International Core Journal of Engineering Volume 5 Issue 11, 2019
ISSN: 2414-1895 DOI: 10.6919/ICJE.201910_5(11).0008

# Forecasting of Landslide Stability Based on Gradient Boosting Decision Tree Model

Fei Niu [a], Lianwu Chen [b]

Xi'an University of Science and Technology, Xi'an 710054, China.

[a]714147074@qq.com, [b]comeon123go@126.com

## Abstract

**In order to forecast landslide stability, an ensemble learning method in machine learning, Gradient Boosting Decision Tree, was used to forecast landslide stability. Combined with the collected landslide data, six parameters, including unit weight of landslide material, cohesion force, internal angle of friction, landslide angle, landslide height, pore pressure ratio, were utilized as the input parameters, while factor of safety was used the output parameter, establishing prediction model. The influence of the main parameters of the GBDT algorithm on the training results was analyzed, and the better parameters were selected. The prediction results were compared with the support vector machine regression and BP neural network. The results show that the prediction results of the Gradient Boosting Decision Tree model are reliable, the average relative error is 6.53%. The prediction accuracy is high and the feasibility is strong. As a means of landslide stability prediction, its application prospect is good.**

## Keywords

**Ensemble learning, GBDT, Landslide stability, Parameter selection.**

## 1. Introduction

There will be landslide stability problems in construction projects such as mining and water conservancy. Landslide instability has adversely affected the construction of the project and may threaten people's lives and property and cause significant losses to the country's economic construction. The stability of landslide is affected by the combination of geological factors and engineering factors, and most of the factors have the characteristics of uncertainty such as randomness, ambiguity and variability. Therefore, it is difficult to obtain accurate and effective prediction results when studying the stability of landslides.

Common methods for predicting landslide stability include limit equilibrium method and numerical analysis method. However, due to the complexity of the landslide system itself and the influencing factors, it is difficult to obtain accurate landslide stability results by applying the above method. Machine learning algorithm is a method to realize artificial intelligence, which can solve problems such as classification and prediction. Many scholars apply it to landslide stability prediction and achieve good results. For example, support vector machines, Gaussian process machine learning, random forests, and so on. The gradient ascending decision tree, as an integrated learning method in machine learning, has strong learning ability and accurate prediction. In recent years, it has performed well in data mining and has received extensive attention. In this paper, based on the gradient decision tree algorithm, a prediction model is established to predict the stability of the landslide. The feasibility and effectiveness of the proposed method are verified by an example. The prediction results are compared with support vector machine regression and BP neural network, and good results are obtained.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201910_5(11).0008

## 2. GBDT Algorithm

GBDT (Gradient Boosting Decision Tree), which is called Gradient Lift Decision Tree, is called MART (Multiple Additive Regression Tree), GTB (Gradient Tree Boosting), etc. It is an integrated learning method in the field of machine learning. The basic idea of the GBDT model is that by constructing multiple weak classifiers, the additive model is combined into a strong classifier. At each iteration, the latter weak classifier learns the residual of the previous weak classifier and subtracts the residual. Create a new combined model in a small gradient direction. The following is a detailed introduction:

The known training set is:

$$T = \{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\} \qquad x_i \in \chi \subseteq R^n \tag{1}$$

Among them, $\chi$ is the input space, $y_i \in \gamma \subseteq R$, $\gamma$ is the output space. For the regression problem, the base learner of the GBDT algorithm is a binary regression tree, and the addition model is used to form a strong classifier. The model is as follows:

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m) \tag{2}$$

Among them, $T(x;\Theta_m)$ is the decision tree, $\Theta_m$ is the parameter of the decision tree, and M is the number of trees. To optimize the problem by the forward distribution algorithm, first determine $f_0(x) = 0$, the mth model is:

$$f_m(x) = f_{m-1}(x) + T(x; \Theta_m), m = 1,2,..., M \tag{3}$$

$f_{(m-1)}(x)$ is the current model. The loss function is:

$$L(y, f_m(x)) = L(y, f_{m-1}(x) + T(x; \Theta_m)) \tag{4}$$

When the squared error is used as the loss function, it can be written as:

$$L(y, f_m(x)) = [y - f_{m-1}(x) - T(x; \Theta_m)]^2 \tag{5}$$

Where $y - f_{m-1}(x)$ is the residual. By fitting the residuals, the regression tree $T_m(x;\Theta_m)$ is obtained, and the model is updated to obtain 4. When the error meets the requirements, stop updating and get the final model:

$$f_M(x) = T_1(x) + T_2(x) + ... + T_M(x) \tag{6}$$

When the loss function is more general, the optimization is difficult. Friedman proposed a gradient lifting algorithm. The principle is that the negative gradient of the loss function is used as the residual approximation, that is:

$$r_{mi} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}(x)} \tag{7}$$

Then fit the regression tree.

## 3. GBDT Landslide Stability Prediction Model

The stability of landslide is affected by many factors, and its main controlling factors are:

(1) Rock and soil properties, including soil bulk density, cohesion, internal friction angle, etc.;

(2) landslide topography, including landslide height, slope, geometry, etc.;

(3) The role of water, including rainfall, engineering drainage, groundwater seepage, etc.;

(4) External loads, including earthquakes, artificial activities, etc.;

In this paper, six important factors, such as bulk density, cohesion force, internal friction angle, landslide angle, landslide height and pore water pressure ratio, are selected as inputs, and the safety

ISSN: 2414-1895 DOI: 10.6919/ICJE.201910_5(11).0008

factor is taken as the output. Establish a landslide stability prediction model, steps are as follows:

(1) Based on the collected landslide examples, filter the data to determine the input variables (features) and output variables (markers).

Sample data preprocessing.

Establish GBDT prediction model, analyze the influence of main parameters on training results, and use grid search method to select better parameters.

Verify the model.

(5) Input the test sample into the trained model for prediction.

The specific operation process is shown in Fig. 1 :

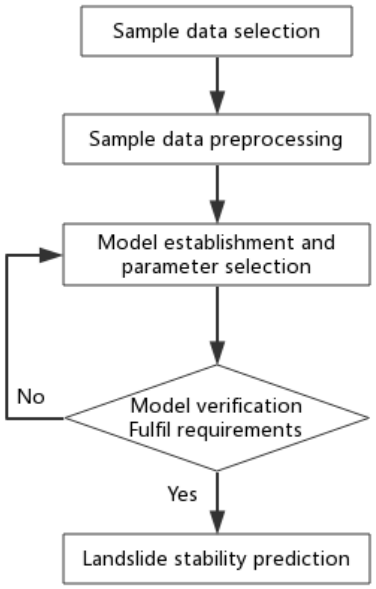


Figure 1. Flow chart of GBDT prediction model

## 4. Instance Application

From the 67 landslide examples, 57 samples were randomly selected as training samples, and 10 samples were used as test samples to establish a GBDT model. The sample data is shown in Table1.

Table 1. Sample data of landslide

| Number | Bulk weight （$kN/m^3$） | Cohesion/kPa | Internal friction angle/(°) | Landslide angle/(°) | Landslide height/m | Pore pressure/kPa | Safety factor |
|---|---|---|---|---|---|---|---|
| 1 | 18.8 | 20 | 10 | 25 | 50 | 0.3 | 0.97 |
| 2 | 20 | 20 | 36 | 45 | 50 | 0.5 | 0.83 |
| 3 | 18.8 | 25 | 20 | 30 | 50 | 0.2 | 1.21 |
| 4 | 19.1 | 10 | 20 | 30 | 50 | 0.4 | 0.65 |
| 5 | 20.6 | 16.28 | 26.5 | 30 | 40 | 0 | 1.25 |
| 6 | 12 | 0 | 30 | 45 | 4 | 0 | 1.44 |
| 7 | 20 | 0 | 24.5 | 20 | 8 | 0.35 | 1.37 |
| 8 | 21.4 | 10 | 30.34 | 30 | 20 | 0 | 1.7 |
| 9 | 22 | 20 | 22 | 20 | 180 | 0.1 | 0.99 |
| 10 | 28.44 | 39.23 | 38 | 35 | 100 | 0 | 1.99 |

ISSN: 2414-1895 DOI: 10.6919/ICJE.201910_5(11).0008

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 11 | 19.1 | 10 | 10 | 25 | 50 | 0.4 | 0.65 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 64 | 28.44 | 29.42 | 35 | 35 | 100 | 0 | 1.78 |
| 65 | 23 | 0 | 20 | 20 | 100 | 0.3 | 1.2 |
| 66 | 18.84 | 14.36 | 25 | 20 | 30.5 | 0.45 | 1.11 |
| 67 | 21 | 20 | 40 | 40 | 12 | 0 | 1.84 |

### 4.1 Sample Data Preprocessing

In order to eliminate the influence of excessive magnitude difference, the iterative convergence speed is accelerated and the data is standardized.

$$X = (X_i - \mu)/\sigma, Y = (Y_j - \mu)/\sigma \quad (8)$$

Where, $X_i$ is the training sample; $Y_i$ is the test sample; $\mu, \sigma$ respectively, the mean and variance of the training sample.

### 4.2 Model Establishment and Parameter Selection

In this paper, the model is established by the machine learning open source framework Scikit-Lear. The GBDT algorithm has many parameters. According to the actual situation of the sample, three important parameter adjustment models: the number of decision trees (the number of iterations), the maximum depth and the learning rate are selected. Other parameters are default parameters. Through the adjustment of three parameters, the impact on the training results is analyzed, and the training results are evaluated by the mean square error. See Fig. 2 ~ Fig. 4.

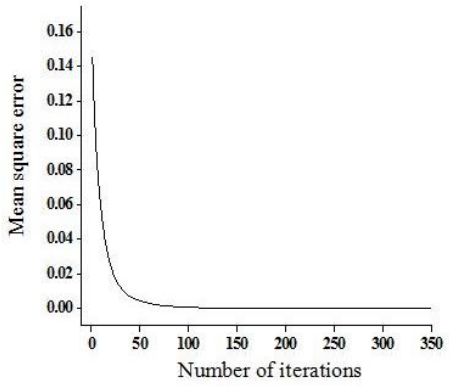


Figure 2. Adjust the number of regression trees

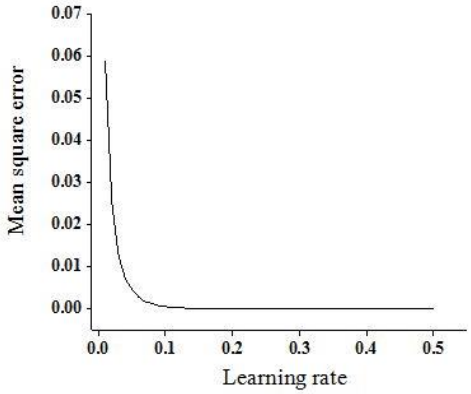


Figure 3. Adjust the learning rate

ISSN: 2414-1895 DOI: 10.6919/ICJE.201910_5(11).0008

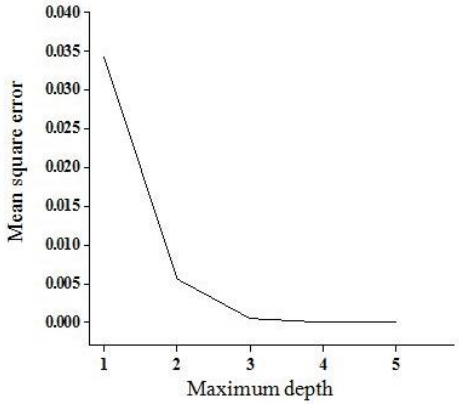


Figure 4. Adjust the maximum depth

It can be seen from Fig. 2 to Fig. 4 that the larger the three parameter values of the number of iterations, the maximum depth and the learning rate, the smaller the mean square error of the training result and the higher the prediction accuracy. It also reflects the strong learning ability of the GBDT model, and generally does not appear under-fitting. However, improper selection of parameters can cause over-fitting. For parameter selection, there is no uniform method today. Combined with the above analysis, the training is carried out under different combinations of iterations of 200~300, maximum depth of 2~5, and learning rate of 0.05~0.1. The learning curve is used to judge the quality of the model, and then the better parameters are selected. The basic idea of the learning curve is to divide the training set into two parts, one for training and the other for verification. When the number of training samples is small, the noise is less and the training results are more closely fitted. However, due to the small number of samples, the trained model has poor generalization ability, and the performance error on the verification set is large. As the number of samples increases, the error on the training set increases, the generalization ability of the model increases, and the error on the verification set decreases. A better model should show that the error curve interval between the training set and the verification set decreases as the number of samples increases. Through experimental analysis, the number of iterations is 260, the maximum depth is 2, the learning rate is 0.075, and the model is better. The learning curve is shown in Fig. 5.

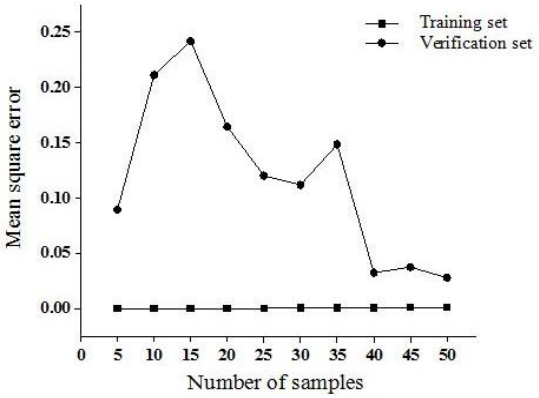


Figure 5. Learning curve

## 4.3 Result Prediction

The test sample is input into the trained GBDT model to obtain the predicted value of the safety factor and compared with other model prediction results. The specific results are shown in Table 2.

Table 2. Comparison of forecast results

| Number | Actual value | Predictive value | | | Absolute error/% | | | Relative error/% | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SVR | BP | GBDT | SVR | BP | GBDT | SVR | BP | GBDT |
| 58 | 1.45 | 1.4519 | 1.4658 | 1.4518 | 0.0019 | 0.0158 | 0.0115 | 0.13 | 1.09 | 0.79 |

ISSN: 2414-1895 DOI: 10.6919/ICJE.201910_5(11).0008

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 59 | 0.96 | 1.1548 | 1.0207 | 0.9499 | 0.1948 | 0.0608 | 0.0101 | 20.29 | 6.33 | 1.05 |
| 60 | 0.89 | 0.7853 | 0.8669 | 0.9256 | 0.1047 | 0.0231 | 0.0356 | 11.77 | 2.60 | 4.00 |
| 61 | 0.8 | 0.9987 | 1.1054 | 0.8991 | 0.1987 | 0.3054 | 0.0991 | 24.83 | 38.17 | 12.39 |
| 62 | 1.08 | 1.1240 | 0.4265 | 1.0791 | 0.0440 | 0.6535 | 0.0009 | 4.07 | 60.51 | 0.08 |
| 63 | 1.13 | 1.3957 | 1.2585 | 1.0324 | 0.2627 | 0.1285 | 0.0976 | 23.24 | 11.38 | 8.63 |
| 64 | 1.78 | 1.9500 | 1.7156 | 1.8091 | 0.1701 | 0.0644 | 0.0291 | 9.55 | 3.62 | 1.63 |
| 65 | 1.2 | 1.3521 | 1.0420 | 1.2000 | 0.1521 | 0.1580 | 0.0096 | 12.67 | 13.17 | 0.80 |
| 66 | 1.11 | 1.3112 | 1.3464 | 1.4184 | 0.2012 | 0.2364 | 0.3084 | 18.13 | 21.30 | 27.78 |
| 67 | 1.84 | 1.6814 | 1.6852 | 1.6892 | 0.1586 | 0.1548 | 0.1508 | 8.62 | 8.41 | 8.19 |
| Average value | | | | | 0.15 | 0.18 | 0.08 | 13.33 | 16.66 | 6.53 |

From Table 2, the GBDT model has an average relative error of 6.53% when predicting 10 sets of data on the test set. It can be used as a means of predicting landslide stability, and the average relative error is significantly lower than the average relative error of the two methods of support vector machine regression and BP neural network.

In order to further verify the learning ability of GBDT, the training results are compared with support vector machine regression (SVR) and BP neural network (BP), as shown in Fig. 6. As can be seen from the figure, when the GBDT model predicts the training set, the output safety factor is closer to the actual value than the other two methods.

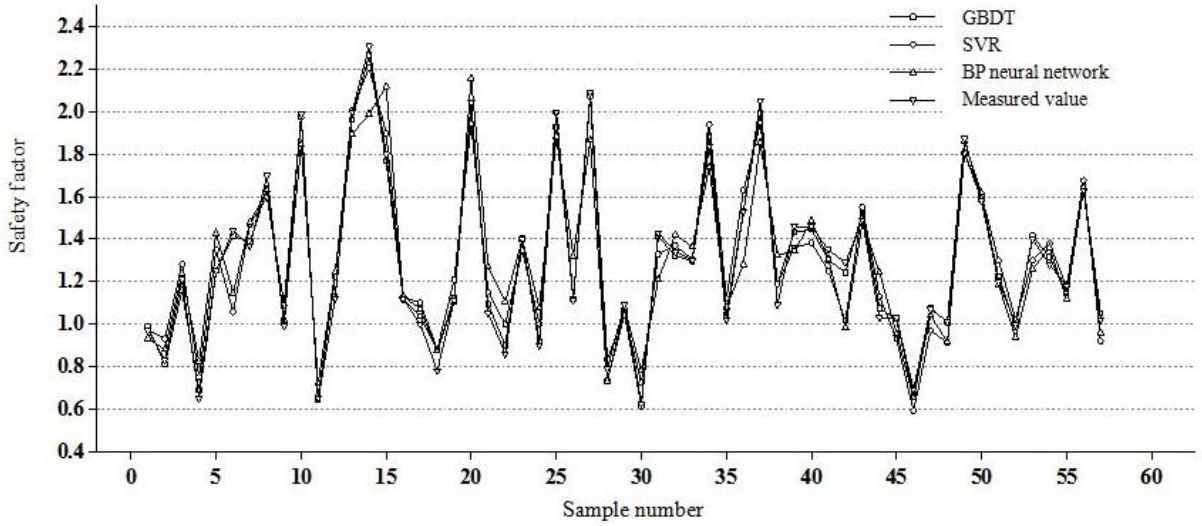


Figure 6. Contrast diagram of training results

## 5. Conclusion

(1) The influence of GBDT main parameters on the training results is analyzed. The grid search method is used to determine the number of decision trees, the maximum depth, and the learning rate parameters. The model is judged and a better model is established. The prediction results of this model are compared with the support vector machine regression and BP neural network prediction results. The results show that the GBDT model is better than the two.

(2) In order to further improve the prediction accuracy of landslide stability, it should be considered according to the local actual situation, how to select the influencing factors of landslide stability, such as seasonal rainfall, artificial activities, earthquakes, etc., and the scientific quantification of indicators needs further the study.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201910_5(11).0008

## References

[1] Y. H. Zhao, G. Liu, et al. Sensitivity analysis for the stability of loess landslide Based on grey correlation degree, Journal of Yangtze River Scientific Research Institute, Vol. 32 (2015) No. 7, p.94-98.

[2] C. Guo, Q. Xu, Y. Wei, et al. Analysis of Propagation and Deposit Characteristics and Liquidity of Jiangliu 4# Landslide in Southern Jingyang Plateau, Shaanxi Province, Science Technology and Engineering,Vol. 17 (2017) No.27, p.15-25.

[3] L.B. Shao, H. Ma, T. X. Wen. Study on landslide stability prediction based on RF—ELM model, Journal of Safety Science and Technology, Vol. 11 (2015) No.03, p.93-98.

[4] S. Su, Y.C. Song, L. B. Yan. Application of Gaussian process machine learning to landslide stability evaluation, Rock and Soil Mechanics, Vol. 30 (2009) No.03, p.675-679.

[5] J .M. Yang, C. Gao, Z. Y. Qu, et al. Random Forest Classification Algorithm Based on Cost-sensitive for Imbalanced Data, Science Technology and Engineering, Vol. 18 (2018) No.06, p.285-290.

[6] Z.Y. Luo, X. J. Yang, X. N. Gong, Support vector machine model in landslide stability evaluation, Chinese Journal of Rock Mechanics and Engineering, (2018) No.01, p.144-148.

[7] H. Li: Statistical learning method (Tsinghua University Press, Beijing, China 2012), p.147-152.

[8] Y. C. Huang: Machine Learning byscikit- learn:Algorithms and Practices (China Machine Press, Beijing, China 2018), p.115-120.