[1]Fei Han

[1]Jingkun Bao

[2]Kun Wang

[2]Jiale An

[1]Zhongcai Gao

[1]Yurong Li

[1]Yongjun Li

Research on Numerical Simulation and Prevention Strategy of Geological Hazards by Integrating Machine Learning and GIS Technology

Abstract: - In order to prevent the occurrence of geologic disasters in a timely manner, and reduce the impact and threat of disasters on society. In this paper, firstly, GIS technology is used to collect data related to geological disasters and analyze the factors triggering geological disasters, so as to facilitate the subsequent establishment of geological disaster prediction models. Secondly, in order to make the collected data more accurate, an interactive iterative cleaning method is used to clean the data to ensure that the data will not affect the establishment of the model. Finally, machine learning is used to establish the geohazard prediction model to complete the prediction of geohazards. In the simulation test, the integration of machine learning and GIS technology disaster loss rate is lower, at about 15%. The predicted value is 0.867, and the discrete index score is around 0.426, which is more discrete. Therefore, combining machine learning and GIS technology can establish a more accurate prediction model of geologic disasters in order to reduce the harm of geologic disasters to human beings.

Keywords: GIS technology; geohazard; interactive iteration; machine learning; numerical simulation

1. Introduction

Geological disaster events have social focus and time urgency, disposal decision-making cannot be separated from scientific and efficient technical support [1]. Numerical simulation technology has the advantages of low-cost, high-efficiency, and multi-case simulation, which has been widely used in the fields of public safety, meteorology, water resources and environment, and its decision-making assistance function is outstanding [2]. However, numerical simulation technology in the field of geohazard emergency response field application is not common, has not yet formed a paradigm of the programmed method, the disaster emergency response to support the effect is very limited [3]. In emergency situations, the application of numerical simulation is limited by technical conditions [4]. In the practice of numerical simulation theory research and application, there has been less consideration of time efficiency for a long time, resulting in a long time-consuming routine simulation, which is a gap with the time-sensitive characteristics of emergency response and disaster relief. Emergency response to sudden-onset geologic disasters is itself characterized by technical integration, method coordination

[1]Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650031, Yunnan, China. *Email：JingkunBao@hotmail.com

[2]Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology, Kunming 650500, Yunnan, China

and the pursuit of effectiveness, and the specificity of the disaster body itself determines that it seems difficult to find universally applicable numerical methods and models. Therefore, it is very important to address the research on numerical modeling and prevention strategies for address disasters [5].

As in previous studies, numerical analysts often have strong subjectivity in software simulation, which reduces the credibility of the results. In this paper, we first analyze the causes of geohazards and use GIS technology to collect geohazard data to provide a basis for the subsequent generation of geohazard models. Secondly, the interactive and iterative data cleaning method is used to clean the collected data to ensure the accuracy of the data. Finally, machine learning is used to model the geohazard prediction model. By analyzing the three geohazard prediction models through multiple indicators, it can be concluded that the geohazard prediction model generated by machine learning can meet the prediction needs of natural disasters. If the prediction accuracy is high, it means that the numerical simulation model can reduce the loss rate of disasters, thus ensuring the smooth implementation of disaster prevention and control work.

2. Related Words

In order to improve the engineering safety level, Li, L et al. for the frequent occurrence of geohazards and the accompanying high risk, in order to explore the influence mechanism of the vulnerability of geohazards, the three characteristics of storm, sensitivity and adaptability are used as the indicators, and the weights of the indicators are obtained by using the sequential diagram method [6]. Li, Z et al. for the southeastern region of the Tibetan Plateau to carry out investigation research, to obtain the valleys and geomorphology of the satellite imagery, and the results of numerical simulation show that that iceberg landslides are important triggers of geologic disasters. Therefore, it is necessary to monitor glacier activities in real time to avoid property losses and casualties [7]. Wang, H et al.'s study considered the sustainable development of the region and analyzed the susceptibility of geologic hazards such as avalanches, landslides, and mudslides in the study area. Ten environmental geohazard factors were extracted through investigation and comprehensive analysis, and combined with GIS technology, an analysis database was constructed to predict the vulnerability areas using the informativeness model and generalized regression neural network [8]. Qian, X established a disaster management system for geohazard emergency response based on association rule data mining, which is supported by a big data platform to visualize the disaster data and warn of geohazards. In the simulation test, it is proved that the built geohazard early warning system can improve the processing speed of the relevant departments by 59.4% [9]. Liu, J et al. pointed out that three-dimensional visual numerical simulation method can be widely used in disaster analysis, through simulating dynamic changes, analyzing static spatial state, and deriving the process of disaster evolution, so as to avoid and prevent the occurrence of disasters [10]. Niu, H et al. proposed to use the GIS technology combined with the information volume model, so as to prevent disasters. GIS technology combined with the informativeness model to obtain basic information such as slope distribution, rock group type, fracture distribution map and other basic information in the study area through hierarchical management of geo-environmental data, in order to effectively predict and prevent the occurrence of landslide disasters [11]. Tan, Q et al. calculated the disaster evaluation and risk level on the basis of theories related to geohazard zoning, used hierarchies to derive the evaluation factor weights and information, combined with a neural network model simulation to study the regional geologic data and proved the effectiveness of the

proposed numerical simulation method [12].

Numerical simulation and prevention of geological disasters is a key part of the effectiveness of disaster prevention and mitigation and reduction of losses, while the actual geological information contains a large amount of data and information, and the processing method at this stage is also accidental and random, and most of them are originated from the traditional technical means of group measurement and group prevention. Numerical simulation integrating machine learning and GIS technology can realize rapid scenario deduction and hazard scope delimitation, which will have intuitive effectiveness and meet the needs of scientific disaster prevention and mitigation.

3. Application of GIS technology

GIS is the abbreviation of geographic information system, which is a kind of spatial information technology system using computer hardware and software technology as well as related science and technology to collect, store, analyze, calculate, and utilize the earth's surface and spatial information [13]. GIS technology is based on computer technology, and it mainly realizes various functions through computer technology. The current GIS technology can collect and organize a large amount of information data, which is of great significance for the prediction, forecasting and post-disaster reconstruction of geological disasters.

3.1 Analysis of geologic hazard factors

3.1.1 Data sources

GIS technology mainly collects and manages spatial data, has strong spatial analysis and mapping ability, and is widely used in the prevention of geologic disasters [14-15]. ArcGIS software is used to collect data related to geohazards and obtain the control factors and triggering factors of geohazards, which is convenient for the training and evaluation of geohazard models, and Table 1 shows the sources of geohazard data. The controlling factors of geohazards generally refer to the internal causes of disasters, and control the disaster-conceiving environment in which geohazards occur, usually with small changes. Common control factors include topography, slope morphology, stratigraphic lithology, meteorological vegetation, and water system distance.

Table 1 Sources of geological disaster data

Type	Data	Resolution/m	Data source
Topography	DEM	30	Digital elevation DEM
Weather and vegetation	Extreme rainfall	60	Geographical survey data
Soil properties	Internal friction angle, cohesion	60	Borehole soil sample analysis

The use of elevation data in ArcGIS software allows for the calculation of slope gradient, slope direction, and terrain relief. Slope gradient is a key factor for slope stabilization. If the slope is too low, it will not be able to provide sufficient downward momentum, and if it is too steep, it will not be able to complete the accumulation

of sediment on the slope, and only when an equilibrium point is reached can the slope provide a supporting force. The geological surface analysis tool in ArcGIS software can be used to calculate the slope gradient and arrive at a suitable equilibrium point. Slope orientation is closely related to rainfall and solar radiation, and has a profound effect on the surface cover and topographic moisture index. Slope orientation maps were generated using ArcGIS tools and divided into nine directions in order, flat, north, northeast, east, southeast, south, southwest, west and northwest. The slope shape of a slope generally refers to the profile curvature, which is mainly used to describe the complexity of the slope topography, and is the size of the slope along the direction of the maximum slope drop. The higher the value of slope form, the higher the degree of curvature change in the vertical direction of the slope, and the more complex the terrain will be. The classification of slope morphology can be done by using elevation data and ArcGIS software can calculate the profile curvature and planar curvature, generally the value of curvature is greater than 0 for convex slopes, equal to 0 for linear slopes, and less than 0 for concave slopes [16].

3.1.2 Causes of disasters

Stratigraphic lithology is also a controlling factor for geohazards, and its data can be obtained by vectorization with ArcGIS software. Geotectonic factors are usually related to the distance of fractures and hazard sites, and the fracture data come from vectorized geological maps. According to the geological map, the fracture distance can be divided into eight intervals: less than 1km, 1 to 2km, 2 to 3km, 3 to 4km, 4 to 5km, 5 to 6km, 6 to 7km and more than 7km. The corresponding fracture distances of geohazard points can be obtained by using the buffer in ArcGIS software. According to the high-level data and ArcGIS software, we can get the water system distance data, and the water system distance can be divided into four intervals of less than 0.6km, 0.6 to 1.2km, 1.2 to 1.8km, and more than 1.8km, and we can get the water system distance corresponding to the disaster point by using the buffer zone in ArcGIS software.

The triggering factors of geohazards are the external causes of geohazards, which are usually in a constant state of change, and when the changes accumulate to a certain extent, they will lead to slope instability and collapse. There are many triggering factors for geohazards, which are mainly categorized into factors such as vegetation cover, precipitation, and land use rate. Vegetation cover is a key factor affecting geohazards, which is usually calculated using Sentinel 2A remote sensing images based on a spatial resolution of 10m. The corresponding Normalized Vegetation Index (NVI) was calculated using ENVI software, and the degree of vegetation cover was calculated using NVI. Higher vegetation cover indicates that the vegetation in the area is well developed and has a strong ability to maintain soil and water, while lower vegetation cover in the area indicates that the vegetation in the area is generally well developed and is difficult to maintain soil and water environments. According to the degree of vegetation in different areas, they are classified as building and other, road, grassland, woodland, artificial excavation land and cultivated land.

Surface moisture is a reflection of the distribution of soil moisture content and groundwater in the vegetation cover, and in general, geologic disasters often occur in areas with high soil moisture. Surface moisture values are obtained by using the raster calculation of the water fraction using the fill and flow tools under the hydrological tools. Extreme rainfall conditions are more prone to geohazards and are an important cause of

induced disasters. Extreme rainfall can cause rapid deformation of slopes in a short period of time, resulting in slope instability and triggering geohazards. Generally speaking, precipitation data comes from global precipitation measurement data, and its spatial resolution is generally 0.1°. ArcGIC software can be used to get the precipitation amount in recent years, so as to prevent the emergence of extreme weather, which can cause geologic disasters and affect the stability of the society.

Based on the above analysis, the corresponding data of geohazards and the discussion of geohazard factors can be collected, which provides data support for the establishment of geohazard prediction model.

3.2 GIS data processing

In order to make the data related to geological hazards collected by GIS technology more accurate and clear, the collected data needs to be processed. Utilize an interactive iterative cleaning method to clean your data to make it more accurate. The collected dataset is represented by D, the corresponding relational model is G, and A is used to represent the set of attributes of G. A contains set $F=\{E_1,\cdots,E_K\}\in A$, which represents the dirty property set, and F corresponds to discrete data, which can be easily modified. $R=A-F=\{W_1,\cdots,W_L\}$ represents the complement of A, which belongs to the clean attribute set, R is composed of clean data, and attribute E_i in set A corresponds to a value range of $dom(E_i)$. Tuple t in dataset D can be divided into two parts, which can be modified for part $t[F]=t[E_1,\cdots,E_K]$ and clean part $t[R]=t[W_1,\cdots,W_L]$, abbreviating $t[R]$ and $t[F]$ to r and f, that is, $t=\langle r,f\rangle$. Dataset D can be divided into two parts based on whether the data has been inspected and cleaned: data that has not been checked or is uncertain $D_e=D-D_c$, and data that has been checked as correct or has been cleaned $D_c\subset D$.

Unlike other cleaning methods, the interactive iterative cleaning method is to start the cleaning of data whose cleanliness is unknown, i.e., $D=D_e$. A cleaning model is built for a small portion of the data D_c obtained by manual cleaning from D_e, and for the tuple of tuples $<r,f>\in D_e$ that may be incorrect, r is brought into the cleaning model to produce the correct data, as follows:

(1) Since an iterative data cleaning method is used, the initial dataset D needs to be divided into multiple chunks. Let the cleanliness of initial dataset D is unknown, both are dirty data, the whole dataset needs to be cleaned and no clean data is referenced, so dataset D will be divided into two parts, one part is dirty data chunks $\{D_e^1,\cdots,D_e^N\}$, which is to be cleaned, and the other part is the initial dirty data training set D_e^t.

(2) Data cleaning needs clean data as a support, so it is necessary to get clean training set D_c^t from the initial

dirty data set D_e by pre-cleaning.Since the cleaning process has manual participation, the divided initial dirty data training set D_e^t is given to the manual unfolding of the cleaning, which results in the clean training set D_c^t. Because the interactive iterative cleaning method contains active learning techniques, the initial D_e^t is smaller, and manual cleaning of the workload is lower, but it also leads to a weaker ability to clean the model initially. In the iterative cleaning with active learning, the manually cleaned data is continuously added to D_c^t and the proportion P_{right}^i of correct values of dirty attributes in the data set is checked in the cleaning set D_e^t using manual inspection, taking P_{right}^i as an intrinsic feature of the data set D expressed in the probability of correctness of each dirty attribute, which is used for determining the degree of certainty of the original data.

(3) Certainty model M can be divided into two parts, one part is a probability classifier, which learns the probability distribution from D_c^t and calculates the probability of occurrence of each outcome. The other part is the certainty screening, which calculates the certainty of the results of the classifier using BvSB as a criterion, while using P_{right}^i to unfold the certainty of the original data, and selecting the certain data as the output of M. Due to the comparative fixing of the certainty, the modification of the data is made rigorous to reduce the operation of error modification.

(4) Select a data block $D_e^i(i=1,2,\cdots,N)$ as the input of M. For each tuple $t=\langle r,f\rangle$ in D_e^i, a suggested modification tuple $t'=\langle r,f'\rangle$ can be obtained in a probabilistic classifier, and the deterministic value C is obtained by unfolding the deterministic value of t' using the deterministic formula, while the original tuple t calculates the original deterministic value C_{init}. When condition $t\neq t'$, the value with the largest certainty is selected as the output of M and inserted into the recommended update dataset $\dot{D}_e^i$. When condition $t=t'$, insert t directly into the dataset to be cleaned $\hat{D}_e$. Since the recommendation ability of the model is not strong in the previous iteration, there is a lot of probability that the predicted value is wrong even if the predicted value is the same as the original value, so it is necessary to summarize this part of the data and start the secondary cleaning in the last round of cleaning using the model with better recommendation ability.

(5) Using the certainty gain formula, the data t' in $\dot{D}_e^i$ is calculated to obtain the certainty gain value C_{gain}. $\dot{D}_e^i$ is sorted in ascending order using the certainty gain and is given to the manual inspection in order. The

manual checking of the remaining data stops when the manual checking of $\dot{D}_e^i$ is considered to be as clean as expected. The Certainty Gain is the difference between the Certainty of the predicted value and the Certainty of the original value, and represents the degree of divergence between the two pre-selected results selected. Therefore, prioritizing and recommending data with small gain values for manual inspection can assist in mediating the modification of divergence and make an assessment of the model's cleaning ability to verify the effectiveness of cleaning and ensure the reliability of data restoration.

(6) The manually checked and cleaned data will be fed back to the clean training set D_c^t, which is a part of the data with large divergence, and added to D_c^t to refine the probability distribution of the data, and compared with simply expanding the training set, the data feedback method based on the certainty gain can improve the cleaning ability of M faster, and the data that are not manually checked or cannot give the correct value after checking will be output to $\hat{D}_e$, waiting for the last round of the secondary cleaning. At this point, one round of cleaning is finished, and when the human is dissatisfied with the cleaning ability of M, it will continue to clean the next dirty data block.

(7) During the process of unfolding the inspection of the data blocks, when the human considers that the cleanliness of the multiple data blocks modified by M has reached the expected level, i.e., M has met the cleaning requirements, the iteration is stopped and the remaining unchecked data blocks are summarized in $\hat{D}_e$. At this point, the final round of cleaning is prepared, including the cleaned D_c^t, the cleaned model M that has been modeled to the satisfaction of the human, and the dirty data $\hat{D}_e$ that has not been checked by the human or is not correctable after the manual check.

(8) Input $\hat{D}_e$ into the cleaning model M, and execute the whole cleaning process again until the manpower expresses satisfaction with the cleanliness of the last batch of proposed modified data, and then summarize the manually inspected and cleaned data, the manually uninspected and uncertain data D_c^t to obtain the final cleaning results and complete the cleaning of the data.

According to the above process, the collected geohazard data will be cleaned to make the data clearer and more accurate, and provide the basis for the subsequent establishment of geohazard model.

4. Numerical modeling of geologic hazards

Based on the geohazard data collected by GIS technology, a geohazard prediction model is built using machine learning to improve the ability of geohazard prediction [17-18].

The XGBoost model is built using machine learning, the weak evaluator with preferences is used as the base learner, and it is combined to start the training learning, so as to get an integrated strong evaluator. Figure 1 shows the flow of the XGBoost model, which is an efficient implementation of GBDT. Unlike the traditional GBDT model, the XGBoost model uses Taylor's second-order expansion to optimize the loss function, and adds a regularization term to control the complexity of the model. Compared with other machine learning algorithms, it greatly improves its computational effectiveness and generalization ability.

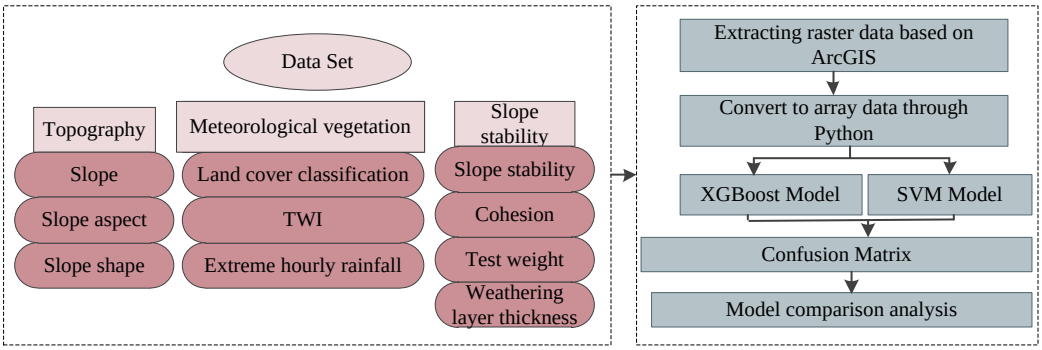


Figure 1 XGBoost model process

The model of XGBoost is represented in the following equation:

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{1}$$

where $\hat{y}$ represents the predicted value, K represents the total number of trees, i represents the i th sample, x_i represents the feature vector corresponding to sample i, F represents the set of all trees, and f_k represents the k th tree produced by the k th iteration as a function in set F.

The objective function of XGBoost is as follows:

$$L^t(y, \hat{y}^t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \Omega(f_t) = \sum_{i=1}^{n} l\left[y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right] + \Omega(f_t) \tag{2}$$

In the formula, the optimization process uses the additive model with forward distribution algorithm, that is, the model prediction value $\hat{y}_i^t$ of the i st sample in the t nd round is based on the model prediction value $\hat{y}_i^{(t-1)}$ in the $t-1$ th round, adding a new tree $f_t(x_i)$, $\sum_{i=1}^{n} l(y_i, \hat{y}_i^t)$ is the prediction error in the t th round, $\Omega(f_t)$ is the regularization term in the t th round, which denotes the complexity of the model in the t th round. Define T as $f_t(x_i)$ the total number of leaf nodes of this tree, and w as the score of each leaf

node, as in the following equation:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

Expanding Taylor's second-order optimization for the objective function and setting $g_i = \partial_{y^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$, $h_i = \partial^2_{y^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$, we have the following equation:

$$\tilde{L}^t \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{4}$$

Simplification of the above equation gives:

$$\tilde{L}^t = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T \tag{5}$$

where $I_j = \{i \mid q(x_i) = j\}$, represents the set of j rd leaf nodes in the tree of $f_t(x_i)$, and γ and λ represent the pre-set hyperparameters.

Let the structure of the tree $q(x)$ be a known term, and let $\sum_{i\in I_j} g_i = G_i$, $\sum_{i\in I_j} h_i = H_i$, minimize the loss function to derive the optimal parameter w_j^* and the loss function $\tilde{L}^t(q)$, as follows in Eq:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{6}$$

$$\tilde{L}^t(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^{\,2}}{H_j + \lambda} + \gamma T \tag{7}$$

where $\tilde{L}^t(q)$ can be used as a function to score the expansion of tree structure $q(x)$, the smaller its value, the more accurate the corresponding model prediction result. The algorithm continuously selects different tree structures through the scoring function to find the tree with the optimal structure. However, it is not realistic to enumerate all possible tree structures, so the XGBoost model uses a greedy algorithm to add splits to the existing leaf nodes. Let I_L and I_R be the nodes after splitting the left and right subtree scores, $I = I_L \cup I_R$, at which time the loss function is:

$$I_s = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \quad (8)$$

The loss function is used to judge whether the tree needs to be partitioned or not.

The above geohazard model is used to learn and train the collected geohazard data to obtain the final geohazard numerical simulation results.

5. Numerical simulation of geologic hazards and analysis of prevention strategies

5.1 Disaster loss ratio test

Disaster loss rate test is to evaluate the proportion of financial loss or resource value caused by disasters, reflecting the degree of damage of disasters to human society. Usually, after using the geohazard prediction model, the lower the disaster damage rate the higher the font integrity, which proves that the geohazard model's effect and performance is better, and it can remind people to pay attention to the geohazard in time and reduce the impact of the geohazard on the human society, and the higher the disaster damage rate,, which proves that the geohazard model's effect and performance is worse, and it can't remind people to pay attention to the geohazard in time and reduce the people's property loss, so it is necessary to start the disaster damage rate test for XGBoost model, LightGBM model and Catboost model. The results of the disaster loss rate test are shown in Table 2, and the XGBoost model using machine learning has a lower disaster loss rate. The disaster loss rate of 66% before use is reduced to 15%, which proves that the model can remind people of geologic disasters in time and reduce the damage of geologic disasters to the social economy and human property, and the model's performance and effect are better, while the disaster loss rate of LightGBM model and Catboost model is higher, around 30% and 35%, although it is reduced compared to the disaster loss rate before use, but the reduced value is less. But the reduced value is less, which proves that the model is difficult to remind human beings of geologic disasters in time, and the performance of the model is poor.

Table 2 Disaster loss rate test

Years	Before using geological disaster prediction model/%	XGBoost model loss rate/%	LightGBM model loss rate/%	Catboost model loss rate/%
2020	65%	15%	30%	35%
2021	67%	14%	32%	35%
2022	65%	16%	29%	38%
2023	67%	15%	30%	33%
2024	66%	15%	30%	35%

5.2 Geologic Hazard Prediction Testing

The geohazard prediction test is a measure of the accuracy of the geohazard prediction model in predicting geohazards, which directly reflects the degree of consistency between the warning signal and the actual disaster occurrence. The higher the value of geohazard prediction proves that the prediction effect of the geohazard model is better, and the prediction result is more accurate and readable, which is highly practical. The lower the predicted value of geohazard proves that the prediction effect of the geohazard model is poorer and the prediction result is inaccurate. Therefore, XGBoost model, LightGBM model and Catboost model are tested for geohazard prediction, and the geohazard prediction results are shown in Figure 2. It can be seen that the geological prediction of XGBoost model based on machine learning is more accurate, and its prediction value is around 0.867, due to the interactive iterative cleaning method used to process the data to ensure the accuracy of the data before the establishment of the XGBoost model, so that the XGBoost model in the prediction of geologic hazards by the higher accuracy, better prediction ability and performance, while the LightGBM model and Catboost model have lower values of geohazard prediction, around 0.797 and 0.847, due to the lack of data processing in advance, which leads to inaccurate prediction of geohazards in the generated model, and it is difficult to have better prediction ability and performance, which does not meet the needs of today's era.

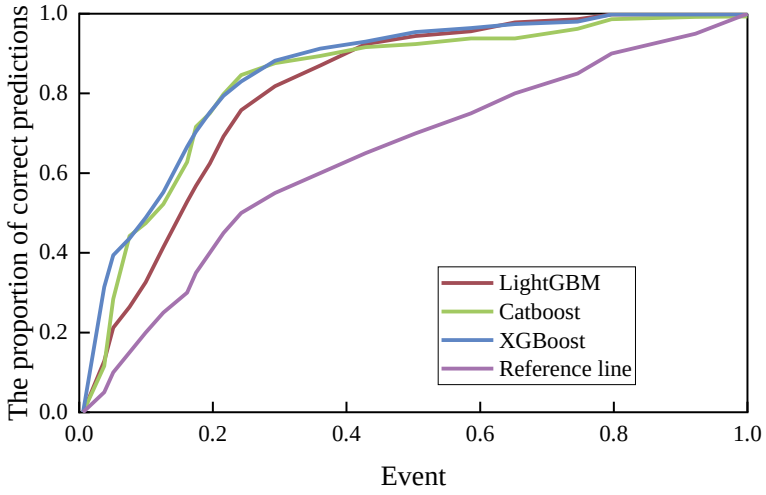
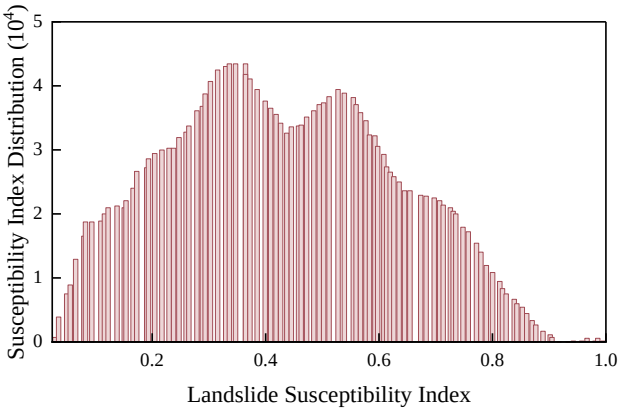


Figure 2 Geological disaster prediction results

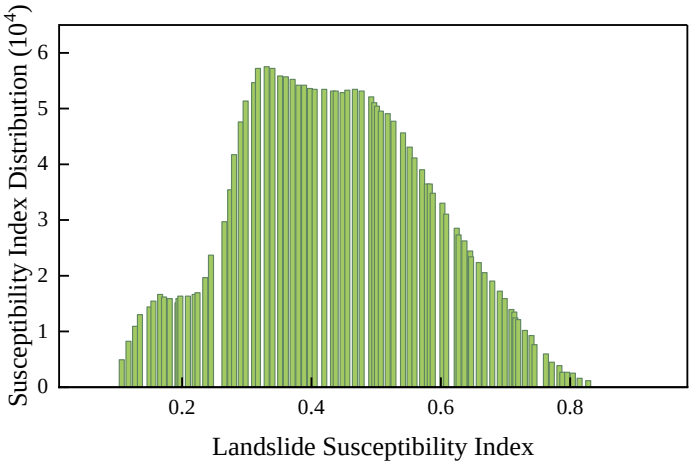
5.3 Geohazard susceptibility index test

The test of geologic disaster susceptibility index mainly evaluates the susceptibility of geologic disasters in different regions by analyzing the factors of geology, environment and climate in different regions, and provides a corresponding basis for geologic disaster early warning and prevention. In general, the higher the discrete degree of the geohazard susceptibility index, the better the model proves the differentiation of geohazard susceptibility, which can reflect the index of geohazard susceptibility in different regions, and the results of the susceptibility index test are presented in Fig. 3.The XGBoost model, as shown in Fig. 3(a), has an index distribution of 4×10^4, and an index score of about 0.426. It shows that the XGBoost model has a higher degree of discretization, and has a better differentiation of the susceptibility of geohazards, which can clearly reflect the susceptibility index of geohazards in different areas with high accuracy.The Catboost model is shown in Fig. 3(b), with an index distribution of 0.425, which has a lower degree of discretization, and it is difficult to have a better differentiation of the susceptibility of geohazards in different areas. The LightGBM model is shown in Fig.

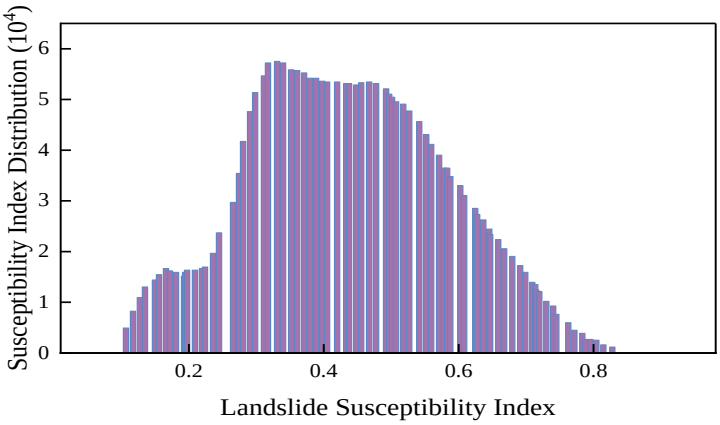
3(c), with the index concentrated around 0.440, which cannot play a preventive role for geohazards and does not meet the demand of numerical simulation of geohazards.



(a) XGBoost Model



(b) Catboost Model



(c) LightGBM Model

Figure 3 Geological disaster susceptibility test

6. Conclusion

This paper collects geohazard data using GIS technology and analyzes the causes of geohazards to pave the way for the subsequent establishment of geohazard prediction models. Then, the interactive iterative data cleaning method is used to clean the collected geohazard data in order to avoid the phenomenon of inaccurate data. Finally, the machine learning method is used to establish the geohazard error model to complete the prediction of geohazards. In order to analyze the ability of the geohazard prediction model, the analysis of the geohazard prediction model is completed from the three indexes of geohazard susceptibility, prediction correctness and coverage, which proves that the geohazard model established by machine learning is more effective, with higher prediction accuracy, and it can remind people to prevent geohazards in time, whereas the geohazard prediction ability of the other two models is worse, and it is difficult to remind people of the occurrence of geohazards in time, and the machine learning model reduces the risk of geohazards. occurrence, and the machine learning model reduces the disaster loss rate, reducing the loss to about 15%, and the other two models are around 30% and 35%, which does not reduce the loss rate of disasters.

To sum up, the geologic disaster model established by machine learning can accurately predict geologic disasters, remind people to prevent geologic disasters in time, reduce the loss rate of disasters, so that the society can be sustained and stable development, and provide a new perspective and field for the prediction of geologic disasters, which is of high practical significance.

References

[1] Zheng, Z., Xie, C., He, Y., Zhu, M., Huang, W., & Shao, T. (2022). Monitoring potential geological hazards with different InSAR algorithms: The case of western Sichuan. *Remote Sensing*, *14*(9), 2049.

[2] Lei, T., Lu, Y., Zhang, C., Wang, J., & Zhou, Q. (2021). Research on the Application of GIS Technology Combined with RBFNN-GA Algorithm in the Delineation of Geological Hazard Prone Areas. *Computational Intelligence and Neuroscience*, *2021*(1), 2677453.

[3] Tian, C., Tian, H., Li, C., & Chen, F. (2022). Stability Evaluation of Massive Landslides Using Ensembled Analysis of Time-Series InSAR and Numerical Simulation along the Yellow River, Northwestern of China. *Geofluids*, *2022*(1), 6546372.

[4] Ren, B., Yuan, L., Mu, W., Zhang, Y., Yu, G., Cao, C., ... & Li, L. (2022). Investigation of a Method to Prevent Rock Failure and Disaster Due to a Collapse Column Below the Mine. *Mine Water and the Environment*, *41*(4), 979-995.

[5] Xu, Y., Qiu, X., Yang, X., Lu, X., & Chen, G. (2020). Disaster risk management models for rural relocation communities of mountainous southwestern China under the stress of geological disasters. *International Journal of Disaster Risk Reduction*, *50*, 101697.

[6] Li, L., Zhi, M., Li, R., Wang, S., & Cao, L. (2022). Simulation of vulnerability to geological disaster in coal mine based on system dynamics. *Geofluids*, *2022*(1), 2248961.

[7] Li, Z., Zhou, F., Han, X., Chen, J., Li, Y., Zhai, S., ... & Bao, Y. (2021). Numerical simulation and analysis of a geological disaster chain in the Peilong valley, SE Tibetan Plateau. *Bulletin of Engineering Geology and the Environment*, *80*, 3405-3422.

[8] Wang, H., Wang, X., Zhang, C., Wang, C., & Li, S. (2023). Analysis on the susceptibility of environmental geological

disasters considering regional sustainable development. *Environmental Science and Pollution Research*, *30*(4), 9749-9762.

[9] Qian, X. (2022). Regional Geological Disasters Emergency Management System Monitored by Big Data Platform. *Processes*, *10*(12), 2741.

[10] Liu, J., Zhao, J., Liu, Q., Su, A., Zhang, Q., Zhang, S., ... & Wang, Z. (2021). Integration and application of 3D visualization technology and numerical simulation technology in geological research. *Environmental Earth Sciences*, *80*, 1-7.

[11] Niu, H., Shao, S., Gao, J., & Jing, H. (2024). Research on GIS-based information value model for landslide geological hazards prediction in soil-rock contact zone in southern Shaanxi. *Physics and Chemistry of the Earth, Parts A/B/C*, *133*, 103515.

[12] Tan, Q., Huang, Y., Hu, J., Zhou, P., & Hu, J. (2021). Application of artificial neural network model based on GIS in geological hazard zoning. *Neural Computing and Applications*, *33*, 591-602.

[13] Wang, X., Wang, C., Jin, X., & Wang, H. (2023). Coordinated analysis of county geological environment carrying capacity and sustainable development under remote sensing interpretation combined with integrated model. *Ecotoxicology and Environmental Safety*, *257*, 114956.

[14] Gabdrakhmanova, N., Fedin, V., Matsuta, B., & Pilgun, M. (2021). The modeling of forecasting new situations in the dynamics of the economic system on the example of several financial indicators. *Procedia Computer Science*, *186*, 512-520.

[15] Curran, E. E., & Bowlick, F. J. (2022). Geographic information science education at Esri development center institutions. *Transactions in GIS*, *26*(1), 341-361.

[16] Yu, G., Bu, L., Wang, C., & Farooq, A. (2022). Composition analysis and distributed assumption GIS model of normal stress on the slope sliding surface. *Frontiers in Earth Science*, *10*, 923620.

[17] Motta, M., de Castro Neto, M., & Sarmento, P. (2021). A mixed approach for urban flood prediction using Machine Learning and GIS. *International journal of disaster risk reduction*, *56*, 102154.

[18] Raihan, A. (2023). A comprehensive review of the recent advancement in integrating deep learning with geographic information systems. *Research Briefs on Information and Communication Technology Evolution*, *9*, 98-115.

FUNDING

Supported by Yunnan Fundamental Research Projects (Grant No. 202301AU070022)

ABOUT THE AUTHOR

Fei Han

Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650031, Yunnan, China.

E-mail: 1589577617@qq.com

Jingkun Bao

Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650031, Yunnan, China.

E-mail: JingkunBao@hotmail.com

Kun Wang
Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology, Kunming 650500, Yunnan, China.
E-mail:kmwk2016@kust.edu.cn
Jiale An
Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology, Kunming 650500, Yunnan, China.
E-mail: 1870021348@qq.com
Zhongcai Gao
Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650031, Yunnan, China.
E-mail: 2653101881@qq.com
Yurong Li
Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650031, Yunnan, China.
E-mail: 1544781823@qq.com
Yongjun Li
Faculty of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650031, Yunnan, China.
E-mail: 1648154044@qq.com