ISSN:2790-1688 DOI: 10.56028/aetr.3.1.321

Research on regional landslide hazard warning model based on logistic regression

Yang Zhang[1*], Hui Liu[1], Chao Gu[1], Yunmei Zheng[2]

[1]State grid Shandong electric power research institute，Jinan 250003

[2]Unisplendour Software System Corporation Limited, Beijing 100084

*windml90@163.com

Abstract: In the process of urban construction and development, landslide disaster is the main factor affecting the health and safety of local residents. How to put forward effective prevention and management measures according to the previous accumulated experience is the main problem of scientific research. Although the traditional regional geological disaster early warning model played an important role during the work, it did not comprehensively predict and analyze the regional landslide disaster and its impact due to the low prediction accuracy, insufficient application precision and insufficient data collection information. Nowadays, scientific researchers propose to use logistic regression algorithm for optimization and innovation in practice. The regional landslide disaster prediction model with logistic regression as the core can better meet the needs of practical research, master the geological disaster prediction information of the local area, and help the staff of the department to quickly develop solutions. Therefore, on the basis of understanding the application and research status of regional landslide disaster warning model, this paper deeply discusses the regional landslide disaster warning model with logistic regression algorithm as the core according to the basic principle of logistic regression algorithm. The final experimental results show that the prediction model has strong generalization ability and training accuracy.

Keywords: Logistic regression; Landslide hazard; Urban construction; Early warning model

1. Introduction

From the perspective of urban construction and development, under the condition of changing climate condition, geological conditions of all parts of our country under the influence of external factors, can cause a series of landslide geological disasters, and harm local people's personal and property safety. [1.2.3]In order to effectively prevent the landslide geological disasters caused too serious security risks, at present, the meteorological department and relevant scholars integration of the main causes of landslide is studied and the specific influence, gradually strengthen the risk assessment of geological disasters, attaches great importance to the geological disaster risk assessment research, on the basis of clear slope stability, formulate scientific and standardized preventive measures, In order to curb the occurrence of landslide geological disasters from the source, reduce casualties and property losses. From the geological point of view, regional landslide means that the region has the necessary failure mode and geological conditions of landslide geological disaster. Generally speaking, the stability of the slope is mainly affected by the shear strength of rock and soil. When the shear stress of the slope exceeds the shear strength of the structural plane, the slope will have the problem of structural failure, resulting in sliding deformation. Landslide phenomenon refers to the slope rock mass under the action of external force and gravity out of balance, along the weak structural plane is bad phenomena of shear failure, a slope often appear weak structural plane shear failure phenomenon, this is also produced one of the origins of regional landslide disaster, outward development, gradually improve surrounding the slide will form a complete landslide zone.[4.5.6]

According to the experience obtained in the prediction and risk assessment of regional landslide disaster in the recent years, we can know that the landslide type is mainly divided into four kinds: first, refers to the sliding type, this kind of disaster is difficult to use the eyes and observe, need to use professional equipment and instruments to analyze the state; Secondly refers to the slow slope, this kind of disaster daily average sliding range control between a few centimeters to dozens of

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.321
centimeters, the staff can use the naked eye to observe; Thirdly, it refers to the medium speed landslide, where the average daily sliding range of disasters is between tens of centimeters and several meters; And finally, high-speed landslides, where a disaster can slide anywhere from a few meters to tens of meters per second. After mastering the main types of regional landslide disasters, scholars from various countries use artificial intelligence algorithm to carry out innovative research on the basis of integrating previous work experience. Among them, the early warning model is the key element of successful geological disaster meteorological early warning analysis, which presents the technical achievements obtained by domestic and foreign scholars in practice. At present, the research on regional geological disaster early warning model construction is mainly divided into two types, one is statistical early warning model, the other is dynamic early warning model. For example, some scholars have put forward the principle of displaying statistical early warning in practice, and invited many scholars to research and build their own statistical early warning models based on the basic characteristics of different regions, so as to provide technical support for the meteorological early warning work of geological disasters at all levels in our country. It should be noted that under the limitation of technical conditions, the research and application of early warning model still has some problems, such as low calculation accuracy and inaccurate prediction value. At the same time, some scholars proposed to use machine learning algorithm to build a new early warning model. The research results of this algorithm prove that it includes several key steps such as the construction of training sample set, optimization modeling, early warning output, etc. The composition process of training sample set is shown in Figure 1 below:

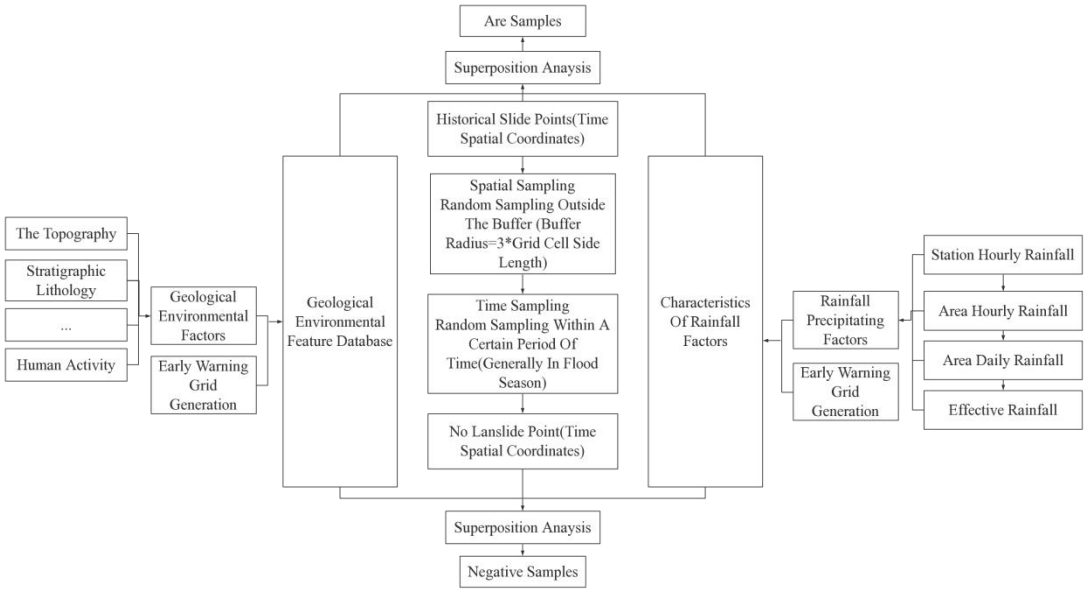

FIG. 1 Flowchart of training sample set composition

By using the construction method of the regional landslide training sample set, the negative samples were randomly sampled under the spatio-temporal constraints, so as to obtain the perfect training sample set. The best performance of each random forest algorithm, followed by logistic regression algorithm, artificial neural network algorithm and decision tree algorithm. Using logistic regression algorithm as an example, this paper mainly analyzes the application effect of landslide disaster warning model in a certain area, and provides effective basis for meteorological early warning work.

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.321

2. Method

2.1 Logistic regression algorithm

This kind of algorithm, also known as Logistic regression analysis, belongs to the supervised learning model in machine learning. The specific structure is shown in Figure 2 below:

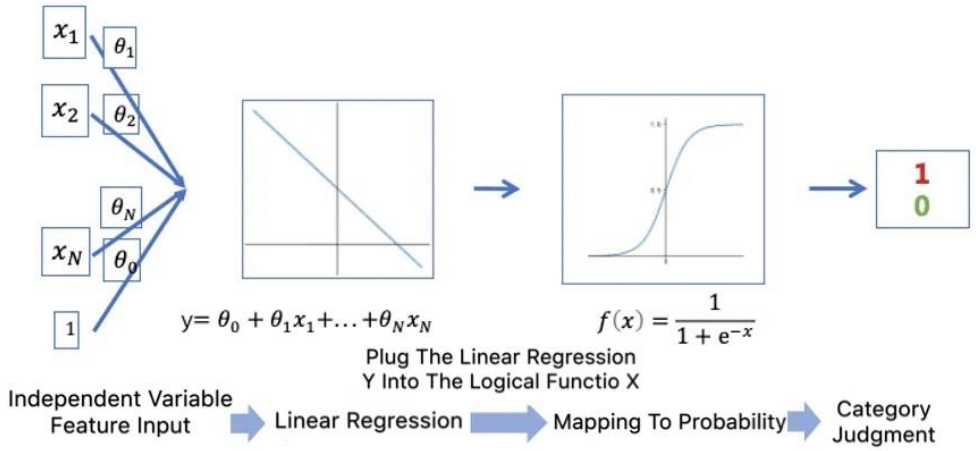

FIG. 2 Structure diagram of logistic regression algorithm

When dealing with binary problems, because there are two classes, one class is labeled as 0 and the other is labeled as 1. Choose a function that maps to a number between 0 and 1 for each set of input data. And if the value of the function is greater than 0.5, it is judged to belong to 1, otherwise it belongs to 0. In addition, undetermined parameters are needed in the function, which can accurately predict the data in the training set through sample training.[4-6]

This function is the sigmoid function of the form $\sigma(x)=\frac{1}{1+e^{-x}}$. In practical research, suppose the function is

$$h(x^2)=\frac{1}{1+e^{-(w^Tx^i+b)}}$$

In the above formula, xi represents the ith data of the test set and is the p-dimensional column vector $(x_1^i,x_2^i\ldots x_p^i)^T$; w represents the p-dimensional column vector $(w_1,w_2\ldots w_p)^T$, and is the parameter to be obtained; b is a number, and it's a parameter.

Combined with practical research, it is found that under the condition of w^Tx+b, the actual result is:

$$w_1x_1+w_2x_2+\ldots+w_px_p+b$$

Therefore, by converting w to $\begin{pmatrix} w_1,w_2\ldots w_p & b\end{pmatrix}^T$ and x^i to $\begin{pmatrix} x_1^i,x_2^i\ldots x_p^i & 1\end{pmatrix}^T$, we can get:

$$h(x^i)=\frac{1}{1+e^{-w^Tx^i}}$$

In the process of solving parameters, this paper mainly discusses the maximum likelihood estimation method. As one of the most common contents in the mathematical statistical analysis of data, it means that if a thing happens, the probability of it happening is the maximum. The actual formula is as follows:

$$\prod_{i=1}^{i=k}h(x_i)\prod_{i=k+1}^{n}(1-h(x_i))$$

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.321

In the above formula,i from 0 to k is the number of people belonging to category 1, and i from k+1 to n is the number of people belonging to category 0, n-k. Since y is the label 0 or 1, the above can also be written as:

$$\prod_{i=1}^{n} h(X_i)^{y_i}\left(1-h(X_i)\right)^{1-y_i}$$

So whether y is 0 or 1, one of them is always going to be raised to the 0 power, which is 1, which is the same thing as the first one.

And just to make it easier to study, let's calculate the logarithm of the logarithm. Because I'm maximizing this expression, I can convert it to multiplying it by minus 1, and then minimizing it. At the same time, for n data, the cumulative value will be large, and gradient descent will easily lead to gradient explosion. So you can divide it by the total number of samples.

$$L(w)=\frac{1}{n}\sum_{i=1}^{n}-y_i\ln\left(h(X_i)\right)-\left(1-y_i\right)\ln\left(1-h(X_i)\right)$$

There are many methods to find the minimum value, and gradient descent series methods are commonly used in machine learning. Or you could do Newton's method, or you could do w at zero.

2.2 Analysis of early warning model

On the one hand, get the cleaning data. Taking a certain area as an example, this paper mainly investigates the geological disaster and environmental survey results, encrypted rainfall monitoring data, supplementary survey and monitoring results of the early warning experimental area collected in recent decades. Among them, the number of training sample sets is 1826, and the actual characteristics and parameters are shown in Table 1 below:[7-9]

Table 1 Describes the collected data information

The serial number	The input features	Input characteristic parameter
1	Slope/(°)	1.0-10;2.10-20;3.20-30;4.30-40;5.≥40
2	Slope to/(°)	1.0-90；2.90-180；3.180-270；4.270-360
3	Elevation/M	1.0-800;2.800-1200;3.1200-1600;4.1600-2000，5.≥2000
4	Geomorphic types	1. Middle and low mountains, 2. Middle Mountains, 3. High mountains
5	Formation lithology	1. Loose accumulation, 2. Weak - semi - hard thin - medium rock formation, 3. Semi - hard thin - medium rock formation, 4. Hard - semi - hard medium-thick bedded rock group, 5. Unknown lithology
6	Distance from fracture /M	1.0-500;2.500-1000;3.1000-1500;4.1500-2000，5.≥2000
7	Annual rainfall/MM	1.0-500;2.500-800;3.800-1000;4.1000-1200，5.≥1200
8	Distance from house /M	1.0-200,2.≥200
9	Distance from road /M	1.0-200,2.≥200
10	Distance from ravine /M	1.0-200,2.≥200
11	Historical disaster points per grid unit	1.0,2.1,3.2,4.3-4,5.5-7
12	Daily rainfall /MM	1.<10,2.10-25,3.25-50,4.50-100,5.>100
13	Rainfall of the previous day /MM	1.<10,2.10-25,3.25-50,4.50-100,5.>100
14	Previous 2 days rainfall /MM	1.<10,2.10-25,3.25-50,4.50-100,5.>100
...	...	...
15	First 15 days rainfall /MM	1.<10,2.10-25,3.25-50,4.50-100,5.>100

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.321

After the training sample set is clearly defined, the missing qualities and abnormal values of the data should be identified and processed to ensure that the early warning model has stronger generalization ability and identification accuracy.[10-12]

On the other hand, build the model. Logistic regression algorithm, as one of the most common nonlinear binary dependent variable regression statistical models, can use maximum likelihood estimation method to study parameters and has consistent asymptotic normality. The corresponding algorithm flow is shown in Figure 3 below:

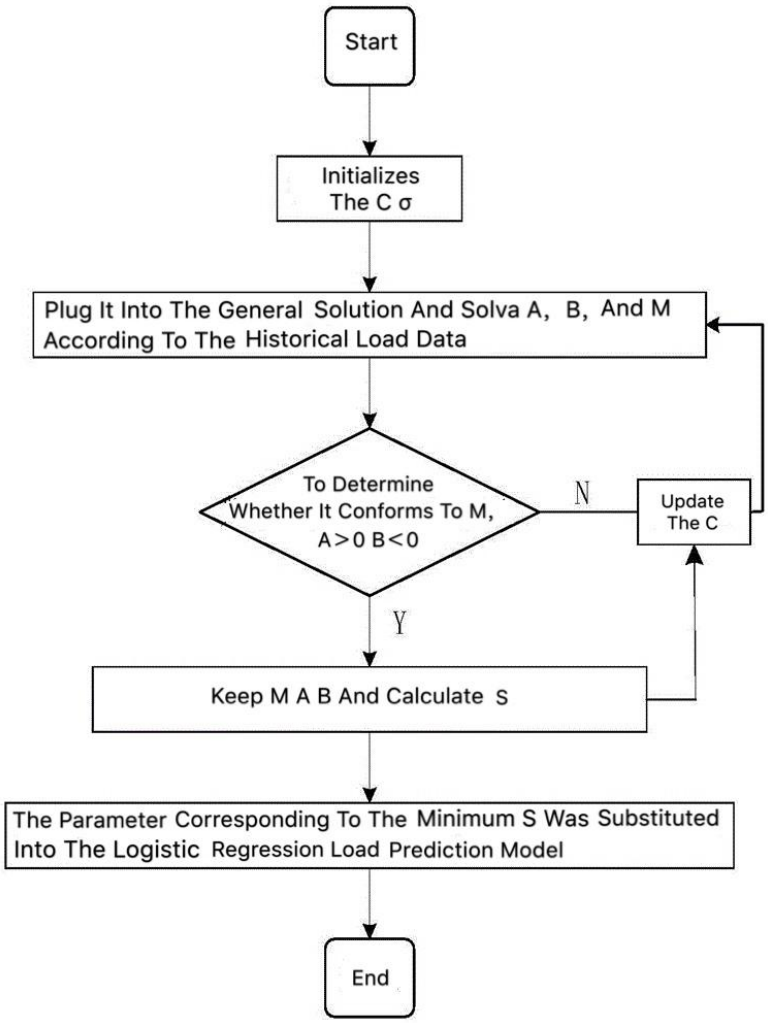

FIG. 3 Flow chart of logistic regression algorithm

In the research experiment of this paper, the training sample set was divided into the training set and the test set according to the ratio of four to one, and then the training analysis was carried out according to the expected experimental process. The optimization model parameters were verified by using the Bayesian optimization algorithm and the 50 fold cross validation. At present, the most common model parameter optimization methods can be divided into two kinds. On the one hand, it refers to the traditional method, whose optimization accuracy and speed are inversely proportional; On the other hand, it refers to the hyperparameter optimization algorithm, the most representative of which is the Bayesian optimization algorithm, which will use Gaussian process to increase the number of training samples, so as to fit the distribution of the objective function, and carry out optimization analysis in the cross verification. Each iteration results in a hyperparameter, which is optimized in the search for the optimal value.[13-15]

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.321

3. Result analysis

In this research experimental model, mainly from three aspects of verification analysis: first, accuracy. It refers to the accuracy of the application of the model, mainly studying the ratio between the number of correctly classified samples predicted by the model and the total number of samples. Secondly, the ROC curve and AUC value represent the generalization ability of the model. The higher the AUC value is, the stronger the application performance of the model is proved. Finally, the learning curve intuitively describes the fit level of the model. Among them, the classification results of the model are shown in Table 2 below:

Table 2 Model classification results

	Accurate rate	The recall rate	Since the score
0	0.949	0.975	0.957
1	0.950	0.883	0.915
Accuracy			0.943
Macro average	0.944	0.929	0.936
The weighted average	0.943	0.943	0.942

The learning curve is shown in Figure 4 below:

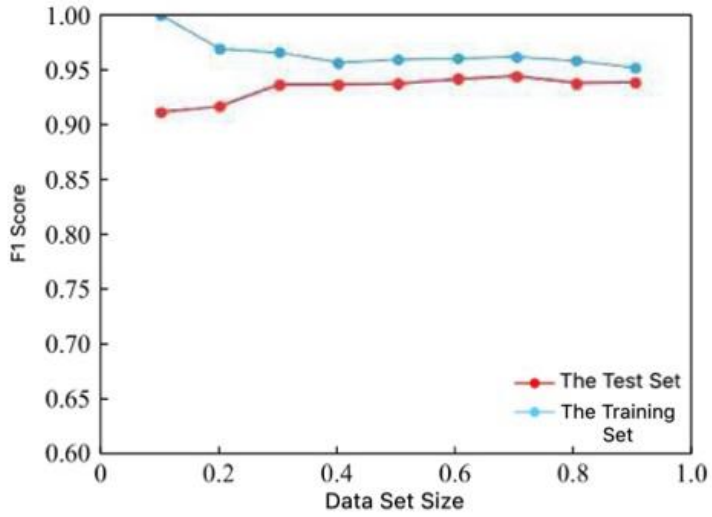

Figure 4 Learning curve

Combined with the result analysis shows that using a logistic regression model, construction of regional landslide disaster early warning model, to guarantee the precision of the numerical simulation prediction and scientific warning hierarchies are completed to improve early warning information systems staff provide specification, make sure that they can make effective protective measures according to their own accumulated experience. In the research experiment of this paper, if the output probability is greater than or equal to 40% and less than 60%, it belongs to the yellow warning. If the output probability is greater than or equal to 60% and less than 80%, it belongs to the orange warning. If the output probability is greater than or equal to 80%, it is a red alert. In this paper, 1,826 training samples were obtained, and the logistic regression algorithm was used to learn and train the landslide disaster early warning model. The actual numerical changes and curves can prove that the constructed model has strong generalization ability and prediction accuracy. Therefore, in the technological innovation of science and technology development in the future, research scholars in our country should continue to explore represented by logistic regression of machine learning and artificial intelligence algorithms such as cluster analysis, integrating it applied to the regional landslide disaster early warning analysis work, both to build application model with specific functions, and to change the traditional early warning management pattern, Pay attention to reduce the probability of landslide disaster from the basis.

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.321

4. Conclusion

To sum up, based on logistic regression algorithm can meet the demand of regional landslide disaster early warning model, not only in the practical work to integrate the local geological information collection and meteorological data, even after the process analysis, according to the algorithm processes the input operation early warning model, it is concluded that effective protective measures as soon as possible, so as to avoid landslide disaster health effects to residents' personal property. Therefore, in the development of modern science and technology innovation, Chinese scholars should continue to discuss with logistic regression as the core area of landslide disaster early warning model, pay attention to change the traditional working mode, fully displays the application value of the logistic regression algorithm, solve the problems existing in the practical prediction work, reduce the staff working pressure, improve regional landslide disaster early warning and protection.

Acknowledgements

Project Name and No.: (Project Name and Number) Research and application of wide-area monitoring and early warning technology of power grid facilities based on multi-source satellite, 520626210009

Reference

[1] Jiahui Xu, Hong Zhang, Wen Haijia, et al. Landslide susceptibility regionalization in Wushan County based on logistic regression [J]. Journal of Chongqing Normal University: Natural Science Edition, 2021, 38(2):9.

[2] Qiang He, Zengwu Wang, Dequan Lu. Zoning and evaluation of landslide disaster susceptibility in Nanchong City [J]. Journal of Chengdu University of Information Technology, 2021, 036(001):118-128.

[3] Yushan Xie. Review of financial risk Theory and early warning model [J]. Science & Technology Economics Guide, 2020, v.28; No.705(07):183+185.

[4] Xia Guo, Pengxiang Liu, Guojian Li. Logistic model analysis of disaster warning of loess landslide [J]. Energy and Environmental Protection, 2022, 44(4):7.

[5] Lujun Wang. Risk identification of Chinese stock market Bubble based on LSTAR Model [J]. 2021(2018-12):102-112.

[6] Dong Wang, Han Du, Qianling Wang. Research and Application of landslide warning Method Based on System clustering-Weighted Markov Coupling model [J]. Journal of China Coal Society, 2020, v.45; No.308(05):233-244.

[7] Zhaohua Wang, Jixian Zhang, Shuwen Yang, et al. Study on early warning of rain-type loess landslide based on Logistic model in Lanzhou City [J]. Science of Surveying and Mapping, 2020, 45(4):8.

[8] Jiazhu Wang, Renji Ba, Hua Ge, et al. Research on prediction model of gradual landslide impending slip based on MACD index [J]. Hydrogeology and Engineering Geology, 2022, 50:1-9.

[9] Shudong Chen. Landslide prediction model based on Logistic regression algorithm [J]. Microprocessor, 2021, 42(3):4.

[10] Lifeng Li, Xiaohu Zhang, Huilin Deng, et al. Landslide hazard susceptibility evaluation based on coupled entropy index and logistic regression model: A case study of Lantian County [J]. Science Technology and Engineering, 2020, 20(14):8.

[11] Jiali Bai. Study on Financial Risk Early Warning of Chinese Listed Companies: Based on Principal Component Analysis and Logistic Regression Model of Financial Risk early warning [J]. Business and Management, 2022(8):8.

[12] Jie Xiao, Hui Zhao, Yi Zhang, et al. Financial early warning of listed forestry companies based on Logistic model [J]. China Forestry Economics, 2021(4):5.

ISSN:2790-1688 DOI: 10.56028/aetr.3.1.321
[13] Research on Financial early warning Model of Listed Companies -- Empirical Analysis Based on Logistic Regression Model [J]. China Chief Accountant, 2022(5):3.

[14] Mingbo Li, Ping Chen, Zhihua Chen, et al. Research on comprehensive warning model of loose soil landslide based on multi-module [J]. Journal of Northwest Normal University: Natural Science Edition, 2020, 56(2):7.

[15] Yunli Wang, Zhenzhen Han, Wenhuan Yang, et al. Research on Risk early warning Model of Qualification Maintenance of High-tech Enterprises -- A Case study of high-tech Enterprises in Hebei Province [J]. Journal of Hebei Academy of Sciences, 2022(003):039.