ORTHOGONAL LEAST SQUARES ALGORITHM FOR TRAINING MULTI-OUTPUT RADIAL BASIS FUNCTION NETWORKS

S. Chen, P.M. Grant and C.F.N. Cowan†

University of Edinburgh, U.K.; †Loughborough University of Technology, U.K.

1. Introduction

The radial basis function (RBF) network [1] offers a viable alternative to the two-layer neural network in many signal processing applications. A novel learning algorithm for RBF networks [2-4] has been derived based on the orthogonal least squares (OLS) method operating in a forward regression manner [5]. This is a rational way to choose RBF centres from data points because each selected centre maximizes the increment to the explained variance of the desired output and the algorithm does not suffer numerical ill-conditioning problems. This learning algorithm was originally derived for RBF networks with a scalar output. The present study extends this previous result to multi-output RBF networks. The basic idea is to use the trace of the desired output covariance as the selection criterion instead of the original variance in the single-output case. Reconstruction of PAM signals and nonlinear system modelling are used as two examples to demonstrate the effectiveness of this learning algorithm.

2. Multi-output RBF network

The RBF network with n inputs and m outputs depicted in Fig.1 implements a mapping $f_r:\mathbf{R}^n \rightarrow \mathbf{R}^m$ according to

$$f_{r_i}(x)=\sum_{j=1}^{n_h}\lambda_{ji}\phi(\|x-c_j\|),\ 1\le i\le m, \tag{1}$$

where $x\in\mathbf{R}^n$, $\phi(.)$ is a given function from $\mathbf{R}^+$ to $\mathbf{R}$, $\|.\|$ denotes the Euclidean norm, λ_{ji} are the weights, $c_j\in\mathbf{R}^n$ are known as the RBF centres and n_h is the number of centres. Two typical choices of $\phi(.)$ are the thin-plate-spline function

$$\phi(v)=v^2\log(v), \tag{2}$$

and the Gaussian function

$$\phi(v)=exp(-v^2/\sigma), \tag{3}$$

where σ is a width constant. The centres are fixed points in $\mathbf{R}^n$ and are normally selected from the input data points. If a set of the inputs and the corresponding desired outputs $\{x(t),d(t)\}_{t=1}^N$ is provided, the weights λ_{ji} can be determined using the LS method.

The RBF network (1) is a special case of the multi-output linear regression model

$$d_i(t)=\sum_{j=1}^{M}p_j(t)\theta_{ji}+e_i(t),\ 1\le i\le m, \tag{4}$$

where $d_i(t)$ is the ith desired output, θ_{ji} are the parameters, $p_j(t)$ are known as the regressors which are some fixed functions of the input $x(t)$ and $e_i(t)$ is the ith error signal. It is apparent that a fixed centre c_j with a given nonlinearity $\phi(.)$ represents a regressor in (4).

Define

$$\mathbf{d}_i=[d_i(1)\cdots d_i(N)]^T,\ 1\le i\le m, \tag{5}$$

$$\mathbf{e}_i=[e_i(1)\cdots e_i(N)]^T,\ 1\le i\le m, \tag{6}$$

$$\mathbf{p}_j=[p_j(1)\cdots p_j(N)]^T,\ 1\le j\le M. \tag{7}$$

Then for $t=1$ to N (4) can be collectively written as

$$[\mathbf{d}_1\cdots\mathbf{d}_m]=[\mathbf{p}_1\cdots\mathbf{p}_M]\begin{bmatrix}\theta_{11} & \cdots & \theta_{1m}\\ \vdots & & \vdots\\ \theta_{M1} & \cdots & \theta_{Mm}\end{bmatrix}+[\mathbf{e}_1\cdots\mathbf{e}_m] \tag{8}$$

or more concisely in the matrix form

$$\mathbf{D}=\mathbf{P}\Theta+\mathbf{E}. \tag{9}$$

3. Multi-output OLS algorithm

The OLS method involves the transformation of the set of basis vectors $\mathbf{p}_j$ into a set of orthogonal basis vectors by decomposing $\mathbf{P}$ into

$$\mathbf{P}=\mathbf{W}A, \tag{10}$$

where

$$A=\begin{bmatrix}1 & \alpha_{12} & \cdots & \alpha_{1M}\\ 0 & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \alpha_{M-1M}\\ 0 & \cdots & 0 & 1\end{bmatrix} \tag{11}$$

and

$$\mathbf{W}=[\mathbf{w}_1\cdots\mathbf{w}_M], \tag{12}$$

with orthogonal columns that satisfy

$$\mathbf{W}^T\mathbf{W}=\begin{bmatrix}\mathbf{w}_1^T\mathbf{w}_1 & & 0\\ & \ddots & \\ 0 & & \mathbf{w}_M^T\mathbf{w}_M\end{bmatrix}. \tag{13}$$

The space spanned by the set of $\mathbf{w}_j$ is the same space spanned by the set of $\mathbf{p}_j$, and (9) can be rewritten as

$$\mathbf{D}=\mathbf{W}G+\mathbf{E}. \tag{14}$$

The OLS solution

$$G=\begin{bmatrix} g_{11} & \cdots & g_{1m} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ g_{M1} & \cdots & g_{Mm} \end{bmatrix} \quad (15)$$

and the ordinary LS solution Θ satisfy the triangular system

$$A\Theta=G. \quad (16)$$

The classical and modified Gram-Schmidt methods [6] can be employed to derive A and G and thus to solve for Θ from (16). The Householder transformation method [7] can alternatively be used to obtain a similar orthogonal decomposition. In the case of RBF networks, each data point $x(t)$ is a candidate centre which corresponds to a candidate regressor. The number of data points is often very large and, therefore, the number of all the candidate regressors M can be very large. An adequate modelling however may only require $M_s(<<M)$ significant regressors. The OLS algorithm [5] offers a simple and effective means to select these significant regressors. Because the error matrix $\mathbf{E}$ is orthogonal to $\mathbf{W}$, after some simple calculation, the trace of the covariance of $d(t)$ is

$$trace(\mathbf{D}^T\mathbf{D}/N)=\sum_{j=1}^{M}(\sum_{i=1}^{m}g_{ji}^2)\mathbf{w}_j^T\mathbf{w}_j/N + trace(\mathbf{E}^T\mathbf{E}/N). \quad (17)$$

The error reduction ratio due to $\mathbf{w}_j$ can be defined as

$$[err]_j=\frac{(\sum_{i=1}^{m}g_{ji}^2)\mathbf{w}_j^T\mathbf{w}_j}{trace(\mathbf{D}^T\mathbf{D})}, \quad 1\le j\le M. \quad (18)$$

The following forward regression procedure can be used to select a subset of significant regressors. At the kth step, a regressor is selected if it produces the largest value of $[err]_k$ from amongst the rest of the $M-k+1$ candidates. The selection is terminated when

$$1-\sum_{k=1}^{M_s}[err]_k<\rho \quad (19)$$

where $0<\rho<1$ is a chosen tolerance. This gives rise to a subset model containing M_s significant regressors.

The first term in the right-hand side of (17) is the part of the trace of the desired output covariance which can be explained by the regressors and the second term is the unexplained trace of the desired output covariance. Thus

$$(\sum_{i=1}^{m}g_{ji}^2)\mathbf{w}_j^T\mathbf{w}_j/N \quad (20)$$

is the increment to the explained trace due to $\mathbf{w}_j$, and the above learning procedure has a property that each selected centre (regressor) maximizes the increment to the explained trace of the desired output covariance. The selection of centres is therefore directly linked to the reduction of the error covarianc trace. This is clearly superior to a random selection of centres proposed originally in [1]. The detailed selection procedure is exactly the same as for the single-output case described in [5,8] except for some obvious alterations required by the multi-output natrure. Another advantage of this algorithm is that numerical ill-conditioning can easily be avoided. It is straightforward to show that $\mathbf{w}_k^T\mathbf{w}_k=0$ implies that $\mathbf{p}_k$ is a linear combination of $\mathbf{p}_1$ to $\mathbf{p}_{k-1}$. Therefore if $\mathbf{w}_k^T\mathbf{w}_k$ is less than a small pre-set threshold, the regressor $\mathbf{p}_k$ will not be selected and this ensures that the LS solution is well-conditioned. The desired choice of ρ is discussed in [5,8].

4. Reconstruction of PAM signals

A general digital communications system is shown in Fig.2, where the channel is modelled as a finite impulse response filter with additive white Gaussian noise $e(t)$. The task of the equaliser is to reconstruct input symbol based on the channel observation vector $[y(t)\cdots y(t-\eta+1)]^T$. The integers η and τ are known as the equaliser order and delay respectively. In the present study, $s(t)$ is asumed to be a 4-ary PAM signal taking values from the set $\{\pm1,\pm3\}$. Equalisation is a nonlinear classification problem, and this is best illustrated using a simple example where the channel transfer function is

$$H(z)=1.0+0.5z^{-1}$$

and the equaliser has a structure of $\eta=2$ and $\tau=0$. In the absence of noise, channel output vectors are some discrete points. Each of these points is shown in Fig.3 using one of the 4 symbols {✱,◆,×,□}, which correspond to the input set $\{-3,-1,1,3\}$. When noise is added, some probability distribution is introduced, giving rise to a 4-state classification problem. For a noise variance 0.0625, the optimal decision boundaries are plotted in Fig.3.

A two-output RBF network can be trained to approximate this optimal equaliser solution. The nonlinearity $\phi(.)$ was chosen as (3), where σ was set to twice large of the noise varinace. 740 points of training data were generated and this gave rise to about 740 candidate centres. A RBF network of 74 centres was selected using the OLS learning algorithm and the decision boundaries of this RBF network are also shown in Fig.3. Fig.4 compares the performance of the optimal equaliser with that of the selected RBF network.

5. Nonlinear system modelling

1000 samples of simulated time series were generated using the nonlinear model

$$y_1(t)=(0.8-0.5exp(-y_1^2(t-1)))y_1(t-1) -(0.3+0.9exp(-y_1^2(t-1)))y_1(t-2) +0.1sin(y_2(t-1))+e_1(t)$$

$$y_2(t)=0.6y_2(t-1)+0.2y_2(t-1)y_2(t-2) +1.2tanh(y_1(t-2))+e_2(t)$$

where the Gaussian noise $e(t)$ had a covariance

$$\begin{bmatrix} 0.01 & 0.0 \\ 0.0 & 0.01 \end{bmatrix}$$

A two-output RBF network was employed to model this nonlinear process. Let

$x(t)=[y^T(t-1)y^T(t-2)]^T$

The idea is to use the RBF network

$\hat{y}(t)=f_r(x(t))$

as the one-step-ahead predictor for $y(t)$. $\phi(.)$ was chosen to be (2). The OLS algorithm identified a RBF network with 50 centres. The observations and the selected centres are plotted in Fig.5. The selected network is a very good one-step-ahead predictor for the simulated system. This RBF model was also used to produce iteratively the network output

$\hat{y}_d(t)=f_r(x_d(t))$

where $x_d(t)=[\hat{y}_d^T(t-1)\hat{y}_d^T(t-2)]^T$. Even though the RBF model was identified using noisy observations, the iterative network outputs closely match to the outputs from the autonomous system ($e(t)=0$) as can be seen in Fig.6. This confirms that the selected RBF model indeed captures the underlying dynamics of the system.

6. Conclusions

An orthogonal least squares algorithm has been developed for the construction of multi-output radial basis function networks. This learning strategy provides a systematic approach linking the selection of RBF centres from the data set to the reduction of the error covariance trace. Application to two different areas of signal processing has been demonstrated.

References

[1] D.S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks", *Complex Systems*, Vol.2, 1988, pp.321-355.

[2] S. Chen, S.A. Billings, C.F.N. Cowan and P.M. Grant, "Non-linear systems identification using radial basis functions", *Int. J. Systems Sci.*, Vol.21, No.12, 1990, pp.2513-2539.

[3] S. Chen, S.A. Billings, C.F.N. Cowan and P.M. Grant, "Practical identification of NARMAX models using radial basis functions", *Int. J. Control*, Vol.52, No.6, 1990, pp.1327-1350.

[4] S. Chen, C.F.N. Cowan and P.M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks", *IEEE Trans. Neural Networks*, Vol.2, No.2, 1991, pp.302-309.

[5] S. Chen, S.A. Billings and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification", *Int. J. Control*, Vol.50, No.5, 1989, pp.1873-1896.

[6] A. Björck, "Solving linear least squares problems by Gram-Schmidt orthogonalization", *Nordisk Tidskr. Informations-Behandling*, Vol.7, 1967, pp.1-21.

[7] G. Golub, "Numerical methods for solving linear least squares problems", *Numerische Mathematik*, Vol.7, 1965, pp.206-216.

[8] S.A. Billings and S. Chen, "Extended model set, global data and threshold model identification of severely non-linear systems", *Int. J. Control*, Vol.50, No.5, 1989, pp.1897-1923.

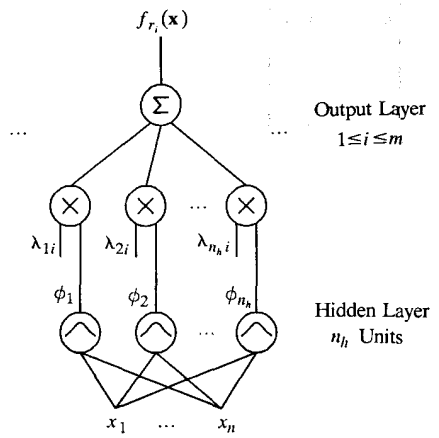

Fig.1. Schematic of RBF Network.

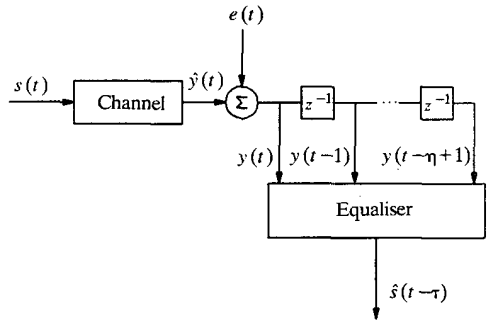

Fig.2. Schematic of Data Transmission System.

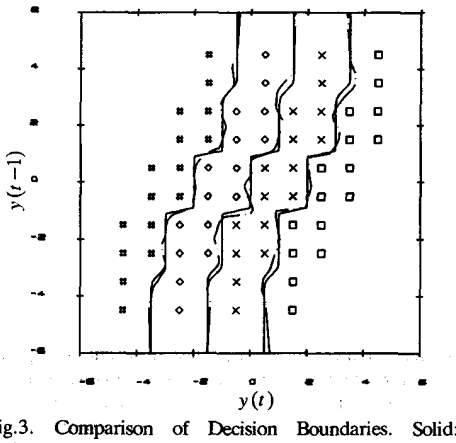

Fig.3. Comparison of Decision Boundaries. Solid: optimal, dashed: RBF network, noise variance 0.0625.

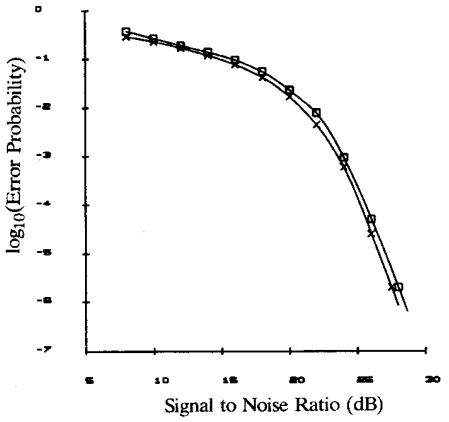

Fig.4. Comparison of Performance. —×— optimal, —□— RBF network.

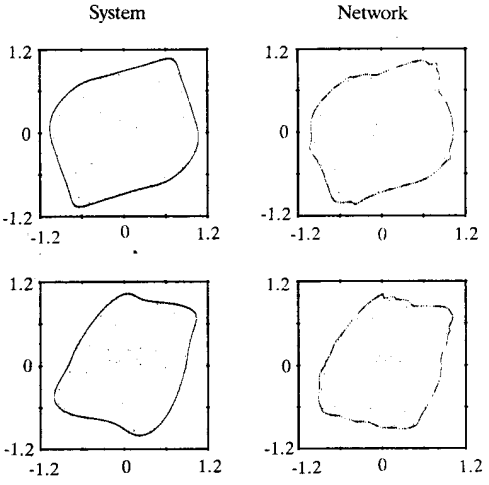

Fig.6. Response of Autonomous System and Iterative Network. Project to two subspaces, 1000 samples.

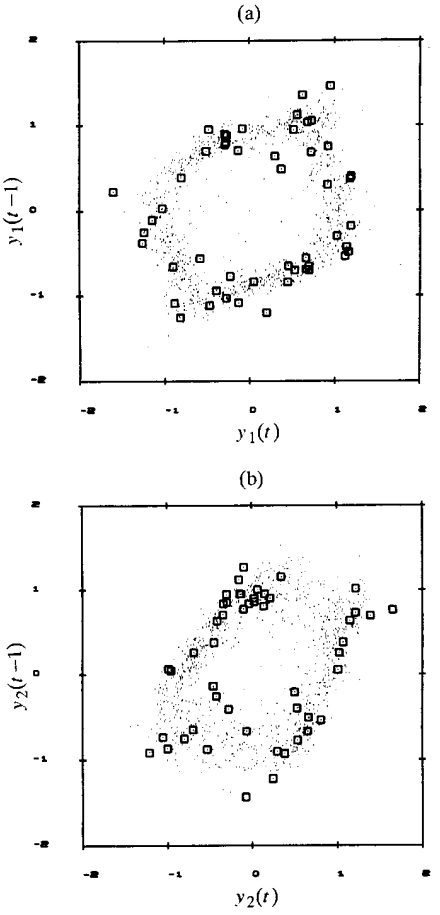

Fig.5. Subspace Projection of Observations (·) and Centres (□).