

Research of using RF model to drought forecast on Huaihe River

J Wu^{1,3} and Y F Chen²

¹Water Resources Department of Hebei Province, Shijiazhuang 050000, China

²College of Hydrological and Water Resource, Hohai University, Nanjing 210098, China

E-mail: 805483939@qq.com

Abstract. Random Forest is a combination classification model based on classification and regression tree. RF model can deal with nonlinear problem, therefore, it is able to predict drought. In this paper, RF model is used to forecast drought on the Huaihe River Basin where more and more serious droughts occur. The runoff of twenty-one stations can reflect the overall drought level on Huaihe River. First, the drought has been divided into three levels depending on the SPI criteria. Then, the most important thirty factors are selected as model variables according to the Incnodepurity index after selecting preliminary screening sets. Furthermore, analysis of the drought level of the hydrological stations during the period from 1963 to 2013 is carried out with RF model. The result shows the average accuracy rate of prediction is 73%. The RF drought model is demonstrated as a new effective drought prediction model to reduce or eliminate the loss caused by drought.

1. Introduction

The frequency and intensity of global drought had an increasing tendency under the global climatic changing and human activities. Therefore, taking effective prevention measures to reduce the damage and losses of drought in agriculture, industry and ecology is very important. The most direct way is to predict the drought accurately and effectively. Drought prediction has two mainly method, one based on the physical and the other based on statistical theory. Due to the implicated factors such as climate changes, hydrologic mechanism and so on, leading both of the traditional method which only based on linear smoothly and multivariate model which based on stochastic simulation technology have limitations which are shortcomings on the nonlinear characteristics of forecast of meteorological elements and weather phenomenon. To describe the nonlinearity on forecasting drought climate, Artificial Neural Network (ANN), Grey System (GM) and Support Vector Machine (SVM) have been used to predict drought. Although ANN has strong ability of nonlinear fitting, but its inherent weaknesses are running the prone process, network structure and learning parameters. Moreover, the momentum factor is difficult to select, leading to an over fitting problem and affecting the generalization ability of the network [1]. GM model GM (1, 1) has many advantages, such as small amount of input data, simple principle, moderate calculation and the high prediction precision, but it is only applied to short-term drought forecasting. The prediction accuracy decreases as the predicted time scales increases [2]. In order to improve the accuracy of precipitation forecast, Li [3] combining the BP Neural Network and Gray Model, but it is also applied to short-term drought forecasting. Furthermore, the result of forecast was influenced by the size of the input of the training sample and



the way of training screening. G F Fan [4] studied the SVM to predict drought. SVM shovelled the problems of the excessive input varies and over-fitting. However, the optimization of its intrinsic parameters and kernel function is very difficult. It did not satisfy the demand of drought prediction so that the result was only divided into two parts, and we must standardize or normalize the sample and factors in classification and prediction.

Random Forest is an assembly-itemized model that is based on CART. The model can solve nonlinear problem and do not need pre-processing the data. Through the summary on lots of classification tree to improve the precision of the predicted model, it can replace some of traditional machine learning methods such as ANN, GM. In addition, it has been applied more and more in all the aspects of the society also has been used in hydrological analysis and got better effect; Zhao [5] has taken random forests model into the water runoff forecast in Yangtze River.

Random Forests can itemize impact factors and rate the importance of each impact factor that the score as the basis for screening important factors, applying a vote system following the majority rule to determine the final sort. This report is expected by establishing the random forests to classifying drought prediction model. The data during the period from 1963 to 2013 of twenty-one representative station at Huaihe River were used to analysis meteorological drought, and using a meteorological drought index: standardized precipitation index (SPI) analysing drought grade. Firstly preliminary choose 372 factors as the factor set, secondly use RF model selecting 30 importance factors, then use RF drought model to test and forecast at last.

2. Selection of drought classification index

Drought is a natural phenomenon that rainfall, runoff and other natural water in a certain period is less than the normal level in a certain geographic range, water deficit in rivers and lakes. Drought refers to meteorology, hydrology, agriculture and the social and economic disciplines. In addition, its essence is the shortage of water, and what mostly affects the drought is the rainfall. Meteorological drought is associated most closely with precipitation, and is also the most important among the four types of drought. According to the important factors, meteorological drought index is selected to evaluate drought. Meteorological drought index chooses standardized precipitation index (SPI). SPI is considered the quantity of the standard normal distribution, corresponding to the same cumulative frequency of a rainfall value as the standardization of the precipitation rainfall index value (see Appendices).

Because SPI assumes that among all sites where drought and flood happens at the same probability, it is unable to distinguish frequent drought region. This paper respectively divides 21 drought classifications, then we can independently analyse each stand true drought conditions. The scale of analysis used the month, in order to prevent the longer time step indicators being used in region where precipitation is relatively small.

Based on the definition of SPI and analysing of current situation, drought can be divided into three categories based on the threshold of SPI. The SPI value that is corresponding to the frequency of the rainfall is less than 30.9% will be defined as the threshold of drought, rainfall frequency more than 30% of the SPI value will be defined as a threshold of waterlogging, the month's SPI value between two thresholds is normal. Table 1 shows results of drought classification.

Table 1. Result of drought classification based on SPI with 21 stations of Huaihe basin.

Station Name	Level 1 (waterlogging)	Level 2 (normal)	Level 3 (drought)
Wang Jiaba	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Gu Zhen	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Gu Zhen sluice	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)

Gao Liangjian sluice	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Hua Wu	($\sim$,0.49]	(-0.48,0.49)	[-0.48, $\sim$)
Ling Bi	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Ming Guang	($\sim$,0.50]	(-0.46,0.50)	[-0.46, $\sim$)
Nan Wan	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Shang Qiu	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
She Yangzhen	($\sim$,0.48]	(-0.48,0.48)	[-0.48, $\sim$)
Shen Qiu	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Shiliang River Reservoir	($\sim$,0.47]	(-0.50,0.47)	[-0.50, $\sim$)
Shi Mantan	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Su Qian sluice	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Tong Bai	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Xiang Hongdian	($\sim$,0.49]	(-0.52,0.49)	[-0.52, $\sim$)
Xin Yanggang sluice	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Xin Zhenf	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Xu Zhou	($\sim$,0.45]	(-0.50,0.45)	[-0.50, $\sim$)
Zhang Shan sluice	($\sim$,0.50]	(-0.50,0.50)	[-0.50, $\sim$)
Zhu Madian	($\sim$,0.50]	(-0.51,0.50)	[-0.51, $\sim$)

Take the March's drought grade of Wang Jiaba station for instance to show the year of drought grade based on SPI classification in table 2.

Table 2. Result of March month drought classification on Wang Jiaba station.

Level	Year
Level 1	1963, 1972, 1977, 1980, 1987, 1991, 1992, 1993, 1997, 1998, 2002, 2003, 2007, 2010
Level 2	1964, 1965, 1966, 1967, 1968, 1969, 1971, 1973, 1974, 1978, 1979, 1982, 1985, 1988, 1990, 1994, 1996, 1999, 2004, 2009, 2012, 2013
Level 3	1970, 1975, 1976, 1981, 1983, 1984, 1986, 1989, 1995, 2000, 2001, 2005, 2006, 2008, 2011

3. Research method

3.1. Principle of random forest

Random Forest (RF) [6] is a kind of statistical learning theory which was put forward by Leo Breiman in 2001. Comparing with other methods, Random Forest's forecasting is more highly accurate and more quick operation which is compared with other machine learning algorithms; it integrates characteristics of split two methods between Bagging and the characteristics of random; it

has a good generalization performance and good robustness on abnormal noise and it has achieved good application in different fields. Random forests contain random forest regression and random forest classification.

Definition: A random forest class is a multi-decision classifier consisting of many CART classification models $\{h(X, \theta_k), k=1, 2, \dots\}$, which $\{\theta_k\}$ are the dependent random vectors of identically distribution [7].

The final classification result of the input X is the optimal classification result selected by each decision tree. The basic process of RFC has three steps.

Firstly, take k sample sets from the original training set through the Bootstrap as shown in figure 1, and keep the same size between the original training set and each sample set. Then, train each of k sample sets to set up k decision tree. Finally, the result of RFC is the final classification for all of the models, using the majority rule in the classification vote as shown in figure 2.

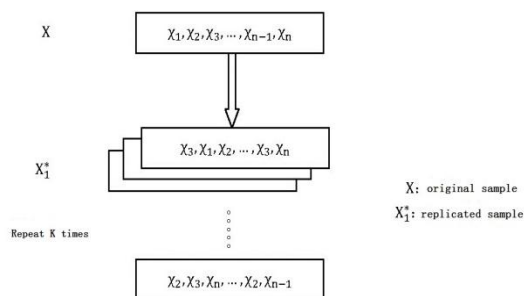


Figure 1. Schematic diagram of bootstrap resampling.

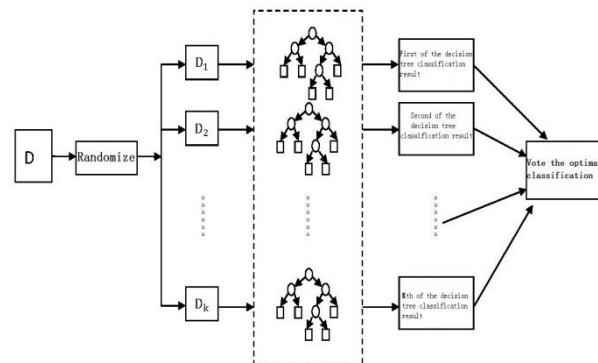


Figure 2. Schematic diagram of random forest structure.

Bagging method is used to form a new training set called RF model, it can divide characteristics randomly, makes random forests be able to tolerate the noise better, and can reduce the correlation between the single tree, not pruning single tree will get lower deviation, guaranteeing classification efficiency on the classification tree. Training set is generated by Bagging methods, the original sample concentration that is close to 37% of the sample will not appear in the training set, the data are called outside the Bag (Out - Of - Bag, OOB) data. Meanwhile, it is available that OOB data can be used to estimate the generalization error. For each decision tree, an error estimation corresponding OOB can be got, by using the random forest OOB error estimates for all the decision trees in average, which will obtain the random forest generalization error estimates. Breiman has shown that OOB error is an unbiased estimation, and relatively to the cross validation by experiments, OOB estimation is effective, and the result is similar to the results of the cross validation. Therefore, the research model is performance evaluation methods ground on generalization error estimation on the OOB estimation method [8].

4. Case study

4.1. Step of the RF model

This report respectively researched and forecasted every month's drought condition of 21 stations in Huaihe River Basin. The model predictor used 12 months 'antecedent precipitation and 74 atmospheric circulation indices that were issued by the National Meteorological Centre. 51 sets of rainfall data from 1963 to 2013 that conducted "three characteristics review" were also used in this research. In the case of the March's drought condition of Wang Jiaba station, the drought prediction is as follows:

- Preliminary selection of forecast factors: Improve the ability of influencing factors selecting and the accuracy of prediction of the model. The selection of explanatory variables should pay attention to the physical or meteorological correlation between the factors and the forecast object. Taking full account of the influence of factors on the spatial and temporal state of the atmosphere and the characteristics, we conducted a preliminary screening of 35 factors from the 74 hydrological and meteorological characteristics, which were related to the formation of meteorological drought.

Table 3. Result of preliminary screening influence factors of random forest model.

List of 24 preliminary screening influence factors			
the area of subtropical high in Northern Hemisphere (5 E-360)	Northern Hemisphere subtropical high ridge (5 E-360)	the intensity of North Hemisphere polar vortex (Region 5, 0-360)	
the area of subtropical high in South China Sea (100 E-120 E)	the north boundary of subtropical high on the Northern Hemisphere (5 E-360)	Eurasia zonal circulation index (IZ, 0-150 E)	
the area of subtropical high in North America (110 W-20 W)	the north boundary of subtropical high on the South China Ocean (100 E-120 E)	Asia zonal circulation index (IZ, 60 E-150 E)	
the area of Subtropical high in Pacific Ocean (110 E-115 W)	the north boundary of subtropical high on the Pacific Ocean (110 E-115 W)	Asia meridional circulation index (IM, 60 E-150 E)	
the intensity index of Northern Hemisphere subtropical high (5 E-360)	the area index of Asia Region polar vortex (Region 1, 60 E-150 E)	The Tibetan Plateau Index_B (TPI_B) (30 N-40 N, 75 E-105 E)	
the intensity index of West Pacific Ocean subtropical high (110 E-180)	the area index of Northern Hemisphere polar vortex (Region 5, 0-360)	India - burma trough (15 N-20 N, 80 E-100 E)	
the intensity index of North Africa Atlantic North America subtropical high (110 W-60 E)	the intensity of Asia Region polar vortex (Region 1, 60 E-150 E)	numbered typhoons	
the intensity index of South China Sea subtropical high (100 E-120 E)	Southern Oscillation Index	sunspot	

- Screening on prediction factors. To reduce the influence of noise added in the evaluation on the accuracy of forecasting random forests, we chose Incnodepurity index to evaluate the importance of the factors, by which we compared the importance of the factor using the impurity reduction value calculated by the segmentation variables of all tree nodes, and the higher values indicated the importance of the factors in the prediction of RF. According to this principle, the 30 factors ordered by the importance of factors on the establishment of the nonlinear relationship can be achieved as shown in table 3. We selected 30 kinds for a total of 372 monthlyobserved hydrology-meteorological data as alternative explanatory variables factor, sequence number of precipitation in the previous 12 months came to 12 and monthly screening of 12 months of 30 atmospheric circulation characteristics arrangement, serial number row to 372. Setting the annual drought level of the Wang Jiaba station as the target variable and all the samples as trained sample sets, we evaluated the importance of the explanatory variables based on the random forest model. According to the explanatory variables importance in descending order, the 30 variables were selected as the final forecasting factor. The impact factor is shown in figure 3, and the prediction of drought in the month is shown in table 4.

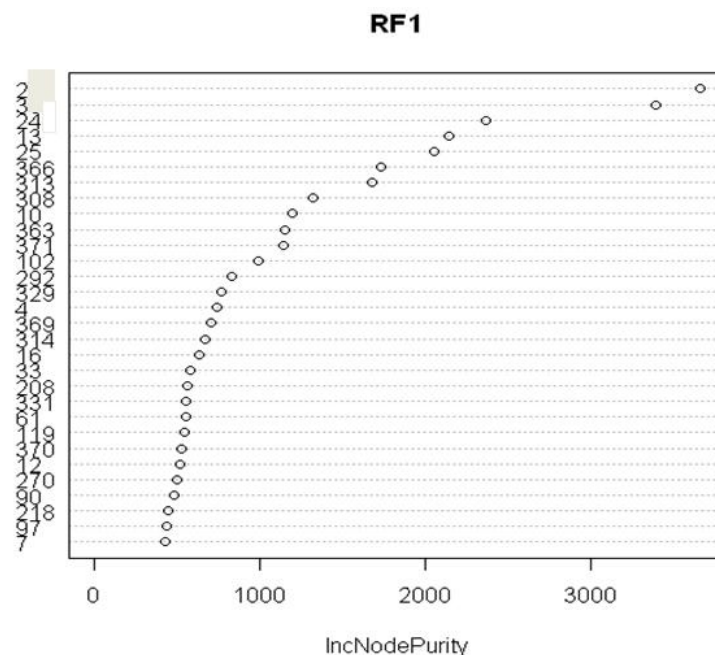


Figure 3. Ranking variable importance that associated with drought classification by random forest.

Table 4. Prediction factors based on RF of March month at Wang Jiaba station.

The forecast factors of March at Wang Jiaba station's RF drought forecast model		
August rainfall in the last year	The Tibetan Plateau Index-B (January)	india-burma trough (March)
December rainfall in the last year	The Tibetan Plateau Index-B (February)	india-burma trough (July)
January rainfall in the current year	the area of subtropical high in North America (January)	the intensity of Pacific Region polar vortex (July)
February rainfall in the current year	the area of subtropical high in North America (September)	Asia meridional circulation index (August)
November rainfall in the last year	the area of Subtropical high in Pacific Ocean (January)	Asia zonal circulation index (March)
February rainfall in the last year	the area of subtropical high in Northern Hemisphere (January)	Asia zonal circulation index (April)
Southern Oscillation Index (October)	the area of subtropical high in Northern Hemisphere (April)	Eurasia zonal circulation index (July)
Southern Oscillation Index (November)	the intensity of North Hemisphere polar vortex (June)	the intensity index of Pacific Ocean subtropical high (June)
Southern Oscillation Index (March)	the area index of Northern Hemisphere polar vortex (February)	the intensity index of North Africa Atlantic North America subtropical high (January)
Southern Oscillation Index (May)	the area index of Asia Region polar vortex (April)	the intensity index of North Africa Atlantic North America subtropical high (June)

4.2. Training and prediction results of RF model

March of Wang Jiaba station of RF drought prediction model took the 1963-2007 data as the model training data, and the error of OOB which was building by model is shown in table 5.

Table 5. The error of OOB at training period of RF model.

Level		Prediction			class rate (%)
		Level 1	Level 2	Level 3	
Real	Level 1	10	2	1	23
	Level 2	2	16	1	16
	Level 3	1	3	9	31

It can be seen from table 5 that the level 1 to 3 of OOB prediction error is 23%, 16% and 31 %, respectively. Average of OOB error of the model is 23%, much better. The result of forecast the grade 2 (normal condition) is the best, because the biggest sample size, or predictors that has not been selected may have little effect on its drought classification. The result of forecast the grade 3 (drought condition) is the worst. It may be related to the sample size and factors that affect the RF model to select the classification level of drought.

Reference climate forecasting business criteria [9], Ts scoring according three grades (Ts = total score/total number of forecasts; see table 6).

Table 6. Ts score standard of 3 level (%).

Level		Prediction		
		Level 1	Level 2	Level 3
Real	Level 1	100	50	0
	Level 2	50	100	50
	Level 3	0	50	100

According to the establishment of the RF model, the results of forecast the drought category from 2008 to 2013 are shown in table 7.

Table 7. Class forecast result of drought from 2008 to 2013 in Wang Jiaba station.

Year	Real Level	Forecasting level	Forecast accuracy
2008	3	2	50
2009	2	2	100
2010	1	2	50
2011	3	2	50
2012	2	2	100
2013	2	2	100

As the table 7 shows, there are 3 years rightly predicted while 3 years predicted differently one-grade. Based on the scoring criteria of table 6, the average forecast accuracy of the random forest model in March is 75%. According to the above prediction process, we made an assay of the drought conditions and obtained the forecast results of drought grade about the 21 stations in Huaihe River Basin as shown in table 8.

Table 8. Drought prediction result of every month of 21stations at the Huaihe River Basin based on RF model (%).

Station Name	January	February	March	April	May	June	July	August	September	October	November	December	Accuracy of station (%)
Wang Jiaba	58.3	66.7	75	83.3	83.3	100	91.7	91.7	75	75	58.3	66.7	77.1
Gu Zhen	58.3	58.3	75	75	66.7	91.7	83.3	75	75	66.7	66.7	58.3	70.8

Gu Zhen sluice	66.7	58.3	66.7	66.7	75	91.7	91.7	75	66.7	66.7	66.7	58.3	70.9
Gao Liangjia n sluice	66.7	66.7	75	66.7	75	91.7	83.3	83.3	75	75	66.7	66.7	74.3
Hua Wu	50	58.3	58.3	75	75	83.3	83.3	75	66.7	66.7	58.3	58.3	67.4
Ling Bi	58.3	58.3	66.7	66.7	83.3	91.7	91.7	83.3	75	66.7	58.3	58.3	71.5
Ming Guang	66.7	66.7	66.7	75	66.7	83.3	83.3	83.3	75	75	66.7	66.7	72.9
Nan Wan	66.7	58.3	66.7	75	66.7	83.3	91.7	83.3	75	66.7	66.7	58.3	71.5
Shang Qiu	58.3	66.7	66.7	75	83.3	91.7	83.3	83.3	75	75	66.7	66.7	74.3
She Yangzhe n	58.3	58.3	66.7	66.7	75	83.3	83.3	83.3	75	75	66.7	66.7	71.5
Shen Qiu	58.3	66.7	66.7	75	83.3	100	91.7	75	75	75	66.7	58.3	74.3
Shiliang River Reservoi r	58.3	58.3	66.7	66.7	75	91.7	91.7	75	75	75	58.3	66.7	71.5
Shi Mantan	50	58.3	66.7	75	75	83.3	83.3	83.3	75	66.7	58.3	58.3	69.4
Su Qian sluice	58.3	66.7	75	75	66.7	91.7	91.7	83.3	66.7	66.7	66.7	58.3	72.2
Tong Bai	66.7	66.7	66.7	66.7	75	91.7	91.7	91.7	83.3	75	58.3	58.3	74.3
Xiang Hongdia n	58.3	58.3	66.7	75	83.3	91.7	91.7	75	75	66.7	58.3	58.3	71.5
Xin Yanggan g sluice	58.3	58.3	66.7	75	83.3	83.3	83.3	83.3	75	75	66.7	58.3	72.2
Xin Zhenf	66.7	66.7	75	75	83.3	91.7	91.7	83.3	66.7	75	58.3	66.7	75.0
Xu Zhou	58.3	66.7	66.7	83.3	83.3	91.7	83.3	83.3	75	66.7	58.3	58.3	72.9
Zhang Shan sluice	58.3	66.7	66.7	75	83.3	91.7	91.7	83.3	75	66.7	58.3	66.7	73.6
Zhu Madian	58.3	58.3	66.7	66.7	83.3	83.3	91.7	83.3	75	66.7	66.7	58.3	71.5
Accurac y of month (%)	59.9	62.3	68.3	73.0	77.4	89.7	88.1	81.7	73.8	70.7	62.7	61.5	72.4

The average forecast accuracy of RF drought prediction model for 21 stations in Huaihe River Basin is 72.4%. Among them, the annual average forecast accuracy of Wang Jiaba station is 77.1%, which is the highest while the annual average forecast accuracy of Huawu station is 67.4%, which is the lowest. The highest of the average month forecast accuracy of each station is June with the accuracy rate 89.7%, while the lowest is January with the accuracy rate 59.9%.

Through the analysis, for accurate rate higher sites and months, the influence factors of RF model were selected to have a close relationship with drought in the area, where the cause and effect relationship was larger, so the model could be more accurate to predict the drought level. On the contrary, for accurate rate lower sites and months, the cause of the drought mechanism is affected by more factors, such as the January drought level more affected by winter temperatures, ENSO, etc. It may be also the poor application of drought classification based on SPI in a certain region and in the month.

5. Conclusion

- Basing on the SPI, the paper predicted the drought grade by RF drought prediction model of twenty-one stations during the period from 1963 to 2013 in Huaihe River, choosing 30 factors from the prediction factor set closely related to drought based on physical causes and statistical characteristics as the final prediction factors to inspect and predict by RF model [10].
- The overall average forecast accuracy of RF drought prediction is 72.4%, better than the climate system of the weather forecast accuracy (65%). The RF drought model is demonstrated as a new effective drought prediction model to reduce or eliminate the loss caused by drought.
- The RF model can draw a relatively high prediction accuracy rate, so it can be applied to different areas. However, the classification model based on just for meteorological drought index SPI analysis, in the future, we can consider blending the indexes from hydrologic drought and agriculture drought, so that the drought prediction is more scientific and persuasive from different viewpoints to level the drought [11].

Acknowledgments

Authors wish to acknowledge assistance and encouragement from all of our colleagues, specially Y Kang. We also thank the editors and anonymous reviewers.

Appendices

SPI: its formula is as follows:

$$SPI = S - \frac{t - (c_2 t + c_1)t + c_0}{[(d_3 t + d_2)t + d_1]t + 1.0} \quad (1)$$

t : Cumulative probability function, $t = \sqrt{\ln \frac{1}{F^2}}$, F : Cumulative probability of rainfall; S : Coefficient, when $F > 0.5$, S is the value of positive, otherwise take negative value, c, d are coefficients: $c_0 = 2.515517$; $c_1 = 0.802853$; $c_2 = 0.010328$; $d_1 = 1.432788$; $d_2 = 0.189269$; $d_3 = 0.001308$.

References

- [1] Wang C X 2011 *Weather Forecast Method Based on Support Vector Machine (SVM)* (Nanjing University of Information Science & Technology)
- [2] Liu D Y, Liang Z M and Zhao W M 2012 Application of grey system theory on drought prediction *Water Power* **38** 10-2
- [3] Li X H, Yang Y and Yang H W 2014 Combining BP neural network with gray model to achieve drought predicting *J. Shenyang Agric. Univ.* **45** 253-6
- [4] Fan G F, Zhang Y and Yang M 2011 Study of drought prediction based on support vector machine *Chinese J. Agric.* **32** 475-8
- [5] Zhao T T G, Yang D W and Cai X M 2012 Predict seasonal low flows in the upper Yangtze River using random forests model *J. Hydroelectric Eng.* **31** 18-24
- [6] Li X H 2013 Using "random forest" for classification and regression *Chinese J. Appl. Entomol.* **50** 1190-7
- [7] Gislason P O, Benediktsson J A and Sveinsson J R 2006 Random forests for land cover classification *Pattern Recogn. Lett.* **27** 294-300
- [8] Li H Y, Xie M and Jiang S 2012 Recognition method for mid-to long-term runoff forecasting factors based on global sensitivity analysis in the Nenjiang River Basin *Hydrol. Process.* **26** 2827-37

- [9] Yang J, Lu Z G and Li S P 2009 Hydrological model year forecast based on SVM method *Y R* **31** 29-30
- [10] Kang Y, Chen Y F and Gu S H 2014 Assessment of sustainable utilization of regional water resources based on random forest *Water Res. Power* **32** 34-8
- [11] Wu J, Chen Y F and Yu S N 2016 Research on drought prediction based on random forest model *China Rural Water Hydropower* **17** 17-22