Transactions on Environment, Energy and Earth Sciences
ISSN: 2960-1924, eISSN: 2960-2300 | Volume 3
GRET 2024

MLP-Based Research in Flood Probability Prediction

Yanzheng Wang*

School of Chemical Engineering, Ocean and Life Sciences, Dalian University of Technology, Dalian, China, 116024

* Corresponding Author Email: wyz_0320@foxmail.com

Abstract. Flood is a kind of natural disaster with great harm. Probability prediction can reduce economic loss effectively. This paper optimizes the prediction of flood probability based on MLP model to reduce the loss. Based on the historical flood data set, a comprehensive correlation analysis using Pearson's correlation coefficient was conducted to find out the influence of each potential factor on the probability of flood occurrence. Based on this, the potential causes of flooding by these factors are examined in detail by taking into account various theoretical and practical aspects, with the aim of revealing the underlying mechanisms and patterns. Two different methods, linear regression and MLP model, are used to establish the corresponding flood probability prediction models. The classical linear regression method is firstly used to predict the probability of flood occurrence, and after rigorous calculation and verification, the obtained prediction accuracy is 99.9250%. In order to further optimize this prediction model to achieve higher accuracy and reliability, MLP is skillfully applied based on five key indicators. Through a series of calculations, the prediction accuracy was successfully increased to 99.9254%. This shows that the improvement is effective and provides a more accurate and efficient method for predicting the probability of flooding. This result not only verifies the rationality and effectiveness of the prediction model, but also provides a solid foundation and strong support for further research and application.

Keywords: Pearson Correlation Coefficient; Linear Regression; MLP.

1. Introduction

Flooding is a natural disaster that poses a serious threat to human society, and accurate prediction of the probability of flood occurrence is of great significance in mitigating its negative impact and formulating effective flood prevention measures[错误!未找到引用源。] . Existing research covers a wide range of studies from traditional statistical models to modern machine learning methods[2]. Extreme value theory has been introduced in statistical modeling, and the Gumbel distribution model has been proposed for predicting extreme flood events. And statistical analysis methods for hydrological data, parameter estimation and model calibration further advance the application of statistical models in flood prediction. In addition, regression models and mixed models have been introduced to explore the linkages among multiple factors. However, existing studies have the limitation of insufficient treatment of nonlinear relationships. Therefore, in this paper, a flood probability prediction model based on MLP is developed to find the pattern of nonlinear relationships and improve the prediction accuracy.

In this paper, we identify the key influencing factors by combining Pearson's coefficient, and then use MLP and linear regression models for modeling respectively, and select the model that is more suitable for analyzing flood probability prediction based on the model prediction accuracy.

In this paper, for the first time, multiple methods such as Pearson's coefficient, MLP (multilayer perceptron) and linear regression are combined to establish a new mathematical model to predict the flood probability, which makes up for the limitations of the traditional methods in flood prediction, and has a better performance for dealing with the data with nonlinear relationship, and improves the accuracy of the prediction.

2. Introduction to related theories

This paper focuses on the application of MLP (Multilayer Perceptron Machine) in flood probability prediction. First, the Pearson correlation coefficient is used to measure the strength of the linear relationship between data features and the probability of flood occurrence. Next, the base model for predicting flood probability is established by linear regression model and MLP respectively, and its prediction performance is evaluated (in this paper, the prediction accuracy of different models is evaluated and compared by mean absolute percentage error (MAPE) to ensure the practical application value of the models). In this paper, these methods are comprehensively applied to optimize the prediction of flood probability and enhance the effectiveness of the early warning system.

2.1. Pearson's correlation coefficient

Pearson's correlation coefficient is a statistical indicator of the strength and direction of the linear relationship between two variables. This paper explores the magnitude of the influence of the factors on the prediction of flood probability, and does not consider the positive or negative influence, so all the data will be substituted into the calculation by taking the absolute value of the posterior in the data processing. Then the value is between 0 and 1, where 1 indicates a completely linear relationship, while 0 indicates no linear relationship[2].

This coefficient is derived by calculating the ratio of the product of the covariance between two variables and their respective standard deviations with the formula:

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y-\overline{Y})^2}} \tag{1}$$

Of these, the X_i and Y_i denote the sample data, and $\overline{X}$ and $\overline{Y}$ denote the sample means.

2.2. Linear regression algorithm

Linear regression modeling is a common method of statistical analysis used to predict a linear relationship between a dependent variable (probability of flooding) and one or more independent variables (influencing factors).

The basic form of the linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \tag{2}$$

Where Y is the dependent variable, and X_n is the independent variable, and β_0 is the intercept term, and β_n is the regression coefficient, and ε is the error term[4].

The least squares method is applied to estimate the regression coefficients.The objective of OLS is to minimize the sum of squares of the errors between the predicted and actual observations, i.e.:

$$Minimize \sum\left(Y_i - \widehat{Y_i}\right)^2 \tag{3}$$

Where $\widehat{Y_i}$ is the flood probability predicted by the regression model.

2.3. MLP

MLP (Multilayer Perceptron) is a feed-forward artificial neural network suitable for dealing with complex nonlinear problems.The basic structure of MLP consists of an input layer, one or more hidden layers and an output layer[5]. The neurons in each layer are nonlinearly transformed by an activation function.The mathematical expression of the MLP model is given below:

$$\hat{Y} = f(W_2 * f(W_1 * X + b_1) + b_2) \quad (4)$$

Where X is the input feature vector, and b_1, b_2 is the bias term.

3. Experiments

In the flood probability prediction model, we analyzed the correlation of 20 potential influencing factors on the probability of flood occurrence based on the Pearson correlation coefficient. These factors were paired with historical flood data and substituted into Equation (1), and the Pearson correlation coefficient was calculated for each factor (the analysis was performed by taking the absolute value of each data in order to focus on the magnitude of the correlation only, without considering the facilitating or inhibiting effect of the factor's influence on the probability of flooding), and the obtained results were plotted as a descending bar chart of the magnitude of the correlation to visualize the correlation of each factor with the probability of flooding. flood probability correlations. These coefficients allow us to identify the factors that are significantly correlated with flood probability and thus filter out the most important ones. This step helps us to more accurately identify the variables that have a significant impact on flood probability and provides key data support for subsequent modeling. Based on the model data prediction, this paper divides the dataset into a training set and a test set. Twenty percent of the data set is used to train the model's training set, and the remaining eighty percent of the data set is used to test the model's generalization ability. And the value of the random seed is set to 0 to ensure that the sequence of random numbers generated is the same every time the code is run, so as to ensure the repeatability of the experiments and facilitate the reproduction of the experimental results. We then defined a list containing all the feature and target variables with the names of the columns used for the analysis. The last column 'Flood Probability' is assumed to be the target column and the rest of the columns are considered as feature columns. Feature column X is extracted from the read data by selecting all the columns in the list except the last element. Also, the column with column name 'Flood Probability' is directly extracted as the target column y. Then a model is initialized with modelLiner and modelMLP, respectively, and the model is fitted on the training set and predicted on the test set, respectively. (where, based on Eq. (4) when making MLP predictions, we set the size of the hidden layer to (100,50), the size of the random seed to 1, and set the number of iterations to a maximum of 500)A linear regression model based on equation (2). The X_n is the corresponding nth influencing factor, which is the input characteristics of the model (including rainfall, soil moisture, terrain height, etc.). The β_0 intercept term, which indicates the base flood probability when all feature values are zero. The β_n is the regression coefficient, which indicates the magnitude of the influence weight of each influencing factor on the flood probability. (If the β_n the value is larger, it means that the influence of the feature on the flood probability is more significant) ε is the error term, which indicates the difference between the predicted and actual values. From this, we can obtain the formula for calculating the flood probability based on the linear regression model:

$$Flood\ probability = \beta_0 + \beta_1 Factor_1 + \ldots + \beta_{20} Factor_{20} + \varepsilon \quad (5)$$

Based on equation (3), the model adjusts its parameters by minimizing the difference between the predicted and actual values. Fitting the model based on the divided dataset continuously optimizes the values of these regression coefficients to find the optimal weight size values (i.e., the β_n value), so that the predicted flood probability is as close as possible to the actual observed value. The magnitude of the regression coefficients is analyzed to understand the strength of the influence of each feature on the flood probability[5]. The MLP model is built based on Eq. (4) and Eq. (6) is obtained:

$$Flood\ probality = f(W_2 * f(W_1 * Factor + b_1) + b_2) \quad (6)$$

Where the input features X (e.g., rainfall, soil moisture, etc.) are direct inputs to the input layer of the MLP model. Two weighting matrices are created W_1 and W_2, the W_1 in the hidden layer affect how the input features are mapped to the activation values of the hidden layer, the W_2 at the output layer determines how the hidden layer activation values affect the final flood probability prediction. The bias vectors b_1 and b_2 respectively, add additional bias to the corresponding layers, allowing the model to fit the data better and also allowing each neuron to produce a non-zero output even in the absence of input features. (i.e., the underlying probability that a flood occurs when it is unaffected by other additional factors) The activation function f performs a nonlinear transformation of the output of each hidden layer allows the model to capture more complex features and relationships[9].

In the MLP regression model, individual flood probability influences enter the model through the input layers and are processed through the weights and biases between the layers. The weight matrices and bias terms are model parameters that together determine the influence of the input features on the final prediction. The activation function, on the other hand, introduces nonlinear properties that allow the model to capture more complex patterns and relationships[7].

Once training is complete, the model can be used to make predictions on new data. In this paper, a linear regression model is trained using the data X_train and the corresponding target value y_train from the training set to learn patterns and relationships in the data. The trained model is used to predict the features X_test of the test set, and the predicted target value y_pred is obtained as the result of the model's prediction of the data in the test set. The performance of the model can be evaluated by comparing the predicted results with the actual values.

In this paper, the accuracy of the model is evaluated by the MAPE value of the calculated results[0] .(MAPE measures the percentage error of the prediction result relative to the actual value, the lower the value the better the prediction performance of the model) The MAPE values of the prediction results of model Liner and model MLP are used to compare the models by model Liner and model MLP respectively.

The prediction accuracy of the model is calculated by the following formula[9]:

$$Accuracy = 100 - mape \tag{7}$$

And in order to be able to more intuitively show the specific performance of the two different prediction models in the process of predicting the probability of flooding, we can draw a line graph comparing the predicted value and the accurate value based on a small amount of sample data respectively. In this paper, we set up a data frame based on 50 initial sample data. The sample data in the data frame are plotted into corresponding line graphs based on the prediction results obtained by the two models respectively for comparison.

4. Results

As shown in Figure 1, we can see the correlation analysis between a series of factors and the flood probability index. These factors are ranked in descending order of magnitude of correlation with flood probability and the top five influencing factors are 'Deterioration of Infrastructure', 'Monsoon Intensity', 'Quality of Dams', 'Topographic Drainage', 'River Management' .

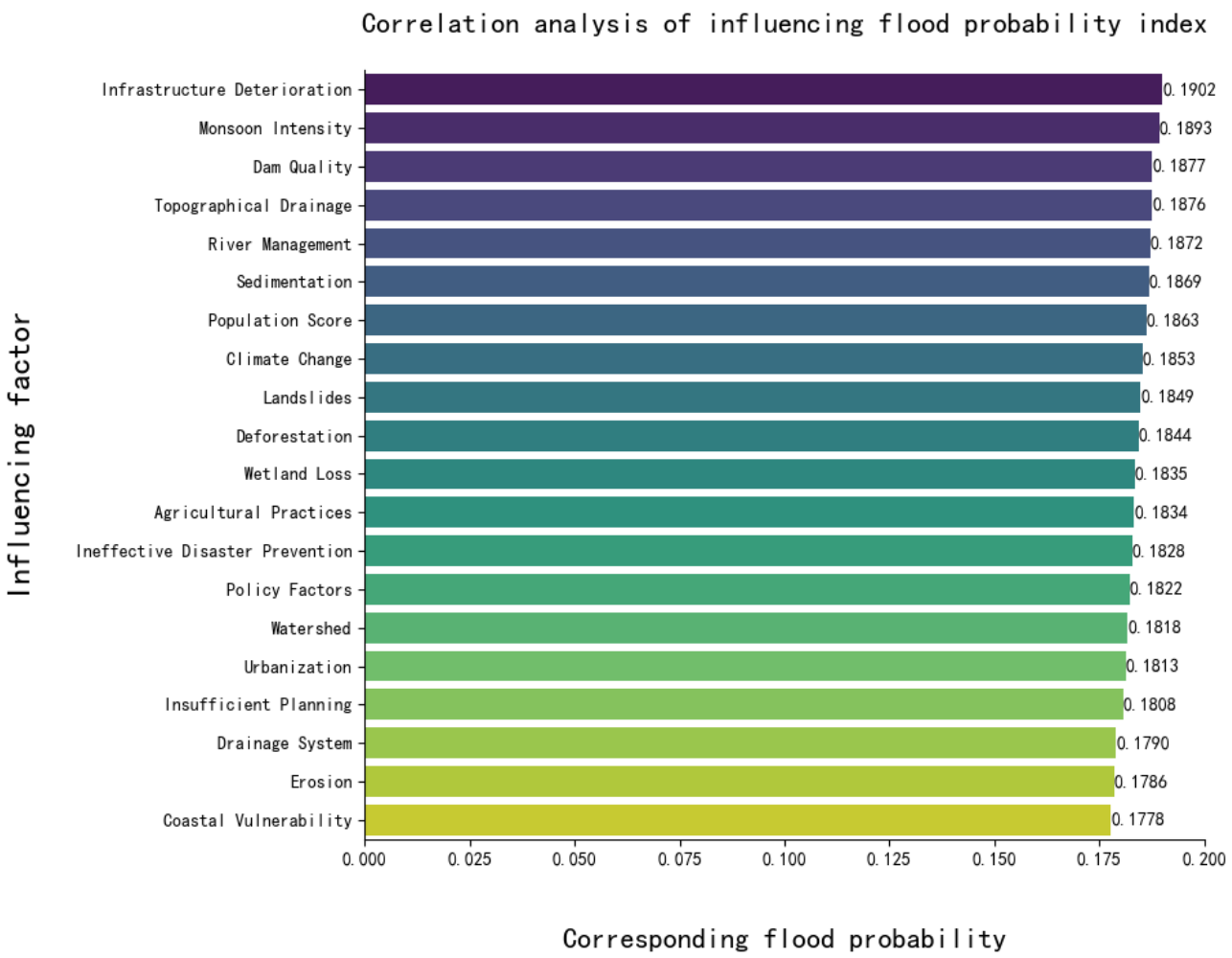

Figure 1. Column analysis of correlation between flood probability and influencing factors

The gradual decrease in correlation values from around 0.19 to around 0.15 indicates that there is a significant difference in the extent to which these factors influence flood probability. Factors with correlations close to 0 or below 0.15, such as coastal vulnerability and erosion, have relatively little impact on flood probability.

The gradual decrease in correlation values from around 0.19 to around 0.15 indicates that there is a significant difference in the extent to which these factors influence flood probability. Factors with correlations close to 0 or below 0.15, such as coastal vulnerability and erosion, have relatively little impact on flood probability.

These data can be used to construct flood probability prediction models, which can be used to better understand and predict flood risk by analyzing the relationship between these factors and flood occurrence. In practical applications, these factors can be used as model input variables to predict the probability of flooding in a given area or condition.

It is also possible to derive some of the potential problems that need to be prioritized for better flood prevention. Such as based on the five most influential factors 'Infrastructure Deterioration', 'Monsoon Intensity', 'Dam Quality', 'Topographic Drainage', 'River Management' can be addressed by enhancing the maintenance and upgrading of infrastructure (especially drainage systems and flood defenses to ensure that they function properly in the event of flooding), regularly checking the structural integrity of the dams and embankments to ensure that they are able to withstand the impact of extreme weather events. Rationalize land use planning through measures such as river dredging, riverbank stabilization

and watershed vegetation restoration, and avoid high-risk development activities, such as the construction of residential or industrial facilities, in flood-prone areas. Establish and improve disaster early warning systems to improve the ability to predict and respond to natural disasters such as floods. Reduce the risk of flooding. Consider the impact of climate change on the probability of flooding and develop adaptation strategies, such as enhancing the flood resilience of infrastructure and improving water resource management. Develop and enforce stringent policies and regulations to reduce flood risk, such as building standards, environmental protection regulations and land use planning.

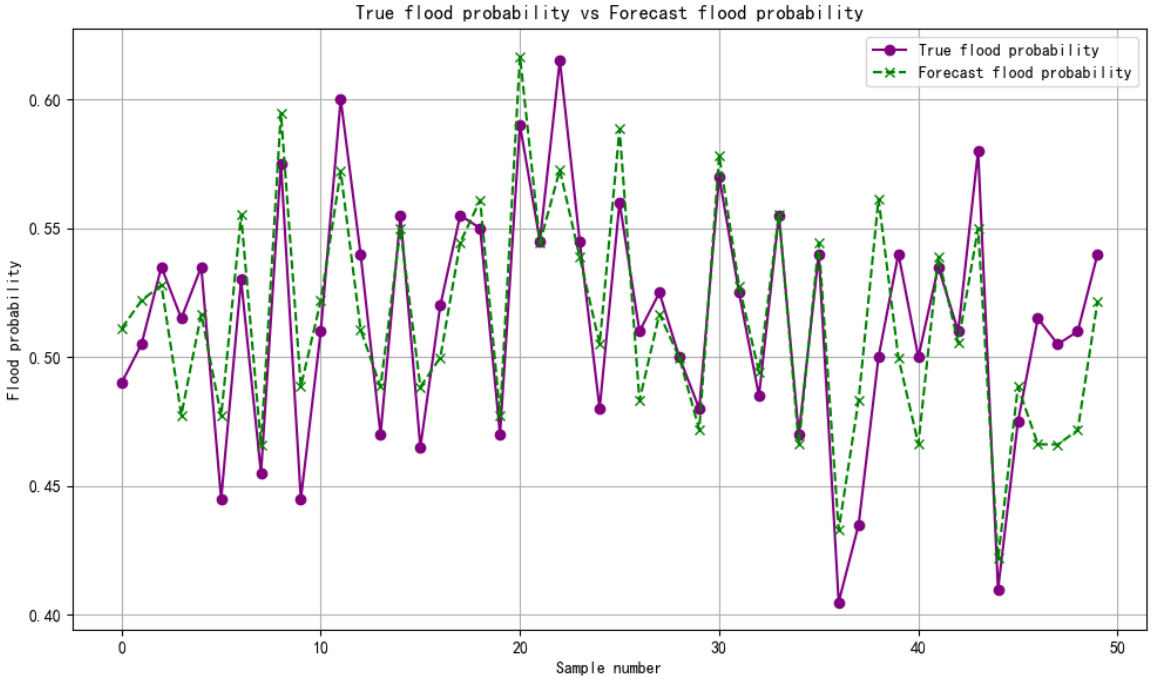

Figure 2. Predicted results with linear regression model with 20 factors

Figure 2 shows the comparison between the prediction results of the prediction model based on linear regression and the actual results, and the prediction accuracy of the linear regression model using 20 key indicators is 99.9345%.

Figure 3 shows the comparison between the prediction results of the MLP-based model and the actual results, and the prediction accuracy of the MLP model using 20 key indicators is 99.9688%.

Based on the analysis of the accuracy outcome metrics, the MLP model is more accurate and better adapted to complex real-world scenarios of flood occurrence.

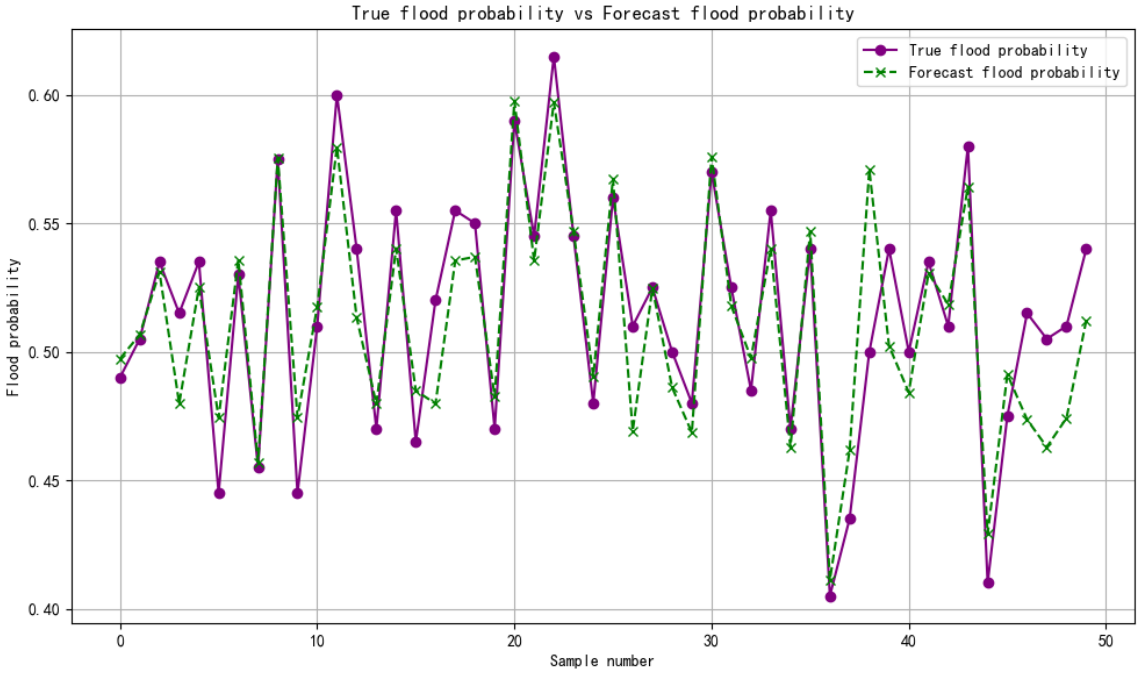

Figure 3. Predicted results with the MLP model with 20 factors

5. Conclusions

Based on the flood history dataset, the correlation between these features and flood occurrence was analyzed by calculating the Pearson correlation coefficient. On this basis, features with high correlation were selected, redundant information was removed, and model input data were optimized. Then, different modeling perspectives for flood prediction were provided by comparing the linear regression model and the MLP model. Linear regression is suitable for situations where the relationship between features and target variables is simple, while MLP is able to capture more complex nonlinear relationships. By comparing the performance of linear regression and MLP models, it was determined that the MLP model is more advantageous for probabilistic flood prediction. The improved flood prediction model can more accurately screen features related to flood occurrence and reduce the interference of irrelevant features. This feature selection method helps to improve the prediction accuracy of the model, which is of great significance to the accuracy of the flood warning system[11]. This paper not only enriches the construction method of the flood prediction model theoretically, but also promotes the progress of flood prevention technology in practical application, provides more accurate and effective tools for the industry, and has important innovative significance.

References

[1] LUO Dan,CHEN Xiaohong,ZHANG Yongzheng,et al. Impact assessment of typhoon on nearshore compound flooding [J]. Hydrology,2024,44 (02):8-18.

[2] Osman A S ,Das J .A robust ensemble of hybrid and bivariate statistical models for flood prediction mapping in Lower Damodar River Basin of India [J].Geosystems and Geoenvironment,2024,3 (4):100312-100312.

[3] Hot Spring,Yu Yuhuan,Zhuang Shangde,et al. Waterway freight volume forecasting based on SHAP and multi-strategy optimization TSO-XGBoost model [J/OL]. Journal of Water Resources and Water Transportation Engineering,1-13 [2024-09-02].

[4] Wang Bo. Analysis of the impact of RMB exchange rate fluctuation on international portfolio investment [J]. China Business Journal,2024,33 (16):118-121.

[5] WANG Zehua,LIU Xiaoming. Research on the test of abrasive belt grinding process parameters and the prediction of test results[J]. Machine Tools and Hydraulics,2024,52 (16):32-39.

[6] ZHAO Peng,WEN Gang,HE Zhanchang,et al. Evaluation of shallow landslide susceptibility in Jinsha River basin based on machine learning [J/OL]. Water Conservancy and Hydropower Technology(in English and Chinese),1-23[2024-09-02].

[7] ZHANG Dageng,WANG Xi-han,GAO Quan-fu. A digital cultural resources recommendation method integrating knowledge graph and interest preference [J/OL]. Computer Technology and Development,1-9 [2024-09-02].

[8] Hui Li,Zixian Cui,I hope you can help me. A study on paper recommendation based on academic knowledge graph with biased random wandering [J/OL]. Data Analysis and KnowledgeDiscovery,1-19 [2024-09-02].

[9] HUA Zhiheng,ZHANG Jinpeng,YIN Bo,et al. An integrated prediction method for sea clutter amplitude distribution in complex spatio-temporal scenarios [J/OL]. Journal of Radio Wave Science,1-8 [2024-09-02].

[10] HU Yehui,WANG Yuhan,JI Yulei. Calculation method of temperature-dependent friction coefficient in simulation modeling of biaxial shoulder stir friction welding [J]. Mechanical Design and Research,2024,40 (04):192-197+206.

[11] Baggio T ,Martini M ,Bettella F , et al.Debris flow and debris flood hazard assessment in mountain catchments [J].Catena,2024,245108338-108338.