Exploration of Mathematical Thinking Methods in Machine Learning Algorithms

Ruyun Liu

Shanghai Starriver Bilingual School, Shanghai, 201108, China

Abstract: *Artificial intelligence and statistics development have given rise to machine learning about data analysis. This emerging discipline is a key direction for researchers in the field of data analysis to explore. Moreover, machine learning refers to the acquisition of new experiences and knowledge by computers through inherent regular information, thus enhancing the intelligence of computers for the purpose of making decisions like humans. With the advancement of computer science, the exploration and application of machine learning has made great achievements. Additionally, studying the mathematical theory of machine learning plays an important role in the advancement of computers. Therefore, in this context, this paper explores the mathematical thinking related to machine learning, starting from several popular machine learning techniques.*

Keywords: *Machine learning; Algorithms; Mathematical Thinking Methods*

1. Introduction

In the era of artificial intelligence, data-based artificial intelligence technology is applied to many areas of society, and the development of artificial intelligence technology is increasingly dependent on the progress of applied mathematics disciplines, and mathematical theory is increasingly used in practical problem solving. In this paper, we take several major algorithms of machine learning as the research object to explore the mathematical thinking behind the algorithms.

But with the rapid development of artificial intelligence, GPS (Global Positioning System) and the leap forward in computer computing performance, the advantages of computers are more and more deeply into the field of thinking, so the computer will be used in the practical use of advanced mathematical theory, very effective in solving many practical problems

2. Types of machine learning algorithms

2.1 Supervised learning

The predicted target or output variable is known. These algorithms generate a function that maps the inputs to the output variables. Moreover, regression and classification algorithms fall into this category.[1] In regression, the output variable is continuous, while in classification, the output variable contains two or more discrete values. Some supervised learning algorithms include linear regression, logistic regression, random forests, support vector machines (SVM), decision trees, Naive Bayes, and neural networks.

2.2 Unsupervised learning

The target or output variables are unknown. These algorithms usually analyze the data and generate data clusters. Moreover, association, clustering, and dimensionality reduction algorithms fall into this category. K-means clustering, principal component analysis (PCA), Apriori algorithm, and others are unsupervised learning algorithms.[2]

2.3 Semi-supervised learning

Semi-supervised learning combines supervised and unsupervised learning methods, which use known data to train themselves and then label unknown data.[3]

2.4 Reinforcement learning

A machine or agent is trained to learn from a "trial and error" process. The machine learns from past decision experiences and uses its learning to predict the outcome of future decisions. Furthermore, examples of reinforcement learning algorithms are Q-Learning, Temporal Difference, and others.[4]

3. Typical mathematical thinking methods in machine learning algorithms

3.1 Linear regression

Linear regression is used to predict the outcome of continuous variables by fitting the best straight line on the data points. The best-fitted line defines the relationship between the dependent and independent variables. The algorithm attempts to find the line that best predicts the value of the target variable. Moreover, the best-fit line is achieved by minimizing the sum of squares of the differences between the data points and the regression line (As show in figure 1).[5]

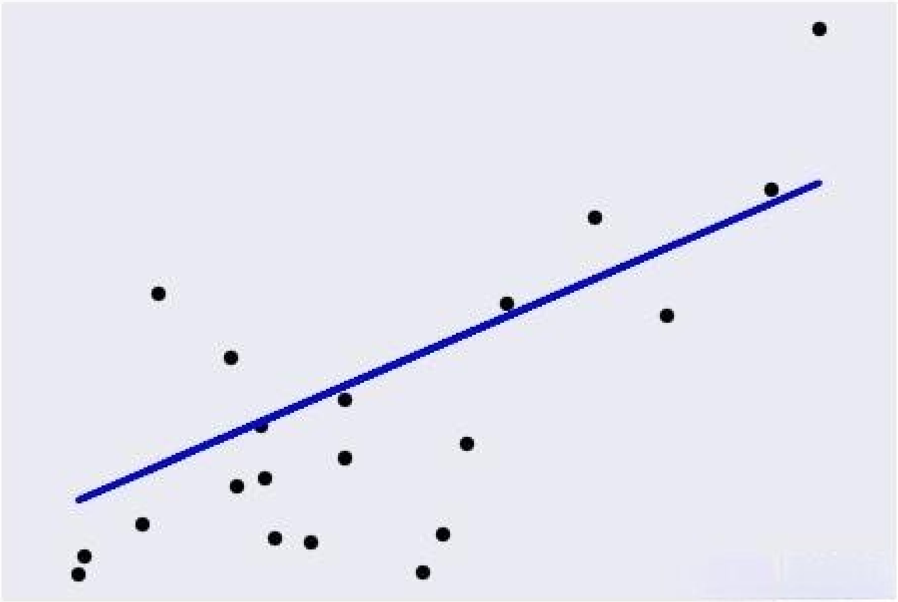


Figure 1. Linear regression

Formula: $Y = c + m_1X_1 + m_2X_2 + \ldots.. +m_nX_n$

Firstly, one of the most important mathematical thoughts for linear regression is that linear algebra is a very basic and widely used discipline in mathematics, which focuses on studying systems of linear equations and their solutions. In machine learning, for example, linear equations need to be solved to get the optimal linear model. Therefore, the knowledge of linear algebra is very important for applying linear regression.

Secondly, matrix theory is also an essential mathematical knowledge in machine learning. High-dimensional data representations, such as matrices of eigenvectors, are often used in machine learning. In this case, matrix theory can help us better understand the essential features of these data structures, thus improving our analytical skills.

Additionally, probability theory is also an essential part of machine learning. Probabilistic methods can be used in machine learning to describe various random phenomena, such as sample distribution, error rate, and so on.

3.2 Logistic regression

Logistic regression is a classification algorithm that estimates the outcome of a categorical variable based on the independent variable. It predicts the probability of an event occurring by fitting the data to a logistic function. The coefficients of the independent variables in the logistic function are optimized by maximizing the likelihood function. The decision boundary is optimized so that the cost function is minimized. The cost function can be minimized using the gradient descent method (As show in figure 2).[6]

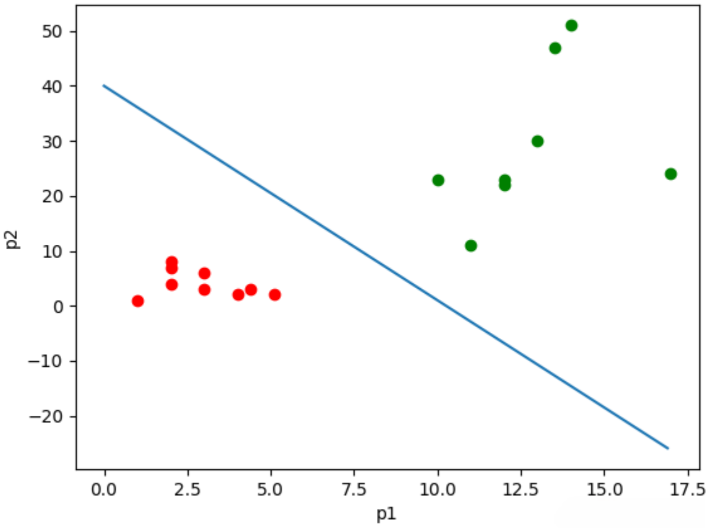


Figure 2. Logistic regression

In logistic regression, common mathematical thinking methods include linear models, nonlinear models, and mixed models. The linear model is the most basic one, which adds the input variables and the characteristic variables as two vectors to get the final result, respectively. However, the linear model does not fully satisfy the practical needs because the input variables may have higher-order structures or complex relationships between the characteristic variables. To better describe the relationship between input variables and characteristic variables, nonlinear models are also widely used in logistic regression.

3.3 Random forest

A random forest consists of multiple decision trees operating as a set. A whole consists of a set of models that are used to predict outcomes rather than a single model. In a random forest, each decision tree predicts a class outcome, and the class outcome with the most votes becomes the prediction of the random forest. Moreover, the correlation between decision trees should be minimal to make accurate predictions. There are two ways to ensure this: Bagging and feature selection. Bagging is a technique for selecting a random sample of observations from a dataset. Feature selection allows decision trees to be modeled on only a random subset of features, which prevents individual trees from using the same features for prediction (As show in figure 3).[7]

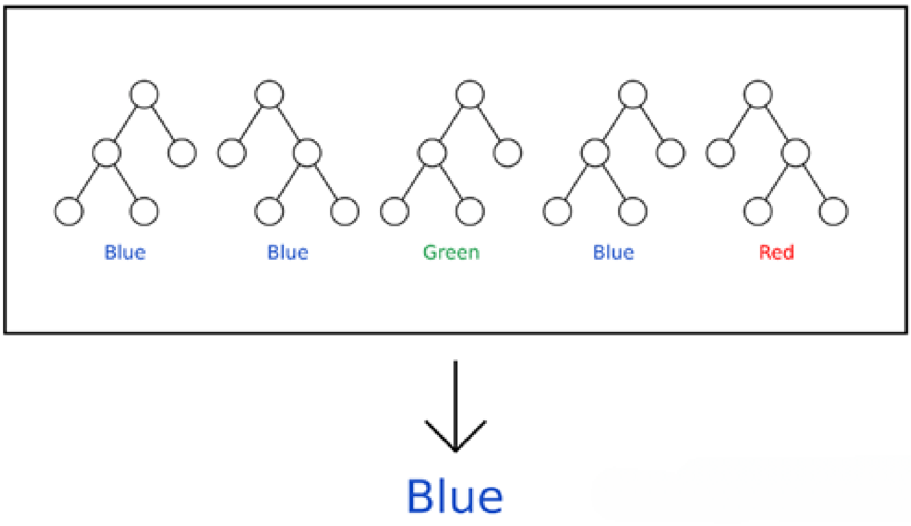


Figure 3. Random forest

The mathematical thinking method of the random forest consists of the following aspects: Firstly, it uses a method called the Bootstrap technique to generate the sample set. This method allows random samples to be selected from the original data as training data without changing the original data. Secondly, random forest uses a technique called Bagging to improve classification accuracy. This technique combines multiple decision trees together to form a more powerful integrated model. Finally, random forests also use a technique called Randomized Shuffling to avoid the problem of overfitting. This technique randomly rearranges the training data during the training process to reduce the repetition and dependency of the training data. In conclusion, the mathematical thinking method of random forest mainly focuses on how to use random sampling, cross-validation, and random recombination techniques to improve classification accuracy.

3.4 Support vector machine

SVM is also a supervised learning algorithm that can be used for classification and regression problems. Support vector machines try to find an optimal hyperplane in N-dimensional space (N refers to the number of features) to help classify different classes. It uses the Hinge loss function to find the optimal hyperplane by maximizing the margin distance between class observation values. The dimensionality of the hyperplane depends on the number of input features. If the number of features is N, the dimension of the hyperplane is N-1.[8]

Loss function: t→Target variable, w→Model parameter, x→Input variable

$$\ell(y) = \max\left(0, 1 + \max_{y \neq t} \mathbf{w}_y \mathbf{x} - \mathbf{w}_t \mathbf{x}\right)$$

Firstly, the most fundamental problem for the SVM training process is finding the appropriate hyperplane. To achieve this, researchers usually use the gradient descent method or the Newton iteration method to obtain the solution. Both methods are based on the basic principle of linear programming, which means that the optimal value of the objective function is approximated by continuously updating the parameters. Additionally, some other optimization methods can be used to solve hyperplane problems, such as random search and genetic algorithms, among others.

3.5 Decision tree

Decision trees are mainly used for classification problems, but they can also be used for regression. In this algorithm, the dataset is partitioned into two or more isomorphic sets based on the attributes that most efficiently partition the dataset. Moreover, one of the methods to select the attributes that will partition the dataset is to calculate the entropy and information gain. Entropy reflects the amount of impurities in the variables. The information gain is the sum of the entropy of the parent node minus the entropy of the child nodes. The attribute that provides the greatest information gain is selected for segmentation. Furthermore, the Gini index can also be used as an impurity criterion to segment the dataset.[9]

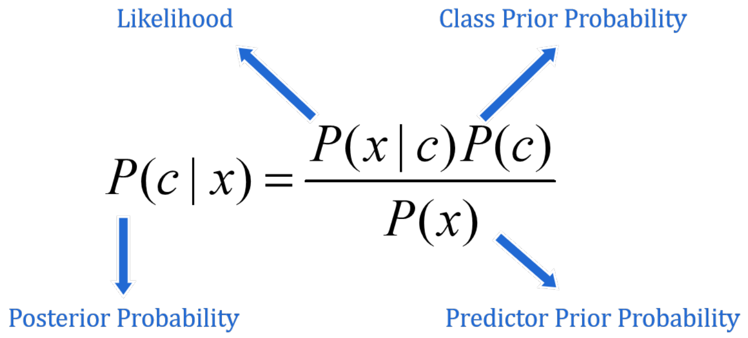
$$Entropy = \sum_{i=1}^{c} -p_i * \log_2\left(p_i\right)$$

Entropy: c→number of classes

In the decision tree, each node represents a condition or attribute, while each leaf represents a category. When input into the decision tree, it finds the final result along the optimal path. This result can be obtained by calculating the probability of all possible branches. Moreover, the most important part of the decision tree is the selection of appropriate features and rules. These rules can be a simple logical expressions or a complex function. The complexity of the rules and their impact on the classification needs to be taken into account when selecting the rules. In addition, it is necessary to consider how to handle outliers and other noise factors.

3.6 Naive Bayes

Naive Bayes is a classification method based on Bayes' theorem, which assumes that the independent variables are not directly or indirectly related to each other in any way but rather determine the correlation between them by observing the degree of similarity between them. Therefore, Bayes' theorem can be used to construct a frequency table to better predict the behavior of various predictors. The posterior probability of each category can be estimated by applying the naive Bayes equation, and the naive Bayesian classifier provides the most accurate results, thus giving the highest probability of these categories.[10]



$$P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}$$

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Feature extraction and representation of sample data are required in naive Bayes, and then the probability of each category is determined by calculating the conditional probability distribution. Among them, the conditional probability distribution is a method to describe the relationship between known attributes and unknown attributes in probabilistic form. Some mathematical thinking methods can improve its performance in naive Bayesimprove its performance in naive Bayes. For example, a weighted average method can be used for a dataset with n features to get the probability value of each category. Additionally, the entropy function can be used to optimize the parameter settings of the conditional probability distribution. These mathematical thinking methods can effectively improve the performance of the naive Bayes classifier.

3.7 Neural network

The artificial neural network is a complex nonlinear information processing system that consists of a large number of independent processing units with self-adaptive and intelligent features. It draws on the research results of modern neuroscience and aims to simulate the behavior of neural networks in the brain for better processing and memory of information (As show in figure 4).[11]

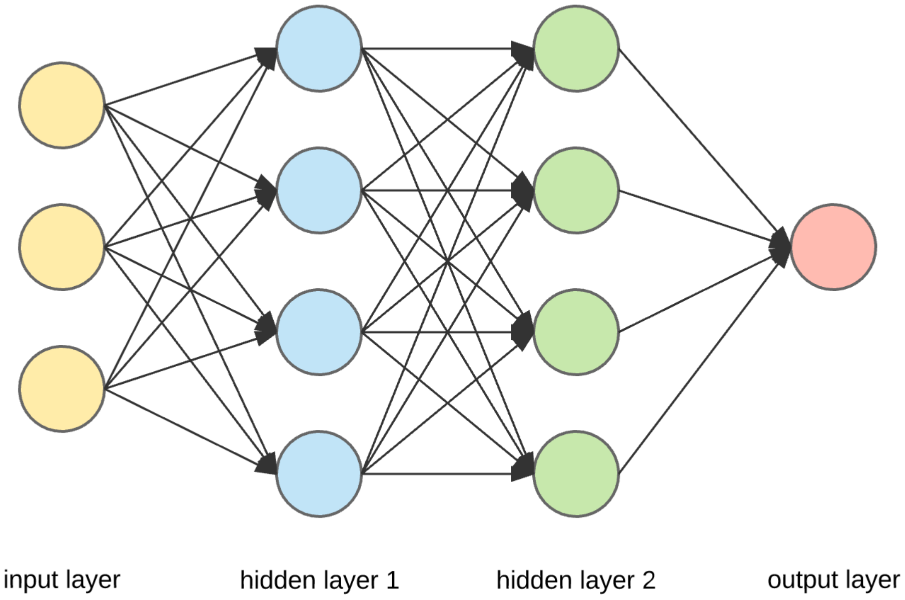
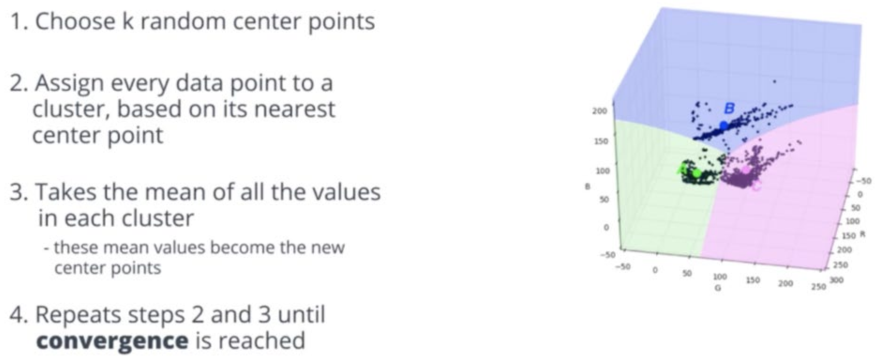


Figure 4. Neural network

In neural networks, the method of mathematical thinking is to achieve intelligence by modeling and training the data. Among them, the most commonly used model is the artificial neuron network (ANN). ANN is a calculator structure based on a simulated biological nervous system, which consists of an input layer, a hidden layer, and an output layer. Each node has a weight and an activation function, and these parameters determine the function of the node. When the input signal goes through multiple layers, the final result got is determined by the nodes in the output layer. In the design process of ANN, many factors need to be considered, such as choosing the appropriate activation function, adjusting the weight size, etc. Additionally, to improve the performance of ANN, the backpropagation method can be used for optimization. Besides ANN, there are other types of neural network models, such as recurrent neural networks (RNN), long short-term memory networks (LSTM), etc. These models are characterized by their ability to handle serial data or time series data.

3.8 K-means clustering

The K-means algorithm aims to construct a data cluster with higher accuracy by unsupervised learning, which contains K independent locations and is considered as the center of the cluster to better capture and compare different features for higher accuracy and higher efficiency. After a series of iterations, the closest cluster is finally identified, and the center of mass of the cluster is determined for better management and control of the data. This process needs to be performed repeatedly to ensure that the structure and characteristics of the clusters are always the same and that the centers of the clusters are always stable. The iterations are repeated to ensure that the algorithm will achieve the desired results and thus terminate the operation (As show in figure 5).[12]



1. Choose k random center points
2. Assign every data point to a cluster, based on its nearest center point
3. Takes the mean of all the values in each cluster
   - these mean values become the new center points
4. Repeats steps 2 and 3 until **convergence** is reached

Figure 5. The principle of K-means clustering

In the implementation of the K-means clustering algorithm, two problems need to be solved: How to choose the initial center point and how to update the position of each center point. For the first problem, the initial center points are usually determined by random selection, while for the second problem, a method called "discrete optimization" is used for updating. Specifically, at each iteration, the position of each center point is fine-tuned to bring it as close as possible to the nearest data point. This approach effectively avoids the situation of fixed center points in one position.

3.9 Principal component analysis (PCA)

PCA is an effective dimensionality reduction technique that maximizes the sample variance by replacing the original n features with fewer m features and combining them linearly. Furthermore, there is no correlation between these new features.[13]

The essence of principal component analysis is vector permutation. X is the original m-dimensional data (m features, n sample points), P is the eigenvector of the covariance matrix, and w represents the matrix consisting of the first k eigenvectors with the largest features.

$$X=\begin{pmatrix} X_1^{(1)} & X_1^{(2)} & \cdots & X_1^{(m)} \\ X_2^{(1)} & X_2^{(2)} & \cdots & X_2^{(m)} \\ \vdots & \vdots & \vdots & \vdots \\ X_n^{(1)} & X_n^{(2)} & \cdots & X_n^{(m)} \end{pmatrix}_{n\times m} \qquad P=\begin{pmatrix} P_1^{(1)} & P_1^{(2)} & \cdots & P_1^{(m)} \\ P_2^{(1)} & P_2^{(2)} & \cdots & P_2^{(m)} \\ \vdots & \vdots & \vdots & \vdots \\ P_m^{(1)} & P_m^{(2)} & \cdots & P_m^{(m)} \end{pmatrix}_{m\times m} \qquad w=\begin{pmatrix} w_1^{(1)} & w_1^{(2)} & \cdots & w_1^{(k)} \\ w_2^{(1)} & w_2^{(2)} & \cdots & w_2^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ w_m^{(1)} & w_m^{(2)} & \cdots & w_m^{(k)} \end{pmatrix}_{m\times k}$$

Convert the n-dimensional feature space to k-dimensional (this is the process of dimensionality reduction)

- **Principle:** Convert n-dimensional sample data to k-dimensional data

• **Operation:** Multiply an n-dimensional sample of the data set X with the matrix W to obtain k-dimensional data;

$$X_{(n\times m)}W_{(m\times k)}=X'_{(n\times k)}$$

In the PCA algorithm, the mathematical thinking method is the key to achieving effective dimensionality reduction. Specifically, the PCA algorithm scales the feature space of the dataset by calculating the PCA, thus making the points with higher similarity in the dataset be clustered together and separated from the data points with different properties. Therefore, how to choose the appropriate number of principal components and select the correct direction of principal components play a pivotal role in the effectiveness of the PCA algorithm. In order to solve this problem, PCA algorithms usually use a method called "Eigenvalue-based". This method is based on the eigenvalues of the sample data matrix, which are analyzed and processed to derive the direction of the principal components and the corresponding coefficients. Among them, determining the number of principal components is the most important step. In general, the number of principal components should be selected to retain the maximum amount of relevant information possible while avoiding excessive redundant information. Additionally, it is necessary to consider whether the interrelationships among the principal components are stable enough. If there is a large error or instability between principal components, then the number of principal components needs to be readjusted, or a more suitable direction of principal components needs to be selected.

4. Conclusion

In conclusion, the advancement of artificial intelligence technology is increasingly dependent on the progress of the discipline of applied mathematics, and the importance of mathematical learning to the cultivation of innovative talents in the era of artificial intelligence is self-evident. This paper combines the author's understanding of machine learning to form a discussion of the mathematical thinking methods related to machine learning. Therefore, this study can deepen the understanding of mathematical thinking methods and help use mathematical knowledge to solve practical problems.

References

[1] Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms[C]//2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). Ieee, 2016: 1310-1315.
[2] Celebi, M. Emre, and Kemal Aydin. Unsupervised learning algorithms[M]. Cham: Springer, 2016.
[3] Zhou X, Belkin M. Semi-supervised learning[M]//Academic Press Library in Signal Processing. Elsevier, 2014, 1: 1239-1269.
[4] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. MIT press, 2018.
[5] Maulud D, Abdulazeez A M. A review on linear regression comprehensive in machine learning[J]. Journal of Applied Science and Technology Trends, 2020, 1(4): 140-147.
[6] Stoltzfus J C. Logistic regression: a brief primer[J]. Academic emergency medicine, 2011, 18(10): 1099-1104.
[7] Biau G, Scornet E. A random forest guided tour[J]. Test, 2016, 25: 197-227.
[8] Suthaharan S, Suthaharan S. Machine learning models and algorithms for big data classification: thinking with examples for effective learning[J]. Support vector machine , 2016: 207-235.
[9] Song Y Y, Ying L U. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.
[10] Webb G I, Keogh E, Miikkulainen R. Naïve Bayes[J]. Encyclopedia of machine learning, 2010, 15: 713-714.
[11] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE transactions on neural networks and learning systems, 2021.
[12] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.
[13] Hasan B M S, Abdulazeez A M. A review of principal component analysis algorithm for dimensionality reduction[J]. Journal of Soft Computing and Data Mining, 2021, 2(1): 20-30.