Nonlinear Support Vector Machine Visualization for Risk Factor Analysis Using Nomograms and Localized Radial Basis Function Kernels

Baek Hwan Cho, Hwanjo Yu, Jongshill Lee, Young Joon Chee, In Young Kim, *Member, IEEE*, and Sun I. Kim, *Member, IEEE*

***Abstract*—Nonlinear classifiers, e.g., support vector machines (SVMs) with radial basis function (RBF) kernels, have been used widely for automatic diagnosis of diseases because of their high accuracies. However, it is difficult to visualize the classifiers, and thus difficult to provide intuitive interpretation of results to physicians. We developed a new nonlinear kernel, the localized radial basis function (LRBF) kernel, and new visualization system visualization for risk factor analysis (VRIFA) that applies a nomogram and LRBF kernel to visualize the results of nonlinear SVMs and improve the interpretability of results while maintaining high prediction accuracy. Three representative medical datasets from the University of California, Irvine repository and Statlog dataset—breast cancer, diabetes, and heart disease datasets—were used to evaluate the system. The results showed that the classification performance of the LRBF is comparable with that of the RBF, and the LRBF is easy to visualize via a nomogram. Our study also showed that the LRBF kernel is less sensitive to noise features than the RBF kernel, whereas the LRBF kernel degrades the prediction accuracy more when important features are eliminated. We demonstrated the VRIFA system, which visualizes the results of linear and nonlinear SVMs with LRBF kernels, on the three datasets.**

***Index Terms*—Decision support systems, feature selection, localized radial basis function (LRBF) kernel, nomograms, support vector machines (SVMs), visualization.**

I. Introduction

RECENT progress in data mining and machine learning has promoted computer-based approaches to solve medical problems, e.g., computer-aided diagnosis, expert systems, and prognostic studies. Support vector machines (SVMs) [1], [2], one of the most actively developed classifiers in the machine learning community, have been successfully applied to a number of medical problems [3]–[8]. For example, SVMs can build classifiers for diseases from history data, and use them to diagnose a new patient. Although SVMs support nonlinear classifiers for accurate prediction, such nonlinear classifiers present difficulty for visualization, and it is thus difficult for physicians to clearly interpret results. As a consequence, many researchers in medicine still rely on linear classifiers, such as logistic regression, for prediction problems. These are usually not as powerful as nonlinear classifiers, but are easy to interpret.

Another disadvantage of nonlinear classifiers is its limitation in feature selection. Feature selection is another important task that ranks or identifies the features that mostly affect the classification results. In a practical clinical situation, physicians may want to find risk factors for a disease, that is, they want to understand how a prediction result would change when a feature value changes. There are various feature selection methods for linear classifiers. However, they are essentially not applicable to nonlinear classifiers. To apply nonlinear modeling techniques in the medical domain, model visualization and feature selection are critical, and much appreciated by physicians.

This paper introduces a new visualization system, visualization for risk factor analysis (VRIFA), which applies a nomogram, and a localized radial basis function (LRBF) kernel in order to visualize the results of nonlinear SVMs and improve the interpretability of results while maintaining high prediction accuracy. Feature selection can also be performed using the nomogram and an SVM with the LRBF kernel. We evaluated our method on three representative medical datasets from the University of California, Irvine (UCI) repository and Statlog dataset—breast cancer, diabetes, and heart disease datasets. We demonstrated the VRIFA system, which visualizes the results of SVM classifiers, and performed feature selections on the datasets.

We first overview related feature selection methods in Section II, and the nomogram visualization method for SVM is described in Section III. Section IV explains the LRBF kernel for visualizing the nonlinear effects of features in nomograms. We take numerical experiments to show the characteristics of the kernel in Section V, and conclude with the significance of the proposed method and future plans in Section VI.

II. Feature Selection Methods

Feature selection methods can be roughly divided into three main approaches—filter, wrapper, and embedded approaches [9]. Whereas the wrapper approach ranks features using classification results, the filter approach selects highly ranked features based on a statistical score as a preprocessing step (i.e., filter methods do not consider classifiers). The embedded approach interacts with a classifier in building a classification model.

Manuscript received October 20, 2006; revised April 2, 2007. This work was supported by a Nanobiotechnology Development Project, in part by the Ministry of Science and Technology, and in part by the Republic of Korea Under Project 2005-01249.

B. H. Cho, J. Lee, Y. J. Chee, I. Y. Kim, and S. I. Kim are with the Department of Biomedical Engineering, Hanyang University, Seoul 133-605, Korea (e-mail: uranus@bme.hanyang.ac.kr; netlee@hanyang.ac.kr; yjchee@bme.hanyang.ac.kr; iykim@hanyang.ac.kr; sunkim@hanyang.ac.kr).

H. Yu was with the Department of Computer Science, University of Iowa, Iowa City, IA 52242 USA. He is now with the Department of Computer Science and Engineering, Pohang University of Science and Technology, Pohang 790-784, Korea (e-mail: hwanjoyu@postech.ac.kr).

Digital Object Identifier 10.1109/TITB.2007.902300

TABLE I
ALGORITHM OF THE RELIEFF METHOD

Line	Code
Inputs	All the instances and their class labels
1	Set all weights W[*f*] = 0
2	Set arbitrary iteration number *a*
3	**for** i = 1 **to** *a* **do**
4	Randomly select an instance *Ii*
5	Find the *k* nearest hits *Hj*
6	Find the *k* nearest misses *Mj*(*C*) for each class *C*≠*class*(*Ii*)
7	**for** *f* = 1 **to** the number of features **do**
8	$\mathbf{W}[f] = \mathbf{W}[f] - \sum_{j=1}^{k} \text{diff}(f, I_i, H_j)/(m \cdot k) + \sum_{C \neq class(I_i)} \left[\frac{P(C)}{1 - P(class(I_i))} \sum_{j=1}^{k} \text{diff}(f, I_i, M_j(C)) \right] /(m \cdot k)$
9	**end for**
10	**end for**
Outputs	Weight vector **W**[*f*]

TABLE II
ALGORITHM OF THE SENSITIVITY ANALYSIS METHOD

Line	Code
Inputs	A classifier *F(x)*
1	Set all weights W[*f*] = 0
2	Set ***O***[*a*] = 0
3	**for** *f* = 1 **to** *the number of features* **do**
4	Initialize an instance **x** = [x_1 = mean(x_1), x_2 = mean(x_2), …, x_n = mean(x_n)]
5	**for** *j* = min(x_f) **to** max(x_f) **do**
6	Set $x_f = j$
7	Set ***O***[*j*] = *F*(***x***)
8	**end for**
9	**W**[*f*] = max(***O***) − min(***O***)
10	**end for**
Outputs	Weight vector **W**[*f*]

TABLE III
ALGORITHM OF THE SVM-RFE METHOD

Line	Code
Inputs	Training instances $\boldsymbol{X}_0$ and their class labels ***y***
1	Initialize subset of surviving features ***s*** = [*1,2,...,n*]
2	Initialize feature ranking list ***r*** = **[ ]**
3	**while** (***s*** ≠ **[ ]**)
4	Restrict training instances to the subset of surviving features $\boldsymbol{X} = \boldsymbol{X}_0(:,\boldsymbol{s})$
5	Train the SVM with the restricted instances and their class labels
6	Compute the wight vector of the SVM classifier $\mathbf{w} = \sum_{i}^{sv} y_i \alpha_i \mathbf{x}_i$
7	Compute the ranking criteria $C_k = (w_k)^2$, for all *k*
8	Update the feature ranking list according to the criteria
9	Eliminate features with the lowest ranking
10	**end while**
Outputs	Feature ranking list ***r***

The wrapper and filter approaches are usually more efficient in computation than the embedded approach, as their feature selection is independent of the classification method. However, embedded methods produce more accurate results in general because they take advantage of properties of the classification method to maximize the accuracy of feature selection.

The next section reviews the main ideas of the three existing feature selection methods, ReliefF (a filter method), sensitivity analysis (a wrapper method), and recursive feature elimination (RFE) with an SVM (SVM-RFE, an embedded method).

A. ReliefF Method

A key idea in the ReliefF method is to estimate the power of each feature in increasing the interclass difference and the intraclass similarity [10], [11]. Table I represents the pseudocode of the ReliefF method. The algorithm randomly selects a sample and looks for *k* nearest "hits" (i.e., samples from the same class) and "misses" (i.e., samples from different classes) that are closest to the sample in the feature space. Then, it updates the weight for each feature f as follows

$$\mathbf{W}[f] = P(\text{different value of } f \,|\text{different class}) - P(\text{different value of } f \,|\text{same class}).$$

After several iterations, the features of high weights are considered important, as they have the greatest influence on the increase in the interclass difference and intraclass similarity.

B. Sensitivity Analysis

Sensitivity analysis is another method to rank input features in terms of their contribution to the deviation of the output [12], [13]. The pseudocode is shown in Table II. As it varies the value of a feature over a reasonable range with the other features fixed, it observes the relative changes in the outputs of the classifier. Features that produce a larger deviation in the output are considered important.

C. SVM-RFE

SVM-RFE builds (or trains) an SVM classifier, from which it computes the weight of each feature, and then it removes features of low weights [14], [15], as such features affect the classifier the least. Table III presents the outline of the algorithm. As one iterates this process of training and feature elimination, SVM-RFE finds a small subset of features that also provides an accurate SVM classifier. However, this algorithm is practically limited to the linear kernel, because it is hard to compute the weight vector from nonlinear kernels due to the kernel characteristics of the implicit mapping.

III. NOMOGRAM VISUALIZATION WITH SVM

This section explains the background and related concepts of a nomogram, such as the general concept of using a nomogram for visualizing classifiers (Section III-A), application of a nomogram to SVMs (Section III-B), and the nomogram-based feature selection method (Section III-C).

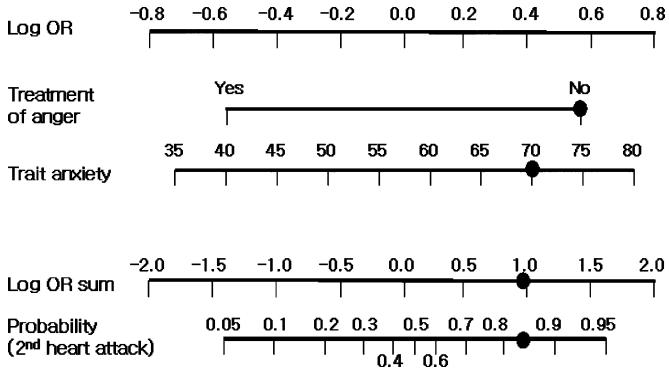

Fig. 1. Nomogram example of a classifier that predicts the probability of having a second heart attack within 1 year.

A. General Concept of a Nomogram in Classifiers

Fig. 1 illustrates an example of a nomogram visualizing a classifier. Consider that we want to predict whether a patient will have a second heart attack within 1 year. For simplicity of explanation, we consider two features only—one is whether the patient completed a treatment consisting of anger control practices, and the other is a score on a trait anxiety scale (a higher score means more anxious). In order to predict using a nomogram, the contribution of features on the scale of log odds ratios [(Log OR) (topmost axis of the nomogram)] are summed, and used to estimate the probability of having a second heart attack (bottommost axis of the nomogram). In this example, the effect of not having any anger control practice is 0.57 on the Log OR, and the trait anxiety 70 is 0.41 on the Log OR. Thus, the Log OR sum becomes $0.57 + 0.41 = 0.98$, which corresponds to a probability of 0.84 that the patient will have a second heart attack within a year.

Via the Log OR line, we can see how much each feature has impacts on the Log OR sum and the target probability. For example, if the patient has a higher score of anxiety such as 80, the probability will become around 0.9. We can also see that longer features on the nomogram will have a wider range of Log OR score, and thus, have higher impacts on the target prediction probability. For example, the trait anxiety line is longer than the anger treatment line, which implies that the probability of a second heart attack is more associated with the anxiety score. Therefore, a feature selection method based on the nomogram determines more important features according to the lengths of the lines.

B. How to Draw a Nomogram With an SVM

Jakulin *et al*. introduced a nomogram approach for visualizing an SVM that can graphically expose its internal structure, and draw the effect of each feature by means of an OR such as logistic regression [16]. The distance from a data sample $(x,\ y)$ to the separating hyperplane of the SVM is considered as an independent variable and denoted as $\delta(\mathbf{x})$. Given a kernel function $K(\mathbf{x},\ \mathbf{z})$ that returns a similarity between $\mathbf{x}$ and $\mathbf{z}$, the distance can be replaced by the decision function in the SVM as follows

$$\delta(\mathbf{x}) \cong b + \sum_{j=1}^{N} y_j \alpha_j K(\mathbf{x}, \mathrm{z}_j) \tag{1}$$

where b is the bias, α is the coefficients of support vectors $\mathbf{z}$ in the SVM, and N is the number of support vectors. When the kernel is linearly decomposable with respect to each feature, the distance becomes

$$\delta(\mathbf{x}) \cong b + \sum_{k=1}^{M} [\mathbf{w}]_k \tag{2}$$

and

$$[\mathbf{w}]_k = \sum_{j=1}^{N} y_j \alpha_j K(x_k,\ z_{j,k}) \tag{3}$$

where M is the number of features, $x_k$is $k\mathrm{th}$ feature of data vector $\mathbf{x}$, and $z_{j,k}$ is $k\mathrm{th}$ feature of $j\mathrm{th}$ support vector.

Then, by considering its class label as a dependent variable, the probability that the sample $\mathbf{x}$ belongs to the positive group (in binary classification problem) is denoted as

$$P\ (y = 1|\ \mathbf{x}) = \frac{1}{1 + e^{-(A + B \times \delta(\mathbf{x}))}}. \tag{4}$$

The parameters A and B can be calculated by optimizing the log likelihood function, as also done in the logistic regression. A cross validation is internally performed to prevent from overfitting [17].

After optimizing the parameters A and B, we can revise (4) as

$$P\ (y = 1|\ \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{k=1}^{M} [\beta]_k)}} \tag{5}$$

where $\beta_0 = A + B \times b$ and $[\beta]_k = B \times [\mathbf{w}]_k \beta_0$ is an intercept, a constant delineating the prior probability in the absence of any features, and $[\beta]_k$ is the effect vector that maps the value of the $k\mathrm{th}$ feature into a point score, which finally represents the line of the Log OR for the feature in a nomogram.

However, it is difficult to interpret the results for SVM nonlinear kernels such as the RBF kernel (RBF kernels generate highly flexible classification functions, and thus, have been most popular in practice [18], [19]). Note that a linear kernel is feasible for decomposing itself by each feature, whereas neither a polynomial kernel nor an RBF kernel is linearly decomposable.

C. Nomogram-Based RFE (Nomogram-RFE)

Since the probability mainly depends on the effect vector, we can deduce that a feature is more important when the length of its line in the nomogram is longer, as described in Fig. 1. Consequently, we can easily see the effect of each feature by drawing a nomogram with an effect vector and perform feature selection by estimating the importance of the features.

From an SVM classifier trained at each iteration rounds, nomogram-RFE calculates the lengths of lines that correspond to their features in nomograms. Similar to SVM-RFE, nomogram-RFE also recursively removes features that have a low effect

on the prediction output (i.e., short length of the line). During the iterative training process, it finds a subset of features that provides an accurate classifier.

IV. LRBF Kernel

This section explains the background of the LRBF kernel (Section IV-A), and describes the VRIFA system for visualizing SVMs with LRBF kernels (Section IV-B).

A. Definition of an LRBF Kernel

An RBF kernel has been most popularly used for nonlinear classification due to its high flexibility. SVMs with RBF kernels have proven to have infinite Vapnik–Chervonenkis (VC) dimension, that is, an SVM predication (or classification) function with an RBF kernel can shatter the dataset of an arbitrary number of data points [18]. However, SVMs with RBF kernels are hard to visualize on a nomogram because they are not additively decomposable for each feature, as shown

$$K(\mathbf{x},\mathbf{z}) = e^{(-\gamma\|\mathbf{x}-\mathbf{z}\|^2)} = \prod_{k=1}^{M} e^{(-\gamma(x_k - z_k)^2)}. \quad (6)$$

Intuitively, an RBF kernel computes a similarity between two vectors by multiplying each feature similarity [i.e., $\prod(\cdot)$], and the feature similarity is expressed by a reciprocal of exponential distance between two features [i.e., $e^{(-\gamma(x_k - z_k)^2)}$]. When it is used within an SVM, the kernel computes a similarity between a data vector $\mathbf{x}$ and a support vector $\mathbf{z}$. In other words, an SVM with an RBF kernel decides the class of a data vector based on its distance to support vectors in the feature space.

However, due to the multiplication characteristic of an RBF kernel, it is not linearly decomposable, and thus, cannot be visualized via a nomogram (note that each feature value is summed via the Log OR in the nomogram). Thus, to overcome the limitations of an RBF kernel, we apply an LRBF kernel to nomograms. This is represented as follows

$$K(\mathbf{x},\mathbf{z}) = \sum_{k=1}^{M} e^{(-\gamma(x_k - z_k)^2)}. \quad (7)$$

An LRBF kernel mimics an RBF kernel except that it localizes the effect of input data on the kernel output on each feature. An LRBF kernel is the summation of each feature similarity, whereas an RBF kernel is the multiplication; thus, it is linearly decomposable and can be visualized via a nomogram. We show that the LRBF kernel is applicable as an SVM kernel and has infinite VC dimension in the Appendix.

B. Application of an LRBF Kernel to Nomogram Visualization

Now that an LRBF kernel is additively decomposable, it is possible to apply it to nomograms. In nomograms with an LRBF kernel, the effect vector generates curves of the Log OR for the features rather than lines due to its nonlinear characteristics. Therefore, for a nomogram-RFE with an LRBF kernel, it is necessary to calculate the range of the effect values for each feature, trying to select features with wider ranges. Whereas SVM-RFE is a representative embedded feature selection method for linear SVMs, a nomogram-RFE using the VRIFA also belongs to the embedded approach category, but it works for an SVM nonlinear kernel (i.e., an LRBF kernel).

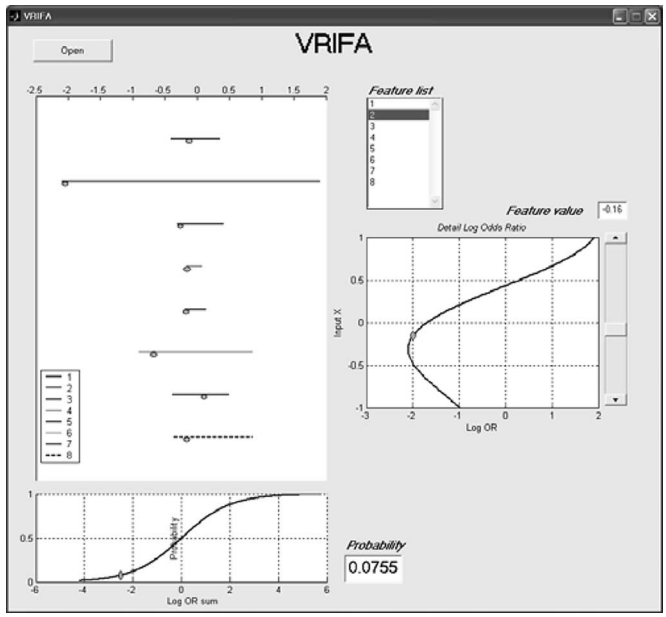

Fig. 2. VRIFA application to predict diabetes. The upper left panel shows the range of effect values for every feature and the right panel shows the detail effect values of a selected feature (feature 2 is selected in this figure). Note that an LRBF kernel is adopted in this classifier. The probability map and the final probability output with an instance are shown at the bottom.

We developed the VRIFA system using a MATLAB interface. Fig. 2 shows a screen shot of the VRIFA system on a diabetes dataset (this will be covered in detail in Section V) that visualizes the classifier constructed with an LRBF kernel. The upper left part of the system shows the range of effect values for every feature (note that there are eight features). We observe that the features have different range widths corresponding to their importance. The exact effect value of a selected feature at a specific input value can be found on the right-hand panel. The effect values of the selected feature form a curve since an LRBF kernel is employed. By applying all the effect values of the features and the intercept in (5), we can calculate the probability of having diabetes at the bottom of the screen.

V. Numerical Tests

A. Experimental Method and Metrics

We performed several experiments to evaluate the performance of an LRBF kernel and present the screen shots of the VRIFA in medical applications. A trial test was firstly carried out to compare the robustness of an LRBF kernel and an RBF kernel with synthetic datasets that include noise features in Section V-B. We measured the accuracies only as an evaluation metric in the trial test because the number of positive and negative samples were the same.

We also took numerical tests with three representative medical datasets from the UCI repository [20] and the Statlog datasets

TABLE IV
ALGORITHM TO CREATE AN ARTIFICIAL DATASET BASED ON EITHER AN RBF OR LRBF KERNEL

Line	Code
1	Create m samples with n random feature values
2	Choose 10% of the total samples for SVs
3	Put the random α [0,1] to the half of the SVs with positive label
4	Put the same α to the rest half with negative label
5	**for** $i = 1$ **to** m **do**
6	Calculate SVM output $\Sigma(\alpha \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_{sv}))$ by either RBF or LRBF kernel
7	**if** SVM output > 0 **then**
8	Put the positive label to the ith sample
9	**else**
10	Put the negative label to the ith sample
11	**end if**
12	**end for**
Outputs	Artificial dataset of m samples with n features and their class labels

[21]—breast cancer, diabetes, and heart disease datasets—in Section V-C. We compared the performance of an LRBF kernel and other kernels using various feature selection methods including ReliefF, sensitivity analysis, SVM-RFE, and nomogram-RFE. We measured the area under the curve (AUC) in the receiver operating characteristic (ROC) analysis as a main evaluation metric, as it has been widely used in the medical domain to evaluate classification algorithms or systems. We used Weka [22] and LIBSVM [19] for implementations of ReliefF and SVM.

Given the limited amount of data, we used 10-fold cross validation in all tests and employed the backward elimination method for feature selection. This removes the most unimportant feature iteratively for each learning step.

B. Trial Test With Artificial Datasets

We generated datasets artificially and labeled them by SVM classifiers randomly generated with LRBF and RBF kernels. Table IV shows the data generation algorithm.

First, we randomly created a dataset using an RBF kernel having 500 instances with five base features. Then, we added noise features one by one (the values of noise features are randomly set), trained two SVM classifiers—one with an RBF kernel and the other with an LRBF kernel—and computed the generalization performance of the classifiers. We repeated this process 30 times on 30 datasets randomly created using an RBF kernel. Fig. 3(a) shows the mean and standard deviation of the accuracies. As expected, the RBF kernel shows higher accuracy when no noise feature is added. However, as more noise features are added, the accuracy of the RBF kernel drops rapidly and becomes lower than that of the LRBF when seven noise features are added. On the other hand, the LRBF kernel shows relatively stable results with the increasing number of noise features.

We repeated the aforementioned experiments on datasets created using an LRBF kernel. Fig. 3(b) shows the results on the datasets created using an LRBF kernel. The LRBF kernel consistently outperformed the RBF kernel with any number of noise features added.

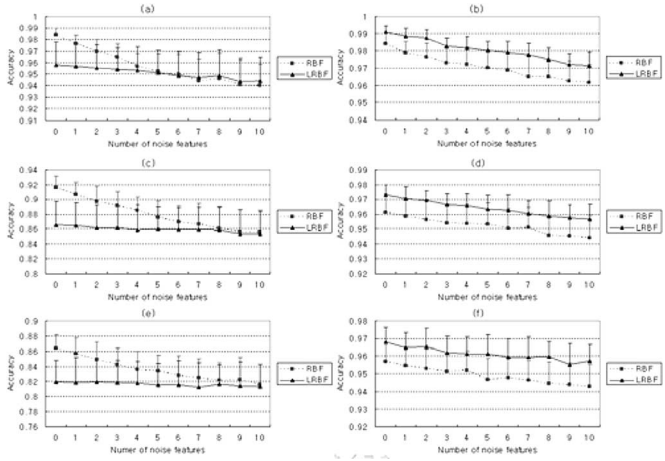

Fig. 3. The averages and standard deviations of the generalization performances for each group (30 datasets) of the artificial datasets with additive random noise features. The original number of features and base kernel, which are used to generate the dataset, for each group are as follows. (a) 5, RBF. (b) 5, LRBF. (c) 15, RBF. (d) 15, LRBF. (e) 25, RBF. (f) 25, LRBF.

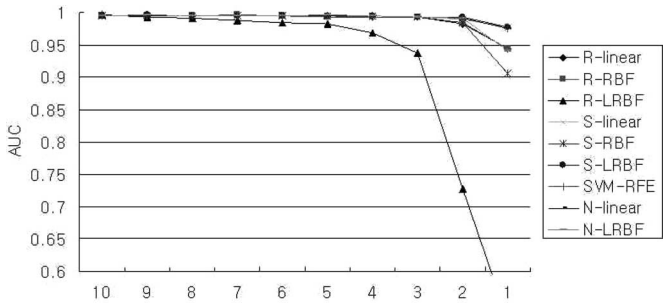

Fig. 4. Performance (AUC) variations of various feature selection methods on the breast cancer dataset. In the legend are ReliefF (R), sensitivity analysis (S), and nomogram-RFE (N).

We also repeated the aforementioned process with different numbers of base features such as 15 and 25. Fig. 3(c) and (d) shows the results with 15 base features of datasets created using an RBF and LRBF, respectively, and Fig. 3(e) and (f) shows the results with 25 base features. They all provide consistent results, that is, the LRBF kernel is less sensitive to noise features, and it may produce better generalization performance than an RBF kernel in some cases.

C. Numerical Tests With Real World Datasets

1) Breast Cancer Dataset: Fig. 4 illustrates the performance of the kernels with various feature selection methods on the breast cancer dataset (from the UCI repository), which contains 239 positive and 444 negative samples (although the original dataset has 699 samples, we excluded 16 samples that contained any missing variables) with 10 input features (sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses). From the AUC variation, we can see that all the methods can distinguish the two

TABLE V
PERFORMANCE COMPARISON IN THE BEST CASE FOR THE BREAST CANCER DATASET

Method	Kernel	No. of features	AUC	Accuracy	Sensitivity	Specificity
ReliefF	Linear	10	**0.996**	0.968	0.945	0.980
	RBF	7	**0.996**	0.968	0.950	0.977
	LRBF	10	**0.996**	0.968	0.950	0.977
Sensitivity Analysis	Linear	6	**0.996**	0.966	0.937	0.982
	RBF	10	**0.996**	0.972	0.962	0.977
	LRBF	9	**0.996**	0.969	0.950	0.980
SVM-RFE	Linear	8	**0.996**	0.968	0.946	0.980
Nomogram-RFE	Linear	10	**0.996**	0.968	0.945	0.980
	LRBF	5	**0.996**	0.968	0.945	0.980

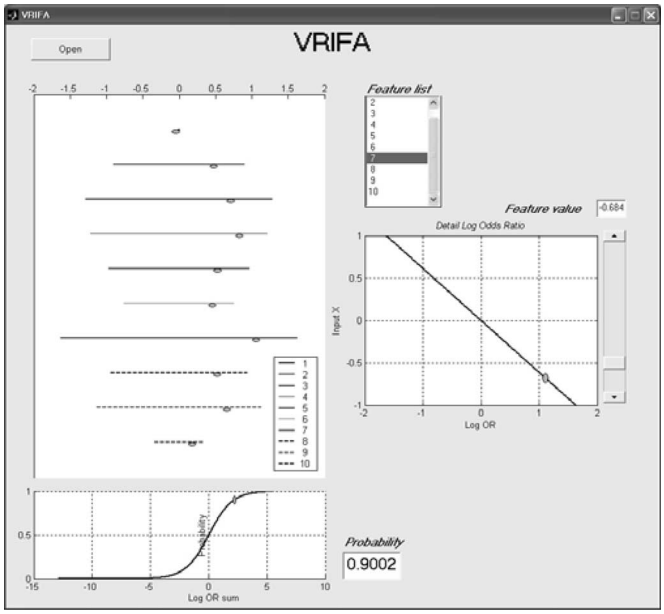

Fig. 5. VRIFA system for predicting breast cancer. The upper left panel shows the range of effect values for features and the right panel shows the detailed effect values of a selected feature (feature 7 is selected in this figure). A linear kernel is adopted in this classifier. The probability map and the final probability output with an instance are shown in the lowest panel.

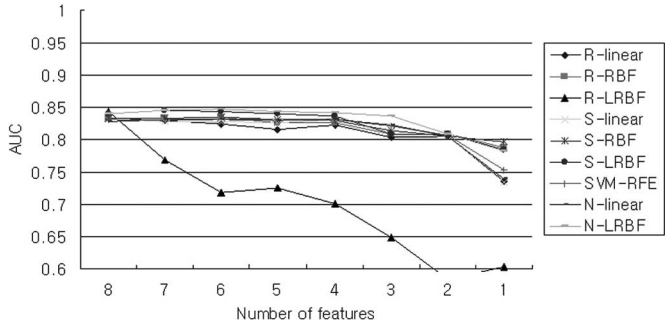

Fig. 6. Performance (AUC) variations of various feature selection methods on the diabetes dataset. In the legend are ReliefF (R), sensitivity analysis (S), and nomogram-RFE (N).

groups well, with only a small number of features. Only ReliefF with the LRBF kernel (R-LRBF in Fig. 4) does not perform as well as the other methods. Table V summarizes the best results where every method shows the same AUC, that is, where the number of selected features and other measurements differ only slightly, without statistical significance.

However, only linear and LRBF kernels (N-linear and N-LRBF in Fig. 4) can be visualized using a nomogram. Fig. 5 shows a screen shot of the VRIFA system on the breast cancer dataset, which shows the result of the SVM with a linear kernel using a nomogram. In the upper left part of the system (note that it shows all 10 features), one can observe that feature 7 has the widest range (i.e., the most important), whereas feature 1 has the narrowest range (i.e., the least important). One can easily understand that feature 1, which corresponds to the sample code number (patient ID number), is irrelevant to the prediction in the figure. In the right panel, the effect value (i.e., Log OR) of feature 7 is approximately 1.1 with the given feature value of -0.684 after normalization. Since we use a linear kernel, the effect value of feature 7 monotonically decreases when the feature value increases from -1.0 to 1.0. This linearity can be seen in all other features. With all the effect values of the features and the intercept, we can see the probability of having breast cancer in the bottom of the screen (it is 0.9002 in this case). If the patient has a score of 1.0 for feature 7, the effect value for the feature is around -1.6 and the final probability would be less than 0.5. However, any variation for feature 1 would not affect the probability much, and thus, it will be eliminated in the next training round.

2) Diabetes Dataset: This Pima Indians diabetes dataset, which is also from the UCI repository, includes 768 samples (268 positives and 500 negatives) with eight input features (number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-h serum insulin, body mass index, diabetes pedigree function, age) [23]. In Fig. 6, except for the R-LRBF method, all AUCs maintain relatively stable scores until the number of features reaches four. Table VI shows the highest AUC for each method. The LRBF kernel shows the highest score among the three kernels, and the nomogram-RFE with an LRBF kernel (N-LRBF) outperforms the others.

TABLE VI
PERFORMANCE COMPARISON IN THE BEST CASE FOR THE DIABETES DATASET

Method	Kernel	No. of features	AUC	Accuracy	Sensitivity	Specificity
ReliefF	Linear	8	0.833	0.764	0.876	0.558
	RBF	6	0.835	0.763	0.882	0.543
	LRBF	8	**0.843**	0.773	0.886	0.566
Sensitivity Analysis	Linear	8	0.833	0.770	0.887	0.549
	RBF	6	0.835	0.755	0.853	0.574
	LRBF	7	**0.845**	0.772	0.881	0.571
SVM-RFE	Linear	6	0.832	0.778	0.891	0.563
Nomogram-RFE	Linear	7	0.832	0.766	0.884	0.547
	LRBF	6	**0.847**	0.776	0.887	0.568

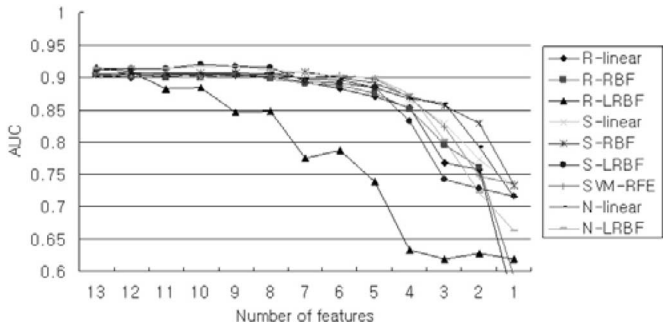

Fig. 7. Performance(AUC) variations of various feature selection methods on the heart disease dataset. In the legend are ReliefF (R), sensitivity analysis (S), and nomogram-RFE (N).

We have already seen the visualized classifier on this dataset in Fig. 2. As shown in the upper left part of the VRIFA, feature 2 (plasma glucose concentration) apparently shows the widest range among all eight features. Feature 4 (triceps skin fold thickness) has the narrowest range. As shown in the right part, feature 2 has an effect value of around -2.0 because the feature value is -0.16. Since an LRBF kernel is applied in this case, it exposes the nonlinear characteristics of the features as the feature value varies within the range -1.0 to 1.0, the effect value decreases firstly, and then, it starts increasing after the feature value goes beyond around -0.3. By summing up all the effect values and the intercept, the final probability of having diabetes is 0.0755 in this case.

D. Heart Disease Dataset

This dataset (from the Statlog dataset, available at http://www.liacc.up.pt/ML/old/statlog/datasets.html) has 120 positives and 170 negatives with 13 input features (age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels by fluoroscopy, thalassemia defect). Fig. 7 shows the performances of the methods, where the R-LRBF method also gives a more rapid decrease than the others. In Table VII, the LRBF kernel outperforms other kernels again for this dataset. Although the sensitivity analysis with the LRBF kernel (S-LRBF) shows the highest AUC, the nomogram-RFE also presents a good result.

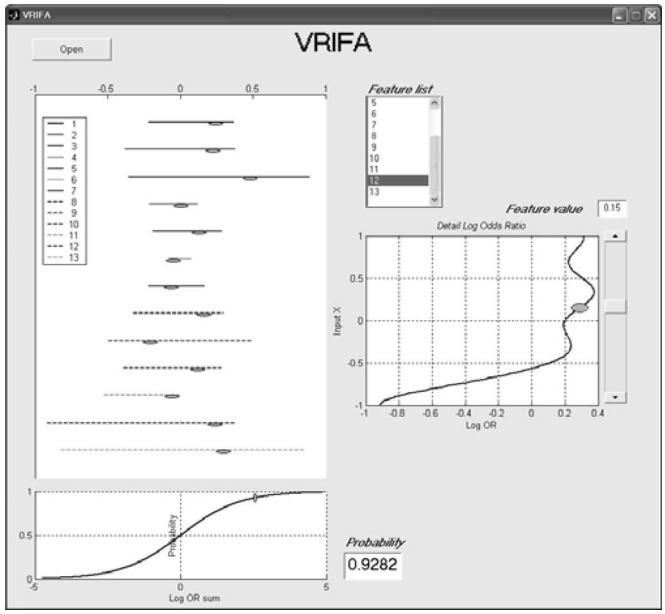

Fig. 8. VRIFA application to predict heart disease. The upper left panel shows the range of effect values for every feature and the right panel shows the detailed effect values of a selected feature (feature 12 is selected in this figure). Note that an LRBF kernel is adopted in this classifier. The probability map and the final probability output with an instance are shown at the bottom of the figure.

Fig. 8 shows the VRIFA screen shot with an LRBF kernel on this dataset. From the screen, we can intuitively interpret that feature 3 (chest pain type), 12 (number of major vessels), and 13 (thalassemia defect) are the most important, while feature 6 (fasting blood sugar) is the least important. As seen in the figure, for feature 12, the effect value increases rapidly as the feature value increases from -1.0 to -0.5. Subsequently, the effect value oscillates within a small range as the feature value increases to 1.0 due to the nonlinear characteristics of the

TABLE VII
PERFORMANCE COMPARISON IN THE BEST CASE FOR THE HEART DISEASE DATASET

Method	Kernel	No. of features	AUC	Accuracy	Sensitivity	Specificity
ReliefF	Linear	13	0.906	0.844	0.808	0.873
	RBF	12	0.907	0.830	0.792	0.860
	LRBF	13	**0.916**	0.844	0.800	0.880
Sensitivity Analysis	Linear	9	0.907	0.848	0.800	0.887
	RBF	7	0.908	0.844	0.817	0.867
	LRBF	10	**0.920**	0.844	0.800	0.880
SVM-RFE	Linear	11	0.905	0.848	0.800	0.887
Nomogram-RFE	Linear	8	0.906	0.844	0.808	0.873
	LRBF	10	**0.917**	0.848	0.792	0.893

LRBF kernel. Finally, with all the effect values, the probability of having heart disease is 0.9282 in this case.

VI. DISCUSSION AND CONCLUSION

In this paper, we have proposed the LRBF kernel method in SVM classification and visualization. An LRBF kernel copies the similarity function of an RBF function, but localizes and decomposes itself for each feature. Namely, it assumes intrafeature nonlinearity and interfeature independence; thus, it is easy to visualize via nomograms, while it captures nonlinearity of the classification function. We also prove that an LRBF kernel has an infinite VC dimension like RBF kernels. Trial tests with artificial datasets suggest that an LRBF kernel shows relatively stable performance even when some noise features are included in the classification, indicating it is less sensitive to noise features than an RBF kernel. Moreover, the trial test pointed out that an LRBF kernel could be superior to an RBF kernel for some datasets. In our numerical experiments with three medical datasets, a classifier using an LRBF kernel combined with various feature selection methods presented a better performance than linear kernels or an RBF kernel. An exception was for the breast cancer dataset, where all the three kernel methods were equal in performance in terms of the AUC in the ROC curve.

Among the feature selection methods, filter methods such as ReliefF have advantages in computation because they do not interact with classifiers. By contrast, although the wrapper and embedded methods are computationally expensive, they consider classifier design, and thus, have a higher chance of estimating the effects of the features on the classification more correctly than filter methods. The performance variation results from the three datasets commonly represented that the R-LRBF (the ReliefF method with the LRBF kernel) gave the most rapid decrease in terms of the AUC, whereas other kernels with the ReliefF method showed relatively stable performances. This implies that the LRBF kernel degrades the prediction accuracy more when important features are eliminated.

VRIFA, a visualization system using a nomogram in SVMs, has been designed and developed for a prototype system. It gives intuitive information of the effect of each feature in SVM classification and a probabilistic output. This information may be very useful for physicians to develop effective treatment strategies because they can easily understand which factors play an important role for determining patient susceptibility to a disease. Combined with the LRBF kernel, the nomogram approach can depict the nonlinear effect of features that are not susceptible to the linear kernel. This must be powerful in cases where there is a nonlinear relationship between an input feature and the target output, e.g., both high and low blood pressure are highly correlated with mortality. In addition to its graphical output, we can take advantage of the nomogram approach for a feature selection method. Even when we included an irrelevant feature (the patient ID in the breast cancer dataset) to the prediction in learning SVM, we could easily find that it had almost no effect on the classifier. In particular, the feature selection method with the LRBF kernel performed well or exceeded the performance of the other methods in our experiments.

As future work, we plan to apply the LRBF kernel method and the VRIFA to other problems in the medical domain, such as the prediction of the onset of diabetic complications. The use of an electronic medical record system has increased over the past decade, and it will make it easier to collect clinical data such as laboratory and physical examination data. The prognostic research for diseases in the area of internal medicine may need to provide practical information for risk factors. In this aspect, we expect our VRIFA system to provide detailed insights of the risk factors to medical practitioners and help them to plan efficient and proper treatment strategies.

APPENDIX

To show the applicability of a function as an SVM kernel, one must show that the function satisfies *Mercer's theorem* [1]. That is, a function must be symmetric and positive semidefinite (has nonnegative eigenvalues), in order to be used as a kernel. The LRBF is a symmetric function on $\mathbf{x}$ and positive semidefinite (has nonnegative eigenvalues), and thus, it can be used as an SVM kernel.

To show that a prediction (or classification) function with an LRBF kernel [i.e., (1), where K is (7)] is as flexible as that with an RBF kernel [i.e., (1) where K is (6)], here we prove that a classification function with an LRBF kernel also has an infinite VC dimension. That is, a classifier using an LRBF kernel can also shatter a dataset of an arbitrary number of data points [18].

Theorem 1: Assume that a dataset can be chosen arbitrarily from $\mathbf{R}^d$. Then, the family of classifiers using SVMs with LRBF kernels, and for which the error penalty is allowed to take all values, has an infinite VC dimension.

To prove this, we can consider an SVM decision function on a sample $\mathbf{s}_j$ that uses the LRBF kernel

$$f(\mathbf{s}_j) = \sum_{i}^{N} \alpha_i y_i \left(\sum_{k=1}^{M} \exp^{(-\gamma(s_{ik} - s_{jk})^2)} \right) + b. \tag{8}$$

If we choose training samples such that the smallest distance between any pair is very large, the sum on the right-hand side will be dominated when $\mathrm{s}_i = \mathbf{s}_j$. For the SVM solution, we again assume every training sample becomes an SV to let the following equalities hold:

$$\begin{aligned} \mathbf{s}_j \mathbf{w} + b &= +1, \qquad \text{for } y_j = +1 \\ \mathbf{s}_j \mathbf{w} + b &= -1, \qquad \text{for } y_j = -1. \end{aligned} \tag{9}$$

Let there be $N_+(N_-)$ positive (negative) polarity samples. If we assume that all positive (negative) samples have the same Lagrange multiplier $\alpha_+(\alpha_-)$, we can get the following equalities from the Karush–Kuhn–Tucker (KKT) condition:

$$\begin{aligned} \alpha_+ + b &= +1 \\ -\alpha_- + b &= -1 \\ N_+\alpha_+ - N_-\alpha_- &= 0 \end{aligned} \tag{10}$$

and

$$\begin{aligned} \alpha_+ &= \frac{2N_-}{N_- + N_+} \\ \alpha_- &= \frac{2N_+}{N_- + N_+} \\ b &= \frac{N_+ - N_-}{N_- + N_+}. \end{aligned} \tag{11}$$

Since the Lagrange multipliers are positive and all the KKT conditions are satisfied, we can learn the SVM without any training errors. This means the VC dimension of the LRBF kernel is infinite because the number of training samples and their class labels are arbitrary.

REFERENCES

[1] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[2] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learning*, vol. 20, pp. 273–297, 1995.

[3] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, pp. 906–914, 2000.

[4] K. Takeuchi and N. Collier, "Bio-medical entity extraction using support vector machines," *Artif. Intell. Med.*, vol. 33, pp. 125–137, 2005.

[5] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, "Learning from imbalanced data in surveillance of nosocomial infection," *Artif. Intell. Med.*, vol. 37, pp. 7–18, 2006.

[6] M. E. Mavroforakis, H. V. Georgiou, N. Dimitropuoulos, D. Cavouras, and S. Theodoridis, "Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers," *Artif. Intell. Med.*, vol. 37, pp. 145–162, 2006.

[7] T. Arodz, M. Kurdziel, E. O. D. Sevre, and D. A. Yuen, "Pattern recognition techniques for automatic detection of suspicious-looking anomalies in mammograms," *Comput. Methods Programs Biomed.*, vol. 79, pp. 135–149, 2005.

[8] L. Ramirez, N. G. Durdle, V. J. Raso, and D. L. Hill, "A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topology," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 84–91, Jan. 2006.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learning Res.*, vol. 3, pp. 1157–1182, 2003.

[10] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," presented at the ECML 1994, Catania, Italy, Apr.

[11] M. Kukar, I. Kononenko, and T. Silvester, "Machine learning in prognosis of the femoral neck fracture recovery," *Aritif. Intell. Med.*, vol. 8, pp. 431–451, 1996.

[12] M. Stevensen, R. Winter, and B. Widrow, "Sensitivity of feed forward neural networks to weight errors," *IEEE Trans. Neural Netw.*, vol. 1, no. 1, pp. 71–80, Mar. 1990.

[13] M. J. Embrechts, F. A. Arciniegas, M. Ozdemir, and R. H. Kewley, "Data mining for molecules with 2-D neural network sensitivity analysis," *Int. J. Smart Eng. Syst. Design*, vol. 5, pp. 225–239, 2003.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learning*, vol. 46, pp. 389–422, 2002.

[15] J. H. Oh, J. Gao, A. Nandi, P. Gurnani, L. Knowles, J. Schorge, and K. P. Rosenblatt, "Multicategory classification using extended SVM-RFE and markov blanket on SELDI-TOF mass spectrometry data," presented at the IEEE Symp. CIBCB 2005, San Diego, CA, Nov. 14–15, pp. 1–7.

[16] A. Jakulin, M. Mozina, J. Demsar, I. Bratko, and B. Zupan, "Nomograms for visualizing support vector machines," presented at the KDD 2005, Chicago, IL, Aug.

[17] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 1999.

[18] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, pp. 121–167, 1998.

[19] C.-C. Chang and C.-J. Lin (2001). *LIBSVM: A library for support vector machines* [Online]. Available: http://www.csie.ntu.edu.tw/˜cjlin/libsvm

[20] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, *UCI Repository of Machine Learning Databases*. Irvine, CA: Dept Inf. Comput. Sci., Univ. California, 1998.

[21] D. Michie, D. J. Spiegelhalter, and C. C. Yaylor, *Machine Learning, Neural and Statistical Classification*. London, U.K.: Ellis Horwood, 1994.

[22] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufman, 2005.

[23] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knoler, and R. S. Johannes, "Using ADAP learning algorithm to forecast the onset of diabetes mellitus," presented at the 12th Symp. Comput. Appl. Med. Care., Washington, D.C. Nov. 1988.

Baek Hwan Cho received the B.Sc. degree in electronics, communication and radio engineering, and the M.Sc. and Ph.D. degrees in biomedical engineering from Hanyang University, Seoul, Korea, in 1999, 2001, and 2007, respectively.

He is currently a Postdoctoral Researcher in the Department of Biomedical Engineering, Hanyang University. From 2001 to 2002, he was a Principal Researcher at the Psychotech Company, Ltd., Seoul, Korea, where he was engaged in research on attention rehabilitation systems. From 2005 to 2006, he was a Visiting Scholar in the Department of Computer Science, University of Iowa, Iowa City. His current research interests include medical decision support systems, data mining with heterogeneous datasets, and machine learning and visualization for various medical applications such as predicting diabetic complications.

Hwanjo Yu received the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign (UIUC), Urbana-Champaign, in 2004.

In 2004, he became an Assistant Professor at the University of Iowa, Iowa City. He is currently with the Department of Computer Science and Engineering, Pohang University of Science and Technology, Pohang, Korea. He is an Associate Editor of *Neurocomputing* and served on the National Science Foundation (NSF) panel in 2006. His current research interests include data mining, database, information systems, machine learning, support vector machines, and bioinformatics.

Dr. Yu was the recipient of two International Business Machines Corporation (IBM) Research Student Scholarship Awards in 2002 and 2003, the Association for the Advancement of Artificial Intelligence (AAAI) Student Scholarship Award in 2003, the Conference on Information and Knowledge Management (CIKM) Student Scholarship Award in 2003, and the UIUC Data Mining Research Gold Award in 2003. He has served on the program committees of the IEEE International Conference on Data Mining (ICDM), the Association for Computing Machinery (ACM) CIKM, and the Society for Industrial and Applied Mathematics (SIAM) Data Mining.

Jongshill Lee received the B.Sc., M.Sc., and Ph.D. degrees in electronic engineering from Inha University, Incheon, Korea, in 1995, 1997, and 2005, respectively.

He is currently a Research Professor in the Department of Biomedical Engineering, Hanyang University, Seoul, Korea. From 2001 to 2005, he was a Lecturer in the Department of Electronic Engineering, Korea Polytechnic University, Siheung, Korea, where he was engaged in research on digital signal processing and digital image processing. His current research interests include digital signal processing, biosignal processing, rehabilitation system, and robot vision system.

Young Joon Chee received the Ph.D. degree in biomedical engineering from Seoul National University, Seoul, Korea, in 2005.

He developed medical equipment for endoscopic surgery during his industrial working period from 1998 to 2004. He is incharge of the biomedical signal processing team at Hanyang University, Seoul, especially for cardiovascular analysis. His current research interests include instrumentation of physiological signals and its processing.

In Young Kim (M'90) received the M.D. degree from the School of Medicine, Seoul National University, Seoul, Korea, in 1989, and the Ph.D. degree from the Department of Biomedical Engineering, Seoul National University, in 1994.

He is currently an Associate Professor in the Department of Biomedical Engineering, Hanyang University, Seoul. He was a Principal Researcher in Samsung Advanced Institute of Technology, Seoul. His current research interests include medical informatics, mobile healthcare systems, and neural engineering.

Sun I. Kim (M'89) received the B.Sc. and M.Sc. degrees in electrical engineering from Seoul National University, Seoul, Korea, in 1976 and 1978, respectively, and the Ph.D. degree from the Department of Biomedical Engineering, Drexel University, Philadelphia, PA, in 1987.

He is currently a Professor and Director of the Department of Biomedical Engineering, Hanyang University, Seoul. From 1987 to 1988, he was a Research Associate in Mayo Clinic, Rochester, MN. His current research interests include virtual reality in medicine, 3-D analysis of medical images, and brain modeling.