WCCI 2006 Presentation

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 $\{(\mathbf{x}_k, y_k)\}_{k=1}^N$, where $y_k \in \{-1, +1\}$ is class label for m-dimensional pattern vector $\mathbf{x}_k$, construct RBF classifier

$$\tilde{y}_k = \mathrm{sgn}(\hat{y}_k) \quad \text{with} \quad \hat{y}_k = f_{\mathrm{RBF}}^{(M)}(\mathbf{x}_k) = \sum_{i=1}^{M} w_i g_i(\mathbf{x}_k),$$

  where $\tilde{y}_k$ is estimated class label for $\mathbf{x}_k$, $f_{\mathrm{RBF}}^{(M)}(\bullet)$ denotes RBF classifier with M units, and $\mathrm{sgn}(y) = -1$ if $y \le 0$, $\mathrm{sgn}(y) = +1$ if $y > 0$

- We consider general tunable RBF unit of form

$$g_i(\mathbf{x}) = K\left(\sqrt{(\mathbf{x} - \boldsymbol{\mu}_i)^T \, \boldsymbol{\Sigma}_i^{-1} \, (\mathbf{x} - \boldsymbol{\mu}_i)}\right)$$

  where $\boldsymbol{\mu}_i$ is centre vector of the ith RBF unit, whose diagonal covariance matrix is $\boldsymbol{\Sigma}_i = \mathrm{diag}\{\sigma_{i,1}^2, \cdots, \sigma_{i,m}^2\}$, and $K(\bullet)$ is basis function

RBF Model

- Regression model of RBF classifier

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

where $\mathbf{w} = [w_1 \; w_2 \cdots w_M]^T$ and $\mathbf{g}(k) = [g_1(\mathbf{x}_k) \; g_2(\mathbf{x}_k) \cdots g_M(\mathbf{x}_k)]^T$

- Define $\mathbf{y} = [y_1 \; y_2 \cdots y_N]^T$, $\mathbf{e} = [e_1 \; e_2 \cdots e_N]^T$, and $\mathbf{G} = [\mathbf{g}_1 \; \mathbf{g}_2 \cdots \mathbf{g}_M]$ with $\mathbf{g}_k = [g_k(\mathbf{x}_1) \; g_k(\mathbf{x}_2) \cdots g_k(\mathbf{x}_N)]^T$, $1 \le k \le M$
- Regression model over training data set:

$$\mathbf{y} = \mathbf{G}\mathbf{w} + \mathbf{e}$$

Note that $\mathbf{g}_k$ denotes kth column of $\mathbf{G}$ while $\mathbf{g}^T(k)$ is kth row of $\mathbf{G}$

- Let an orthogonal decomposition of regression matrix $\mathbf{G}$ be $\mathbf{G} = \mathbf{P}\mathbf{A}$. Then RBF model can alternatively be expressed

$$\mathbf{y} = \mathbf{P}\boldsymbol{\theta} + \mathbf{e}$$

Misclassification Rate

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

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

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

- Define signed decision variable

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

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

$$\mathcal{M}_r = \frac{1}{N}\sum_{k=1}^{N}\mathcal{I}_d(s_k) \quad \text{where} \quad \mathcal{I}_d(y) = \begin{cases} 1, & y \le 0 \\ 0, & y > 0 \end{cases}$$

Leave-One-Out Cross Validation

- Denote kth modelling error of n-unit RBF classifier, identified using the entire $\{(\mathbf{x}_k, y_k)\}_{k=1}^N$, as $e_k^{(n)} = y_k - f_{\text{RBF}}^{(n)}(\mathbf{x}_k) = y_k - \hat{y}_k^{(n)}$
- Let $f_{\text{RBF}}^{(n,-k)}(\bullet)$ be n-unit RBF classifier identified using $\{(\mathbf{x}_k, y_k)\}_{k=1}^N$ but with its kth data point being removed
- Test output of this n-unit RBF classifier at kth data point not used in training is computed by $\hat{y}_k^{(n,-k)} = f_{\text{RBF}}^{(n,-k)}(\mathbf{x}_k)$
- 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

$$\phi_k^{(n)} = \phi_k^{(n-1)} + y_k\, \theta_n\, p_n(k) - \frac{p_n^2(k)}{\mathbf{p}_n^T \mathbf{p}_n + \lambda}$$

and

$$\eta_k^{(n)} = \eta_k^{(n-1)} - \frac{p_n^2(k)}{\mathbf{p}_n^T \mathbf{p}_n + \lambda}$$

- Proposed algorithm constructs RBF units one by one by minimising J_n

Positioning and Shaping RBF Unit

- At nth construction stage, determine nth RBF unit by minimising J_n

$$\min_{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n} J_n\left(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right)$$

- Construction procedure is automatically terminated when

$$J_M \le J_{M+1}$$

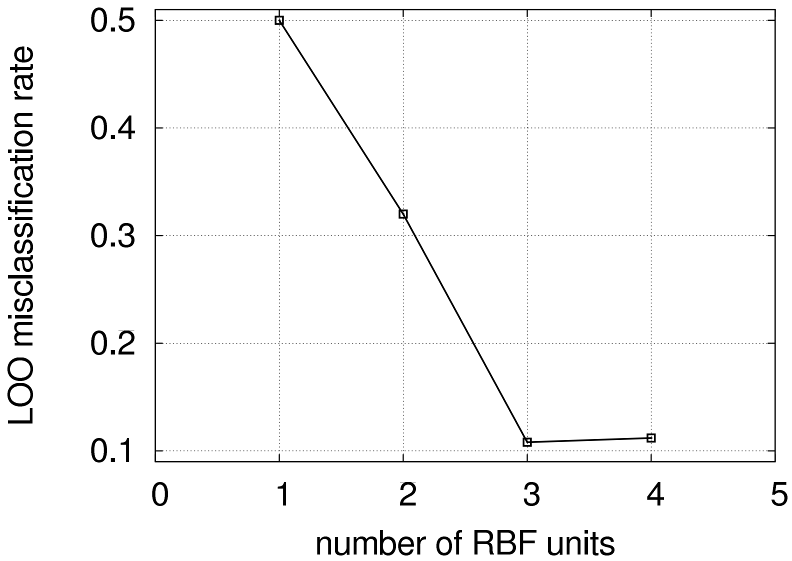
  yielding M-term RBF classifier

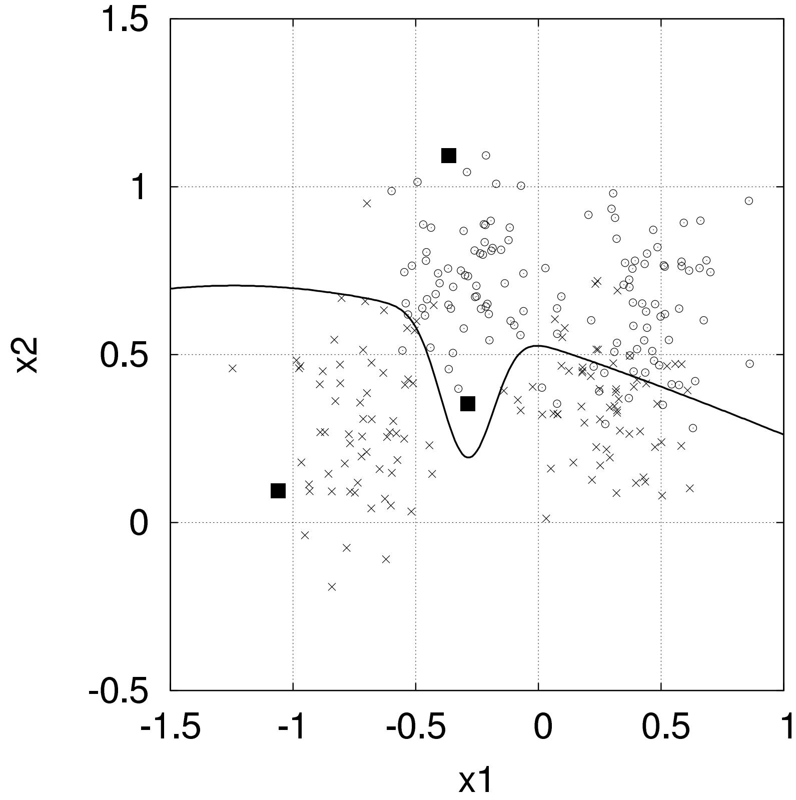
- Note that LOO criterion J_n is at least locally convex, and there exists an "optimal" M such that: for $n \le M$ J_n decreases as model size n increases while the above condition holds

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

Synthetic Two-Class Problem

B.D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996. `http://www.stats.ox.ac.uk/PRNN/`





algorithm	model size	test error rate
SVM	38	10.6%
RVM	4	9.3%
Proposed	3	8.0%

SVM and RVM quoted from M.E. Tipping, *J. Machine Learning Research*, vol.1, pp.211–244, 2001.

Breast Cancer Data Set

Average classification test error rate in % over 100 realizations

method	test error rate	model size
RBF-Network	27.64 ± 4.71	5
AdaBoost with RBF-Network	30.36 ± 4.73	5
LP-Reg-AdaBoost (-”-)	26.79 ± 6.08	5
QP-Reg-AdaBoost (-”-)	25.91 ± 4.61	5
AdaBoost-Reg (-”-)	26.51 ± 4.47	5
SVM with RBF-Kernel	26.04 ± 4.74	not available
Kernel Fisher Discriminant	24.77 ± 4.63	not available
Proposed	24.49 ± 3.28	3.1 ± 1.2

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

method	test error rate	model size
RBF-Network	24.29 ± 1.88	15
AdaBoost with RBF-Network	26.47 ± 2.29	15
LP-Reg-AdaBoost (-"-)	24.11 ± 1.90	15
QP-Reg-AdaBoost (-"-)	25.39 ± 2.20	15
AdaBoost-Reg (-"-)	23.79 ± 1.80	15
SVM with RBF-Kernel	23.53 ± 1.73	not available
Kernel Fisher Discriminant	23.21 ± 1.63	not available
Proposed	22.16 ± 1.47	4.0 ± 1.6

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

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

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