Journal of Applied Sciences 10 (11): 1006-1010, 2010
ISSN 1812-5654

Comparison of Neural Network and K-Nearest Neighbor Methods in Daily Flow Forecasting

[1]Alireza Eskandarinia, [2]Hadi Nazarpour, [3]Mehdi Teimouri and [4]Mirkhalegh Z. Ahmadi
[1]Jihad-e-Agriculture Organization, Hamedan, Iran
[2]Babol Noshirvani University of Technology, Iran
[3]Ferdowsi University of Mashhad, Iran
[4]University of Sari Agriculture and Natural Resources Sciences, Iran

Abstract: This study illustrates the application of Multilayer perceptron (MLP) Neural Network (NN) for flow prediction of a Bakhtiari River. Since measurement of variables is time consuming and defining the efficient variable is essential for better performance of NN, alternative method of flow forecasting is needed. The K-Nearest Neighbor (K-NN) method which is a non-parametric regression methodology as indicated by the absence of any parameterized analytical function of the input-output relationship is used in this study. The implementation of each time series technique is investigated and the performances of the models are then compared. It is concluded that discharge in one day-ahead and Antecedent Precipitation Index (API) for seven days-ahead are the most important inputs and NN model has little better result than nearest neighbor method.

Key words: River flow, nonparametric model, discharge, precipitation, Bakhtiari

INTRODUCTION

Rivers in southwest of Iran have unique challenges for managers to characterize their flow patterns and behaviors, because there are many dams and other hydraulic structures that participate important role in people life. Also, stream discharges as are important for characterizing the nature of the stream, for measuring the severity of storm events and for finding the fresh water amount that needs to be stored in reservoirs and dams. Among the rainfall-runoff models, black-box models has simplified manner and can be formulated easily in an adaptive framework, because do not describe the hydrological processes occurring within the catchments (Brath *et al*., 2002). One of the black-box techniques that have been widely used in hydrology and water resources application is Artificial Neural Network (ANN). An ANN is an adaptable system that learns relationships from the input and output data sets and then is able to predict a previously unseen data set of similar characteristics to the input data (Haykin, 1999; ASCE Task Committee, 2000; Sahoo and Ray, 2006). The application of ANN in stream flow forecasting has already been discussed by many researchers like Maier and Dandy (2000), Birikundavyi *et al.* (2002), Solaimani and Zahara (2008) and Gholizadeh and Darand (2009). However, their investigations with ANN are different from each other in terms of the flow patterns of the streams, the number of variables in the input and output data sets, the measurements of ANN performances and the types of NN chosen and their training algorithms.

Another method in rainfall runoff modeling is K-nearest neighbor (K-NN), a non-parametric regression methodology, not implying any structured interaction, but exploiting the closeness between the most recent observation and K similar sets of observations chosen in any adequately large training sample (Toth *et al*., 2000). This study tests the daily forecasting improvement brought about by the inclusion of some inputs to ANN models like previous discharge (one and two day- ahead), precipitation in some stations and API in previous days (1, 7 and 15 days). In second step, daily forecasting with K-NN method obtained and two methods compared with each other. This study includes only the comparison of two approaches (ANNs and K-NNs). Bakhtiari River in South-west of Iran with continuous regime is used as a case study.

MATERIALS AND METHODS

Study area and data set description: The study catchment is the Bakhtiari River, a tributary of the Dez River in southwest part of Iran, with a drainage area of 6495 km^2 (Eskandarinia *et al*., 2009). The sources of river flow are mostly originated from snowmelt, springs and also individual rain events. Mean annual precipitation of watershed is about 1024 mm. The used data was hydrometric and climatic daily data with 21 years from 1981 to 2002. For the mentioned models 16 years data were applied for calibration and other 5 years for verification. Both daily datasets, which include stream flow Q, rainfall P have already been exploited for ANN R-R

Corresponding Author: Mehdi Teimouri, Ferdowsi University of Mashhad, P.O. Box 94615-147, Shirvan, Iran
Tel: 98 585 6353660 Fax: +98 585 6353663

Table 1: Statistical characteristics of the daily variables

Characteristics	Training	Testing
Dates	1981-1997	1998-2002
$\bar{Q}$ (cms)	176	126
$\bar{Q}_{max}$ (cms)	2546	1220
$\bar{P}$ (mm)	3.5	2.73
$\bar{P}_{max}$ (cm)	172	207

The over bar denotes a daily mean variable

forecasting, the former by Eskandarinia *et al.* (2009). A statistical comparison of the datasets and the selected models training and testing periods are identified in Table 1.

ANN methodology: There are many ways of using ANNs in the context of RR models (Anctil *et al.*, 2004) like network topology, training algorithm, input selection and network size optimization and each of them has a priori experience-based assumptions. Neural Networks distribute computations to processing units called neurons, grouped in layers. Three different layer types can be distinguished: input layer, connecting the input information to the network, one or more hidden layer and acting as intermediate computational layers between input and output and output layer, producing the final output (Toth *et al.*, 2000). Each one of the entering values is multiplied by a connection weight. Such products are then all summed with neuron-specific parameters, called bias, used to scale the sum of products into a useful range. Multi layer perceptron network with biases, a single hidden layer and an output layer are the most commonly used network topologies in the field of water resources (Coulibaly *et al.*, 2000; Maier and Dandy, 2000; Anctil *et al.*, 2004), provided that the training is sufficient. This topology is used here in accordance with the Levenberg-Marquardt backpropagation algorithm (Hagan and Menhaj, 1994), a second order non-linear optimization technique, since it is usually faster and more reliable than other back propagation variants (Masters, 1995; Bertsekas and Tsitsiklis, 1996).

Antecedent Precipitation Index (API): API has an old concept and is intended to reflect the fluctuating infiltration capacity of the soil associated with the frequency and depth of previous rainfall events (Anctil *et al.*, 2004). Heggen (2001) expressed the API as a precipitation moving average with decay (Eq. 1):

$$API_i(t) = \sum_{j=1}^{i} P_{t-j} k^{-j} \quad (1)$$

where, i is the number of antecedent days, k is the decay constant and P_{t-j} is the precipitation total for day t-j. The decay coefficient k is a watershed and seasonal parameter. The application of API in practice suggests that k should generally lie between 0.8 and 0.98 (Heggen, 2001; Anctil *et al.*, 2004). Here, k is set at 1.00 to reflect the common practice in the ANN community where non-decayed moving averaged is assumed. In this study API_i time series calculated with 1, 7 and 15 antecedent days.

The Nearest-Neighbor method (K-NN): The K-NN method is a non-parametric statistical pattern recognition procedure and among the various non-parametric techniques is the most intuitive, but nevertheless possesses powerful statistical properties (Toth *et al.*, 2000). Karlsson and Yakowitz (1987) extended the K-NN method, originally a pattern recognition procedure, to Time-series and forecasting problems, constructing a robust Theoretical base for the K-NN method and introducing it into the hydrological research world, where successful Forecasting applications were developed (Galeati, 1990; Shamseldin and O'Connor, 1996; Todini, 2000; Brath *et al.*, 2002). The prediction of a time series is based on a local approximation, making use of only the nearby observations. Define the composition of the feature vector D_t of dimension d, e.g., D_t: (xt_{-1}, x_{t-2}), d = 2; denote the current feature vector as D_t and determine its k nearest neighbors among the D_t, using the weighted Euclidean distance (Eq. 2):

$$r_{it} = \left(\sum_{j=1}^{d} w_j \left(v_{ij} - v_{tj} \right)^2 \right)^{1/2} \quad (2)$$

where, v_{tj} is the jth component of D_t and w_j are scaling weights (e.g., 1 or $1/s_j$, where s_j is some measure of scale such as the standard deviation or range of v_j).

In the next step a discrete kernel (k ($j_{(i)}$)) for resampling of the $x_{j(i)}$ define as follow (Eq. 3):

$$K\left(j_{(i)}\right) = \frac{1/j}{\sum_{j=1}^{k} 1/j} \quad (3)$$

where, K ($j_{(i)}$) is the probability with which $x_{j(i)}$ is resample. This resampling kernel is the same for any i and can be computed and stored prior to the start of the simulation (Lall and Sharma, 1995).

Evaluation criteria for prediction: The performances of the ANN and K-NN are measured with three efficiency terms. Each term is estimated from the predicted values of the models and the measured discharges as follows:

- The correlation coefficient (R-value) has been widely used to evaluate the goodness-of-fit of hydrologic and hydrodynamic models (Legates and Mccabe Jr, 1999). This is obtained by performing a linear regression between the ANN and K-NN predicted values and targets and is computes by Eq. 4:

$$R = \frac{\sum_{i=1}^{N} t_i P_i}{\sqrt{\sum_{i=1}^{N} t_i^2}\sqrt{\sum_{i=1}^{N} P_i^2}} \quad (4)$$

where, N is the number of samples, $t_i = T_i - T_{avg}$, $p_i = P_i - P_{avg}$, T_i and P_i are the target and predicted values for i = 1,…,N, T_{avg} and P_{avg} are the mean values of the target and predicted data set, respectively (Sahoo and Ray, 2006)

- The ability of the models predicted values to match measured data is evaluated by the root mean square error (RMSE). It is defined as (Eq. 5) (Schaap and Leij, 1998):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i - P_i)^2} \quad (5)$$

- The %VE statistic measures the percent error in volume under the observed and simulated flows, summed over the data period defined as Eq. 6:

$$VE\ (\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{T_i - P_i}{T_i}\right| \quad (6)$$

Overall, the models responses are more precise if R, RMSE and VE are found to be close to 1, 0 and 0, respectively

RESULTS AND DISCUSSION

ANN: Trial-and-error processes: In the present study the input dimensions are determined by the input variables and the lag time. Input selection is a crucial step in ANN implementation. The lack of pertinent input impairs the network capacity to map the input into a close estimate of the observed stream flow; the same is true for the use of unnecessary or redundant inputs. MLP networks are not engineered to eliminate superfluous inputs. It has been shown on many occasions that their performance is optimized by selecting only the inputs pertinent to the mapping (Anctil *et al*., 2004).

To determine an appropriate ANN structure forecasting the stream flow at time, we develop nine different models as follow:

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t)\} \quad (7)$$

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t), Q(t-1)\} \quad (8)$$

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t), Q(t-1), Q(t-2)\} \quad (9)$$

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t), Q(t-1), API_1\} \quad (10)$$

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t), Q(t-1), API_7\} \quad (11)$$

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t), Q(t-1), API_{15}\} \quad (12)$$

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t), Q(t-1), Q(t-2), API_1\} \quad (13)$$

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t), Q(t-1), Q(t-2), API_7\} \quad (14)$$

$$Q(t) = f\{P_{ka}(t), P_{za}(t), P_{gh}(t), P_{ta}(t), Q(t-1), Q(t-2), API_{15}\} \quad (15)$$

where, Q (t) is predicted runoff, Q (t-1), Q (t-2) are runoff for the time step of (t-1) and (t-2). P_{ka}, P_{za}, P_{gh} and P_{ta} are daily rainfall data of Kazemabad, Zardfahre, ghalian and Tangepanj rain gauge stations. $API_{1,\ 7,\ 15}$ are total antecedent precipitation index in the 1, 7 and 15 days ahead, respectively.

Data preprocessing, network training and testing: To ensure that all variables receive equal attention during the training process, they should be normalized (Maier and Dandy, 2000). At first step, time series candidates are compared according to the level of improvement they provide as MLP inputs measured against a reference performance obtained from optimized MLPs relying solely on stream flow and precipitation time series.

The performance efficiencies of a MLP with 3 hidden layer (an architecture of 6-3-1 represents 6 neurons in the input, 3 hidden layer and one neuron in the output layer) is presented in Table 2 (corresponding to Eq. 11). Figure 1 the predicted flows of NN with architectures 6-3-1 are compared with corresponding measured flows.

Nearest Neighbor Method: trial-and-error process: A trial-and-error procedure was implemented for a Number of nearest neighbors, K, ranging from 5 to 100 and a dimension of the feature vector, d, ranging from 2 to 12. The improvement of the performance with an increasing number of nearest neighbors is less noticeable for more than 20 neighbors and there is no marginal improvement in the overall performance when increasing K beyond 50. Small values (from 2 to 4) of the feature dimension gave

Table 2: Performance efficiencies of different network architectures in training and testing stages

	Training/calibration data			Testing/validation data		
NN architecture	RMSE	VE	R	RMSE	VE	R
4-3-1	2.54	95.65	0.379	2.82	136.98	0.246
5-3-1	0.84	7.90	0.934	1.09	9.30	0.936
6-5-1	0.86	8.59	0.938	1.06	11.63	0.933
6-5-1	0.83	9.27	0.939	1.04	11.38	0.939
6-3-1	0.69	7.80	0.940	1.03	8.28	0.957
6-2-1	0.83	8.10	0.929	1.14	8.75	0.937
7-5-1	0.83	7.99	0.941	1.03	9.09	0.937
7-3-1	0.86	8.14	0.936	1.07	9.86	0.935
7-5-1	0.84	7.44	0.943	1.04	8.94	0.936

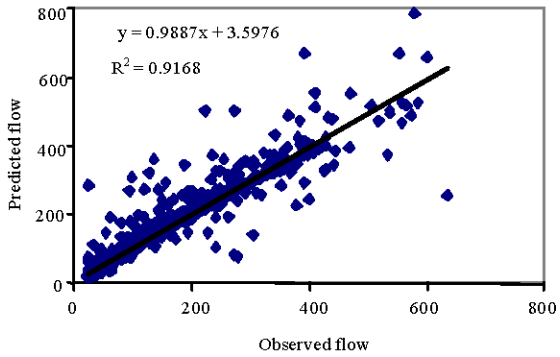

Fig. 1: Scattering graph of predicted against measured flow (cms) in ANN model

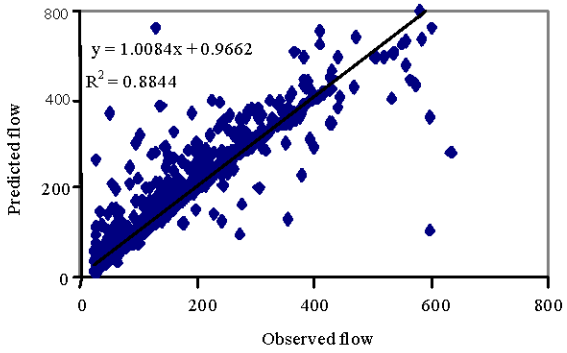

Fig. 2: Scattering graph of predicted against measured flow (cms) in K-NN model

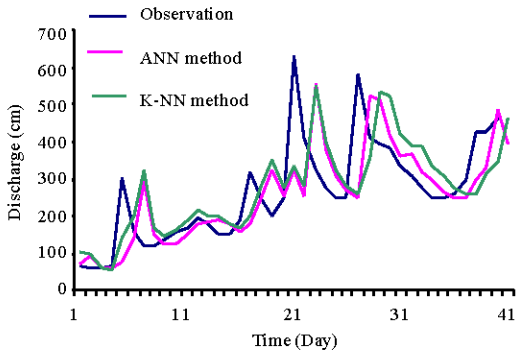

Fig. 3: Comparison of observed data with predicted data using ANN and K-NN models

the most satisfactory results for each given Number of neighbor vectors K.

Model performance level measured by R, RMSE and VE and result show its values as 0.94, 0.853 and 9.033, respectively. Furthermore, computed flows by K-NN model is compared with the corresponding observed values and illustrated by Fig. 2 and 3 show the

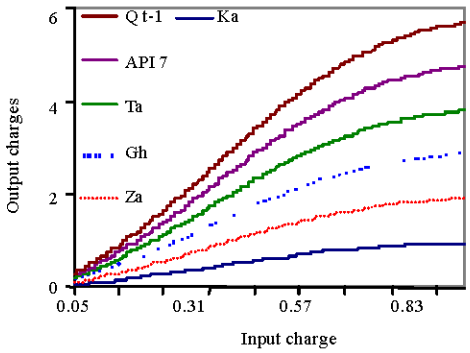

Fig. 4: Sensitivity Analysis of ANN model to changing the variables

relationship between flow data at gauge station and predicted data using the best models from ANN and K-NN for partial year of 1999.

Sensitivity analysis: The crucial tasks in MLP NN modeling are to design a network with a specific number of layers, each having a certain number of neurons and to train the network optimally. ANN model sensitivity and ration of input changes to output changes are shown in Fig. 4. As have seen input changes in one day-ahead discharge, API_7 and amount of rain in tangepanj station have more output changes to input changes than other variables.

CONCLUSIONS

In this study, MLP-ANN models trained for river flow forecasting by means of a variety of climatic forcing derived from datasets gathered at Bakhtiari River in southwest of Iran. Nine climatic forcing time series are then used as additional inputs to the optimized ANN models in order to test their significance as inputs. All of these are derived from simple hydro meteorological concepts.

Also nearest neighbor approach was applied for forecasting River flow that occurred in the stream. The forecast performances of each technique were evaluated by comparing observed and predicted River flow data. As discussed by Dibik and Solomatine (2001) in modeling process using ANN, the optimum number of antecedent rainfall and runoff should be investigated first and only then the final simulation could be made. Accordance to previous section, while correlation between observed and predicted data was obtained 0.957 on verification data with the appropriate input pattern, corresponding value of 0.94 were found with the K-NN model and proposed ANN could be applied for stream flow forecasting slightly better performance to the given K-NN model. The results showed that ANN models has

encouraging finding according to the gained results of this study and the previous findings of Tokar and Markus (2000), Lee *et al.* (2002). Also, Toth *et al.* (2000) provide a significant improvement in flood forecasting with ANN. The results of this study has shown that with combination of computational efficiency measures and ability of output variables which describe physical behavior of hydro-climatologic variables, improvement of the model predictability is possible in ANN environment as shown by Anctil *et al.* (2004).

ACKNOWLEDGMENTS

The writers would like to thanks the Sari Agriculture and Natural Resources University, for technical and financial supports.

REFERENCES

Anctil, F., C. Michel, C. Perrin and V. Andreassian, 2004. A soil moisture index as an auxiliary ANN input for stream flow forecasting. J. Hydrol., 286: 155-167.

ASCE Task Committee, 2000. Arti?cial neural network in hydrology. II: Hydrologic applications. J. Hydrol. Eng., 5: 124-137.

Bertsekas, D.P. and J.N. Tsitsiklis, 1996. Neuro-Dynamic Programming. 1st Edn., Athena Scientific, Belmont, MA., ISBN-10: 1886529108, pp: 491.

Birikundavyi, S., R. Labib, H.T. Trung and J. Rousselle, 2002. Performance of neural networks in daily stream flow forecasting. J. Hydrol. Eng., 7: 392-398.

Brath, A., A. Montanari and E. Toth, 2002. Neural networks and non parametric methods for improving real time flood forecasting through conceptual hydrological models. Hydrol. Earth Syst. Sci., 6: 627-639.

Coulibaly, P., F. Anctil and B. Bobee, 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J. Hydrol., 230: 244-257.

Dibik, Y.B. and D.P. Solomatine, 2001. River flow forecasting using artificial neural networks. J. Phys. Chem. Earth, 26: 1-7.

Eskandarinia, A., M. Ahmadi, H. Nazarpour, M. Teimouri and M. Zakermoshfegh, 2009. Investigation of antecedent precipitation in daily flow forecasting using rainfall runoff intelligence modeling. Proceedings of the 8th International Congress on Civil Engineering, May 11-13, Shiraz, Iran.

Galeati, G., 1990. A comparison of parametric and non-parametric methods for runoff forecasting. Hydrol. Sci. J., 35: 79-94.

Gholizadeh, M.H. and M. Darand, 2009. Forecasting precipitation with artificial neural networks (case study: Tehran). J. Applied Sci., 9: 1786-1790.

Hagan, M.T. and M. Menhaj, 1994. Training feedforward networks with the Marquardt algorithm. IEEE Trans. Neural Networks, 5: 989-993.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. 2nd Edn., Prentice Hall, New Jersey, USA., ISBN: 8120323734, pp: 443-484.

Heggen, R.J., 2001. Normalized antecedent precipitation index. J. Hydrol. Eng., 6: 377-381.

Karlsson, M. and S. Yakowitz, 1987. Nearest-neighbor methods for non-parametric rainfall-runoff forecasting. Water Resour. Res., 23: 1300-1308.

Lall, U. and A. Sharma, 1995. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resour. Res., 32: 679-693.

Lee, D.S., C.O. Jeon, J.M. park and K.S. Chang, 2002. Hybrid neural networks modeling of a full scale industrial waste water treatment process. Biotechnol. Bioeng., 78: 670-682.

Legates, D.R. and G.J. Mccabe Jr, 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydro-climatic model validation. Water Resour. Res., 35: 233-241.

Maier, H.R. and G.C. Dandy, 2000. Neural networks for the prediction and forecasting of water resources variables: A review of modeling issues and applications. Environ. Modeling Software, 15: 101-124.

Masters, T., 1995. Advanced Algorithm for Neural Networks: A C^{++} Sourcebook. John Wiley and Sons Inc., New York, ISBN-10: 0471105880, pp: 448.

Sahoo, G.B. and C. Ray, 2006. Flow forecasting for a Hawaii stream using rating curves and neural networks. J. Hydrol., 317: 63-80.

Schaap, M.G. and F.J. Leij, 1998. Database related accuracy and uncertainty of pedotransfer functions. Soil Sci., 163: 765-779.

Shamseldin, A.Y. and K.M. O'Connor, 1996. A nearest neighbors linear perturbation model for river flow forecasting. J. Hydrol., 179: 353-375.

Solaimani, K. and D. Zahra, 2008. Suitability of artificial neural network in daily flow forecasting. J. Applied Sci., 8: 2949-2957.

Todini, E., 2000. Real-time Flood Forecasting: Operational Experience and Recent Advances. In: Flood Issues in Contemporary Water Management, Marsalek, J., W.Ed. Watt, E. Zeman and F. Sieker (Eds.). Kluwer Academic publisher, Netherlands, ISBN-10: 0792364511, pp: 261-270.

Tokar, A.S. and M. Markus, 2000. Precipitation runoff modeling using ANNs and conceptual models. J. Hydrol. Eng., 5: 156-161.

Toth, E., A. Brath and A. Montanari, 2000. Comparison of short-term rainfall prediction models for real-time flood forecasting. J. Hydrol., 239: 132-147.