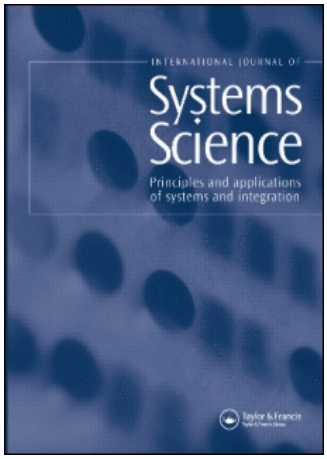


International Journal of Systems Science

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713697751

A fast linear-in-the-parameters classifier construction algorithm using orthogonal forward selection to minimize leave-one-out misclassification rate

X. Hong [a]; S. Chen [b]; C. J. Harris [b]
[a] School of Systems Engineering, University of Reading, Reading, RG6 6AY, UK
[b] School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

Online Publication Date: 01 February 2008
To cite this Article: Hong, X., Chen, S. and Harris, C. J. (2008) 'A fast linear-in-the-parameters classifier construction algorithm using orthogonal forward selection to minimize leave-one-out misclassification rate', International Journal of Systems Science, 39:2, 119 - 125
To link to this article: DOI: 10.1080/00207720701727822
URL: http://dx.doi.org/10.1080/00207720701727822

International Journal of Systems Science
Vol. 39, No. 2, February 2008, 119–125

A fast linear-in-the-parameters classifier construction algorithm using orthogonal forward selection to minimize leave-one-out misclassification rate

X. HONG*†, S. CHEN‡ and C. J. HARRIS‡

†School of Systems Engineering, University of Reading, Reading, RG6 6AY, UK
‡School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

(*Received 4 January 2006; in final form 4 October 2007*)

We propose a simple and computationally efficient construction algorithm for two class linear-in-the-parameters classifiers. In order to optimize model generalization, a forward orthogonal selection (OFS) procedure is used for minimizing the leave-one-out (LOO) misclassification rate directly. An analytic formula and a set of forward recursive updating formula of the LOO misclassification rate are developed and applied in the proposed algorithm. Numerical examples are used to demonstrate that the proposed algorithm is an excellent alternative approach to construct sparse two class classifiers in terms of performance and computational efficiency.

Keywords: Classification; Cross validation; Forward Regression; Regularization; System identification.

1. Introduction

In most supervised learning algorithms, some system input/output mappings are constructed as parametric models, e.g., neural networks, kernel regression/classification models, based on observational data, i.e., pairs of system input/output examples. The two class classification problems can be configured into a regression framework that solves a separating hyperplane for two classes, with the known class labels used as the system output examples for model training. Models are identified according to some objective criteria, e.g., the minimization of model generalization errors. Note that information-based criteria of model generalization, such as the AIC (Akaike 1974), often include a penalty term to avoid an oversized model, which may tend to overfit to the training data set. Parsimonious models are also preferable in engineering applications, since a model's computational complexity scales with its model complexity. Moreover, a parsimonious model is easier to interpret from the viewpoint of knowledge extraction. Consequently, a practical nonlinear modeling principle is to find the smallest model that generalizes well. Modeling techniques on model construction/selection have be widely studied, e.g., support vector machine (SVM), relevance vector machine (RVM), and orthogonal forward regression (OFR) (Vapnik 1995, Hong and Harris 2001, Tipping 2001, Scholkopf and Smola 2002). The orthogonal least square algorithm (Chen *et al*. 1989) was developed as a practical linear-in-the-parameters models construction algorithm. A large class of nonlinear representations, e.g., radial basis functions (RBF) networks and SVM can be classified as the linear-in-the-parameters models. An orthogonal forward selection (OFS) procedure can be applied to construct parsimonious two class classifiers by incrementally maximizing the Fisher ratio of class separability measure (Mao 2002, Chen *et al*. 2006b). Alternatively, the SVM is based on the structural risk minimization (SRM) principle and approximately minimizes an upper bound on the generalization error (Vapnik 1995) via minimizing of the norm of weights in the feature space

*Corresponding author. Email: x.hong@reading.ac.uk

International Journal of Systems Science
ISSN 0020–7721 print/ISSN 1464–5319 online © 2008 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00207720701727822

(Vapnik 1998). The SVM is characterized by a kernel function, lending its solution as that of the convex quadratic programming, such that the resultant model corresponds to a sparse model with a subset of the training data set used as support vectors.

In regression, a fundamental concept in the evaluation of model generalization capability is that of cross validation (Stone 1974). The leave-one-out (LOO) cross validation is often used to estimate generalization error for choosing among different network architectures (Stone 1974). LOO errors can be derived using algebraic operation rather than actually splitting the training data set for linear-in-the-parameters models. The calculation of LOO errors is computational expensive. The generalized cross validation (GCV) (Golub *et al*. 1979) has been introduced as a variant of leave-one-out (LOO) cross validation to improve computational efficiency. 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 (Hong *et al*. 2003, Chen *et al*. 2004).

In this article, the construction of parsimonious linear-in-the-parameters models using LOO cross validation for two class classifiers is considered. An analytic formula for LOO misclassification rate is initially derived, based on the regularized orthogonal least squares (ROLS) parameter estimates (Chen *et al*. 2004). The proposed algorithm shares some common derivations as in Hong *et al*. (2003) and Chen *et al*. (2004), as both use the same orthogonalization procedure. Note that in classification, the modeling objective is often to minimize the number of misclassified samples rather than the MSE and LOO errors. The proposed method extends forward regression procedure in Hong *et al*. (2003) and Chen *et al*. (2004) to classification problem by using the leave-one-out misclassification rate, the true generalization capability of a classifier, for model selection, rather than the direct extension of Hong *et al*. (2003) and Chen *et al*. (2004) of using LOO errors for model selection. Furthermore, it is shown that the orthogonalization procedure brings the advantage of calculating the LOO misclassification rate *via* a set of new forward recursive updating formula at minimal computational expense. Then, a fast two class linear-in-the-parameters classifier construction algorithm is presented using orthogonal forward selection by directly minimizing LOO misclassification rate to optimize the model generalization. Numerical examples are used to demonstrate the efficacy of the proposed approach compared with other current kernel-based classifiers.

2. Problem formulation

Consider a training data set $D_N = \{\mathbf{x}(i), y(i)\}_{i=1}^N$, in which $y(i) \in \{1, -1\}$ denotes the class type for each data sample $\mathrm{x}(i) \in \Re^n$. Let a two class classifier $f(\mathbf{x}) : \Re^n \to \{1, -1\}$ be formed using the data set. The linear-in-the-parameters classifier is given as

$$\hat{y}(i) = \mathrm{sgn}(f(i)) \quad \text{with} \quad f(i) = \sum_{j=1}^{L} \boldsymbol{\theta}_j p_j(\mathbf{x}(i)) \tag{1}$$

where $p_j(\bullet)$ denotes the classifier kernels with a known nonlinear basis function, such as RBF, or B-spline fuzzy membership functions. Model (1) is very general, but the Gaussian kernel functions $p_j(\mathrm{x}) = \exp\{-\|\mathrm{x} - \mathbf{c}_j\|^2/2\sigma^2\}$ are employed in this study, where $\mathbf{c}_j \in \Re^n$ are kernel centers, θ_j are model parameters, L is the number of regressors (kernels), and $\hat{y}(i)$ is the model predicted class label for $\mathrm{x}(i)$.

Taking the complete training data set D_N, denoting $\xi(i) = y(i) - f(i)$ as the modeling residual sequence with zero mean, equation (1) can be written in vector form as

$$\mathbf{y} = \mathbf{P}\boldsymbol{\theta} + \Xi \tag{2}$$

where $\Xi = [\xi(1), \ldots, \xi(N)]^T$ is the residual vector, and $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_L] \in \Re^{N \times L}$ is the regression matrix, with column vectors $\mathbf{p}_j = [p_j(\mathbf{x}(1)), \ldots, p_j(\mathbf{x}(\mathrm{N}))]^T$. Denote the row vectors in $\mathbf{P}$ as $\mathbf{p}(i) = [p_1(i), \ldots, p_L(i)]^T$, $i = 1, \ldots, N$. Geometrically, a set of parameter vectors $\theta = [\theta_1, \ldots, \theta_L]$ defines a hyperplane by

$$\sum_{j=1}^{L} \boldsymbol{\theta}_j p_j(\mathbf{x}) = 0 \tag{3}$$

dividing the data into two classes.

An orthogonal decomposition of $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{W}\mathbf{A} \tag{4}$$

where $\mathbf{A} = \{a_{ij}\}$ is an $L \times L$ unit upper triangular matrix and $\mathbf{W}$ is an $N \times L$ matrix with orthogonal columns that satisfy

$$\mathbf{W}^T\mathbf{W} = \mathrm{diag}\{\kappa_1, \ldots, \kappa_L\} \tag{5}$$

with

$$\kappa_j = \mathbf{w}_j^T\mathbf{w}_j, \quad j = 1, \ldots, L \tag{6}$$

For $\mathbf{W}$, the column vectors are denoted as $\mathbf{w}_j = [w_j(1)), \ldots, w_j(N)]^T, j = 1, \ldots, L$, and the row vectors as $\mathbf{w}(i) = [w_1(i), \ldots, w_L(i)]^T$, $i = 1, \ldots, N$.

Equation (2) can now be expressed as

$$\mathbf{y} = (\mathbf{P}\mathbf{A}^{-1})(\mathbf{A}\theta) + \Xi = \mathbf{W}\boldsymbol{\Gamma} + \Xi \tag{7}$$

in which $\Gamma = [\gamma_1, \ldots, \gamma_L]^T$ is an auxiliary vector, for which the regularized orthogonal least squares (ROLS) parameter estimates (Chen *et al.* 2004) is

$$\gamma_j = \frac{\mathbf{w}_j^T \mathbf{y}}{\kappa_j + \lambda_j}, \quad j = 1, \ldots, L \tag{8}$$

in which λ_j are positive regularization parameters. If all λ_j is set as zero, the parameter estimator is simply the least squares estimator. The original model coefficient vector $\theta = [\theta_1, \ldots, \theta_L]^T$ can then be calculated from $\mathbf{A}\theta = \boldsymbol{\Gamma}$ through back-substitution.

The regularization parameters, λ_j, can be optimized iteratively using an evidence procedure (Mackay 1992, Tipping 2001, Chen *et al.* 2004). The following updating formulae quoted from Chen *et al.* (2004) are used to determine the regularization parameters.

$$\begin{aligned} \lambda_j^{\text{new}} &= \frac{\rho_j^{\text{new}}}{N - \rho^{\text{old}}} \frac{\Xi^T \Xi}{N - \gamma_j^2}, \quad j = 1, \ldots, L \\ \text{where } \rho_j &= \frac{\mathbf{w}_j^T \mathbf{w}_j}{\lambda_j + \mathbf{w}_j^T \mathbf{w}_j} \quad \text{and } \rho = \sum_{j=1}^{L} \rho_j. \end{aligned} \tag{9}$$

3. Leave-one-out misclassification rate

The misclassification rate for a given two class classifier based on (1) can be evaluated based on the misclassified data examples as

$$J = \tfrac{1}{N} \sum_{i=1}^{N} \text{Id}[y(i) f(i)] \tag{10}$$

where Id(•) denotes the misclassification indication function for a data example, and is defined as

$$\text{Id}(v) = \begin{cases} 1 & \text{if } v < 0 \\ 0 & \text{if } v \geq 0 \end{cases}$$

Cross-validation criteria are metrics that measure a model's generalization capability (Stone 1974). One commonly used version of cross-validation is the so called leave-one-out cross-validation. The idea is that, for any predictor, each data point in the estimation data set D_N is sequentially set aside in turn, a model is estimated using the remaining $(N-1)$ data, and the prediction error 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 $f^{(-i)}(i)$. The associated predicted class label is calculated by

$$\hat{y}^{(-i)}(i) = \text{sgn}(f^{(-i)}(i)) \tag{11}$$

It is desirable to derive a classifier with good generalization capability, i.e., to derive a classifier with a minimal misclassification error rate over a new data set that is not used in model estimation. The leave-one-out (LOO) cross validation is often used to estimate generalization error for choosing among different network architectures (Stone 1974). The LOO misclassification rate is computed by

$$J^{(-)} = \frac{1}{N} \sum_{i=1}^{N} \text{Id}[y(i) f^{(-i)}(i)] = \tfrac{1}{N} \sum_{i=1}^{N} \text{Id}[g(i)] \tag{12}$$

in which $g(i)$ denotes $y(i)f^{(-i)}(i)$. If $g(i) < 0$, this means the ith data sample is misclassified, such that the class label produced by the model $f^{(-i)}$ is different from the actual class label $y(i)$.

Instead of directly calculating (11) for predicted class labels, which requires extensive computational effort, it is shown in the following that the LOO misclassification rate can be evaluated without actually sequentially splitting the estimation data set.

4. A forward regression kernel classifier identification algorithm minimizing leave-one-out misclassification rate (LOO + OFS)

The leave-one-out model residual is given by

$$\xi^{(-i)}(i) = y(i) - f^{(-i)}(i) \tag{13}$$

It has been shown that 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 (Myers 1990). For models evaluated using regularized orthogonal least square parameter estimates, it can be shown that the LOO model residuals (Chen *et al.* 2004) are given by

$$\begin{aligned} \xi^{(-i)}(i) &= \frac{\xi(i)}{1 - \mathbf{w}(i)^T [\mathbf{W}^T \mathbf{W} + \Lambda]^{-1} \mathbf{w}(i)} \\ &= \frac{y(i) - f(i)}{1 - \sum_{j=1}^{L} \dfrac{\mathbf{w}_j^2(i)}{\kappa_j + \lambda_j}} \end{aligned} \tag{14}$$

where $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_L\}$. Hence

$$y(i) - f^{(-i)}(i) = \frac{y(i) - f(i)}{1 - \sum_{j=1}^{L} \frac{w_j^2(i)}{\kappa_j + \lambda_j}} \tag{15}$$

Multiplying both sides of (15) with $y(i)$, and applying $y^2(i) = 1, \forall i$, to yield

$$1 - y(i)f^{(-i)}(i) = \frac{1 - f(i)y(i)}{1 - \sum_{j=1}^{L} \frac{w_j^2(i)}{\kappa_j + \lambda_j}} \tag{16}$$

so that

$$y(i)f^{(-i)}(i) = \frac{\sum_{j=1}^{L} \gamma_j w_j(i) y(i) - \sum_{j=1}^{L} \frac{w_j^2(i)}{\kappa_j + \lambda_j}}{1 - \sum_{j=1}^{L} \frac{w_j^2(i)}{\kappa_j + \lambda_j}} \tag{17}$$

In the following, it is shown that computational expense associated with classifier regressors 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 ($k \ll L$), that is, a subset selected from the full model set consisting of L initial regressors (given by (2)) to approximate the system. By replacing L with a variable model size k, and $y(i)f^{(-i)}(i)$ with $g_k(i)$, (17) is represented by

$$\begin{aligned} g_k(i) &= \frac{\sum_{j=1}^{k} \gamma_j w_j(i) y(i) - \sum_{j=1}^{k} \frac{w_j^2(i)}{\kappa_j + \lambda_j}}{1 - \sum_{j=1}^{k} \frac{w_j^2(i)}{\kappa_j + \lambda_j}} \\ &= \frac{\alpha_k(i)}{\beta_k(i)} \end{aligned} \tag{18}$$

where
$\beta_k(i) = 1 - \sum_{j=1}^{k} w_j^2(i)/\kappa_j + \lambda_j$, $\alpha_k(i) = \sum_{j=1}^{k} \gamma_j w_j(i)$
$y(i) - \sum_{j=1}^{k} w_j^2(i)/\kappa_j + \lambda_j$.

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

$$\begin{aligned} \alpha_k(i) &= \alpha_{k-1}(i) + \gamma_k w_k(i) y(i) - \frac{w_k^2(i)}{\kappa_k + \lambda_j} \\ \beta_k(i) &= \beta_{k-1}(i) - \frac{w_k^2(i)}{\kappa_k + \lambda_j} \end{aligned} \tag{19}$$

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

$$J_k^{(-1)} = \frac{1}{N} \sum_{i=1}^{N} \text{Id}[g_k(i)] \tag{20}$$

where $g_k(i) = \alpha_k(i)/\beta_k(i)$. 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. The Gram–Schmidt procedure is used to construct the orthogonal basis $\mathbf{w}_k$ in a forward regression manner (Hong *et al*. 2003, Chen *et al*. 2004). At each regression step, the regressor with the minimal LOO misclassification rate $J_k^{(-)}$ is selected.

4.1 *LOO misclassification rate minimization-based forward Gram–Schmidt subset selection algorithm (LOO + OFS)*

1. Initialize $\alpha_0(i) = 0$, $\beta_0(i) = 1$, for $i = 1, \ldots, N$. Set regularization parameters λ_j as a very small positive value λ.
2. At the kth step where $k \geq 1$, for $1 \leq l \leq L$, $l \neq l_1, \ldots l \neq l_{k-1}$, compute

$$a_{jk}^{(l)} = \begin{cases} 1 & \text{if } j = k \\ \frac{\mathbf{w}_j^T \mathbf{p}_l}{\mathbf{w}_j^T \mathbf{w}_j}, & 1 \leq j < k \end{cases}$$

$$\text{w}_k^{(l)} = \begin{cases} \mathbf{p}_l & \text{if } k = 1 \\ \mathbf{p}_l - \sum_{j=1}^{k-1} a_{jk}^{(l)} \mathbf{w}_j, & k \geq 2 \end{cases}$$

$$\kappa_k^{(l)} = (\mathbf{w}_k^{(l)})^T \mathbf{w}_k^{(l)},$$

$$\gamma_k^{(l)} = \frac{(\mathbf{w}_k^{(l)})^T \mathbf{y}}{\kappa_k^{(l)} + \lambda},$$

$$\alpha_k^{(l)}(i) = \alpha_{k-1}(i) + \gamma_k^{(l)} w_k^{(l)}(i) y(i) - \frac{[w_k^{(l)}(i)]^2}{\kappa_k^{(l)} + \lambda}, \quad (i = 1, \ldots, N)$$

$$\beta_k^{(l)}(i) = \beta_{k-1}(i) - \frac{[w_k^{(l)}(i)]^2}{\kappa_k^{(l)} + \lambda}, \quad (i = 1, \ldots, N)$$

$$g_k^{(l)}(i) = \frac{\alpha_k^{(l)}(i)}{\beta_k^{(l)}(i)}, \quad (i = 1, \ldots, N)$$

$$J_k^{(-,l)} = \frac{1}{N} \sum_{i=1}^{N} \text{Id}(g_k^{(l)}(i)) \tag{21}$$

Find

$$\begin{aligned} l_k = \arg[\min\{&J_k^{(-,l)}, \\ &1 \leq l \leq L, \quad l \neq l_1, \ldots, l \neq l_{k-1}\}] \end{aligned} \tag{22}$$

and select

$$a_{jk} = a_{jk}^{(l_k)}, \quad J_k^{(-)} = J_k^{(-,l_k)} \tag{23}$$

and update

$$\alpha_k(i) = \alpha_k^{(l_k)}(i), \quad \beta_k(i) = \beta_k^{(l_k)}(i), \quad (i = 1, \ldots, N) \quad (24)$$

$$\mathbf{w}_k = \mathbf{w}_k^{(l_k)} = \begin{cases} \mathbf{p}_{l_k} & \text{if } k = 1 \\ \mathbf{p}_{l_k} - \sum_{j=1}^{k-1} a_{jk}\mathbf{w}_j & k \geq 2 \end{cases}$$

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

The above procedure derives a model with $n_\theta \ll L$ regressors. Finally, with a predetermined number of iterations, the procedure as given in (9) (with L replaced by n_θ) is applied to derive the optimized regularization parameters that are used in the final model.

4.2 *Remarks*

1. The computational complexity in above LOO + OFS algorithm is in the order of $O(NL)$. The actual computation cost varies with the final model size, and the smaller the derived model size n_θ, the smaller the computation expense. When N is very large, e.g., over several thousands, a reduced subset of data points can be used so that $L \ll N$ to control the computational complexity. Note that the proposed procedure for regularization parameters optimization is operated based on $n_\theta \ll L$ selected regressors, hence, the additional computation cost involved in regularization parameters optimization is very small at the level $O(Nn_\theta)$.
2. Note that it is generally difficult to perform parameter estimation, so as to optimize the classification performance directly. This is due to the factors, such as unknown probability function of the data distribution or possibly non-differentiable objective functions. In the proposed algorithm and other algorithms (Mao 2002, Chen *et al.* 2006b, Hong *et al.* 2007), the two class classification problem is configured as a regression problem, and the least squares-type parameter estimators have been used for parameter estimation. This brings the advantage that the classifier can be easily obtained. The disadvantage of the regression approach is that models are not directly derived by optimizing the classification performance. However, in the proposed algorithm we initially use least squares-type parameter estimator for generating candidate models, followed by the direct evaluation of these models in terms of classification performance. The model selection step can therefore guarantee that the best model in terms of classification performance is found amongst the candidate models. This means that the aforementioned disadvantage could be alleviated effectively.
3. A closely related method is the kernel matching pursuit (KMP) (Vincent and Bengio 2002). One of the contributions in (Vincent and Bengio 2002) is to adopt variations of loss functions, for either the model term selection or parameter estimation. A key difference and advantage of the proposed algorithm in comparison with KMP is that there is no need to use a separate validation set to terminate the algorithm. In the proposed algorithm, the LOO classification error represents model generation capability for classification and is calculated analytically.
4. Clearly the width σ has a high impact in the performance of the obtained classifier. However, 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 empirically with cross validation, and this approach is used in the illustrative examples below. Obviously, there is an added computational complexity, but this would be equally applicable to any alternative approaches with Gaussian kernels. Alternatively each kernel may have individually tunable width and be optimized (Chen *et al.* 2006a).

5. Illustrative examples

Numerical experiments were performed to demonstrate the modeling results of the proposed LOO + OFS algorithm in comparison to that of several existing classifications algorithms as published in Rätsch *et al.* (2001). Three data sets were experimented: breast cancer, diabetes, and heart, which are available from Rätsch (n.d.). Note that we did not experiment on all the data set provided in Rätsch (n.d.), as our aim is simply to demonstrate the proposed approach can be used as a viable alternative. For the details of alternative methods used in comparison, the readers are referred to Rätsch *et al.* (2001).

The results of first six methods for all examples are quoted from Rätsch *et al.* (2001) and Rätsch (n.d.). Each data set contains 100 realizations of training and test data set respectively. Models are constructed over 100 training data sets and generalization performance is evaluated using the average misclassification rate of the

Table 1. Average misclassification rate in % over 100 realizations of the breast cancer test data set and model size.

	Misclassification rate	Model size
RBF	27.6 ± 4.7	5
Adaboost with RBF	30.4 ± 4.7	5
AdaBoost$_R$*eg*	26.5 ± 4.5	5
LP$_R$*eg*-AdaBoost	26.8 ± 6.1	5
QP$_R$*eg*-AdaBoost	25.9 ± 4.6	5
SVM with RBF kernel	26.0 ± 4.7	Not available
Proposed LOO + OFS	25.74 ± 5	6 ± 2

Table 2. Average misclassification rate in % over 100 realizations of the diabetes test data set and model size.

	Misclassification rate	Model size
RBF	24.3 ± 1.9	15
Adaboost with RBF	26.5 ± 2.3	15
AdaBoost$_R$*eg*	23.8 ± 1.8	15
LP$_R$*eg*-AdaBoost	24.1 ± 1.9	15
QP$_R$*eg*-AdaBoost	25.4 ± 2.2	15
SVM with RBF kernel	23.5 ± 1.7	Not available
Proposed LOO + OFS	23.0 ± 1.7	6 ± 1

Table 3. Average misclassification rate in % over 100 realizations of the heart test data set and model size.

	Misclassification rate	Model size
RBF	17.6 ± 3.3	4
Adaboost with RBF	20.3 ± 3.4	4
AdaBoost$_R$*eg*	16.5 ± 3.5	4
LP$_R$*eg*-AdaBoost	17.5 ± 3.5	4
QP$_R$*eg*-AdaBoost	17.2 ± 3.4	4
SVM with RBF kernel	16.0 ± 3.3	Not available
Proposed LOO + OFS	15.8 ± 3.7	10 ± 3

corresponding models over the 100 test data sets. The Gaussian kernel functions $p_j(\mathbf{x}) = \exp\{-\|\mathbf{x} - \mathbf{c_j}\|^2/2\sigma^2\}$ have been employed in the experiments. Values of σ were predetermined to derive individual models for each realization. For each realization of all three data sets, the full training data sets were used as the RBF centers to form the candidate regressors set. The performance is summarized in tables 1–3 respectively. The results have shown that the proposed approach can construct parsimonious classifiers with competitive classification accuracy for these data sets.

6. Conclusions

Based upon the idea of using the orthogonal forward selection (OFS) procedure to optimize model generalization, a simple and computationally efficient algorithm has been introduced to construct sparse two class linear-in-the-parameters classifiers by directly minimizing the leave-one-out (LOO) misclassification rate. The contribution is to develop a set of forward recursive updating formula of the LOO misclassification rate in the proposed algorithm. Experimental results on three benchmark examples are used to demonstrate the efficacy of the proposed approach.

Acknowledgment

The authors gratefully acknowledge that part of this work was supported by EPSRC in the UK. We also thank the reviewers for their valuable comments.

References

H. Akaike, "A new look at the statistical model identification", *IEEE Trans. Automat. Contr.*, AC-19, pp. 716–723, 1974.

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

S. Chen, X. Hong and C.J. Harris, "Construction of RBF classifiers with tunable units using orthogonal forward selection based on leave-one-out misclassification rate", in: *Proceedings of International Joint Conference on Neural Networks*, Vancouver, BC, Canada, 2006a, pp. 6390–6394.

S. Chen, X. Hong, C.J. Harris and P.M. Sharkey, "Sparse modelling using orthogonal forward regression with PRESS statistic and regularization", *IEEE Trans. Syst. Man Cybernetics, Part B: Cybernetics*, 34, pp. 898–911, 2004.

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 Networks*, 17, pp. 1652–1656, 2006b.

G.H. Golub, M. Heath and G. Wahba, "Generalized cross-validation as a method for choosing good ridge parameter", *Technometrics*, 21, pp. 215–223, 1979.

X. Hong and C.J. Harris, "Nonlinear model structure design and construction using orthogonal least squares and D-optimality design", *IEEE Trans. Neural Networks*, 13, pp. 1245–1250, 2001.

X. Hong, P.M. Sharkey and K. Warwick, "Automatic nonlinear predictive model construction using forward regression and the PRESS statistic", *IEE Proc.–Contr. Theory Appl.*, 150, pp. 245–254, 2003.

X. Hong, S. Chen and C.J. Harris, "A kernel based two-class classifier for imbalanced data sets", *IEEE Trans. Neural Networks*, 18, pp. 28–41, 2007.

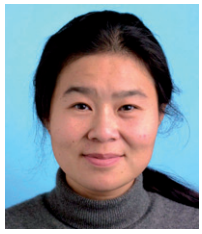
D.J. Mackay, "Bayesian interpolation", *Neural Comput.*, 4, pp. 415–447, 1992.

K.Z. Mao, "RBF neural network center selection based on fisher ratio class separability measure", *IEEE Trans. Neural Networks*, 13, pp. 1211–1217, 2002.

R.H. Myers, *Classical and Modern Regression with Applications*, 2nd ed., Boston: PWS-KENT, 1990.

Rätsch, G. (n.d.). http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm

G. Rätsch, T. Onoda and K.R. Müller, "Soft margins for AdaBoost", *Machine Learning*, 42, pp. 287–320, 2001.
B. Scholkopf and A.J. Smola, *Learning with Kernels: Support Vector Machine, Regularization, Optimization and Beyond*, Cambridge, MA: MIT Press, 2002.
M. Stone, "Cross validatory choice and assessment of statistical predictions", *Appl. Stat.*, 36, pp. 117–147, 1974.
M.E. Tipping, "Sparse Bayesian learning and the relevance vector machine", *J. Machine Learning Res.*, 1, pp. 211–244, 2001.
V. Vapnik, *The Nature of Statictical Learning Theory*, New York: Springer-Verlag, 1995.
V. Vapnik, *Statistical Learning Theory*, New York: J. Wiley, 1998.
P. Vincent and Y. Bengio, "Kernel matching pursuit", *Machine Learning*, 48, pp. 169–191, 2002.



Xia Hong received her university education at National University of Defense Technology, P. R. China (BSc, 1984; MSc, 1987), and University of Sheffield, UK (PhD, 1998), all in automatic control. She worked as a research assistant in 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 at University of Southampton from 1997 to 2001. She is currently a lecturer at the School of Systems Engineering, University of Reading. She is actively engaged in research into nonlinear systems identification, data modeling, estimation and intelligent control, neural networks, pattern recognition, learning theory and their applications. She has published over 80 research papers and coauthored a research book. She was awarded a Donald Julius Groen Prize by IMechE in 1999.

Sheng Chen obtained a BEng degree in control engineering from the East China Petroleum Institute in 1982 and a PhD degree in control engineering from the City University at London in 1986. He joined the University of Southampton in September 1999. He previously held research and academic appointments at the Universities of Sheffield, Edingburgh and Portsmouth. Prof Chen is a Senior Member of IEEE in the USA. His recent research works include adaptive nonlinear signal processing, modeling and identification of nonlinear systems, neural network research, finite-precision digital controller design, evolutionary computation methods and optimization. He has published over 200 research papers.

Chris Harris received university education at Leicester (BSc), Oxford (MA) and Southampton (PhD). He previously held appointments at the Universities of Hull, UMIST, Oxford and Cranfield, as well as being employed by the UK Ministry of Defence. 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, multi-sensor data fusion and systems integration. He has authored or co-authored 12 books and over 400 research papers, and he was the associate editor of numerous international journals including Automatica, Engineering Applications of AI, Int. J. General Systems Engineering, International J. of System Science and the Int. J. on Mathematical Control and Information Theory. He was elected to the Royal Academy of Engineering in 1996, was awarded the IEE Senior Achievement medal in 1998 for his work on autonomous systems, and the highest international award in IEE, the IEE Faraday medal in 2001 for his work in Intelligent Control and Neurofuzzy System.