Changes in Water Resources Systems: Methodologies to Maintain Water Security and Ensure Integrated Management (Proceedings of Symposium HS3006 at IUGG2007, Perugia, July 2007). IAHS Publ. 315, 2007.

Flood forecasting by Coupling Cluster method and Artificial Neural Networks

YIN XIONGRUI[1], **XIA JUN**[2] **& ZHANG XIANG**[1]

1 *State Key Laboratory of Water Resources and Hydropower Engineering Sciences,*
Wuhan University, Wuhan 430072, China
aw_yin@163.com

2 *Key Laboratory of Water Cycle and Related land Surface Processes, IGSNRR,*
Chinese Academy of Sciences, Beijing 100101,China

Abstract Flood forecasting takes a vital role for flood control and water resources management of catchments. However, it is generally accepted that the relationship of rainfall and runoff is highly complicated, and for a basin, the underlying mechanisms of streamflow generation in rain periods are quite different from those in non-rain periods. The flow hydrograph is broken into several segments, then the rainfall–runoff relationship is separately established. In this study, we employ two methods to divide the flow hydrograph into several segments. One is the Fuzzy C Means (FCM) method, and the other is the Self-Organizing Feature Map (SOFM). Based on the two clustering results, multi-layer Feedforward Networks (MFN) was used to simulate the rainfall–runoff relationship of each segment. In this way two hybrid artificial neural networks (FCMMFN and SOMMFN) are established. The methods mentioned above are applied to Wangjiachang Reservoir inflow forecasting, in the Hunan province of China, for three-hour-ahead flood forecasting. Forty-five historical flood processes from 11 years (1982–1992) are applied for calibration whilst 14 flood processes from 3 years (1994–1996) are utilized for validation. Antecedent precipitation and streamflow data is input into FCM and SOFM for flow hydrograph decomposing and clustering. The result shows that FCM and SOFM are both able to find the potential knowledge of flow, and that it is easy to find that flow hydrographs as corresponding output is classified into four different stages: (1) low flow; (2) rising flow; (3) flood peak; and (4) recession. Then, for each segment, a MFN is applied to simulate its rainfall–runoff relationship. Results show FCMMFN and SOMMFN are both superior to MFN, i.e. the two hybrid models can simulate precisely the rainfall–runoff relationship simultaneity in low flow, middle flow and high flow. Moreover, FCMMFN and SOMMFN are investigated and compared, and FCMMFN appears to be better.

Key words Artificial Neural Networks; flood forecasting; Fuzzy C Mean; Self-Organizing Feature Map

INTRODUCTION

It is generally accepted that the relationship between rainfall and runoff is highly complicated, and is usually considered to be nonlinear and seasonal. The responses of discharge in different rain periods show various behaviours, because the underlying mechanisms of runoff generation are probably different in low, middle and high flow stages. Therefore, we are probably able to divide the flow hydrograph into a few segments, and each segment is expected to represent one kind of the mechanisms of streamflow generation. Hence, flood forecasting models will be likely to be built based on each clustering of data. Much literature has been found to simulate the rainfall–runoff relationship based on dividing the complex relationship into a simple one via some techniques, such as the threshold method (Wang & Huang, 2002; Jian & Srinivasulu, 2006), self-organized networks (SOM) (See & Openshaw, 1999; Abrahart & See, 2000), and others (Hsu *et al*., 1995).

In this paper, the Fuzzy C Means clustering method and Self-Organizing Feature Map clustering method are both employed to break the flow hydrograph into several segments, then two hybrid artificial neural networks (FCMMFN and SOMMFN), based on Fuzzy C Means and Self-Organizing Feature Map separately, are applied to simulate the rainfall–runoff relationship. The performance of the model is compared with the single Multi-layer Feedforward Network (MFN).

METHODS

Multi-layer Feedforward Network (MFN)

The Multi-layer Feedforward Network, applied widely to hydrology since the early 1990s, usually consists of three layers, namely input layer, hidden layer and output layer. There is as yet no systematic way to establish a suitable architecture, and the selection of the appropriate number of neurons is basically problem specific, too. The learning process of MFN is iterative and optimizes its parameters by minimizing an object function. The back-propagation algorithm (BP) used in this study, is a learning algorithm for a multilayered neural network in which the weights are modified via propagation of an error gradient backward from the output to the input. The BP algorithm and Multi-layer Feedforward Network is applied most extensively in the field, so we will not describe the structure of this network and the training process in detail, and a specific description can be easily found in the literature, including the references below.

Fuzzy C means

Fuzzy C means, which is well known as ISODATA, is a method based on Fuzzy theory for clustering. FCM divides *n* vectors to *c* groups, and the category each vector belongs to can be determined by comparing the membership degrees. The centre of each group can be obtained via minimizing the object function. For one vector, the sum of total membership degree is 1, this can be expressed as:

$$\sum_{i=1}^{c} u_{ij} = 1, j = 1, \cdots, n \tag{1}$$

Hence, the cost function (or objective function) is:

$$J(U, c_1, \cdots c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \tag{2}$$

where $u_{ij} \in [0,1]$ is the membership degree of the *j*th vector belonged to the *i*th group, and c_i is the fuzzy centre of the *i*th group. $d_{ij} = \left\| c_i - x_j \right\|$ is the Euclidean distance between the *i*th group centre and the *j*th vector. $m \in [1,\infty)$ is a weight index.
The new function can be described as follows:

$$\overline{J}(U,c_1,\cdots,c_c,\lambda_1,\cdots,\lambda_n) = J(U,c_1,\cdots,c_c) + \sum\nolimits_{j=1}^{n} \lambda_j (\sum_{j=1}^{c} u_{ij} - 1)$$

$$= \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} + \sum\nolimits_{j=1}^{n} \lambda_j (\sum_{j=1}^{c} u_{ij} - 1) \tag{3}$$

where, λ_j, $j = 1, \ldots, n$ is the Lagrange-multiplier with n constraints of equation (1). To obtain the minimum value of (2), calculate the partial derivatives to every parameter and make them zeros, then the necessary condition can be express as:

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{4}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{dkj})^{2/(m-1)}} \tag{5}$$

The simple iterative process of the FCM algorithm is: (a) Initialize membership degree vector *U* by randomly generating between 0 and 1, and be sure to satisfy constraint (1). (b) *U* is used to calculate c_i according equation (4); (c) the cost function value can be obtained from equation (2). If this value or its change compared to the previous one is lower than the threshold, then stop. (d) Use equation (5) to calculate the new *U*, and return to (b).

Self-Organized Feature Map (SOFM)

The SOFM method, originally proposed by Prof. Teuvo Kohonen in 1981, is a type of artificial neural networks, and is used for projecting patterns from high-dimensional to low-dimensional spaces (most commonly 2-D). SOFM is an unsupervised classification, used to cluster data set based on statistics. SOFM can adjust the weight vectors of adjacent units in the competitive layer to a similar vector by competitive learning and to approximate the distribution of the target pattern. The neurons of the competitive layer are arranged in a lattice and are connected to all the inputs.

Suppose that vector $X = (x_1, x_2, \cdots x_n)^T \in R^n$ is input, and is connected with every neuron in the competitive layer in the same manner, the weight vector of the *j*th neuron is denoted by $W_j = [w_{j1}, w_{j2}, \cdots w_{jn}]^T \in R^n$. The unsupervised training of SOFM is surmised as follows:

Initialize randomly the weight vectors for each SOFM connection weight.

For the weight vector W_j, we treat each input X_k as follows:

- Calculate the Euclidean distance between the *k*th sample X_k and W_j.

$$d_j = [\sum_{i=1}^{n} (x_i^k - W_j)^2]^{1/2} \quad (j = 1,2\cdots p) \tag{6}$$

- The neuron N_j^* with the minimum distance will be the winner.

$$d_j^* = \min_{j=1,p}\{d_j\} \tag{7}$$

- Adjust the weights of neuron N_j^* as well as the neurons in its geometry neighborhood $Nc_j^*(t)$.

$$\Delta W_{ij} = \eta(t)[x_i^k - W_{ij}] \tag{8}$$

  where t is the current iteration of learning, $N_j^* \in Nc_j^*(t)$, $j = 1,2,\cdots,p$, $i = 1,2,\cdots,n$. $\eta(t)$ is learning rate, which decreases as the learning course. Along with process continues, the adjustment scope of weights gradually reduces, thus leads to the weight vector of competition-to-win neuron represents the essential attribute of some kind of specific pattern.
- Change the input sample and train again, take $t = t + 1$, return to step 2 and calculate repeatedly until all samples are input.

Hybrid models

FCM and SOMF divide up the streamflow into several different segments, then for each segment, a MFN is applied to simulate its rainfall–runoff relationship. When performing flood forecasting with new input, input data is clustered by classifier FCM or SOMF first, then the MFN linked to that cluster is chosen for discharge forecasting. Figure 1 is the structure of the hybrid model.

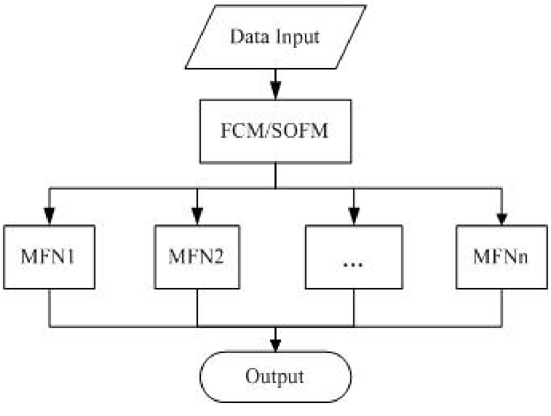

Fig. 1 ANN model based on classification.

APPLICATION OF FCMMFN, SOMMFN AND COMPARISON

The proposed hybrid networks methods for forecasting discharge is applied to the Wangjiachang Reservoir in Hunan province, China. Wangjiachang Reservoir, located in the middle reaches of the Cen River, is a large-scale reservoir. The drainage area is 484 km^2, which is 41% of the whole basin area. The basin is located in a subtropical monsoon zone with rich rainfall and good vegetation cover. The annual precipitation is 1280 mm. However, the temporal distribution of the rainfall during a given year is significantly heterogeneous in this area. The flood events in this area are mainly due to the thunderstorms and 80% of the total rainfall falls between April and August. Six rain-gauged stations selected in this area were used in this study. Forty-five historical floods from 11 years in the Wangjiachang Reservoir are applied for calibration, whilst 14 floods in 3 recent years were utilized for validation.

Using system models to simulate the rainfall–runoff relationship, rainfall is input and the flow hydrograph is output. The discharge in the forecasting period can be considered as the function of antecedent precipitation and antecedent discharge, which is written as:

$$Q_t = f(P_{t-n_x},\cdots,P_{t-1},P_t,Q_{t-1},Q_{t-2},\cdots,Q_{t-n_y}) \tag{9}$$

where f is an unknown nonlinear function, P represents the precipitation, Q is the current and antecedent discharge produced by antecedent precipitation. n_x is the influential interval of antecedent precipitation, n_y is the influential interval of antecedent discharge. At present, there are few effective methods to decide n_x and n_y. The way used in this study is the trial-and-error method. After a number of experiments, we find that when $n_x = 1$ and $n_y = 2$, higher performance and more precise forecasting can be obtained by MFN. So we take $n_x = 1$ and $n_y = 2$ as the influential interval precipitation and discharge separately.

The result of clustering by FCM

The input pattern of the vector in the FCM method is: $x = [P_{t-1},P_t,Q_{t-1},Q_{t-2}]^T$. Because the vector element is different in unit, normalization of each element is necessary before the vector is input to FCM. Four groups are determined to divide by applying FCM, and we can recognize four different flood behaviours from them: (1) Low flow: no rain or less rain, low discharge and the range of flow is small. That is the part of base flow mostly in the hydrograph. (2) Rising flow: rain depth is large and discharge just begins to increase. (3) Flood peak: this segment is including the peak of the hydrograph. (4) Recession: rainfall intensities have reduced considerably, and the hydrograph has begun to recede. The centre of each group is shown in Table 1, and an example of dividing the flood event by using FCM is shown in Fig. 2.

The result of clustering by SOFM

We design three kinds of structure of competitive layer for training, that is 3 × 3, 4 × 4 and 5 × 5. The input pattern, the same as FCM, is [$P_{t-1},P_t,Q_{t-1},Q_{t-2}$]. Via

Table 1 The centre of clustering using FCM.

Group	P_{t-1} (mm)	P_t (mm)	Q_{t-2} (m^3 s^{-1})	Q_{t-1} (m^3 s^{-1})
I	0.30	0.35	20.85	19.80
II	15.38	17.63	54.08	88.80
III	6.43	3.27	427.57	385.36
IV	1.74	1.42	110.82	102.08

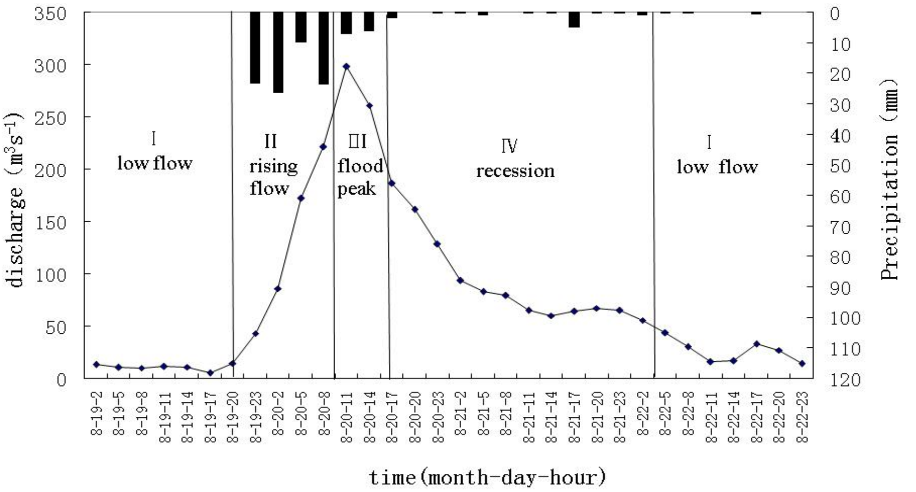

Fig. 2 An example: a flood event in 1987, divided four segments by using FCM.

preprocessing, each element in the vector will be normalized. By a number of experiments, we found 4 × 4 for the structure of competitive layer is the best. So, 16 clusters can be obtained by using SFOM.

By calculating and resampling according to the Euclidean distance, 16 clusters are also finally remerged to four groups. Groups show similar flood behaviour as the result of clustering by FCM. That is: (1) low flow; (2) rising flow; (3) flood peak; and (4) recession. The results suggest that SOFM can also distinguish the antecedent land surface situations of catchments, and intelligently divide the samples into four different behaviours.

Results by applying two neural networks

FCM and SOMF divided up the streamflow into four different regimes, so it is convenient for them to couple with MFN in each group for building FCNMFN and SOMMFN models. When forecasting, a specific MFN model is chosen, depending on which cluster the input data belongs to.

There is extensive literature on model forecasting evaluation indices. The coefficient of efficiency (CE) introduced by Nash and Sutcliffe is still one of the most widely used criteria for the assessment of model performance. CE has the form: $R^2 = 1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \overline{Q})^2}$, where Q_i is the observed value, $\overline{Q}$ is the mean value of the

Table 2 Comparisons of the performances of models.

Model	Periods	CE		RE	ER
FCMMFN	calibration	1982–1992	92.40%	19.29%	77.78%
	validation	1994	89.21%	17.21%	92.86%
		1995	93.39%	18.50%	
		1996	89.86%	10.92%	
SOMMFN	calibration	1982–1992	92.30%	20.61%	68.89%
	validation	1994	88.05%	17.60%	71.43%
		1995	93.16%	19.27%	
		1996	91.92%	10.01%	
MFN	calibration	1982–1992	92.10%	39.43%	62.22%
	validation	1994a	89.10%	32.17%	78.60%
		1995a	95.20%	35.24%	
		1996a	90.10%	14.71%	

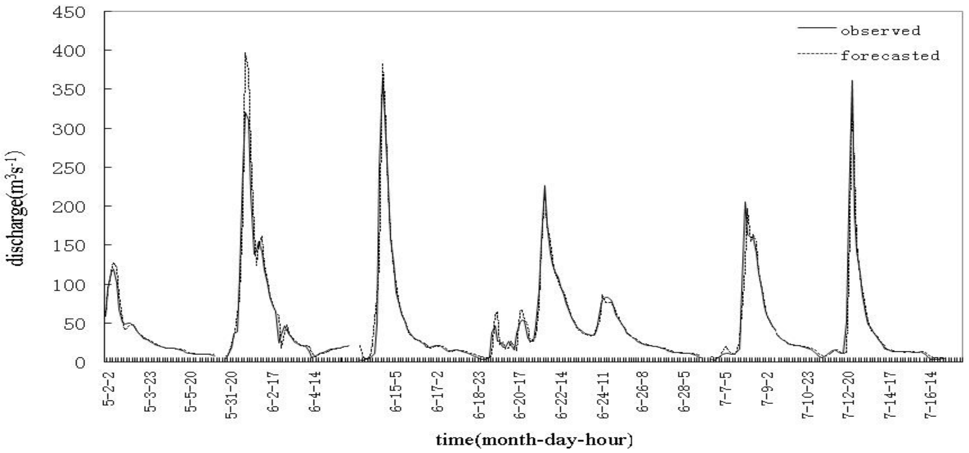

Fig. 3 Comparisons of the observed and forecasted discharge hydrograph for the year 1995 (FCMMFN)

observed data, $\hat{Q}_i$ is the predicted value. Additionally, there are other important performance evaluation indices, such as the mean relative errors of streanmflow (RE=$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Q_i-\hat{Q}_i}{Q_i}\right|$), and eligible rate (ER) relative to peak discharge and peak time. The simulated flood is eligible if absolute percentage error of peak is less than 20%. Table 2 is the comparison of the results of FCMMFN, SOMMFN and MFN.

Examination of Table 2 indicates that three models almost have the same good performance when only considering the coefficients of efficiency, which are all up to or over 90%. But as far as the results of comparison of RE and ER are concerned, the two hybrid models show better precision than single MFN, since the RE values of the FCMMFN and SOMMFN are both lower than MFN. It suggests that both FCMMFN and SOMMFN have better capacity to simulate the streamflow hydrograph, especially low flow. Furthermore, it indicates that MFN based on classification has been notably improved in the performance for flood forecasting.

The comparison of the results of the FCMMFN and the SOMMFN show small differences in the performance in CE and RE in both calibration and validation periods, but for ER, the value of ER of the FCMMFN is higher than that of SOMMFN. This result reveals that the FCMMFN has a better performance in simulating and forecasting the peaks than the SOMMFN. Figure 3 is the forecasted and observed hydrographs for the year 1995 by FCMMFN.

CONCLUSIONS

In this study, two methods for unsupervised clustering, Fuzzy C Means (FCM) and Self-Organized Feature Map (SOFM), are used to divide different flood behaviours according to the antecedent precipitation and discharge. Based on the partitioning results, the FCMMFN and the SOMMFN are built to forecast three-hour ahead discharge of the Wangjiachang Reservoir in the Hunan province of China. For the purpose of comparing the forecasting efficiency, the single multi-layer feed-forward network is selected as the baseline model. The model evaluation results indicate that FCMMFN performs best of the three models. Furthermore, the FCMMFN performs better than SOMMFN in peak forecasting. The two models show better performance than the single MFN. Comparison of the results of three models reveals that the performance of MFN can be greatly improved in simulating the flood peak and hydrograph, by preprocessing the input data effectively.

Acknowledgements The authors wish to thank the projects of the China Natural Sciences Foundation (40671035, 40671035) and Opening Research Foundation of State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (2006412011), Hohai University for their support.

REFERENCES

Abrahart, R. J. & See, L. (2000) Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrol. Processes* **14**, 2157–2172.

Hsu, K., Gupta, H. V. & Sorooshian, S. (1995) Artificial neural network modeling of the rainfall–runoff process. *Water Resour. Res*. **31**, 2517–2530.

Jain, A. & Srinivasulu,. S. (2006) Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. *J. Hydrol.* **317**, 291–306.

See, L. & Openshaw, S. (1999) Applying soft computing approaches to river level forecasting. *Hydrol. Sci. J.* **44**(5), 763–778.

Wang, L & Huang, G. (2002) An Artificial Neural Network Model of forecasting daily runoff based on runoff classification. *Irrig. & Drainage* **21**(4), 45–48.