AN IMPROVED NOVELTY CRITERION FOR RESOURCE ALLOCATING NETWORKS

A M^{c}Lachlan

Neural Computing Research Group
Aston University
UK

ABSTRACT

In this paper, we introduce a new novelty criterion for resource allocating RBF networks (RANs) based on standard signal processing theory. This network growth prescription is considerably less sensitive to noise and outliers than those of previous RANs, and also removes the need for ad-hoc hyperparameters. An added advantage of this novelty criterion is that, as it is independent of the parameters of the extended Kalman filter training algorithm, the filter can be modified for application to slowly varying non-stationary environments without adversely affecting the network's capacity for growth.

We demonstrate the relative improvement of this criterion on two non-stationary real-world problems : electricity load forecasting and exchange rate prediction.

THE RAN NETWORK

The resource allocating network was introduced by Platt [7] as a means of constructing a network of adequate complexity online. It is a Gaussian RBF, whose training requires a novelty criterion (which determines whether a new basis function is to be added) and an online training algorithm (which updates the existing weights).

Considerable work has been performed regarding training algorithms for both stationary and non-stationary data [3, 4, 5], but little has been done regarding novelty criteria, and it is this problem which we shall be addressing in this paper. It should be emphasised that we will be concerning ourselves with improvements to the RAN prescription rather than with comparisons of RAN performance to other techniques. Such analysis can be found elsewhere in the literature [7, 3, 6].

THE EXTENDED KALMAN FILTER

As the networks in this paper will be trained using the extended Kalman filter, we shall briefly describe the algorithm [1] in the context of its application to neural networks before proceeding with the discussion of novelty criteria.

We are interested obtaining an online estimate of the network weights from the data sequence $Y_t = \{\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, ...\}$. The state space equations specify the relation between the measurements $\boldsymbol{y}_t$ and the weights $\boldsymbol{w}_t$ along with the evolution of the weights between timesteps. We view the network as a function of the weights $\boldsymbol{f}_t(\boldsymbol{w}_t)$, the input vectors being represented by the time index on the network function f_t.

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{f}_t(\boldsymbol{w}_t) + \boldsymbol{\nu}_t \\ \boldsymbol{w}_t &= \boldsymbol{w}_{t-1} + \boldsymbol{\mu}_{t-1} \end{aligned} \quad (1)$$

$\boldsymbol{\nu}_t$ and $\boldsymbol{\mu}_t$ are respectively measurement and system noise processes. While the weight evolution between timesteps is generally trivial, the introduction of system noise can be beneficial in speeding up training for stationary systems and in increasing network adaptability in non-stationary systems. On receiving a new datum, our estimate of the weights is obtained by maximising the posterior

$$p(\boldsymbol{w}_t|Y_t) = \frac{p(\boldsymbol{y}_t|\boldsymbol{w}_t)p(\boldsymbol{w}_t|Y_{t-1})}{p(\boldsymbol{y}_t|Y_{t-1})} \quad (2)$$

where the likelihood $p(\boldsymbol{y}_t|\boldsymbol{w}_t)$ contains the information from the datum at time t and the prior $p(\boldsymbol{w}_t|Y_{t-1})$ contains information gathered from previous data Y_{t-1}. The prior is obtained from the posterior at the previous timestep by marginalising the joint distribution $p(\boldsymbol{w}_t\ \boldsymbol{w}_{t-1}|Y_{t-1})$ thus

$$p(\boldsymbol{w}_t|Y_{t-1}) = \int d\boldsymbol{w}_{t-1}\ p(\boldsymbol{w}_t|\boldsymbol{w}_{t-1})p(\boldsymbol{w}_{t-1}|Y_{t-1}) \quad (3)$$

We now approximate the above distributions as Gaussians. in the following, the notation $\hat{\boldsymbol{w}}_{t|t-1}$ signifies our estimate of the weights $\boldsymbol{w}_t$ given the information available up to time $t-1$. A similar notation pertains to estimates of the covariance P_t. The system and measurement noise processes are taken to be white, with covariances Q_t and R_t respectively.

In short

$$\begin{aligned} p(\boldsymbol{w}_t|\boldsymbol{w}_{t-1}) &= N(\hat{\boldsymbol{w}}_{t-1|t-1}, Q_{t-1}) \\ p(\boldsymbol{w}_t|Y_t) &= N(\hat{\boldsymbol{w}}_{t|t}, \hat{P}_{t|t}) \\ p(\boldsymbol{y}_t|\boldsymbol{w}_t) &= N(\boldsymbol{f}_t(\boldsymbol{w}_t), R_t) \\ p(\boldsymbol{w}_t|Y_{t-1}) &= N(\hat{\boldsymbol{w}}_{t|t-1}, \hat{P}_{t|t-1}) \\ p(\boldsymbol{y}_t|Y_{t-1}) &= N(\boldsymbol{f}_t(\hat{\boldsymbol{w}}_{t|t-1}), S_t) \quad (4) \end{aligned}$$

The last line above reveals the evidence to be a distribution of the prediction errors $\boldsymbol{e}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}$ where our one-step-ahead prediction is given by $\hat{\boldsymbol{y}}_{t|t-1} = \boldsymbol{f}_t(\hat{\boldsymbol{w}}_{t|t-1})$. The covariance S_t gives us an estimate of the Bayesian error-bars for our predictions. If we now linearise the network output $\boldsymbol{f}$ about our current estimate of the weights $\hat{\boldsymbol{w}}_{t|t-1}$ thus,

$$\begin{aligned} \boldsymbol{f}_t(\boldsymbol{w}_t) &= \boldsymbol{f}_t(\hat{\boldsymbol{w}}_{t|t-1}) + \\ &\quad \boldsymbol{f}'_t(\hat{\boldsymbol{w}}_{t|t-1})(\boldsymbol{w}_t - \hat{\boldsymbol{w}}_{t|t-1}) \quad (5) \end{aligned}$$

we obtain the (first order) extended Kalman filter algorithm as an approximate MAP solution

$$\begin{aligned} \hat{\boldsymbol{w}}_{t|t-1} &= \hat{\boldsymbol{w}}_{t-1|t-1} \\ \hat{P}_{t|t-1} &= \hat{P}_{t-1|t-1} + Q_{t-1} \\ \hat{\boldsymbol{y}}_{t|t-1} &= \boldsymbol{f}_t(\hat{\boldsymbol{w}}_{t|t-1}) \\ \boldsymbol{e}_t &= \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1} \\ S_t &= R_t + \boldsymbol{f}'_t(\hat{\boldsymbol{w}}_{t|t-1})\hat{P}_{t|t-1}\boldsymbol{f}'_t(\hat{\boldsymbol{w}}_{t|t-1})^T \\ K_t &= \hat{P}_{t|t-1}\boldsymbol{f}'_t(\hat{\boldsymbol{w}}_{t|t-1})^T S_t^{-1} \\ \hat{\boldsymbol{w}}_{t|t} &= \hat{\boldsymbol{w}}_{t|t-1} + K_t\boldsymbol{e}_t \\ \hat{P}_{t|t} &= (I - K_t\boldsymbol{f}'_t(\hat{\boldsymbol{w}}_{t|t-1}))\hat{P}_{t|t-1} \quad (6) \end{aligned}$$

The matrix K_t is known as the Kalman gain.

THE ORIGINAL NOVELTY CRITERION

Platt's novelty criterion is formulated for a single output for simplicity and consists of two distinct parts. Firstly, the prediction error e_t at time t is compared to a critical value e^c : if $e_t < e^c$, then the existing network is taken to be performing adequately and no unit is added. If $e_t > e^c$, then the network may be attempting to extrapolate, and so the Euclidean distance d_t from the input vector to the nearest centre is compared to a critical value d^c, with a new unit being added only if this critical value is exceeded. The critical distance is allowed to decrease exponentially with time in order to inhibit excessive addition of units early in the training while permitting unit addition at a later stage if neccessary.

In summary, for a new unit to be added we must have

$$\begin{aligned} e_t &> e^c \text{ and } d_t^{\text{nearest unit}} > d_t^c \text{ , where} \\ d_t^c &= \max(\gamma^t d_{\max}^c, d_{\min}^c) \text{ , } \gamma \in (0,1) \quad (7) \end{aligned}$$

INCNET

The obvious drawback of the above criterion lies in the arbitrariness of the choice of critical parameters $e^c, d_{\max}^c, d_{\min}^c$ and γ. (If we have more than one output, then there will be a critical error for each.) Minor changes to these parameters can lead to major changes in network growth, in many cases with no significant improvement in performance [4, 5, 6].

With this in mind, Kadirkamanathan introduced IncNet [2]. This RAN trains using the extended Kalman filter and has a novelty criterion which tests the compatibility of the prediction error, e_t (again dealing with one output for simplicity) and the Kalman filter's online estimation of its variance S_t, as shown in equation (6). The statistic $e_t/\sqrt{S_t}$ follows a t distribution, and a new unit is added only if the prediction error exceeds the 95% confidence limit at that timestep. This removes the need for critical parameters, and has been shown to lead to the construction of more compact networks than the original RAN *for stationary data sets.* It can be extended to multiple outputs by testing each component of $\boldsymbol{e}$ against the diagonal elements of S.

A NEW NOVELTY CRITERION

An undesirable feature of both the above novelty criteria is that, by basing the decision on the instantaneous error, the RAN becomes sensitive to outliers. And as the new elements in the Kalman filter's weight covariance matrix $\hat{P}$ require initialisation after unit addition, the filter should be given time to compensate before further assessments of performance are made. The above criteria do not take this into account, and hence on ocassion many unnecessary units are added in consecutive timesteps. A further problem with IncNet comes when RAN networks are applied to non-stationary problems. In order to prevent the Kalman filter from converging, and thus allow the network to be more adaptable without having always to rely on unit addition, zero mean Gaussian system noise $\boldsymbol{\mu}$ with covariance Q is added to widen the prior in equation (6). As this artificially extends the error

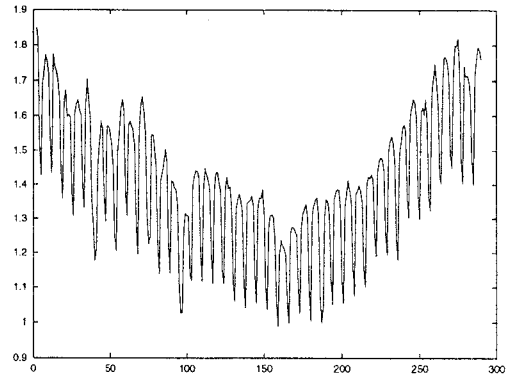

Figure 1: Electricity load.

bars obtained from the matrix S, the IncNet criterion will only be satisfied when the prediction error is large. This leads to the construction of inadequate networks whose prediction performance is unacceptably poor, and can even stop unit addition altogether.

We therefore propose a new novelty criterion with the following rationale. If the model at time t produces an adequate fit to the underlying data generator, then the subsequent output sequence should be distributed as a white noise process. As we do not have access to an ensemble at time t, the best we can do is to estimate the statistics using the outputs at previous timesteps. As we may be dealing with non-stationary systems, then the size of the window we use must be less that the characteristic timescale of the non-stationarity so that we can use this ergodic approximation. Other constraints on the window size are that it must be large enough to allow meaningful sample estimates to be obtained, but not so large that the network becomes slow to react to the non-stationarity.

Our novelty criterion therefore involves testing for zero mean and whiteness at the 95% confidence level. As the network requires time to re-adjust after the addition of a new unit, we do not permit further unit addition for N timesteps after a previous increment. This gives us time to aggregate prediction error information over a window which does not contain error information obtained from the previous architecture, and also allows the filter time to adapt to a potentially unsuitable initialisation of the new components of the weight covariance matrix.

To test for a zero mean sequence we simply construct the t statistic

$$< \hat{e} >_t \sqrt{N/\hat{S}_t} \qquad (8)$$

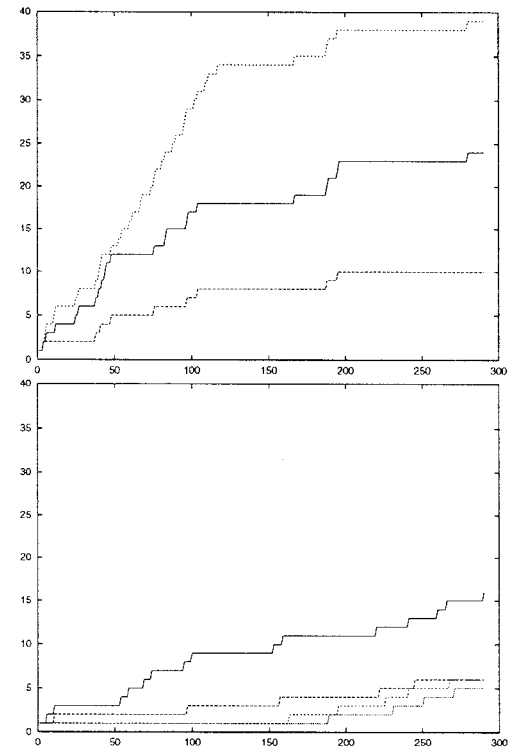

Figure 2: Electricity : growth of RAN using original criterion with varying critical parameters (top), growth of RAN using new criterion with varying window lengths (bottom).

where N is the window length over which we have obtained the sample estimates $\hat{S}$ which replace the artificially noise-corrupted filter estimate S_t. (This can be viewed as a generalisation of the IncNet criterion.) The whiteness test can be performed with the *weighted-sum-squared-residual* (WSSR) statistic [1] which integrates information across all outputs

$$\sum_{k=t-N+1}^{t} e_k^T \hat{S}_k^{-1} e_k \qquad (9)$$

For a network with p outputs, this statistic is distributed as $\chi^2(Np)$.

ELECTRICITY LOAD FORECASTING

This data set represents averaged daily electricity load demand (Figure 1). As there is a weekly cycle superimposed on the seasonality, a window of the previous seven days' load was used as input during training.

Figure 2 and Table 1 illustrate the effects of altering just two critical parameters in the original criterion and of altering the window

$d^c_{\max}$	e^c	Network Size	NPE
0.5	0.1	24	0.376
0.25	0.1	39	0.395
0.5	0.2	10	0.370

window size	Network Size	NPE
5	16	0.391
10	6	0.383
15	6	0.402
20	5	0.400

Table 1: Comparison of old and new criteria for load data. Naive predictor gives NPE = 0.668

size in the new criterion. The sensitivity of the original RAN to the settings of the critical parameters can clearly be seen in Figure 2, as can the bursts of unnecessary unit addition discussed previously.

The new criterion leads to the construction of smaller networks, some of whose performances are comparable to the much larger nets constructed via the original RAN criterion. Table 1 also demonstrates the drawback of calculating the statistics over too large a window : the network becomes slow to react slow to react to the non-stationarity and and thus performance is degraded. However, it is easier to obtain an appropriate value of a single window length parameter than it is to optimise the various hyperparameters of the original criterion.

EXCHANGE RATE FORECASTING

The data in Figure 3 shows daily prices for the Deutsche Mark/French Franc market. As there can be 'day of week' effects in such data, a window of 5 previous values was used as input.

Figure 4 and Table 2 reinforce our conclusions from the load forecasting case. The new procedure consistently gives smaller networks of comparative power, and shows considerably less variability in the size of networks constructed.

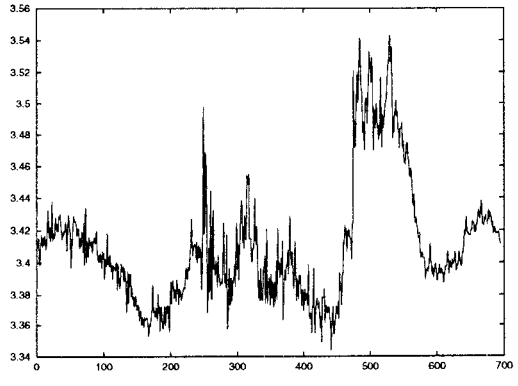

Figure 3: *DM/Fr exchange rate.*

CONCLUSIONS

The original novelty criterion for resource allocating networks is a hyperparameter-intensive algorithm with little statistical motivation which has been demonstrated to give highly variable performance. While the more statistically sound IncNet algorithm is free of hyperparameters, its dependence on Kalman filter estimates of the statistics of the residuals reduces its effectiveness when the filter is modified for application to non-stationary problems. (In both of the above experiments, the addition of system noise was sufficient to effectively disable the IncNet procedure, resulting in the production of extremely small, inadequate networks.)

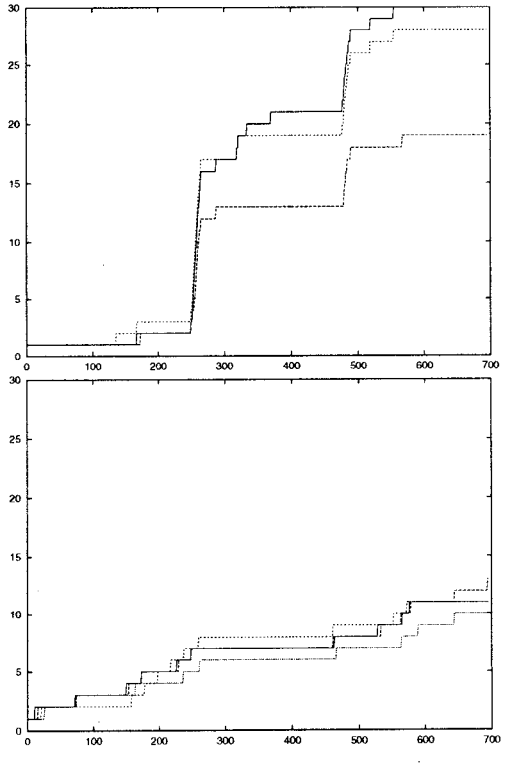

Figure 4: *Exchange rate : growth of RAN using original criterion with varying critical parameters (top), growth of RAN using new criterion with varying window lengths (bottom).*

$d^c_{\max}$	e^c	Network Size	NPE
0.5	0.0075	30	0.332
0.25	0.0075	28	0.332
0.5	0.015	19	0.333

window size	Network Size	NPE
10	11	0.336
15	13	0.334
20	11	0.332
25	10	0.329

Table 2: Comparison of old and new criteria for exchange rate data. Naive predictor gives NPE = 0.357

Both of these criteria suffer from the drawback of being sensitive to noise through their dependence on the instantaneous error at each timestep.

In this paper, we have introduced a new criterion derived from standard signal processing theory which

- reduces the effects of noise by using statistics gathered via a travelling window over the residuals instead of the instantaneous error,
- uses sample estimates of these statistics to remove dependence on parameters of the Kalman filter, and
- contains only a single hyperparameter - the window width.

We have shown that this new criterion can produce much smaller networks than those obtained previously, but with comparable predicitive power.

ACKNOWLEDGEMENTS

This work was supported by EPSRC grant GR/J75425.I would like to thank Midland Electricity plc, UK and Pareto Partners, London, UK for the provision of the data, and David Lowe for helpful comments.

REFERENCES

[1] John V Candy. *Signal Processing : The Model-Based Approach.* McGraw-Hill, 1986.

[2] Visakan Kadirkamanathan. A statistical inference based growth criterion for the RBF network. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, volume IV, pages 12–21, 1994.

[3] Visakan Kadirkamanathan and Mahesan Niranjan. A function estimation approach to sequential learning with neural networks. *Neural Computation*, 5:954–975, 1993.

[4] David Lowe and Alan McLachlan. Modelling of nonstationary processes using radial basis function networks. In *Fourth IEE International Conference on Artificial Neural Networks*, pages 300–305. IEE Conference Proceedings No. 409, 1995.

[5] Alan McLachlan and David Lowe. Tracking of nonstationary time series using resource allocating RBF networks. In R Trappl, editor, *Cybernetics and Systems '96*, pages 1066–1071. Austrian Society for Cybernetic Studies, 1996.

[6] Ian T Nabney, Alan McLachlan, and David Lowe. Practical methods of tracking nonstationary time series applied to real world data (invited talk). In S K Rogers and D W Ruck, editors, *AeroSense '96 - Applications and Science of Artificial Neural Networks II*, pages 152–163. SPIE Publications Vol. 2760, 1996.

[7] John C Platt. A resource allocating network for function interpolation. *Neural Computation*, 3:213–225, 1991.