Proc. of the 8th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Jinsong Huang, D.V. Griffiths, Shui-Hua Jiang, Anna Giacomini and Richard Kelly
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-5182-7_08-007-cd

ISGSR 2022

A Machine Learning Prediction Model for Rockburst Based on Oversampling Algorithm and Bayesian-XGBoost

Qing Kang[1] and Yong Liu[2]

[1]State Key Laboratory of Water Resources and Hydropower Engineering Science,
Wuhan University, 299 Bayi Road, Wuhan 430072, P. R. China.
E-mail: kangqing@whu.edu.cn (Corresponding Author)
[2]State Key Laboratory of Water Resources and Hydropower Engineering Science,
Wuhan University, 299 Bayi Road, Wuhan 430072, P. R. China.
E-mail: liuy203@whu.edu.cn

Abstract: Rockburst is a key source of risk in engineering construction of deep-buried tunnels, which is induced by high in-situ stress and strong dynamic disturbance. Due to the highly complex relation between rockburst and the impact factors, traditional mechanism-based prediction methods have some limitations. Extreme gradient boosting (XGBoost) is herein introduced to predict the rockburst problem, where six hyperparameters are optimized by Bayesian optimization algorithms. This study collected 384 data sets based on real rockburst cases. Rockburst prediction can be divided into four categories: no rockburst, light rockburst, moderate rockburst and high rockburst. The occurrence probability of these four categories is different in actual engineering, which leads to imbalanced samples of the four rockburst categories. The synthetic minority oversampling technique (SMOTE) algorithm is utilized to process the collected data sets as the sample imbalance can affect the prediction accuracy of XGBoost model. The SMOTE-Bayesian-XGBoost model proposed in this study does not rely on the internal mechanism of rockburst. It provides a simple but effective model for rockburst prediction, which is of practical significance to reduce the risk during deep-buried tunnel construction.

Keywords: Rockburst prediction; extreme gradient boosting; Bayesian optimization algorithms; synthetic minority oversampling technique.

1 Introduction

In recent years, the continuous improvement of underground engineering facilities has brought many risks to deep engineering excavation, especially rockburst disasters. The mechanism of rockburst is very complicated, mainly due to the formation of high-stress concentration areas in deep-buried areas or the large energy generated by rock mass during failure (Zhang 2022). Therefore, the timely prediction of rockburst is an urgent problem to be solved in underground engineering.

Many machine learning methods have been applied to rockburst prediction. Ahmad et al. (2021) compared the application of J48 algorithm and random forest algorithm in rockburst prediction, and the results show that the random forest method has higher accuracy. Based on analytic hierarchy process (AHP), Yin et al. (2021) proposed a new tree-based algorithm to predict the occurrence of rockburst. Firstly, t-SNE and clustering algorithm were used to reduce and cluster the database, then grouping rules established previously were used to predict the rockburst. Wojtecki et al. (2021) used a wide range of machine learning methods to predict rockburst in a deep coal mine in Poland, which proves the effectiveness of machine learning methods in rockburst prediction. Zhang et al. (2021) used the risk synthesis index method to evaluate the rockburst in the deep coal seam group. Liang et al. (2019) introduced the multi-attributive border approximation area comparison (MABAC) method to rockburst risk assessment and obtained the specific rockburst risk level in a fuzzy environment. However, these methods have certain limitations: (1) Sample disequilibrium; the amount of data for each type of rockburst is uneven in the dataset is unbalanced, which will affect the prediction results of rockburst. (2) Hyperparametric optimization; most of the existing methods are based on search grid or local optimization, which is not only computationally expensive but also prone to local optimization.

In this study, a Bayesian-XGBoost rockburst prediction model based on oversampling technique is proposed. SMOTE algorithm solves the imbalance of samples in each category in rockburst prediction. The Bayesian method is used to optimize parameters, which can optimize the next iteration according to the previous calculation results and greatly improve the optimization process. The main work of this study is as follows: Firstly, the rockburst database is established by collecting data from literature. Then, the proposed model is trained based on the collected database. Finally, the validity of the proposed model is verified based on an actual engineering case.

2 Methods

2.1 Machine learning algorithms

2.1.1 EXtreme Gradient Boosting (XGBoost)

XGBoost is a tree-based integration algorithm. The main idea is to combine multiple classification tree models to build a model with higher accuracy.

1. The basic model of XGBoost is shown below:

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

where $\hat{y}_i$ represents the predicted results of sample x_i, K is the number of trees. $i = 1, 2, 3.....,n$, n is the number of the samples. F is the set of trees and f_k is one of the functions.

2. The loss function f_{obj} can be expressed as:

$$\begin{cases} L(\theta) = l(\hat{y}_i, y_i) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \\ \Omega(\theta) = \sum_{k=1}^{K} \Omega(f_k) \end{cases} \rightarrow f_{obj} = L(\theta) + \Omega(\theta) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where $L(\theta)$ represents the error term, y_i represents the true value. $\Omega(\theta)$ represents the regularization item, $\Omega(f_k)$ represents the regularization term of the kth tree.

Since XGBoost model is trained by addition, the loss function will change every time a new function $f(x_i)$ is added, which can be expressed as:

$$f_{obj}^{(t)} = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{t-1} + f_t(x_i)))^2 + \Omega(f_t) + C \tag{3}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{4}$$

where $f_t(x_i)$ represents the function added at the tth time; C is the constant term. T represents the number of leaf nodes; ω represents the fraction of each leaf node. γ and λ control the number and fraction of leaf nodes, respectively.

Taylor expansion was performed on Eq. (3) by combining Eq. (4):

$$\begin{aligned} f_{obj}^{(t)} &= \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \end{aligned} \tag{5}$$

where g_i and h_i represent the first partial derivatives and second partial derivatives, respectively.

Eq. (6) and Eq. (7) are used to simplify Eq. (5), and the simplified formula is shown in Eq. (8):

$$G_j = \sum_{i \in I_j} g_i \tag{6}$$

$$H_j = \sum_{i \in I_j} h_i \tag{7}$$

$$f_{obj}^{(t)} = \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2} \left(H + \lambda_j \right) \omega_j^2 \right] + \gamma T \tag{8}$$

where $I_j = \{ i | q(x_i) = j \}$, $q(x)$ represents the leaf node of sample x, I_j represents the set of samples of every leaf in the jth tree.

3. To get the optimal objective function, take the partial derivative of ω_j with respect to Eq. (8), the optimal objective function can be solved:

$$G_j + (H_j + \lambda)\omega_j = 0 \tag{9}$$

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}, f_{Obj} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{10}$$

2.1.2 Other machine learning algorithms

In order to compare with the proposed model, six algorithms are introduced in this study, including support vector machine classification (SVC); decision tree (DT); random forests (RF); Adaptive boosting (Adaboost); K-Nearest-Neighbors (KNN); Multilayer perceptron (MLP).

2.2 Oversampling technique (SMOTE)

In classification prediction problems, the number of categories per category may be unbalanced due to limited data sets. In the process of prediction, this may lead to the excessive prediction of the model for the categories with a large amount of data, and the prediction results will have a large deviation.

Machine learning algorithms usually take the maximum accuracy as the objective function, which will lead to the algorithm paying too much attention to the majority of samples. Therefore, the predictive performance of the algorithm will decrease for a few samples. In order to solve this problem, SMOTE method is introduced to increase the minority sample. The specific steps of SMOTE are as follows:

(1) For minority classes, the Euclidean distance is used to calculate the distance from each sample to all samples in the minority sample set.

(2) According to the imbalance ratio of each category, set a sampling ratio, and multiple samples are randomly selected from the k nearest neighbor of sample x of each minority category, assuming that the selected nearest neighbor is x_n.

(3) For the selected neighbor x_n, a new sample is constructed according to the following formula:

$$x_{new} = x_n + rand(0,1) \times (\tilde{x}_n - x_n) \tag{11}$$

where x_{new} represents the newly generated sample of the minority class; x_n and $\tilde{x}_n$ represent the minority class sample and the selected nearest neighbor points, respectively; rand (0,1) represents random numbers between (0,1).

2.3 Bayesian optimization

In order to speed up the training speed of the neural network and improve the computational efficiency, it is necessary to optimize the hyperparameters of the neural network. The traditional parameter tuning methods include grid search and random search, but these two methods have a large amount of computation and are easy to fall into the local optimal situation. In the process of searching the optimal parameters, these two methods will not be adjusted based on the past evaluation results, which has certain limitations. In order to speed up the process of parameter optimization, a large number of parameter optimization methods have been developed in recent years. The three most common methods are Bayesian optimization; particle swarm optimization; genetic algorithm.

In this study, the Bayesian optimization method is used to optimize parameters. The main principle of the Bayesian algorithm is to determine the optimal hyperparameter of the model through global optimization. The Bayesian optimization algorithm can combine the previous results to optimize the next calculation. The algorithm core is the probability proxy model and the acquisition function. In this study, the common Gaussian model and the improved probability acquisition method are adopted.

3 Construction of SMOTE-Bayesian-XGBoost Model for Rockburst Prediction

3.1 Database description

3.1.1 Database establishment

A total of 384 sets of rockburst data are collected in this study. They are mainly derived from Zhou et al. (2016); Xue et al. (2019); Dong et al. (2013); Wang et al. (2013); Zhou et al. (2013). According to the intensity of rockburst, it can be divided into four grades: no rockburst damage (0), slight rockburst (1), medium rockburst (2) and strong rockburst (3). Six indexes are selected in this project to evaluate the rockburst level: maximum tangential stress (MTS); stress concentration factor: ratio of maximum tangential stress to uniaxial compressive strength (SCF); brittleness coefficient: ratio of uniaxial compressive strength to uniaxial tensile strength (BC); uniaxial compressive strength (UCS); uniaxial tensile strength (UTS), elastic energy index (W_{et}).

3.1.2 Database processing

In order to ensure the prediction effect of the model, it is necessary to preprocess the data before training. In the collected dataset, the distribution of the four types of rockburst data is uneven. The four rockburst categories are no rockburst (71 cases), light rockburst samples (114 cases), moderate rockburst (137 cases) and high rockburst (62 cases).

It can be seen that the data amount of slight rockburst and moderate rockburst accounts for a large proportion. To solve this problem, random sampling technique is used to increase the frequency of a few types of samples, so that the sampling frequency of different types of rockburst cases can reach a balance. After data processing, the data amount of each sample reached 137 cases, a total of 548 groups.

3.2 Model training

In order to avoid over-fitting and increase the generalization effect of the model, the 10-fold cross validation method is adopted here. Figure 1 shows the principle of 10-fold cross validation. The final calculation result is the average value of 10 times, and the core in the figure refers to the accuracy of prediction.

In this study, accuracy rate is used to represent the prediction effect of the model, and accuracy rate can be defined as:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} l(\hat{y}_i, y_i) \tag{12}$$

$$l(\hat{y}_i, y_i) = \begin{cases} 1 & \hat{y}_i = y_i \\ 0 & \hat{y}_i \neq y_i \end{cases} \tag{13}$$

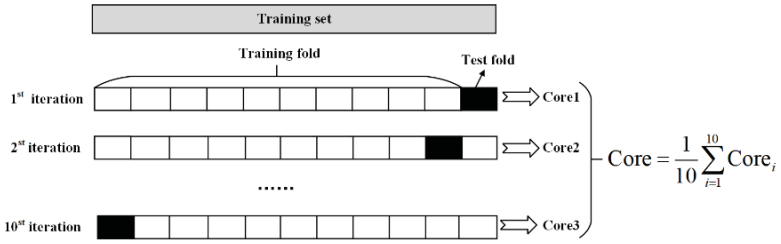

Figure 1. Schematic diagram of cross validation.

In order to compare with the prediction effect of XGBoost, six algorithms were selected to train the dataset. Table 1 indicates the predicted results under XGBoost model and the other six algorithms. Figure 2 shows the rockburst prediction results of 7 algorithms in the original data and oversampling datasets more intuitively.

Table 1. The rockburst prediction accuracy based on raw data and oversampled data.

Algorithm	Accuracy rate (%) (Raw data)	Accuracy rate (%) (Oversampled data)
XGBoost	74.15	82.70
Adaboost	70.23	78.70
K-NN	61.34	76.14
DT	63.94	73.02
RF	73.08	82.15
SVC	63.18	76.32
MLP	64.27	75.76

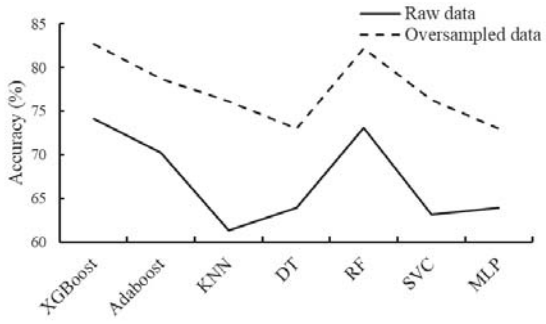

Figure 2. Rockburst accuracy of various algorithms based on raw data and oversampled data.

The following phenomena are shown in Figure 2 and Table 1: (1) XGBoost model has the highest prediction accuracy, followed by Adaboost and RF. The prediction accuracy of these three models is more than 70% in the original data set and about 80% after over-sampling. This is because these three algorithms are integrated classification algorithms, while the remaining four algorithms are single classification methods. (2) After oversampling the data set, the accuracy of each model increases greatly. This proves that it is necessary to carry out sample equalization on the data set.

In order to improve the prediction efficiency of XGBoost, this study uses Bayesian optimization algorithm to optimize the hyperparameters of XGBoost model. For the two penalty coefficients L1 and L2 of XGBoost model, the default value 0 is generally adopted and these two parameters are not optimized here. Table 2 shows the optimal values for the remaining six parameters. After hyperparametric optimization of XGBoost, the accuracy of prediction is improved to 84.14%. It proves that Bayesian optimization parameters can improve the accuracy of the model.

Table 2. XGBoost model parameters optimization.

Parameter	Meaning	Optimal value	Accuracy (%)
n_estimator	number of classifiers	145	84.14
eta	shrinkage step	0.24	
max_depth	maximum depth of a tree	5	
min_child_weight	sum of sample weights of minimum leaf nodes	5.46	
max_leaf_nodes	maximum incremental step for each tree's weight estimation	3	
subsample	random sampling ratio	0.65	

4 Application in The Practical Engineering of Models

In order to verify the validity of the proposed model, part of the diversion tunnel section of the riverside hydropower station is selected as a validation case. The diversion tunnel is located in the area of medium and high in-situ stress, which is prone to rockburst and brings serious harm. The 10 groups of data measured on-site (Xue et al. 2020) are shown in Table 3. Also, Table 3 indicates the predicted results based on the Bayesian-XGBoost model. It can be seen that the rockburst level is correctly predicted for each group of samples which can verify the validity of the model.

Table 3. Practical engineering applications of the SMOTE-Bayesian-XGBoost.

No.	MTS	UCS	UTS	SCF	BC	W_{et}	Actual level	Predicted level
1	91.43	157.63	11.96	0.58	13.18	6.27	3	3
2	19.14	106.31	2.76	0.18	38.52	2.03	0	0
3	58.05	147.85	6.98	0.39	21.18	3.62	2	2
4	34.89	151.7	7.47	0.23	20.31	3.17	1	1
5	51.5	132.05	6.33	0.39	20.86	4.63	2	2
6	35.82	127.93	4.43	0.28	28.9	3.67	1	1
7	9.74	88.51	2.16	0.11	40.98	1.77	0	0
8	33.94	117.48	4.23	0.29	27.77	2.37	1	1
9	18.32	96.41	2.01	0.19	47.93	1.87	0	0
10	110.35	167.19	12.67	0.66	13.2	6.83	3	3

Based on the XGBoost algorithm, Figure 3 shows the importance of each input parameter to the rockburst level. It can be seen that MTS and UCS have the greatest influence on the rockburst level. Compare the first set of data with the second set in Table 3, MTS (No.1) is much larger than MTS (No.2) and UCS (No.1) is also larger than UCS (No.2). The results indicate that MTS and UCS may have a positive correlation between the intensity of rockburst. We can pay more attention to these two parameters of rock mass in practical engineering, which is of great practical significance for timely warning of rockburst.

5 Conclusions

A new rockburst prediction model based on SMOTE and Baysian-XGBoost is proposed in this study. In order to solve the problem of class imbalance in the dataset, SMOTE algorithm is introduced to synthesize samples of the minority class. The hyperparameters of XGBoost model are optimized by Bayesian optimization algorithm. In order to compare the effect of the proposed model, six other machine learning algorithms are also introduced for

comparison. The new model is trained based on the dataset collected, and the correctness of the proposed model is verified by an actual project.

In this study, the SMOTE has been developed to deal with the classification of imbalanced datasets, the amount of data for each category can be consistent by resampling the instances of the minority class. It avoids the problem of inaccurate prediction caused by unbalanced dataset samples. The results show that the accuracy of the proposed model is greatly improved after the dataset is processed.

There are many hyperparameters in XGBoost model, and the selection of hyperparameters has a significant influence on the prediction accuracy of the model. In this study, the Bayesian algorithm is introduced to optimize 6 important parameters in XGBoost, and the optimal values of 6 parameters are obtained. The prediction accuracy of XGBoost model after Bayesian optimization is improved from 82.70% to 84.14%, which proves that Bayesian optimization can improve the model prediction efficiency to a certain extent.

In the case study, the proposed SMOTE-Bayesian-XGBoost model is used to predict a real engineering example. Ten groups of rockburst data are selected and the prediction results of the proposed model are in good agreement with the actual results. The model proposed in this study extends the application of machine learning in rockburst prediction and can provide some guidance for practical engineering.

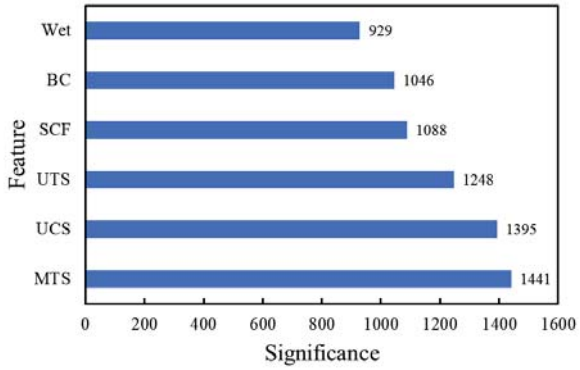

Figure 3. Significance ranking of influential factors of rockburst.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant No. 52079099).

References

Ahmad, M., Hu, J.L., Hadzima-Nyarko, M., Ahmad, F., Tang, X.W., Rahman, Z.U., Nawaz, A., and Abrar, M. (2021). Rockburst Hazard Prediction in Underground Projects Using Two Intelligent Classification Techniques: A Comparative Study. *Symmetry-Basel*, 13(4), 632.

Dong, L.J., Li, X.B., and Peng, K. (2013). Prediction of rockburst classification using Random Forest. *Transactions of Nonferrous Metals Society of China*, 23, 472-477.

Liang, W.Z., Zhao, G.Y., Wu, H., and Dai, B. (2019). Risk assessment of rockburst via an extended MABAC method under fuzzy environment. *Tunnelling and Underground Space Technology*, 83, 533-544.

Wang Y., Xu, Q., Cai, H.J., Liu, L., Xia, Y.C., and Wang, X.D. (2013). Rock burst prediction in deep shaft based on RBF-AR model. *Journal of Jilin University (Earth Science Edition)*, 43(6), 1943-1949.

Wojtecki, L., Iwaszenko, S., Apel, D.B., and Cichy, T. (2021). An Attempt to Use Machine Learning Algorithms to Estimate the Rockburst Hazard in Underground Excavations of Hard Coal Mine. *Energies*, 14(21), 6928.

Xue, Y.G., Li, Z.Q., Li, S.C., Qiu, D.H., Tao, Y.F., Wang, L., Yang, W.M., and Zhang, K. (2019). Prediction of rock burst in underground caverns based on rough set and extensible comprehensive evaluation. *Bulletin of Engineering Geology and the Environment*, 78(1), 417-429.

Xue, Y.G., Bai, C.H., Qiu, D.H., Kong, F.M., and Li, Z.Q. (2020). Predicting rockburst with database using particle swarm optimization and extreme learning machine. *Tunnelling and Underground Space Technology*, 98, 103287.

Yin, X., Liu, Q.S., Pan, Y.C., and Huang, X. (2021). A novel tree-based algorithm for real-time prediction of rockburst risk using field microseismic monitoring. *Environmental Earth Sciences*, 80(16), 504.

Zhou K.P., Lei, T., and H, J.H. (2013). RS-TOPSIS model of rockburst prediction in deep metal mines and its application. *Chinese Journal of Rock Mechanics and Engineering*, 32(2), 3705-3711.

Zhou, J., Li, X.B., and Mitri, H.S. (2016). Classification of rockburst in underground projects: comparison of ten supervised learning methods. *Journal of Computing in Civil Engineering*, 30(5), 4016003.

Zhang, Q.M., Wang, E.Y., Feng, X.J., Wang, C., Qiu, L.M., and Wang, H. (2021). Assessment of rockburst risk in deep mining: an improved comprehensive undex method. *Natural Resources Research*, 30(2), 1817-1834.

Zhang, M.C. (2022). Prediction of rockburst hazard based on particle swarm algorithm and neural network. *Neural Computing & Applications*, 34(4), 2649-2659.