MAPLE TREE
PUBLISHING HOUSE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION (IRJMT)

http://www.mapletreejournals.com/index.php/IRJMT
Received 04 August 2019 ISSN 2582-1040
Accepted 24 September 2019 2019; 1(6); 1-6
Published online 02 November 2019

Forecasting Monthly Discharge Using Machine Learning Techniques

Bharthavarapu Srikanth[1]*, A. Geetha Selvarani[2], Bibhuti Bhusan Sahoo[3]

[1] Research Scholar Civil Department Vel Tech Rangarajan Dr. sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, 600062, India.

[2] Professor, Civil Department Vel Tech Rangarajan Dr.sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, 600062, India.

[3] Associate Professor Department of Civil Engineering MVR college of engineering & Technology, Paritala (V), Kanchikacherla (M), Krishna District, Andhra Pradesh, 521180, India.

**Corresponding author E-Mail ID: srikanthbharthavarapu@gmail.com*

Doi: https://doi.org/10.34256/irjmtcon1

ABSTRACT

Discharge prediction methods play crucial role in providing early warnings and helping local people and government agencies to prepare well before flood or managing available water for various purposes. The ability to predict future river flows helps people anticipate and plan for upcoming flooding, preventing deaths and decreasing property destruction. Different hydrological models supporting these predictions have different characteristics, driven by available data and the research area. This study applied two different types of Machine learning techniques to the Tikarpara station present in the lower end of the Mahanadi river basin India. The two Machine learning techniques include Multi-layer perception (MLP) and support vector regression (SVR) MLP has shown great deal of accuracy as compared to SVR across the cases used in the study; based on available data and the study area, MLP showed the best applicability, compared to SVR techniques. MLP out performed SVR model with r2 = 0.75 and lowest RMSE = 0.58.MLP can be used as a promising tool for forecasting monthly discharge at the selected station.

Keywords: MLP, SVR, Forecasting, Time series

Article Highlights

- ***Two Machine learning models are applied for monthly river flow forecasting***
- ***MLP utilizes a supervised learning technique called backpropagation for training***
- ***One month ahead forecasting of monthly discharge can be suitable done with the help of developed MLP model.***

INTRODUCTION

Forecasting hydrological time series is an important issue in operational hydrology. Forecasting discharge (Q) plays a crucial role in many water resources management practices. Numerous data-driven modeling techniques were proposed for the forecast and simulation of the stream flow series in past few years [1]. Throughout the literature, traditional and machine learning models have been applied to this task. Some of the traditional techniques for hydrological time series forecasting includes autoregressive (AR), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and seasonal ARIMA (SARIMA). These models are made a great success in stream flow forecasting [2, 3] with an assumption that a time series is originated from a stochastic process with an infinite number of degrees of freedom. With increase in computational capability the use of data driven machine learning models increases in the last decades by various researchers across the globe.

The scope of this study is to compare two forecast models, Multilayer Perception (MLP) and Support Vector Regression (SVR) and develop an optimal model for monthly stream flow prediction. This paper is organized in the following manner. Section 2 presents the stream flow data used in this study and study area description. Section 3 first describes the Methodology of MLP and SVR. The implementation and development of the forecast models, including data preparation and selection of parameters, is discussed in Section 4. Forecast results are described in Section 5 and conclusions of the study are presented in Section 6.

2. STUDY AREA AND DATA COLLECTION

The daily stream flow data were collected from central water commission India for the station of Tikarpara located in the end stream of the river. The monthly stream flow series spanned from June 1972to May 2007 is used in this study which is derived from the daily stream flow data. Figure 1 shows the selected gauging station over the Mahanadi river basin. Figure 2 shows the monthly discharge time series

3. METHODOLOGY

3.1 Multilayer Perceptron (MLP)

MLP is one of the most widely used machine learning algorithm for discharge prediction in river. A MLP is a feed forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses back propagation for training the network. A MLP having a single hidden layer, with 4 input and 1 output node shown is shown in Fig.3.

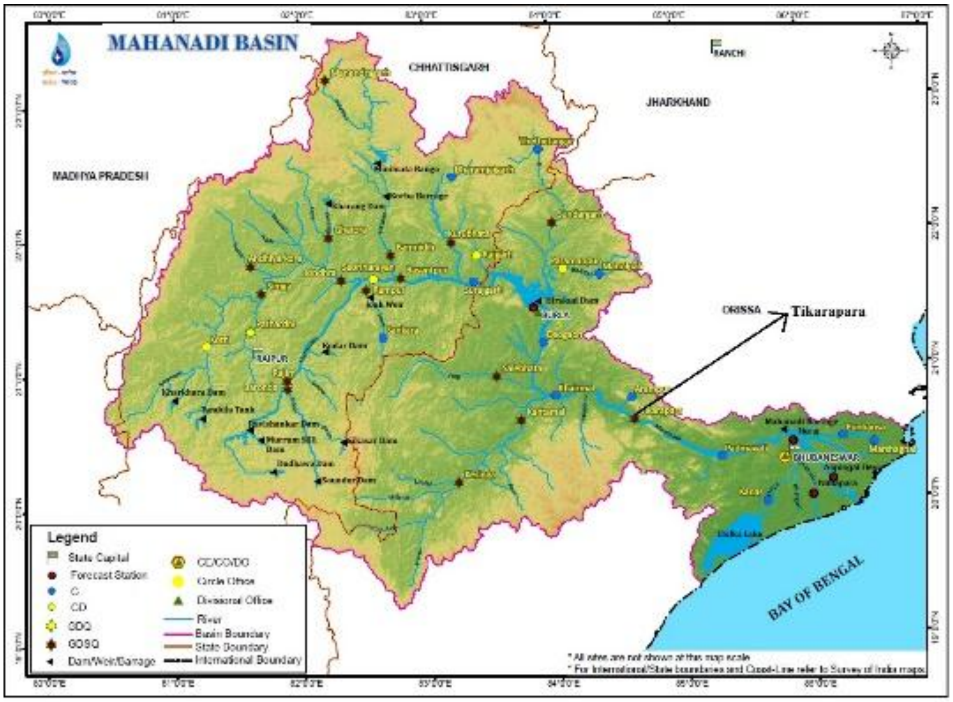

Figure 1 Showing the location of Tikarpara in Mahanadi river basin

(*Source: http://www.india-wris.nrsc.gov.in/wrpinfo/images/5/53/Mahanadi_basin.png*)

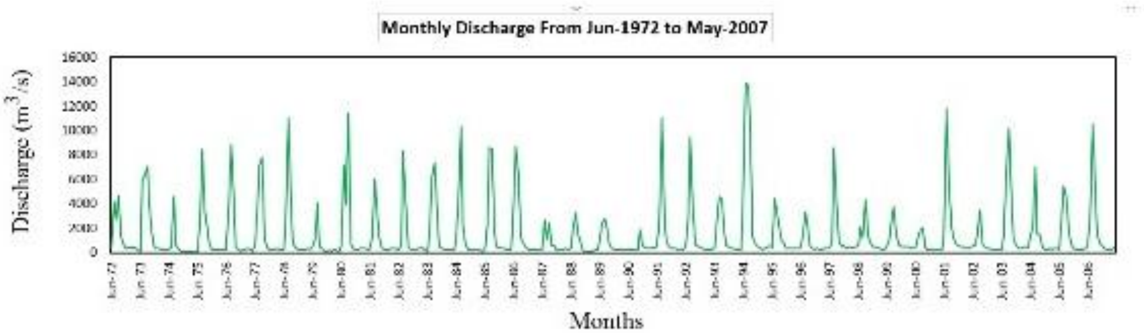

Figure 2 Monthly Discharge time series of Tikarpara

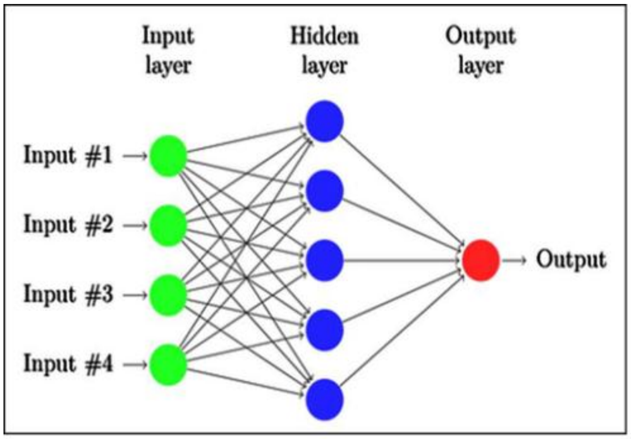

Figure 3 Architecture of a simple MLP

Each input unit of the training data set is passed through the network from the input layer to output layer. The network output is compared with the desired target output and output error (E) is computed using Eq. (1). This error is propagated backward through the network to each neuron, and the connection weights are adjusted based on Eq. (1).

$$E = 1/2 \sum_{i=1}^{N} (Q_i - Q_i^*)^2$$

Where the observed rainfall for ith is sample and is the predicted discharge for ith sample.

3.2 Support Vector Regression

SVR a robust and efficient algorithm developed by Vapnik [4] based on Statistical learning theory. It became more popular due to its successful application in classification [5, 6] and regression tasks to get minimum regression error [6] .The support vector regression (SVR) technique, instead, aims at finding the simplest function that can fit all the data while minimizing the sum of prediction errors above a predefined threshold. A review of the concepts and characteristics of these techniques with specific reference to hydrology is provided by [7, 8].

For a given training data with N number of samples, represented by $(x_1, y_1), \ldots \ldots, (x_N$ where is an input vector and is a corresponding output value, SVM estimator ( ) on regression can be represented by:

$$f(x) = w \cdot \varphi(x) \qquad (2)$$

Where   is a weight vector,   is a bias, ‘‘ ’’ denotes the dot product and   is a non-linear mapping function. A smaller value of w indicates the flatness of equation ( ), which can be obtained using minimizing the Euclidean norm as defined by   .Vapnik (1998) introduced the following convex optimization problem with the ε-insensitive loss function can be defined as follows.

$$L(y) = 0 \quad \text{For} \quad |f(x) - y| < \varepsilon$$

$$\text{Otherwise} \quad L(y) = |f(x) - y| - \varepsilon \qquad (3)$$

Eq. (3) defines a tube which is represented by ε in (Fig.4). The forecasted value has no loss when all the forecasted value within the tube (ε) otherwise forecasted loss is estimated by modulus of their deviation (forecast value -actual value) minus epsilon (ε).

This nonlinear regression problem can be expressed as in (Fig.4) that shows the generalized concept of SVR corresponding to Eq. (7)

$$\text{Minimize} \quad \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^* \right) \qquad (4)$$

Subjected to

$$y_i - \left(w \cdot \phi\left(x_i + b\right)\right) \leq \varepsilon + \xi_i$$

$$\left(w \cdot \phi\left(x_i + b\right)\right) - x_i \leq \varepsilon + \xi_i^* \qquad (5)$$

$$\xi_i, \xi_i^* \geq 0 \quad i = 1,2,3,........n$$

where ξ_i and ξ_i^* are slack variables introduced to evaluate the deviation of training samples outside the ε-insensitive zone, the distance from the training data from where the errors less than ε are ignored and index i labels the n training cases with x_i is the independent variable.

Hence the dual form of nonlinear of SVR can be formulated as using the kernel trick is expressed as follow

$$f(x) = \sum_{i,\, j=1}^{n} \left(\alpha_i - \alpha_i^* \right) \left(\phi\left(x_i\right) \cdot \phi\left(y_j\right) \right) + b \qquad (7)$$

Where α_i, α_i^* Lagrange multipliers variables constraints which lead to the construction of the dual optimization problem.

4. MODEL DEVELOPMENTS

4.1Input variable selection for forecasting and model development:

In hydrological time series forecasting models that compute the output from input (predictor) based on historical records which commonly uses the combination of different time lag [9,10] of the variable as an input parameter. For constructing these AI models, there exists no certain universally accepted guideline [11, 9].However the combination of different time lag seen as a common procedure reported by different researchers [9-12] in time series forecasting. The focus of the study is to predicting discharges i.e. the monthly flow using different time lags values to build up a model of the following form:

$$A^m = f\left(B^m\right) \qquad (8)$$

Where B^m is an m-dimensional input vector consisting of variables $b_1 \ldots, b_i \ldots b_m$, and A^m is

The output variable, consisting of the subsequent variables of interest $^{a_1 \ldots, a_i \ldots a_m}$.

We conducted Augmented Dickey–Fuller (ADF) test to make the time series stationary. The ADF test concluded that a lag of 7 is suitable for making the time series stationary. Thereafter we used up to lag 7 to predict the next month discharge in the selected station. The analysis carried out with 7 time lag of the low flow in the input vector, to build MLP and SVR model is built (Table1).

Table 1. MLP & SVR models used in the course of the analysis with the corresponding predictor variable as input

MLP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5}, Q_{t-6}, Q_{t-7}\right)$
SVR	$Q_t = f\left(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5}, Q_{t-6}, Q_{t-7}\right)$

5. RESULTS AND DISCUSSION

The creditableness of MLP model for forecasting monthly Q for the selected station over Mahanadi river basin, India is examined as a case study, where the models used the different lag of Q time series data. To assess the predictability of the MLP model, a comprehensive comparison with SVR models is performed using the several statistical metrics of forecasted and observed values of Q.Figure.5 (a), (b) shows a scatterplot of forecasted discharge versus the observed discharge value for the data analyzed in the testing period from MLP and SVR models

5.1 Performance Assessment

The performance of the MLP and SVR model was assessed by the comparison of the observed Q and the forecasted Q in the training and testing period using

1. Nash-Sutcliffe coefficient (ENS):

$$ENS = 1 - \left[\frac{\sum_{i=1}^{N}\left(Q_{obs} - Q_{for}\right)^2}{\sqrt{\sum_{i=1}^{N}\left(Q_{obs} - \bar{Q}_{for}\right)^2}}\right], \quad -\infty \le ENS \le 1$$

2. Coefficient of dete^{r}mination (r^2)

$$r^2 = \left[\frac{\sum_{i=1}^{N}\left(Q_{obs} - \bar{Q}_{obs}\right)\left(Q_{for} - \bar{Q}_{for}\right)}{\sqrt{\sum_{i=1}^{N}\left(Q_{obs} - \bar{Q}_{obs}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Q_{for} - \bar{Q}_{for}\right)^2}}\right]^2$$

3. Root-mean-square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Q_{obs} - Q_{for}\right)^2}{N}}$$

4. Mean absolute error (MAE)

$$MAE = \frac{\sum_{i=1}^{N}\left|Q_{obs} - Q_{for}\right|}{N}$$

Where = observed discharge; = forecasted discharge; = average observed discharge; = average forecasted discharge; N = number of data points (75% for training and 25% for testing of the data).

Table 2. shows various error calculated during testing period

	MLP	SVR
r^2	0.75	0.62
RMSE	0.58	0.74
MAE	0.36	0.52
ENS	0.68	0.54

6. CONCLUSIONS

Mahanadi river basin has many water resources. Water resource fluctuations due to climate change make integrated water management vital. This study used data from Tikarpara, one of the gauging stations, to better understand river discharge projections in Mahanadi river basin. Accurate river flow forecasts are a vital component of sustainable water management. The purpose of this study attempts to determine a relative optimal forecast model for monthly stream flow data. Two methods namely MLP and the SVR, were employed. Model performance was assessed using coefficient of determination (r2), Root mean square error (RMSE), Mean Absolute Error (MAE) and the Nash Sutcliffe model efficiency coefficient (ENS). MLP out performed SVR model with r2 = 0.75 and lowest RMSE = 0.58. In conclusion, MLP's can be used to predict river flows by using the historical flow data.

REFERENCES

[1] Marques C, Ferreira J, Rocha A, Castanheira J, 1. Marques C, Ferreira J, Rocha A, Castanheira J, Melo-Gonçalves P, Vaz N, Dias J (2006) Singular spectrum analysis and forecasting of hydrological time series. Physics and Chemistry of the Earth, Parts A/B/C 31 (18):1172-1179

[2] Carlson RF, MacCormick A, Watts DG (1970) Application of linear random models to four annual streamflow series. Water Resources Research 6 (4):1070-1078

[3] Box G, Jenkins G (1970) Time series analysis; forecasting and control. Holden-Day, San Francisco(CA).

[4] Vapnik V (1998) Statistical learning theory. 1998. Wiley, New York,

[5] Osuna E, Freund R, Girosi F (1997) Support vector machines: Training and applications.

[6] Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2 (2):121-167

[7] Deka PC (2014) Support vector machine applications in the field of hydrology: a review. Applied soft computing 19:372-386

[8] Pai P-F, Lin K-P, Lin C-S, Chang P-T (2010) Time series forecasting by a seasonal support vector regression model. Expert Systems with Applications 37 (6):4261-4265

[9] Sang Y-F, Wang D, Wu J-C, Zhu Q-P, Wang L (2009) The relation between periods' identification and noises in hydrologic series data. Journal of Hydrology 368 (1):165-177

[10] Nayak PC, Sudheer K, Rangan D, Ramasastri K (2004) A neuro-fuzzy computing technique for modeling hydrological time series. Journal of Hydrology 291 (1):52-66

[11] Sudheer K, Gosain A, Ramasastri K (2002) A data-driven algorithm for constructing artificial neural network rainfall-runoff models. Hydrological processes 16 (6):1325-1330

[12] Cheng C-T, Lin J-Y, Sun Y-G, Chau K (2005) Long-term prediction of discharges in Manwan Hydropower using adaptive-network-based fuzzy inference systems models. Advances in natural computation:434-434

Funding

This study was not funded by any grant

Conflict of interest

None of the authors have any conflicts of interest to declare.

About the License