PROJECTION PURSUIT REGRESSION MODEL BASED ON REAL-CODED GENETIC ALGORITHM FOR FLOOD FORECASTION

[1]YU-FENG LIANG, [2*]HONG-WEI ZHOU
[1]Ph. D. Student, College of Water Resources & Hydropower, Sichuan University, Chengdu 610065
[2*]Lecture, State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu 610065
E-mail: [1]371289809@qq.com, [2]64821321@qq.com

ABSTRACT

Combining the advantages of genetic algorithm (GA) and projection pursuit regression (PPR), this article firstly uses improved Hermit polynomial as the ridge function of projection pursuit regression model. And then adopted the real-coded genetic algorithm to optimize the projection direction, a forecasting model for peak flow of short flood forecasting is presented. Applied the presented model to forecasting the flood of the Wujiang River at Wulong station, and compared with the BP neural network method. Computing results show that, the presented model has a strong advantage of dimensionality reduction adaptability than BP network in dealing with the poor fitting data of one dimension space, the forecasting accuracy is improved, and can be applied in hydrological simulation and forecasting.

Keywords: *Projection Pursuit Regression, Genetic Algorithm, Neural Networks, Flood Forecasting*

1. INTRODUCTION

Disaster of flood has a huge impact on the safety of people's lives and property, reasonable prediction of flood significant help in the formulation of disaster prevention and mitigation measures for flood [1]. Because the flood forecasting is a complicate problem, the prediction precision is very important [2]. In recent years, with the rapid development of science and technology and the arrival of information age, artificial intelligence model has been widely applied in various fields, and has achieved good results for several areas, especially in the flood forecasting problem [3].

There are lots of forecasting models for river flood, in these models, the projection pursuit regression model is widely used [4]-[6]. It has some advantages in flood forecasting [7]. Some researchers applied projection pursuit regression model in flood forecasting and achieved good results in Zipingpu reach, and then present a new method of parameters projection regression [8]. Artificial neural network (ANN) has obvious advantages in simulation and prediction of the nonlinear system, especially BP (Back Propagation) neural network had aroused much attention of researchers [9]-[11]. This paper applied projection pursuit regression in Wujiang River and combined with its river features, uses Hermit polynomial as the ridge function of projection pursuit regression model, and adopt real-coded genetic algorithm to optimize the projection direction [12], and compares the result with more mature BP neural network on condition of MATLAB in discussion of result reasonability, and demonstrates that dimensionality reduction advantage of projection pursuit regression can better excavate data sample in dealing data and is more adaptable in fitting poor data and more reasonable in forecasting results.

2. PROJECTION PURSUIT REGRESSION MODEL

2.1. Theory And Algorithms

Select P projection directions, and seek projection pursuit model between each projection direction and variable *y* which can be expressed as:

$$y=\sum_{k=1}^{p}\beta_k f_k\left(\sum_{j=1}^{m}a_{kj}x_j\right)+\varepsilon \qquad (1)$$

where a_{kj} is the k_{th} projection direction in m dimensional spaces, $\sum_{j=1}^{m}a_{k\,j}x_j$ is projection amount of the k_{th} projection direction of observation vector $(x_1, x_2, \ldots, x_{\text{m}})$; f_k is projection function of the k_{th} projection direction, which is usually called ridge function, and reflects relationship between the k_{th}

projection amount or characteristics and the dependent variable y; β_k is weight that the k_{th} ridge function f_k shares in relationship between the y and the factor (x_1, x_2, …, x_m).

According to principles of projection pursuit, seeking method of p best projection directions, p ridge functions and p weight coefficients constitutes the PP regression algorithm. The algorithm is as follows:

(1) Set n group observation data y_i (x_{i1}, x_{i2}, …, x_{im}); Choose an initial projection direction $a_1 = (a_{11}$, a_{12}, ..., $a_{1m})$, project the observation vector group in the direction a_1, and get n one-dimensional projection data in the direction a1, which is

$$\omega_i = \sum^{m} a_{1i} x_y, (i = 1,2\cdots,n) \tag{2}$$

(2) Fit the objective function: $\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left[z_i - f_1(\sum_{i=1}^{m} a_{1j}x_{ij})\right]^2$, find ridge function f_1 of the minimum objective function by optimization algorithm.

(3) Change the projection direction, repeat (1) and (2), until you find the best projection direction a_1 and the best ridge functions, and test whether model accuracy meets the requirements, if model accuracy meets the requirements and calculation terminates; otherwise, add the second projection direction a_2, and start the next step calculation.

(4) Calculate residual sequence: $y^{(1)} = \{y_i^{(1)} \mid i = 1,2,\cdots,n\}$, where $y^{(1)}$ is optimization residual variables of the first step, and $y^{(1)}$ (i=1,2,…,n) is the corresponding residual observations.

$$y^{(1)} = y - f_1(\sum_{j=1}^{m} a_{1j} x_j) \tag{3}$$

(5) Use y_1 to instead of y, and establish relationships between y and the factor observation data

$$y^{(1)} = f_2(\sum_{j=1}^{m} a_{2j} x_j) \tag{4}$$

Repeat Eqs. (1)-(4) until you find the second best projection direction a_2 and the best ridge function.

(6) Use the least square method to calculate weight by $y = \sum_{k=1}^{2}\beta_k f_k(\sum_{j=1}^{m} a_{kj} x_j) + \varepsilon$.

After optimizing β_1 and β_2 , also need to check whether model accuracy meets the requirements, if model accuracy meets the requirements and calculation terminates; Otherwise, add the third projection direction, and start the next step calculation.

(7) Suppose to have optimized the best projection direction of the first λ , if need to increase the projection direction, then calculate residual sequence of the next step, $y^{(\lambda)} = \{y_i^{(\lambda)}\}$ (i=1, 2, …, n). Use $y^{(\lambda)}$ to instead of y and find the best direction and best ridge function of the first λ+1 by

$$y^{(\lambda)} = f_{\lambda+1}(\sum_{j=1}^{m} a_{\lambda+1,j} x_j) \, . \tag{5}$$

(8) When the above steps optimize P the best projection directions, the best ridge functions and the weight coefficient, if the model accuracy meets the requirements, and terminate the calculation.

Above principle and algorithm of projection pursuit regression model determine a steady projection pursuit regression model which should have two basic requirements, namely, efficient optimization algorithms and appropriate ridge functions [6].

2.2. Ridge Function Selection

At present, there are many types of ridge functions to select. This article selects polynomial which has a composition of Hermite function to approximate non-linear ridge functions, whose mathematical expression of R order Hermite function is:

$$h_y(\omega) = (y!)^{-\frac{1}{2}}\pi^{\frac{1}{4}}2^{-\frac{y-1}{2}}H_y(\omega)\varphi(\omega) \, , \quad -\infty < \omega < +\infty \tag{6}$$

where y is order of Hermite function; ω is one dimensional variable; PP regression model $\omega = (\sum_{j=1}^{m} a_j x_j)$; φ is standard Gaussian function; $H_y(\omega)$ is defined by the recurrence relation:

$$\begin{aligned} &H_0(\omega) = 1, H_1(\omega) = 2\omega, \\ &H_y(\omega) = [\omega H_{y-1}(\omega) - (y-1)H_{y-2}(\omega)] \end{aligned} \tag{7}$$

R order Hermite polynomial can be expressed as:

$$g(\omega)=\sum_{y=0}^{R} c_y h_y(\omega) \tag{8}$$

where c_y is polynomial coefficient, use R order Hermite polynomials to approximate the first k ridge, and get

$$y=\sum_{k=1}^{p}\beta_k f_k(\sum_{j=1}^{m} a_{knj}x_j)+\varepsilon \tag{9}$$

Then transform into

$$\begin{aligned} y &= \sum_{k=1}^{p}\beta_k[\sum_{y=0}^{R} C_{k,y}h_{k,y}(\omega)] \\ &= \sum_{k=1}^{p}\beta_k[\sum_{y=0}^{R} C_{k,y}h_{k,y}(\sum_{j=0}^{m} a_{k,j}x_j)] \end{aligned} \tag{10}$$

2.3. Real-Coded Genetic Algorithm

This article uses real-coded genetic algorithm [7] and its implementation process is as follows:

(1) Coded. Encode a real number x into a real in [0, 1], u is a real number in [0, 1].

(2) The parent group initialization. Set that population size is np, generate np random numbers $\{u_i \mid i=1,2,\cdots np\}$ in [0, 1], substitute each random number u_i into $x=a+u(b-a)$ to get np original variables values $\{x^{(i)} \mid i=1,2,\cdots np\}$, record whose corresponding value u_i into $y^{(i)}$, and put them as initial parent groups.

(3) Fitness evaluation of parent individual. First, calculate the fitness value of each individual, second, calculate the selection probability:

$$p(x^{(i)})=\frac{F'(x^{(i)})}{\sum_{i=1}^{np}F'(x^{(i)})} \tag{11}$$

(4) Selection of parent individual. Let:

$$p_1=\sum_{k=1}^{i} p(x^{(i)});(i=1,2,\cdots,np) \tag{12}$$

Its sequence $\{p_i \mid i=1,2,\cdots np\}$ divides [0, 1] into np subintervals, these intervals correspond np parent individuals $y^{(i)};(i=1,2,\cdots,np)$, and generate np —5 random numbers $\{u_k \mid k=1,2,\cdots np-5\}$ in [0, 1]. If u_k locates in $[p_{i-1},p_i]$ ($u_k\in[p_{i-1},p_i]$), the first k individual is selected, and select np —5 individuals in all. Meanwhile, select five numbers of the largest probability to directly join into selected individual set, which is so called immigration operation, namely, excellent individuals immigrate into new groups directly, and new parent groups are called contemporary parent groups.

(5) Hybrid of parent group. Gene is each real number corresponding to the binary bits for binary encoding, individual is directly seen as single gene individual for decimal encoding, whose hybrid operation follow the following methods:

Randomly pair np contemporary parent groups of the choice, denote $(y^{(i_1)},y^{(i_2)})$, generate three uniform random numbers in [0, 1], and generate two offspring individuals by random linear combination as follows:

$$\left.\begin{array}{l} y^{(k_1)}=u_1y^{(i_1)}+(1-u_1)y^{(i_2)},u_3<0.5 \\ y^{(k_2)}=u_2y^{(i_1)}+(1-u_2)y^{(i_2)},u_3\geq 0.5 \end{array}\right\} \tag{13}$$

(6) Offspring individual mutation. Hybrid operation produces offspring which mutates on the mutation probability. Let mutation probability is $pm_k=1-p(x^{(k)})$, and generate np random numbers $\{u_k \mid k=1,2,\cdots np-5\}$ in [0, 1]; mutation operation according to the following formula:

$$\left.\begin{array}{l} y^{(k)}=u_k,u_k\leq pm_k \\ y^{(k)}=y^{(k)},u_k>pm_k \end{array}\right\} \tag{14}$$

Thus, mutation probability of offspring individual resulting from hybrid is the smaller, whose mutation possibility is the smaller. Because selection probability $p_k(x^{(k)})$ of good parent individual is large, whose mutation probability of corresponding offspring is small, therefore, good parent individual genes are retained.

(7) Get 3np offspring individuals by evolution iteration from step three to step six, rank them according to their fitness values in descending order, and take get np offspring individuals in the front as new parent groups. Algorithm turns to step three, start the next round evolution, and repeat evolution iteration until the iteration meets the accuracy or preset iteration steps [8].

3. BP NEURAL NETWORK BASED ON MATLAB

MATLAB is commercial mathematical software developed by Mathworks Company of USA, which is applied for algorithm development, data visualization, data analysis, high-level technical computing language of numerical calculation and interactive environment. It is widely used and much loved software environment, whose basic data element is matrix, provides a variety of matrix operations and operating, and has strong graphics capabilities.

Neural network toolbox is one of the toolbox, developed under the MATLAB environment. It is based on artificial neural network theory and uses MATLAB language to construct typical activation function of neural network, so that output calculation of the selected network changes into activation function call. In addition, using MATLAB, according to various rules of typical amendment network weight and the network training process, compile a variety of network weight training subprogram, which can be called directly and increase computing efficiency and quality [6].

4. THE EXAMPLE ANALYSIS

4.1. Overview Of The Study Watershed

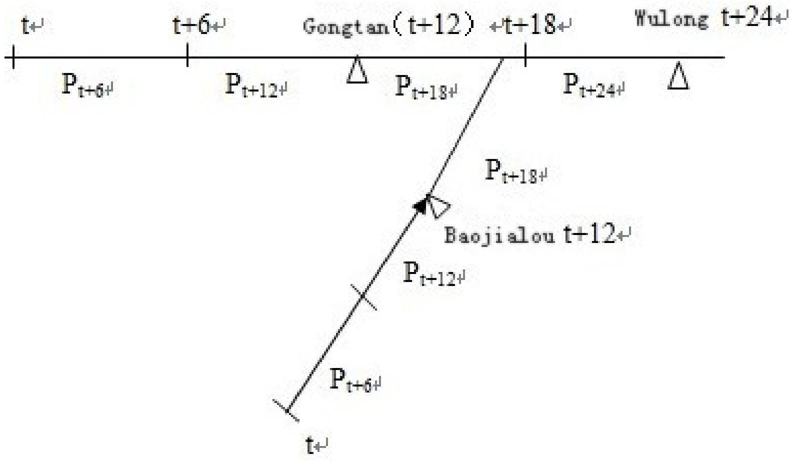

Figure 1: Schematic Of The Study River Section

Wujiang River watershed system is distributed in feather, whose river network density is large and has the 58 first level tributaries. Among them, the watershed area more than 300 km^2 has 42 tributaries; more than 1000 km^2 has 16 tributaries. This article combines characteristics and law of floods propagation of Wujiang River (Figure 1), and respectively establishes projection pursuit regression model and BP network model, and does the flood forecasting simulation of Wulong station.

4.2. Projection Pursuit Regression

According to the characteristics of Wujiang River, with τ as predictable period, because runoff of Baojialou tributary has a great impact on the flow of Wulong station, therefore, after comprehensive analysis of flood peak propagation time of each station, put Baojialou, Gongtan, Wulong stations as affecting forecast factors, and foresee period is scheduled to 6h.

To simulate each stations of Wujiang River from August 1, 2008 to September 30, 2008, and forecasting results are shown in Figure 2.

As shown in Figure 2, total number of model samples is 488, qualified samples is 431, qualified rate is 88.31%, relative error is 11.69%, and forecast accuracy is class A according to L250-2000 "hydrological information forecasting standards".

4.3. BP Neural Network

According to the actual characteristics of Wujiang River, also considering Baojialou, Gongtan, and Wulong stations as basic input of network training, namely, the flow of export cross section of each stations from time to t-n, establish three-layer BP neural network. Using MATLAB neural network toolbox to forecast flow data from August 1, 2008 to September 30, 2008, and calculation results is shown in Figure 3.

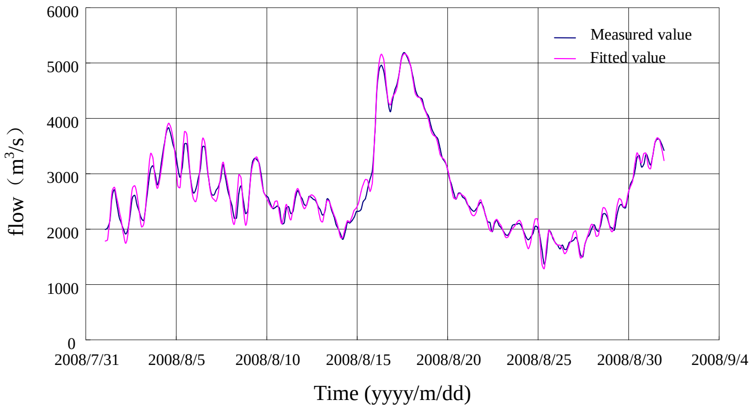

Figure 2: Simulation Figure Of Projection Pursuit Model Of Wulong Station

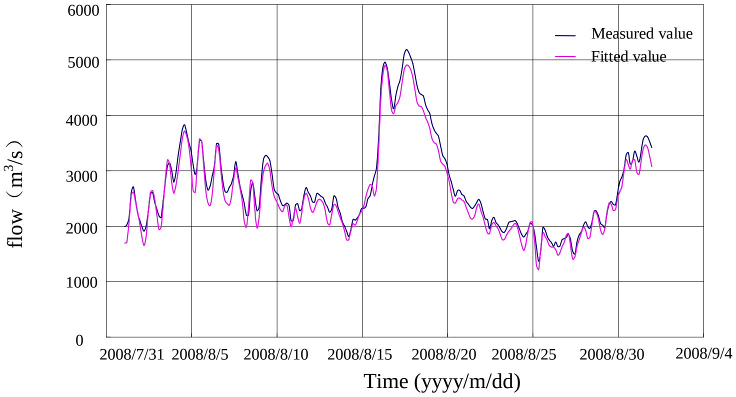

Figure 3: Simulation Figure Of BP Neural Network Model

We can see from figure 3: total number of model samples is 488, qualified samples is 391, qualified rate is 80.12%, relative error is 19.88%, and forecast accuracy is class according to L250-2000 "hydrological information forecasting standards".

4.4. Comparison of forecast results

Comparison of forecast results and accuracy of simulating flood peak based on two kinds of intelligent models, projection pursuit regression of genetic algorithm (GA-PPR) and BP neural network of MATLAB (BP-ANN), as shown in Table1 and Table2.

Table 1: Comparison Of Forecast Results And Accuracy Of Two Kinds Of Intelligent Models

Time	Measured peak (m^3/s)	PPR Fitting peak (m^3/s)	Relative error (%)	Peak current error (h)	BP Fitting peak (m^3/s)	Relative error (%)	Peak current error (h)
2009/8/14 2:00	2246	2133	5.0%	0	2197	2.2%	0
2009/8/14 23:00	2519	2316	8.0%	1	2408	4.4%	0
2009/9/21 23:00	2639	2735	3.6%	0	2691	1.9%	0
2009/9/22 2:00	2699	2786	3.2%	0	2698	0%	0
2009/9/22 5:00	2512	2431	3.2%	0	2389	4.9%	0

Table 2: Comparison Of Forecast Results And Accuracy Of Two Kinds Of Intelligent Models

Total number of samples	PPR Qualified samples	PPR Qualified rate (%)	PPR Relative error (%)	BP Qualified samples	BP Qualified rate (%)	BP Relative error (%)
488	431	88.31%	11.69%	391	80.12%	19.88%

As shown in Table 1 and Table 2, two kinds of models, GA-PPR and BP-ANN, both of them can better reflect the degree of peak simulation and meet accuracy requirement of L250-2000 "hydrological information forecasting standards". Although accuracy of flood peak simulation based on projection pursuit model is slightly lower than accuracy based on BP network model, they are all within random permission errors.

5. CONCLUSIONS

Calculation and analysis by example suggest that two kinds of models, GA-PPR and BP-ANN, better meet forecast requirements for non-linear regression, but projection pursuit regression model having a strong advantage of dimensionality reduction can project input variables in one dimension space and effectively reflect excavation of sample information; Meanwhile, the number of ridge function of projection pursuit model is easily identifiable than hidden layers and nodes of BP network in difficult aspects of the model constitute, therefore, in terms of relative average error and efficiency, we can get even better forecast accuracy. Overall, the projection pursuit regression model has better adaptability than BP network in dealing with poor fitting data of one dimension space.

In addition, flood is random, uncertainties and complex, which makes data sample be poor representative in the part in the forecast process, often results in low accuracy of rate model and error too large of forecast result in simulating and fitting data, and this is also common defects of black box model.

ACKNOWLEGEMENTS

This work is supported by the Sichuan University Youth Foundation (2012SCU11063).

REFRENCES:

[1] H.Y. Li, J. Zhao, A. Wang, Z. Han, and Y.X. Wang, “Muskingum parameter optimization through extension field search genetic algorithm and its application”, *Journal of Jilin University*, Vol. 41, No. 3, 2011, pp. 861-865.

[2] S. Maskey, V. Guinot, and R.K. Price, “Propagation of precipitation uncertainty through a flood forecasting model”, IAHS-AISH, Vol. 282, No. 3, 2003, pp. 93-100.

[3] M. Nasseri, K. Asghari, and M.J. Abedini, “Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network”, *Expert Systems with Applications*, Vol. 35, No. 3, 2008, pp. 1415-1421.

[4] J.W. Zhou, W.Y. Xu, X.G. Yang, C. Shi, and Z.H. Yang, “The 28 October 1996 landslide and analysis of the stability of the current Huashiban slope at the Liangjiaren Hydropower Station, Southwest China”, *Engineering Geology*, Vol. 114, No. 1-2, 2010, pp. 45-56.

[5] S.K. Tasoulis, M.G. Epitropakis, D.K. Tasoulis, and V.P. Plagianakos, “Density based projection pursuit clustering” *2012 IEEE Congress on Evolutionary Computation*, 2012, 10.1109/CEC.2012.6253006.

[6] C.X. Jia, “Analysis of Transient laminar flow in Swro with hybrid membrane channels”, *Journal of Theoretical and Applied Information Technology*, Vol. 45, No. 2, 2012, pp. 491-501.

[7] Y. Su, S. Shan, X. Chen, and W. Gao, “Classifiability-based discriminatory projection pursuit”, *IEEE Transactions on Neural Networks*, Vol. 22, No. 12, 2011, pp. 2050-2061.

[8] J. Touboul, “Projection pursuit through relative entropy minimization”, *Communications in Statistics: Simulation and Computation*, Vol. 40, No. 6, 2011, pp. 854-878.

[9] F.G. Xu, H.W. Zhou, J.W. Zhou, and X.G. Yang, “A Mathematical Model for Forecasting the Dam-Break Flood Routing Process of a Landslide Dam”, *Mathematical Problems in Engineering*, Article ID 139642, 2012, 16 pages, doi:10.1155/2012/139642.

[10] S.L. Marie-Sainte, A. Berro, and A. Ruiz-Gazen, “An efficient optimization method for revealing local optima of projection pursuit indices” *Lecture Notes in Computer Science*, Vol. 623, 2010, pp. 60-71.

[11]G.V. Nadiammai, and M. Hemalatha, “An enhanced rule approach for network intrusion detection using efficient data adapted decision tree algorithm”, *Journal of Theoretical and Applied Information Technology*, Vol. 47, No. 2, 2013, pp. 426-433.

[12] P. Bousfield, C. Zanottelli, F. Lapolli, S. Schossland, C. Ganske, and L. Chapuis, “Development of a flood forecasting system in hydrographic basins by means of Artificial Neural Networks”, *Proceedings of the IADIS International Conference Intelligent Systems and Agents*, 2010. pp. 114-118.