Neural networks for water systems analysis: from fundamentals to complex pattern recognition

SANDHYA SAMARASINGHE

Centre for Advanced Computational Solutions (C-fACS), Department of Environmental Management, Lincoln University, Christchurch, New Zealand

sandhya.samarasinghe@lincoln.ac.nz

Abstract Accurate river flows are crucial for effective water resource management. However, estimating flows in ungauged rivers, particularly those in difficult to access terrains, is a challenging problem for water scientists and managers. As a solution, hydrological regionalisation (HR) has been proposed to estimate river flows based on proxy-basin, interpolation and regression methods. Recently, neural networks have been shown to produce improved estimates. In this study, HR-based artificial neural networks (ANN) models were developed for estimating monthly flows in ungauged rivers in New Zealand using hydrological and geomorphological attributes. After rigorous input selection, multilayer perceptron (MLP) networks were first developed by trial and error. Then, a new MLP method, not involving trial and error, was developed by clustering the correlated hidden neurons in a trained MLP to simplify the model structure; this produced overall better results than the trial-and-error MLP and a genetic algorithm optimised MLP. Results show that accurate and parsimonious MLP models can be developed for flow estimation based on HR using the new method. Therefore, the study presents the hydrological community with improved neural networks tools based on HR to estimate flows in ungauged rivers for more effective water management.

Key words river flows; hydrological regionalisation; neural networks; network pruning; New Zealand

INTRODUCTION

Hydrological processes are characterised by high complexity, dynamism, and nonlinearity in both spatial and temporal scales. Lack of physical understanding of these processes has hampered the development of efficient models to study their behaviour and manage water resources effectively. The last decade has seen ANN applications in all areas of water resources (Adeli, 2001), mainly due to their ability to nonlinearly relate input and output variables and capture temporal dynamics in complex dynamical systems without needing a detailed understanding of the physics of the processes involved. There is a variety of existing conceptual and mathematical hydrological rainfall–runoff models, but there is always the difficulty in choosing, calibrating and validating parameters; therefore, incremental parameter estimation by ANN from historical data offers an attractive advantage.

The three major types of ANN used in water resources are: Multiple Layer Perceptrons (MLP) – a powerful multivariate nonlinear regressor; Recurrent Neural Networks (RNN) – a nonlinear autoregressive network for time-series forecasting, and Self Organising Maps (SOM) – a nonlinear unsupervised clustering approach that reveals clusters in the data while preserving cluster proximity; and hybridisations of the above network types as well as their newer variants (Samarasinghe, 2006).

When rivers are not gauged, HR assumes that hydrological and geomorphological similarity with nearby basins can be used for estimating their flows (Bormannet *et al.*, 1999) for managing water resources. The simplest approach is the direct transfer of model parameters to ungauged basins from nearby basins using, for example, the proxy-basin method (Xu, 1999), linear interpolation methods (Guo *et al.*, 2001) and kriging interpolation methods (Vandewiele & Elias, 1995). Another approach includes a two-step regression procedure where models relating hydrological variables to flows are developed for each gauged river in the first step, and the model parameters are related to the geomorphological characteristics of basins in the second step to obtain accurate parameters for the ungauged rivers (Tung *et al.*, 1997). Recently, a single step model involving ANN using both hydrological and geomorphological variables together has been shown to be superior to both one step and two-step regression (Cutore *et al.*, 2007). However, there are several modelling issues to be addressed to make ANN more efficient in flow prediction from HR as well as in other applications. One of these issues is the optimisation of hidden layers of ANN.

This paper highlights some current issues in ANN and presents a new network pruning approach based on SOM clustering of redundant MLP neurons for effective HR based flow estimation of an ungauged river in a seven-river basin in New Zealand.

NEURAL NETWORKS

In a neural network, a number of computational neurons are organised in layers in a highly parallel network in such a way that inputs flowing through the network can interact nonlinearly developing complex input-output mapping functions. Figure 1 shows a typical MLP network with input, hidden and output layers connected to each other with weights (model parameters). The output z is a function of inputs and model parameters:

$$z = f(X, a_{ij}, b_j) = f_1\left(\sum_{j=0}^{n} b_j y_j\right) = f_1\left(\sum_{j=0}^{n} b_j f_2 \sum_{i=0}^{m} x_i . a_{ij}\right) \quad (1)$$

where X is the input vector, a_{ij} is the input-hidden layer weights and b_j is the hidden-output neuron weights with a_{0j} and b_{0j} representing the bias weights. A hidden neuron j produces its output y_j by transforming the weighted sum of its inputs through function $f_2(.)$ and the output neuron computes the network output z by transforming the weighted sum of its inputs with $f_1(.)$. Nonlinear functions in the hidden and output neurons give the network nonlinear capability, and the number of hidden neurons provides the ability to represent an arbitrarily complex function to any desired accuracy. The ANN must be trained incrementally by repeatedly processing historical data while adjusting model parameters based on one of several training methods until the network output converges to the target output. An error criterion such as RMSE is used to measure the prediction accuracy.

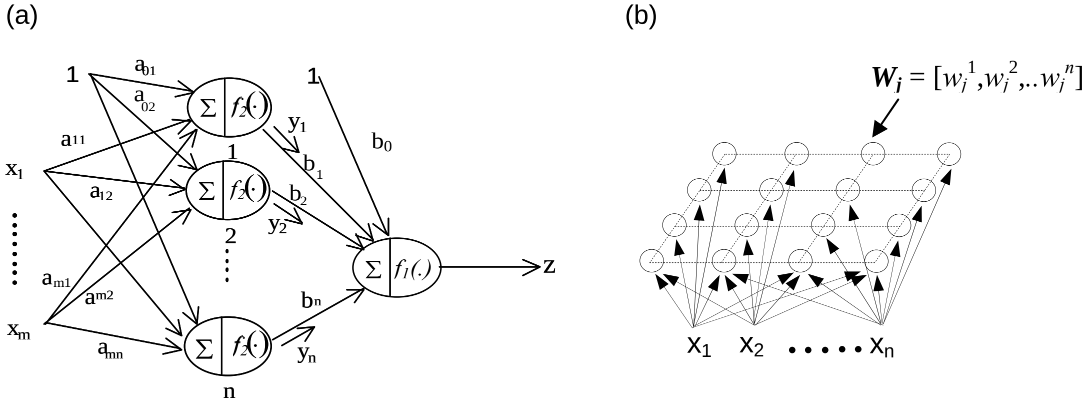

Fig. 1 (a) MLP neural network and (b) Self Organising Map network.

Figure 1(b) shows the structure of SOM that consists of n neurons arranged in a grid. Input vectors ($\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$) are presented to neurons with initially randomized weight vectors ($\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$). First, the winning neuron closest to each input vector is found from the minimum distance d between all weights and the input (equation (2)). Then, the weights of the winner and s neurons in its neighbourhood are adjusted in each iteration t using a neighbourhood function $N(s,t)$ to preserve the proximity of data using a learning rate $\eta(t)$ (equation (2)). Over repeated iterations, where both $\eta(t)$ and $N(s,t)$ functions decrease with time, SOM learns to represent the inputs accurately showing clusters and their spatial relations:

$$d = \min\{\, \| X - W_i^t \| \,\}; \qquad W_i^{t+1} = W_i^t + \eta(t)\, N(s,t)\, d \quad (2)$$

One of the major issues in MLP is finding the simplest model with consistent model parameters for a particular problem. This can be addressed by selecting crucial inputs and optimizing the hidden layer. Use of independent inputs that are the most relevant to the output simplifies the ANN structure and various statistical methods, including simple and partial correlation, principal components analysis, etc., have been used. Optimising the hidden layer is more complex and the current methods, such as optimal brain damage (OBD), require pruning and training of the reduced network in several iterations until the optimum structure is obtained. Another popular method, genetic algorithms (GA), also uses extensive searches in optimizing the hidden layer (Samarasinghe, 2006).

This paper presents a new approach to automatically find the number of neurons in an MLP based on the idea that redundant neurons in a network have correlated activations (Samarasinghe, 2007). An MLP with a relatively large network is developed first and weighted hidden neuron activations $b_j y_j$ are used as input to an SOM trained with correlation distance measure, followed by Ward clustering (Samarasinghe, 2006) to cluster correlated hidden neuron activations (SOMCNA). The number of clusters indicates the required number of neurons in the MLP. The method is applied to estimate flows in an ungauged river in a large basin in New Zealand.

FLOW ESTIMATION OF UNGAUGED RIVERS

Seven river basins in the Canterbury Region in New Zealand were selected: Waipara, Ashley, Halswell, Selwyn, South Ashburton, North Ashburton and Rangitata rivers. Waipara River was selected to depict an ungauged river. In developing an ANN, a series of hydrological and physical concepts can be rationally applied to inform the model of important interactions underlying the dynamic process. For example, rainfall runoff is the major factor affecting river flows. Runoff is affected by drainage area; average ground slope; rainfall interception mainly by forest canopy and land use; evaporation and evapotranspiration affected by soil type; a basin’s shape or form (basin length/width ratio) and drainage density (ratio of the total channel-segment lengths to basin area).

For each basin, monthly average flow and rainfall time series were computed from daily data recorded for periods ranging from 5 to 29 years by Environment Canterbury (Ecan, New Zealand). The following data were obtained from GIS databases: drainage area, average slope, drainage density, basin form, vegetation type and land-use categories, such as urban, pasture, or preserved areas. A new variable, compound factor was introduced as (total area – forested area).

A statistical data pre-processing stage was implemented to clean the data, study relationships and trends, and contrast data from different basins for consistency. Average monthly flows in the seven rivers followed the expected seasonal precipitation patterns in the region. Correlation and partial correlation were performed in three stages, incrementally incorporating several flow and precipitations lags in each stage, to select key inputs that have the highest correlation to flow and least correlation to themselves. The selected final set of three inputs and their partial and simple correlations with the flow are: previous month’s flow (0.3518, 0.845), current precipitation (0.5573, 0.675), and the compound factor (0.3304, 0.858).

ANN model development

Networks (MLP) were developed on Matlab (2009) to predict Waipara River flows (validation set with 63 data points). The rest of the data were divided into training (70%) and calibration (30%) with 1079 and 269 data points, respectively. The Levenberg Marquardt method was used for training with data scaled between 0 and 1, and early-stopping was used to preventing overfitting. The ANN models with sigmoid hidden neuron and linear output functions were developed using: trial and error, SOMCNA and GA; and an MLR model was also developed. The best ANN from each method was re-run with 10 random weight initialisations and R^2, RMSE and mean error were used for assessing their performance. Results are shown in Table 1 and Fig. 3 and discussed below.

Networks were developed in two stages by trial and error: first with the three selected inputs and then sequentially adding secondary variables, as partial correlation only reveals linear relationships. Introduction of basin form thus improved network results. The best network had four inputs (previous flow, current precipitation, basin form and compound factor) and 70 hidden neurons.

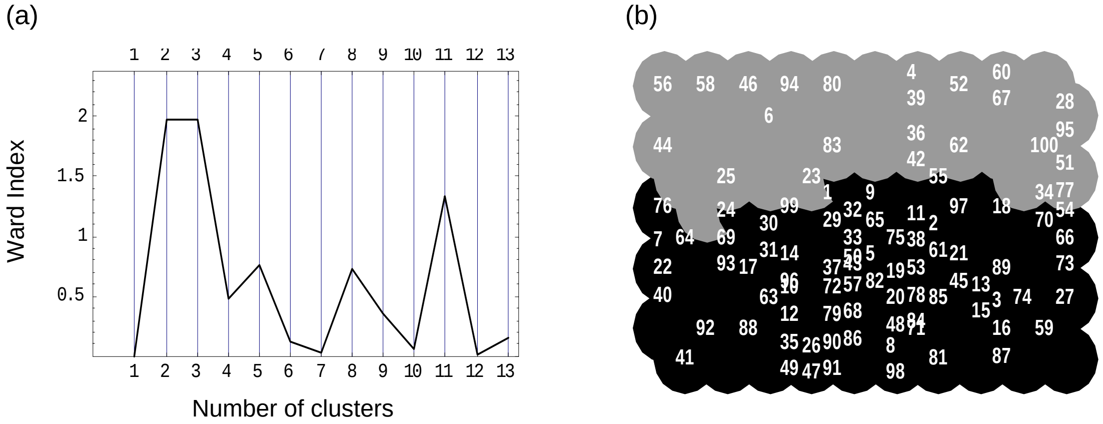

Fig. 2 (a) Ward likelihood index against potential number of clusters, and (b) two optimum clusters found by SOMCNA.

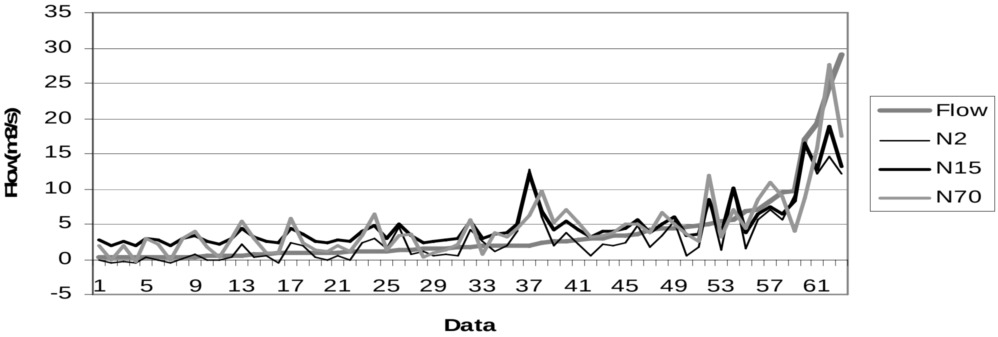

Fig. 3 Estimated flows for validation set from ANN obtained from trial and error, SOMCNA and GA superimposed on actual flows sorted in ascending order.

Table 1 Performance measures of flow estimation models on validation data.

Measure/Method	MLR	SOMCNA(2N)	GA (15N)	70N
R^2	0.15	0.62	0.61	0.67
ME	0.0067	–0.0037	0.002	–0.0028
RMSE	0.0162	0.011	0.011	0.010

ANN structure optimization

In SOMCNA, an MLP with 100 hidden neurons was trained with the four inputs and SOM was applied to cluster the hidden neurons. Results revealed two optimum clusters as shown by the maximum Ward likelihood index in Fig. 2(a) with the resulting two clusters depicted on the trained SOM in Fig. 2(b) where the numbers indicate neurons on the original SOM. GA based optimization using 20 populations and 100 generations for 100 epochs with double point crossover, uniform mutation with probability of mutation of 0.1 and tournament selection (Synapse, 2006) resulted in 15 optimum hidden neurons. Results from trial and error, SOMCNA and GA and MLR indicate that all ANN models have performed better than MLR and SOMCNA has produced the simplest model (Table 1) and better performance overall than the two larger networks (thin line in Fig. 3) indicating the efficacy SOMCNA. As Fig. 3 shows, SOMCNA results, for the most part, more closely follow the actual flow than the predictions from the other models. RMSE is the same across the networks and other parameters are similar with different networks producing the maximum. A sensitivity analysis on the networks revealed the following contributions: precipitation (36%), previous flow (24%), compound factor (22%) and basin form (18%).

SUMMARY AND CONCLUSIONS

Estimating flows in ungauged rivers, particularly those in difficult-to-access terrains, is a challenging problem for water scientists and managers. Hydrological regionalisation (HR) has been proposed to address this issue using several approaches, including proxy-basin, linear interpolation and regression methods (Vandewiele & Elias, 1995; Xu, 1999; Guo *et al.*, 2001). Recently, neural networks have been shown to produce improved estimates (Cutore *et al.*, 2007). In this study, an improved ANN model based on SOM-based clustering of hidden neurons was successfully developed and validated for estimating flows in an ungauged river using HR. The pruned network had a similar performance to that obtained from trial and error and GA optimisation. However, the new method requires training of only one MLP and one application of the pruning method thereby eliminating trial and error in model development and iterative pruning in structure optimisation. Therefore, the study presents the hydrological community with an efficient approach to develop neural networks based on HR to estimate flows in ungauged rivers for more effective water management.

REFERENCES

Bormann, H., Diekkruger, B., Renschler, C. & Richter, O. (1999) Regionalization concept for the prediction of large-scale water fluxes. In: *Regionalization in Hydrology* (ed. by B. Diekkruger, M. J. Kirkby & U. Schroder), (13–22). IAHS Publ. 254. IAHS Press, Wallingford, UK.

Cutore, P., Cristaudo, G., Campisano, A., Modica, C., Cancelliere, A. & Rossi, G. (2007) Regional models for the estimation of stream flow series in ungauged basins. *Water Resour. Manage.* **21**(5), 789–800.

Guo S., Wang, J. & Yang, J. (2001) A semi-distributed hydrological model and its application in a macroscale basin in China. In: *Soil–Vegetation–Atmosphere Transfer Schemes and Large-scale Hydrological Models* (ed. by A. J. Dolman, A. J. Hall, M. L. Kavvas, T. Oki & J. W. Pomeroy), 167–174. IAHS Publ. 270. IAHS Press, Wallingford, UK.

Samarasinghe, S. (2006) *Neural Networks for Applied Sciences and Engineering:* F*rom Fundamentals to Complex Pattern Recognition*. CRC Press, USA.

Samarasinghe, S. (2007) Optimum structure of feed forward neural networks by SOM clustering of neuron activation. In: *Proc. the International Congress on Modelling and Simulation* (ed. by D. Kulasiri & L. Oxley), 2278–2284. Modelling and Simulation Society of Australia and New Zealand. ISBN: 978-0-9758400-4-7.

Synapse v.1.25 (2007) Peltarion HB, Slipgatan 2, 117 39 Stockholm, Sweden.

Tung, Y. K., Yeh, K. C. & Yang, J. C. (1997) Regionalization of unit hydrograph parameters: 1. Comparison of regression analysis techniques. *Stochast. Hydrol. Hydraul.* **11**, 145–171.

Vandewiele, G. L. & Elias, A. (1995) Monthly water-balance of ungauged catchments obtained by geographical regionalization. *J. Hydrol.* **170**(1-4), 277–291.

Xu, C. Y. (1999) Operational testing of a water balance model for predicting climate change impacts. *Agric. For. Met*. **98–99**, 295–304.