RBF Neural Network Based on K-means Algorithm with Density Parameter and Its Application to the Rainfall Forecasting

Zhenxiang Xing, Hao Guo, Shuhua Dong, Qiang Fu, Jing Li

RBF Neural Network Based on K-means Algorithm with Density Parameter and its Application to the Rainfall Forecasting

Zhenxiang Xing[1,2,3,a], Hao Guo[1,b], Shuhua Dong[4,c], Qiang Fu[1,2,3,d], Jing Li[1,e]

[1]College of Water Conservancy &Civil Engineering, Northeast Agricultural University, Harbin 150030, China; [2] Collaborative Innovation Center of Grain Production Capacity Improvement in Heilongjiang Province, Harbin 150030, China;[3] The Key lab of Agricultural Water resources higher-efficient utilization of Ministry of Agriculture of PRC, Harbin 150030,China;
[4]Heilongjiang Province Hydrology Bureau, Harbin, 150001
[a] zxxneau@hotmail.com, [b]ghneau@sina.cn, [c]shdshuiwen@sina.com, [d]fuqiangneau@sina.cn, [e] lijingneau@sina.com

Abstract. The Radial Basis Function (RBF) neural network is a feed-forward artificial neural network with strong approximation capability. A K-means algorithm based on density parameter was introduced to determine clustering center aimed to improve the training rate of the RBF. It could reduce sensitivity of traditional K-means algorithm for initial clustering centers. A rainfall forecasting model of RBF based on K-means algorithm was built, which was applied to forecast monthly rainfall in Shuangyashan City during the flood season, aiming to test the effectiveness of this model. The case study showed that the mean relative error of rainfall forecasting in flood season (from June to September) of the year 2006, 2007 and 2008 was 10.81%, and the deterministic coefficient was 0.95. It demonstrated a higher forecasting accuracy comparing to a RBF model based on a standard K-means algorithm and BP (Back Propagation) model, and the rainfall forecasting results satisfied the requirements of hydrologic prediction.

Keywords: Rainfall forecasting; Radial Basis Function Neural Network; Density parameter; K-means

1 Introduction

The rainfall is an important process with higher uncertainty in natural water cycle. A rainfall forecasting method with high-accuracy could predict the change amount of precipitation, which could provide important significance to decision-making of flood control and disaster reduction. There are a lot of methods for rainfall forecasting, such as the regression analysis [1], the grey prediction [2], the fuzzy prediction [3], and artificial neural network [4.5] and so on.

The artificial neural network has many advantages of rainfall forecasting, which has strong ability to deal with nonlinear problem and high generalization ability. So the BP neural network and the RBF neural network is widely used network model. Compared with the BP network, the number of hidden layer of RBF network can be

adaptive adjusted in training phase. In addition, input layer and hidden layer of RBF use linear connection instead of weights. It seems that the RBF has advantages compared with BP, which can greatly improve the convergence speed of network. RBF also has better approximation capability of nonlinear function. In this paper, a RBF model was trying to be used to forecast the precipitation in Naolihe catchment in Sanjiang plain.

2 Radial Basis Function Neural Network

The Radial Basis Function Neural Network (RBF) is a three-layer feed forward network with single hidden layer [6], such as input layer, hidden layer, output layer. The network also can approximate any continuous function with arbitrary precision theoretically.

In RBF network is, the hidden layer space is constructed by the RBF, so an input vector can be directly (do not need the weights) mapped into hidden space. The mapping relationships between hidden layers to output layers were described as a linear function and the outputs of network are linear weighted sum of hidden unit output [7].

Gaussian function is commonly used as the radial basis function in RBF network. The expression of activation function is:

$$\phi(x_p - c_i) = \exp\left[-\frac{1}{2\sigma^2} \| x_p - c_i \|^2\right] \tag{1}$$

where ϕ is activation function; $\| x_p - c_i \|$ is the European norm; $x_p = \left(x_1^p, x_2^p, \ldots, x_m^p\right)^T$ is the p^{th} input samples; p=1,2,...,P (P is the total number of samples); c_i are a center of Gaussian function; σ are a variance of Gaussian function.

The outputs of network are obtained through the RBF structure

$$y_j = \sum_{i=1}^{h} w_{ij} \exp\left[-\frac{1}{2\sigma^2} \| x_p - c_i \|^2\right] \tag{2}$$

where w_{ij} denote weights between the hidden layer to output layer; i=1,2,...,h (is number of nodes in a hidden layer); y_j denote actual outputs from the j^{th} of output corresponding with input sample; and other symbols have the same meaning as above.

3 K-means Algorithm Based on Density Parameter

In an RBF neural network, three parameters are needed to be solved, which are the centers of basis function in hidden layer, the variance and the weights of the hidden layer to output layer. The key to build a good network is to select an adaptive basis function center. There are many methods, such as the randomly selected algorithm, the self-organization selected algorithm, the clustering analysis algorithm and the orthogonal least squares algorithm to do this job. Among the methods above, the K-means clustering algorithm is one of the fairly effective learning algorithms.

The K-means algorithm is a clustering algorithm based on distance, i.e., a distance is the assessment criteria of the comparability. The traditional K-means algorithm is

easy to understand, which is easier for programming. However, it is obvious that K-means algorithm is sensitive to the initial clustering center and clustering results fluctuate when given different initial input. It will infect the final characteristics of sample groups. The K-means algorithm based on density parameters can reduce the influence caused by initial clustering centers to clustering results comparing to a traditional algorithm. Therefore, the K-means algorithm based on density parameter of RBF is applied to forecast the rainfall in this paper.

3.1 The Concept of Density Parameter

The aggregation of sample data: S= $\{x_1, x_2, \ldots, x_n\}$, the initial clustering centers: z_1, $z_2, \ldots, z_k$.

Define 1 an Euclidean distance between each two samples

$$d(x_i, x_j) = (|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)^{1/2} \quad (3)$$

where $x_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ and $x_j = \{x_{j1}, x_{j2}, \ldots, x_{jp}\}$ are samples with p-dimension.

Define 2 an average distance between samples

$$MeanDist = \frac{1}{C_n^2} \times \sum d(x_i, x_j) \quad (4)$$

where n is the total number of samples.

Define 3 the density parameter [8]

In the density space, neighborhood of any point is defined as the region which point p is the center and *MeanDist* is the radius. The number of points in region is known as the density parameter based on *MeanDist*, called *density* (p, *MeanDist*).

3.2 K-means Algorithm Based on Density Parameter

The Euclidean distance is used as similarity measure of the K-means algorithm. The mutual farthest k data objects are more representative than random k data. It is considered that noisy data are often mixed in practical data. In order to avoid selecting the noisy points, we can get k points in the high density area as the initial clustering center by following steps [8]:

(1) Calculate the distance between any two data according to formula (3), called as d (x_i, x_j);

(2) Calculate the average distance of total data according to formula (4), called as *MeanDist*;

(3) Calculate the density parameters of all data, named as *density* (p, *MeanDist*), and composing a data set named D.

(4) Find the maximum value from D, $z_k = \max\{density(p_i, MeanDist) \mid i \in (1,2\ldots, n) density(p_i, MeanDist) \in D\}$. If $d(p_i, z_k) < MeanDist$, the *density* (p, *MeanDist*) is deleted from D; z_k is the k^{th} of cluster center.

(5) Repeat the step (3) and step (4) until finding cluster centers, and the number of cluster centers is k.

Therefore, cluster centers can be obtained in accordance with the method, that is, the final center of the basis function of RBF network.

4 The Rainfall Forecasting Model Based on RBF

4.1 The Structure of RBF Network

4.1.1 The number of input layer neurons

It has much significance to choice the number of input neurons for RBF network. The unreasonable choice will affect training ability of the network, and even lead to model crash. Therefore, the node number of input neurons of RBF network is determined by autocorrelation analysis technology in this article.

4.1.2 The number of hidden layer neurons

The hidden layer neurons are used to store connection weights and thresholds between input layer and hidden layer. It is characterized in terms of reflecting inherent law between training sample and expected output. There are no accurate and scientific methods to determine the hidden layer neurons at present, and as well as more determined by empirical formula and numerical test. The traditional method is also used in this paper, and the empirical formulas are [9]:

$$n_2 = \sqrt{n_1 + m} + a \tag{6}$$

$$n_2 = \log 2^{n_1} \tag{7}$$

where n_2 is the number of hidden layer neurons, n_1 is the number of output layer neurons, m is the number of input layer neurons, a is a constant between [0, 1].

4.2 The Rainfall Forecasting Model

The network model is built by RBF tool box of MATLAB, the program language as follows:

net =newrb(p,t,goal,spread,mn,df)

where p and t are the input vector and the output vector of sample respectively; *goal* is the mean square error for the purpose of preventing excessive fitting. In this paper, *goal* was set as 0.001; *mn* is the maximum number of neurons; *df* is the number of neurons to add between displays; *spread* is the width of the radial basis function, and in this paper it is determined by the distance between each center of hidden layer neurons, and the formula is

$$spread = b \times d_i \tag{8}$$

where b is the overlap coefficient, and usually this value are set as an integer greater than 1[10]; d_i is the minimum distance of basic function centers of the hidden layer, the basis function center is obtained by K-means algorithm based on density parameters.

5 Case Study

5.1 The Experimental Data

The rainfall forecasting model of RBF based on K-means algorithm was built, which was taken precipitation in Shuangyashan City of Sanjiang plain in each flood season from 1955 to 2005 (June to September) as an example. The precipitation of the year from 1955 to 2003 was used for network training, and that from 2004 to 2005 was used for network testing, forecasting the rainfall of the year 2006, 2007 and 2008.

5.2 Data Preprocessing

Input and output data of RBF network should be within [0, 1], so the precipitation need to be normalized before training model according to following formula,

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{9}$$

where x_i and y_i is the rainfall before and after normalizing respectively; x_{min} and x_{max} are the minimum and maximum values of all rainfall.
The processed data will be used as the samples of the RBF neural network model.

5.3 Identifying and Training RBF Network

The key point to build RBF neural network is to determine number of nodes in input layer and hidden layer, which is based on the characteristics of rainfall. The correlation analysis of rainfall series in flood season in Shuangyashan was built, and we can estimate that the rainfall series is suitable for the AR (p) model. The order of Markov process should be four by analyzing the AIC criterion. That is to say, the input layer neurons of RBF network are four. According to the empirical formulas and trial calculation, number of nodes in hidden layer is eight. Therefore, the structure of network in this paper is 4:8:1.
To achieve converging, the rainfall of the year from 1955 to 2003 is trained 172 times by using neural network which has been constructed above, and the fitting error of network is 0.001. The rainfall of the year 2004 and 2005 is used to test and verify the generalization ability of the RBF network model. The mean prediction relative error is 10.9736% and the deterministic coefficient is 0.95. In conclusion, the constructed RBF network could be applied to forecast rainfall in Naolihe catchment.

5.4 Results and Analysis

The calculation in Table 1 show that the mean relative error of rainfall forecasting in 2006, 2007 and 2008 is 10.81% by using established RBF network model, the deterministic coefficient is 0.95; the relative error using K-means algorithm is 14.30%, and the deterministic coefficient is 0.86. The mean relative error of BP network model is 14.38%, the deterministic coefficient is 0.87.Meanwhile, fitting time and convergence times of the established RBF network model are reduced, which are 80s and 172 times respectively; and the standard K-means algorithm's fitting time and

convergence times are 97s and 235 times, while the BP network model's are 125s and 302 times. Therefore, compared with RBF model of standard K-means and BP model, the computing speed of the RBF network that has been constructed in this paper are increased by 18% and 36% respectively, and the mean relative error are reduced by 24% and 25%.The fitting figure between observed rainfall and forecasted rainfall in training period, testing period and forecasting period was shown in Fig 1. (The first 182 series is training period, series of183^{th} to 190^{th} is testing period, and other series is forecasting period).

Table 1. The results of rainfall forecasting compared impoved RBF network with standard K-means RBF network and BP network

Time	Observed rainfall /mm	a		b		c	
		Forecas-ted rainfall /mm	relative error /%	Forecas-ted rainfall /mm	relative error /%	Forecas-ted rainfall /mm	relative error /%
0606	127.9	105.76	-17.31	106.68	-16.59	99.86	21.92
0607	148.7	130.88	-11.98	125.03	-15.92	120.86	18.72
0608	64.9	70.00	7.86	72.05	11.01	70.14	8.08
0609	17.2	20.07	16.70	18.80	9.30	20.00	16.29
0706	65.9	70.22	6.56	75.01	13.82	73.03	10.83
0707	35.8	41.04	14.64	40.37	12.77	42.08	17.53
0708	156.1	142.37	-8.79	132.50	-15.12	140.66	9.89
0709	39.6	35.40	-10.60	45.09	13.87	43.53	9.92
0806	109.9	100.56	-8.50	95.65	-12.97	104.40	5.00
0807	48.4	52.37	8.21	42.38	-12.44	40.07	17.21
0808	101.0	110.20	9.11	118.00	16.84	117.84	16.68
0809	29.3	32.05	9.39	35.46	21.03	35.30	20.49
MRE	—	—	10.81	—	14.30	—	14.38
DC	—	0.95	—	0.86	—	0.87	—

a. The improved RBF network that has been built in this paper;
b.RBF network based on standard K-means; c. BP network;
MRE is the mean relative error; DC is the deterministic coefficient

6 Conclusions

The K-means algorithm is commonly used as learning algorithm in RBF neural network. The initial clustering centers are selected randomly based on the traditional K-means algorithm, which result in different convergence abilities. The method of

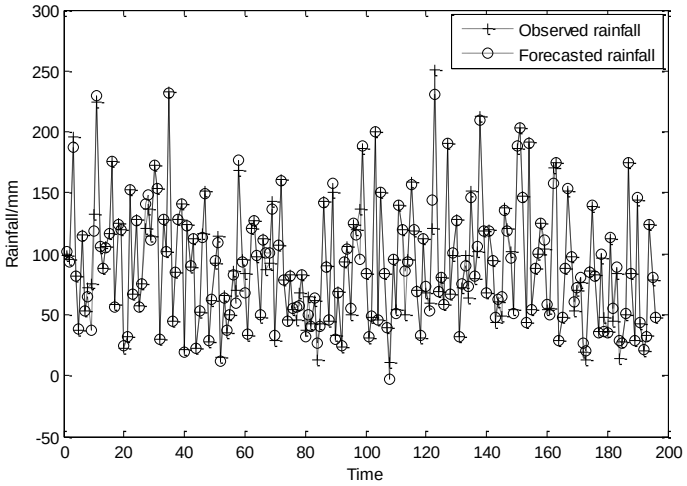

Fig.1. The fitting figure between observed rainfall and forecasted rainfall

density parameters was trying to be used to reduce the sensitivity of selecting randomly for clustering centers in this paper, and to calculate the width of basis function of the RBF network. And then a rainfall forecasting model of RBF based on K-means algorithm was built, which was applied to forecast monthly rainfall in Shuangyashan City in Heilongjiang Province. The case study results and conclusions were shown as follows:

1) Comparing with the traditional K-means algorithm, the RBF network model based on the density parameter had its advantages for determining the number of clustering results and the centers of basis function. And the model had higher convergence rate and forecast accuracy.

2) The accuracy of RBF network model on account of rainfall forecasting during flood season in Shuangyashan City had been improved greatly compared with BP network model. The mean relative error of monthly rainfall in 2006, 2007 and 2008 were reduced by 24% and the deterministic coefficient was increased by 10%. So the model built in this paper could be used as an effective method for rainfall forecasting or other time series forecasting with strong uncertainties.

3) In this paper, number of nodes in the network was determined by empirical formula and numerical test. In order to further increase the calculation precision of model forecasting, some intelligent optimization algorithms such as genetic algorithm could be adopted to optimize number of nodes in hidden layers in the future research.

Acknowledgment

Funds for this research was provided by the national natural science foundation (51109036; 51179032), the specialized research fund for the doctoral program of higher university of the Ministry of Education (20112325120009) the special scientific research funds for Ministry of water resources public welfare industry

(201301096), The leader talent echelon back-up headman funding projects of Heilongjiang province (500001), the studying funds of postdoctoral in Heilongjiang province (LBH-Q12147), the science and technology project of Heilongjiang Provincial Education Department(11541022).

References

1. Chang Qing, Zhang Xin, Cai Huanjie et al. A Precipitation Forecast Model Based on Regression Analysis and Time Series Analysis [J]. Journal of Soil and Water Conservation Bulletin, 2009, 29(1):88-91.
2. Li Caiyuan, Gu Yonggang. Grey Prediction Model in the Application of the Upper Yangtze River Basin Surface Rainfall Forecast [J]. Journal of Climate Science, 2003, 31 (4): 223-225.
3. Sun Caizhi, Lin Xueyu. Research on fuzzy Markov Chain Model with Weights and its Application in Predicting the Precipitation State. [J]. Journal of Systems Engineering, 2003, 17 (4): 294-299.
4. Feng Yi, Wu Boqiang, Cui Lingzhou. Study on Forecasting Typhoon Rainfall Based on BP Neural Network [J].Research of Soil and Water Conservation, 2012, 12 (3): 289-293.
5. Li Xiangyang, Cheng Chuntian, Lin Jianyi. Bayesian Probabilistic Forecasting Model Based on BP ANN [J]. Journal of Hydraulic Engineering, 2006 37(03): 354-359.
6. Yi Yanping, Lu Wenxi, Zhang Yun et al. Study on the Surrogate Model of Groundwater Numerical Simulation Model Based on Radial Basis Function Neural Network [J]. Research of Soil and Water Conservation, 2012, 19(4):265-269.
7. Zhang Defeng. Neural Network Application Design of Matlab [M]. Beijing: China Machine Press, 2009.
8. Zhang Jianhui. Research and Application of K-means Clustering algorithm [D]. Wuhan University of Technology, 2007.
9. Zhu Ming. Data Mining [M].Hefei: Press of University of Science and Technology of China, 2002.
10. Su Meijuan. Research on Radial Basis Function Neural Network Learning Algorithms [D].Suzhou: Soochow University, 2008.