Neural Computing and Applications (2022) 34:3443–3455
https://doi.org/10.1007/s00521-021-05757-6

SPECIAL ISSUE ON MULTI-MODAL INFORMATION LEARNING AND ANALYTICS ON BIG DATA

Flood disaster risk assessment based on random forest algorithm

Zijiang Zhu[1,2] · Yu Zhang[3]

Received: 13 October 2020 / Accepted: 19 January 2021 / Published online: 21 March 2021

Abstract
With the frequent occurrence of natural disasters, timely warning of flood disasters has become an issue of concern. This research mainly discusses flood disaster risk assessment based on random forest algorithm. This study uses the special functions of GIS to collect, manage, and analyze data to propose a method of flood disaster risk assessment based on GIS. This method is based on the characteristics of natural disaster-causing factors in the study area, selects an appropriate grid size, and finally realizes the function of visual expression of regional disaster risk. First, use ArcGIS10.1 to analyze and integrate each hazard factor into the flood disaster report index model. Second, the random forest algorithm is used as the weight of each parameter of the flood disaster index model. Finally, use ArcGIS spatial analysis tool map algebra function to model, carry out flood risk assessment in different periods, and use spatial analysis function to extract the median value to point function to extract the flood inundation depth of the study area in a specific scenario. In the experimental part, this research uses layer overlay to determine the number and types of affected areas. Using the natural break point method of ArcGIS 10.1 platform, the study area is divided according to the magnitude of the flood disaster risk value. At the same time, there are a total of 85 samples that have experienced flood disasters, of which only six have been misjudged as no flood disasters. Generally speaking, the model prediction accuracy is high. The research results show that the combination of random forest algorithm and GIS technology is convenient for analyzing the spatial pattern and internal laws of flood risk, and has good applicability.

Keywords Random forest algorithm · Flood disaster · GIS · Hazard factors

1 Introduction

Due to the randomness, suddenness, high destructiveness, and wide distribution of time and space of flood disasters, traditional loss assessment methods are low in efficiency and costly, and the assessment results are not ideal. With the continuous and in-depth application of GIS technology in flood disaster monitoring, it is possible to carry out research on flood disaster risk assessment and crop loss assessment based on GIS. GIS technology performs real-time monitoring of flood disasters. Random forest algorithm can carry out regionalization and loss assessment of agricultural flood disaster risk. The combination of random forest algorithm and GIS technology can not only grasp the possibility of flood disaster in time, predict the degree of loss, and temporal and spatial distribution of losses can also help relevant flood prevention and mitigation departments and agricultural departments to actively respond to flood disasters, and formulate flood prevention and mitigation plans in advance according to the risk levels of different regions.

The ability of crops to withstand flood disasters is poor, especially early crops. The climate zone in which my country is located makes the frequency of flood disasters high and the intensity of floods is high. Flood disasters have become one of the main disasters that cause crop production reduction or no harvest in my country, which is a serious obstacle. And destroy food security and social production. The real-time monitoring of GIS technology

✉ Yu Zhang
201701003@jyu.edu.cn

Zijiang Zhu
zzjgwng@sina.com

[1] School of Information Science and Technology, South China Business College, Guangdong University of Foreign Studies, Guangzhou 510545, Guangdong, China

[2] Institute of Intelligent Information Processing, South China Business College, Guangdong University of Foreign Studies, Guangzhou 510545, Guangdong, China

[3] School of Geographic Science and Tourism, Jiaying University, Meizhou 514015, Guangdong, China

allows us to take measures in advance to minimize the economic loss caused by flood disasters, and to turn the flood disaster prevention work from passive to active, from low efficiency to high efficiency. The use of random forest algorithm to study the risk analysis and evaluation of flood disasters not only provides important reference value for disaster prevention command and dispatch, rescue and relief, emergency response and agricultural flood disaster insurance, but also for crop production management and agricultural disaster compensation, planting structure adjustment, yield forecasting is of great significance.

The random forest algorithm is very versatile. Polan investigated and optimized the random forest algorithm for automatic organ segmentation, explored the limitations of the random forest algorithm suitable for CT environments, and proved the accuracy of segmentation in feasibility studies for pediatric and adult patients. He used Matlab and FIJI's TWS plug-in to classify seven materials: the random forest algorithm uses 200 trees, and each node randomly selects two features. He analyzed the dice similarity coefficient (DSC) calculation between the manually segmented images from 21 patient image parts and the random forest algorithm segmented images. His research process is too complicated [1]. Xu proposed a real-time traffic classification method for power services based on an improved random forest algorithm. On the basis of analyzing the characteristics of real-time traffic in the secure access of electric power services, the traditional random forest algorithm is improved. He pruned the random forest based on the edge weight to improve the real-time classification, and edited the new sample data to improve the classification accuracy. Based on the improved algorithm, he designed the real-time traffic classification process in the secure access of power services. Finally, the feasibility and effectiveness of this method are verified by taking a province electric power enterprise's safe access to real-time traffic classification as an example. His research lacks data [2]. Joshuva believes that wind energy is converted into electrical energy using rotating blades connected to a generator. He uses structural health monitoring to diagnose downtime regularly, which can reduce downtime. These are considered as pattern recognition problems, which include three stages, namely feature extraction, feature selection and feature classification. He extracted statistical features from vibration signals, used J48 decision tree algorithm for feature selection, and used random forest algorithm for feature classification. His research lacks data [3]. He believes that the study of Mendelian disease and the identification of its causative genes are of great significance in the field of genetics. He calculated the genetic pathogenicity prediction (GPP) score through a machine learning method (random forest algorithm) to evaluate the pathogenicity of genes. Apply the GPP score to the test gene set. His research lacks data [4]. Wang Y believes that inspecting the steel surface is important to ensure the quality of the steel. He proposed an improved random forest algorithm (OMFF-RF algorithm) with optimal feature set fusion for distributed defect recognition. He extracted and merged the histogram of the oriented gradient (HOG) feature set and the gray-level co-occurrence matrix (GLCM) feature set to describe local and global texture features, respectively. Secondly, in view of the small number of samples of the distributed defect images and the high dimensionality of the extracted feature set, a random forest algorithm is introduced for defect classification. His research lacks data [5].

This research first clarifies the background and significance of the topic. Then, for the principles and methods of regional flood disaster risk assessment, relevant theories are analyzed, summarized and refined, the principles and methods of flood disaster risk assessment are explained, the evaluation model and method of this research are proposed, the research area is introduced, and the research is explained. Using data and sources, preprocessing the relevant data, and establishing a spatial database under the ArcGIS platform. Finally, for the construction and index analysis of the flood disaster risk evaluation index system, analyze the factors that affect the flood disaster risk, introduce the principles and methods for selecting the evaluation index, analyze the flood disaster risk evaluation index, determine the composition of the evaluation index system, and analyze the selected items. The indicators are analyzed and standardized.

2 Flood disaster risk assessment

2.1 Flood disaster

The remote sensing monitoring method of rainstorm and flood disaster is relatively mature at present, and the method has the characteristics of diversified, deep and comprehensive. In the research of this article, it is found that the method of extracting the information of rainstorm and flood disaster only realizes the determination of the flood disaster area by extracting the distribution range of the flood disaster water body, and has not studied the specific dynamic process of crop damage in the cultivated area [6, 7]. And more research is limited to a single remote sensing data source, and more attention is paid to the use of satellite remote sensing data to detect and extract information about sudden changes in flooded cultivated land caused by heavy rains and floods [8].

Traditional manual survey and monitoring of disasters are not only time-consuming, labor-intensive and expensive, but also the survey data cannot reflect the dynamic

information of crop damage. Remote sensing technology can make up for many shortcomings of traditional disaster monitoring, and realize real-time, macro, and large-scale ground monitoring. Obtain information on the dynamic changes of crop growth affected by the disaster area of a wide range of cultivated land and heavy rain disasters [9, 10]. The management and analysis functions of GIS technology combined with multi-source satellite data can quickly and comprehensively analyze disaster information. During the heavy rain and flood disaster, the weather is bad, and the remote sensing data obtained by satellites are mostly affected by the cloud layer, which leads to the extremely unsatisfactory quality and quantity of the remote sensing data that captures changes in the ground state [11].

2.2 Application of GIS in flood disaster risk assessment

Geographic information system (GIS) provides a powerful technical means for spatial data processing, analysis, management, and mapping. It is widely used in the research and practice of flood disaster risk assessment [12, 13]. Geographical information system can effectively manage and update various spatial data in flood disaster risk assessment. A large amount of data is involved in flood disaster risk assessment. These data include historical flood data, socio-economic data, remote sensing images, geographic thematic data, and various vector data and raster data. Most of these data have both attributes and space. Location, these data are not only complex but also massive. Geographical information systems can store these data in sub-databases, layers, types, and plots [14]. If the image function is converted into a binary image to represent the shape area of the GIS partial discharge map, the image $p + q$ moment is defined as:

$$M_{pq} = \sum_{x}\sum_{y} x^p y^p f(x,y)\ (p,q = 0,1,2,\ldots) \tag{1}$$

Among them, (x, y) is the center of mass [15, 16]. Geographic information system provides powerful analysis tools for flood disaster risk assessment research. In flood disaster risk assessment, sometimes discrete precipitation monitoring data are spatially gridded, and the spatial interpolation analysis function of GIS can quickly realize this demand. Sometimes, it is necessary to comprehensively calculate various risk indicators according to a certain model to obtain the final risk result. The grid mathematical operation function of GIS is quite powerful, which can quickly calculate a large amount of grid data. The spatial statistical analysis of GIS can help us to make statistics according to plots and fields. At present, it is relatively mature to use a series of analysis functions of GIS technology to carry out flood disaster risk assessment and zoning research [17]. Geographic information system is the best method for making flood disaster risk maps. GIS software can directly display and output flood disaster risk related data and evaluation results as two-dimensional maps or three-dimensional images, and produce various flood disaster risk maps [18, 19].

2.3 Random forest algorithm

In the random forest algorithm, a large number of decision tree classifiers are constructed first, and then, these classifiers are used to vote on the test samples, and finally the final decision result is made through the principle of majority dominance. The model selection of random forest is to select a specified number of suitable classifiers from the many decision tree classifiers constructed, and then use these classifiers to vote on the test samples [20].

There are two main steps in the random forest algorithm. The first is to build a large number of decision tree classifiers, and the second is to integrate these decision tree classifiers into voting [21, 22]. In the original random forest (the original random forest, referred to as RF) model algorithm, it uses the method of all selection integration when the model is selected. That is to say, if k decision trees are built, then these k. All decision trees are selected for integration, and then, the random forest model containing k decision tree classifiers is used to vote and finally vote [23]. Suppose $S = \{(x_1, y_1)\ldots(x_N, y_N)\}$ is a sample set, and the classification error of the classifier f on the sample set is:

$$g\_\text{error} = \frac{1}{N}\sum_{(x_i,y_i)\in S} 1_{\{f(x_i)\neq y_i\}} + P_\text{D} \tag{2}$$

Among them, N is the size of the sample set (total number of samples), and $1_{\{f(x_i)\neq y_i\}}$ is the indicator function [24, 25]. Because in the static model selection method, the model selection is only performed once, and the built integrated model does not need to be changed. Therefore, the calculation amount of this method in the model selection stage is relatively low, and the time complexity will be relatively small. In the static model selection method, the decision tree classifier is generated based on training samples at the beginning, so these classifiers are more dependent on the training samples in the probability distribution. When several classifiers are selected from these decision tree classifiers for integration, this ensemble model will to some extent be more biased toward correctly classifying samples with the same probability distribution as the training samples [26]. The strength of the classifier set $\{h(X)\}$ is the expectation of the marginal function of the random forest:

$$s = E(mr(X, Y)) = \frac{1}{n}\sum_{i=1}^{n}\left(Q(x_i, y) - \max Q^c\left(x_i, y_j\right)\right) \quad (3)$$

If there is a large difference between the test sample and the training sample in the probability distribution, the ensemble model may make wrong classification predictions for the test sample. And because the ensemble model does not change after construction, when there are large differences between the test samples, then using the same ensemble model for prediction may also lead to incorrect classification. For a given classifier, the input vector x and its corresponding output y, define the interval function of the sample point (x, y) as follows:

$$mg(x, y) = av_k I(h_k(x) = y) - \max_{j \neq y} av_k I(h_k(x) = j) \quad (4)$$

Among them, $I(h_k(x) = y)$ is the indicator function.

3 Flood disaster risk assessment experiment

3.1 Data sources

The data used in this study includes remote sensing data, geographic basic data, and statistical data. Remote sensing data mainly includes land use data, which is obtained by unsupervised classification and interpretation of Landsat remote sensing images in 2000 using EDRAS software. The basic geographic data includes digital line maps (including administrative area boundaries, traffic, rivers, and other elements), topographic data (contour lines, elevation points) and topographic data are vectorized from the 1:250,000 topographic map of our province. Statistics include rainfall data, population data, economic data, and agricultural data.

3.2 Modeling of random forest model

In the algorithm package random forest of R language, two important parameters need to be carefully selected. These two parameters are the number of preselected feature attributes mtry and the number of decision trees in the random forest model ntree. Generally speaking, ntree is usually set to 500, the value of mtry is generally the root of the sample format, the number of mtry is too small, and a decision tree is generated, which will cause the classifier to overfit and cause the classification accuracy to be low. If it is too large, it will cause the running speed to become much slower. Simply put, if the mtry gets smaller and smaller, the correlation between the trees will become smaller and smaller, and at the same time, the classification accuracy will decrease. Therefore, in the random forest model, the number of decision trees must be set reasonably. If the number of decision trees is small, the training will be insufficient, and if the number of decision trees is relatively large, the model will be greatly increased. The amount of calculation. The algorithm is based on the randomForest() function of the R language machine learning package. Therefore, in this experiment, 375 samples are randomly selected for replacement, and six features are randomly selected each time, and 100 decision trees are selected to build each the decision tree is fully grown. Natural disaster risk assessment based on GIS is the use of special functions of GIS: collecting, managing, analyzing data, etc. This method selects the appropriate grid size according to the characteristics of the natural disaster-causing factors in the study area, and finally realizes the visual expression of the disaster risk in the scenic area.

3.3 Establishment of flood disaster index model

This study uses the flood disaster index model to evaluate the flood risk in different periods in Guanzhong area. Firstly, use ArcGIS10.1 to analyze and integrate each hazard factor into the flood disaster report index model. Secondly, experts in related fields (hydrology, geographic information system, urban science) are invited to decide the relative importance of hazards. Random forest algorithm is used as a method to determine the weight of each parameter of the flood disaster index model. Finally, use ArcGIS spatial analysis tool map algebra function to model, carry out flood risk assessment in different periods, and use spatial analysis function to extract the median value to point function to extract the flood inundation depth of the study area in a specific scenario. Determine the number and types of affected areas by overlaying layers. Using the natural break point method of ArcGIS 10.1 platform, the study area is divided according to the magnitude of the flood disaster risk value. The risk is divided into five levels: disaster loss in extremely low-risk areas is defined as $< 5\%$, low-risk areas are defined as 5–10%, medium-risk areas are defined as 10–50%, and high-risk areas are defined as disaster losses. It is defined as 80–100% for 50–80%, high-risk area disaster loss. Obtaining flood risk prevention and control maps through evaluation is helpful to flood control and management decision-making. Finally, according to the spatial analysis function of GIS, the disaster area statistics and calculation are carried out, and finally the disaster level evaluation is carried out.

3.4 Feature selection

In this paper, rainfall, elevation, and rainfall are selected as features. The rainfall is divided into 12 dimensions such as January rainfall, February rainfall up to December rainfall,

and a total of 15 features. Before the model calculation, in order to verify will there be any correlation between the features of our fifteen dimensions, whether there is a connection between them, if there is a correlation, we need to carry out preliminary data cleaning work, such as reducing the dimensionality of the data, so then we need to diagnose multiple collinearity for these characteristic attributes.

3.5 Model sample data selection

The model sample data select the historical flood disaster data of 34 districts and counties from 2000 to 2010. The points where the flood disaster has occurred are marked as 1, and the points that have not occurred are marked as 0, so a total of 375 sample data. 15 features constitute the final sample data set. Generalization ability is the most direct manifestation of model effect.

3.6 Early warning model process design

The early warning model process is as follows:

(1) Sort out the historical example samples of fault tripping under flood disasters, construct the misclassification cost function according to the actual distribution of the samples, and calculate the misclassification cost of the fault samples and the normal samples.
(2) Using the Bagging method to extract the sample set for the sample sample set to form k sample subsets and the corresponding out-of-bag data set.
(3) For k sample subsets:
    ① Select m features from the feature space of the original data set to form a feature subset.
    ② Calculate the misclassification cost reduction value Rcc of each feature in the feature subset.
    ③ Select the largest feature attribute of Ree to split the node.
    ④ Repeat ②–③ until the samples in the sample subset are classified or reach the maximum node level, and finally a decision tree model is generated based on cost-sensitive learning and the flood fault warning model of random forest.
(4) Each decision tree uses the out-of-bag data set corresponding to the sample subset for classification testing.
(5) For the prediction sample set, the final classification result is obtained through the weight voting of k decision trees.

3.7 Model training

First, the normal samples in the example sample set are down-sampled to extract 500 normal samples, and the failure samples are not expanded. The cost function is determined according to the actual distribution of 538 samples. Subsequently, Borderine-SMOTE oversampling was performed on 20 of the 38 fault samples, which expanded the number of minority samples to 216; the non-fault samples were denoised, and the non-fault samples were further reduced to 216. There is a total of 432 samples in the training sample set based on the hybrid sampling algorithm. Since the method in this paper has not been deployed in the actual system, some historical samples are used as test samples to verify the effectiveness of the method in this paper. In the calculation example, 292 samples (non-fault sample: fault sample = 1:1) are input into the model to train the model, and the remaining 140 samples including 18 instance fault samples are used as subsequent prediction samples. Therefore, in this experiment, the training sample and the test sample set are distinguished at a ratio of 7:3. The number of training sample sets is 260, of which the number of positive samples (marked as samples that have experienced landslides) is 190. The number of negative samples (marked as samples that have not experienced landslides) is 70: the number of test sample sets is 115, of which the number of positive samples (marked as samples that have experienced landslides) is 85, and the number of negative samples (marked as the number of samples that have experienced landslides is 30): some data of the test sample set are shown in Table 1.

3.8 Establish a risk analysis model

Use the sklearn machine learning package in the Python language to implement the random forest algorithm. 70% of the sample set is used for training, and 30% is used to test the prediction ability of the trained random forest model on new samples. In the process of evaluating the accuracy of the model, the accuracy of the model is 98.1% on the training set and 94.3% on the test set, which means that the model has good predictive ability. According to the prediction results of the model, ArcGIS software is used to

Table 1 Partial data of the test sample set

Type	X_1	X_2	X_4	X_5
1	13	3	31	110
1	10	1	31	183
0	14	3	19	133
0	11	3	31	138

generate the prediction map of the railway network failure probability affected by the heavy rainfall disaster.

3.9 Selection of risk samples

According to detailed disaster site survey data, various evaluation indicators and relevant government data (reports, planning documents, etc.) to determine the disaster risk level of the sample. First, calculate the disaster impact coefficient of each county in the study area according to the attributes of the disaster scale, the density of the disaster point, the number of people threatened, the threatened property, the number of damaged houses, and the area of the damaged road in the detailed investigation data of the impact coefficient. The law preliminarily delineates the disaster risk level of each county: high, high, medium, and low risk, and counties without flood disasters are classified as low-risk areas. Secondly, through various evaluation indicators, relevant government data and previous research results, the classification risk level is revised. Then, input various indicators and risk levels of the selected samples into the model to form disaster risk classification rules. Finally, according to the above rules, all the data to be tested in the study area are re-input into the RF model to predict the flood disaster risk level of the study area. The weights and classification criteria are implicit in the inherent rules of the data. The modeling and calculation of the random forest model are realized by Rstuo software platform using R language programming.

3.10 Urban expansion and mountain disaster risk coupling

Taking the flood disaster risk change rate of the study area as the explanatory variable, taking the urban expansion intensity and the effective grain size change rate as the explanatory variable, the spatiotemporal geographic weighted regression model is used to further analyze the spatiotemporal coupling relationship between urban expansion and disaster risk. The calculation of regression coefficients is implemented in the GTWR plug-in in ArcGIS 10.4 software. What needs to be clear is that the disaster risk evaluation results in this article are risk levels, which are type variables. Participating in the model regression directly by assigning values may cause the regression results to have lower accuracy. Therefore, this study uses the contribution of each index obtained by the random forest model as the index weight, combined with each index value, and uses the weighted sum method to calculate the flood disaster risk value of each county unit in the study area.

3.11 Flood disaster risk assessment design

Compared with other disaster risk studies, this study starts from the dynamics of disaster risk, comprehensively considers the dynamic change characteristics of disaster factors in the study area, and incorporates the static and dynamic factors of each county into the study area. In the disaster evaluation system, it can objectively reflect the temporal and spatial evolution of flood disaster risk in the study area. It should be pointed out that in the process of selecting and assigning levels to disaster risk samples for each year, it is inevitable that they will be subjectively affected by individuals, which may have an important impact on the evaluation results. Therefore, this study first carried out the disaster risk of the study area in 2015 by constructing a random forest evaluation model, and used the correlation between the indicators built in the model and the correlation with disaster risk to compare the data in 2000, 2005, and 2010. The index data are also incorporated into the random forest model, and the index data of each county in 2015 are used as training and testing samples, so that the results of disaster risk zoning in each year can be obtained.

4 Flood disaster assessment analysis

4.1 Analysis of flood disaster assessment results

The disaster risk samples of the study area in 2019 are selected and assigned corresponding risk levels. There are 10 disaster risk sample counties for each level, for a total of 50 counties. Based on the fivefold cross-validation method, the selected disaster risk samples are input into the random forest model. After many debugging and parameter analysis, the classification tree of the random forest model is set to 1000, and the variables at the nodes are set to 4 according to the model parameter setting requirements. Run the random forest model to obtain the accuracy evaluation results of the model. From the training results of the random forest model, the RF model has a good effect in flood disaster risk assessment, with an AUC value of 0.952, indicating that the RF model has good robustness and generalization ability and high evaluation accuracy. A total of 278 points to be tested for various indicators of each county were input into the trained random forest model, and the results of flood disaster risk zoning in the study area in 2019 were further obtained. Taking the 2019 evaluation results as the training and testing samples, the random forest model was run again to obtain the flood disaster risk classification results of the study area in 2016,

2017, and 2018. The flood disaster risk classification results are shown in Table 2.

On the whole, the flood disaster risk in the study area is roughly low in the west and high in the east. The high-value areas are roughly distributed along the fault zone, river valleys, and low hills. The high-risk areas and high-risk areas are mainly distributed in the long fault zone, along the river valley, hills and ridge-valley areas, and river valley areas. The reason is that the above-mentioned areas are located near broken rivers, with active rainfall, at the same time, the population and economic industry layout are relatively dense, the infrastructure is relatively complete, and they are easily affected by flood disasters. The medium-risk area is adjacent to the high-risk area and the higher-risk area, mainly concentrated in hills, river valleys and some districts and counties at the edge of the river valley; the lower-risk areas are in the low mountain and hilly area, the plateau area and the northwest plateau area. There are distributions, and low-risk areas are mainly distributed in plateau districts and counties, and they are widely distributed. Although the plateau area has the conditions for the development of flood disasters, due to its sparse population, low level of economic development, and imperfect infrastructure, it is not easy to be threatened by flood disasters, so it has become a low-risk area for flood disasters. From the perspective of time change, from 2016 to 2019, the dynamic change characteristics of flood disaster risk in the study area are very obvious. In 2016, the number of high-risk and high-risk counties in the study area accounted for 25.93% of the total counties. Medium-risk areas dominate the total number of counties, accounting for 33.81%. There are more counties in the lower-risk districts, accounting for about 26.62% of the total study area. The low-risk area accounts for approximately 14.39% of the total study area. In 2017, the proportion of counties with various levels of disaster risk was basically the same as that in 2016, and the spatial distribution of various types of risks changed little. From 2017 to 2018, the high-risk and high-risk areas have increased rapidly, and their rising rates were 18.75% and 4%, respectively, and the proportion of the two in the total counties rose rapidly to 7%. The number of other risk-level counties showed a downward trend. The newly added high-risk areas are mainly distributed along fault zones and river valleys, and the disaster risk of mountain ranges close to rivers has also increased significantly. This is mainly due to the great damage caused by the earthquake to the counties in the affected area. After the disaster, aftershocks continued, and secondary flood disasters such as landslides, collapses, and mudslides occurred frequently, which further increased the flood disaster risk in the earthquake area. From 2018 to 2019, the high-risk and high-risk areas in the study area have declined, and their proportion in the total counties has dropped to 28%. The proportions of counties in the medium-risk area, low-risk area and low-risk area have all increased, with increasing rates of 22.78%, 5.17%, and 13%, respectively.

4.2 Flood disaster forecast analysis

The specific error between the predicted value and the true value is shown in Fig. 1. It can be seen from the figure that the number of real flood disasters is 70, but due to model errors, 30 samples are judged as flood disasters, and the error rate is about 40%. At the same time, a total of 190 have occurred. Of the samples of flood disasters, only 6 of them were misjudged as having never experienced flood disasters. Therefore, due to the complexity and diversity of flood disasters, there are certain errors. However, in general, especially for floods that have occurred. The prediction accuracy of disasters is higher than that of flood disasters. The analysis of the reasons, a large part of which is because there are 190 positive samples (referring to samples marked as flood disasters) in the sample, and negative samples (referring to marking as no flood disasters). There are only 70 samples, the positive and negative distribution of the samples are not uniform, and the samples contain noisy data, which leads to the prediction accuracy of negative samples is not high enough, but overall, the model prediction accuracy is high. The overall OOB error analysis rate of the final model was 17.69%. It was found that when 100 trees were selected, the residual error of the model was small. Too many trees increase the amount of model calculation, and the effect is not better. Therefore, the random forest model of this study is selected 100 trees. According to the training sample data set of OOB error, the estimated error and prediction error are basically the same, and the trend of the two errors with the value of K is also exactly the same, which shows that the OOB error can be used to estimate the maximum number

Table 2 Results of flood disaster risk classification

Years	High-risk area	Higher risk area	Medium-risk area	Lower risk area	Low-risk area
2016	13	57	95	75	50
2017	16	57	95	76	53
2018	35	70	79	57	36
2019	25	55	97	61	50

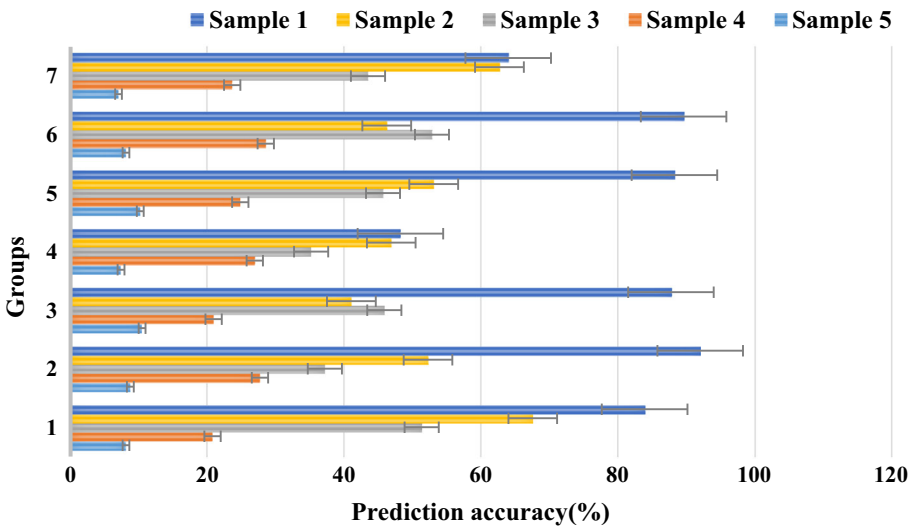

Fig. 1 The error between the specific predicted value and the true value

of trees. A good value range, as can be seen from Fig. 1, OOB error tends to be stable as ntree increases. On the other hand, it shows that the random forest algorithm has better convergence. When choosing between 3 and 8 trees, the error will be relatively low. It can be seen from Fig. 1 that the number of real unflooded disasters is 30, and the predicted value is also 30. The accuracy of predicting flood disasters is 93%. At the same time, there are a total of 85 samples that have experienced flood disasters, of which only 6 was misjudged as no flood disaster. Generally speaking, the model has a higher prediction accuracy. The model can also evaluate the importance of features. In the r language, analyze and explain the importance of features according to the importance command, so that you can understand which features are the more important factors in the process of flood disasters. You can see the ranking of the importance of the variables used in the model, the importance of the variables, that is, assuming that the variable is not in the model, how much influence the error of the model is, slope, rainfall and altitude are more important factors affecting flood disasters.

4.3 Macro coupling analysis of urban expansion and flood disaster risk

In order to better analyze the influence of each characteristic factor on the risk of flood disasters, in the research, the curve of the effect of rainfall on the risk of flood disasters is drawn for the first-ranked factor slope as shown in Fig. 2. The importance of rainfall is ranked the highest. It can be seen in Fig. 2 that the abscissa in Fig. 2 corresponds to the value of the factor, and the ordinate corresponds to the factor's contribution to the probability of flood hazard, that is, the larger the ordinate, the more likely it is to occur. With the increase in rainfall, the possibility of flood disasters gradually increases, which is basically consistent with the laws of geology, and the risk of flood disasters also increases. In the prediction results of the random forest model, flood disasters mostly occur when rainfall is between 15 and 30 mm, and flood disasters rarely occur with low rainfall of 0–15 mm. In summary, compared with the regional natural disaster risk, the natural disaster risk assessment of scenic spots has its own uniqueness. This uniqueness is mainly reflected in the following aspects. First, the subject of natural disaster risk assessment in the study area mainly tourists, and the disaster-bearers of the regional natural disaster risk assessment are the permanent population. Tourists are unstable for the regional scenic spots; secondly, the value of tourism resources will not be considered separately in the regional natural disaster risk assessment. On the contrary, tourism resources play a very important role in the natural disaster risk assessment of scenic spots; third, the acquisition of data for natural disaster risk assessment of scenic spots is more difficult than that of regional natural disaster risk assessment. Therefore, natural disasters in scenic spots. It is difficult to fully quantify the risk assessment. The risk assessment of regional natural disasters generally requires 20–30 years of data. However, most of the tourist attractions in China are not open for a long time. Therefore, whether it is the data record of natural disasters or scenic tourism the statistical data of usually fail to meet the continuity requirements. This urges the natural disaster risk assessment of scenic spots to not completely rely on historical statistics of disaster loss data, but to be flexible. Based on the estimation results of the spatial–temporal geographic weighted regression model, in-depth analysis of the temporal and spatial differences of the impact of urban expansion characteristic measurement indicators on mountain disaster risk. The fitting coefficients of the explanatory variables of

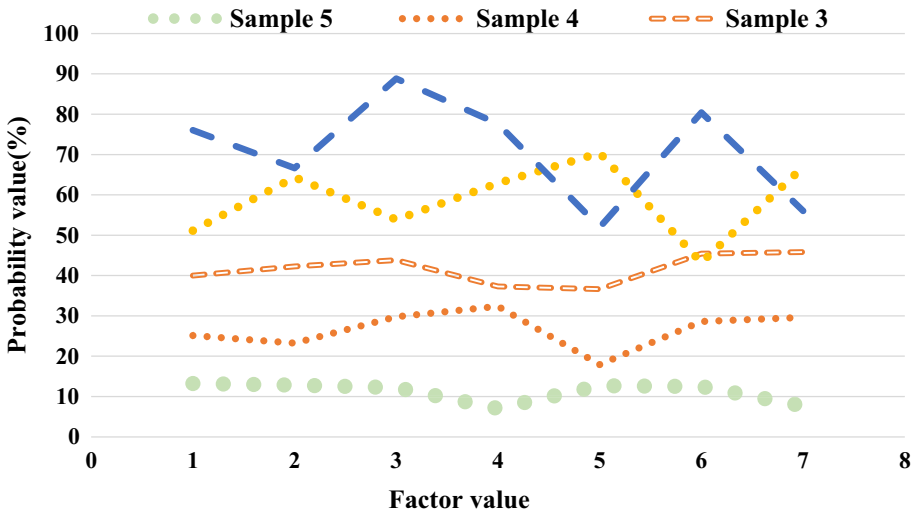

Fig. 2 The effect of rainfall on the risk of flood disasters

the GTWR model vary with space and time, which can express the coupling relationship between the explanatory variables and disaster risk. The statistical results show that the effective grain size change rate of the urban landscape and the urban expansion intensity index have different effects on the disaster risk changes of different counties in different periods. The positive and negative effects coexist, and the regression coefficient varies greatly. As for a certain single-influencing factor, its impact on disaster risk changes also has temporal and spatial differences. Therefore, it is necessary to consider the temporal and spatial heterogeneity of the impact of various indicators on regional disaster risk changes from a local perspective.

4.4 Analysis of risk assessment results

The random forest model has been used to evaluate the risk of landslides in the study area, and a landslide danger zone map based on the slope unit has been obtained. On this basis, the kilometer grid unit and the risk zone layer are superimposed, and the grid processing, assign each grid cell according to the risk value corresponding to the grid, and obtain the landslide risk evaluation of the research area based on the grid cell. Using the grid calculator tool in ArcGIS 10.2, it will be based on the grid. The landslide risk evaluation layer of the unit and the vulnerability evaluation layer based on the grid unit are multiplied to obtain the landslide risk of all grid units in the research area, and then, the natural breakpoint method is used to divide the risk level to obtain. Based on the grid cell-based landslide risk zone map in the study area, the risk assessment result is shown in Fig. 3.

It can be seen that most of the slope units near the main stream of the study area are high-risk and extremely high-risk, while the slopes farther from the study area are less dangerous. The extremely high-risk areas are mainly concentrated in the slopes near the two banks of the study area waters about 0–30 km away from the dam site, the slopes near the right bank of the study area waters 30–90 km away, and a very small slope area at the end of the reservoir. The high-risk areas are mainly concentrated on the slopes on both sides of the waters in the study area at the head of the reservoir, the slopes on the right bank 30–90 km away from the dam site and the slopes on both sides of the waters in the study area 90–120 km. The medium-risk areas are mainly concentrated in the part of the slope on the left bank of the reservoir near the waters of the study area, and most of the other slopes far away from the waters of the study area are low-risk areas and extremely low-risk areas.

Statistics of the area distribution of each dangerous level area and the distribution data of landslide disasters within each different level range, as can be seen from Fig. 3, the area of the extremely high-risk zone, the high-risk zone, the medium-risk zone, and the low-risk zone are 672.20 km^2, 49,447 km^2, respectively, 409.41 km^2, 115,748 km^2, and 904.27 km^2 accounted for 18.48%, 13.59%, 11.25%, 31.82%, and 24.56% of the total area of the study area, respectively. The medium-risk and below areas accounted for 67.63% of the total area of the study area, indicating that most of the slopes in the reservoir area were less affected by flood disasters. Among them, the high-risk area and the extremely high-risk area accounted for 32.37% of the total area, but 76.33% of the study area was included. The number of flood disasters shows that the distribution of flood disasters in these areas is very dense, and the degree of danger is naturally higher, which is consistent with the actual situation. And with the increase in the degree of danger, the corresponding proportion of flood disasters increased from 0.55 to 12.79%. The distribution density of flood disaster points gradually increases with the degree of danger. There is a good positive correlation between the

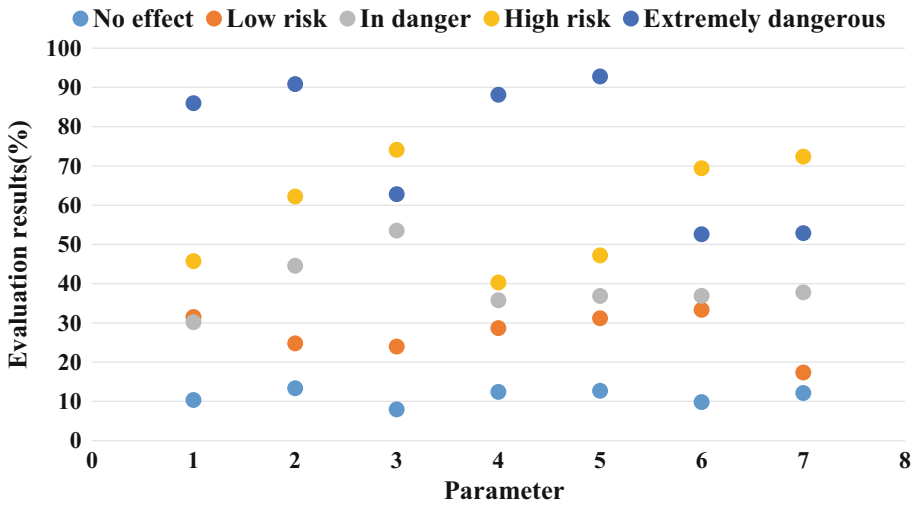

Fig. 3 Results of risk assessment

two, and the correlation coefficient reaches 0.8569, which conforms to the principle of classification in theory. There are 885 grid cells in the extremely high-risk area, accounting for 24.67% of the total area of the study area. These areas are densely populated areas, and the density of buildings and cultivated land is relatively large. Some of them are also built with factories, and most of these areas are also located in high-risk areas of landslides close to the main stream of the study area, so they are affected by floods. The risk of harm is also higher. There are 446 grid cells in the high-risk area of flood disaster, accounting for 12.43% of the total area of the study area. These areas are mainly some areas with relatively dense human activities far away from the two banks of the study area. There are 1015 grid cells in the medium risk area of flood disaster, accounting for 28.29% of the area of the study area. The medium-risk area is mainly distributed in some areas relatively far from the main stream of the study area, and the population density in this area is very low. Human activities are relatively weak, and even if there is a flood disaster, the damage caused is relatively small. There are 1242 grid units in the low-risk area of landslides, accounting for 34.61% of the total area of the study area. The main feature of these areas is that the altitude is usually high, and there is basically no human activity, so there are basically no buildings or cultivated land. The economic loss caused by the flood disaster is very small.

4.5 Analysis of vulnerability zone statistics results

Many research areas are simultaneously affected by flood disasters, but the depth of submergence is obviously different. Then, extract the water depth information of the flood disaster raster data to the point vector data of the research area. The process is: spatial analysis tool in ArcGIS10.2 software → extraction analysis tool → value extraction to point tool, and finally obtain the affected research area Use the "sampling" function in the ArcGIS natural break point method to ignore the submerged depth value of the non-coastal study area, and then use it for five-level division, that is, the analysis result is extremely high (2.62–4.49 m), medium (0.89–1.53 m), low (0.47–0.89 m), very low (0–0.47 m) five grades, assigned 5, 4, 3, 2, and 1, respectively. This natural breakpoint method is based on the characteristics of the data itself, and the classification breakpoints are selected based on the principle of the minimum sum of variation in each level. Therefore, the data classification effect is good, and it is also widely used in GIS analysis. The result of risk assessment is the basis for formulating measures to reduce flood disaster risk in coastal research areas. The main content of the assessment is to classify flood disaster risk. According to the flood hazard and the vulnerability of the disaster-bearing body in the coastal research area obtained in the first two sections of this chapter, the basic model of risk assessment is adopted: risk = risk × vulnerability, and then, the risks mentioned above are passed. The matrix table determines the flood disaster risk level, and finally uses ArcGIS to spread the assessment results under the two scenarios to show the spatial differences of risks. The analysis of the statistical results of the vulnerability zone is shown in Fig. 4. Through the analysis of the statistical results of the vulnerability zone, it can be seen that the area of low vulnerability and low vulnerability of the flood disaster board in the study area is 1767 km^2, accounting for 49.22% of the total area of the study area, which indicates that nearly half of the study area is vulnerable to landslides. The damage is low, and most of these areas belong to uninhabited areas in the study area or areas more than 2 km away from main traffic roads. The vulnerable area in the flood disaster is 1015 km^2, accounting for about 28.27% of

the total area of the study area. Most of these areas are distributed in areas less than 2 km away from roads, but less densely populated. The area with extremely high vulnerability and high vulnerability to flood disasters is 808 km^2, which accounts for about 22% of the total area of the study area. The areas near the main waters of the study area of 10–30 km and 50–80 km. These towns and regions have relatively high population densities, and there are a large number of residential houses and industrial plants that are closely related to human production and life. Once floods occur in these areas, it will cause great losses to humans and buildings, and the natural vulnerability is higher. If experiencing strong rainfall in a short period of time, it is very easy to cause secondary disasters such as landslides and mudslides, blocking railway traffic in mountainous areas, and the inherent geological weakness in southwestern China also increases the risk of secondary disasters affecting line operations; in addition, the northwest, southwest, and north China regions are all in the active fault zone. The surface soil accumulates loose sediments and lacks vegetation protection. If extreme rainfall is encountered, it is easy to cause disasters such as mudslides and landslides, threatening the safety of the route. Combined with the above analysis, it is necessary to strengthen effective early warning for areas with harsh natural conditions. In general, the intensity of human activities in areas with extremely high risk of flood disasters and high-risk areas is high, and human production activities are frequent. Once a flood disaster occurs, it may cause heavy casualties and property losses. Therefore, we should focus on the extremely high and high landslides. Risk areas should do a good job in disaster prevention, mitigation, monitoring and early warning, and establish corresponding protection projects for key dangerous landslides; in areas with moderate risk of flood disasters, the intensity of human activities is weak, and there are fewer load-bearing bodies. Even if a flood disaster occurs, it will cause threat is also small, and preventive, monitoring, and early warning measures can be taken. For the low-risk areas of flood disasters, they basically belong to areas with no human activities, and the losses caused by flood disasters are very small, but corresponding disaster prevention and mitigation work is also needed.

5 Conclusion

In general, the intensity of human activities in areas with extremely high risk of flood disasters and high-risk areas is high, and human production activities are frequent. Once a flood disaster occurs, it may cause heavy casualties and property losses. Therefore, we should focus on the extremely high and high landslides. In risk areas, disaster prevention and mitigation, monitoring and early warning should be done well, and corresponding protection projects should be established for key dangerous landslides. In view of the characteristics of flood disaster risk assessment, focusing on key issues such as the time range, spatial scale, and flood disaster type of flood disaster risk assessment research, it is concluded that compared with the current mainstream geological disaster risk assessment, the space for flood disaster risk assessment scope is relatively special, and it is generally a striped area; the time range of the evaluation and the type of flood disaster are relatively clear, and the evaluation is carried out separately according to the rainfall scheduling cycle. The flood disaster objects studied need to consider the specific boundary to analyze the water level.

This research established GIS-based data to obtain more accurate comprehensive information of storm disasters, and

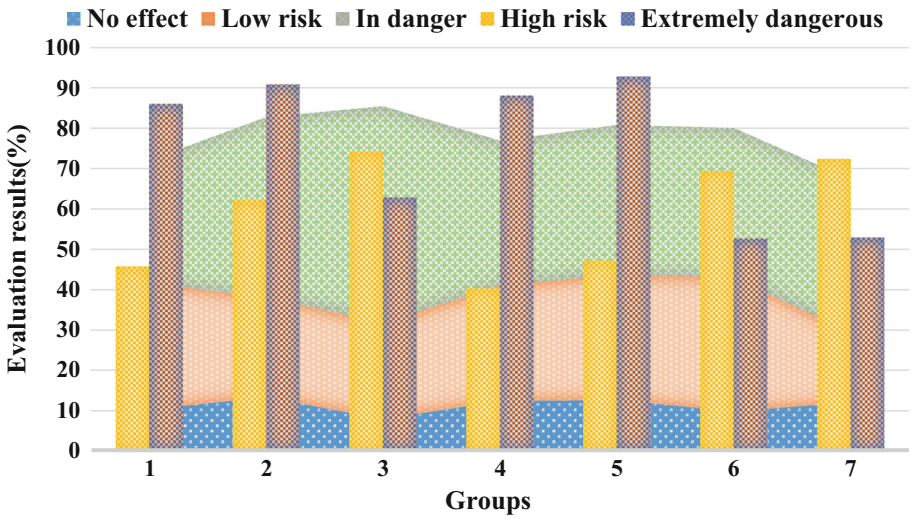

Fig. 4 Vulnerability zone statistics results

realized comprehensive utilization of information to obtain comprehensive information (scope, disaster starting period, duration, and post-disaster recovery status). Based on the time series change data of the random forest algorithm, the calculation rules are established and the disaster feature extraction is performed. It can extract rich spatiotemporal dynamic information including the starting period and duration of the disaster affected, and the coverage is large. It is very useful for large-scale regional disaster monitoring and evaluation.

This research first clarifies the background and significance of the topic. Then, for the principles and methods of regional flood disaster risk assessment, the connotations of relevant terms are explained, related theories are analyzed, summarized, and refined, and the principles and methods of flood disaster risk assessment are described, and the evaluation model and method of this research are proposed, introduce the general situation of the research area, explain the data and sources used in the research, preprocess the relevant data, and establish a spatial database under the ArcGIS platform. Finally, for the construction and index analysis of the flood disaster risk evaluation index system, analyze the factors that affect the flood disaster risk, introduce the principles and methods for selecting the evaluation index, analyze the flood disaster risk evaluation index, determine the composition of the evaluation index system, and analyze the selected items. The indicators are analyzed and standardized.

Acknowledgements This work was supported by the Special Projects in Key Areas (New Generation of Information Technology) of Colleges and Universities in Guangdong Province (2020ZDZX3046), the Characteristics innovation project of colleges and universities of Guangdong Province (Natural Science, No. 2019KTSCX235, 2019), and the Higher Education of the Ministry of Education of the People's Republic of China has the first batch of "industry-academic cooperation, collaborative education" projects in 2019 (No. 201901070016), Characteristic Innovation Projects of Guangdong Province Education Program (2018KTSCX209, 2019GKTSCX092); Science and Technology Program of Guangdong Province (2020B121201013); Science and Technology Special Fund Program of Guangdong Province (2020A0102009); Rural Science and Technology Commissioner Program of Guangdong Province (KTP20200278); Collaborative Innovation Center of Big Data Research and Application, JYU and GMIP (130B0310); Research Achievement Award Cultivation Project, Jiaying University. The Special Projects in Key Areas (New Generation of Information Technology) of Colleges and Universities in Guangdong Province (CN) (Grant No. 2020ZDZX3046).

Compliance with ethical standards

Conflict of interest These no potential competing interests in our paper. And all authors have seen the manuscript and approved to submit to your journal. We confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

References

1. Polan D, Brady S, Kaufman R (2016) Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study. Med Phys 43(6):3330–3331
2. Xu Y, Zhang J, Gong X et al (2016) A method of real-time traffic classification in secure access of the power enterprise based on improved random forest algorithm. Power Syst Protect Control 44(24):82–89
3. Joshuva A, Sugumaran V (2017) Fault diagnosis for wind turbine blade through vibration signals using statistical features and random forest algorithm. Int J Pharm Technol 9(1):28684–28696
4. He S, Chen W, Liu H et al (2019) Gene pathogenicity prediction of Mendelian diseases via the random forest algorithm. Hum Genet 138(6):673–679
5. Wang Y, Xia H, Yuan X et al (2018) Distributed defect recognition on steel surfaces using an improved random forest algorithm with optimal multi-feature-set fusion. Multimed Tools Appl 77(13):16741–16770
6. Mohammady M, Pourghasemi HR, Amiri M (2019) Land subsidence susceptibility assessment using random forest machine learning algorithm. Environ Earth Sci 78(16):1–12
7. Guo J, Wang J, Li Q et al (2019) Construction of prediction model of neural network railway bulk cargo floating price based on random forest regression algorithm. Neural Comput Appl 31:8139–8145
8. Rewade AD, Mohod SW (2018) Content based alternate medicine recommendation by using random forest algorithm a review. Int J Comput Sci Eng 6(10):770–775
9. Wu Q, Wang H, Yan X et al (2019) MapReduce-based adaptive random forest algorithm for multi-label classification. Neural Comput Appl 31:8239–8252
10. Zhang X, Huang W, Lin X et al (2020) Complex image recognition algorithm based on immune random forest model. Soft Comput 24:12643–12657
11. Levantesi S, Nigri A (2020) A random forest algorithm to improve the Lee–Carter mortality forecasting: impact on q-forward. Soft Comput 24:8553–8567
12. Rewade AD, Mohod SW, Bargat SP (2019) Content based alternate medicine recommendation by using random forest algorithm. Int J Comput Sci Eng 7(4):1163–1168
13. Lin P, Yang L (2019) Urban classification based on random forest algorithm. Int J Adv Res 7(11):844–849
14. Zhang Q, Sun X, Feng K et al (2017) Predicting citrullination sites in protein sequences using mRMR method and random forest algorithm. Comb Chem High Throughput Screen 20(2):164–173
15. Shevchik SA, Saeidi F, Meylan B et al (2017) Prediction of failure in lubricated surfaces using acoustic time-frequency features and random forest algorithm. IEEE Trans Industr Inf 13(4):1541–1553
16. Kim A, Myung J, Kim H (2020) Random forest ensemble using a weight-adjusted voting algorithm. J Korean Data Inf Sci Soc 31(2):427–438
17. Chen J, Li K, Tang Z et al (2017) A parallel random forest algorithm for big data in a spark cloud computing environment. IEEE Trans Parallel Distrib Syst 28(4):919–933
18. Li N, Cheng X, Guo H et al (2016) Recognizing human interactions by genetic algorithm-based random forest spatio-temporal correlation. Pattern Anal Appl 19(1):267–282
19. Bharati S, Podder P, Paul PK (2019) Lung cancer recognition and prediction according to random forest ensemble and RUSBoost algorithm using LIDC data. Int J Hybrid Intell Syst 15(2):91–100
20. Wang Y, Li T (2018) Improving semi-supervised co-forest algorithm in evolving data streams. Appl Intell 48:3248–3262

21. Xue L, Wang L (2018) Video tracking algorithm based on particle filter and online random forest. Wirel Pers Commun 102:3725–3735
22. Yao D, Zhan X, Kwoh CK (2019) An improved random forest-based computational model for predicting novel miRNA-disease associations. BMC Bioinform 20:624
23. Kumar S, Sahoo G (2017) A random forest classifier based on genetic algorithm for cardiovascular diseases diagnosis. Int J Eng Trans B 30(11):1723–1729
24. Cao Y, Fan X, Guo Y et al (2020) Multi-objective optimization of injection-molded plastic parts using entropy weight, random forest, and genetic algorithm methods. J Polym Eng 40(4):360–371
25. Peng X, Li J, Wang G et al (2019) Random forest based optimal feature selection for partial discharge pattern recognition in HV cables. IEEE Trans Power Deliv 34(4):1715–1724
26. Song J, Li C, Zheng C et al (2017) MetalExplorer, a bioinformatics tool for the improved prediction of eight types of metal-binding sites using a random forest algorithm with two-step feature selection. Curr Bioinform 12(6):480–489