Available online at www.sciencedirect.com

ScienceDirect

Journal of Hydro-environment Research 9 (2015) 325–339

Journal of
Hydro-environment
Research

www.elsevier.com/locate/jher

Research paper

Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers

Sung Eun Kim, Il Won Seo*

Department of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-744, South Korea

Received 6 April 2014; revised 6 September 2014; accepted 26 September 2014
Available online 22 April 2015

Abstract

The Artificial Neural Network (ANN) is a powerful data-driven model that can capture and represent both linear and non-linear relationships between input and output data. Hence, ANNs have been widely used for the prediction and forecasting of water quality variables, to treat the uncertainty of contaminant source, and nonlinearity of water quality data. However, the initial weight parameter problem and imbalanced training data set make it difficult to assess the optimality of the results obtained, and impede the performance of ANN modeling. This study attempted to employ the ensemble modeling technique to estimate the performance of the ANN without the influence of initial weight parameters on the model results, and to apply several clustering methods, to alleviate the imbalance of the training data set. An ANN ensemble model was developed, and applied to forecast the water quality variables, *pH*, *DO*, *turbidity* (*Turb*), *TN*, and *TP*, at Sangdong station, on the Nakdong River. The optimal ANN models for each water quality variable could be selected from the ensemble modeling. The optimal ANN models for *pH*, *DO*, *TN*, and *TP*, of which the training target data set was distributed evenly, showed good results, with R squared higher than 0.90. But the ANN model for *Turb*, of which the training data set was imbalanced, showed large RMSE (11.8 NTU), and low R squared (0.58). The training data set of *Turb* was partitioned into several classes, by conjunctive clustering methods according to the patterns of data set for each number of clusters. The ANN ensemble models for *Turb* with the clustered training data set (clustered ANN models) were then developed. All clustered ANN models for *Turb* showed better results, than the model without clustering. In particular, the three-clustered ANN model showed an increase of R squared from 0.58 to 0.88, and a decrease of total RMSE from 11.8 NTU to 6.3 NTU.

Keywords: Artificial Neural Network; Ensemble modeling; Clustering; Water quality forecasting; Nakdong River

1. Introduction

The Artificial Neural Network (ANN) has become a new tool and an efficient model for the prediction and forecasting of various water quality variables in river systems, due to the inherent uncertainties of contaminant source and water quality data. However, despite many researches conducted using the ANN model to predict water quality variables, the model building process has been poorly treated, which has made it difficult to assess the optimality of the results obtained. In addition to the major criticism that ANNs lack transparency, ANNs still suffer from limitations and problems; and a significant research effort is needed to address these deficiencies of ANNs. It is well known that the deficiency of ANNs come from the initial weight parameter and imbalance of training data set in the ANN development process (Maier et al., 2010; Kasiviswanathan et al., 2013).

The ANN model gives different results for the same original inputs, depending on the initial weight parameter set, before training the neural network. Problems with the initial weight parameters often force the ANN modelers to select a single "good" result, and accept it as the final result, omitting explanation of the optimal initial weight parameter. Using a

* Corresponding author.
E-mail address: seoilwon@snu.ac.kr (I.W. Seo).

http://dx.doi.org/10.1016/j.jher.2014.09.006
1570-6443/© 2015 International Association for Hydro-environment Engineering and Research, Asia Pacific Division. Published by Elsevier B.V. All rights reserved.

Monte–Carlo experiment, Kolen and Pollack (1990) showed that the training algorithm was very sensitive to the initial weight parameters. Yam and Chow (1995) presented an algorithm based on linear algebraic methods for determining the optimal initial weight parameters, and showed that with the optimal initial weight parameters, the initial network error can be greatly reduced. Other methods involving the genetic algorithm (GA) have been implemented to find the optimal initial weight parameters, and have enhanced the accuracy of the ANN model (Venkatesan et al., 2009; Chang et al., 2012; Mulia et al., 2013). These researches agree that the optimal initial weight parameters were very sensitive to the training algorithms and data structures, and there were no fixed optimal initial weight parameters that were universally applicable to the varieties of data structures and training algorithms. For this reason, ensemble techniques have been applied, due to the basic fact that the selection of weights is an optimization problem, with many local minima (Hansen and Salamon, 1990). Laucelli et al. (2007) applied ensemble modeling and genetic programming to hydrological forecasts, and showed the error due to the variance is effectively eliminated, by using an average model (ensemble model), as the resultant model of many runs. Boucher et al. (2009) developed the one-day ahead ensemble ANN model, for streamflow forecast. This study showed that random initialization of the weight parameters mainly accounted for the uncertainty linked to the optimization of the model's parameters; and ensemble modeling could reduce the uncertainty, using the proper assessment tools for the performance of ensemble models. Zamani et al. (2009) developed an ensemble ANN model with a Kalman Filter that corrects the outputs of the ANNs, to find the best estimate of the wave height; and showed the prediction results were improved, as the number of ensemble members increased. Khalil et al. (2011) developed the ensemble ANN model for the estimation of the mean values of four selected water quality variables. The results showed that the ensemble ANN model provided better generalization ability, than the single best ANN model. These researches indicate that the ANN model cannot guarantee that the model will produce an optimal result, without considering appropriate methods for the initial weight parameters.

The imbalance of the training data set is one of the fundamental problems in ANN modeling, and has recently drawn much attention (Zhou and Liu, 2006; Alejo et al., 2007; Yoon and Kwek, 2007; Nguyen et al., 2008). The imbalance (uneven distribution) of water quality data sets is common, where the number of training instances of a minority class is much smaller, compared to other majority classes (Nguyen et al., 2008). As a result, the neural network has difficulty in learning from imbalanced data sets, since the network tends to ignore the minority class, and treats it as noise, due to the overwhelming training instances of the majority class (e.g. Murphey et al., 2004; Nguyen et al., 2008). To alleviate the problem of the imbalanced training data set, Lu et al. (1998), Berardi and Zhang (1999), and Yoon and Kwek (2007) have attempted to employ resampling methods, such as oversampling and under-sampling, and modification of the training algorithms. However, their methods included modifying the probability or distribution of the training data set, which led to loss of information of the data set, and increase of the training time.

The objective of this study is to reduce the modeling errors of ANN in water quality prediction caused by the initial weight parameter problems and imbalanced training data set, by employing an ensemble modeling technique, and clustering methods. Ensemble modeling was applied to estimate the ANN performance, by removing the effect of initial weight parameters on the variance of ANN model results. In order to alleviate the imbalance of the training data set, several clustering methods were applied to separate the training data set, according to the patterns in the training data set, without the process modifying the probability or distribution of the data set. In this study, each one-step ahead water quality forecasting ANN ensemble model for *pH*, *DO*, turbidity (*Turb*), *TN*, and *TP* was developed. ANN ensemble models with clustered training data sets (clustered ANN models) were developed for *Turb*, of which the training data set was highly imbalanced.

2. Models and methods

2.1. ANN ensemble modeling

The ANN consists of a very simple and highly interconnected processor called a neuron. A neuron is an information-processing unit that is fundamental to the operation of a neural network, and consists of a weight and an activation function (Fig. 1). The weights are the most important parameters acting as the memory of ANN, and the activation function provides nonlinear mapping potential with the network. The manner in which the neurons of ANNs are structured determines the architecture of ANNs (Haykin, 1999). In general, there are three fundamentally different classes of network architecture. The first is a single-layer feedforward network, without hidden layers. The second is a multilayer feedforward network, with more than one hidden layer. The third is a recurrent neural network, with at least one feedback loop. In this study, the multilayer feedforward neural network (MFNN) with one hidden layer was used, because it is able to approximate most of the nonlinear functions demanded by practice (Mulia et al., 2013).

The weight parameters on the links between neurons are determined by the training algorithm. The most common and standard algorithm is the backpropagation training algorithm, the central idea of which is that the errors for the neurons of the hidden layer are determined by back-propagation of the error of the neurons of the output layer, as shown in Fig. 1. There are a number of variations in backpropagation training algorithms on the basic algorithm that are based on other standard optimization techniques, such as the steepest descent algorithm, conjugate gradient algorithm, and Newton's method. Among various backpropagation methods, the Levenberg–Marquardt (LM) algorithm has been very successfully applied to the training of ANN to predict streamflow and water