A Kernel-Based Two-Class Classifier for Imbalanced Data Sets

Xia Hong, *Senior Member, IEEE*, Sheng Chen, *Senior Member, IEEE*, and Chris J. Harris

***Abstract*—Many kernel classifier construction algorithms adopt classification accuracy as performance metrics in model evaluation. Moreover, equal weighting is often applied to each data sample in parameter estimation. These modeling practices often become problematic if the data sets are imbalanced. We present a kernel classifier construction algorithm using orthogonal forward selection (OFS) in order to optimize the model generalization for imbalanced two-class data sets. This kernel classifier identification algorithm is based on a new regularized orthogonal weighted least squares (ROWLS) estimator and the model selection criterion of maximal leave-one-out area under curve (LOO-AUC) of the receiver operating characteristics (ROCs). It is shown that, owing to the orthogonalization procedure, the LOO-AUC can be calculated via an analytic formula based on the new regularized orthogonal weighted least squares parameter estimator, without actually splitting the estimation data set. The proposed algorithm can achieve minimal computational expense via a set of forward recursive updating formula in searching model terms with maximal incremental LOO-AUC value. Numerical examples are used to demonstrate the efficacy of the algorithm.**

***Index Terms*—Forward selection, imbalanced data sets, kernel classifier, leave-one-out (LOO) cross validation, receiver operating characteristics (ROCs).**

I. Introduction

MODEL EVALUATION in terms of good generalization performance is essential in the development and analysis of data-based learning algorithms for the construction of object classifiers. A fundamental concept in the evaluation of model generalization capability is that of cross validation [1], e.g., leave-one-out (LOO) cross validation is often used to estimate generalization error by choosing amongst different network architectures [1]. The study of classifiers for improving model generalization capability has been widely researched [2]–[10]. In most model combinatory approaches, it is assumed that a set of different classifiers with some reasonable performances are readily available. Cross validation, as required in most algorithms for model generalization evaluation, contributes significantly to computational cost. Clearly, the overall computational cost is intensive for most model combinatory approaches. Alternatively, in order to produce an individual model with good generalization for regression/classification, there has been significant research on kernel-model-based construction/selection approached, such as support vector machine (SVM), relevance vector machine (RVM), and orthogonal forward regression (OFR), etc., [11]–[14]. The efficient construction of a sparse model representation is crucial in generating a model that is easy to use and generalizes well. A class of forward orthogonal least squares (OLS) algorithms to select model regressors in a forward regression manner by virtue of their contributions to the model, which are measured by some objective function for model construction, have been developed [14]–[19]. For the construction of a sparse regression model that generalizes well, regressors are incrementally appended in an efficient forward regression procedure while minimizing the LOO errors [18], [19]. For the two-class classification problem, sparse kernel classifiers can be constructed via incrementally maximizing the Fisher ratio of class separability using the orthogonal forward selection (OFS) procedure [20], [21]. Alternatively, the data-structure-preserving criterion has been introduced as a neuron selection criterion to improve model generalization [22]. Recent work [23] has developed an OFS procedure based on minimizing the LOO misclassification rate for constructing a two-class classifier. It has been shown in [18], [19], and [23] that LOO cross validation for good model generalization can be performed efficiently, without resort to either actually splitting the estimation data set or utilizing an additional validation data set due to orthogonalization decomposition used in forward selection.

Many techniques on classifier construction, including kernel classifier of standard SVM, RVM, or OFS, are based on classification accuracy as a performance metric. This type of performance metric may break down if the class distribution is not well balanced, or if the two types of misclassification costs are skewed [9], [24]. A common problem in learning the imbalanced data sets using classification accuracy as modeling objective function is that a trained classifier tends to classify all examples to the major class [25]. Boosting algorithms have been developed [25], [26] based on the optimal setting of the margin given by imbalanced data sets, in which the parameter estimator uses unequal weights for data samples based on their margins. There has been considerable interest in cost-sensitive classification [9], [25]–[27]. An essential characteristic of cost-sensitive classifiers is that they can successfully handle imbalanced data and/or skewed misclassification cost. In [27], the cost is defined as a general performance metric including the costs of misclassification errors and experiments. In the context of SVM kernel classifiers, techniques have been introduced for imbalanced training data sets including a control sensitivity loss function [28], a kernel boundary alignment (KBA) algorithm [29],

Manuscript received July 7, 2005; revised June 12, 2006. This work was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC).

X. Hong is with the Cybernetic Intelligence Research Group, School of Systems Engineering, University of Reading, Reading RG6 6AY, U.K. (e-mail: x.hong@reading.ac.uk).

S. Chen and C. J. Harris are with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.

Digital Object Identifier 10.1109/TNN.2006.882812

and a proximal SVM algorithm [30]. A novel concept of using autoassociator neural network to be trained only by one particular class data samples has been proposed by Japkowicz [31]. The associator is applied to new data samples by comparing its reconstruction error to a threshold. In the context of k-nearest neighbor classifiers [2], a neighbor-weighted k-nearest neighbor (NKKNN) algorithm has been introduced and applied to imbalanced text categorization [32]. A theoretically sophisticated model combinatory approach known as random forest (RF) is shown to be the most accurate classifier for many benchmark data sets [5]. One remarkable feature of the RF is that it does not overfit since its generalization error converges asymptotically to a limit as the number of trees in the forest becomes large. There are other valuable original approaches developed for learning from imbalanced data sets through sampling techniques [33], [34]. The techniques of cost-sensitive learning and sampling has been introduced in an RF classifier for imbalanced data sets [35] with superior classification performance.

Fundamental to the modeling of imbalanced data sets is the receiver operating characteristics (ROCs) which is a classical methodology from signal detection theory [36], [37]. The ROC analysis has been widely used in medical diagnosis [37] and is receiving considerable attention in the machine learning research community [9]. It has been shown that the ROC approach is a powerful tool both for making practical choices and for drawing scientific conclusions [24]. The cost of misclassification is described in [9], in which the ROC convex hull (ROCCH) method has been introduced to manage (via combining or voting) classifiers. One of the performance metrics used in the ROC analysis [9], [36], [37] is to maximize the area under curve (AUC) of an ROC, which is equivalent to the probability that for a pair of randomly drawn data samples from the positive and negative groups respectively, the classifier ranks the positive sample higher than the negative sample in terms of "being positive."

In all the aforementioned OLS or OFS algorithms [14]–[21], [23], the model parameters are based on a least squares or a regularized lease squares estimator, with equal weighting for all data points in estimation. Intuitively, this type of estimator, if used for imbalanced data, will produce unfavorable classification results for the minority class. As the forward selection model construction algorithms are computationally efficient algorithms in producing parsimonious kernel models, it is then desirable to develop new forward selection model construction algorithms for building two-class classifiers from imbalanced data sets, and this is the objective of this paper. Note that in forward-orthogonal-selection-based algorithms, two cost functions are generally utilized. These are as follows: 1) a parameter estimation cost function which is used to derive parameters for candidate models, such as least squares parameter estimators; and 2) model selective criteria, which are used to decide amongst the candidate models, i.e., which term to be included into the model in each forward regression step. From previous analysis, it is seen that there are two feasible approaches in order to identify a classifier suitable for imbalanced data sets: i) for model performance criteria, the criteria as in ROC analysis are preferable to classification accuracy as model selective criteria; and ii) for parameter estimators, applying unequal weighting factors for data points instead of equal weighting as used in most least squares solutions in forward regression algorithms.

Consequently, we derive the proposed algorithm containing two elements specifically for effectively handling imbalanced data sets. First, by combining the concepts of LOO cross validation and AUC, the model generalization metrics via leave-one-out area under curve (LOO-AUC) is used as the objective function to select the model kernels in a forward selection manner. For a two-class classification problem with imbalanced data samples, and/or imbalanced misclassification cost, it is helpful to build up a parameter estimator using some cost function that is sensitive to the data sample's importance for classification [25], [26]. Another element of this paper is to introduce a new forward regression model structure selection algorithm, the forward regularized orthogonal weighted least squares algorithm (FROWLS), in which the parameter estimation cost function is sensitive to their class labels, i.e., by assigning more weights on the error due to a data sample in the minority class, but less weights on the error due to a data sample in the majority class.

This paper is organized as follows. Section II initially introduces the ROC plane and LOO-AUC which forms the basic idea of model generalization evaluation for imbalanced data sets. Section III introduces the new kernel classifier identification algorithm for imbalanced data sets, by using forward selection algorithm based on a new regularized orthogonal weighted least squares algorithm as parameter estimator and a maximal LOO-AUC as model selective criterion. It is shown that LOO-AUC can be calculated via analytic formula based on a new regularized orthogonal weighted least squares algorithm as a parameter estimator, rather that actually splitting the estimation data set owing to orthogonalization procedure. The proposed algorithm can achieve minimal computational expense via a set of forward recursive updating formulas in the search of terms with maximal LOO-AUC. Finally, the proposed two-class kernel classifier construction algorithm is presented using OFS by directly maximizing the LOO-AUC in order to optimize the model generalization for imbalanced data sets. Numerical examples are used to demonstrate the efficacy of the algorithm in Section IV and conclusions are given in Section V.

II. ROC Plane and AUC

Consider a data set $D_\pm = \{D_+, D_-\}$ consisting of N n-dimensional data samples that belong to a two-class set, i.e., $D_\pm = \{\mathbf{x}(i), t(i)\}_{i=1}^N$, where $t(i) \in [1, -1]$ is used to denote the class type for each data sample $\mathbf{x}(i) \in \Re^n$. Assume that there are N_+ positive data samples with $t(i) = 1$ and N_- negative data samples $t(i) = -1$, respectively $(N = N_+ + N_-)$. In this paper, the minority class is referred to as the positive class D_+. Let a two-class classifier $\hat{t}(\mathbf{x}) : \Re^n \rightarrow [1, -1]$ be formed using the data set. The performance of the classifier may be evaluated using the counts of data samples $\{a, b, c, d\}$ defined via the confusion matrix of Table I.

TABLE I
CONFUSION MATRIX OF A CLASSIFIER (COUNTS OF DATA SAMPLES)

	Predicted positive	Predicted negative
Actual positive	a	b
Actual negative	c	d

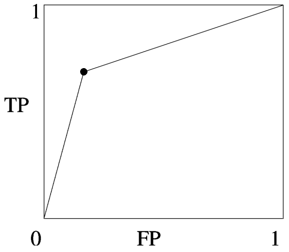

Fig. 1. ROC of a single classifier.

Clearly $N_+ = a + b$ and $N_- = c + d$. The true positive rate (TP) is the proportion of positive data samples that were correctly identified, as given by [37]

$$\text{TP} = \frac{a}{a+b} = \frac{a}{N_+}. \tag{1}$$

The false positive rate (FP) is the proportion of the negatives data samples that were incorrectly classified as positive, as calculated using [37]

$$\text{FP} = \frac{c}{c+d} = \frac{c}{N_-}. \tag{2}$$

Note that maximizing classification accuracy (CA) of

$$\text{CA} = \frac{a+d}{a+b+c+d} = \frac{a+d}{N} \tag{3}$$

is a commonly used performance metric, which is equivalent to minimizing the misclassification rate of the classifier.

Alternatively, a classifier can be mapped as a point in two-dimensional (2-D) ROC plane with coordinates as {FP, TP}, as shown in Fig. 1. By connecting this point with $\{0, 0\}$ and $\{1, 1\}$, we get the ROC curve for a nonprobabilistic or hard classifier. ROC analysis is commonly applied in visualizing model performance, decision analysis, and model combinations [9] with extensive scope and applications [36], [37]. The ROC is a graph showing the conditional probability of classifying positive samples as positive plotted against the conditional probability of classifying negative samples as positive [37]. From left to right, the ROC represents the variation of performance of the classifier if the criterion of choosing positive becomes more lenient. The ROC curve shows the tradeoff between TP and FP that acknowledges the fact that the capacity of any classifier cannot increase TP without also increasing FP. The area under the ROC curve AUC is one of the performance metrics in the ROC analysis [9], [36], [37], which is equivalent to the probability that, given a pair of randomly drawn data samples from the positive and negative groups, respectively, the classifier ranks the positive sample higher than the negative sample in terms of "being positive." Note that an AUC of 0.50 means that the diagnostic accuracy is equivalent to a pure random guess, and an AUC of 1 means that the classier distinguishes class examples perfectly. One way of selecting classifier is based on using a classifier which maximizes the AUC of a ROC, which can be calculated by

$$\text{AUC} = \frac{1 + \text{TP} - \text{FP}}{2} \tag{4}$$

for the ROC curve of a single hard classifier as shown in Fig. 1.

Equation (4) can be easily adjusted to cope with cost-sensitive classification if the misclassification costs are skewed [9], or other performance metrics derived from TP and FP are used. Clearly, the AUC of (4) is a classifier metric with a tradeoff between high TP and low FP. Note that if the data set is completely balanced with $N_+ = N_-$, then AUC = CA [see (3)]. However, for imbalanced data sets, this equivalence no longer holds. By separating the performance of a classifier into two terms that represent the performance for two classes, respectively, in comparison to the classification accuracy of (3) which only has a single term, this enables the possibility to manage the classification performance for imbalanced data.

Alternatively, cross-validation criteria are metrics that measure model's generalization capability [1]. One commonly used version of a cross validation is the LOO cross validation. The idea is that, for a given classifier model structure, each data point in the estimation data set D_N is sequentially set aside in turn, a classifier is estimated using the remaining $(N-1)$ data, and the predicted label is derived for the data point that was removed. By excluding the ith data example in estimation data set, the output of the model for the ith data example using a model estimated by using remaining $(N-1)$ data examples is denoted as $\hat{t}^{(-i)}(\mathbf{x}(i))$, or in short $\hat{t}^{(-i)}(i)$. We can use the LOO-AUC as the metrics for model generalization in terms of AUC performance, as calculated by

$$\text{AUC}^{(-)} = \frac{1 + \text{TP}^{(-)} - \text{FP}^{(-)}}{2} \tag{5}$$

where

$$\begin{aligned} \text{TP}^{(-)} &= \frac{1}{N_+} \sum_{i=1}^{N} \text{IdT}\left(\hat{t}^{(-i)}(i) \times t(i), t(i)\right) \\ \text{FP}^{(-)} &= \frac{1}{N_-} \sum_{i=1}^{N} \text{IdF}\left(\hat{t}^{(-i)}(i) \times t(i), t(i)\right) \end{aligned} \tag{6}$$

in which the indicator functions $\text{IdT}(u, v)$ for TP and $\text{IdF}(u, v)$ for FP are defined as

$$\text{IdT}(u, v) = \begin{cases} 1, & \text{if } u \geq 0 \text{ and } v = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{IdF}(u, v) = \begin{cases} 1, & \text{if } u \leq 0 \text{ and } v = -1 \\ 0, & \text{otherwise.} \end{cases}$$

In order to optimize the model structure of a classifier in terms of maximal LOO-AUC, the model selective criterion by using the maximization AUC in (5) is used in this paper. For linear in the parameter model, it has been shown [18], [19], [23] that LOO cross validation for better model generalization can be performed efficiently based on the orthogonal decomposition due to the fact that LOO errors can be derived using algebraic operations rather than actually splitting the training data set. Consequently, we achieve good model generalization performance without resort to an additional validation data set. In Section III, a new forward orthogonal weighted least squares (FOWLS) classifier construction algorithm is introduced based on the model selective criterion of maximizing LOO-AUC given by (5), in which a set of forward recursive updating formula for fast calculation of LOO-AUC is developed based on the kernel classifier model representation in orthogonal form.

III. FROWLS for Classifier Construction

Consider the kernel classifier

$$\hat{t}(i) = \mathrm{sgn}(y(i)) \text{ with } y(i) = \sum_{j=1}^{M} w_j g_j(\mathbf{x}(i)) \tag{7}$$

where $g_j(\mathbf{x}(i))$ denotes the classifier kernels. M is the number of kernels and $\hat{t}(i)$ is the estimated class label for $\mathbf{x}(i)$. $w_j, j = 1, \ldots, M$ denotes the classifier weights. The Gaussian kernel functions $g_j(\mathbf{x}) = \exp\left[-\left(\left(\|\mathbf{x}-\mathbf{c}_j\|^2\right)/(\sigma^2)\right)\right]$ are used in this paper, where the centers $\mathbf{c}_j$'s are to be selected from the full set of input vectors as a subset, and σ is the width parameter assumed to be appropriately chosen by the user.

Given the training data set $D_\pm = \{\mathbf{x}(i), t(i)\}_{i=1}^N$, define the kernel matrix $\mathbf{G} = [\mathbf{g}_1, \ldots \mathbf{g}_j, \ldots \mathbf{g}_M]$ in which $\mathbf{g}_j = [g_j(\mathbf{x}(1)), \ldots, g_j(\mathbf{x}(N))]^T$. Over the training data set, (7) can be written in vector form as

$$\mathbf{t} = \mathbf{y} + \mathbf{e} = \mathbf{G}\mathbf{w} + \mathbf{e} \tag{8}$$

where $\mathbf{t} = [t(1), \ldots, t(N)]^T$, and $\mathbf{e} = [e(1), \ldots, e(i), \ldots, e(N)]^T$ is model residual vector with $e(i) = t(i) - y(i)$. $\mathbf{w} = [w_1, \ldots w_M]^T$ is the classifier's weight vector. $\mathbf{y} = [y(1), y(2), \ldots, y(N)]^T$ is the classifier output vector. Geometrically, a set of weight vectors $\mathbf{w}$ of the kernel model defines a hyperplane of

$$\sum_{j=1}^{M} w_j g_j(\mathbf{x}) = 0 \tag{9}$$

dividing the data into two classes.

The LOO predicted label $\hat{t}^{(-i)}(i)$ of a kernel classifier of (7), is given by $\hat{t}^{(-i)}(i) = \mathrm{sgn}(y^{(-i)}(i))$, in which $y^{(-i)}(i)$ is the model prediction evaluated using the ith data sample, from a model estimated from the data set $D_\pm \setminus [\mathbf{x}(i), t(i)]$, i.e., the ith data sample $[\mathbf{x}(i), t(i)]$ is removed from $D_\pm$ for estimation. By definition, we have $\mathrm{sgn}(\hat{t}(i)) = \mathrm{sgn}(y(i))$ and $\mathrm{sgn}(\hat{t}^{(-i)}(i)) = \mathrm{sgn}(y^{(-i)}(i))$, so that

$$\mathrm{TP}^{(-)} = \frac{1}{N_+}\sum_{i=1}^{N} \mathrm{IdT}(h(i), t(i))$$
$$\mathrm{FP}^{(-)} = \frac{1}{N_-}\sum_{i=1}^{N} \mathrm{IdF}(h(i), t(i)) \tag{10}$$

with $h(i) = \hat{y}^{(-i)}(i)t(i)$.

The basic idea in the proposed algorithm is that the efficient evaluation of LOO-AUC can be achieved through the fast calculation of $h(i)$. In the following, it will be shown that the calculation of $h(i)$ can be performed efficiently without actually splitting the training data set. An analytic formula of $h(i)$ is given in Section III-B, based on a new regularized orthogonal weighted least squares (ROWLS) classifier estimator as introduced in Section III-A.

A. ROWLS Parameter Estimation

For a two-class problem, sparse kernel classifiers can be constructed using the OFS procedure [20], [21], [23]. A common feature of these algorithms is that least squares type parameter estimators have been used for parameter estimation. Note that parameter estimators that directly optimize classification performance are generally difficult to find due to the factors such as unknown probability function of the data distribution, or possibly nondifferentiable objective functions. The advantage of using a least-squares-type parameter estimator is that the classifier can be easily obtained, but the disadvantage is that they are not directly derived by optimizing the results of classification, or the cost of classification. The OFS procedure [20], [21], [23] can effectively alleviate this disadvantage as we initially use least squares type parameter estimator for generating candidate models, followed by direct evaluation of these models in terms of classification performance, i.e., minimizing the LOO misclassification rate [23]. The model selection step can, therefore, guarantee that the best model in terms of classification performance is found amongst the candidate models.

A common scenario in learning the imbalanced data sets is that the trained classifier tends to classify all examples to the major class [25]. In practice, it is often very important to have accurate classification for the minority class, e.g., in the application of abnormality detection. For this purpose, in the proposed algorithm, the maximization of LOO-AUC is used as the model selection criterion to choose from candidate models. However, in order to guarantee that the final model has a high value of LOO-AUC, it is necessary to make sure that the parameter estimation for candidate models is appropriate for imbalanced data. Intuitively, for imbalanced data, least squares or regularized lease squares estimator with equal weighting for all data points will produce unfavorable classification results for the minority class, which has a much smaller influence than the majority class to the resultant parameter estimates, and hence the decision boundary. A basic idea is that it is helpful to build up a parameter estimator using some cost function that is sensitive to data sample's importance for classification [25], [26] in order to produce better candidate models. In this paper, the parameter estimator is based on the following cost function:

$$J = \rho \sum_{\mathbf{x}(i)\in D_+} e^2(i) + \sum_{\mathbf{x}(i)\in D_-} e^2(i) \tag{11}$$

in which $\rho \geq 1$ is the weight for minority class as defined by the user through trials. Note that if $\rho = 1$, then the aforementioned cost function for parameter estimator becomes the same as used in [20], [21], and [23]. The cost function (11) gives more weights to data points in the minority class to alleviate the potential problem of classifying all examples to the major class, because, as demonstrated in Fig. 2, the previous parameter cost function has a similar effect as that of [26] in that the hyperplane as given by (9) is forced away from the minority class.

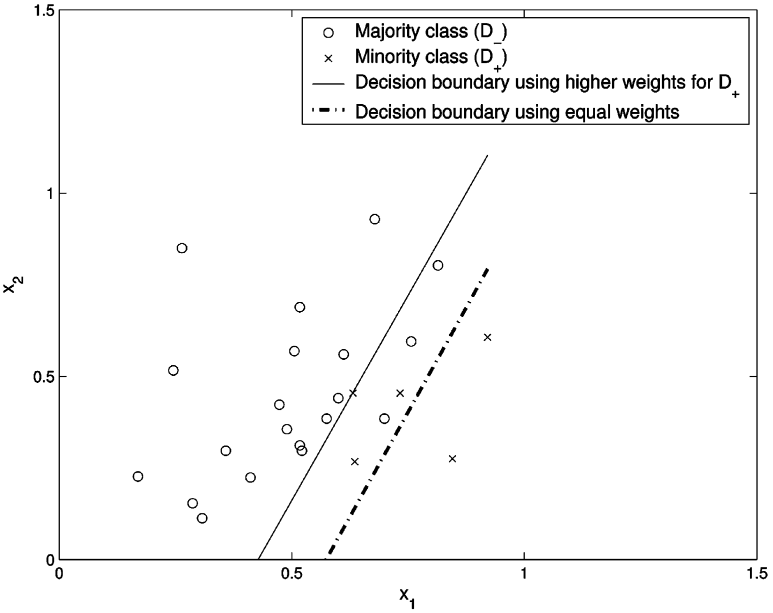

Fig. 2. Effect of using (11) as a cost function; here, a linear discriminant function is used for illustration.

Let $\mathbf{\Lambda} = \mathrm{diag}\{\lambda(1), \dots \lambda(i), \dots \lambda(N)\}$, in which

$$\lambda(i) = \begin{cases} \rho, & \text{if } \mathbf{x}(i) \in D_+ \\ 1, & \text{if } \mathbf{x}(i) \in D_- \end{cases}.$$

Then, (11) can be expressed in vector form

$$\begin{aligned} J &= \mathbf{e}^T \mathbf{\Lambda} \mathbf{e} \\ &= [\mathbf{t} - \mathbf{G}\mathbf{w}]^T \mathbf{\Lambda} [\mathbf{t} - \mathbf{G}\mathbf{w}] \\ &= [\mathbf{\Lambda}^{1/2}\mathbf{t} - \mathbf{\Lambda}^{1/2}\mathbf{G}\mathbf{w}]^T [\mathbf{\Lambda}^{1/2}\mathbf{t} - \mathbf{\Lambda}^{1/2}\mathbf{G}\mathbf{w}] \\ &= [\mathbf{\Lambda}^{1/2}\mathbf{t} - \tilde{\mathbf{G}}\mathbf{w}]^T [\mathbf{\Lambda}^{1/2}\mathbf{t} - \tilde{\mathbf{G}}\mathbf{w}] \end{aligned} \tag{12}$$

in which the equation shown at the bottom of the page holds, and the nonsingular square root matrix of $\mathbf{\Lambda}$ is denoted by $\mathbf{\Lambda}^{1/2} = \mathrm{diag}\{\sqrt{\lambda(1)}, \dots, \sqrt{\lambda(i)}, \dots \sqrt{\lambda(N)}\}$.

Denote $\tilde{\mathbf{G}} = [\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_M]$. The column and row vectors of $\tilde{\mathbf{G}}$ are represented by $\tilde{\mathbf{g}}_k = [\tilde{g}_k(1), \dots, \tilde{g}_k(N)]^T$, and $\tilde{\mathbf{g}}(i) = [\tilde{g}_1(i), \dots, \tilde{g}_M(i)]^T$, respectively. Let an orthogonal decomposition of $\tilde{\mathbf{G}}$ be $\tilde{\mathbf{G}} = \mathbf{P}\mathbf{A}$, where $\mathbf{A} \in \Re^{M \times M}$ is an upper triangular matrix. $\mathbf{P}$ is a $N \times M$ matrix with orthogonal columns that satisfy

$$\mathbf{P}^T \mathbf{P} = \mathrm{diag}(\kappa_1, \dots, \kappa_M) \tag{13}$$

with

$$\kappa_j = \mathbf{p}_j^T \mathbf{p}_j, \qquad j = 1, \dots, M. \tag{14}$$

The column and row vectors of $\mathbf{P}$ are represented by $\mathbf{p}_k = [p_k(1), \dots, p_k(N)]^T$, and $\mathbf{p}(i) = [p_1(i), \dots, p_M(i)]^T$, respectively. The aforementioned orthogonal decomposition can be realized through the Gram–Schmidt procedure by transferring nonorthogonal vectors $\tilde{\mathbf{g}}_j$, $j = 1, \dots, M$ to orthogonal vectors $\mathbf{p}_j$, $j = 1, \dots, M$. Note that

$$\begin{aligned} \mathbf{y} &= \mathbf{G}\mathbf{w} \\ &= \mathbf{\Lambda}^{-1/2} \mathbf{P} \boldsymbol{\theta} \end{aligned} \tag{15}$$

where $\boldsymbol{\theta} = \mathbf{A}\mathbf{w} = [\theta_1, \dots, \theta_M]^T$ is a weight vector in orthogonal space. Equation (15) is equivalent to

$$y(i) = \frac{1}{\sqrt{\lambda(i)}} \sum_{j=1}^{M} \theta_j p_j(i) \qquad \forall\, i. \tag{16}$$

The ROWLS parameter estimator is the parameter vector that minimizes the cost function $J_R(\boldsymbol{\theta}) = (J + \mu \boldsymbol{\theta}^T \boldsymbol{\theta})$, where $\mu \geq 0$ is the regularization parameter as set by the user

$$J_R(\boldsymbol{\theta}) = [\mathbf{\Lambda}^{1/2}\mathbf{t} - \mathbf{P}\boldsymbol{\theta}]^T [\mathbf{\Lambda}^{1/2}\mathbf{t} - \mathbf{P}\boldsymbol{\theta}] + \mu \boldsymbol{\theta}^T \boldsymbol{\theta}. \tag{17}$$

By setting $(\partial / \partial \boldsymbol{\theta}) J_R(\boldsymbol{\theta}) = \mathbf{0}$, θ_j's are obtained as

$$\theta_j = \frac{\mathbf{p}_j^T \mathbf{\Lambda}^{1/2} \mathbf{t}}{\kappa_j + \mu}, \qquad j = 1, \dots, M. \tag{18}$$

The solution for the classifier's weight vector $\mathbf{w}$ is readily available from $\mathbf{A}\mathbf{w} = \boldsymbol{\theta}$ using backward substitution.

$$\tilde{\mathbf{G}} = \mathbf{\Lambda}^{1/2} \mathbf{G} = \begin{bmatrix} \sqrt{\lambda(1)} g_1(1) & \sqrt{\lambda(1)} g_2(1) & \cdots & \sqrt{\lambda(1)} g_M(1) \\ \sqrt{\lambda(2)} g_1(2) & \sqrt{\lambda(2)} g_2(2) & \cdots & \sqrt{\lambda(2)} g_M(2) \\ \cdots & \cdots & \cdots & \cdots \\ \sqrt{\lambda(N)} g_1(N) & \sqrt{\lambda(N)} g_2(N) & \cdots & \sqrt{\lambda(N)} g_M(N) \end{bmatrix}$$

B. Analytic Formula for Evaluating $h(i)$

Based on the regularized orthogonal least squares model parameters given by (18), the model residuals can be represented by

$$\begin{aligned} e(i) &= t(i) - \frac{1}{\sqrt{\lambda(i)}}\sum_{j=1}^{M}\theta_j p_j(i) \\ &= t(i) - \frac{1}{\sqrt{\lambda(i)}}\boldsymbol{\theta}^T\mathbf{p}(i) \qquad \forall\, i. \end{aligned} \tag{19}$$

The LOO model residual is given by

$$e^{(-i)}(i) = t(i) - y^{(-i)}(i). \tag{20}$$

It has been shown [19] that for regularized orthogonal least squares estimator, the LOO model residuals can be derived using an algebraic operation rather than actually splitting the training data set based on the Sherman–Morrison–Woodbury theorem [38]. For models evaluated using regularized orthogonal weighted least square parameter estimates, it can be shown that the LOO model residuals are given by (see Appendix)

$$e^{(-i)}(i) = \frac{e(i)}{1-\mathbf{p}^T(i)[\mathbf{P}^T\mathbf{P}+\mu\mathbf{I}]^{-1}\mathbf{p}(i)} \tag{21}$$

which has the same form as in regularized orthogonal least squares parameter estimates [19].

Consequently

$$\begin{aligned} e^{(-i)}(i) &= t(i) - y^{(-i)}(i) \\ &= \frac{e(i)}{1-\mathbf{p}(i)^T[\mathbf{P}^T\mathbf{P}+\mu\mathbf{I}]^{-1}\mathbf{p}(i)} \\ &= \frac{t(i)-y(i)}{1-\sum_{j=1}^{M}\frac{p_j^2(i)}{\kappa_j+\mu}}. \end{aligned} \tag{22}$$

Hence

$$t(i) - y^{(-i)}(i) = \frac{t(i)-y(i)}{1-\sum_{j=1}^{M}\frac{p_j^2(i)}{\kappa_j+\mu}}. \tag{23}$$

Multiplying both sides of (23) with $t(i)$, and applying $t^2(i) = 1, \forall\, i$ yields

$$1 - t(i)y^{(-i)}(i) = \frac{1-t(i)y(i)}{1-\sum_{j=1}^{M}\frac{p_j^2(i)}{\kappa_j+\lambda}} \tag{24}$$

so that

$$h(i) = t(i)y^{(-i)}(i) = \frac{\frac{t(i)}{\sqrt{\lambda(i)}}\sum_{j=1}^{M}\theta_j p_j(i) - \sum_{j=1}^{M}\frac{p_j^2(i)}{\kappa_j+\mu}}{1-\sum_{j=1}^{M}\frac{p_j^2(i)}{\kappa_j+\mu}} \tag{25}$$

in which (16) was applied. The simplicity of (25) is due to: 1) the linear-in-the-parameters model structure so that (21) is valid; and 2) the orthogonal procedure to enable the matrix inversion in (21) to be applied to the diagonal matrix $[\mathbf{P}^T\mathbf{P}+\mu\mathbf{I}]$.

C. FOWLS Maximal LOO-AUC Classifier Construction Algorithm

In the following, it is shown that computational expense associated with kernel classification determination can be significantly reduced by utilizing the forward regression process via a recursive formula. In the forward regression process, the model size is configured as a growing variable k. Consider the model construction by using a subset of k regressors ($k \ll M$), that is a subset selected from the full model set consisting of M initial regressors [given by (7)] to approximate the system. By replacing M with a variable model size k, and $t(i)y^{(-i)}(i)$ with $h_k(i)$, (25) is represented by

$$h_k(i) = \frac{\frac{t(i)}{\sqrt{\lambda(i)}}\sum_{j=1}^{k}\theta_j p_j(i) - \sum_{j=1}^{k}\frac{p_j^2(i)}{\kappa_j+\mu}}{1-\sum_{j=1}^{k}\frac{p_j^2(i)}{\kappa_j+\mu}} = \frac{\alpha_k(i)}{\beta_k(i)} \tag{26}$$

where $\beta_k(i) = 1 - \sum_{j=1}^{k}\left(p_j^2(i)/\kappa_j+\mu\right)$, $\alpha_k(i) = (t(i)/(\sqrt{\lambda(i)}))\sum_{j=1}^{k}\theta_j p_j(i) - \sum_{j=1}^{k}\left(\left(p_j^2(i)\right)/(\kappa_j+\mu)\right)$.

$\alpha_k(i), \beta_k(i)$ can be represented using the following recursive formula:

$$\begin{aligned} \alpha_k(i) &= \alpha_{k-1}(i) + \frac{\theta_k p_k(i)t(i)}{\sqrt{\lambda(i)}} - \frac{p_k^2(i)}{\kappa_k+\mu} \\ \beta_k(i) &= \beta_{k-1}(i) - \frac{p_k^2(i)}{\kappa_k+\mu}. \end{aligned} \tag{27}$$

Thus, the LOO-AUC for a new model with size increased from $(k-1)$ to k is calculated by

$$\begin{aligned} \mathrm{TP}_k^{(-)} &= \frac{1}{N_+}\sum_{i=1}^{N}\mathrm{IdT}(h_k(i), t(i)) \\ \mathrm{FP}_k^{(-)} &= \frac{1}{N_-}\sum_{i=1}^{N}\mathrm{IdF}(h_k(i), t(i)) \\ \mathrm{AUC}_k^{(-)} &= \frac{1+\mathrm{TP}_k^{(-)}-\mathrm{FP}_k^{(-)}}{2} \end{aligned} \tag{28}$$

where $h_k(i) = ((\alpha_k(i))/(\beta_k(i)))$. This is advantageous in that, for a new model whose size is increased from $(k-1)$ to k, we only need to adjust both numerator $\alpha_k(i)$ and the denominator $\beta_k(i)$ based on that of the model of size $(k-1)$, with a minimal computational effort. In order to construct a sparse k-term classifier that maximizes the value of LOO-AUC as given by (5), a FOWLS model construction approach is applied that incrementally adds a kernel per forward regression step. The Gram–Schmidt procedure is used to construct the orthogonal basis $\mathbf{p}_k$ in a forward regression manner [18], [19]. At each regression step, the regressor with the maximal LOO-AUC $\left(\mathrm{AUC}_k^{(-)}\right)$ is selected.

LOO-AUC maximization-based forward Gram-Schmidt subset selection algorithm (LOO-AUC+OFS)

1) Construct $\mathbf{G}$, and $\mathbf{\Lambda}$ for given ρ. Initialize $\alpha_0(i) = 0$ and $\beta_0(i) = 1$, for $i = 1, \ldots, N$. Form $\tilde{\mathbf{G}} = \mathbf{\Lambda}^{1/2}\mathbf{G}$.

2) At the kth step where $k \geq 1$, for $1 \leq l \leq M$, $l \neq l_1, \ldots l \neq l_{k-1}$, compute

$$\begin{aligned}
a_{jk}^{(l)} &= \begin{cases} 1, & \texttt{if } j = k \\ \frac{\mathbf{p}_j^T \tilde{\mathbf{g}}_l}{\mathbf{p}_j^T \mathbf{p}_j}, & 1 \leq j < k \end{cases} \\
\mathbf{p}_k^{(l)} &= \begin{cases} \tilde{\mathbf{g}}_l, & \texttt{if } k = 1 \\ \tilde{\mathbf{g}}_l - \sum_{j=1}^{k-1} a_{jk}^{(l)} \mathbf{p}_j, & k \geq 2 \end{cases} \\
\kappa_k^{(l)} &= \left(\mathbf{p}_k^{(l)}\right)^T \mathbf{p}_k^{(l)} \\
\theta_k^{(l)} &= \frac{\left(\mathbf{p}_k^{(l)}\right)^T \mathbf{\Lambda}^{1/2}\mathbf{t}}{\kappa_k^{(l)} + \mu} \\
\alpha_k^{(l)}(i) &= \alpha_{k-1}(i) + \frac{\theta_k^{(l)} p_k^{(l)}(i) t(i)}{\sqrt{\lambda(i)}} \\
&\quad - \frac{[p_k^{(l)}(i)]^2}{\kappa_k^{(l)} + \mu}, \qquad (i = 1, \ldots, N) \\
\beta_k^{(l)}(i) &= \beta_{k-1}(i) - \frac{\left[p_k^{(l)}(i)\right]^2}{\kappa_k^{(l)} + \mu}, \qquad (i = 1, \ldots, N) \\
h_k^{(l)}(i) &= \frac{\alpha_k^{(l)}(i)}{\beta_k^{(l)}(i)}, \qquad (i = 1, \ldots, N) \\
\texttt{TP}_k^{(-,l)} &= \frac{1}{N_+} \sum_{i=1}^{N} \mathrm{IdT}\left(h_k^{(l)}(i), t(i)\right) \\
\texttt{FP}_k^{(-,l)} &= \frac{1}{N_-} \sum_{i=1}^{N} \mathrm{IdF}\left(h_k^{(l)}(i), t(i)\right) \\
\texttt{AUC}_k^{(-,l)} &= \frac{1 + \texttt{TP}^{(-,k)} - \texttt{FP}^{(-.k)}}{2}.
\end{aligned} \tag{29}$$

Find

$$l_k = \arg[\max\{\texttt{AUC}_k^{(-,\ l)}, 1 \leq l \leq M,\ \ l \neq l_1, \ldots l \neq l_{k-1}\}] \tag{30}$$

and select

$$a_{jk} = a_{jk}^{(l_k)} \quad \texttt{AUC}_k^{(-)} = \texttt{AUC}_k^{(-,l_k)} \tag{31}$$

and update

$$\begin{aligned}
\alpha_k(i) &= \alpha_k^{(l_k)}(i) \quad \beta_k(i) = \beta_k^{(l_k)}(i), \qquad (i = 1, \ldots, N) \\
\mathbf{p}_k &= \mathbf{p}_k^{(l_k)} = \begin{cases} \tilde{\mathbf{g}}_{l_k}, & \texttt{if } k = 1 \\ \tilde{\mathbf{g}}_{l_k} - \sum_{j=1}^{k-1} a_{jk} \mathbf{p}_j, & k \geq 2 \end{cases}
\end{aligned} \tag{32}$$

3) The procedure is monitored and terminated at the derived $k = n_\theta$ step, when $\texttt{AUC}_k^{(-)} \leq \texttt{AUC}_{k-1}^{(-)}$. Otherwise, set $k = k + 1$, and go to step 2).

This algorithm is presented with some predetermined parameters ρ, σ, and μ. Note that the setting of user-defined parameters more or less affect the classification performance. It seems that if the ultimate goal is the classification performance, then all parameters including ρ, σ, and μ should be also optimized. However, the task of both model structure identification of the kernel classifier and nonlinear/nondifferentiable parameters optimization simultaneously is to solve an intractable problem. In order to obtain a tradeoff between model performance and algorithmic complexity, it is a common practice to set some parameters empirically because suboptimal approaches are often preferable in practical data modeling algorithms/applications.

Remarks:

1) The parameter ρ is an important parameter introduced to increase the flexibility of dealing with different degrees of imbalances in the data sets. The effects of ρ will be investigated in Section IV through simulations. It will be shown that the empirical results conform to the initial analysis in Section III-A, e.g., models obtained by increasing ρ from 1 will increase both TP and FP until a balanced model is found.
2) The classification performance is quite robust to the width σ, as long as this is chosen in a wide range in the same scale of the input data set. Note that the input data samples should be standardized if the input variables are not in the same range. A simple way of locating a good choice of σ is to use a simple grid search.
3) Note that the regularization parameters μ may be optimized iteratively using the evidence procedure [12], [39], [19] for regression and balanced data set classification. Further research on the suitability of the procedure over imbalanced data set is necessary as the underlying assumption of the procedure may no longer be suitable for imbalanced data sets. Nevertheless, it may be useful to set the regularization parameter μ as a small positive parameter [40] to improve numerical stability and overcome overfitting for some noisy data and/or small data set.
4) An estimate of the computational cost of the algorithm at the training stage $O(n_\theta NM)$, with $n_\theta \ll M$ and $M \leq N$. For a large data set, a subset of M randomly drawn data samples from N training data samples may be used to control the computational cost. Note that the proposed classifier generally has a minimal model size bringing the advantage of the minimal computational cost when applying to new data set.

IV. Illustrative Examples

Illustrative examples are reported in the following to examine the operation of the proposed algorithm in deriving single classifiers based on incrementally maximizing LOO-AUC performance in a forward regression manner. Simulation results using k-fold cross validation are used to indicate model generalization capabilities based on multiple specifications. It is shown that the results from the proposed approach are comparable with other approaches.

Example 1: For this synthetic data set, an estimation data set with two features x_1 and x_2 was generated, with the majority class (D_-) of mean vector $[0, 0]^T$ and the covariance matrix as the identity matrix. The minority class (D_-) has a mean vector of $[2, 2]^T$ and the same identity matrix as the covariance matrix. The estimation data set consists of 100 data samples from D_- and ten data samples from D_+. A test data set with the same

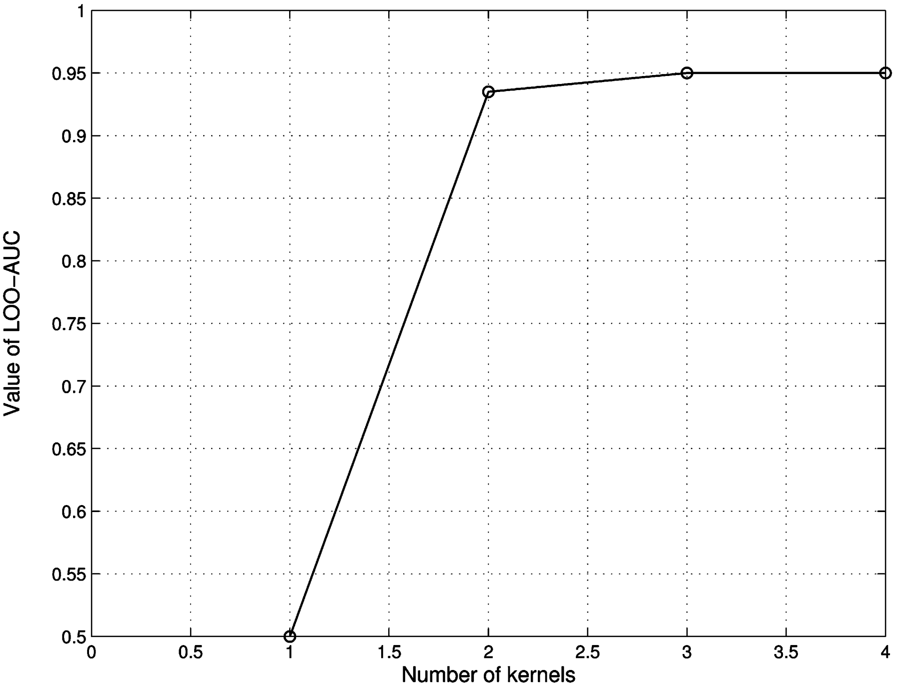

Fig. 3. Evolution of LOO-AUC versus the classifier size for the synthetic data set using the proposed algorithm.

TABLE II
GENERALIZATION PERFORMANCE OF CLASSIFICATION FOR SYNTHETIC DATA SET

	TP	FP	Precision	F-measure	G-mean	Wt.Accuracy
KRLS with all data as centres	0.840	0.037	0.694	0.760	0.899	0.901
1-NN	0.830	0.047	0.638	0.722	0.889	0.892
3-NN	0.780	**0.022**	**0.780**	0.780	0.873	0.879
LOO-AUC+OFS (ρ=1)	0.860	0.049	0.637	0.732	0.904	0.905
LOO-AUC+OFS (ρ=5)	0.840	0.028	0.75	**0.792**	0.903	0.906
LOO-AUC+OFS (ρ=10)	**0.90**	0.063	0.588	0.712	**0.918**	**0.919**
LOO-AUC+OFS (ρ=15)	0.870	0.046	0.654	0.747	0.911	0.912
LOO-AUC+OFS (ρ=20)	0.870	0.049	0.640	0.737	0.909	0.911

distribution was also generated, providing 1000 data samples for the majority class and 100 data samples for the minority class.

By using the estimation data set, the Gaussian kernel function $g_j(\mathbf{x}) = \exp[-\|\mathbf{x} - \mathbf{c}_j\|^2/\sigma^2]$ is used as basis functions, where σ is set as 1. All estimation data examples were used to form the candidate center set $\mathbf{c}_j$. The regularization parameter was set as $\mu = 1\times 10^{-3}$. Taking $\rho = 10$ for illustration, the proposed maximal LOO-AUC-based forward selection algorithm (LOO - AUC+OFS) was applied. The modeling process is illustrated in Fig. 3, in which it is seen that LOO-AUC increases over the forward regression step up to model size 3.

To illustrate the effect of the proposed algorithm of achieving a more balanced classification results for both classes, the full set of 110 kernels was used to construct a kernel classifier with regularized least squares parameter estimator (KRLS). The regularization parameter $1\times\ 10^{-3}$ was applied. Other algorithms used for comparison are the k-nearest neighbor classifiers (1-NN and 3-NN). For imbalanced data sets, there are several performance metrics widely used for comparison. All of them are based on values in the confusion matrix of Table I. The performance on TP and FP can be used to generate a single performance metric as geometric mean (G-mean), defined by

$$\text{G-mean} = \sqrt{\text{TP} \times (1 - \text{FP})}. \tag{33}$$

The G-mean value is a metric which favors balanced classification performance for two classes. In addition to TP, FP, and G-mean, there are precision [41] and F-measure [41] as defined by

$$\text{Precision} = \frac{a}{a + c} \tag{34}$$

and

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{TP}}{\text{Precision} + \text{TP}}. \tag{35}$$

Simulation results of the three methods over the test data set that has not been involved in model construction are shown in Table II and Fig. 4. From Table II, it is seen that the proposed algorithm has better tradeoff between TP and FP than both the k-nearest neighbor classifier and the conventional KRLS parameter estimator based on F-measure and G-mean values. From Fig. 4, it is clearly seen that the decision boundary of the three-term kernel classifier as derived by the proposed algorithm is further away from the minority class D_- than that of the KRLS. This example shows that the proposed algorithm has the effect of achieving better classification for the minority class and more balanced classification results for both classes. Note that the main computational cost of KRLS is the matrix inversion estimated as $O(N^3)$ for the full data sets as centers. This means

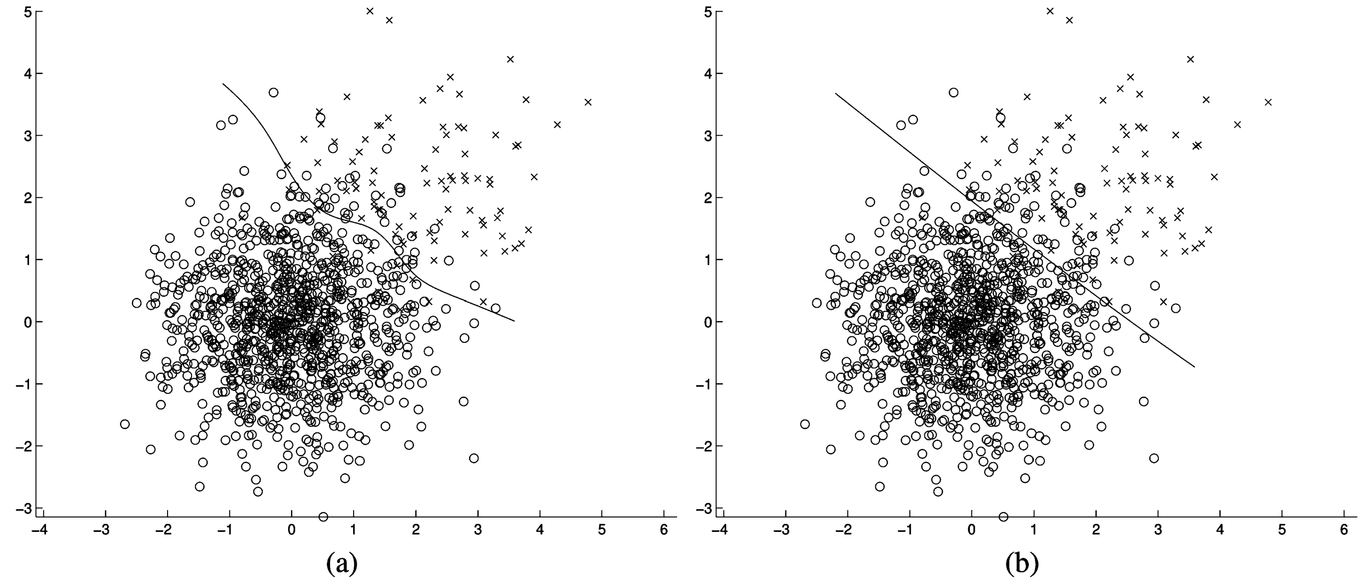

Fig. 4. Synthetic test data set and the decision boundary of (a) the regularized least squares 110-center kernel classifier with equal data weighting and (b) the derived three-term kernel classifier by the proposed algorithm.

TABLE III
EIGHTFOLD CROSS-VALIDATION CLASSIFICATION PERFORMANCE FOR PIMA INDIAN DIABETES DATA SET

	TP	FP	Precision	F-measure	G-mean	Wt.Accuracy
KRLS with all data as centres	0.56 ±0.05	0.14 ±0.04	0.68 ±0.07	0.61 ±0.04	0.69 ±0.03	0.71 ±0.02
1-NN	0.54 ±0.04	0.21 ±0.04	0.58 ±0.06	0.56 ±0.04	0.65 ±0.02	0.66 ±0.02
3-NN	0.58 ±0.06	0.17 ±0.06	0.65 ±0.07	0.61 ±0.04	0.69 ±0.04	0.70 ±0.03
LOO-AUC+OFS (ρ=1)	0.58 ±0.03	**0.13** ±0.05	**0.70** ±0.09	0.63 ±0.05	0.71 ±0.03	0.72 ±0.03
LOO-AUC+OFS (ρ=1.5)	0.68 ±0.06	0.20 ±0.07	0.65 ±0.08	0.66 ±0.05	0.73 ±0.04	0.74 ±0.04
LOO-AUC+OFS (ρ=2)	0.73 ±0.05	0.24 ±0.07	0.62 ±0.07	**0.67** ±0.05	**0.74** ±0.04	**0.74** ±0.04
LOO-AUC+OFS (ρ=2.5)	**0.77** ±0.05	0.31 ±0.06	0.57 ±0.05	0.66 ±0.07	0.73 ±0.03	0.73 ±0.03

that it is impractical for KRLS to be applied to very large data samples. In addition, both KRLS and k-nearest neighbor classifier will certainly be more computational expensive when evaluating new data samples as they all involve calculation through the full estimation data set.

Example 2: The Pima Indians diabetes data set obtained from the University of California at Irvine (UCI) repository [42] contains 768 samples from two classes with 500 negative samples and 268 positive samples. The positive class is interpreted as "tested positive for diabetes." There are eight input features for the data samples. Initially, all eight input features are normalized to the range $[0,1]$ using the operation: $((x_s(i) - \min(x_s))/(\max(x_s) - \min(x_s))), \rightarrow x_s(i)$, $\forall\ i$, for $s = 1, 2, \ldots 8$. Then, the Gaussian kernel function $g_j(\mathbf{x}) = \exp[-((\|\mathbf{x} - \mathbf{c}_j\|^2)/(\sigma^2))]$ is used as the basis functions, where σ is set as 1. As shown in Fig. 5, initially taking $\rho = 2$ and based on the whole data set, the model construction process using LOO-AUC model term selection is demonstrated by automatically deriving a kernel classifier with nine centers. The regularization parameter $\mu = 1 \times 10^{-4}$ was applied.

Then, the eightfold cross validation was used to investigate the effectiveness of the proposed algorithm. The proposed algorithm with various values of ρ were experimented with regularization parameter as $\mu = 1 \times 10^{-4}$, in comparison with alternative approaches. The results of k-nearest neighbor (1-NN and 3-NN) and a kernel classifier were used for comparison; the latter is based on the regularized least squares algorithm using all estimation data as centers and the regularization parameter of 1×10^{-4}. The results of the eightfold cross validation are shown in Table III. For models derived using equal weighting for data samples, the models have a low TP, i.e., weak detection capability for diabetes. The variation in performance in terms of TP and FP with respect to that of parameter ρ is examined. ρ is increased from 1 to 2.5 at a step of 0.5; it increased the detection capability at the cost of an increased FP. Based on G-mean and F-measures, it can be concluded that the choice of ρ around two produced better models.

The models derived using the (LOO - AUC+OFS) algorithm in eightfold cross-validation experiments have model sizes in the range of 4–24 centers.

Example 3: Haberman's survival data set obtained from the UCI repository [42] is from a study on the survival of patients who had undergone surgery for breast cancer. The data set contains data samples of 306 patients with 225 negative samples

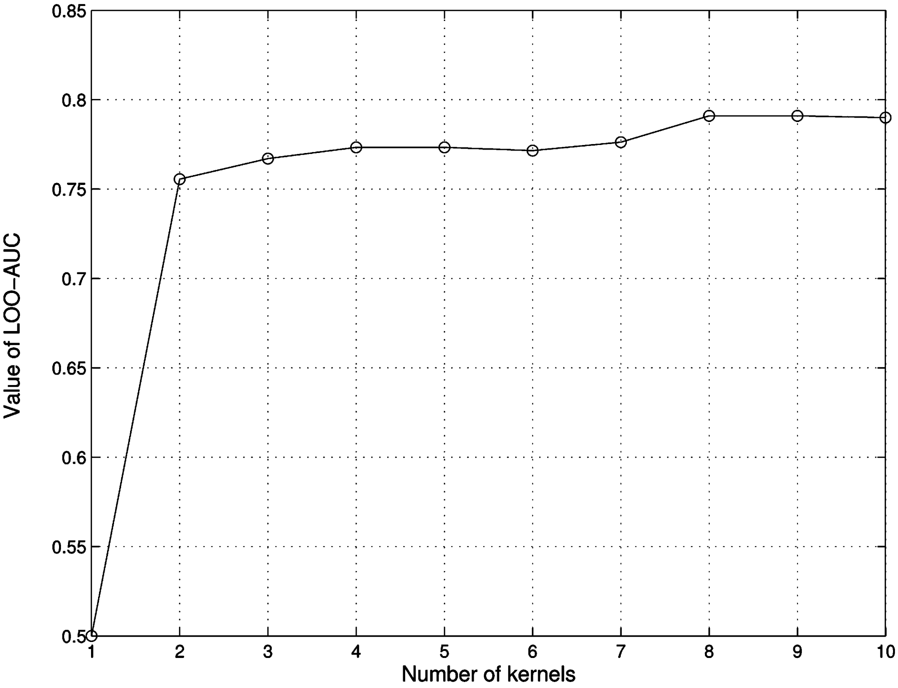

Fig. 5. Evolution of LOO-AUC versus the classifier size for Pima Indians diabetes data set using the proposed algorithm.

TABLE IV
TWOFOLD CROSS-VALIDATION CLASSIFICATION PERFORMANCE FOR HABERMAN DATA SET

	TP	FP	Precision	F-measure	G-mean	Wt.Accuracy
KRLS with all data as centres	0.33 ±0.05	0.11 ±0.01	**0.63** ±0.07	0.41 ±0.05	0.54 ±0.04	0.61 ±0.03
1-NN	0.32 ±0.03	0.21 ±0.02	0.36 ±0.01	0.38 ±0.02	0.50 ±0.02	0.56 ±0.01
3-NN	0.17 ±0.06	0.15 ±0.06	0.30 ±0.07	0.22 ±0.04	0.38 ±0.04	0.51 ±0.03
LOO-AUC+OFS (ρ=1)	0.21 ±0.02	**0.05** ±0.01	0.61 ±0.05	0.31 ±0.03	0.45 ±0.02	0.58 ±0.01
LOO-AUC+OFS (ρ=2)	0.38 ±0.08	0.13 ±0.02	0.51 ±0.02	0.44 ±0.06	0.57 ±0.05	0.63 ±0.03
LOO-AUC+OFS (ρ=3)	0.62 ±0.08	0.27 ±0.03	0.45 ±0.05	**0.52** ±0.06	**0.67** ±0.05	**0.68** ±0.05
LOO-AUC+OFS (ρ=4)	**0.67** ±0.02	0.42 ±0.08	0.36 ±0.03	0.47 ±0.02	0.62 ±0.03	0.62 ±0.03

and 81 positive samples. The positive class represents "the patients who died within five years." There are three input features for the data samples. The data set was randomly split into 50/50 for the twofold cross-validation experiments. Initially, all three input features are normalized to the range $[1, 1]$ using the operation: $((x_s(i) - \min(x_s))/(\max(x_s) - \min(x_s))), \rightarrow x_s(i)$, $\forall\ i$, for $s = 1, 2, \ldots 3$. Then, the Gaussian kernel function $g_j(\mathbf{x}) = \exp\left[-\left(\left(\|\mathbf{x} - \mathbf{c}_j\|^2\right)/(\sigma^2)\right)\right]$ is used as a basis function, where σ is set as 1. The proposed algorithm with various values of ρ was experimented with regularization parameter as $\mu = 1 \times 10^{-4}$, in comparison with the k-nearest neighbor (1-NN and 3-NN) and a kernel classifier that are based on the regularized least squares algorithm using all estimation data as centers and the regularization parameter of 1×10^{-4}. The results of the twofold cross validation are shown in Table IV. For models derived using equal weighting for data samples, the models have a low TP, i.e., weak detection capability for the positive class. Experiments on the proposed algorithm were carried out by increasing ρ from 1 to 4 at a step of 1. It is clear that TP is increased at the cost of an increased FP. Based on the G-mean and F-measures, it can be concluded that the choice of ρ around three produced better models.

Example 4: The austempered ductile iron (ADI) material data set for automotive camshaft application [43] is used to study why fatigue cracks are initiated from the graphite nodules within the microstructure. There are nine features and two class labels ("crack" and "no crack"). The data set is very imbalanced with a total of 2923 samples in which 116 samples are "crack class," and 2807 samples are "no crack class." A cost-sensitive support vector machine (CS-SVM) and a cost-sensitive SUPANOVA model [43] were applied to investigate the data set [43]. The cost-sensitive SUPANOVA model used one-norm regularization to derive a reduced model set trading model interpretability with classification performance. They used a reduced training data set of 90 "crack" and 700 "no crack" data samples.

We used random drawn data samples consisting of 90 "crack" and 700 "no crack" data samples from the data set, and the process was repeated 16 times to obtain eight

TABLE V
GENERALIZATION PERFORMANCE OF CLASSIFICATION FOR ADI DATA SET

	TP	FP	Precision	F-measure	G-mean	Wt.Accuracy
SVM $(C+ = 1000, C- = 1000)$	0.34	0.10	0.30	0.32	0.55 ±0.03	0.62
CS SVM $(C+ = 1, C- = 0.1)$	0.72	0.23	0.29	**0.42**	0.74 ±0.02	0.75
SUPANOVA $(C+ = 1, C- = 0.1)$	0.80	0.53	0.18	0.29	0.64 ±0.03	0.67
LOO-AUC+OFS (ρ=1)	0.21 ±0.03	**0.01** ±0.01	**0.67** ±0.08	0.32 ±0.04	0.46 ±0.03	0.60 ±0.02
LOO-AUC+OFS (ρ=5)	0.55 ±0.09	0.14 ±0.02	0.33 ±0.02	0.41 ±0.04	0.68 ±0.05	0.70 ±0.04
LOO-AUC+OFS (ρ=8)	0.71 ±0.07	0.23 ±0.03	0.29 ±0.01	0.41 ±0.02	0.74 ±0.03	0.74 ±0.02
LOO-AUC+OFS (ρ=10)	0.71 ±0.05	0.22 ±0.03	0.30 ±0.01	**0.42** ±0.01	0.74 ±0.02	0.74 ±0.02
LOO-AUC+OFS (ρ=12)	0.77 ±0.02	0.25 ±0.02	0.28 ±0.02	0.41 ±0.02	**0.76** ±0.01	**0.76** ±0.01
LOO-AUC+OFS (ρ=15)	0.83 ±0.02	0.29 ±0.02	0.27 ±0.01	0.40 ±0.02	**0.76** ±0.01	**0.76** ±0.01
LOO-AUC+OFS (ρ=20)	**0.88** ±0.03	0.36 ±0.04	0.24 ±0.02	0.37 ±0.02	0.75 ±0.02	0.75 ±0.02

pairs of training/testing data set in order to make a comparison with results of [43]. Initially, all nine input features are normalized to the range $[0,1]$ using the operation: $((x_s(i) - \min(x_s))/(\max(x_s) - \min(x_s))), \rightarrow x_s(i)$, $\forall i$, for $s = 1,2,\ldots 9$. Then, the Gaussian kernel function $g_j(\mathbf{x}) = \exp\left[-\left(\left(\|\mathbf{x}-\mathbf{c}_j\|^2\right)/(\sigma^2)\right)\right]$ is used as a basis function, where σ is set as 1. For each pair of training/testing data set, all the data samples in the training data set were used to form the candidate center set $\mathbf{c}_j$. The regularization parameter was set as $\mu = 1 \times 10^{-4}$. The proposed maximal LOO-AUC-based forward selection algorithm $(\mathrm{LOO-AUC{+}OFS})$ was applied while ρ was varied from 1 to 20. The classification results of the eight trials were listed in Table V. The results are quoted from [43] for comparison. The CS-SVM approach in [43] used a control sensitivity loss function [28] including weights $C+, C-$ for slacking variables of two classes, respectively. Lee [43] used a grid search by varying $C+, C-$ in the range [0.01, 10 000], and the best model in terms of best G-mean was found the model with ($C+ = 1$, $C- = 0.1$). To provide the overall performance on other criteria on precision, F-measure, G-mean, and weighted accuracy for SVM, CS-SVM, and SUPANOVA, these values are added based on Table I, (1), (2), and (33)–(35). It can be concluded that the choice of ρ around 10–12 produced better models with competitive performance with the best CS-SVM ($C+ = 1$, $C- = 0.1$). For $\rho = 12$, the model size is between 3–7.

Example 5: The Satimage data set from the UCI repository [42] obtained from [44] contains 36 attributes that are numerical values for nine neighborhood pixels in four frequencies, and six class labels [44] denoting types of soil (or crop) of the centered pixel. There are 9.73% samples of the data set from class 4 (damp grey soil) as the least prevalent class, and this was chosen to be classification target in this study (D_+). Data with other class labels were combined as a major class "not class 4," containing 90.27% samples of the data set (D_-). The total number of data samples is 6435. The data set was split into ten pairs of training/test data sets using tenfold cross validation. The size of each training data set is, therefore, $N = 5791$. For each training data set in turn, the experiments were repeated to obtain results for ρ in the range of between 1–10.

Initially, all 36 input features are normalized to the range $[0,1]$ using the operation: $((x_s(i) - \min(x_s))/(\max(x_s) - \min(x_s))), \rightarrow x_s(i), \forall i$, for $s = 1,2,\ldots 36$. Then, the Gaussian kernel function $g_j(\mathbf{x}) = \exp\left[-\left(\left(\|\mathbf{x}-\mathbf{c}_j\|^2\right)/(\sigma^2)\right)\right]$ is used as a basis function, where σ is set as 1. In order to reduce the computational cost, 20% randomly drawn training data samples ($M \approx 1158$) rather than all the training data samples were used to form the candidate center set $\mathbf{c}_j$. The regularization parameter was set very small as $\mu = 1\times 10^{-9}$ since the training data set size is large. The proposed maximal LOO-AUC-based forward selection algorithm $(\mathrm{LOO-AUC{+}OFS})$ was repeatedly applied to each training data set with different ρ. The derived model size are between 15–30. In Table VI, the tenfold cross-validation results of the proposed approach were listed together with some other imbalanced data classifiers quoted from [35] for reference. It is worth mentioning that the RF approach used in [35] is a sophisticated classification tree approach with unparallel classification accuracy [5]. Note that the model derived from LOO - AUC+OFS algorithm with a proper ρ tends to achieve high performance over minority class trading off performance in the majority class. We do not claim the superiority over other imbalanced data modeling methods, as the best tradeoff between TP and FP is dependent upon the application. The purpose of the paper is to investigate the applicability of the algorithm for imbalanced data, especially in its capability to achieve a good balanced performance through the objective function (5). Note that it is easy to modify the algorithm by using other model selective criteria, e.g., F-measure (35) in order to derive models with better performance for the specific measure of interest. The example indicates that the proposed algorithm can achieve good classifiers with simple model structure.

V. CONCLUSION

Classification algorithms have often been developed with the aim to obtain good classification accuracy, i.e., to minimize the

TABLE VI
TENFOLD CLASSIFICATION PERFORMANCE FOR SATIMAGE DATA SET

	TP	FP	Precision	F-measure	G-mean	Wt.Accuracy
Standard RIPPER	0.4743	0.0241	0.6792	0.5550	0.6803	0.7251
SMOTE 100	0.6517	0.0554	0.5588	0.5997	0.7846	0.7982
SMOTE 200	0.7489	0.0871	0.4808	0.5826	0.8268	0.8309
SMOTE-Boost 100	0.6388	**0.0198**	**0.7771**	**0.7012**	0.7913	0.8095
SMOTE-Boost 300	0.6787	0.0275	**0.7268**	**0.7019**	0.8124	0.8256
BRF cutoff1	0.6709	0.0403	0.6422	0.6562	0.8024	0.8153
BRF cutoff2	0.7700	0.0644	0.5631	0.6505	**0.8488**	**0.8528**
WRF weight1	0.6933	0.0329	0.6944	0.6938	0.8188	0.8302
WRF weight2	0.7748	0.0544	0.6055	0.6798	**0.8560**	**0.8602**
LOO-AUC+OFS ($\rho = 1$)	0.4603	**0.0230**	0.6866	0.5497	0.6689	0.7187
LOO-AUC+OFS ($\rho = 2$)	0.6337	0.0616	0.5279	0.5754	0.7708	0.7861
LOO-AUC+OFS ($\rho = 3$)	0.7626	0.0924	0.4733	0.5832	0.8315	0.8315
LOO-AUC+OFS ($\rho = 4$)	0.8236	0.1446	0.3827	0.5209	0.8389	0.8395
LOO-AUC+OFS ($\rho = 5$)	0.8492	0.1523	0.3763	0.5205	0.8482	0.8484
LOO-AUC+OFS ($\rho = 8$)	**0.8563**	0.1584	0.3732	0.5177	0.8486	0.8490
LOO-AUC+OFS ($\rho = 10$)	**0.8642**	0.1687	0.3602	0.5072	0.8470	0.8478

misclassification rate over all data samples. In real applications, imbalanced data sets are common where the objective is to achieve a high accuracy for minority class as well as that of majority class. Conventional classifiers tend to produce unfavorable classification results for minority class unless the issues caused by the imbalance between classes are addressed in an appropriate manner. A new two-class kernel classifier construction algorithm uses OFS in order to optimize the model generalization for imbalanced data sets. The new kernel classifier identification algorithm is based on a new ROWLS and the model selective criterion of a maximal LOO-AUC of the ROCs. The new regularized orthogonal weighted least squares algorithm as parameter estimator is developed in the framework of forward orthogonal selection. Based on the orthogonalization procedure, the computing of LOO-AUC is performed using analytic formula rather than actually splitting the estimation data set. Consequently, the proposed algorithm can achieve minimal computational expense via a set of forward recursive updating formula in the search of terms with maximal LOO-AUC value. Several examples have been provided to demonstrate the proposed algorithm is a viable, highly computational, and efficient alternative approach for building two-class classifier for imbalanced data sets.

APPENDIX
LOO MODELING RESIDUALS OF REGULARIZED OWLS

Following (18), the regularized parameter vector estimator based on orthogonal weighted least squares algorithm is

$$\boldsymbol{\theta} = [\mathbf{P}^T\mathbf{P} + \mu\mathbf{I}]^{-1}\mathbf{P}^T\boldsymbol{\Lambda}^{1/2}\mathbf{t} = \mathbf{D}^{-1}\mathbf{P}^T\boldsymbol{\Lambda}^{1/2}\mathbf{t} \tag{36}$$

where $\mathbf{I}$ is a unit matrix. Substitute (36) into (19) of the model residual representation to yield

$$\begin{aligned} e(i) &= t(i) - \frac{1}{\sqrt{\lambda(i)}}\boldsymbol{\theta}^T\mathbf{p}(i) \\ &= t(i) - \frac{1}{\sqrt{\lambda(i)}}\mathbf{t}^T\boldsymbol{\Lambda}^{1/2}\mathbf{P}\mathbf{D}^{-1}\mathbf{p}(i). \end{aligned} \tag{37}$$

If the data sample indexed at i is deleted from estimation data set, the regularized LOO model parameter vector based on OWLS is given by

$$\begin{aligned} \boldsymbol{\theta}^{(-i)} &= \{[\mathbf{P}^{(-i)}]^T\mathbf{P}^{(-i)} + \mu\mathbf{I}\}^{-1}[\mathbf{P}^{(-i)}]^T\{\boldsymbol{\Lambda}^{1/2}\mathbf{t}\}^{(-i)} \\ &= [\mathbf{D}^{(-i)}]^{-1}[\mathbf{P}^{(-i)}]^T\tilde{\mathbf{t}}^{(-i)} \end{aligned} \tag{38}$$

where $\tilde{\mathbf{t}}$ denotes the weighted output vector $\boldsymbol{\Lambda}^{(1/2)}\mathbf{t}$, with its ith element as $\sqrt{\lambda(i)}t(i)$. $\mathbf{P}^{(-i)}$ and $\tilde{\mathbf{t}}^{(-i)}$ denote the LOO regression matrix and weighted output vector, respectively. By derivation, it can be shown that

$$\mathbf{D}^{(-i)} = \mathbf{D} - \mathbf{p}(i)\mathbf{p}^T(i) \tag{39}$$

$$\begin{aligned} [\tilde{\mathbf{t}}^{(-i)}]^T\mathbf{P}^{(-i)} &= \tilde{\mathbf{t}}^T\mathbf{P} - \sqrt{\lambda(i)}t(i)\mathbf{p}^T(i) \\ &= \mathbf{t}^T\boldsymbol{\Lambda}^{1/2}\mathbf{P} - \sqrt{\lambda(i)}t(i)\mathbf{p}^T(i). \end{aligned} \tag{40}$$

The LOO modeling errors evaluated at i, based on the regularized orthogonal weighted least squares, are given by

$$\begin{aligned} e^{(-i)}(i) &= t(i) - y^{(-i)}(i) \\ &= t(i) - \frac{1}{\sqrt{\lambda(i)}}[\boldsymbol{\theta}^{(-i)}]^T\mathbf{p}(i) \\ &= t(i) - \frac{1}{\sqrt{\lambda(i)}}[\tilde{\mathbf{t}}^{(-i)}]^T\mathbf{P}^{(-i)}[\mathbf{D}^{(-i)}]^{-1}\mathbf{p}(i). \end{aligned} \tag{41}$$

Equation (39) and using the matrix inversion lemma yields

$$\begin{aligned} [\mathbf{D}^{(-i)}]^{-1} &= [\mathbf{D} - \mathbf{p}(i)\mathbf{p}^T(i)]^{-1} \\ &= \mathbf{D}^{-1} + \frac{\mathbf{D}^{-1}\mathbf{p}(i)\mathbf{p}^T(i)\mathbf{D}^{-1}}{1 - \mathbf{p}^T(i)\mathbf{D}^{-1}\mathbf{p}(i)} \end{aligned} \tag{42}$$

and

$$[\mathbf{D}^{(-i)}]^{-1}\mathbf{p}(i) = \frac{\mathbf{D}^{-1}\mathbf{p}(i)}{1 - \mathbf{p}^T(i)\mathbf{D}^{-1}\mathbf{p}(i)}. \tag{43}$$

Substituting (40) and (43) into (41) yields

$$\begin{aligned} e^{(-i)}(i) &= t(i) - \frac{1}{\sqrt{\lambda(i)}}[\mathbf{t}^T\boldsymbol{\Lambda}^{1/2}\mathbf{P} - \sqrt{\lambda(i)}t(i)\mathbf{p}^T(i)] \\ &\quad \times \frac{\mathbf{D}^{-1}\mathbf{p}(i)}{1 - \mathbf{p}^T(i)\mathbf{D}^{-1}\mathbf{p}(i)} \\ &= \frac{t(i) - \frac{1}{\sqrt{\lambda(i)}}\mathbf{t}^T\boldsymbol{\Lambda}^{1/2}\mathbf{P}\mathbf{D}^{-1}\mathbf{p}(i)}{1 - \mathbf{p}^T(i)\mathbf{D}^{-1}\mathbf{p}(i)}. \end{aligned} \tag{44}$$

Applying (37) to (44) yields

$$e^{(-i)}(i) = \frac{e(i)}{1-\mathbf{p}^T(i)\mathbf{D}^{-1}\mathbf{p}(i)} = \frac{e(i)}{1-\mathbf{p}^T(i)[\mathbf{P}^T\mathbf{P}+\mu\mathbf{I}]^{-1}\mathbf{p}(i)}\ . \tag{45}$$

ACKNOWLEDGMENT

The authors would like to thank Dr. P. A. S. Reed and Dr. K. K. Lee for providing the austempered ductile iron (ADI) data set and the reviewers for their valuable comments.

REFERENCES

[1] M. Stone, "Cross validatory choice and assessment of statistical predictions," *J. Roy. Stat. Soc., Ser. B*, vol. 36, pp. 117–147, 1974.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[3] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[4] ——, "Arcing classifiers," *Ann. Stat.*, vol. 26, no. 3, pp. 801–849, 1996.

[5] ——, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–12, 2001.

[6] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hint, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, pp. 79–87, 1991.

[7] C. Ji and S. Ma, "Combinations of weak classifiers," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 32–42, Jan. 1997.

[8] E. M. Kleinberg, "Stochastic discrimination," *Ann. Math. Artif. Intell.*, vol. 1, no. 1–4, pp. 207–239, 1990.

[9] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Mach. Learn.*, vol. 42, pp. 203–231, 2001.

[10] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

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

[12] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[13] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machine, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press, 2002.

[14] X. Hong and C. J. Harris, "Nonlinear model structure design and construction using orthogonal least squares and D-optimality design," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1245–1250, Sep. 2001.

[15] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their applications to non-linear system identification," *Int. J. Contr.*, vol. 50, pp. 1873–1896, 1989.

[16] S. Chen, Y. Wu, and B. L. Luk, "Combined genetic algorithm optimization and regularized orthogonal least squares learning for radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1239–1243, Sep. 1999.

[17] S. Chen, X. Hong, and C. J. Harris, "Sparse kernel regression modelling using combined locally regularized orthogonal least squares and D-optimality experimental design," *IEEE Trans. Autom. Control*, vol. 48, no. 6, pp. 1029–1036, Jun. 2003.

[18] X. Hong, P. M. Sharkey, and K. Warwick, "A robust nonlinear identification algorithm using press statistic and forward regression," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 454–458, Mar. 2003.

[19] S. Chen, X. Hong, C. J. Harris, and P. M. Sharkey, "Sparse modeling using orthogonal forward regression with press statistic and regularization," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 34, no. 2, pp. 898–911, Apr. 2004.

[20] K. Z. Mao, "RBF neural network center selection based on fisher ratio class separability measure," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1211–1217, Sep. 2002.

[21] S. Chen, X. X. Wang, X. Hong, and C. J. Harris, "Kernel classifier construction using orthogonal forward selection and boosting with Fisher ratio class separability," *IEEE Trans. Neural Netw.*, 2006, to be published.

[22] K. Z. Mao and G. B. Huang, "Neuron selection for RBF neural network classifier based on data structure preserving criterion," *IEEE Trans. Neural Netw.*, vol. 16, no. 6, pp. 1531–1540, Nov. 2005.

[23] X. Hong, S. Chen, and C. J. Harris, "Fast kernel classifier construction algorithm using orthogonal forward selection to minimize leave-one-out misclassification rate," in *Proc. 2nd Int. Conf. Intell. Comp. Pt. I*, D. S. Huang, K. Li, and G. W. Irwin, Eds., Kuming, China, 2006, pp. 106–114.

[24] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, 1998, pp. 443–453.

[25] J. Leskovec and J. Shawe-Taylor, "Linear programming boost for uneven datasets," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Washington, DC, 2003, pp. 456–463.

[26] G. Karakoulas and J. Shawe-Taylor, "Optimizing classifiers for imbalanced training sets," in *Proc. Neural Inf. Process. Workshop (NIPS'98)*, Denver, CO, 1998, pp. 253–259.

[27] P. D. Turney, "Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm," *J. Artif. Intell. Res.*, no. 2, pp. 369–409, 1995.

[28] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machine," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI99)*, Stockholm, Sweden, 1999, vol. Workshop ML3, pp. 55–60.

[29] G. Wu and E. Y. Chang, "Aligning boundary in kernel space for learning imbalanced dataset," in *Proc. 4th IEEE Int. Conf. Data Mining (ICDM 2004)*, Brighton, U.K., 2004, pp. 265–272.

[30] G. M. Fung and O. L. Mangasarian, "Multicategory proximal support vector machine classifier," *Mach. Learn.*, vol. 59, no. 1–2, pp. 77–97, 2005.

[31] N. Japkowicz, "Supervised versus unsupervised binary-learning by feedforward neural networks," *Mach. Learn.*, vol. 42, no. 1–2, pp. 97–122, 2001.

[32] S. Tan, "Neighbor-weighted k-nearest neighbor for unbalanced text corpus," *Expert Syst. Appl.*, vol. 28, pp. 667–671, 2005.

[33] M. Kubat, R. Holte, and S. Matwin, "Learning when negative examples abound," in *Proc. Eur. Conf. Mach. Learn. (ECML 97)*, Prague, Czech Republic, 1997, pp. 1467–153.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority oversampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[35] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data" Dept. Statistics, Univ. California Berkeley, Tech. Rep. 666, 2004.

[36] J. P. Egan, *Signal Detection Theory and ROC Analysis*. New York: Academic, 1975.

[37] J. A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Mahwah, NJ: Lawrence Erlbaum Associates, 1996.

[38] R. H. Myers, *Classical and Modern Regression With Applications*, 2nd ed. Boston, MA: PWS-Kent, 1990.

[39] D. J. Mackay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.

[40] D. W. Marquardt, "Generalised inverse, ridge regression, biased linear estimation and nonlinear estimation," *Technometrics*, vol. 12, no. 3, pp. 591–612, 1970.

[41] C. van Rijsbergen, *Information Retrieval*. London, U.K.: Butterworths, 1979.

[42] S. Hettich, C. L. Blake, and C. J. Merz, Repository of Machine Learning Databases Univ. California Irvine, 1998 [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[43] K. K. Lee, C. J. Harris, S. R. Gunn, and P. A. S. Reed, "Classification of imbalanced data with transparent kernel," in *Proc. INNS/IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Washington, DC, 2001, pp. 2410–2415.

[44] UCL Machine Learning Group, Elena Database 2004 [Online]. Available: http://www.dice.ucl.ac.be/mlg/?page=elena

Xia Hong (SM'02) received the B.Sc. and M.Sc. degrees from the National University of Defense Technology, Hunan, P.R. China, in 1984 and 1987, respectively, and the Ph.D. degree from the University of Sheffield, Sheffield, U.K., in 1998, all in automatic control.

She worked as a Research Assistant at Beijing Institute of Systems Engineering, Beijing, China, from 1987 to 1993. She worked as a Research Fellow in the Department of Electronics and Computer Science, University of Southampton, Southampton, U.K., from 1997 to 2001. She is currently a Lecturer at the School of Systems Engineering, University of Reading, Reading, U.K. She is actively engaged in research of nonlinear systems identification, data modeling, estimation and

intelligent control, neural networks, pattern recognition, learning theory, and their applications. She has published over 60 research papers, and coauthored a research book.

Dr. Hong was awarded a Donald Julius Groen Prize by Institution of Mechanical Engineers (IMechE) in 1999.

Sheng Chen (SM'97) received the B.Eng. degree in control engineering from the East China Petroleum Institute, Shandong, China, in 1982, the Ph.D. degree in control engineering from the City University at London, London, U.K., in 1986, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in 2005.

He joined the University of Southampton, Southampton, U.K., in September 1999. He previously held research and academic appointments at the Universities of Sheffield, Edingburgh, and Portsmouth, U.K. He has published over 200 research papers. His recent research work includes adaptive nonlinear signal processing, modeling and identification of nonlinear systems, neural network research, finite-precision digital controller design, evolutionary computation methods, and optimization.

Chris J. Harris received the B.Sc. degree from the University of Leicester, Leicester, U.K., in 1967, the M.A. degree from Oxford University, Oxford, U.K., in 1976, and the Ph.D. degree from the University of Southampton, Southampton, U.K., in 2002, all in electrical engineering.

He is now with the University of Southampton, Southampton, U.K. He previously held appointments at the Universities of Hull, UMIST, Oxford, and Cranfield, and he was employed by the U.K. Ministry of Defence. He has authored or coauthored 12 books and over 400 research papers. His research interests are in the area of intelligent and adaptive systems theory and its application to intelligent autonomous systems, management infrastructures, intelligent control and estimation of dynamic processes, multisensor data fusion, and systems integration.

Dr. Harris was elected to the Royal Academy of Engineering in 1996. He was awarded the Institution of Electrical Engineers (IEE) Senior Achievement medal in 1998 for his work on autonomous systems, as well as the highest international award in IEE, the IEE Faraday Medal in 2001 for his work in intelligent control and neurofuzzy system. He was the Associate Editor of numerous international journals including *Automatica*, *Engineering Applications of Artificial Intelligence*, *International Journal of General Systems Engineering*, *International Journal of System Science*, and the *International Journal on Mathematical Control and Information Theory*.