Monthly Streamflow Prediction Based on Random Forest Algorithm and Phase Space Reconstruction Theory

Feng Peng¹, Jie Wen¹, Yao Zhang¹ and Junying Jin¹*

¹College of Resources and Environment, Southwest University, Chongqing, China.
*Corresponding author e-mail: junyingjin@126.com

Abstract. In order to find an effective time series forecasting method, a random forest prediction model based on phase space reconstruction theory is proposed in this paper. The Cao method and mutual information function method are used to reconstruct the phase space, and the parameters of the random forest (RF) model are discussed. The feasibility of the model is verified by analyzing the monthly runoff data of Pingshan hydrological station in Jinsha River Basin. Compared with BP neural network (BPnet) model and traditional support vector machine (SVM) model which optimizes parameters by grid algorithm, random forest model has higher prediction accuracy and less calculation in dealing with complex nonlinear hydrological time series as shown by the result.

1. Introduction

Hydrological system is a nonlinear and highly complex dynamic system. As hydrological process is affected by the coupling effect of human activities and natural conditions such as climate, landform, soil and vegetation, hydrological time series always show nonlinearity. Due to the nonlinearity and uncertainty of hydrological time series, it is very difficult to accurately predict the discharge. Many linear statistical prediction models such as AR model [1] and ARMA model [2] have been widely used in flow prediction for a long time in the past, and their accuracy is often disappointing. In recent decades, with the rise of machine learning algorithm, nonlinear prediction models such as artificial neural network (ANN) and support vector machine (SVM) have been applied to hydrological field [3,4], which improves the prediction accuracy of time series to a certain extent.

With the development of nonlinear prediction technology such as chaos theory, it has become an effective method to reconstruct hydrological time series into a phase space equivalent to hydrodynamics system to study the evolution of hydrological system [5]. At present, some scholars have used ANN to have a short-term prediction of discharge in the Namakan Lake sub basin after reconstructing the phase space, and good prediction results have been achieved [3]. The SVM, an alternative method of ANNs, is also applied to runoff prediction [4]. The results show that the performance of SVM for monthly runoff prediction is better than that of ANNs. Compared with the over fitting and slow learning of ANNs, SVM is sensitive to parameters and kernel functions, so we try to make up for the shortcomings of both by random forest algorithm. Therefore, a random forest model is constructed and the prediction performance of it based on phase space reconstruction theory is explored in this paper.

2. Methods

Previous studies have shown that there is chaos in hydrological systems [6,7]. Therefore, it is an effective tool to analyze streamflow time series by using chaos theory.
2.1. Reconstruction of Phase Space

For one-dimensional time series, Takens [8] and Packard et al. [9] proposed a method of reconstructing phase space. The method considers that it is only necessary to study one component and treat the observed values at some fixed time delay points as new dimensions to reconstruct an equivalent phase space and restore the original dynamic system in this phase space. For time series \( x_t(t = 1,2,3, ..., N) \), the phase space can be reconstructed as

\[
Y_j = (x_j, x_{j+\tau}, x, ..., x_{j+(m-1)\tau})
\]

(1)

Where, \( j = 1,2, ..., N - (m - 1)\tau \), \( m \) is the embedding dimension, \( \tau \) is the delay time.

2.1.1. Calculation of delay time. Mutual information function method (MIF) is adopted in this paper, which is better to reflect the nonlinear relationship between runoff time series. For chaotic time series \( x_t(t = 1,2,3, ..., N) \), the mutual information is defined as:

\[
I(\tau) = \sum_{t=1}^{N-\tau} P(x_t, x_{t+\tau}) \cdot \log_2 \left( \frac{P(x_t, x_{t+\tau})}{P(x_t) \cdot P(x_{t+\tau})} \right)
\]

(2)

Where, \( P(x_t) \) is the probability density of \( x_t \), \( P(x_t, x_{t+\tau}) \) is the joint probability density of \( x_t \) and \( x_{t+\tau} \). In the case of joint probability density \( I(\tau) = 0 \), \( x_t \) and \( x_{t+\tau} \) are independent. But when they are completely related, \( I(\tau) = \infty \)[10]. In order to select the best delay time, the MIF must be as small as possible. It usually takes the first minimum as the delay time.

2.1.2. Calculation of embedding dimension. Cao method [11] is an improvement of the false neighbors method. It retains the calculation principle of the false neighbors method, but it simplifies the calculation amount. Meanwhile, the selection of embedding dimension is more accurate. The specific contents about Cao method can refer to [11].

2.2. Random Forest Algorithm

Random forest is a combined classifier which contains multiple non-pruning classification regression trees (CART), which is the inheritance and development of traditional decision tree method [12]. The combined classifier introduces independent and identically distributed random variables \( \theta \). Then using training set data and \( \theta \) to generate decision tree \( h(x, \theta) \). Finally, all decision trees are combined by ensemble learning. For regression problem, the prediction result of random forest algorithm is the mean of all the predicted values of the decision trees.

\[
h(X) = \frac{1}{K} \sum_{k=1}^{K} h(x, \theta_k)
\]

(3)

Where \( k = 1, ..., K \). \( K \) is the number of regression trees. \( x \) is the input vector. \( \theta \) is an independent, uniformly distributed random vector, which is used by the prediction tree as a numerical value.

In random forest algorithm, \( ntree \) (the number of maximum regression trees), \( mtry \) (the best decomposition score), and \( minleaf \) (the minimum size of leaf nodes) are the main influencing factors on the prediction results. This paper mainly explores the influence of \( ntree \) and \( minleaf \) on the prediction.

In addition, BP neural network model and traditional support vector machine model are used to compare the prediction performance of random forest model [13,14].

3. Study Case

3.1. Study Area

The Jinsha River Basin is selected as the research object. Located in the upper reaches of the Yangtze River, Jinsha River originates from the Qinghai-Tibet Plateau, and is regarded as the starting point of the Yangtze River. It is 3481 km long and covers an area of 502000 km². With the development of national economy, the construction of water conservancy facilities on the Jinsha River has been
gradually developed, including flood control, hydropower generation, agricultural production, urban and industrial water supply and so on. Therefore, it is of great significance to forecast the medium and long-term streamflow of Jinsha River.

3.2. Model Building

Based on random forest toolbox of R language, monthly streamflow data of Pingshan station of Jinsha River control station from 1954 to 1986 is selected for prediction. 25 years of monthly streamflow data from 1954 to 1978 are used as training samples and the rest are verification samples. For the reconstructed phase space, the input vector of the model is \( Y_t = (x_t, x_{t+1}, x_{t+2}, \ldots, x_{t+(m-1)}) \), the corresponding expectation is \( x_{t+1} \).

Random forest model is not sensitive to the internal range of the sample, so it’s not necessary to normalize like using ANNs and SVM. Here are the specific steps for building a random forest model.

Step 1: Determine the best embedding dimension and delay time. Firstly, according to Cao method, the E1 and E2 distribution map with the maximum embedding dimension of 50 is calculated. The distribution diagram of E1 and E2 with embedding dimension of monthly runoff time series after calculation is shown in Fig.1.

According to Fig.1, the best embedding dimension of monthly streamflow time series is 17.

Secondly, the delay time is calculated by mutual information, and the mutual information function of monthly runoff time series is drawn as shown in Fig.2.

It can be seen from Fig.2 that the delay time of monthly streamflow time series is 5.

Step 2: Phase space reconstruction. According to the monthly streamflow time series with embedding dimension of 17 and delay time of 5 \((n = 396)\), it can be reconstructed into the following phase space:

\[
Y = \begin{bmatrix}
x_1 & x_6 & \ldots & x_{86} \\
x_2 & x_7 & \ldots & x_{87} \\
\vdots & \vdots & \ddots & \vdots \\
x_{315} & x_{318} & \ldots & x_{396}
\end{bmatrix}
\]

After phase space reconstruction, the number of vectors in 17-dimensional phase space is 315. In this paper, the first 219 vectors are used for model fitting and the others are used for model verification.

Step 3: Model fitting. The effects of the \textit{minleaf} and the \textit{ntree} on the mean variance error (MSE) in the model fitting are evaluated. The \textit{ntree} is 1000 and the optimal number of splits (\textit{mtry})
is the square root of the dimension of the input vector, and the minleaf is 1, 2, 5, 10, 20, 50, 100 for the model fitting. Fig. 3 shows the MSE with increasing number of regression trees under different minleaf.

It can be seen from Fig.3 that when the number of regression trees is 200, the MSE is relatively stable. With the increase of the number of regression trees, the MSE of different minleaf does not change much. It can be seen that MSE curves with minleaf of 20 and 50 are close to each other, and the MSE of both curves is greater than that of minleaf less than 10. The MSE of minleaf is the largest MSE of the curve of 100.

**Figure 3.** Predicted results using different models

Based on the above results, construction of the random forest model uses the following values of the parameters in this paper: ntree = 200, mtry = 5, minleaf = 5.

Step 4: Different models are used for streamflow simulation. In the past, ANNs and SVM are widely used in the prediction model. While in this paper, the BPnet model and the traditional SVM model are selected as the control groups to compare the prediction performance of the random forest models.

BPnet model parameters are set as follows: the maximum number of iterations is 2000 and the mean square error of termination criteria is \( e = 1 \times 10^{-5} \). The learning rate is \( lr = 0.05 \).

SVM model parameters are set as follows: The kernel function of SVM model is Gaussian radial basis function, and the grid search algorithm is used to search for the best value of parameter \( c \) and \( \sigma \). The values of the parameters are penalty coefficient \( c = 1.741 \), the gamma in kernel function \( \sigma = 0.1895 \), and loss function value \( \varepsilon = 0.01 \). The mean square error of the termination criteria is \( e = 1 \times 10^{-5} \).

3.3. Error Evaluation

Four indicators are used to evaluate the fitting and prediction ability of different models in this study, they are: root mean square error (RMSE), Nash Sutcliffe efficiency coefficient (NSE), determination coefficient (\( R^2 \)) and average relative error (MRE). As for RMSE and MRE, the smaller the better; while when it comes to NSE and \( R^2 \), the closer the indicator value is to 1, the better.

4. Results and Analysis

In this paper, the random forest model based on phase space reconstruction, BPnet model and traditional SVM model are used to predict the monthly streamflow of Pingshan station. The statistical results of prediction accuracy of different models are summarized in Table 1.
Table 1 shows that the random forest model has the best prediction performance. Its RMSE, NSE and R² have the best indicators, but its MRE is slightly lower than that of SVM model. BPnet model has the lowest prediction performance, and its prediction results are inferior to the other two models in all indicators. SVM model and RF model have the similar performance, and both their NSE are above 0.8, which shows that the predicted values of the two models are in good agreement with the measured values. But, the random forest model has better results in R² and RMSE than SVM model.

| Model | RMSE    | NSE    | R²     | MRE    |
|-------|---------|--------|--------|--------|
| BPnet | 1833.05 | 0.703  | 0.741  | 30.00% |
| SVM   | 1474.24 | 0.808  | 0.809  | 15.95% |
| RF    | 1416.37 | 0.823  | 0.828  | 16.45% |

Fig. 4 shows the prediction effect of the three models. It can be seen that all the three models have good fitting effect for the prediction of measured monthly streamflow.

![Influence of different MinLeaf on prediction accuracy](image)

Figure 4. MSE of Random Forest models with varied numbers of regression trees and minleaf

In addition, the calculation amount of the model is usually an object of concern, and its size can be expressed by the running time of the model. Under the environment of MATLAB 9.1 and CPU of inter i7 8th, the running time of the three models respectively are 3.981s for BPnet model, 3.067s for SVM model, and 0.894s for RF model. As the results, RF model has a faster convergence speed, in the application of a large amount of data time series, for example, the daily runoff time series, the RF model obviously has more advantages.

5. Conclusions

In order to find an effective streamflow prediction model, a random forest prediction model based on phase space reconstruction is proposed in this paper, and the influence of its parameters on the prediction performance is evaluated. In this study, for the data sample with data volume of 400 and input vector dimension of 17, good prediction results can be obtained when the number of regression trees is 200, the optimal split number is 5, and the minimum leaf size is 5.

Besides, BP neural network model and traditional support vector machine model based on grid optimization are used to compare with random forest model. The results show that the root mean
square error, Nash efficiency coefficient, determination coefficient and average relative error of random forest model are 1416.37, 0.823, 0.828 and 16.45% respectively. The prediction performance of BP neural network is the worst, and that of support vector machine model and random forest model is similar. However, prediction performance of the random forest model among the latter two is better. What’s more, the calculation time of random forest model is less than that of support vector machine, and it can be more advantageous in processing time series with high data volume.

6. Acknowledgments
This research is financially supported by Undergraduate Science and Technology Innovation “Guangjiong” Program of Resources and Environment College (201911) and Southwest University Teaching Reform Program (2019JY148), and the supports were gratefully acknowledged.

7. Reference
[1] Wensheng W, Juliang J and Yueqing L 2007 Adv Water Sci. 05 768-775.
[2] Salas J D and Obeysekera J T B 1982 Water Resour Res. 18 1011-1021.
[3] Tokar and Markus 2000 J Hydrol Eng. 5 156-161
[4] Lin J Y, Cheng C T and Chau K W 2006 Hydrolog Sci J. 51 599-612
[5] Kantz H and Schreiber T 1997 Nonlinear Time Series Analysis. (Cambrige: Cambridge University Press)
[6] Sivakumar B 2000 J Hydrol. 227 1-20
[7] Khatibi R, Sivakumar B, Ghorbani M A, Kisi O, Kocak K and Zadeh D F 2011 J Hydrol. 414 108-117.
[8] Takens, F 1980 Lect Notes Math 898 366–381
[9] Packard N H, Crutchfield J P, Farmer J D and Shaw R S 1980 Phys Rev Lett. 45 712–716.
[10] Özlem B and Kasim K 2011 J Hydrol. 508 356-363
[11] Cao L 1997 Physica D. 110 43–50.
[12] Breiman L 2001 Mach Learn. 45 5–32.
[13] Sadeghi BHM 2000 J Mater Process Tech.103 411-416.
[14] Lin C J 2003 A practical guide to support vector classification (Taipei: Department of Computer Science & Information Engineering National Taiwan University) p 1-14.