Expert Systems with Applications 37 (2010) 5856–5862

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic

Roohollah Noori [a,b,*], Amir Khakpour [b,c], Babak Omidvar [b], Ashkan Farokhnia [a,d]

[a] Department of Water Resources Research, Institute of Water Researches, Ministry of Energy, Tehran, Iran
[b] Department of Environmental Engineering, Graduate Faculty of Environment, University of Tehran, Tehran, Iran
[c] Director of Civil, Environmental, Laboratory and Consulting Engineering (CELCO) Company, Tehran, Iran
[d] Department of Water Resources Engineering, Kerman Graduate University of Technology, Kerman, Iran

ARTICLE INFO

Keywords:
Multivariate linear regression
Artificial neural networks
Monthly flow
Developed discrepancy ratio statistic

ABSTRACT

Predicting the stream flow is one of the most important steps in the water resources management. Artificial neural network (ANN) has been suggested and applied for this purpose by many of researchers. In such studies for verification and comparison of ANN results usually the popular methods such as multivariate linear regression (MLR) is used. Unfortunately, the presented methodology in some researches is faced with some problems. Thus, in this paper we have tried to find out the deficiencies of them and subsequently to present a correct the MLR methodology based on principal component analysis (PCA) for prediction of monthly stream flow. Then, assessment of different training functions on ANN operation is investigated and the best training function for optimizing the ANN parameters is selected. Afterward, the imperfections of the discrepancy ration (DR) statistic are remedied and a proper DR statistic is developed. Finally, the error distribution for testing stage of MLR and ANN models are calculated using developed DR statistic. The results of comparison show that the presented methodology in this research has improved the MLR operation. Also, comparing with the MLR, the ANN model possesses satisfactory predicting performance.

1. Introduction

Stream flow discharge forecasting has been considered as an important challenge for the researchers in the two past decades. For the purpose of stream flow discharge modeling different approaches such as regression (Kuligowski & Barros, 1998; Adeloye & Munari, 2006), conceptual (Jain & Srinivasulu, 2006; Xu, Seibert, & Halldin, 1996) and black box (Hsu, Gupta, & Sorooshian, 1995; Muller-Wohlfeil, Xu, & Iversen, 2003) models are used. Multivariate linear regression (MLR) model is on of the customary statistical models that used along with the artificial neural network (ANN) models in the hydrological modeling. Besides, the simpler application of it than conceptual models has caused MLR could be even used by no experts in field of water resources management (WRM). Applequist, Garhs, and Pfeffer (2002) compared five different techniques (ANN, linear regression, discriminant analysis, logistic regression and a classifying system) for rainfall forecasts. They used meteorological variables for training over central and eastern areas of the USA. The logistic regression model in their study had the best performance. Ramirez, Velho, and Ferreira (2005) compared two ANN and MLR techniques for rainfall forecasting. They reported that ANN forecasts were superior to the ones obtained by the linear regression model. Dawson, Abrahart, Shamseldin, and Wilby (2006) developed flood estimation model at ungauged sites using ANN and MLR models and stated that ANN provide better results than MLR model. High correlation of independent variables to each other is the famous customary in the hydrological process when we have very input variables. It caused the multicolinearity problem for MLR model that unfortunately neglected in the some researches (Applequist et al., 2002; Dawson et al., 2006; Kuligowski & Barros, 1998; Ramirez et al., 2005).

On the other hand, a survey on literatures of the past two decades show that ANN modeling studies in WRM are continued and every year we can see some papers about different aspects of this model which have innovative solutions for challenging problems in this field. Most of them have been done by feedforward backpropagation neural network (Karunanithi, Grenney, Whitley, & Bovee, 1994; Kisi, 2004; Noori, Farokhnia, Morid, & Riyahi, 2009a). The standard backpropagation algorithm (SBPA) has some problems include; the training convergence speed is very slow and easy

* Corresponding author. Tel.: +98 9374320526.
E-mail address: roohollahnoori@gmail.com (R. Noori).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.02.020

entrapment in a local minimum (Haykin, 1994). The researchers in during last decades have tried to find the optimum solution for these problems and improve the ANN operation. Chau (2006) has used the particle swarm optimization as a training function to optimize the network weights and biases for prediction the water level in Shing Mun River. He compared the results with the SBPA and reported the superiority of his model. Rogers, Dowla, and Johnson (1995) proposed the genetic algorithm instead of SBPA. Also, Wang, Gelder, Vrijling, and Ma (2006) have used gradient descent with momentum training function to optimize the network parameters for daily flow prediction. This training function often provides faster convergence than SBPA and momentum allows a network to respond not only to the local gradient, but also to recent trends in the error surface (Hagan, Demuth, & Beale, 1996). In another research Ramirez et al. (2005) proposed the resilient backpropagation (RP) training function for network training to predict the rainfall in Sao Paulo, Brazil. Using RP can improve the results. Multilayer networks usually use sigmoid transfer functions in the hidden layers. These functions are often called "squashing" functions. These functions compress an infinite input range into a finite output range. Sigmoid functions slopes must approach zero, as the input gets large. So when you use steepest descent to train a multilayer network with sigmoid functions some problem will be accrued, because the gradient can have a very small magnitude and it causes small changes in the weights and biases. RP can eliminate these harmful effects. Also, some researchers proposed the Levenberg–Marquardt algorithm suggested with Levenberg (1944) and Marquardt (1963) (TRAINLM) as a training function for SBPA.

In this paper we used two methods, ANN and MLR models. We have proposed a new application of principal component analysis (PCA) for using in process of feed data in MLR model. In addition, we investigated the effect of some important training functions which use heuristic and optimization methods to update the ANN weights and biases. Finally for results comparison of the models a proper statistic index is developed based on discrepancy ratio (DR) statistic presented by White, Milli, and Crabbe (1973).

2. Material and methods

2.1. Case study and data

Alavian dam is located on Sofichay River, in 120 km far from to Tabriz southwest. Input discharge to this dam is the combination of Sofichay River and a small branch that is named Esfestanj. Sofichay River with 4.6 m^3/s annual average discharge originate from Sahand mountains and after passing from west part of Tabriz flows down to Oromie Lake. Data sets consist of the monthly rainfall (R), discharge (Q), sun radiation (Rad) and temperature {all as minimum (T_{min}), maximum (T_{max}) and average (T_{ave})} data for 18 years (in the middle of 1983 close to the end of 2004) have been used. Mentioned data with the three temporal delays all together formed 18 variables which used as inputs. The most important statistic properties of data sets are listed in Table 1.

Table 1
Statistical characteristics of the data sets.

Statistical characteristics	R	Q	Rad	T_{min}	T_{max}	T_{ave}
Maximum	28.40	35.00	26.60	609.92	23.20	101.34
Minimum	−5.40	−1.40	−10.00	130.20	0.15	0.00
Standard deviation	9.47	10.72	8.43	149.89	4.54	24.89
Average	12.59	17.92	7.52	381.77	3.66	26.71
Median	12.50	17.90	7.60	372.35	1.57	21.60

2.2. Principal component analysis

PCA is one of the multivariate statistical methods which can use to reduce input variables complexity when we have a huge volume of information and we want to have a better interpretation of variables (Camdevyren, Demyr, Kanik, & Keskyn, 2005). By using of this method, input variables change into principal components (PCs) that are independent and linear compound of input variables (Lu, Wang, Wang, Xu, & Leung, 2003). Instead of direct use of input variables, we change them into PCs and then we can use them as input variables. In this method, the information of input variables will present with minimum losses in PCs (Helena et al., 2000). Details for mastering the art of PCA is published elsewhere (Noori, Abdoli, Ameri, & Jalili-Ghazizade, 2009b; Noori, Abdoli, Jalili-Ghazizade, & Samifard, 2009c; Noori, Kerachian, Khodadadi, & Shakibayinia, 2007; Tabachnick & Fidell, 2001; Wackernagel, 1995).

2.3. Multivariable linear regression

Regression model in matrix form can be shown as:

$$Y = X\beta + e \tag{1}$$

In Eq. (5), β is regression coefficient matrix, e is fitting error matrix and Y is response matrix. By solving Eq. (5) for β we will have:

$$\beta = (X'X)^{-1}(X'Y) \tag{2}$$

In Eq. (6), X' is transpose of X. For calculating inverse of $(X'X)$, it is necessary that the independent variables have not high relativity, because in this situation $(X'X)$ matrix cannot become inverse and we will have more error. To solve this problem, we should remove the multicolinearity between independent variables with PCA method. The variance inflated factor (VIF) criterion is usually applied to check the results. The ideal value for VIF is one. The higher VIF values, the more multicolinearity between independent variables exist.

2.4. Neural network

ANNs customary architecture compose of three layers. Theoretical works and many experimental results have shown that a single hidden layer is sufficient for ANNs to approximate any complex nonlinear function (Hornik, Stinchcombe, & White, 1989; Jalili-Ghazi Zade & Noori, 2008; Noori, Hoshyaripour, Ashrafi, & Nadjar-Araabi, 2010a; Noori, Abdoli, Farokhnia, & Abbasi, 2009d). A major reason for this is that intermediate cells not directly connected to output cells will have very small weight changes and will learn very slowly (Gallant, 1993). In this study a model based on a feedforward neural network with a single hidden layer is used. The backpropagation (BP) algorithm is used to train the network. The BP algorithm is essentially a gradient descent technique that minimizes the network error function (ASCE Task Committee, 2000; Haykin, 1994). There are two different ways in which this gradient descent algorithm can be implemented: incremental mode and batch mode. In this paper the batch mode was used for ANN training. Negative point, which exists in some studies, is that sigmoid transfer functions is used in the output layer of ANN (Nilsson, Uvo, & Ronny, 2006; Sahoo & Ray, 2006; Sha, 2007). This can have harmful effect on the results of these researches. If the last layer of a multilayer network has sigmoid neurons, then the outputs of the network are limited to a small range. If linear output neurons are used, the network outputs can take on any value (Haykin, 1994). With to take into consideration in this paper tangent sigmoid and purline transfer functions selected in hidden and output layers, respectively. There are different training functions to optimize the network weights and biases in the BP algorithm. They can divide in two categories. The first category uses heuristic techniques and the second category uses standard numerical optimization techniques.