Hybrid Neural Network Model Application in Annual Precipitation Forecast

Li Ma [1, 2, 3], Xuelian Li [1, 2] and Jin Wang[1, 2]

[1] *Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044*

[2] *School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044*

[3] *Key Laboratory of Meteorological Disaster of Ministry of Education, Nanjing University of Information Science & Technology, Nanjing 210044*

mali1775088@163.com; lixuelian191@126.com; wangjin@nuist.edu.cn

Abstract

When applied to precipitation forecasting, the mean generating function - optimal subset regression (MGF-OSR) model is limited by its low accuracy and high error, while the back propagation (BP) neural network model has difficulty in learning for matrix selection. This paper proposes a new MGF-OSR-BP model, which uses a MGF to extend original data, an OSR to select the best series as the BP neural network input node and learning matrix, and the resultant data for training. The training procedure determines the number of hidden layers and uses an optimal number of hidden layers for model training. This paper uses the MGF-OSR-BP model to analyze precipitation data from Hangzhou, China, for 53 years, from 1956 to 2008. The 1956-2006 precipitation data are used as the training sample, and the 2007-2008 data are used as the test set data to verify the practicality of the forecast system. A fitting verification is performed using the forecasted data against field measurement data, and the results show that the forecast accuracy is better than that of the MGF-OSR model or the MGF stepwise multiple regression model.

Key words*: precipitation forecast, neural network (NN), mean generating function (MGF), optimal subset regression (OSR)*

1. Introduction

China is constantly affected by its monsoon climate, and meteorological disasters emerge frequently. Among these disasters, droughts and floods have the greatest effect on agriculture. To mitigate the damage resulting from droughts and floods as much as possible, it is important to obtain accurate precipitation forecasts [1].

The artificial neural network method has many prominent characteristics, including good self-adaptive learning ability and non-linear mapping capability. Furthermore, it does not need to know the internal structure of a forecast system, and instead emphasizes a non-linear mapping relation between the model input and output. Currently, neural-network-based precipitation forecast systems are the subject of much research. Jin and Chen used a neural network integrated forecast method to study spring precipitation in Nanjing, and their results show that a neural network integrated forecast model results in a better fit and a more

accurate forecast than other traditional integrated forecast formulas; however, it does not provide a good solution for selecting a neural network model structure and parameters and determining hidden nodes [2]. Li, et al., used a back propagation (BP) neural network to build a flood season precipitation forecast model, and their results show that the BP method has a better fitting of historic samples and better forecast results than the stepwise regression model; however, the intrinsic limitations of BP neural network model have not been thoroughly investigated [3]. Jin et al. used a neural network combined with the mean generating function (MGF) to build a hybrid forecast model and conducted a forecast experiment on the precipitation in the northern, central, and southern regions of the Guangxi province in June, and their results show that this forecast method is better than the MGF regression forecast model and forecast factor regression forecast model. However, in that work, the model built upon a MGF regression is not always the global optimum, while the advantage of optimal subset regression (OSR) is the ability to select the globally optimal subset [4]. Huang et al. conducted research on a principal-component-based neural network model and applied it to a water level forecast. This model is prominently better than a regression factor neural network forecast model; however, its historic samples have a less accurate fit than the traditional neural network model [5]. Sun performed research on combining the MGF and OSR to build a model and used the OSR modeling method to calculate the error series for a forecast formula optimization. The results show that the MGF model has some degree of reliability for a hydrological factor long-term forecast [6].

This paper uses a combination of MGF, OSR, and neural networks to build a new hybrid MGF-OSR-BP forecast model. This new forecast model considers both the model and the learning matrix construction, sets model parameters, determines an optimal hidden node number and uses OSR to select a global optimal subset as a learning matrix, which overcomes the weakness of MGF in selecting a local subset. Our experimental results show that the MGF-OSR-BP model is clearly better at fitting historic samples and forecasting independent samples than the MGF-OSR model and the MGF stepwise multiple regression model.

2. Data and Methods

2.1 Data Selection

This paper uses 1956-2008 annual precipitation data from Hangzhou (Table 1) as calculation samples to build the forecast model; the 1956-2006 annual precipitation data are used as the training sample and the 2007-2008 data are used as the verification sample for forecast verification.

Table 1. Annual Precipitation Series from 1956 to 2008 in Hangzhou (mm)

year	annual precipitation					
1956-1961	1644	1442	1362	1692	1627	1643
1962-1967	1480	1399	1253	1454	1448	1208
1968-1973	1286	1645	1667	1377	1460	2070
1974-1979	1533	1844	1367	1781	1054	1184
1980-1985	1632	1612	1346	1963	1348	1465
1986-1991	1276	1686	1329	1756	1695	1547
1992-1997	1533	1832	1583	1665	1702	1643
1998-2003	1663	1964	1360	1656	1925	1254
2004-2008	1252	1286	1332	1381	1579	

2.2 MGF Extension

MGF calculates the mean for one-dimensional time series observed values at a certain time interval [7]. In this section, data in 51 original series are used by the MGF to generate 68 extension series. Suppose that the standard-deviation-normalized time series is as follows [1]:

$$x^{(0)}(t) = \{x(1), x(2), \cdots, x(N)\}. \quad (1)$$

A first-order differential operation is applied to the series:

$$\Delta x(t) = x(t+1) - x(t), t = 1,2,\cdots, N-1.$$

Thus, the following first-order differential series is obtained:

$$x^{(1)}(t) = \{\Delta x(1), \Delta x(2), \cdots, \Delta x(N-1)\}. \quad (2)$$

Next, a second-order differential operation is applied:

$$\Delta\Delta x(t) = \Delta^2 x(t) = \Delta x(t+1) - \Delta x(t), t = 1,2,\cdots, N-2.$$

Thus, the following second-order differential series is obtained:

$$x^{(2)}(t) = \{\Delta^2 x(1), \Delta^2 x(2), \cdots, \Delta^2 x(N-2)\}. \quad (3)$$

We use the following formula,

$$\bar{x}_l(i) = \frac{1}{n_l}\sum_{j=0}^{n_l-1} x(i + jl). \quad (4)$$

to perform a MGF calculation on the original series, first-order differential series, and second-order differential series, in which i=1,2,⋯,l,1 < =l<=M and nl=INT(N/l) and M can be INT(N/2) or INT(N/3) depending on the sample size, where INT indicates the integer part [1]. In this example, suppose that the sample series maximum period length M_{max}=INT(N/3)=17 , the MGF at time interval 17 is then as follows:

$$\overline{X}_1(1) = 1/51(1644 + 1442 + \ldots + 1332) = 1534$$
$$\overline{X}_2(1) = 1/25(1644 + 1362 + \ldots + 1252) = 1473$$
$$\overline{X}_2(2) = 1/25(1422 + 1692 + \ldots + 1286) = 1603$$
$$\vdots$$
$$\overline{X}_{17}(1) = 1/3(1644 + 2070 + 1695) = 1803$$
$$\overline{X}_{17}(2) = 1/3(1422 + 1533 + 1547) = 1507$$
$$\vdots$$
$$\overline{X}_{17}(17) = 1/3(1460 + 1756 + 1332) = 1516$$

Next, formula (5) is used to perform a periodic extension calculation on series MGF.

$$f_l(t) = \bar{x}_l\left[t - IINT\left(\frac{t-1}{l}\right)\right]. \quad (5)$$

Where t=1, 2,…,N; l=1,2,…,M, and the original series constructed MGF extension matrix is as follows:

$$H_{51\times17} = \begin{bmatrix} 1534 & 1473 & 1567 & 1462 & \cdots & 1803 \\ 1534 & 1603 & 1586 & 1667 & \cdots & 1507 \\ 1534 & 1473 & 1448 & 1503 & \cdots & 1580 \\ 1534 & 1603 & 1567 & 1562 & \cdots & 1630 \\ 1534 & 1473 & 1586 & 1462 & \cdots & 1664 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1534 & 1473 & 1448 & 1503 & \cdots & 1516 \end{bmatrix}$$

Similarly, a periodic extension calculation is performed on the first- and second-order differential MGF series.

$$f_l^{(3)}(t) = X(1) + \sum_{i=1}^{t-1} f_l^{(1)}(i+1)(t = 2,3,\cdots,n; l = 1,2,\cdots,m). \quad (6)$$

We can then use formula (6) to generate the accumulated extension series. Finally, we derive the 4M (68) MGF extension series from the original series for independent variable selection.

2.3 MGF Stepwise Multiple Regression Equation

When there are a large number of independent variables, some factors may not have a significant impact on the dependent variables, and the independent variables x may not be completely independent from each other. In this case, a stepwise regression analysis can be used for x factor selection. In this way, an established multiple-regression model can yield a better forecast result.

During a stepwise regression analysis, a main regression equation is first established, including all of dependent variables y and independent variables x, and a hypothesis test is performed on the main equation and each independent variable. If the main equation is not obvious, then the multiple regression equation does not have a linear relation. Any independent variable with a weak effect on y should be eliminated to re-establish multiple regression equations without this factor. Select the factor with the obvious effect as the independent variable and establish an "optimal" regression equation [8].

Earlier, we obtained a 68 forecast factor series. Using fi(t) as the forecast factor and x(t) as the forecast object, according to a stepwise regression procedure [9], we obtain the following forecast model:

$$X(t) = a_0 + \sum_{i=2}^{M} a_i f_i(t) + e(t) \quad t = 1,2,\cdots,N,$$

in which a0 and ai are undetermined coefficients and e(t) is white noise.

Using a stepwise regression analysis to obtain coefficients a0 and ai, we obtain the following fitting and forecast equations [9]:

$$X(t) = a_0 + \sum_{i=2}^{M} a_i f_i(t) \cdot$$

$$X \bullet (N + p) = a_0 + \sum_{i=2}^{M} a_i f_i(N + p) \quad p = 1,2,\cdots,$$

respectively, where p is a forecast point. The MGF extension series is used to establish the stepwise multiple regression equation, which is as follows:

$$Y=-2357.476+0.533X_3+0.709X_7+0.545X_{4+}0.386X_5+0.357X_6.$$

Substituting each factor with its value in the regression equation yields the fitting results shown in Figure 1 and the forecast results shown in Table 4.

2.4 Establishment of an OSR Forecast Equation

Establishing an OSR forecast equation means finding the OSR from all possible regressions. Theoretically, it has been shown that with a given forecast factor, a stepwise regression cannot obtain an optimal regression equation. Conceptually, a stepwise regression follows a branched tree and runs a selection using one factor along one branch; if that one factor is not selected properly, there is no guarantee of obtaining an optimal, or even sub-optimal, regression equation. In addition, another difficulty in stepwise regression is that the F-test critical value cannot determine α and has a high degree of randomness [7]. Meanwhile, when an analysis uses an optimal subset method to construct a BP neural network learning matrix, it is also required to establish an optimal regression forecast equation.

Based on the above method, 68 MGF extension series were generated as independent variables for screening. A simple regression for the time of each extension series and original series was established, and the couple score criterion (CSC) value was calculated. Selecting series satisfying $CSC > \chi_{\alpha}^{2}$ as initial forecast factors, it was assumed that all P extension series had been selected. Next, all possible 2Pregression subsets were calculated [8]. According to the couple score variable selection criterion, one optimal subset was selected as the forecast equation.

Table 2 shows the optimal subset for different numbers of independent variables, CSC values, and root mean square error (RMSE) values. According to the optimal subset combination of a different number of independent variables for annual precipitation in Hangzhou and the CSC and RMSE values, one optimal regression subset was selected as the forecast equation. The forecast equation is as follows:

$$Y=-3606.115+0.646X_2+0.603X_3+0.542X_4+0.440X_5+0.426X_6+0.687X_7+0.501X_{9.}(7)$$

Using formula (7) to perform a forecast that fits the data for the original series, as shown in Table 4, yields the fitting results shown in Figure 1 and Figure 2. Meanwhile, the selected MGF extension series with 7 independent variables was used to construct a BP neural network training set, and the training set was loaded into a 3-layer feed forward network input end.

Table 2. Optimal Subset of Different Numbers of Independent Variables

k	Optimal subset	CSC	RMSE
1	x3	33.32	183.87
2	x3 x7	41.66	161.56
3	x3 x4 x7	53.52	143.84
4	x3x4x5x7	73.82	130.77
5	x3x4x5x6x7	75.4	124.62
6	x3x4x5x6x7x9	72.99	121.16
7	x2x3x4x5x6x7x9	81.75	114.97
8	x2x3x4x5x6x7x8x9	81.13	114.6
9	x1x2x3x4x5x6x7x8x9	70.5	114.11
10	x1x2x3x4x5x6x7x8x9x12	69.77	114.05
11	x1x2xx3x4x5x6x7x8x9x11x12	69.04	113.99
12	x1x2x3x4x5x6x7x8x9x10x11x12	68.3	113.97

2.5 BP Neural Network

MGF selection is mainly determined by a classified forecast information entropy calculation and a linear correlation calculation. In fact, regarding whether a MGF has only a linear correlation with a forecast object, different forecast objects may yield different results. While a notable characteristic of neural networks is that prior knowledge of the internal structure of the forecast system is not necessary, it emphasizes a non-linear mapping relation between the model input and output [10].Therefore, we use the annual precipitation in Hangzhou as the forecast object and the BP neural network combined with the MGF-OSR as the annual precipitation forecast.

The BP neural network includes an input layer, hidden layer, and output layer. The BP network algorithm learning procedure uses a gradient descent method to modify the network connection weight (weight and threshold) to minimize the sum of the squared error for the network. The signal from the input layer is processed by each hidden layer and is transmitted to the output layer. The neuron state of each layer only affects the next layer's neuron state. If the error between the output layer's actual output and the expected output is larger than the defined error standard, then the error signal is transmitted along the original connection route in reverse to modify the network's original weight and minimize the error.

Based on the above analysis, 7 independent variables were selected as the 3-layer BP neural network input and the original series as the network output. The key of the BP neural network node determination is to determine the hidden layer node. There are various methods for network hidden layer node number determination, such as an empirical method and an equation method. In this example, the formula $N_1 = \sqrt{n+m} + a$ (a = [1, 10], n is the input layer node, and m is the output layer node) is used to determine the hidden layer node. In the example, the hidden layer node range is [4,13], and by training, we obtain 9 hidden layer nodes and 1 output layer, establish a MGF-OSR-BP neural network model, and use it for training and solving. The training parameter settings are as follows: the training function is trainlm for 3000 iterations, and learning is complete when the convergence error reaches 0.0001. The results of form fitting the MGF-OSR-BP neural network model and MGF-OSR model with the 51-year samples are shown in Figure 2, and the forecast results are shown in Table 4.

3. Fitting and Forecast Result Analysis

We used the MGF stepwise multiple regression method and the MGF-OSR method to process the 1956-2006 precipitation data samples from Hangzhou, and we compared and analyzed the respective fitted values with those from the original sample data. Similarly, we used an MGF-OSR-BP neural network method to process the 51-year sample data, and we compared and analyzed the calculated results and fits of the original data and the MGF-OSR model. Below is a detailed analysis.

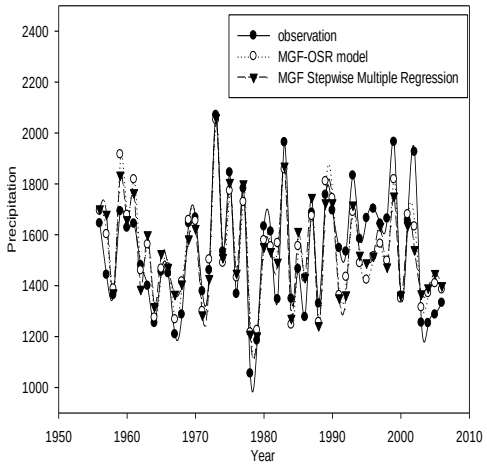

Figure 1. Fittings of the MGF Stepwise Multiple Regression and MGF-OSR Model

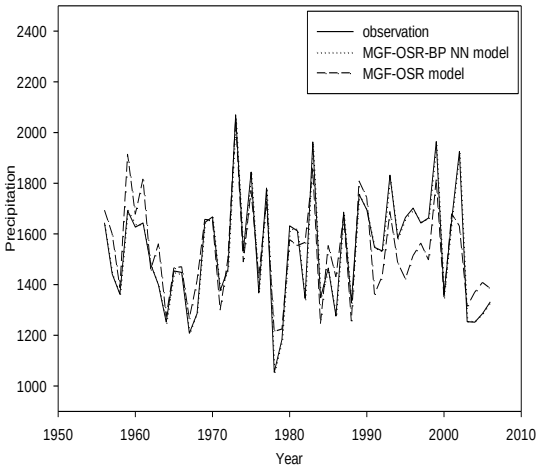

Figure 2. Fittings of the MGF-OSR Model and MGF-OSR-BP NN Model

Figure 1 shows the variation curves of the actual values of annual precipitation in Hangzhou and the fitted values from the MGF stepwise multiple regression model and the MGF-OSR model. This figure shows that the two models yield good precipitation fitting results, with the MGF-OSR model having better precipitation fitting than the MGF stepwise multiple regression model. This result is not surprising because the MGF-OSR model uses

OSR for modeling factor selection and extracts the global optimal subset variables to establish the model. As a result, its model fitting accuracy is better than that of the regression model.

Figure 2 shows the variation curves of the actual precipitation values in Hangzhou and the fitted value of the MGF-OSR model and the MGF-OSR-BP neural network model. It is clear that the MGF-OSR-BP neural network model has better precipitation fitting than the MGF-OSR model. This superior fitting is exactly the advantage of neural networks: it can approximate a non-linear function within any degree of accuracy. As a result, it has a better fitting accuracy.

To compare the fitting results of the three models in a quantitative approach, the following 4 indices are defined [11]:

(1) Mean absolute percentage error (MAPE)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{|y_t|}. \quad (8)$$

(2) Mean squared error (MSE)

$$MSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}. \quad (9)$$

(3) Mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|. \quad (10)$$

(4) Correlation coefficient (PR)

$$PR = \frac{\sum_{t=1}^{n}(y_t - \bar{y}_t)(\hat{y}_t - \bar{\hat{y}}_t)}{\sqrt{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2}\sqrt{\sum_{t=1}^{n}(\hat{y}_t - \bar{\hat{y}}_t)^2}}. \quad (11)$$

In formulas (8)-(11), y_t and $\hat{y}_t$ represent the actual and fitted values, respectively. The index statistic results of the three models are shown in Table 3.

Table 3. Comparison of the Fitting Accuracy of the Three Models

model	MAPE	MSE	MAE	PR
MGF-OSR-BP NN model	0.23	3.59	4.93	0.99
MGF-OSR model	6.10	114.97	92.34	0.86
MGF stepwise multiple regression model	6.68	124.62	99.96	0.83

Table 3 also shows that all three models have good precipitation fitting results, while the MGF-OSR-BP neural network model provides a better fit than the other two models, with a perfect result. The MGF stepwise multiple regression model has the lowest fitting accuracy, and the MGF-OSR model fitting accuracy is in between those of the other two models.

Above is a comparison of the fitting results of the different models. The MGF-OSR-BP neural network model has a better fitting accuracy than the other two models; however, having a strong forecast model fitting capability does not necessarily indicate a better

practical forecast capability. Although the fitting results are one aspect of model evaluation, the forecast results are more important. We selected independent data samples from 2007-2008 and used the three models to run forecast tests for the independent samples, and the results are shown in Table 4.

Table 4. Comparison of the Predicting Accuracy of Independent Samples of the Three Models

(O: observation, P: prediction, A: absolute error, R: relative error, AVA: average.)

Year	O(mm)	MGF stepwise multiple regression model			MGF-OSR model			MGF-OSR-BP NN model		
		P	A	R/%	P	A	R/%	P	A	R/%
2007	1381	1807.2	426.2	30.8	1724.7	343.74	24.9	1439.4	58.4	4.2
2008	1579	1509.2	-69.8	-4.4	1538.2	-40.84	-2.6	1653.6	74.6	4.7
AVA			248.0	17.6		192.29	13.8		66.5	4.5

The results in Table 4 show that, using an MGF stepwise multiple regression model for the 2-year independent sample forecast, the mean absolute error is 248.02. Again, the regression method is not good at annual forecasting. When using the MGF-OSR model for the forecast of 2-year independent samples, the error is moderate, and the mean absolute error is 192.29. The MGF-OSR-BP neural network model forecast yields markedly better results than the other two models, and its forecast for 2-year independent samples has a mean absolute error of 66.5, which indicates a better forecast capability. Further analysis shows that MGF-OSR-BP neural network model precipitation forecast of Hangzhou in 2007 is the closest to the actual data and that it has a better forecast accuracy than the other two models.

4. Conclusions

This paper uses a mean generating function (MGF) for data extension, based on an optimal subset regression (OSR) to select the optimal data series as backpropagation (BP) neural network input factors, and establishes a new MGF-OSR-BP neural network model. This model has a better fitting accuracy and forecast result than the other two models. It fully utilizes the advantages of MGF and OSR in global optimal learning matrix selection, and in modeling, it properly utilizes the excellent performance of neural networks in self-adaptive learning and non-linear mapping. The improvement in forecast capability provides a new method to extend the application of neural networks in future forecast research areas and provides a reference for similar middle- and long-term forecast research based on elements of time series data. It also has promising potential future applications.

References

[1] L. Jin, Editor, “Neural network forecasting model theory methods and application”, China Meteorological Press, Beijing **(2004)**.
[2] L. Jin and N. Chen, “Acta Meteorologica Sinica”, vol. 57, **(1999)**, pp. 198.
[3] Y. H Li, D. Liu and L. Jin, Scientia Meteorological Sinica, vol. 22, **(2002)**, pp. 461.
[4] L. Jin, Y. Luo and Y. H. Wang, Plateau Meteorology, vol. 22, **(2003)**, pp. 618.
[5] H. H. Huang, C. Z. Sun and L. Jin, Nanjing institute of meteorology, vol. 28, **(2005)**, pp. 58.
[6] Y. Y. Sun, Water Resources and Power, vol. 27, **(2009)**, pp. 14.
[7] Q. X. Wen, G. D. Sun and C. J. Zhang, Journal of catastrophology, vol. 15, **(2000)**, pp. 11.

[8] F. Wei, Editor, "Modern climate statistic diagnostics prediction technology", China Meteorological Press, Beijing, **(1999)**.
[9] Z. N. Qin, Journal of Guangxi Meteorology, vol. 24, **(2003)**, pp. 15.
[10] L. Jin, Y. Luo and Z. S. Lin, Acta Meteorologica Sinica, vol. 11, **(1997)**, pp. 364.
[11] Y.H. Li, L. Jin and Q. L. Miao, NanJing Institute of Meteorology, vol. 28, **(2005)**, pp. 549.

Authors

Li MA

She received her B.S. degree in 1985 from the Chengdu Institute of Meteorology and her Ph. D degree in 2011 from Nanjing University of Information Science and Technology. She is a professor and tutor for graduates in Nanjing University of Information Science and Technology. Her main research interests include image processing, pattern recognition, and meteorological information processing and data assimilation.

Xuelian LI

She received her bachelor degree in the Nanjing University of Information Science and Technology, 2010. Currently she is a Master candidate in the Nanjing University of Information Science and Technology. Her areas of interest are meteorological information processing and data assimilation.

Jin Wang

He received the B.S. and M.S. degree in the Electronical Engineering from Nanjing University of Posts and Telecommunications, China in 2002 and 2005, respectively. He received Ph.D. degree in the Ubiquitous Computing laboratory from the Computer Engineering Department of Kyung Hee University Korea in 2010. Now, he is a professor in the Computer and Software Institute, Nanjing University of Information Science and technology. His research interests mainly include routing protocol and algorithm design, performance evaluation and optimization for wireless ad hoc and sensor networks. He is a member of the IEEE and ACM.