width in the RBF kernel. Moreover, the norms of samples in the feature space are all one, since $\bar{\kappa}(x, x, \beta) = 1$ for all $x \in R^d$. This means that the samples in the feature space are on the surface of a hypersphere. Therefore, the magnitudes of distances of samples in H are of the same order of magnitudes of the angles of samples in H, and the similarity, the cosine values of angles between samples can be determined by the normalized kernel function value. Furthermore, different value of the parameter β indicates that different corresponding mapping ϕ, and, hence, the corre-

sponding feature space H is also adopted. Based on this property, we propose an automatic method for determining which value of the parameter β is the best for discrimination.

2.3 Soft-Margin Support Vector Machines

Soft-margin SVM is to find a hyperplane in the feature space, a Hilbert space H, in the middle of the most separated margins between two classes, and this hyperplane can be applied for classifying the new testing samples [1-7]. Let $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ be a set of training samples and the corresponding label set, respectively. The soft-margin SVM algorithm is performed by the following constrained minimization optimal problem:

$$\min_{w,\xi_i} \frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i, \tag{5}$$
$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, 2, \ldots, N$$

where w is a vector normal to the hyperplane in H, b is a constant such that $b/\|w\|$ represents the distance of hyperplane from the origin, ξ_i's are slack variables to control the training errors, and $C \in R^+ - \{0\}$ is a penalty parameter that permits to tune the generalization capability.

In general, an equivalent dual representation by using the Lagrange optimization is used to find the optimizer. The corresponding dual Lagrange function is defined as:

$$\max_{\alpha_i} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j), \tag{6}$$
$$\text{subject to } \sum_{i=1}^{N} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, N.$$

Once α_i are determined, any new test pattern $z \in R^d$ is associated with a forecasting label y_z,

$$y_z = \text{sgn}(\sum_{i=1}^{N} y_i \alpha_i \kappa(x_i, z) + b) \tag{7}$$

where b is chosen so that

$$y_j \left(\sum_{i-1}^{N} y_i \alpha_i \kappa(x_i, x_j) + b \right) = 1, \tag{8}$$

for any x_j with $0 < \alpha_j < C$.

There are two parameters, β and C, for soft-margin SVMs with the normalized kernel function. Which are the best for a given problem is unknown beforehand. To identify good β and C so that the classifier can accurately predict unknown samples is the main goal. A "grid-search" on β and C of soft-margin SVMs with the k-fold cross-validation (CV) is often used and prevents the overfitting problem [6, 7]. However, this approach is extremely time-consuming, especially for the large training data set situation or the high-

dimensional dataset situation. Moreover, the range and fineness of the grid could also affect the quality of the selected parameter value. Hence, in the next section, an automatic way for determining the value of β is proposed for solving this parameter selection problem.

3. PROPOSED METHOD: AUTOMATIC PARAMETER SELECTION METHOD

Suppose Ω_i is the set of training samples in class i, $i = 1, 2, \ldots, L$. There are two important properties of the normalized kernel function: (1) $\bar{\kappa}(x_i, x_i, \beta) = 1$, $\forall i = 1, \ldots, n$, *i.e.*, the norm of every sample in the feature space is 1, and (2) $-1 \le \bar{\kappa}(x_i, x_j, \beta) \le 1$, $\forall i, j = 1, \ldots, n$, *i.e.*, the cosine value of two training samples x_i and x_j in the feature space can be computed by $\bar{\kappa}(x_i, x_j, \beta)$ and it determines the similarity between these two samples.

Based on the above two observations and the concepts, two properties are desired and described as follows: (1) The samples in the same class should be mapped into the same area in the feature space and (2) the samples in the different classes should be mapped into the different areas. We want to find a proper parameter β such that

(1) $\bar{\kappa}(x, z, \beta) \approx 1$, if $x, z \in \Omega_i$, $i = 1, \ldots, L$ and
(2) $\bar{\kappa}(x, z, \beta) \approx -1$, if $x \in \Omega_i$, $z = \Omega_j$, $i \neq j$.

We use RBF kernel to illustrate these properties. If the parameter σ is close to 0, then the corresponding kernel function values are all close to 0, the lower bound of the RBF kernel function. This means that all samples in a feature space are all approximately mutually perpendicular. When σ increases, the values of the RBF kernel function with respect to the samples which are closer by applying the Euclidean distance in the original space increase fast. As σ is close to infinity, the corresponding kernel function values are all close to 1. So the samples in the feature space are close to a fixed point. Fig. 1 shows the image of the ideal distribution in the ideal feature space.

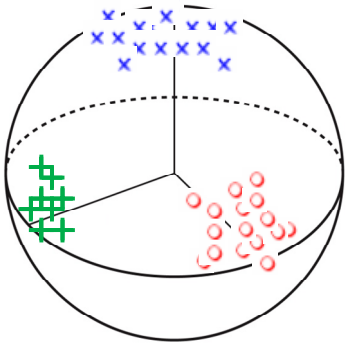


Fig. 1. The ideal distribution in the desirous feature space.

In this paper, two criteria are proposed for measuring these properties. First one is

the mean of values applied by the normal kernel function on the samples in the same class:

$$\omega(\beta) = \frac{1}{\sum_{i=1}^{L} |\Omega_i|^2} \sum_{i=1}^{L} \sum_{x \in \Omega_i} \sum_{z \in \Omega_i} \bar{\kappa}(x, z, \beta), \tag{9}$$

where $|\Omega_i|$ is the number of training samples in class i. The parameter β should be determined such that $\omega(\beta)$ closes to 1. Second one is the mean of values applied by the normalized kernel function on the samples in the different classes:

$$b(\beta) = \frac{1}{\sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} |\Omega_i| |\Omega_j|} \sum_{i=1}^{L} \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{x \in \Omega_i} \sum_{z \in \Omega_j} \bar{\kappa}(x, z, \beta). \tag{10}$$

So β should be determined also such that $b(\beta)$ closes to -1. Hence, the optimal β^* can be obtained by solving the following optimization problem:

$$\min_{\beta} J(\beta) \equiv (1 - \omega(\beta)) + (1 + b(\beta)) = 2 - \omega(\beta) + b(\beta). \tag{11}$$

The time complexity for computing $J(\beta)$ is $O(N^2 d)$ which depends on the number of training samples N and the dimensionality d.

Note that if $\bar{\kappa}(x, z, \beta)$ is differentiable, *e.g.*, the based kernel is RBF kernel function, with respect to β, the gradient descent method [10],

$$\beta_{n+1} = \beta_n - \gamma_n \nabla J(\beta_n),\ \gamma_n > 0,\ n = 1, 2, \ldots \tag{12}$$

is used to solve the proposed optimization problem, where

$$\nabla J(\beta_n) = \frac{\partial}{\partial \beta} b(\beta_n) - \frac{\partial}{\partial \beta} \omega(\beta_n), \tag{13}$$

and γ_n is the step size at the nth iteration.

Otherwise, if the parameter β is discrete, *e.g.*, the based kernel is polynomial kernel, then we can find the best β^* such that

$$\beta^* = \arg \min_{\beta} \{J(\beta) \,|\, \beta = 1, 2, \ldots, s\} \tag{14}$$

where s is an integer and should be pre-determined.

The general steps of the proposed method to the soft-margin SVM can be listed as follows.

Step 1: Find the best parameter β^* such that

$$\min_{\beta} J(\beta) \equiv (1 - \omega(\beta)) + (1 + b(\beta)) = 2 - \omega(\beta) + b(\beta). \tag{15}$$

Step 2: Use the k-fold cross-validation (CV) to find the proper parameter C of the soft-margin SVM by applying the normalized kernel function $\overline{\kappa}(x, x, \beta)$ with β^*.

4. EXPERIMENTS

In this section, the experiments in order to investigate the multiclass classification performances of the soft-margin SVMs with the normalized kernel function by applying our proposed method (OP) and the selected parameters the 5-fold cross-validation (CV) in experiments 1 and 2. Both the parameters C by applying OP and CV to the soft-margin SVMs should still be selected via 5-fold cross-validation. The polynomial kernel function and the RBF kernel function are used for as the based kernels. This study employs two measures of classification accuracy to investigate classifier performance: (1) overall classification accuracy (the percentage of the correctly classified samples for all classes) and (2) overall kappa coefficient (the percentage of the kappa coefficient for all classes).

4.1 Experiment 1

Two real data sets are applied to compare the performances in the experiment 1. They are the Indian Pine Site image, a mixed forest/agricultural site in Indiana, and the Washington, DC Mall hyperspectral image [11] as an urban site. The first one of these data sets was gathered by a sensor known as the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). The Indian Pine Site image, mounted from an aircraft flown at 65000 ft altitude and operated by the NASA/Jet Propulsion Laboratory, with the size of 145 × 145 pixels has 220 spectral bands measuring approximately 20 m across on the ground. Since the size of samples in some classes are too small to retain enough disjoint samples for training and testing, only eight classes, Cornmin, Corn-notill, Soybean-clean, Grass/Pasture, Soybeans-min, Hay-windrowed, Soybeans-notill, and Woods, were selected for the experiments.

The other dataset, Washington, DC Mall from an urban area, is a Hyperspectral Digital Imagery Collection Experiment airborne hyperspectral data flight line over the Washington, DC Mall. Two hundred and ten bands were collected in the 0.4-2.4 m region of the visible and infrared spectrum. Some water-absorption channels are discarded, resulting in 191 channels [11]. There are seven information classes, roofs, roads, trails, grass, trees, water, and shadows, in the data set.

In this experiment, for investigating the influences of training sample sizes to the dimension, three distinct cases, $|\Omega_i| = 20 < n < d$ (case 1), $|\Omega_i| = 40 < d < n$ (case 2), and $d < |\Omega_i| = 300 < n$ (case 3), will be discussed. The case 1 is a so called ill-posed classification situation, which means data dimensionality exceeds the total number of independent training samples in every class. Moreover, the case 2 is a so called poorly-posed classification situation, which means data dimensionality is greater than or comparable to the number of (independent) per-class representative training samples, but smaller than the total number of representative samples. In case 3, there are enough independent training samples. The MultiSpec [11] was used to select training and testing samples. In the Indian Pine Site dataset, all samples in the image are used as the testing samples to compute the validation measures, overall accuracies and overall kappa accuracies. In the Wash-

ington, DC Mall dataset, 100 testing samples per class, in our experiments which is the same method in [11-13].

If the RBF kernel is the based kernel, then the best β by applying CV is found in the given sets $\{2^7, 2^8, \ldots, 2^{16}\}$ and $\{2^{15}, 2^{16}, \ldots, 2^{24}\}$ in Indian Pine Site and Washington, DC Mall data sets, respectively. The parameters C by applying OP and CV in these two data sets should be selected on the set $\{2^0, 2^1, \ldots, 2^{15}\}$. If polynomial kernel is the based kernel, then the best β by applying CV is found in the given set $\{1, 2, \ldots, 32\}$.

The shape of the function $J(\beta)$ by using the Indian Pine Site dataset and the RBF kernel as the based kernel in case 1 is shown in Fig. 2. The horizontal and vertical axes are the values of the parameter β and the corresponding $J(\beta)$, respectively. This graph indicates that $J(\beta)$ has only one minimum value which is the desired selected value of σ in the proposed method. Fig. 3 shows the accuracies and kappa accuracies of testing samples and all samples in the Indian Pine Site Image at different β by applying soft-margin SVMs with a fixed C in case 1.

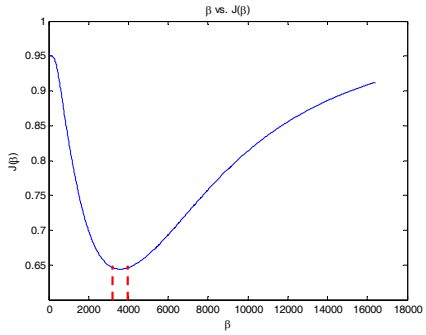


Fig. 2. The optimizer locates in the range [3500, 4000].

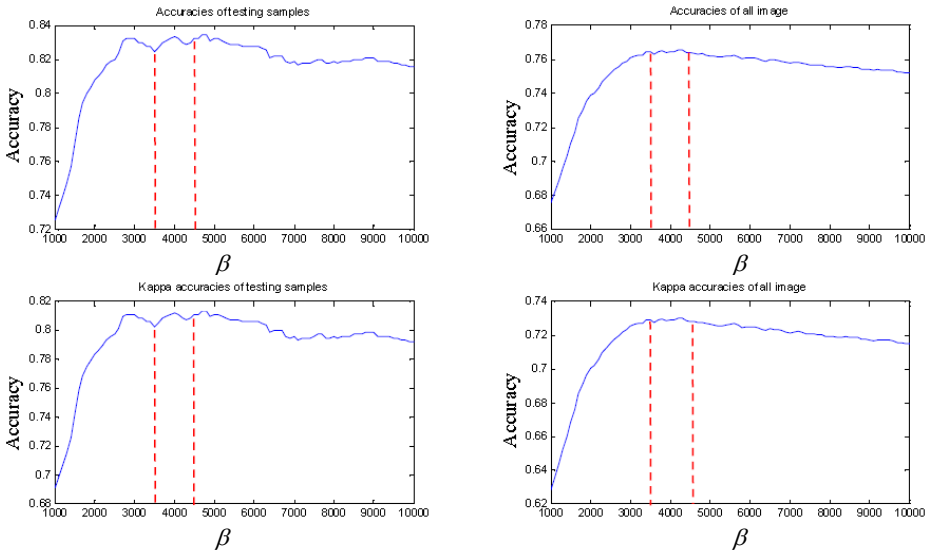
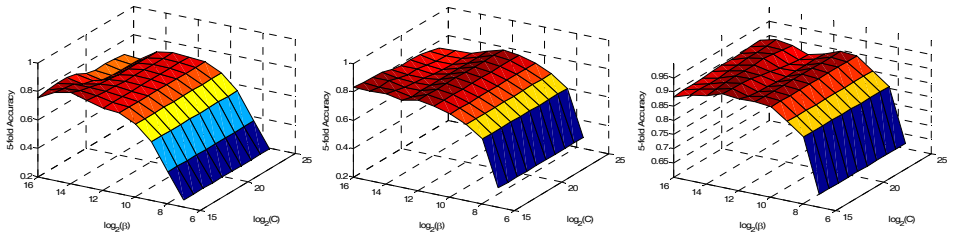


Fig. 3. There are accuracies and kappa accuracies of testing samples and all samples in the Indian Pine Site Image at different β by applying soft-margin SVMs with a fixed C in the case 1. The near optimal performances occur in the rage [3500, 4500].

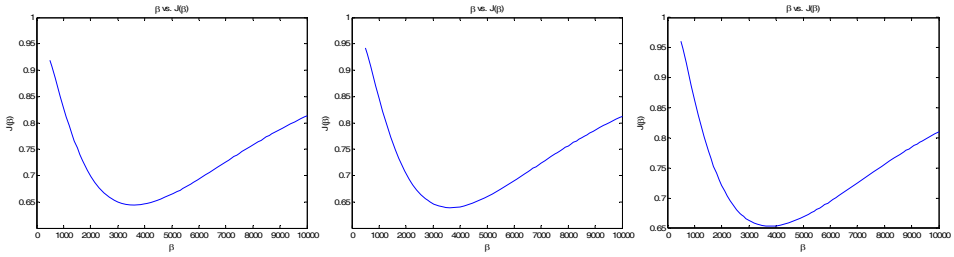
One can note that the minimum of $J(\beta)$ in Fig. 2 locates in the range [3500, 4000] and the near optimal overall and kappa accuracies of testing samples and all samples in the Indian Pine Site Image by applying SVMs with a fixed C occur in the rage [3500, 4500]. These two figures show that the proposed method obtains a proper parameter which the overall classification accuracy and kappa accuracy are near the best.

Moreover, Fig. 4 shows the 5-fold accuracies of training samples in three cases. The more are training samples, the smoother are the surfaces. Moreover, Fig. 5 shows the "β versus $J(\beta)$" in three cases. The optimizers (also in the range [3500, 4000]) are similar among three cases. However, the optimizers obtained from k-fold cross-validation (Fig. 4) are very different among three cases, since it depends on the random partitions. Hence, the proposed method could find a more suitable parameter than k-fold cross-validation, especially in the small sample size. From Fig. 4, we observe that the tendency near the optimizer is similar in every case. Therefore, it is reasonable to seek the optimal parameter β and then seek the parameter C which is following [16].



(a) (b) (c)

Fig. 4. (a), (b), and (c) are the 5-fold accuracies of training samples in cases 1, 2, and 3, respectively.



(a) (b) (c)

Fig. 5. (a), (b), and (c) are the "β versus $J(\beta)$" in cases 1, 2, and 3, respectively.

Tables 1 and 2 are the overall and kappa accuracies in Indian Pine Site dataset and Washington, DC dataset by applying RBF kernel as the based kernel function, respectively. Tables 3 and 4 are the overall and kappa accuracies in Indian Pine Site dataset and Washington, DC dataset by applying polynomial kernel as the based kernel function, respectively. One can find that the cost of time for proposed method is much less than the 5-fold cross-validation on both two datasets. Moreover, the classification results show that the soft-margin SVMs by using OP to find the parameter can obtain more accurate in the small sample size. By the way, the optimal values by OP in three cases are similar. So we can use small number of samples to approximate the optimal value.

Table 1. Overall and kappa accuracies in the Indian Pine Site dataset (the based kernel is the RBF kernel).

N_i	Method	CPU Time (sec)	β	Overall Accuracy	Overall Kappa Accuracy
20	CV	197.50	8192	0.749	0.712
	OP	21.22	3622.80	0.768	0.733
40	CV	531.25	8192	0.811	0.781
	OP	58.78	3615.36	0.831	0.804
300	CV	22859.95	4096	0.928	0.915
	OP	2416.61	3795.66	0.928	0.916

Table 2. Overall and kappa accuracies in the Washington, DC Mall dataset (the based kernel is the RBF kernel).

N_i	Method	CPU Time (sec)	β	Overall Accuracy	Overall Kappa Accuracy
20	CV	91.56	524288	0.826	0.80
	OP	9.91	178600.96	0.844	0.82
40	CV	249.64	131072	0.886	0.87
	OP	27.91	177898.80	0.881	0.86
300	CV	14191.69	2097152	0.961	0.96
	OP	1474.45	182370.06	0.951	0.94

Table 3. Overall and kappa accuracies in the Indian Pine Site dataset (the based kernel is the polynomial kernel).

N_i	Method	CPU Time (sec)	β	Overall Accuracy	Overall Kappa Accuracy
20	CV	1020.14	31	0.753	0.716
	OP	42.20	32	0.754	0.717
40	CV	4138.03	32	0.806	0.777
	OP	170.66	32	0.806	0.777
300	CV	295986.27	32	0.918	0.904
	OP	12159.09	32	0.918	0.904

Table 4. Overall and kappa accuracies in the Washington, DC Mall dataset (the based kernel is the polynomial kernel).

N_i	Method	CPU Time (sec)	β	Overall Accuracy	Overall Kappa Accuracy
20	CV	1619.03	16	0.850	0.825
	OP	50.844	4	0.857	0.833
40	CV	4848.56	2	0.874	0.853
	OP	148.906	3	0.879	0.858
300	CV	433692.27	2	0.956	0.948
	OP	14291.23	3	0.950	0.941

After comparing Tables 1-4, we can find that, in some case, the performance by using the polynomial kernel as the based kernel is better than by using the RBF kernel as the based kernel. For example, in Table 4, the overall accuracy of the Washington, DC Mall dataset with OP in case 1 is 0.857. But the overall accuracy of the Washington, DC Mall dataset with OP in case 1 in Table 2 is 0.844. Hence, in upcoming research, we can extend our proposed method based on the composite kernel, *i.e.*, a combination of the

RBF kernels and the polynomial kernels, to combine the advantages of these two basic kernel functions for classification.

4.2 Experiment 2

The four real data sets, "Ionosphere," "Monk1," "Pima," and "Iris" are described in Table 5. These data sets are available from the FTP server of the UCI [23] data repository. Similar to the experiment 1, the RBF kernel and the polynomial kernel are used as the based kernels. The grid-search ranges by applying CV method are also showed in Table 5. The columns $\beta = \sigma$ and $\beta = \gamma$ indicate the based kernel functions are the RBF kernel and the polynomial kernel, respectively. Again, the parameters C by applying OP and CV in these data sets should be selected on the sets which show in the last column of Table 5.

Table 5. Descriptions of three real data sets and the corresponding sets for CV (# of FEs means the number of features).

Dataset	Classes	# of Samples	# of FEs	$\beta = \sigma$	$\beta = \gamma$	C
Ionosphere	2	351	34	$\{2^{-5}, 2^{-4}, \ldots, 2^{14}\}$	$\{1, 2, \ldots, 7\}$	$\{2^{-30}, 2^{-29}, \ldots, 2^{30}\}$
Monk1	2	432	6	$\{2^{-3}, 2^{-2}, \ldots, 2^{5}\}$	$\{1, 2, \ldots, 30\}$	$\{2^{0}, 2^{1}, \ldots, 2^{10}\}$
Pima	2	768	8	$\{2^{-8}, 2^{-7}, \ldots, 2^{8}\}$	$\{1, 2, \ldots, 26\}$	$\{2^{0}, 2^{1}, \ldots, 2^{10}\}$
Iris	3	150	4	$\{2^{-8}, 2^{-7}, \ldots, 2^{10}\}$	$\{1, 2, \ldots, 100\}$	$\{2^{0}, 2^{1}, \ldots, 2^{20}\}$

Table 6. Overall and kappa accuracies in UCI data sets (the based kernel is the RBF kernel).

Dataset	Method	CPU Time (sec)	β	Overall Accuracy	Overall Kappa Accuracy
Ionosphere	CV	32.22	2	0.980	0.931
	OP	2.53	1.94	0.980	0.931
Monk1	CV	11.81	1	0.884	0.769
	OP	1.48	1.66	0.894	0.787
Pima	CV	318.88	128	0.755	0.489
	OP	25.02	73.83	0.786	0.557
Iris	CV	34.80	0.5	0.973	0.960
	OP	2.48	1.45	0.987	0.980

Table 7. Overall and kappa accuracies in UCI data sets (the based kernel is the polynomial kernel).

Dataset	Method	CPU Time (sec)	β	Overall Accuracy	Overall Kappa Accuracy
Ionosphere	CV	95.83	7	0.960	0.869
	OP	13.46	3	0.980	0.931
Monk1	CV	60.56	22	0.857	0.713
	OP	1.90	11	0.887	0.773
Pima	CV	2159.78	2	0.734	0.454
	OP	65.86	5	0.729	0.439
Iris	CV	556.02	11	0.987	0.980
	OP	5.69	51	0.987	0.980

From Tables 6 and 7, the proposed method OP spends much less time than CV. Moreover, the accuracies by applying OP are higher than or similar to the accuracies by applying CV.

5. CONCLUSION

In this paper, an automatic method for selecting the parameter of the normalized kernel was proposed, and we have compared it and the *k*-fold cross-validation experimentally. The experiments of two hyperspectral images and four UCI data sets show that the time cost of the proposed method is much less than the *k*-fold cross-validation. However, the classification performance by applying the soft-margin SVM with our proposed method is higher than or similar to that with the *k*-fold cross-validation. Furthermore, we will try to develop the framework to other kernel functions and apply the proposed method to kernel-based feature extraction methods, *e.g.*, generalized discriminant analysis [14] and kernel nonparametric weighted feature extraction [15].

ACKNOWLEDGMENT

The authors would like to thank Prof. Landgrebe for providing the Indian Pine Site and the Washington DC Mall data sets.

REFERENCES

1. S. T. John and C. Nello, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004.
2. F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 42, 2004, pp. 1778-1790.
3. G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 43, 2005, pp. 1351-1362.
4. G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 46, 2008, pp. 1822-1835.
5. G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*, John Wiley and Sons, Ltd., London, 2009.
6. G. Camps-Valls, L. Gómez-Chova, J. Calpe, E. Soria, J. D. Martín, L. Alonso, and J. Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 42, 2004, pp. 1530-1542.
7. C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.
8. H. L. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Optimizing the kernel in the em-

pirical feature space," *IEEE Transactions on Neural Networks*, Vol. 16, 2005, pp. 460-474.
9. B. Chen, H. Liu, and Z. Bao, "Optimizing the data-dependent kernel under a unified kernel optimization framework," *Pattern Recognition*, Vol. 41, 2007, pp. 2107-2119.
10. E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, 3rd ed., John Wiley and Sons, Inc., New York, 2008.
11. D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, Wiley, Hoboken, New Jersey, 2003.
12. J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 43, 2005, pp. 480-491.
13. B. S. Sebastiano and M. Gabriele, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 45, 2007, pp. 484-495.
14. G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, Vol. 12, 2000, pp. 2385-2404.
15. B. C. Kuo, C. H. Li, and J. M. Yang, "Kernel nonparametric weighted feature extraction for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 47, 2009, pp. 1139-1155.
16. K. P. Wu and S. D. Wang, "Choosing the kernel parameters for support vector machines by theinter-cluster," *Pattern Recognition*, Vol. 42, 2009, pp. 710-717.

Cheng-Hsuan Li (李政軒) received the B.S. and M.S. degrees from National Chung Hsing University, Taichung, Taiwan, in 2001 and 2003, respectively. He is currently working toward the Ph.D. degree in Institute of Electrical Control Engineering, National Chiao Tung University, Hsinchu, Taiwan. He is currently a research assistant with the Graduate Institute of Educational Measurement and Statistics, National Taichung University of Education, Taichung. His research interests include pattern recognition and machine learning.

Hsin-Hua Ho (何省華) received the B.S. degree from National Chung Hsing University, Taichung, Taiwan, in 2002, and the M.S. degree from National Taichung University of Education, Taichung, Taiwan in 2006. He is currently working toward the Ph.D. degree in Department of Electrical Engineer, National Chung Hsing University, Taichung, Taiwan. Her research interests include pattern recognition and machine learning.

Yu-Lung Liu (劉育隆) received the B.S. degrees from St. John's University, Taipei Campus, Taiwan, in 2005, and M.S. degree from School of Graduate Institute of Educational Measurement and Statistics, National Taichung University, in 2007. He is currently a Ph.D. candidate in the Computer Science and Information Engineering at the Asia University, Taiwan. His research interests are computerized adaptive testing, control item-exposure, U-learning, Bayesian network and machine learning.

Chin-Teng Lin (林進燈) received the B.S. degree from the National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1986, and the Ph.D. degree in Electrical Engineering from Purdue University, West Lafayette, IN, in 1992. He is currently the Chair Professor of Electrical and Computer Engineering at NCTU. His research interests include biologically inspired information systems. He has authored or coauthored more than 120 journal papers in the areas of neural networks, fuzzy systems, multimedia hardware/software, and cognitive neuroengineering, including approximately 74 IEEE journal papers.

Bor-Chen Kuo (郭伯臣) received the B.S. degree in Mathematics Education and the M.S. degree in Statistics and Assessment from National Taichung Teachers College, Taichung, Taiwan, in 1993 and 1996, respectively, and the Ph.D. degree in Electrical Engineering from Purdue University, West Lafayette, IN, in 2001. From 2002 to 2005, he was an Associate Professor with National Taichung University of Education, Taichung, where he is currently a Professor and the Director of the Graduate Institute of Educational Measurement and Statistics. His research interests include pattern recognition, feature extraction and classification of high-dimensional data, computerized adaptive testing, and nonparametric functional estimation. Dr. Kuo is a member of the Chinese Association of Psychological Testing AdCom and the Editor of the Journal of Educational Measurement and Statistics, Taiwan. He was the recipient of an Outstanding and Excellence Research Award from the Taiwan Education and Research Society in 2009.

Jin-Shiuh Taur (陶金旭) received the B.S. and M.S. degrees in Electrical Engineering from the National Taiwan University, Taipei, Taiwan, in 1987 and 1989, respectively, and the Ph.D. degree in Electrical Engineering from Princeton University, Princeton, NJ, in 1993. He was a member of the Technical Staff with Siemens Corporate Research, Inc. He is currently a Professor with the National Chung Hsing University, Taichung, Taiwan. His research interests include neural networks, pattern recognition, computer vision, and fuzzy-logic systems. Dr. Taur was the recipient of the 1996 IEEE Signal Processing Society's Best Paper Award.