A Statistical Input Pruning Method for Artificial Neural Networks Used in Environmental Modelling

G. B. Kingston, H. R. Maier and M. F. Lambert
Centre for Applied Modelling in Water Engineering, School of Civil and Environmental Engineering, University of Adelaide, Adelaide, SA, 5005, Australia. Email: gkingsto@civeng.adelaide.edu.au

Abstract: Artificial neural networks (ANNs) provide a useful and effective tool for modelling poorly understood and complex processes, such as those that occur in nature. However, developing an ANN to properly model the desired relationship is not a trivial task. Selection of the correct causal inputs is one of the most important tasks faced by neural network practitioners, but as knowledge regarding the relationships modelled by ANNs is generally limited, selecting the appropriate inputs is also one of the most difficult tasks in the development of an ANN. Many of the methods available for assessing the significance of potential input variables do not consider the uncertainty or variability associated with the input relevance measures used and, consequently, this important factor is neglected during hypothesis testing. In this paper a model-based method is presented for pruning ANN inputs, based on the statistical significance of the relationship between the input variables and the response variable. The approach uses Bayesian methods to estimate the input relevance measure such that the uncertainty associated with this parameter can be quantified and hypothesis testing can be carried out in a straightforward and statistical manner. The proposed methodology is applied to a synthetically generated data set and it is found to successfully identify the 3 relevant inputs that were used to generate the data from 15 possible input variables that were originally entered into the ANN.

Keywords: Artificial neural networks; Input selection; Pruning; Bayesian; Environmental modelling

1. INTRODUCTION

Artificial neural networks (ANNs) provide a useful and effective tool for modelling the complex and poorly understood processes that occur in nature, as they are able to extract functional relationships between model inputs and outputs from data without requiring explicit consideration of the actual data generating process. However, in order to achieve a good representation of the data-generating relationship, an ANN needs to contain all information relevant to the problem. Therefore, selection of the correct causal inputs is one of the most important tasks faced by neural network practitioners.

Knowledge about exact environmental relationships is generally lacking and, consequently, it is difficult to select the correct set of inputs that are relevant to the process. Often, little consideration is given to this task as it has been assumed that, because ANNs are a data driven approach, the relevant inputs will be determined automatically during the modelling process (Maier and Dandy 2000). However, the number of potential inputs can be large for complex environmental systems, particularly when the process under study is dynamic and requires the inclusion of time-lagged input variables. Presenting all potential inputs to an ANN increases the size and complexity of the network, which slows training and increases the amount of data required to estimate the free parameters, or weights, of the network. Moreover, the inclusion of irrelevant inputs can confuse the training process, resulting in spurious correlations in the data being modelled, rather than the actual underlying process.

To help ensure that a good representation of the underlying process is obtained, it is necessary to consider methods for assessing the statistical significance of potential inputs. This is particularly important when the model is used to acquire knowledge about the system, rather than being used solely for predictive purposes. In this paper a model-based method is presented for pruning ANN inputs, based on the statistical significance of the relationship between the inputs and the response variable. This approach uses Bayesian methods to estimate the input relevance measure such that the uncertainty associated with

this parameter can be quantified and hypothesis testing can be carried out in a straightforward manner. The method is applied to a synthetically generated data set in order to demonstrate its application.

2. BACKGROUND

2.1 Input Significance Testing

According to Refenes and Zapranis (1999), determining the significance of a potential ANN input involves the 3 following stages:

1. Defining the relevance of the input to the model.
2. Defining the variance of the relevance measure.
3. Testing the hypothesis that the input is irrelevant to the model.

There have been a number of methods proposed in the literature for addressing the first stage of this problem. These include sensitivity analyses (Lek et al. 1996), assessing the weights of the trained network (Garson 1991), and stepwise methods where the importance of an input is determined by the change in predictive error when it is added to or subtracted from the network (Maier et al. 1998). Although these methods provide a means of determining the overall influence of a potential input, they are generally based on the single-valued weights of a trained ANN and, therefore, do not facilitate the further two stages of the problem. Consequently, inputs are included or excluded from the model in a subjective manner, depending on their effect on the output or model error, as there is no way to statistically test their significance.

Olden and Jackson (2002) introduced a randomization method for statistically assessing the importance of an input based on the comparison of the input's overall connection weight (*OCW*) with a statistical measure of irrelevance. The overall connection weight of an input is the sum of the products of the weights between an input and the output. With reference to Figure 1, the *OCW* of input 1 can be calculated by determining $c_{A,1}$ and $c_{B,1}$, which are the contributions of input 1 via hidden nodes A and B, respectively, and summing them to obtain OCW_1 as follows:

$$\begin{aligned} c_{A,1} &= w_{A,1} \times w_{O,A} \\ c_{B,1} &= w_{B,1} \times w_{O,B} \\ OCW_1 &= c_{A,1} + c_{B,1} \end{aligned} \qquad (1)$$

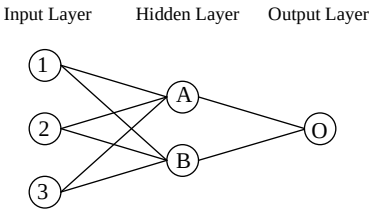

Figure 1. Example ANN structure

Under this paradigm the statistical measure of irrelevance is determined by removing any functional structure between the model inputs and outputs and using a bootstrap procedure to obtain a probability density function (pdf) of the input's *OCW* when there was no remaining relationship between it and the output. Inputs were considered irrelevant if the original *OCW* of the input was not significantly different from the *OCW* when the relationship had been removed.

Although the method of Olden and Jackson (2002) addresses each of the 3 stages of input significance testing, its success is reliant on finding a single set of optimal weights that correctly approximate the underlying function. Due to complications during training and the inherent variability of the underlying process itself, it is unlikely that a single optimal weight vector will be found, particularly when irrelevant inputs are included in the model. It is therefore important to consider a distribution of the network weights such that the uncertainty associated with finding an optimal weight vector can be incorporated into the input significance tests. By describing the weights as distributions a range of possible weight values is considered, preventing one, possibly incorrect weight vector, from completely dominating the calculated *OCW*s, which are fundamental in testing the relevance of the inputs.

2.2 Bayesian Weight Estimation

Bayesian methodology was first applied to estimate the weights of an ANN by Mackay (1992) and Neal (1992). It provides an approach for explicitly handling uncertainty in the weights by considering the weight vector, **w**, as a random variable. Using Bayes' Theorem, the posterior weight distribution, P(**w**|**y**,**x**), may be inferred from the data as follows:

$$\begin{aligned} \mathrm{P}(\mathbf{w}|\mathbf{y},\mathbf{x}) &= \frac{\mathrm{P}(\mathbf{y}|\mathbf{x},\mathbf{w})\mathrm{P}(\mathbf{w}|\mathbf{x})}{\mathrm{P}(\mathbf{y}|\mathbf{x})} \quad \text{or} \\ \mathrm{P}(\mathbf{w}|\mathbf{y},\mathbf{x}) &\propto \mathrm{P}(\mathbf{y}|\mathbf{x},\mathbf{w})\mathrm{P}(\mathbf{w}|\mathbf{x}) \end{aligned} \qquad (2)$$

where **w** is a vector of ANN weights, **y** is a vector of N observations and **x** is a set of N input vectors. In (2), P(**w**|**x**) is the prior weight distribution, which describes any knowledge of **w** before the

data were observed. $P(\mathbf{y}|\mathbf{x},\mathbf{w})$ is known as the likelihood function. This function uses information obtained by comparing the model predictions to the observed data to update the prior knowledge of $\mathbf{w}$. By assuming that each observation is independently drawn from a Gaussian distribution, the likelihood function can be described by:

$$P(\mathbf{y}|\mathbf{x},\mathbf{w}) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i - f(\mathbf{x}_i,\mathbf{w})}{\sigma}\right)^2\right) \quad (3)$$

where $f(\mathbf{x}_i,\mathbf{w})$ is the ANN output for the ith input vector and σ is the standard deviation of the model residuals.

2.2.1 The Metropolis Algorithm

The high dimensionality of the conditional probabilities in (2) makes it difficult to calculate the posterior weight distribution analytically. Consequently, methods have been introduced to approximate (2). Neal (1992) introduced a Markov chain Monte Carlo (MCMC) implementation to sample from the posterior weight distribution such that $P(\mathbf{w}|\mathbf{y},\mathbf{x})$ could be evaluated numerically.

The Metropolis algorithm is a commonly used MCMC approach, which generates samples from the posterior distribution of an unknown variable, e.g. ANN weights. As it is difficult to sample from the complex posterior distribution directly, this method uses a simpler, symmetrical distribution (a multinormal distribution was used in this study), known as the proposal distribution, to generate candidate weight vectors. By employing an adaptive acceptance-rejection criterion the random walk sequence of weight vectors converges towards the posterior distribution over many iterations. Details of the computational implementation of the Metropolis algorithm can be found in Thyer et al. (2002).

The covariance of the proposal distribution has important implications on the convergence properties and efficiency of the Metropolis algorithm. Poor selection of this parameter may result in an insufficient number of generated samples to adequately represent the posterior distribution. Haario et al. (2001) introduced a variation of the Metropolis algorithm that was developed to increase its convergence rate. In this algorithm the proposal distribution continually adapts to the posterior distribution by updating the covariance at each iteration based on all previous states of the weight vector. The adaptation strategy ensures that information about the posterior distribution, accumulated from the beginning of the simulation, is used to increase the efficiency of the algorithm. This algorithm is known as the adaptive Metropolis algorithm.

3. METHODS

The proposed input selection method is a model-based pruning approach, where the initial ANN includes all potential inputs and "irrelevant" inputs are eliminated, or pruned, from the network throughout the process. The method addresses the 3 stages of input significance testing in a systematic and consistent manner by using the Bayesian framework to estimate distributions of the network weights.

The overall connection weight (*OCW*) measure, used by Olden and Jackson (2002), is employed to quantify the input variables' relevance to the model. The *OCW* of an input measures the strength and direction of the relationship between that input and the output. If this measure is approximately equal to zero there is no relationship between the input and the response variable.

The adaptive Metropolis algorithm is used to generate samples from the posterior weight distribution. The corresponding *OCW* values are then calculated for each sampled weight vector, producing empirical distributions of the *OCW*s, which capture the variation in these relevance measures. In this study, a uniform prior distribution over the range [-3,3] was assumed for each weight. After a warm-up period of 30,000 iterations, 100,000 weight vectors were sampled from the posterior weight distribution and the corresponding *OCW*s calculated.

By having distributions of the input relevance measures, Bayesian probability intervals can be formed in order to test the hypothesis that an input is irrelevant to the model (*OCW*=0). The probability intervals are initially formed around the mode of the pdf such that 100(1-*a*)% of the distribution is contained within the interval, where *a* is the significance level. If zero lies within these bounds, the hypothesis that the input is irrelevant to the model is true.

The weights of an ANN are generally small values centred around zero, thus it is likely that the *OCW*s are also relatively close to zero. Moreover, it is expected that the initial *OCW* distributions will be quite variable due to the inclusion of irrelevant inputs in the model. Therefore, it is likely that a number of the initial *OCW* distributions will contain zero regardless of whether the input is

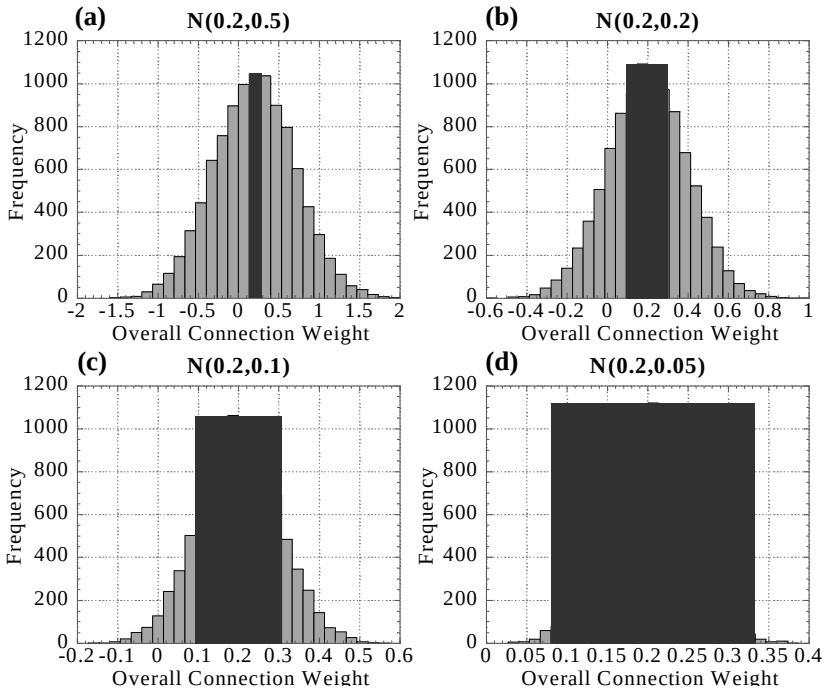

Figure 2 Increasing width of probability intervals (shaded region) at different stages of the pruning process. The standard deviation of the *OCW* reduces from 0.5 in (a) to 0.05 in (d) when the relationship between the input and the output becomes well defined. Eventually the input is tested for its dissimilarity to 0.

relevant to the model or not. To ensure that important inputs are not pruned in these initial stages when the relationship is poorly defined, inputs are only pruned from the network when their *OCW*s are statistically similar to zero at a high significance level (e.g. 95%). As the process continues and irrelevant inputs are pruned from the network, the relationship between inputs and outputs becomes better defined and the variance in the *OCW*s reduces. This means that the pdf of a significant input's *OCW* is less likely to contain zero as more irrelevant inputs are pruned. Therefore, the significance level at which inputs are tested for their similarity to zero may be reduced gradually throughout the process. Eventually inputs are tested for their statistical dissimilarity to zero (*OCW*=0 at 5% $\Leftrightarrow$ *OCW*$\neq$0 at 95%), which ensures that only inputs having a significant relationship with the output are included in the model. This process is illustrated in Figure 2 where the pdf of an *OCW* is Gaussian with a mean of 0.2. The standard deviation of the distribution decreases from 0.5 in Figure 2 (a) to 0.05 in Figure 2 (d) as indicated by the scale on the x-axis. As the variance decreases it becomes more evident that the *OCW* is significantly different from zero. Even though the probability intervals are eventually widened to include 95% of the distribution, zero is never included within this range, indicating that the input is statistically significant. However, if the bounds were set wider when the variance was large this input would have been considered irrelevant.

The following process is carried out until all inputs remaining in the model are statistically significant:

1. Sample 100,000 weight vectors from the posterior weight distribution and calculate the corresponding *OCW*s for each input, forming empirical distributions of the *OCW*s.
2. Test the hypothesis that the inputs are irrelevant (beginning at the 95% significance level) by constructing probability intervals around the mode *OCW* value for each input. If zero is included within these intervals the input is considered irrelevant and is pruned from the network. If no inputs are irrelevant at the current significance level, widen the bounds to include a greater proportion of the distribution (e.g. decrease the significance level by 5-10%).
3. Repeat steps 1 and 2 until the only remaining inputs have *OCW*s that are statistically different from zero at a high significance level (e.g. 95%) or, in other words, that the *OCW* is equal to zero at a low significance level (e.g. 5%).

4. CASE STUDY

4.1 Data

Autoregressive (AR) models are commonly used to model natural systems (e.g. hydrological time series data). The autoregressive model, AR(9),

was used to generate a set of synthetic time series data according to:

$$x_t = 0.3x_{t-1} - 0.6x_{t-4} - 0.5x_{t-9} + \varepsilon_t \quad (4)$$

where ε_t is a random noise component with distribution N~(0,1). This model was selected for demonstrating the proposed input selection method as it depends on more than one input variable and has known dependence attributes. Moreover, the use of synthetic data enabled the generation of as much data as was required. 400 data points were generated as this number was considered to represent a realistic data set size for environmental data, which are generally limiting.

4.2 ANN Model

Although the response variable x_t only depends on inputs x_{t-1}, x_{t-4} and x_{t-9}, 15 inputs from x_{t-1} to x_{t-15} were included in the ANN in order to determine whether the proposed input selection method could identify the 12 irrelevant inputs that needed to be pruned from the model. An ANN with 1 hidden layer with 2 hidden layer nodes was used to model the data. It should be noted that due to the large number of inputs included in the model, and thus the large number of free parameters, it is likely that the model would overfit to noise in the data in the initial stages of the pruning process. This amplifies the need to only prune those inputs that have *OCW*s statistically similar to zero at a high significance level in the initial stages. Initially, testing the hypothesis of input irrelevance began at the 95% significance level (i.e *OCW*=0 with 95% probability). However, there were no irrelevant inputs at this level and the significance was decreased in increments of 5% until there were one or more irrelevant inputs. This occurred at the 85% significance level.

5. RESULTS & DISCUSSION

The results of the input selection process are given in Table 1. The final inputs remaining in the model (relevant at the 95% significance level) were x_{t-1}, x_{t-4} and x_{t-9} which are the correct causal inputs for the AR(9) data. Therefore, the proposed method was able to properly identify the irrelevant inputs such that they could be pruned from the ANN. It can be seen in Table 1 that 7 runs were required to achieve the final model.

Plots of the *OCW* distributions of inputs x_{t-3} and x_{t-9} are shown in Figure 3. Figures 3 (a) and (b) show the *OCW* distributions of x_{t-3} after run 1 and run 6 respectively, while Figures 3 (c) and (d) give the same plots for x_{t-9}. It can be seen that the variances

Table 1 Results of the pruning process. The remaining inputs were x_{t-1}, x_{t-4} and x_{t-9}.

Run no.	No. of initial inputs	Significance level[a]	Irrelevant inputs
1	15	85%	x_{t-6}, x_{t-13}, x_{t-14}
2	12	80%	x_{t-8}
3	11	80%	x_{t-7}
4	10	75%	x_{t-5}
5	9	5%	x_{t-2}, x_{t-10}, x_{t-11}, x_{t-12}, x_{t-15}
6	4	5%	x_{t-3}
7	3	5%	-

[a]with which the *OCW*s of the pruned inputs were similar to 0

of the *OCW* distributions are quite large when 15 inputs were included in the model (Figures (a) and (c)). Additionally, it appears that x_{t-3} is significant to the model at this stage, which indicates that the underlying relationship has been incorrectly approximated due to the inclusion of irrelevant inputs. This demonstrates that an ANN will not necessarily determine which inputs are relevant to the output automatically and highlights the need for analytical methods for this purpose. When the model contained only 4 inputs the relationships between inputs and outputs became better defined as indicated by the reduced spread of the distributions in Figures 3 (b) and (d). Here it has been correctly identified that x_{t-3} is irrelevant to the model and x_{t-9} is relevant.

6. CONCLUSIONS

Selection of the correct causal inputs is vital for ensuring that an ANN model gives a good representation of the underlying function. This is particularly important when the model is used to gain knowledge of the system and an interpretation of the network function is required.

A number of methods have been proposed in the literature for assessing the relevance of potential input variables in predicting the response variable, but few have considered the variability in the relevance measure or the uncertainty in the network weights, both of which are fundamental for assessing input significance. In this paper an input pruning method has been presented which considers both of these factors by using Bayesian methods to estimate the network weights. When the method was applied to a synthetically generated data set it was able to correctly identify the 12 irrelevant input variables that were initially included in the ANN such that these were pruned

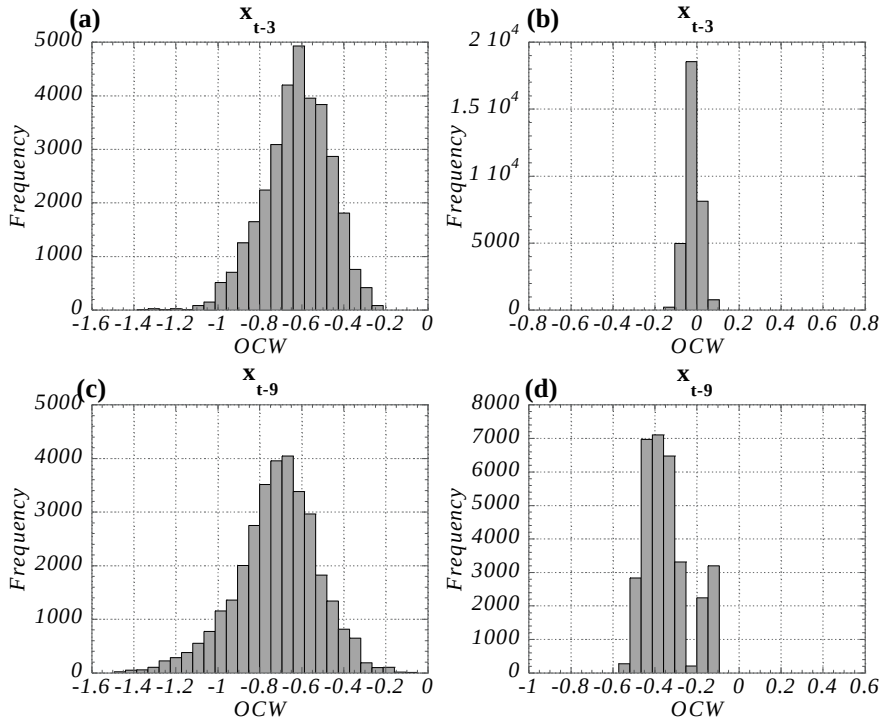

Figure 3 OCW distributions of inputs x_{t-3} and x_{t-9}. (a) and (c) are the distributions obtained after 1 run of the pruning process (15 inputs included), while (b) and (d) are the distributions obtained after 6 runs (4 inputs included)

and the final model only included the 3 correct causal inputs.

A limitation of the proposed method is that the network architecture needs to be specified and this may have implications on the relationship modelled. Future research will consider a method for pruning inputs and hidden nodes concurrently.

7. REFERENCES

Garson, G. D. 1991. Interpreting neural network connection weights. AI Expert **6**:47-51.

Haario, H., E. Saksman, and J. Tamminen. 2001. An adaptive Metropolis algorithm. Bernoulli **7**:223-242.

Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, and S. Aulagnier. 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling **90**:39-52.

Mackay, D. J. C. 1992. A practical Bayesian framework for backpropagation networks. Neural Computation **4**:448-472.

Maier, H. R., and G. C. Dandy. 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software 15:101-124.

Maier, H. R., G. C. Dandy, and M. D. Burch. 1998. Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. Ecological Modelling **105**:257-272.

Neal, R. M. 1992. Bayesian Training of Backpropagation Networks by the Hybrid Monte Carlo Method. Department of Computer Science, University of Toronto.

Olden, J. D., and D. A. Jackson. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling **154**:135-150.

Refenes, A.-P. N., and A. D. Zapranis. 1999. Neural model identification, variable selection and model adequacy. Journal of Forecasting **18**:299-332.

Thyer, M., G. Kuczera, and Q. J. Wang. 2002. Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation. Journal of Hydrology **265**:246-257.