Construction of RBF Classifiers with Tunable Units
Using Orthogonal Forward Selection Based on
Leave-One-Out Misclassification Rate

S. Chen†, X. Hong‡ and C.J. Harris†

† School of Electronics and Computer Science,
University of Southampton, SO17 1BJ, UK

‡ Department of Cybernetics,
University of Reading, RG6 6AY, UK
Outline

- Existing RBF classifier construction methods and motivations for the present work.

- The proposed RBF classifier construction method.

- Experimental investigation of the proposed method and comparison with some existing techniques.
Overview of Existing Methods

- **Nonlinear optimisation approach:** Optimise all parameters (centre vectors, node variances or covariance matrices, weights)
  - Very “sparse” (small size)
  - All problems associated with nonlinear optimisation

- **Linear optimisation approach:** Fix centres to training input data, and seek a “linear” subset model
  - **Orthogonal least squares forward selection**
    - Sparse, good performance, and efficient construction
    - Need to specify RBF variance (via cross validation)
  - **Kernel modelling methods**
    - Sparse (though not as sparse as OLS), good performance
    - Need to specify RBF variance and other kernel hyperparameters (via costly cross validation)
Motivations

- How good a RBF classifier method:
  - Generalisation performance
  - Sparsity level or classifier’s size
  - Efficiency of classifier construction process

- Combine best of both nonlinear and linear approaches
  - Keep OLS selection procedure to pick RBF units one by one
    - Retain efficiency of OLS construction process
  - But each RBF unit is optimised via nonlinear optimisation
    - Determine centre vector and covariance matrix by directly optimising generalisation capability: leave-one-out misclassification rate
    - This nonlinear optimisation carried out by a simple yet efficient global search method: repeated weighted boosting search
Two-Class Classification

- Given training set \( \{(x_k, y_k)\}_{k=1}^{N} \), where \( y_k \in \{-1, +1\} \) is class label for \( m \)-dimensional pattern vector \( x_k \), construct RBF classifier

\[
\tilde{y}_k = \text{sgn}(\hat{y}_k) \quad \text{with} \quad \hat{y}_k = f^{(M)}_{\text{RBF}}(x_k) = \sum_{i=1}^{M} w_i g_i(x_k),
\]

where \( \tilde{y}_k \) is estimated class label for \( x_k \), \( f^{(M)}_{\text{RBF}}(\bullet) \) denotes RBF classifier with \( M \) units, and \( \text{sgn}(y) = -1 \) if \( y \leq 0 \), \( \text{sgn}(y) = +1 \) if \( y > 0 \)

- We consider general tunable RBF unit of form

\[
g_i(x) = K \left( \sqrt{(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)} \right)
\]

where \( \mu_i \) is centre vector of the \( i \)th RBF unit, whose diagonal covariance matrix is \( \Sigma_i = \text{diag}\{\sigma_{i,1}^2, \cdots, \sigma_{i,m}^2\} \), and \( K(\bullet) \) is basis function
RBF Model

- Regression model of RBF classifier

\[ y_k = \hat{y}_k + e_k = g^T(k)w + e_k \]

where \( w = [w_1 \ w_2 \cdots w_M]^T \) and \( g(k) = [g_1(x_k) \ g_2(x_k) \cdots g_M(x_k)]^T \)

- Define \( y = [y_1 \ y_2 \cdots y_N]^T \), \( e = [e_1 \ e_2 \cdots e_N]^T \), and \( G = [g_1 \ g_2 \cdots g_M] \)

with \( g_k = [g_k(x_1) \ g_k(x_2) \cdots g_k(x_N)]^T \), \( 1 \leq k \leq M \)

- Regression model over training data set:

\[ y = Gw + e \]

Note that \( g_k \) denotes \( k \)th column of \( G \) while \( g^T(k) \) is \( k \)th row of \( G \)

- Let an orthogonal decomposition of regression matrix \( G \) be \( G = PA \).

Then RBF model can alternatively be expressed

\[ y = P\theta + e \]
Misclassification Rate

- Weight vector $\mathbf{\theta} = [\theta_1 \, \theta_2 \, \cdots \, \theta_M]^T$ in orthogonal space $\mathbf{P} = [\mathbf{p}_1 \, \mathbf{p}_2 \, \cdots \, \mathbf{p}_M]$ satisfies triangular system $\mathbf{A} \mathbf{w} = \mathbf{\theta}$, where $\mathbf{A}$ is upper triangular.

- RBF model output is equivalently expressed in orthogonal space as

$$
\hat{y}_k = \mathbf{p}^T(k) \mathbf{\theta}
$$

where $\mathbf{p}^T(k) = [p_1(k) \, p_2(k) \, \cdots \, p_M(k)]$ is $k$th row of $\mathbf{P}$.

- Define signed decision variable

$$
s_k = \text{sgn}(y_k) \hat{y}_k = y_k \hat{y}_k = y_k f_{RBF}^{(M)}(\mathbf{x}_k)
$$

- Then misclassification rate over $\{ (\mathbf{x}_k, y_k) \}_{k=1}^{N}$ is

$$
\mathcal{M}_r = \frac{1}{N} \sum_{k=1}^{N} \mathcal{I}_d(s_k) \quad \text{where} \quad \mathcal{I}_d(y) = \begin{cases} 
1, & y \leq 0 \\
0, & y > 0
\end{cases}
$$
Leave-One-Out Cross Validation

- Denote $k$th modelling error of $n$-unit RBF classifier, identified using the entire $\{(x_k, y_k)\}_{k=1}^N$, as $e_k^{(n)} = y_k - f_{RBF}^{(n)}(x_k) = y_k - \hat{y}_k^{(n)}$

- Let $f_{RBF}^{(n,-k)}(\bullet)$ be $n$-unit RBF classifier identified using $\{(x_k, y_k)\}_{k=1}^N$ but with its $k$th data point being removed

- Test output of this $n$-unit RBF classifier at $k$th data point not used in training is computed by $\hat{y}_k^{(n,-k)} = f_{RBF}^{(n,-k)}(x_k)$

- Leave-one-out signed decision variable is defined by $s_k^{(n,-k)} = y_k \hat{y}_k^{(n,-k)}$

- Leave-one-out misclassification rate is computed by
  $$J_n = \frac{1}{N} \sum_{k=1}^N \mathcal{I}_d \left( s_k^{(n,-k)} \right)$$
Efficient Computation

- LOO misclassification rate $J_n$ is a measure of classifier’s generalisation capability.

- $J_n$ can be computed efficiently, as owing to orthogonal decomposition we have:

$$s_{k}^{(n,-k)} = \frac{\phi_{k}^{(n)}}{\eta_{k}^{(n)}}$$

with

$$\phi_{k}^{(n)} = \phi_{k}^{(n-1)} + y_{k}\theta_{n}p_{n}(k) - \frac{p_{n}^{2}(k)}{p_{n}^{T}p_{n} + \lambda}$$

and

$$\eta_{k}^{(n)} = \eta_{k}^{(n-1)} - \frac{p_{n}^{2}(k)}{p_{n}^{T}p_{n} + \lambda}$$

- Proposed algorithm constructs RBF units one by one by minimising $J_n$.
Positioning and Shaping RBF Unit

- At \( n \)th construction stage, determine \( n \)th RBF unit by minimising \( J_n \)

\[
\min_{\mu_n, \Sigma_n} J_n (\mu_n, \Sigma_n)
\]

- Construction procedure is automatically terminated when

\[
J_M \leq J_{M+1}
\]

yielding \( M \)-term RBF classifier

- Note that LOO criterion \( J_n \) is at least locally convex, and there exists an “optimal” \( M \) such that: for \( n \leq M \) \( J_n \) decreases as model size \( n \) increases while the above condition holds

- Nonlinear optimisation is performed using a simple yet efficient global search algorithm called repeated weighted boosting search
Synthetic Two-Class Problem


<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model Size</th>
<th>Test Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>38</td>
<td>10.6%</td>
</tr>
<tr>
<td>RVM</td>
<td>4</td>
<td>9.3%</td>
</tr>
<tr>
<td>Proposed</td>
<td>3</td>
<td>8.0%</td>
</tr>
</tbody>
</table>

## Breast Cancer Data Set

Average classification test error rate in % over 100 realizations

<table>
<thead>
<tr>
<th>method</th>
<th>test error rate</th>
<th>model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF-Network</td>
<td>27.64 ± 4.71</td>
<td>5</td>
</tr>
<tr>
<td>AdaBoost with RBF-Network</td>
<td>30.36 ± 4.73</td>
<td>5</td>
</tr>
<tr>
<td>LP-Reg-AdaBoost (&quot;&quot;-)</td>
<td>26.79 ± 6.08</td>
<td>5</td>
</tr>
<tr>
<td>QP-Reg-AdaBoost (&quot;&quot;-)</td>
<td>25.91 ± 4.61</td>
<td>5</td>
</tr>
<tr>
<td>AdaBoost-Reg (&quot;&quot;-)</td>
<td>26.51 ± 4.47</td>
<td>5</td>
</tr>
<tr>
<td>SVM with RBF-Kernel</td>
<td>26.04 ± 4.74</td>
<td>not available</td>
</tr>
<tr>
<td>Kernel Fisher Discriminant</td>
<td>24.77 ± 4.63</td>
<td>not available</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td><strong>24.49 ± 3.28</strong></td>
<td><strong>3.1 ± 1.2</strong></td>
</tr>
</tbody>
</table>

Data and first 7 results from:

http://ida.first.fhg.de/projects/bench/benchmarks.htm
## Diabetis Data Set

Average classification test error rate in % over 100 realizations

<table>
<thead>
<tr>
<th>method</th>
<th>test error rate</th>
<th>model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF-Network</td>
<td>24.29 ± 1.88</td>
<td>15</td>
</tr>
<tr>
<td>AdaBoost with RBF-Network</td>
<td>26.47 ± 2.29</td>
<td>15</td>
</tr>
<tr>
<td>LP-Reg-AdaBoost (-”-”)</td>
<td>24.11 ± 1.90</td>
<td>15</td>
</tr>
<tr>
<td>QP-Reg-AdaBoost (-”-”)</td>
<td>25.39 ± 2.20</td>
<td>15</td>
</tr>
<tr>
<td>AdaBoost-Reg (-”-”)</td>
<td>23.79 ± 1.80</td>
<td>15</td>
</tr>
<tr>
<td>SVM with RBF-Kernel</td>
<td>23.53 ± 1.73</td>
<td>not available</td>
</tr>
<tr>
<td>Kernel Fisher Discriminant</td>
<td>23.21 ± 1.63</td>
<td>not available</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>22.16 ± 1.47</td>
<td>4.0 ± 1.6</td>
</tr>
</tbody>
</table>

Data and first 7 results from:

http://ida.first.fhg.de/projects/bench/benchmarks.htm
## Thyroid Data Set

Average classification test error rate in % over 100 realizations

<table>
<thead>
<tr>
<th>method</th>
<th>test error rate</th>
<th>model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF-Network</td>
<td>4.52 ± 2.12</td>
<td>8</td>
</tr>
<tr>
<td>AdaBoost with RBF-Network</td>
<td>4.40 ± 2.18</td>
<td>8</td>
</tr>
<tr>
<td>LP-Reg-AdaBoost (-”-”)</td>
<td>4.59 ± 2.22</td>
<td>8</td>
</tr>
<tr>
<td>QP-Reg-AdaBoost (-”-”)</td>
<td>4.35 ± 2.18</td>
<td>8</td>
</tr>
<tr>
<td>AdaBoost-Reg (-”-”)</td>
<td>4.55 ± 2.19</td>
<td>8</td>
</tr>
<tr>
<td>SVM with RBF-Kernel</td>
<td>4.80 ± 2.19</td>
<td>not available</td>
</tr>
<tr>
<td>Kernel Fisher Discriminant</td>
<td>4.20 ± 2.07</td>
<td>not available</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>3.21 ± 1.35</td>
<td>3.9 ± 0.8</td>
</tr>
</tbody>
</table>

Data and first 7 results from:

http://ida.first.fhg.de/projects/bench/benchmarks.htm
Conclusions

A novel construction algorithm has been proposed for RBF classifiers with tunable units

★ Each RBF unit has individually adjusted centre and diagonal covariance matrix

★ RBF units are selected in a computationally efficient orthogonal forward selection procedure

★ Each RBF unit is optimised by minimising leave-one-out misclassification rate, a measure of generalisation capability

Several examples have shown that proposed method compares favourably with existing state-of-the-art
THANK YOU.

S. Chen wish to thank the support of the United Kingdom Royal Academy of Engineering