MODELS OF DYNAMIC COMPLEXITY FOR TIME-SERIES PREDICTION

Visakan Kadirkamanathan *Mahesan Niranjan* *Frank Fallside*
Cambridge University Engineering Department,
Trumpington street, Cambridge CB2 1PZ, UK

ABSTRACT

In this paper, we have developed a model of dynamic complexity, a growing Gaussian radial basis function (GRBF) network, by analysing sequential learning in the function space. The criteria to add a new basis function to the model are based on the angle formed between a new basis function and the existing basis functions and also on the prediction error. When a new basis function is not added the model parameters are adapted by the extended Kalman filter (EKF) algorithm. This model is similar to the resource allocating network (RAN) and hence this work provides an alternative interpretation to the RAN. An enhancement to the RAN is suggested where RAN is combined with EKF. The RAN and its variants are applied to the task of predicting the logistic map and the Mackey-Glass chaotic time-series and the advantages of the enhanced model is demonstrated.

1. INTRODUCTION

Artificial neural networks (ANNs) have emerged as a powerful class of nonlinear models for predicting time-series. In using ANNs, the problem of time-series prediction is reduced to an approximation problem, where, it is assumed that the series value in the future is only a function of the past few values. The potential of ANNs lie in their ability to construct a good approximation to the underlying functional relationship, provided it exists.

The goodness of approximation depends on the complexity[1] of the ANN and the amount of data that is available to estimate the network parameters. A difficulty with using nonlinear models such as ANNs to approximate the underlying function is that the complexity (hence the size) of the network must be determined *a priori.* A network that is too large for the problem is known to suffer from poor prediction performance, since the estimated parameters tend to have a large variance — the problem known in neural network literature as *generalisation.* A smaller network does not have the capacity to approximate the underlying model well enough to give good prediction performance. This observation has lead to recent development of networks that increase their complexity, dynamically, with increasing complexity of the task during learning. The importance of optimal complexity models can also be appreciated from computational considerations.

Time-series prediction with ANNs have largely involved the use of block estimation algorithms, as in [5, 7] and the references therein. The assumption used in such approaches is that the time-series data are available en-bloc. We consider here, prediction problems where the time-series data is received on-line with prediction performance continually evaluated. This demands the use of sequential (or recursive) algorithms where adaptation is based on the prediction performance.

In this paper, we provide a function space approach to sequential learning in nonlinear models and develop a model of dynamic complexity, similar to the resource allocating network (RAN) of Platt [8]. This model and its variants are applied to the task of predicting the *logistic map* and the *Mackey-Glass* chaotic time-series.

2. A FUNCTION SPACE APPROACH TO SEQUENTIAL LEARNING

Learning in nonlinear models with fixed complexity can be viewed as a problem of determining the optimal set of parameter values. This is often posed as an optimisation problem in the parameter space. An alternative approach is to view sequential learning in the function space, observing that neural networks and other nonlinear models provide an input – output mapping. In the function space approach, we are no longer limited to using a network of fixed size which defines the parameter space; models with increasing complexity can be readily analysed.

The principle of $\mathcal{F}$-Projection is developed as a sequential learning method from a function space approach [2]. The principle, subject to an added constraint that the underlying function is smooth and continuous, provides an approximate solution [1], given by,

$$f^{(n)}(\underline{x}) = f^{(n-1)}(\underline{x}) + e_n \phi_n(\underline{x}) \tag{1}$$

where $f^{(n)}$ is the posterior estimate of the underlying function, $f^{(n-1)}$ is the prior estimate, $\phi_n(\underline{x})$ is a spatially localised basis function such as the Gaussian radial basis function (GRBF) and e_n is the prediction error for the present observation $(\underline{x}_n, y_n)$, given by,

$$e_n = y_n - f^{(n-1)}(\underline{x}_n) \tag{2}$$

Suppose $f^{(n-1)}$ is constructed by a linear combination of a set of GRBFs, essentially the mapping constructed by a Gaussian RBF network,

$$f^{(n-1)}(\underline{x}) = \sum_{k=1}^{K} \alpha_k \phi_k(\underline{x}) \tag{3}$$

where $\phi_k(\underline{x})$ is the k^{th} GRBF, given by,

$$\phi_k(\underline{x}) = \exp\left\{ -\frac{1}{\sigma_k^2} \|\underline{x} - \underline{\mu}_k\|^2 \right\} \tag{4}$$

[1] The measure of complexity here is the number of adaptable parameters in the network.

α is the coefficient of the basis function, $\underline{\mu}$ is the centre of the GRBF in the input space and σ is the width of the GRBF. The posterior estimate given in equation (1) is a mapping described by the RBF network with an added hidden unit with values given by,

$$\alpha_{K+1} = e_n \tag{5}$$
$$\underline{\mu}_{K+1} = \underline{x}_n \tag{6}$$
$$\sigma_{K+1} = \sigma_0 \tag{7}$$

The width parameter σ_{K+1} determines the smoothness of the GRBF and therefore the value σ_0 represents the required smoothness.

We thus have a network that learns by growing with each new observation. Since the observations $(\underline{x}_n, y_n)$ are stored implicitly, $\underline{x}_n$ as the centres of the GRBFs and y_n as the coefficients of the basis functions, the solution obtained is similar in spirit to the Parzen window method. A problem with this solution is that the network grows indefinitely when applied to an on-line learning problem.

3. A DYNAMIC COMPLEXITY MODEL

The function space approach can be extended to provide a criteria to limit the growth of the network. In doing so, we arrive at a dynamic complexity model is are complex enough to give good prediction performance but not too complex to suffer from generalisation or large computational overhead.

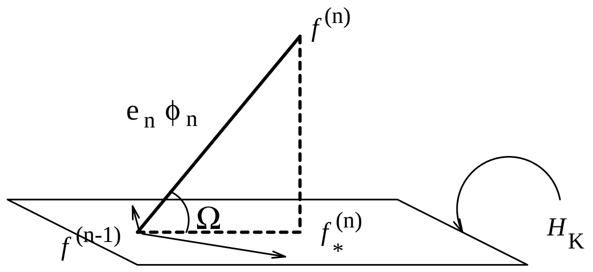


Figure 1: 3-D illustration of the error in approximation due to not adding a new basis function

The K basis functions which are linearly combined, form a K dimensional subspace, $\mathcal{H}_K$, in the infinite dimensional Hilbert space of square integrable real functions. A simplified 3-D illustration is given in Figure 1. The error in approximation due to not adding the basis function ϕ_n is $\|f^{(n)} - f_*^{(n)}\|$, where $\|.\|$ is the L_2 norm in the Hilbert space and $f_*^{(n)}$ is the posterior estimate that can be described by the K basis functions and has the least distance to $f^{(n)}$ in the Hilbert space. This error, E, is given by,

$$E = |e_n|.\|\phi_n\| \sin(\Omega) \tag{8}$$

where Ω is the angle formed by the new basis function ϕ_n to the subspace $\mathcal{H}_K$ defined by the K basis functions in $f^{(n-1)}$. The norm of the basis function ϕ_n depends only on the width σ_0. Hence, the error E depends on the parameters e_n and Ω. The angle lies between 0 and $\frac{\pi}{2}$ and therefore $0 \leq \sin(\Omega) \leq 1$.

The criterion to add a new basis function is based on whether the added complexity of introducing a new basis function is greater than the approximation error that would incur otherwise. The criterion that E is above a threshold can be simplified into the parameters e_n and Ω both exceeding threshold values, *viz.*,

$$e_n > e_{\min} \tag{9}$$
$$\Omega > \Omega_{\min} \tag{10}$$

The angle Ω is difficult to evaluate in general. An approximation is to find the smallest angle between the new basis function and all other existing basis functions. The angle between two GRBFs of the same width σ_0 is given by (in [1]),

$$\Omega_k = \cos^{-1}\left(\exp\left\{-\frac{1}{2\sigma_0^2}\|\underline{x}_n - \underline{\mu}_k\|^2\right\}\right) \tag{11}$$

The angle criterion for growth is then reduced to a criterion on the values of the basis functions to the input $\underline{x}_n$, expressed as,

$$\sup_k \quad \phi_k(\underline{x}_n) < \cos^2(\Omega_{\min}) \tag{12}$$

or equivalently as the criterion based on the distance between the input $\underline{x}_n$ and the GRBF centres $\underline{\mu}_k$, *i.e.*,

$$\inf_k \quad \|\underline{x}_n - \underline{\mu}_k\| > \epsilon_0 \tag{13}$$

Given a new observation $(\underline{x}_n, y_n)$, if the criteria for growth given in equations (9) and (13) are satisfied, a new basis function is added. If a new basis function is not added, the function space approach suggests the use of the principle of $\mathcal{F}$-Projection, which gives a posterior function estimate of the network to be least distant from the prior and also satisfy the constraint that $y_n = f^{(n)}(\underline{x}_n)$.

For a network of fixed complexity, this principle has a close relationship to the extended Kalman filter (EKF) algorithm [3]. Hence, we shall use the EKF algorithm to adapt the coefficients of the basis when a new hidden unit is not added. When the new basis function is added the parameter values are chosen according to equations (5), (6) and (7).

The EKF algorithm adapts a set of parameters $\underline{\theta}$ according to:

$$\underline{\theta}^{(n)} = \underline{\theta}^{(n-1)} + e_n \underline{k}_n \tag{14}$$
$$\underline{k}_n = (\underline{a}^T P_{n-1} \underline{a} + R)^{-1} P_{n-1} \underline{a} \tag{15}$$
$$P_n = [I - \underline{k}_n \underline{a}^T] P_{n-1} \tag{16}$$

where $\underline{a} = \nabla_\theta f$ and $\nabla_\theta f$ is the gradient of the network mapping f (in equation (3)) with respect to the parameters $\underline{\theta}$ estimated with the values at time $(n-1)$. R is the measurement noise variance and $\underline{k}_n$ is known as the Kalman gain. If only the coefficients of the basis functions are adapted, then $\underline{a} = [\phi_1(\underline{x}_n), \ldots, \phi_K(\underline{x}_n)]^T$.

The symmetric matrix P_n is an estimated error covariance matrix for the parameters with dimensionality equalling that of $\underline{\theta}$. In the event of adding a new basis function, the size of P_n must also be increased appropriately. We assign P_n according to:

$$P_n = \begin{pmatrix} P_{n-1} & 0 \\ 0 & p_0 I \end{pmatrix} \tag{17}$$

where I is the identity matrix with dimensionality equal to the number of adaptable parameters introduced to the network by the allocation of a new basis function. The value p_0 reflects the uncertainty in the corresponding parameter values of the new basis function and is set equal to the value of R.

4. THE RESOURCE ALLOCATING NETWORK

The dynamic complexity model developed from the function space approach is similar in form to the resource allocating network (RAN) of Platt [8]. RAN differs from the above model in three aspects. Firstly, the width of the basis function σ is assigned according to,

$$\sigma_{K+1} = \kappa \| \underline{x}_n - \underline{\mu}_{nr} \| \tag{18}$$

where κ is an overlap parameter that determines the spatial overlap in the input space between the two basis functions and $\underline{\mu}_{nr}$ is the nearest GRBF centre to $\underline{x}_n$ in the input space.

Secondly, the growth criterion based on the distances in the input space has the same form as in (13), but the parameter ϵ_0 is decreased exponentially until it reaches a lower bound, *viz.*,

$$\epsilon_0 = \min\left(r_{\min}, r_0 \exp\{-\frac{n}{\tau}\}\right) \tag{19}$$

where τ is a decay constant. The lower bound effectively provides an upper bound for the width of the GRBF, σ, ensuring the smoothness of the basis functions. The exponential decaying of the distance criterion allows fewer basis functions with small widths (smoother basis functions) initially and as time goes, more basis functions with larger widths are allocated to fine tune the approximation.

The third difference is that RAN uses LMS algorithm to adapt the parameters α_k and $\underline{\mu}_k$, instead of the EKF, when a new basis function is not added. The parameters $\underline{\theta} = [\ldots, \alpha_k, \underline{\mu}_k, \ldots]$ are adapted according to,

$$\underline{\theta}^{(n)} = \underline{\theta}^{(n-1)} - \eta \; e_n \nabla_\theta f \tag{20}$$

where η is the adaptation step size. The distance criterion given in equation (13) is also arrived at from the angle criterion for GRBFs with differing widths [1].

The development of RAN as a function interpolating network can be seen to be the extension of the restricted Coulomb energy model of pattern classification [9] and is based on the localisation property of the basis functions in the input space. In contrast, we have shown that RAN has its foundations in the analysis of sequential learning in the function space.

An algorithm developed from the principle of $\mathcal{F}$-Projection has been shown to converge faster than LMS for the GRBF network [4]. It is computationally intense. The relationship it has to the EKF [3] suggests an enhancement to the RAN where the EKF is used in place of LMS, the network being referred to as RAN-EKF.

5. RESULTS ON PREDICTING CHAOTIC SERIES

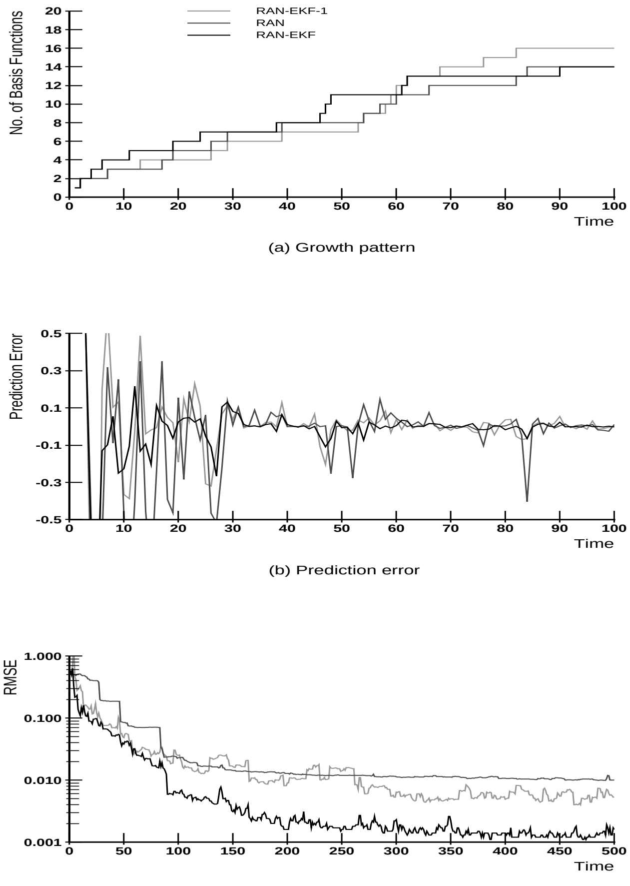
Chaotic time-series have been used to illustrate the advantages of using ANNs for predicting time-series [5, 6]. They are generated by a deterministic low order nonlinear map and pass the statistical tests for randomness. Here, we consider the logistic map and the Mackey-Glass chaotic time-series.

The particular logistic map chosen is generated by a difference equation, given by,

$$s_n = 4 s_{n-1} \left(1 - s_{n-1}\right) \tag{21}$$

The task for the models is to predict one-step ahead, that is to predict the value of s_n based on s_{n-1}. Since there exists an exact relationship via equation (21) good prediction performance can be expected from models that learn to approximate this underlying mapping based on the observations received. Since the prediction error will also indicative of the approximation error, the root mean squared error (RMSE) based on a test series is used as a performance index.

Three different forms of RAN have been used. RAN: as proposed by Platt using the LMS algorithm, RAN-EKF: the RAN with EKF to adapt the coefficients and the centres of GRBF and RAN-EKF1: the RAN with EKF to adapt only the coefficients. The parameters used in the growth of the network are, $r_0^2 = 0.125$, $r_{\min}^2 = 0.00125$, $e_{\min} = 0.05$. The parameters used for adaptation are $\eta = 0.02$, $P_0 = I$, $Q = 0.02I$, $R = 0.01$ and $p_0 = 1.0$. The growth pattern and prediction error for upto 100 samples and RMSE upto 500 samples are shown in Figure 2.



(a) Growth pattern

(b) Prediction error

(c) Approximation error

Figure 2: Prediction results for the logistic map

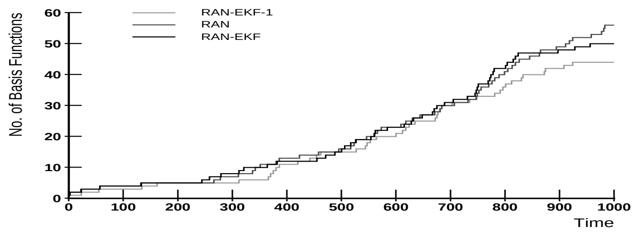
The prediction error and the RMSE clearly demonstrate the advantage of using EKF with RAN instead of the LMS algorithm. The growth pattern of the networks were very similar, indicating that the prediction error criterion did not contribute to the saturation of the network. The RAN-EKF models exhibited very low prediction errors in comparison to RAN and the RMSE of RAN-EKF was an order lower than that of RAN.

The Mackey-Glass time-series is generated a delay differential equation, given by,

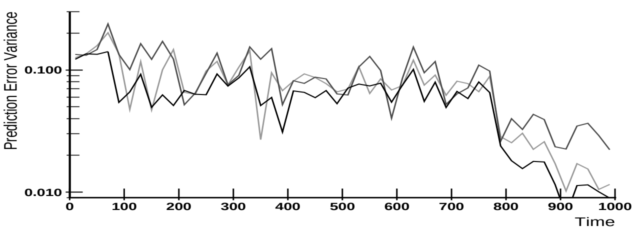
$$\frac{ds(t)}{dt} = \frac{a.s(t-\tau)}{1+s(t-\tau)^{10}} - b.s(t) \qquad (22)$$

with values $a = 0.2$, $b = 0.1$ and $\tau = 17$. Here, the task of prediction is to predict s_{n+85} from s_n, s_{n-6}, s_{n-12} and s_{n-18}. As such, an exact mapping between the input and output may not exist. The models however attempt to find an approximation to this underlying mapping.

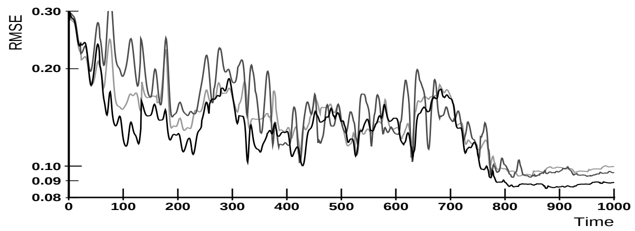
The growth parameters of RAN are chosen with $r_0 = 0.7$, $r_{\min} = 0.07$ and $e_{\min} = 0.02$. The adaptation parameters are $\eta = 0.05$, $P_0 = I$, $Q = 0.01I$ and $R = p_0 = 1.0$. The growth pattern, prediction error and RMSE are shown in Figure 3.



(a) Growth pattern



(b) Prediction error



(c) Approximation error

Figure 3: Prediction results for the Mackey-Glass series

The growth pattern of all three models are similar, with RAN-EKF1 being the most compact with 44 basis functions and RAN the largest with 55. The RMSE is also comparable for all three models, except in the initial stages where RAN-EKF and RAN-EKF1 show faster adaptation. The plot of prediction error is complicated and does not emphasise the low errors achieved by RAN-EKFs. An alternative measure is plotted where, the prediction error is squared, summed and averaged for 20 samples. The prediction performance for RAN-EKF was better than RAN at all times. Significant improvement is achieved only in the initial and final stages of prediction.

6. CONCLUSIONS

We have adopted a function space approach to sequential learning as an alternative to the parameter space approach. This has led to the development of a model that dynamically increases its complexity. For each new observation, a new localised basis function is added to the existing function mapped by the model. For a Gaussian radial basis function (GRBF) network, this amounts to adding a new hidden unit.

The function space approach is extended to derive criteria upon which the decision to add a new basis function is based. The criteria are based on the prediction error and the output of the basis functions for the new observation or equivalently, the distance between the new observation and the existing centres of the GRBFs in the input space. These criteria are similar to those of RAN.

An enhancement to RAN is suggested where RAN uses the EKF algorithm to adapt the parameters when a new basis function is not added, instead of the LMS algorithm. Here, we have showed that RAN with EKF performs better than RAN in predicting the logistic map and Mackey-Glass chaotic time-series. Our current interest is in applying the above work to nonstationary signals and developing criteria for complexity reduction by eliminating basis functions that contribute little to the mapping.

REFERENCES

[1] V. Kadirkamanathan. *Sequential learning in artificial neural networks.* PhD thesis, Cambridge University Engineering Department, October 1991.

[2] V. Kadirkamanathan and F. Fallside. F-Projection: A nonlinear recursive estimation algorithm for neural networks. Tech. Report CUED/F-INFENG/TR.53, Cambridge University Engineering Department, October 1990.

[3] V. Kadirkamanathan and M. Niranjan. Nonlinear adaptive filtering in nonstationary environments. In *Proc. ICASSP*, Toronto, 1991.

[4] V. Kadirkamanathan, M. Niranjan, and F. Fallside. Sequential adaptation of radial basis function neural networks and its application to time-series prediction. In R. P. Lippmann et al. eds., *Neural Information Processing Systems 3.* Morgan Kaufmann, 1991.

[5] A. S. Lapedes and R. M. Farber. Nonlinear signal processing using neural networks: prediction and system modelling. Tech. Report LA-UR-87-2662, Los Alamos National Laboratory, 1987.

[6] J. Moody and C. Darken. Learning with localized receptive fields. Tech. Report YALEU/DCS/RR-649, Dept. of Computer Science, Yale University, 1988.

[7] M. Niranjan and V. Kadirkamanathan. A nonlinear model for time-series prediction and signal interpolation. In *Proc. ICASSP*, Toronto, 1991.

[8] J. Platt. A resource allocating network for function interpolation. *Neural Computation*, 3(2), 1991.

[9] D. L. Reilly, L. N. Cooper, and C. Elbaum. A neural model for category learning. *Biological Cybernetics*, 45:35–41, 1982.