An Ensemble Forecast Method of Rainstorm Based on mRMR and Random Forest algorithms

Hua-sheng Zhao *, Xiao-yan Huang, Ying Huang
Guangxi Research Institute of Meteorological Disasters Mitigation, Nanning China
*Corresponding author’s e-mail: 2006zhaohuasheng@163.com

Abstract. This thesis presents an ensemble method of rainstorm based on maximum relevance minimum redundancy (mRMR) and random forest algorithms called mRMR-RFR. The proposed method is applied to the forecasting results of ensemble numbers from European Centre for Medium-Range Weather Forecasts (ECMWF). The method filtrates the 51 collected forecast members of ECMWF using the mRMR algorithm, and selects the members with maximum relevance and minimum redundancy to the forecasting objects. The selected members are regarded as the input factors of the random forest algorithm for forecast modeling. Experimental forecasting statistics show that, compared with the interpolation method of the collected forecasting members of ECMWF, mRMR-RFR can result in better forecast effect and use the numerical forecast products more effectively. The proposed method is thus suitable for forecasting.

1. Introduction
Heavy rainfall, especially the one heavier than rainstorm, is the critical source of flood and debris flow. The improvement of numerical model forecast in recent years has promoted the intensive use of numerical forecast products in weather forecasting and demonstrated the role such method plays. Given that atmosphere is a highly nonlinear system, the result of numerical weather prediction is very sensitive [1] to even a minimal error in initial conditions. Ensemble forecast is an effective method [2–4] for overcoming the uncertainty of single forecast and has been applied to improve the veracity of numerical weather predictions.

Collection forecast extracts available information from multiple forecast results of numerical models to create an ensemble forecast. A few prediction methods based on the fundamental idea and purpose of ensemble have been proposed, and they present good theoretical foundation and improved effectiveness[5]. Krishnamurti [6] first introduced the weighting forecasting based on the idea of super ensemble forecast to obtain accurate prediction. Chinese scholars argue that collection forecast of equal weighting and weighted arithmetic average from different memberships is better than single control forecast. The reason is that the former method can significantly restrain random error and improve the forecasting result. L. Chen et al. [7] collected the forecast results from the units involved in the rainfall prediction during China’s flood season. Zhao[8] constructed the collected temperature forecasting system of more than 600 Chinese sites using BP networks of neural network approach; this system is based on 2m-high temperature forecast in three weather centers. C. Chen[9] explored the multimodal short-term forecast of super collection in a limited area model and used the methods of related weighting, multiple linear regression, and support vector regression. All the studies mentioned above suggest that collected forecast using support vector is more efficient than the forecast under related weighting and multiple linear regression. Moreover, all collected forecast methods perform...
better than single forecast in most cases. However, no mature and effective collected numerical prediction is available at home and abroad, and forecasting technicians has no clear consensus on the appropriate method.

This thesis presents an ensemble forecasting method based on maximum relevance minimum redundancy (mRMR) and random forest algorithms called mRMR-RFR. The rainfall events of 89 stations in Guangxi Province are treated as the forecast objects. Several members are selected from the 51 collected forecast members of European Center for Medium-Range Weather Forecasts (ECMWF) using mRMR. These members present maximum relevance (max-relevance hereinafter) to the forecast objects, and minimum redundancy (min-redundancy hereinafter) among themselves. The chosen members are used as the input factors of the random forest algorithm for forecast modeling. This study provides a new numerical forecast product for forecasting.

2. Research Materials and Method

2.1 Research Materials

2.1.1 Research Materials

The 24-h and 48-h precipitation daily data of forecast raining members at 8:00 and 20:00 from the collected forecast models of ECMWF are used in constructing the forecast models and conducting forecast experiments. A total of 2100 samples are derived from April 2012 to August 2016 excluding missing data. The data from April 2012 to December 2015 are used as samples for forecast modeling, and the data from January 2016 to September 2016 (a total of 526 samples) are used as independent samples for forecast testing.

2.2 Method

The mRMR algorithm of mutual information and the random forest algorithm are adopted to build a single-station rainstorm forecast model for 89 meteorological stations in Guangxi Province. First, the data of each station from the 51 collected forecast members of ECMWF are filtrated by mRMR, the selected members are then used as the inputs in the random forest algorithm for forecast modeling. The final forecast results can then be obtained.

2.2.1 mRMR Algorithm.

The mRMR algorithm was introduced by Peng[10]. This algorithm relies on max-dependency. Specifically, the method selects a character subset S with m character quantity from the character set. Subsets has the max-dependency with the category c. The formula is expressed as

$$\max D(s,c), \quad D = I\{\{x_i, i = 1, \ldots, m\}\}; c$$  \hspace{1cm} (1)

Given that max-dependency exhibits an approximation relation with max-relevance based on mutual information, max-relevance computes the mutual information average of single variable xi and category C. The formula is expressed as

$$\max D(s,c), \quad D = \frac{1}{S} \sum_{x_i} I(x_i, c)$$  \hspace{1cm} (2)

During the feature selection of max-relevance, the redundant feature can be possibly chosen because of its high redundancy with features. Min-redundancy must thus be considered to obtain a near-optimal feature set.

$$\min R(s), \quad D = \frac{1}{S^2} \sum_{x_i, x_j} I(x_i, x_j)$$  \hspace{1cm} (3)

Combining the two restraints mentioned above is the principle of mRMR. Peng defined the operators that combine the relevant factor D and redundant factor R, and introduced the simplified formula below to optimize D and R simultaneously.

$$\max \Phi(D, R) = D - R$$  \hspace{1cm} (4)
2.2.2 Random Forest Algorithm

The random forest algorithm has two types of application models: classification and regression. Random forest regression (RFR) is a statistical learning theory [11], proposed by Leo Breiman in 2001. The basic steps of RFR are as follows: extracting several samples from the original samples using the bootstrap re-sampling method, conducting the tree modeling for each bootstrap sample, predicting according to the combined decision trees, and deriving the final forecast by averaging[12]. Its basic function is the multiple decision tree models, which conducts prediction by combining several decision trees. The algorithm presents good forecast accuracy and generalization capability, fast convergence rate, and less regulation parameter. RFR can therefore avoid “over-fitting” effectively and is suitable for calculating different data sets. The algorithm also exhibits good robustness and is applied widely in various fields, such as medical science, management science, and economics[13-15].

Random forest forms the \( \{h(X, \theta), k = 1, \ldots, p\} \) composite pattern by random vector \( \theta \) (i.e., regression tree) according to the bootstrap re-sampling technology. The forecast variable is numerical variable, and the generated random forest is polybasic nonlinear regression analysis model. The formation of random forest forecast is the average of \( k \) trees \( \{h(X, \theta_k)\} \). The training set shaping the random forest is independent and selected from random vector \( Y, X \). The average generalization error of numerical forecast vector is expressed as

\[
E_{X,Y}(Y - h(X))
\]  

(5)

The characteristics of RFR are as follows:

(1) When the number of tree in the forest tends to infinity, then

\[
E_{X,Y}(Y - \text{av}_k h(X, \theta_k))^2 \to E_{X,Y}(Y - E_\theta h(X, \theta))^2
\]

(6)

(2) If for all \( \theta \), \( E(Y) = E_\theta h(X, \theta) \), then

\[
PE^*(\text{forest}) \leq \rho PE^*(\text{tree})
\]

(7)

In these equations, \( PE^*(\text{forest}) = E_\theta E_{X,Y}(Y - E_\theta h(X, \theta))^2 \), \( \rho \) is the weight relation between rest \( Y - h(X, \theta) \) and \( Y - h(X, \theta') \), while \( \theta \) is independent. The procedure of RFR is as follows:

(1) The content of the original data sample is \( N \); the number of training set is \( k \) (\( \theta_1, \ldots, \theta_k \)), which will be extracted randomly by bootstrap (the extraction includes recovery); on the basis of every training set, the corresponding decision tree \( \{T(x, \theta_k)\} \) will be formed. The unextracted samples after every bootstrap sampling produce \( k \) out-of-bag data (OOB), which are used the test sample of random forest;

(2) If the variable number of the original data is \( p \), \( \text{mtry} \) variable can be extracted randomly from every node of every regression tree. This variable can be used as the alternative branching variable. The optimal branch will be chosen on the basis of superiority degree. The selection method is the same as the method of constructing regression tree model. In the random tree regression, the parameter tree \( \text{mtry}= p/3; \)

Every regression tree is top-down branching recursively; the minimum size of set leaf node is \( 5 \) (nodesize=5), which is treated as the terminal condition of regression tree growing.

The random tree regression model consists of formed \( k \) regression tree, and the effectiveness valuation of regression is the residual mean square of OOB as presented in Formulas (8) and (9):

\[
MSE_{OOB} = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

(8)

\[
R^2_{RF} = 1 - \frac{MSE_{OOB}}{\sigma_y^2}
\]

(9)
where $y_i$ is the actual value of the dependent variable of OOB; $\hat{y}_i$ is the forecast value of the random forest of OOB; $\sigma^2_y$ is the variance of the forecast value of OOB from random forest.

3. Modeling Test of Collected Rainstorm Forecast

In this study, the precipitation events of 89 meteorological stations in Guangxi Province in the next 24 hours (two times daily: 8:00 and 20:00) are used as the forecast objects. This study determines the precipitation events with magnitude above that of rainstorm.

3.1 Modeling Sample and Forecast Factor Disposition of Collected Rainstorm Forecast in Single Station

The statistical analysis from April 2012 to August 2016 results in a total of 2176 samples from all stations. Notably, many stations show a precipitation with magnitude above that of rainstorm for 4775 times within 24 hours. The period from January 2016 to September 2016 is regarded as the independent sample test time (a total of 526 samples). Several stations present a rainfall magnitude of 50 mm for 967 times within a day. This study determines the frequency of rainfall events with magnitude above that of rainstorm within 24 hours. After constructing the model of single-station collected rainstorm forecast, the modeling sample type adopts classified process to improve the accuracy of rainstorm forecast and select model forecast factors using mRMR. The concrete steps are as follows:

1. For each of the 51 collected forecast members of ECMWF, 48 hours’ calculated precipitation of raining number in previous day is subtracted with 24 hours’ calculated precipitation of raining number to obtained the forecast raining number $R_{24}$ in the next 24 hours.

2. Polynomial interpolation must be adopted to interpolate the rainfall forecast number $R_{24}$ into the 89 meteorological stations in Chart 1. For each forecast object (meteorological station), 51 forecast factors exist (the interpolation of the 51 collected forecast members of ECMWF) denoted as $F_{51}$.

3. For the $k$-th ($k=1,...,89$) forecast object $Y$ (meteorological station), if the $F_{51}$ average of the station is higher than the threshold value $m$ ($m=25$ mm, i.e., above the magnitude of heavy rain), then the modeling sample is the history sample in which the real precipitation is more than $g$ ($g=20$ mm) (all samples before the modeling). The corresponding sample serial number is indexed. Otherwise, the modeling sample will be the sample in which the real precipitation is less than $g$ ($g=50$ mm), and the corresponding serial number is indexed.

4. From the serial number index in Step(3), the modeling sample matrix $X_{51}$ consisting of 51 factors and predicted and sequenced $YY$ can be derived.

5. By introducing $X_{51}$ and $YY$ into mRMR, several factors of the 51 collected forecast members can be computed. These factors present max-relevance and min-redundancy to the predicted and sequenced $YY$. A total of 10 of the $F_{51}$ is used and denoted as $T_{x}$.

6. After selecting the factor matrix in Step (5) and introducing $YY$ into the random forest regression, a collected forecast model will be constructed.

7. $T_{x}$ will be introduced into the forecast model to obtain the rainfall forecast of the station in the next 24 hours.

3.2 Result and Analysis

During the forecast test of single sample from January 2016 to August 2016, the adopted successive forecast is the same with the real forecast. To compare the numerical forecast product with the forecast result, the lattice point data of the collected forecast members of ECMWF are interpolated into the stations by linear interpolation, polynomial interpolation, and spline function interpolation. The averaging forecast method is obtained by interpolating each interpolation to the station. Three averaging forecast methods are used and denoted as M1, M2, and M3. Chart 1 shows the TS score after using mRMR-RFR and that after obtaining the average by interpolating the 51 collected forecast.
members of ECMWF to the station. High score means the corresponding forecast method is highly accurate. The formula of TS is expressed as:

\[
Ts = \frac{Na}{Na + Nb + Nc}
\]

(10)

where \(Na\) is the amount of station with right forecast; \(Nb\) is the amount of useless forecast station; \(Nc\) is the station amount with no forecast;

Table 1. TS score of single-station forecast sample using different methods from January 2016 to August 2016

| Month | mRMR-RFR | M1 | M2 | M3 | Station amount with precipitation≥50 mm |
|-------|---------|----|----|----|----------------------------------------|
| 1     | 0.24    | 0.00 | 0.00 | 0.00 | 60                                     |
| 2     | 0.50    | 0.00 | 0.00 | 0.00 | 1                                      |
| 3     | 0.07    | 0.06 | 0.06 | 0.06 | 11                                     |
| 4     | 0.08    | 0.03 | 0.03 | 0.04 | 105                                    |
| 5     | 0.17    | 0.13 | 0.13 | 0.13 | 203                                    |
| 6     | 0.17    | 0.06 | 0.06 | 0.06 | 231                                    |
| 7     | 0.13    | 0.01 | 0.03 | 0.03 | 114                                    |
| 8     | 0.24    | 0.18 | 0.19 | 0.18 | 206                                    |
| 9     | 0.07    | 0.00 | 0.00 | 0.00 | 36                                     |
| 1–9   | 0.15    | 0.08 | 0.09 | 0.10 | 967                                    |

According to table 1, the TS scores of the three interpolation methods show no significant difference in the independent forecast for nine months. They obtain a TS score of 0.19 in August, the highest among their results. They derive a rainstorm TS of 0 in three months. The TS score of the proposed mRMR-RFR forecast model does not obtain 0 in the entire study period. The TS of the proposed model reaches 0.24 in August, which is stably higher than the TS score of M1, M2, and M3. Except the TS score in April, the TS score of mRMR-RFR from April to August (in this period, more than 100 stations obtain ≥50mm precipitation) is higher than 0.10. From January to September, the total TS score of the mRMR-RFR forecast model improves by over 50%, compared with the TS score of the three interpolation forecast methods: M1, M2, and M3.

In order to understand the specific forecast performance of the mRMR-RFR forecasting model, the paper selected two forecasts in the 2016 business trial for analysis. One of them was rainstorm precipitation caused by the impact of Typhoon No. 4 in 2016. The time period was from 20:00 on August 2, 2016 to 20:00 on March 03 (the total rainfall of 27 stations in this period reached 50 mm or more). The specific forecast comparison chart of the method is shown in Figure 1. The other is the general heavy precipitation process from August 20th to 13th, 2016 at 20:00 (the total rainfall of 27 stations is more than 50 mm). The specific forecast comparison of the two forecasting methods is shown in Figure 2.
Figure 1 Comparison of typhoon rainstorm sample from 20:00 on August 2, 2016 to 20:00 on August 3, 2016

Figure 2 Comparison of rainstorm sample not caused by typhoon from 20:00 on August 12, 2016 to 20:00 on August 13, 2016

Figure 1 shows that the raining areas with over 50mm rainfall obtained by mRMR-RFR and M2 are generally affected by typhoon. The raining areas nearly cover all the stations. Thus, the effect of mRMR-RFR and M2 is similar.

Figure 2 indicates that M2 of the collected forecast members of ECMWF predicts the rain band; however, the rain magnitudes of all stations are less than 50mm and many stations exhibit precipitation magnitude of nearly 50mm. After conducting the mRMR-RFR classified fitting on history samples, some adjustments and improvements on the forecast magnitude given by ECMWF collected forecast are made. In the rainstorm forecast on the actual precipitation of 27 stations, mRMR-RFR provides accurate forecast for 13 stations, empty forecast for 16 stations, and missing data for 14 stations. The total TS score of the proposed method is 0.30.

The forecast capability of mRMR-RFR is generally better than that of the interpolation method of the collected forecast members of ECMWF, especially in forecasting rainstorm not caused by typhoon. The proposed method can significantly improve the effectiveness of the collected forecast members of ECMWF on rainstorm. This performance is attributable to the adopted factor selection method, modeling sample classification, and forecast model construction. The proposed method is thus suitable
for forecasting heavy rainfall and can eliminate samples with small rainfall magnitude from the forecasted modeling samples with heavy rain and reduce disturbance in model samples.

4. Summary and Discussion

In the factor selection, the mutual information of max-relevance and min-redundancy is adopted. This method can optimally improve the forecast information of the selected factor and guarantee max-relevance with forecast quantity. The random forest of the model presents few adjustable parameters, fast computing speed, and good fitting capability and generalization performance.

The analysis using independent forecast sample shows that, compared with the interpolation method of the collected forecast members of ECMEF, mRMR-RFR can better estimate the areas and ranges of rainstorm, especially the rainstorm not caused by typhoon.

The proposed method can accurately interpret numerical forecast products. The method is simple, contains no parameter, and is accessible to most practical forecasters. Therefore, the method contributes to the interaction between vocation work and scientific research and the improvement of forecast level.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 41575051,41765002) and the Guangxi Key National Natural Science Foundation (Grant No. 2017GXNSFDA198030).

References

[1] Lorenz E N (1965). A Study of the Predict Ability of a 28-variable Atmospheric Model .Tell us,17:321–333.
[2] Du J. (2002). Present Situation and Prospects of Ensemble Numerical Prediction. Journal of Applied Meteorological Science, 13(1):16-28.
[3] Du J. and Chen J. (2010). The Corner Stone in Facilitating the Transition from Deterministic to Probabilistic Forecasts--Ensemble Forecasting and its Impact on Numerical Weather Prediction. 36(11):1-11.
[4] Li Z. CH. and Chen D. H. (2002). The Development and Application of the Operational Ensemble Prediction System at National Meteorological Centre. Journal of Applied Meteorological Science.
[5] Ma Q. (2008). Study of the Bias correction and Multi-model Combine of Mesoscale Ensemble Forecast [D]. Nanjing: Nanjing University of information science & technology.
[6] Krishnamurti TN ,Kishtawal CM, LaRow TE,et al.(1999). Improved Weather and Seasonal Climate Forecasts from Multimodel Super ensemble. Science. 285(5433):1548-1550.
[7] Chen L. J., Xu L., Wang Y. G. (2005). Application of Super ensemble to Precipitation Prediction in China during Flood Season [J]. Meteorological Monthly, (5): 52-54.
[8] Zhao SH. R. (2006). Multi-model Consensus Fore-cast for Temperature. Journal of Applied Meteorological Science, 17 (1): 52-58.
[9] Chen Ch. H, Li Ch. Y., Tan Y. K. ,et al.(2010). Research of the Multi-model Super-ensemble Prediction Based on Cross-validation. Acta Meteorologica Sinica, 68 (4): 464-476.
[10] Peng H. Long F H, Ding C.(2005). Feature selection based on mutual information: criteria of max-dependency , max-relevance , and min-redundancy . IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (8): 1226-1238.
[11] Breiman L. (2001). Random Forests. Machine Learning. 45(1):5-32.
[12] Fang K. N. (2012). Random Forest Combination Forecasting Theory and Its Application in Finance. Xiamen: Xiamen University Press.
[13] Zou L, Huang Q, Li A, et al. (2012). A Genome-wide Association Study of Alzheimer’s Disease Using Random Forests and Enrichment Analysis. Sci China Life Sci, 55: 618–625.
[14] Li J. G. and Gao ZH. SH. (2009). Random Forest: An Important Feature Genes Selection Method of Tumor. Acta Biophysica Sinica. 25(1): 51-56.

[15] Fang K. G., Zhu J. P., Xie B. Ch. (2010). A Research into the forecasting of Fund Return Rate Direction and Trading Strategies Based on Random Forest Method. Economic Survey. (2): 61-65.