Feasibility study of typhoon disaster economic loss assessment based on random forest

Fei Wang
Wuhan University of Technology, Wuhan Hubei 430070, China
*Corresponding author’s e-mail: 2469442824@qq.com

Abstract: Based on the disaster-causing factors and the actual disaster-causing factors of typhoon disasters, this paper selects 9 assessment indicators, obtains an effective sample set through noise processing, and uses a random forest algorithm to set up a typhoon disaster economic loss assessment model. The model verification and results show that the random forest model has very obvious advantages in both single-level evaluation accuracy and total evaluation accuracy, and can meet the consistency with the actual evaluation standards. The GINI index reduction value obtained through model feedback can be used as an indicator of importance, and the indicator of the number of disaster victims is the most important factor of economic loss.

1.Introduction
Typhoon is one of the most serious natural disasters in China, especially as the global warming has led to significant warming of the climate, resulting in the strengthening of typhoons in recent years. Super typhoons and strong typhoons have become common occurrences[1], and typhoon disasters are frequent. It has caused a severe impact on the ecological environment and economic development of coastal areas in China. For example, the super typhoon Samny in 2006, Typhoon Mesa in 2005, and the recent super typhoon Mangosteen all caused serious casualties and economic losses. Under the influence of the strongest typhoon Sangmei in China since 1949, more than 500 people died and the affected population was 6.65 million, causing direct economic losses of 19.65 billion yuan. Therefore, it is very important to further strengthen early warning research on typhoon disasters and post-disaster data analysis and learning. China started late in the study of typhoon disasters. Wang Jingai et al. studied the regional chain laws of typhoon disasters in the southeastern coast[2]. Wang Xiurong conducted an assessment study on the comprehensive level of typhoon disasters[3]. This article will start with the economic losses caused by the typhoon and its related influencing factors[4], explore the impact of these influencing factors on the economic losses of the typhoon, and classify and evaluate the severity of the economic losses.

2.Establishment of a random forest model
The random forest model is composed of many decision trees, so the Bagging method is used to obtain the sub-training set {D1, D2, ..., DK} from the total training set D to construct each decision tree[5]. This method uses Bootstrap to repeat Independent random sampling, and then use the sub-training set to build a decision tree. The decision tree is a sub-model of random forest. The final classification result of random forest is the total evaluation of the voting results of each decision tree. The core algorithms for building decision trees currently include ID3, C4.5, and CART. The CART algorithm is used in this article. The CART algorithm is a standard binary tree classification algorithm. The value
of each classification node feature is "yes" and "no", so that the decision tree is equivalent to recursively dichotomize these features. CART uses the GINI coefficient minimization criterion to Perform feature selection to generate a binary tree, the calculation method is as formula (1):

$$Gini(t) = \sum_{j=1}^{k} p_{j|t} (1 - p_{j|t}) = 1 - \sum_{j=1}^{k} p_{j}^2$$

(1)

Where $Gini(t)$ is the Gini index at node t; $p_{j|t}$ is the probability of economic loss grade occurring at node t, and k value represents the number of loss grades.

![Figure 1: A single classification decision tree](image1)

The random forest model can calculate the importance of each sub-indicator\(^6\), that is, by calculating the Gini index reduction value $DGini$ of the economic loss feature at its corresponding node, and then summing the $DGini$ value of the feature in the forest and then comparing this value with The feature $DGini$ value is more important than the feature, such as formula 2 and formula 3:

$$DGini = Gini(t_{w-1}) - Gini(t_w)$$

(2)

$$Q_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} DG_{ij}}{n \sum_{k=1}^{n} \sum_{j=1}^{k} DG_{kij}} \times 100\%$$

(3)

Where $w$ is the order number of nodes, $n$ is the number of random forests, $t$ is the number of nodes in the decision tree, $Q_k$ is the percentage of the k-th indicator in all indicators, and $m$ is the overall number of related risk indicators. The final random forest model is shown in Figure 2:

![Figure 2 Random forest model generation steps](image2)

3. Case analysis
This article selects tropical cyclone data from 1985 to 2012 for research, and at the same time carries out noise processing on typhoon data, and finally retains all typhoon samples that landed in or
generated in China, and analyzes the characteristics of disaster factors in the sample that may affect economic losses, making the model classification results can better fit the real situation.

### 3.1 Index Table for Evaluation of Economic Loss in Wind Disasters

Determining the characteristics of disaster factors that affect the level of economic loss is an important part of model construction. It is known that there are many disaster factors that cause economic loss levels, so there are also many choices as feature classes. Usually when we research something, we usually work on two aspects: the internal attributes of the object and the external environmental impact. Therefore, this article refines the characteristics of the typhoon internal hazard and disaster factors to obtain the typhoon accompanying attributes and typhoon Four third-level indicators of internal attributes, personnel disaster indicators, and the number of property losses. Finally, 9 representative four-level indicators are selected as the feature class of the model sample, as shown in Table 1:

| First-level indicators | Secondary indicators | Third-level indicators                      |
|------------------------|----------------------|--------------------------------------------|
| Typhoon disaster       |                      | Maximum precipitation in the region        |
| economic loss level     |                      | Maximum daily rainfall in the area         |
|                        |                      | Number of days                             |
|                        |                      | Minimum central air pressure               |
|                        |                      | Login speed                                |
|                        | Disaster factor      | Number of victims                          |
|                        |                      | Casualties                                 |
|                        |                      | Affected area of crops                     |
|                        |                      | Number of houses collapsed                 |

Considering the single index grading standards for China's typhoon disaster levels are catastrophes, catastrophes, moderate disasters, minor disasters and minor disasters, the economic loss levels are divided into four categories according to the actual situation of the samples over the years, considering the economic loss span of minor disasters and minor disasters not large, so the small disasters and minor disasters are unified into the fourth level, as shown in Figure 2:

### 3.2 Parameter setting and model establishment

In the randomforest software package environment of matlab, first import the training set data to establish a random forest model, and at the same time optimize the model parameters, namely ntree and mtry values, these two parameters represent the number of decision trees in the random forest and the formation of decision trees. The number of branches in the process is compared with the OOB error value under different parameter values, and finally the parameter values of ntree = 1000 and mtry = 5 are selected. At this time, the total generalization error OOB value of the model is 0.1194.
3.3 Index importance calculation
The average Gini index decrease value calculated by the model during the establishment of the training set can represent the result of the importance of the indicator[8].

Table 3 The importance of indicators on economic loss

| Index                        | Average Gini index decrease | % Of importance |
|------------------------------|-----------------------------|-----------------|
| Maximum total precipitation  | 4.2092                      | 3.54            |
| Landing wind speed           | 11.0686                     | 9.31            |
| Minimum air pressure         | 8.5470                      | 7.19            |
| Number of days               | 2.9802                      | 2.51            |
| Maximum daily rainfall       | 4.4265                      | 3.72            |
| Number of victims            | 48.5894                     | 40.85           |
| Casualties                   | 6.4605                      | 5.43            |
| Affected area of crops       | 15.1803                     | 12.76           |
| Number of collapsed houses   | 17.4791                     | 14.70           |

According to the data in the table, it can be known that the number of victims is the most important evaluation index of typhoon disaster economic loss[9], with an importance ratio of 40.85%, followed by crop disaster area of 12.76%, house collapse number of 14.70%, landing wind speed 9.31% and minimum air pressure 7.19% It is also an important evaluation index, and the impact of maximum total precipitation 3.54%, duration 2.51%, maximum daily rainfall 3.72% and casualties 5.43% on the model evaluation is not very obvious. It can also be found that the disaster-affected factors are more important than the typhoon internal disaster factors in the typhoon disaster assessment model.

3.4 Model comparison
In this paper, the support vector machine model and the improved random forest model are compared in the typhoon disaster economic loss level evaluation. The same sample is used for testing. From the table, the total evaluation accuracy of the support vector machine model is 91.67%, while the improved random forest model is 68.06%, the accuracy of the improved random forest model is significantly higher than that of the support vector machine model, and the assessment accuracy of the second, third, and fourth levels of the improved random forest model is significantly higher than the support vector machine model. It shows that the improved random forest model has better self-validation function, no deviation evaluation, and more efficient calculation.

Table 4 Comparison of model confusion matrix

| Confusion matrix | Improved random forest model | Support Vector Machine Model |
|------------------|------------------------------|------------------------------|
|                  | 1    | 2    | 3    | 4    | Accuracy | 1    | 2    | 3    | 4    | Accuracy |
| 1                | 4    | 0    | 0    | 0    | 100%     | 4    | 0    | 0    | 0    | 100%     |
| 2                | 0    | 17   | 2    | 0    | 89.47%   | 7    | 12   | 0    | 0    | 63.16%   |
| 3                | 0    | 2    | 11   | 1    | 78.57%   | 0    | 5    | 9    | 0    | 64.29%   |
| 4                | 0    | 0    | 1    | 34   | 97.14%   | 0    | 1    | 10   | 24   | 68.57%   |
| Total accuracy   |      | 91.67% |      |      |          |      | 68.06% |      |      |          |

4. Conclusion
The model's own verification can be used to obtain the importance of various indicators on the impact of economic losses. For example, in this case, the number of victims who accounted for 40.85% of the importance and 14.70% of the number of house collapses should be paid more attention to and control the resettlement. The number of people and maintenance of damaged houses and buildings can greatly reduce the index of economic loss[10].
After comparing with the support vector machine model, the overall assessment accuracy of SMOTE improved random forest model and the assessment accuracy of each level have been greatly improved, which also proves that the improved random forest model has better assessment stability, the assessment of the economic losses of typhoon disasters has a better effect.

References
[1] Han R, Zhou B, An L. Quantitative assessment of enterprise environmental risk
[2] Wang Jingai, Lei Yongdeng, Zhou Hongjian. Study on the Regional Laws and Adaptation Strategies of Typhoon Disaster Chain in the Southeast Coast of China [J]. Journal of Beijing Normal University (Social Science Edition) (02): 132-140.
[3] Wang Xiurong, Wang Weiguo, Ma Qingyun, et al. Comprehensive evaluation model of typhoon disaster and its application [J]. Meteorology, 2010, 36 (1): 66-71.
[4] Zhang Bin [1], Chen Haiyan [2], Gu Junqiang [1]. Design and development of GIS-based typhoon disaster assessment system [J]. Disaster Science, 2008 (1).
[5] Huang Yulin, Zhao Yongqian, Cao Zheng. Research on Dengue Fever Risk Assessment Based on Random Forest Regression Model [J]. South China Journal of Preventive Medicine, 2019 (1): 26-31.
[6] Chen Junfei, Dong Ran. Research on flood disaster risk assessment based on random forest algorithm [J]. Water Resources Economics, 2019 (3).
[7] X. Chen. Vulnerability diagnosis and assessment of typhoon disaster system at coastal regions[J][J]. Journal of Catastrophology, 2007, 22 (3):6-10.
[8] Kim J M, Son K, Kim Y J. Assessing regional typhoon risk of disaster management by clustering typhoon paths[J]. Environment, Development and Sustainability, 2018.
[9] Hanaoka K, Sugiyasu K, Murao O. Impacts of Typhoon Haiyan and the post-disaster recovery:Analysis based on a situation report and fieldwork in Philippines[C]// General Meeting of the Association of Japanese Geographers. The Association of Japanese Geographers, 2014.
[10] Jianxiong T, Lilai X U, Yanmin L I, et al. Assessing the damage caused by typhoon on urban green space ecosystem service based on UAV remote sensing[J]. journal of natural disasters, 2018.