Wavelet Neural Networks Model Used for Runoff Forecast Based on Fuzzy C-Means Clustering

Zhong Wei¹, Ma Xiaoyu¹, Xu Jin¹*

I School of Management
Tianjin University of Technology
Tianjin 300384, Tianjin, China

Abstract — Considering various seasons differs greatly in runoff distribution, a new runoff forecasting method based on fuzzy clustering analysis on forecasting factor set is presented in this paper. Firstly, the historical runoff data are classified as four categories by fuzzy C-means clustering. Then partial forecasting models between the factor set and measured data are respectively established by using wavelet neural network model. A network model categorized recognizer is adopted, which can automatically search a compatible partial forecasting model. Comparison between simple wavelet neural model and integrated forecasting model proposed in this paper is made by illustration. The results demonstrate that the proposed integrated model is of higher forecasting accuracy than the simple one.

Keywords - runoff forecast; wavelet neural network model; fuzzy C-means clustering; genetic algorithm

I. INTRODUCTION

Through statistical analysis on different regions' runoff over years, it is commonly discovered that various seasons differs greatly in runoff distribution. The traditional numerical methods for runoff forecasting generally applies a simple model to runoff series as a whole. The simple model always takes means as variables and parameters, which is not conducive to simulating the reservoir's great influx differences caused by different conditions of various seasons. Therefore, taking the runoff process as a whole that could be averaged by different phases is obviously inappropriate.

In this paper, fuzzy C-means clustering[1], wavelet neural network [2-4] and pattern recognition technology are integrately applied in runoff forecasting. Firstly, make use of FCM method to classify the historical runoff data, and then respectively establish partial forecasting models between the forecasting factors set and measured data with wavelet neural network model. By a categorized recognizer of network model to automatically search a compatible partial forecasting model. The established model has been applied to a reservoir’s daily runoff forecasting. The result indicates that the forecasting model established by the proposed method is provided with high speed and high accuracy.

II. FUZZY C-MEANS CLUSTERING (FCM) THEORY

Fuzzy C-means clustering (FCM) is a clustering algorithm that can determine the degree of each data point falls into a certain clustering by degree of membership [5]. We suppose that there are n samples Xk(k=1,2,…,n), the samples are divided into C categories, uik signify the membership degree of sample xk on category i, the value of uik is defined between [0, 1] and meet the formula (1):

\[ \sum_{i=1}^{c} u_{ik} = 1, \forall k = 1, 2, ..., n \]  

FCM's objective function is shown as formula (2):

\[ J(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^p d_{ik}^2 \]  

In the formula (2), U={uik}(i=1,2,…,c, k=1,2,…,n) is the membership degree set, V={v1,v2,…,vc} is the center sets of C clusters. p ∈ [1, ∞ ) is the weighted index. dik is the Euclidean distance between sample k and the center of cluster i. dik can be figured out by formula (3):

\[ d_{ik}^2(x_i, v_i) = \| x_i - v_i \|^2 = (x_i - v_i)' A (x_i - v_i) \]

In the formula (3), A is a positive definite matrix.

FCM is realized by the minimizing membership vector set and objective function Jm(U,V) of clustering center set V. Its calculation steps are as follows:

Step 1: Randomly generated initial membership set U={uik} within [0, 1] to fit the condition of formula (1)

Step 2: Calculate the center of cluster.

\[ v_i = \frac{\sum_{k=1}^{n} (u_{ik})^p x_i}{\sum_{k=1}^{n} (u_{ik})^p} \]  

Step 3: Using formula (2) to calculate objective function value Jm. If the result is less than the fixed convergence precision, or the result relative to changes in the last result is less than the convergence precision, the algorithm stop.

Step 4: Update the set of membership degree.

\[ u_{ik} = \left[ \frac{\sum_{j=1}^{c} \left( \frac{d_{jk}(x_i, v_j)}{d_{ik}(x_i, v_i)} \right)^{\frac{2}{p-1}} }{\sum_{j=1}^{c} \left( \frac{d_{jk}(x_i, v_j)}{d_{ik}(x_i, v_i)} \right)^{\frac{2}{p-1}} } \right]^1 \]
Ⅲ. Wavelet Neural Network Model Used for Forecasting

A. Structure of Wavelet Neural Networks Model

To design the wavelet neural network model used for forecasting, which is a multi-input & single-output system, commonly 3-layer network structure is applied. Its input layer has m nodes, hidden layer has n nodes and output layer has one node, as shown with figure 1:

B. Calculation Flow of the Model

The parameters of Ukj, Wj, aj, bj, θ in the network model can be trained by repeatedly approaching the genetic algorithms[7]. Training parameters of the model to be designed in sequence composed of string-type structure parameters vector, each vector is a genetic manipulation of chromosomes. The initial value of parameters determined by the following methods:

1) Define the initial value of parameter aj and bj

The initial value for the parameter θ to determine runoff forecasting, which is a multi-input & single-output system, commonly 3-layer network structure is applied. Its input layer has m nodes, hidden layer has n nodes and output layer has one node, as shown with figure 1:

2) Define the initial value of parameter Ukj and Wj

In these formulas mentioned above, number of the hidden layer of network model is n; yjmax and yjmin are the input layer’s No. neuron j in training sample maximum and minimum values.

3) Define the initial value of parameter θ

The initial value of the weight from network input layer to hidden layer - Ukj and the weight from hidden layer to output layer - Wj are randomly selected from the range of [-1,1], and both the value of Ukj and Wj can not be zero.

4) Standardized the training sample data

The training sample data should be went through standardized treatment before the network training. According the formula (9), so as to make the initial sample within the range of [0,1].

In the formula (9), the value of xk(max) is the smallest number in the input data; and the value of xk(min) is the biggest number in the input data.

5) The network training

In the training process, each parameter is selected as follow: wavelet neurons function is set as Morlet wavelet[9]. Its population size L=100; the largest iteration number of evolution Jt=300; network training convergence precision e=2×10-3; selection probability Ps=0.65; crossover probability Pc=0.8; mutation probability Pm=0.05; the fitness value E can be figured out by formula (10);

In above formula(6), x=(x1,x2,…,xm)T is the input vector; Ukj is the connection weights from the node k of the input layer to the node j of the hidden layer;Ψj is the activation function of neuron j from the hidden layer; Wj is the connection weights from node j of the hidden layer to the output layer; bj is the translational parameter of the wavelet function; aj is the stretching parameter of the wavelet function; θ is the mean estimation of the output timing; Q(x) is the output value of the network.

In the formula (10), the fitness value E is smaller, then the effect of network training is better; the value of Q(Xt) is the result of calculation by formula (6); the value of Qt is actual runoff data; the value of T is number of sample runoff data. The entire network training process is below:

Step 1: At the first of network training, the initial species P1,P2,…,PL are determined in according to the parameters of the value methods mentioned above, and the value of J is set as 0.

Step 2: The fitness value E is Calculated and the result is chosen the fitness function value of the smaller.

Step 3: The higher fitness individuals is select to enter the next generation after crossover and mutation operation.

Step 4: Convergence judgment.

Step 5: Most optimized collection of the network parameters is figure out; the end of training network.

The process is shown in figure 2.
The initial species $P_1, P_2, ..., P_L$ are determined in accordance to the parameters of the value methods mentioned above, and the value of $J$ is set as 0.

The fitness value $E$ is calculated and the result is chosen the fitness function value of the smaller.

The higher fitness individuals is select to enter the next generation after crossover and mutation operation.

$\frac{E_{\text{max}} - E_{\text{min}}}{E_{\text{avg}}} <= \varepsilon$

The optimize collection of the network parameters is figure out.

At the beginning

The initial species $P_1, P_2, ..., P_L$ are determined in accordance to the parameters of the value methods mentioned above, and the value of $J$ is set as 0.

The fitness value $E$ is calculated and the result is chosen the fitness function value of the smaller.

The higher fitness individuals is select to enter the next generation after crossover and mutation operation.

$\frac{E_{\text{max}} - E_{\text{min}}}{E_{\text{avg}}} <= \varepsilon$

The optimize collection of the network parameters is figure out.

The end of training network

Figure 2. Network training process

IV. CALCULATION EXAMPLES AND ANALYSIS

A. Application of the model

In order to test the forecasting validity of the model presented in the paper, it will be applied to runoff forecasting for a reservoir located in Southwest China. We will choose the daily average flux as statistical indicators for the analysis and collect the reservoir’s daily average runoff flux data in year 2002 ~ 2006, in which data of year 2002 ~ 2005 will be used as training samples while data of year 2006 will be taken as the samples to be tested.

According to the ideas proposed previously, corresponding calculation program is worked out, which calculation steps are as follow:

Step 1: Classify the training samples by FCM as four season types. And obtain the cluster center of each season type’s runoff samples. Based on the cluster center, it is judged that cluster center $Q_{\text{avg}} < V_1$ signifies dry season, $Q_{\text{avg}} > V_4$ signifies wet season, $V_2 \leq Q_{\text{avg}} \leq V_4$ signifies usual season (hereinto $V_2 \leq Q_{\text{avg}} < V_3$ is called the usual season I while $V_3 \leq Q_{\text{avg}} \leq V_4$ is called the usual season II). $V_1, V_2, V_3, V_4$ signify the samples’ cluster center o figured out by formula (4):

Step 2: Respectively build up the partial forecasting model for each season type and make mode identification toward each partial model. According to the formation mechanism of runoff, the flux of the runoff forecasting period is mainly affected by the rainfall of t periods and runoff flux [10]. Therefore, the nodes of network’s input layer are determined by the runoff correlation analysis between period t and period t-k. The input mode of each model is as follow: $Q(t-1), ..., Q(t-k)$ means runoff flux during the day before forecasting to the foregoing k days. $Q(t)$ means the runoff output quantity of the forecasting day. Integrated optimization [11] method is adopted for determining the number of hidden layer’s nodes. The structure of partial forecasting model corresponding to each of the four season types is demonstrated in table 1.

**TABLE I STRUCTURE OF PARTIAL FORECASTING MODEL CORRESPONDING TO EACH OF THE FOUR SEASON TYPES**

<table>
<thead>
<tr>
<th>Season Types</th>
<th>Input Mode</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry Season $\rightarrow$ CWNNR1</td>
<td>$Q(t-1), ..., Q(t-3)$</td>
<td>$Q(t)$</td>
</tr>
<tr>
<td>Usual Season I $\rightarrow$ CWNNR2</td>
<td>$Q(t-1), ..., Q(t-4)$</td>
<td>$Q(t)$</td>
</tr>
<tr>
<td>Usual Season II $\rightarrow$ CWNNR3</td>
<td>$Q(t-1), ..., Q(t-5)$</td>
<td>$Q(t)$</td>
</tr>
<tr>
<td>Wet Season $\rightarrow$ CWNNR4</td>
<td>$Q(t-1), ..., Q(t-7)$</td>
<td>$Q(t)$</td>
</tr>
</tbody>
</table>

Step 3: Input the sample value of the model. Based on the Step 1, identify which season type it belongs to. Select the corresponding partial forecasting model. Input the sample vector into the network model to make emulational simulation. The output result is the forecasting result of the forecasting object $Q(x)$.

B. Analysis to the forecasting result

For making analysis to the forecasting result, choose the runoff measured data of year 2006 as the samples to be tested. Adopt the Nash Efficiency Coefficient to evaluate the accuracy of the model: expressed by $R^2($%), peak relative error: expressed by $RP($%) and the relative error of total runoff amount: expressed by $RE($%), shown as formula below:

$$R^2 = \frac{\sum_{i=1}^{N} (Q(t) - \bar{Q}(t))^2}{\sum_{i=1}^{N} (Q(t) - \bar{Q}(t))^2} \times 100\%$$ (12)
It is found out from the table 2 used for evaluating model accuracy that both the flood peak and flood quantity of wet season are of the highest relative error, which RP and RE are 4.3% and 6.8% respectively; both the flood peak and flood season are of the lowest relative error, which RP and RE are 10.4% and 10.6%. With regard to the model efficiency coefficient R2, 88.5% of the wet season is the best one while 80.7% of the dry season is the worst one. R2 of several times of flood simulations are above 80%, RP and RE is controlled to less than 11%, which validates the forecasting results can meet our satisfaction.

Then, we compare the simple wavelet neural network model(WNN) with the integrated model, their forecasting performance are listed in table 3.

The data in table 3 reveal that the performances of different forecasting models established by different network forecasting methods are quite dissimilar when the daily runoff forecasting is done to the same reservoir. Various estimation indexes of the integrated model proposed by this paper are obviously better than that of the simple forecasting model of wavelet neural network. Model efficiency coefficient, flood peak relative error and total runoff relative error of the integrated model are 85.7%, 6.82% and 8.55%, which excels 80.2%, 13.4% and 16.3% from the simple wavelet neural network model. Moreover, the average training time of the integrated model is shorter than that of the simple wavelet neural network model, which is good for greatly reducing network training time.

### TABLE II VERIFICATION RESULT OF DIFFERENT NEURAL NETWORK MODELS TO PREDICT RUN-OF

<table>
<thead>
<tr>
<th>Season types</th>
<th>Corresponding model</th>
<th>R2(%)</th>
<th>RP(%)</th>
<th>RE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry Season</td>
<td>CWNRR1</td>
<td>85.7</td>
<td>8.4</td>
<td>8.6</td>
</tr>
<tr>
<td>Usual Season I</td>
<td>CWNRR2</td>
<td>88.2</td>
<td>6.4</td>
<td>7.6</td>
</tr>
<tr>
<td>Usual Season II</td>
<td>CWNRR3</td>
<td>90.4</td>
<td>6.2</td>
<td>7.2</td>
</tr>
<tr>
<td>Wet Season</td>
<td>CWNRR4</td>
<td>88.5</td>
<td>4.3</td>
<td>6.8</td>
</tr>
<tr>
<td>Integrated Model</td>
<td></td>
<td>88.2</td>
<td>6.32</td>
<td>7.55</td>
</tr>
</tbody>
</table>

### TABLE III FORECASTING PERFORMANCE COMPARISON

<table>
<thead>
<tr>
<th>Models</th>
<th>Training time(s)</th>
<th>R2(%)</th>
<th>RP(%)</th>
<th>RE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNN</td>
<td>237</td>
<td>80.2</td>
<td>13.4</td>
<td>16.3</td>
</tr>
<tr>
<td>Integrated model</td>
<td>112</td>
<td>85.7</td>
<td>6.32</td>
<td>7.55</td>
</tr>
</tbody>
</table>

V. SUMMARY

Considering various seasons differs greatly in runoff distribution, this paper studied the method of utilizing fuzzy C-means clustering analysis combined with wavelet neural network to build runoff forecasting model, which is proved to be feasible through testing. Furthermore, the comparison on forecasting results of the proposed model and the simple wavelet neural network model reach the following conclusions: the established integrated model has distinct improvement on enhancing forecasting accuracy and reducing training time.

The above-mentioned study found that in the process of continuous runoff forecasting, the characteristic of various seasons differs greatly in runoff distribution should be taken into full account, the data of similar formation cause and relatively means should be extracted from the mass data, so as to build the runoff forecasting model of better targeted and better reflect any phase’s runoff characteristics.

ACKNOWLEDGMENTS

This research is sponsored by the Humanities and Social Science Foundation of the Ministry of Education of China (Grant No. 14YJC630211) and the National Major Science and Technology Research Project (Grant No. 2013ZX07304 - 003).

REFERENCES