Journal of Hydrology 528 (2015) 52–62

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Prediction and structural uncertainty analyses of artificial neural networks using hierarchical Bayesian model averaging

Nima Chitsazan [a,1], Ata Allah Nadiri [b,2], Frank T.-C. Tsai [a,*]

[a] Department of Civil and Environmental Engineering, Louisiana State University, 3418G Patrick F. Taylor Hall, Baton Rouge, LA 70803, USA
[b] Department of Earth Sciences, Faculty of Science, University of Tabriz, 29 Bahman Boulevard, Tabriz, East Azarbaijan 5166616471, Iran

ARTICLE INFO

Article history:
Received 8 September 2014
Received in revised form 2 June 2015
Accepted 3 June 2015
Available online 11 June 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of K.P. Sudheer, Associate Editor

Keywords:
Bayesian model averaging
Artificial neural network
Uncertainty
Ensemble method

SUMMARY

This study adopts a hierarchical Bayesian model averaging (HBMA) method to analyze prediction uncertainty resulted from uncertain components in artificial neural networks (ANNs). The HBMA is an ensemble method for prediction and is used to segregate the sources of model structure uncertainty in ANNs and investigate their variance contributions to total prediction variance. Specific sources of uncertainty considered in ANNs include the uncertainty in neural network weights and biases (model parameters), uncertainty of selecting an activation function for the hidden layer, and uncertainty of selecting a number of hidden layer nodes (model structure). Prediction uncertainties due to uncertain inputs and ANN model parameters are represented by within-model variance. Prediction uncertainties due to uncertain activation function and uncertain number of nodes for the hidden layer are represented by between-model variance. The method is demonstrated through a study that employs ANNs to predict fluoride concentration in the aquifers of the Maku area, Azarbaijan, Iran. The results show that uncertain inputs and ANN model parameters produces the most prediction variance, followed by prediction variances from uncertain number of hidden layer nodes and uncertain activation function.

1. Introduction

Artificial neural networks (ANNs) are mathematical approximators with strong ability to find the non-linear correlation between input and output data without requiring a comprehensive knowledge about the physical system (Zhang et al., 2009). ANNs have been applied to various hydrologic problems (Govindaraju, 2000a,b; Dawson and Wilby, 2001), for example river flood forecasting (Sudheer et al., 2003; Wei et al., 2012), rainfall forecasting (Hsu et al., 1995; Kumar et al., 2005; Valverde Ramírez et al., 2005), evaporation modeling (Sudheer et al., 2002), groundwater level modeling (Coppola et al., 2005; Daliakopoulos et al., 2005; Nourani et al., 2008a,b; Jha and Sahoo, 2015), and groundwater remediation optimization (Rogers and Dowla, 1994). Despite wide applications of ANNs in the field of hydrology, over-parameterization in weights and biases and consequently overfitting to "in-sample" data is a major limitation of ANNs, which may occur in the process of maximizing likelihood functions or alternatively minimizing fitting errors (Kingston et al., 2008). This may reduce prediction accuracy for "out-sample" data that are not used for training ANNs (Tetko et al., 1995; Maier and Dandy, 2000). In other words, overfitting can reduce the "generalization" ability of ANN models. Generalization is the ability of an ANN model to represent an underlying system for an "out-sample" data that are not used in the training procedure. Therefore, it is crucial to select an ANN structure based on the parsimony principle (Box, 1976; Stone, 1981) to balance fitting errors and the number of unknown parameters. To address this issue, Franses and Draisma (1997) proposed the use of information criteria to choose the best ANN structure. Information criteria tend to select a simpler ANN structure unless a more complicated structure provides a better understanding of the underlying system. However, Qi and Zhang (2001) empirically showed that the selected ANN based on information criteria does not necessarily have an acceptable performance for fitting "out-sample" data. They concluded that "to improve generalization performance of neural network models, one may need to go beyond the model selection methods."

An alternative to the model selection approach is the artificial neural network ensemble (Shu and Burn, 2004). In this approach, outputs from several ANN models, which are trained by the same

* Corresponding author. Tel.: +1 225 578 4246; fax: +1 225 578 4945.
E-mail addresses: nchits2@lsu.edu (N. Chitsazan), Nadiri@tabrizu.ac.ir (A.A. Nadiri), ftsai@lsu.edu (F.T.-C. Tsai).
[1] Formerly with Department of Civil and Environmental Engineering, Louisiana State University, 3418G Patrick F. Taylor Hall, Baton Rouge, LA 70803, USA. Tel.: +1 225 578 4246; fax: +1 225 578 4945.
[2] Tel.: +98 04113392731; fax: +98 0411335029.

http://dx.doi.org/10.1016/j.jhydrol.2015.06.007
0022-1694/

data set, are combined to generate an ensemble output. Several methods for combining the outputs of multiple artificial intelligence models are proposed in the literature (Ahmad and Zhang, 2002; Shu and Burn, 2004), such as the committee machine (CM) (Perrone and Cooper, 1993; Chen and Lin, 2006; Kadkhodaie-Ilkhchi et al., 2009; Labani et al., 2010), the supervised committee machine (SCM) (Drucker, 1997; Hu and Tsoukalas, 2003; Nadiri et al., 2013), and the Bayesian model averaging (BMA) (Nadiri et al., 2014). Although the committee machine methods are shown to improve the generalization ability of ANNs (Shu and Burn, 2004), they are subjected to two major shortcomings. First, they are not able to quantify the prediction uncertainty. Second, they do not consider the parsimony principle in evaluating the combination of individual ANN models.

ANN prediction uncertainty stems from two main sources: (1) the uncertainty in ANN inputs, weights and biases and (2) the uncertainty in ANN structures. Bayesian neural networks (BNNs) (Bishop, 1995; Neal, 1995; Müller and Insua, 1998) were used to train ANN models and to quantify uncertainties from ANN weights and biases (Kingston et al., 2005; Khan and Coulibaly, 2006) as well as uncertainties from model structures (Zhang et al., 2009). The BNN approach prescribes ANN parameters by relatively broad prior probability distributions. Then, during the calibration (training) period, prior distributions are sampled by Markov chain Monte Carlo (MCMC) (Geyer, 1992) or Metropolis algorithm (Haario et al., 2001) to update to posterior distributions given observation data. Finally, the posterior distributions of ANN parameters are used to evaluate prediction means and variances. However, the BNN method can be overwhelmed by a high number of ANN parameters. In addition, it is crucial to identify the contribution of individual sources of uncertainty to total prediction uncertainty (Wagener and Gupta, 2005).

In this study, we adopt the hierarchical Bayesian model averaging (HBMA) (Chitsazan and Tsai, 2015a; Tsai and Elshall, 2013) to analyze the contribution of individual sources of uncertainty to ANN predictions. Bayesian model averaging (Draper, 1995; Hoeting et al., 1999) is a well-known ensemble approach based on the law of total probability and Bayes' theorem. BMA derives overall prediction mean and variance by averaging all models at once (Nadiri et al., 2014). However, it does not quantify contributions of individual uncertain model elements to total estimated variance. The HBMA extends the Bayesian model averaging method by segregating multiple sources of model structure uncertainty in a hierarchical order (Chitsazan and Tsai, 2015b). The HBMA derives BMA prediction mean and variance by averaging subset models that are developed by an uncertainty source. By doing so, the HBMA is able to identify the contribution of each source of uncertainty to total prediction variance. This study adopts the HBMA to analyze the structural uncertainty as well as parameter uncertainty in the ANNs. The advantages of the HBMA in comparison to other neural network ensemble methods are investigated via the study of predicting fluoride concentration in aquifers of the Maku area, Azarbaijan, Iran.

2. Methodology

2.1. Artificial neural network (ANN)

Artificial neural networks (ANNs) have been applied to various large-scale problems, including pattern recognition, classification, parameter estimation and prediction (Anderson and McNeill, 1992; Ablameyko et al., 2003). The most widely used ANNs to simulate hydrological systems are the multilayer perceptron (MLP), a three-layer feedforward neural network (Govindaraju, 2000a,b; Sharma et al., 2003). The MLP consists of an input layer, a hidden layer, and an output layer. Fig. 1 shows a schematic of a typical multilayer perceptron ANN structure. This study uses four types of input data (Na^+, K^+, Ca^{2+}, HCO_3^-). The processing elements in each layer are called neurons or nodes. The input layer consists of input variables for the network. The output layer sends out predictions (outputs) of the network. The output function of the output layer is a linear function that aggregates the input signals of the output layer. The hidden layers are placed between the input layer and the output layer to transform and deliver signals. A vital part of the hidden layer which has a main role in transferring the network inputs to output is an activation function, which is typically a continuous and bounded nonlinear function such as hyperbolic-tangent-sigmoid (Tansig) or logarithm-sigmoid (Logsig) functions (Cybenko, 1989).

Training algorithms are needed in order to train ANNs to behave similar to the system of interest. Data used for training ANNs are referred to as "in-sample" data. Data that are not used in the training procedure is referred to as "out-sample" data. A brief overview of the MLP, including its mathematical aspects and implementation details, can be found in Govindaraju (2000a). The mathematical expression of a three-layer feedforward ANN for prediction is

$$O_{jk} = f_1(b_j + \sum_i W_{ij} I_{ik}) \quad (1)$$

$$O_k = b + \sum_j W_j O_{jk} \quad (2)$$

where f_1 is the activation function for the hidden layer, I_{ik} is the ith input for the kth sample point, O_{jk} is the output of jth node of the hidden layer, W_{ij} and W_j are the weights that control the strength of connections between layers and b_j and b are the biases that are used to adjust the mean value for the hidden layer and the output layer, respectively. The ANN output O_k is the kth predicted fluoride concentration. In the ANN training step, we use the Levenberg–Marquardt (LM) algorithm (Sahoo and Ray, 2006; Sahoo et al., 2006) as a supervised learning algorithm to estimate the weights (W_{ij} and W_j) and the biases (b_j and b) (Daliakopoulos et al., 2005).

There are two common sources of uncertainty in the three-layer feedforward ANN structure: (1) which activation function should be used for the hidden layer? (2) How many nodes are needed in the hidden layer? This study considers Tansig and Logsig activation functions and two to nine nodes in the hidden layer, combinations of which result in $2 \times 8 = 16$ ANN structures for fluoride concentration prediction.

2.2. Hierarchical Bayesian model averaging for neural networks

To segregate and prioritize the sources of structural uncertainty in ANNs, this study introduces the hierarchical Bayesian model averaging (HBMA) for prediction and uncertainty analysis. Consider p sources of uncertainty in a hierarchical order. Each uncertainty source may suggest a number of propositions (or alternatives). For example, in this study the uncertainty in choosing an activation function for the hidden layer suggests two propositions: Tansig function and Logsig function. The uncertainty in choosing the number of nodes in the hidden layer suggests eight propositions: from two nodes to nine nodes. The combinations of all considered sources of structural uncertainty form the base models for fluoride concentration prediction.

We segregate and prioritize the sources of uncertainty in a hierarchical order and denote the base models at level p and name this level as base level. Aggregation of base models through the BMA over different propositions under the same source of uncertainty forms BMA models at one level up, named level $p-1$. Aggregation of BMA models at level $p-1$ through the BMA over different propositions under the same source of uncertainty forms another BMA models at one level up, named level $p-2$.