Research on Quality Control of Ore Processing Based on Xgboost and Decision Tree Algorithm

Xiyuan Liang

College of Science, North China Institute of Science and Technology, Hebei, China, 065200

1685448643@qq.com

Abstract. Improving the quality of ore processing can directly or indirectly save non-renewable mineral resources and the energy required for processing, thus promoting energy saving and emission reduction and helping to achieve the goal of "double carbon". Ore processing is a complex process, in which voltage, water pressure and temperature, as important factors affecting ore processing, directly affect the quality of ore products. In order to explore the influence of temperature and other factors on ore quality and qualification rate, the corresponding model is established to solve the model of predicting four product quality indexes by system temperature, and fully consider the influence of other uncertain factors to predict the highest possible product quality index at similar temperature.

Keywords: XGBoost algorithm, decision tree algorithm, machine learning regression.

1. Introduction

In this paper, we build corresponding models to solve the model of predicting four product quality indexes by system temperature based on the underlying data, and fully consider the influence of other uncertainties to predict the product quality index with the highest probability at similar temperatures and fill in the table [1]. The specific process is to use XGBoost algorithm, decision tree algorithm, to calculate the quality indicators, system temperature, pass rate, etc [2]. The time dimension of the data was unified to remove outliers and missing values, with special attention to the problem that one ore processing takes two full hours, i.e., there is a two-hour time difference between the system temperature and the quality index of the same ore product. The system temperature, raw ore parameters, etc. were defined as independent variables and quality indicators as dependent variables, and the processed data were applied to linear regression, XGBoost algorithm, random forest algorithm and other models for comparison, and the best fit was obtained with the XGBoost algorithm with the best fit R^2=0.99. The decision tree algorithm model was used for prediction. On this basis, it is made to reach the corresponding temperature deduced through the developed quality index [3]. According to the model, the dependent independent variables are changed to obtain the corresponding solutions.

2. Model assumptions and notation

2.1. Assumptions

1) Assume that the voltage and water pressure during the ore processing are constant.
2) Assume that the origin of the ore has no effect on the quality of the ore.

2.2. Notations

Important notations used in this paper are listed in Table 1.

Table 1. Notations

Serial number	Symbols	Meaning
1	F	CART decision tree structure set
2	q	Sample mapping to the tree structure of the leaf nodes
3	T	Number of leaf nodes
4	w	Real fraction of leaf nodes
5	L	Error function term
6	Ω	Model Complexity Function Term
7	Obj	Scoring functions
8	S	Data sample set
9	Ci	Sample Type
10	si	Sample size of Ci
11	A	Property Indicators
12	v	Product Qualification Rate
13	D_i	Product qualification interval
14	n_j	Qualified product set
15	M	Number of regression trees
16	I(x)	Instruction functions
17	θ	Beyond Parameters

3. Model construction and solving

3.1. Preparation of the model

(1) Unification of data dimensions

By observing the data, we can find that the temperature of the data, the raw ore parameters and the time corresponding to the four indicators are different [4]. By observation, the values of raw ore parameters are more stable and less fluctuating throughout the day, so here we use the same raw ore parameters for all the time periods around the clock. The temperatures in Annex I are collected by minute, so the hourly temperature averages can be used to fit the four quality indicators collected by hour. As we know from the question, the ore is processed in two hours, so for the same processed ore there is a time difference between the system temperature and the quality indicators in the data, with a difference of two hours.

(2) Model theory

XGBoost is the abbreviation of "Extreme Gradient Boosting" (EGBoost), and XGBoost algorithm is a class of synthetic algorithm that combines basis functions and weights to form a good fit to data. Since XGBoost model has the advantages of strong generalization ability, high scalability, and fast computing speed, it has been popular in the fields of statistics, data mining, and machine learning since it was proposed in 2015 [5].

3.2. Model building

Set the independent variables X1=Temperature of system I, X2=Temperature of system II, X3=Mineral parameter 1, X4=Mineral parameter 2, X5=Mineral parameter 3, X6=Mineral parameter 4. parameter 3 (Mineral parameter 3), X6=Mineral parameter 4 (Mineral parameter 3). Set the dependent variable Y1=Index A (index A), Y2=Index B (index B), Y3=Index C (index C), Y4=Index D (index D) [6].

The following four models were used here: linear regression, random forest, XGBoost, and GBDT. the input processed data were operated to observe the fitting effect of different models, and the best fit of the four models corresponding to the four indicators was considered comprehensively to select the model with the best fit.

Among them, the XGBoost model was focused on and considered. In XGBoost for a data set containing n m-dimensional entries, the XGBoost model can be expressed as.

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F(i = 1,2,\dots n) \tag{1}$$

Is the set of CART decision tree structures, q is the tree structure of samples mapped to leaf nodes, T is the number of leaf nodes, and w is the real number fraction of leaf nodes. When constructing the XGBoost model, it is necessary to find the optimal parameters according to the principle of minimizing the objective function to build the optimal model. The objective function of the XGBoost model can be divided into the error function term L and the model complexity function term Ω. The objective function can be written as.

$$Obj = L + \Omega \tag{2}$$

$$L = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{3}$$

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

When training the model optimally using the training data, it is necessary to keep the original model unchanged and add a new function f to the model so that the objective function is reduced as much as possible by the following process.

$$\widehat{y_i}^{(0)} = 0 \tag{5}$$

$$\widehat{y_i}^{(1)} = \widehat{y_i}^{(0)} + f_1(x_i) \tag{6}$$

$$\widehat{y_i}^{(2)} = \widehat{y_i}^{(1)} + f_2(x_i) \tag{7}$$

$$\widehat{y_i}^{(t)} = \widehat{y_i}^{(t-1)} + f_t(x_i) \tag{8}$$

At this point the objective function is expressed as.

$$Obj^{(t)} = \sum_{i=1}^{n}(y_i - (\widehat{y_i}^{(t-1)} + f_i(x_i)))^2 + \Omega \tag{9}$$

In the XGBoost algorithm, in order to quickly find the parameter that minimizes the objective function, a second-order Taylor expansion is performed on the objective function to obtain the approximate objective function.

$$Obj^{(t)} \approx \sum_{i=1}^{n}[(y_i - \widehat{y_i}^{(t-1)})^2 + 2(y_i - \widehat{y_i}^{(t-1)})f_i(x_i) - h_i f_t^2(x_i)] + \Omega \tag{10}$$

When the constant term is removed, it is known that the objective function is only related to the first- and second-order derivatives of the error function. At this point, the objective function is expressed as.

$$Obj^{(t)} \approx \sum_{i=1}^{n}[g_i w_{q(x_i)} + \frac{1}{2}h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\sum_{j=1}^{T} w_j^2 \tag{11}$$

$$= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T$$

If the structural part q of the tree is known, the objective function can be used to find the optimal Wj and obtain the optimal objective function value. Its essence can be categorized as a quadratic minimax solution problem. The solution is.

$$w_j^* = \frac{-\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{12}$$

Obj is a scoring function that can be used to evaluate the model, and the smaller the Obj value, the better the model. By recursively calling the above tree building method, we can obtain a large number

of regression tree structures, and use Obj to search for the best tree structure and put it into the existing model, so as to build the optimal XGBoost model.

3.3. Solving of the model

For the dependent variable Y1, the goodness-of-fit indices of these four models are obtained. Y1 model goodness of fit is shown in Table 2.

Table 2. Y1, Y2, Y3, Y4 model goodness of fit

Linear regression (Y1)	0.173041
Random Forest (Y1)	0.902857
XGBoost (Y1)	0.991978
GBDT (Y1)	0.785472
Linear regression (Y2)	0.151663
Random Forest (Y2)	0.895885
XGBoost (Y2)	0.983886
GBDT (Y2)	0.783620
Linear regression (Y3)	0.381973
Random Forest (Y3)	0.926962
XGBoost (Y3)	0.994905
GBDT (Y3)	0.850677
Linear regression (Y4)	0.076967
Random Forest (Y4)	0.887423
XGBoost (Y4)	0.996342
GBDT (Y4)	0.799177

When the goodness of fit of the four models under the four dependent variables is observed comprehensively, XGBoost has the highest goodness of fit, so the model chooses XGBoost for the subsequent prediction analysis. Get X Collection.

X={1404.89, 859.77, 52.75, 96.87, 46.61, 22.91 1151.75, 859.77, 52.75, 96.87, 46.61, 22.91}

Bring into the model for training and get the result as shown in

Table 3. Results

Time	System I set temperature	System II set temperature	Indicator A	Indicator B	Indicator C	Indicator D
2022-01-23	1404.89	859.77	80.32	23.06	11.52	17.48
2022-01-23	1151.75	859.77	79.16	23.02	12.13	15.56

3.4. Model after redefining the dependent and independent variables

Decision tree is a tree structure, which starts from the root node, tests the data samples, and divides the data samples into different subsets of data samples according to different results, and each subset of data samples constitutes - a child node [7]. The purpose of constructing a decision tree is to find out the relationship between attributes and categories, and use it to predict the categories of records of unknown categories in the future.

There are two types of decision trees: categorical trees for discrete variables and regression trees for continuous variables. A decision tree is a tree in which the root node is the entire data set space and each branch node is a test for a single variable that splits the data set space into two or more blocks. Each leaf node is a record belonging to a single category. The process of constructing a decision tree is as follows [8]: First, the initial split is found. The entire training set is used as the set to generate the decision tree, and each record in the training set must already be classified. Decide which attribute field (Field) is the best indicator for classification at the moment. The general practice

is to exhaust all attribute fields, quantify the goodness of each attribute field split, and compute the best one [1].

The criteria for calculating attribute domain splits are not quite the same for different algorithms. The first step is repeated until all the records in each leaf node belong to the same class, growing to a complete tree.

3.4.1 Model building

The model construction is still performed using the data obtained after processing. The independent variables X1=index A (index A), X2=index B (index B), X3=index C (index C), X4=index D (index D), X 5=mineral parameter 1 (Mineral parameter 1), X6=mineral parameter 2 (Mineral parameter 1), X7=mineral parameter 3 (Mineral parameter 3), X8=Mineral parameter 4 (Mineral parameter 3). Set the dependent variable Y1=Temperature of system I and Y2=Temperature of system II.

The model toolbox that comes with SPSSPRO was used because the fit obtained using linear regression in the first problem was not satisfactory, so a new model decision tree was tried here for the fit. The following four models were used in this problem: decision tree, random forest, XGBoost, and GBDT, and the data were input to perform the operation, observe the fitting effect of different models, consider the four models corresponding to the four indicators, and select the model with the best fitting degree [9].

The focus was on the decision tree model. The ID3 algorithm exists in the decision tree model, which uses a criterion called gain to select the attributes to be tested, and is based on the concept of entropy in information theory. Let S be the set of data samples for this problem, and assume that the class

Suppose that the attributes labeled as raw mineral parameters and quality have m different values, and m different classes Ci (i=1, ..., m) are defined. Let si be the number of samples in class Ci. The desired information required to classify a given sample is given by the following equation:

$$I(S_1, S_2, \ldots, S_m) = -\sum_{i=1}^{m} P_i \log_2(P_i) \tag{13}$$

Where pi=si/s is the probability that any sample belongs to Ci. Because the information is binary encoded, the logarithm function is based on 2.

For the attribute index, let there exist attribute A. with v different values {a1, a2,..., av}. The attribute A can be used to partition S into v subsets {S1, S2,..., Sv}, where the samples in Sj have the same value aj (j=1, 2,..., v) on the attribute. Let sij be the number of samples of class Ci in subset Sj. The entropy or information expectation of partitioning into subsets by this attribute A is given by the following equation [10]:

$$E(A) = \sum_{j=1}((S_{ij} + S_{2j} + \ldots + S_{mj})/s) * I(S_{1j} + S_{2j} + \ldots + S_{mj}) \tag{14}$$

The smaller the entropy value, the higher the purity of subset partitioning. For a given subset Sj, its information expectation is:

$$I(S_{1j} + S_{2j} + \ldots + S_{mj}) = -\sum_{i=1}^{m} P_{ij} \log_2(P_{ij}) \tag{15}$$

The information gain that will be obtained by branching on attribute A is.

$$Gain(A) = I(S_1, S_2, \ldots, S_m) - E(A) \tag{16}$$

The gain ratio is used as the attribute selection criterion for selecting attributes at each level of the decision tree using C4.5 algorithm.

$$SplitInformation(A, S) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{17}$$

$$GainRatio(A, S) = \frac{Gain(S,A)}{SplitInformation(S,A)} \tag{18}$$

For the SLIQ, SPRINT, and PUBLIC algorithms, gini index is used instead of Information as the attribute selection criterion. The gini index has better performance than information and is easy to compute. For a dataset S containing n classes, gini(S) is defined as.

$$\text{gini}(S)=1-\sum p_j*\text{p}_\text{j} \tag{19}$$

Where, pj is the frequency of the jth class of data in S. The smaller the gini, the larger the Information Gain, and the corresponding attribute selection can be made [3].

The goodness-of-fit indices of these four models are then obtained based on the dependent variables.

4. Conclusion

In order to improve the quality of ore processing and product qualification rate, after fully considering the influence of four important factors such as temperature and voltage on the quality and qualification rate of ore, we constructed a model to predict the corresponding four indicators based on the system set temperature and a model to predict the corresponding system set temperature based on the four indicators by using XGBoost algorithm and decision tree algorithm. Through the two models in this paper, the purpose of providing the reference data of the corresponding influencing factors for the rational use of energy and improving the product quality of ore production line is achieved, which helps to save energy and reduce emissions and achieve the goal of double carbon.

References

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11) [Online Application Software]. Retrieved from https://www.spsspro.com.2022-05-03.

[2] X.B. Yang, J. Zhang. Decision tree algorithm and its core technology [J]. Computer Technology and Development, 2007(01): 43-45.

[3] Lin Sen, Guo Guijhen, Liu Beibei. A gradient boosted decision tree (GBDT) algorithm-based model for assessing house collapse in southern floods [J]. Disaster Science, 2020, 35(01): 109-112+130.

[4] Pan Shaowei, Wang Chaoyang, Zhang Yun, Cai Wenbin. Lithology identification based on long and short-term memory neural network complementary logging curves and hybrid optimized XGBoost [J]. Journal of China University of Petroleum (Natural Science Edition), 2022, 46(03).

[5] Shen Yaqian, Huang Zhijia, Zhou Tao. Thermal comfort prediction model based on XGBoost classification algorithm [J]. Journal of Anhui University of Technology (Natural Science Edition), 2022, 39(01).

[6] Wang F, Zhang WJ, Liu J, Chen LF, Li ZN. Flight delay prediction model based on classification regression decision tree algorithm [J]. Journal of Civil Aviation University of China, 2022, 40(03).

[7] Xie XIN, Zhang HYANYONG, Yang JILIN. Improved ID3 Decision Tree Algorithm Induced by Neighborhood Equivalence Relation. Computer Application Research. 2022, 39(01).

[8] Zhang Lulu, Zhang Junhua, Cheng Nianfu. Automatic classification of logging lithology based on decision tree algorithm and its application. The Chinese Geophysics 2012 The Chinese Geophysics 2012 Proceedings of the Chinese Geophysical Society.

[9] Li Hongbo, Luo Pingya, Bai Yang, Li Daoxiong, Chang Shuang, Liu Xinguo. Overview of machine learning algorithms and their application in drilling engineering. Xinjiang Oil and Gas. 2022, 18(01).

[10] Zhu, S. & Li, H. Yao. (2022). A theft crime prediction model based on the joint PCA-XGBoost algorithm. Computer Applications and Software (05), 98-103.