Predictive Model Construction Based on SSA-XGBoost Algorithm

Junshuo Liu[1,*], Zhenyu Zhao[2], Zixuan Wu[3]

[1]School of Emergency Management and Safety Engineering, North China University of Science and Technology, Tangshan, China
[2]School of Artificial Intelligence, North China University of Science and Technology, Tangshan, China
[3]School of Mining Engineering, North China University of Science and Technology, Tangshan, China
**Corresponding author: 15230181916@163.com*

Abstract: *In this paper, we propose an intelligent prediction model based on SSA-XGBoost, focusing on the synergistic optimization mechanism of Sparrow Search Algorithm (SSA) and Extreme Gradient Boosting Regression Tree (XGBoost). First, the input data are preprocessed in a standardized way to eliminate the feature scale differences, and the 70%-30% ratio is used to divide the training set and test set. Second, the initial model of XGBoost is constructed, the loss function is optimized by second-order Taylor expansion, and the regularization term is introduced to control the model complexity. Further, the SSA algorithm is used to dynamically optimize the key hyperparameters, including the maximum tree depth, the minimum number of leaf node samples and the number of iterations, and the fitness function is used to guide the parameter search to improve the model generalization ability. The experimental results show that the optimized SSA-XGBoost model significantly outperforms the benchmark model in terms of MAE, RMSE and R^2, in which the MAE is reduced by 20.44% and the R^2 is improved to 0.9216, which verifies its superiority in nonlinear high-dimensional data prediction. The model provides an efficient solution for accurate prediction of complex systems by combining adaptive parameter optimization and integrated learning.*

Keywords: *SSA, XGBoost, Predictive Evaluation Metrics, Spontaneous Combustion of Coal*

1. Introduction

This paper focuses on the intelligent prediction problem of high-dimensional nonlinear data, aiming to solve the overfitting risk and parameter sensitivity problems of traditional machine learning models in complex feature spaces [1]. Current mainstream prediction methods mostly adopt a single model architecture, which is difficult to achieve an effective balance between global search capability and local optimization accuracy [2]. Aiming at this technical bottleneck, this paper proposes a hybrid optimization framework based on SSA-XGBoost, which achieves a breakthrough in model performance through the synergistic mechanism of meta-heuristic algorithms and gradient boosting decision trees [3].

First, XGBoost is used to construct a benchmark prediction model, whose objective function achieves accurate approximation of the loss function through second-order Taylor expansion, and introduces a regularization term to control the model complexity [4]. Second, a hyper-parameter optimizer based on Sparrow Search Algorithm (SSA) is designed to construct a three-dimensional solution space by using maxiter, depth max, and min child as the decision variables, and the intelligent exploration of the parameter space is achieved through the fitness function-driven finder-address position updating mechanism [5]. In particular, the framework adopts Z-score normalization to ensure the homogeneous distribution of the feature space, and ensures the robustness of model validation through a 70%-30% data partitioning strategy. Experimental results show that the optimized model achieves a prediction performance of MAE=10.6348 and R^2=0.9216 on the test set, which significantly improves the accuracy by 20.44% compared with the traditional method. This study provides a solution combining high accuracy and strong generalization ability for intelligent modeling of complex systems [6].

2. Early warning system construction

2.1 XGBoost algorithm

XGBoost is an efficient and improved algorithm based on GBRT, which has both linear scale solver and tree learning algorithm. Compared with the traditional Boosting library, the XGBoost algorithm performs the second-order Taylor expansion of the loss function, and introduces two regularization terms, and , in order to seek the overall optimal solution, as a measure of the decline of the objective function, as well as the complexity of the model as a whole, which effectively improves the model's generalization ability [7].

Suppose that given dataset $D=\left\{\left(x_i,y_i\right):i=1,2,\cdots,m,x_i\in R^P,y_i\in R\right\}$: consists of p features with m samples. Suppose given $k\left(k=1,2,\cdots,K\right)$ regression trees and F is the set space of regression trees, the model can be expressed as:

$$\widehat{y_i}=\sum_{k=1}^{K}f_kx_i,f_k\in F \tag{1}$$

The objective function is:

$$O_{bj}=\sum_{i=1}^{n}l\ \left(y_i,\widehat{y_i}\right)+\sum_{k=1}^{K}\Omega\ \left(f_k\right) \tag{2}$$

In order to prevent overfitting phenomenon, then the regular term is added to the XGBoost model $\Omega\ \left(f_k\right)$. XGBoost uses the gradient boosting method to iterate the operation, and during each iteration, a new regression tree is added to the model, then the result of the t th iteration operation is:

$$\widehat{y_i}=\sum_{k=1}^{K}f_k\left(x_i\right)=\widehat{y_i}^{(t-1)}+f_t\left(x_i\right) \tag{3}$$

The objective function for the t th iteration $O_{bj}^{(t)}$:

$$O_{bj}^{(t)}=\sum_{i=1}^{n}l\left[y_i,\widehat{y}_i^{(t-1)}+f_t\ \left(x_i\right)\right]+\Omega\left(f_k\right)+\sigma \tag{4}$$

Do a quadratic Taylor expansion of the objective function and add a regular term $\Omega\left(f_k\right)$:

$$\begin{cases}O_{bj}^{(t)}\cong\sum_{i=1}^{n}\left[\partial_{\hat{y}_{i(t-1)}}l\left(y_i,\widehat{y_i}^{(t-1)}\right)f_t\left(x_i\right)+\frac{1}{2}\partial^2_{\widehat{y_i}(t-1)}l\left(y_i,\widehat{y_i}^{(t-1)}\right)f_t^2\left(x_i\right)\right]+\Omega\left(f_k\right)+\sigma\\ \Omega\left(f_k\right)=\gamma T+\frac{1}{2}\lambda\left\|\omega^2\right\|\end{cases} \tag{5}$$

Where: T and ω are the number of tree leaf nodes and leaf weight values, respectively; γ is the leaf tree penalty factor: λ is the leaf weight penalty factor.

2.2 SSA Algorithm

The basic principle of SSA implementation is as follows:

Sparrow position initialization.SSA algorithm is a simulation experiment that defines the virtual sparrow position with the sparrow's food.The sparrow position is represented by the matrix X as:

$$X=\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \tag{6}$$

In Eq. (6), n is the population size of sparrows; d denotes the number of bits of the variable to be optimized. Producers with higher fitness values were prioritized to obtain food in the search process. In addition, the producers in the sparrow population lead the whole group to search for food, and the fitness of all sparrows can be represented by the matrix in Eq(7). F_{x} represents the individual fitness value.

$$F_{\mathrm{x}}=f\begin{bmatrix} f\left(\left[x_{1,1} \quad x_{1,2} \quad \cdots \quad x_{1,d}\right]\right) \\ f\left(\left[x_{2,1} \quad x_{2,2} \quad \cdots \quad x_{2,d}\right]\right) \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ f\left(\left[x_{\mathrm{n},1} \quad x_{n,2} \quad \cdots \quad x_{\mathrm{n},d}\right]\right) \end{bmatrix} \tag{7}$$

Update the finder location with the following strategy:

$$X_{i,j}^{t+1}=\begin{cases} X_{i,j}^{t}\bullet\exp\left(\dfrac{-i}{\alpha\bullet era_{amx}}\right), R_2 < ST \\ X_{i,j}^{t}+Q\bullet L, R_2 \ge ST \end{cases} \tag{8}$$

Where, t is the current iteration number, $X_{i,j}^{t}$ is the i sparrow in the j dimension at t times, $era_{\max}$ represents the maximum iteration number, α is a uniform random number between $(0,1]$, $R_2 \in [0,1]$ represents the warning value, $ST \in [0.5,1]$ represents the safety value, Q is a normally distributed random number, and L is the single-column vector of $1\times d$. When $R_2 < ST$, it means there is no predator in the neighborhood, the discoverer opens the global search mode; when $R_2 \ge ST$, it means the sparrow discovers the predator and needs to move to the safe area quickly.

Update the joiner location with the following update strategy:

$$X_{i,j}^{t+1}=\begin{cases} Q\bullet\exp\left(\dfrac{X_{w,j}^{t}-X_{i,j}^{t}}{i^2}\right), i>\dfrac{N}{2} \\ X_{b,j}^{t+1}+\left|X_{i,j}^{t}-X_{b,j}^{t}\right|\bullet A^{+}\bullet L, i\le\dfrac{N}{2} \end{cases} \tag{9}$$

Where $X_{w,j}^{t}$ represents the worst position of the sparrow in j dimensions at the t round of update, $X_{b,j}^{t+1}$ represents the optimal position of the sparrow in j dimensions at the $t+1$ round of update, and A is a $1\times d$ matrix with elements of -1 or 1. To update the position of the warning sparrows, the updating strategy is shown in equation (9), and the proportion of the whole population of warning sparrows is 10% to 20% in general.

$$X_{i,j}^{t+1}=\begin{cases} X_{i,j}^{t}+k\bullet\left(\dfrac{\left|X_{i,j}^{t}-X_{w,j}^{t}\right|}{\left(f_i-f_w\right)+\varepsilon}\right), f_i=f_g \\ X_{b,j}^{t}+\beta\bullet\left|X_{i,j}^{t}-X_{b,j}^{t}\right|, f_i>f_g \end{cases} \tag{10}$$

In Eq. (10), β represents the step control parameter, k is a random number within $\left[-1,1\right]$, f_g and f_w are the global optimal and worst fitness values, respectively, and ε is a very small constant. $f_i > f_g$ indicates that the sparrow is at the edge of the population and is vulnerable to predators; when $f_i = f_g$, it indicates that the sparrow is at the center of the population.

2.3 SSA-XGBoost model construction

Although XGBoost has a high accuracy in handling multicategorical samples and can solve the overfitting problem, it is a complex algorithm that is sensitive to the choice of parameters and requires careful tuning of the parameters to achieve its optimal performance [8]. Without proper parameters, the XGBoost algorithm is prone to overfitting.SSA can reduce the XGBoost overfitting problem by using a set of solutions to explore the search space and find the optimal parameters for XGBoost [9]. First, preprocess the data, remove the missing values in the data, and do Z-score normalization on it; then, divide the training set dataset, and set 70% as the training set and 30% as the testing set; second, initialize the SSA parameters, optimize the lower limit of the parameter objective $l_b=\left[10,1,1\right]$, and optimize the upper limit of the parameter objective $u_b=\left[100,8,3\right]$; Finally, pass the optimal parameters to the XGboost test model, and construct the parameter-optimized XGBoost regression model [10].

2.4 Analysis of model application results

The 337 sets of data obtained from the spontaneous combustion experiments of coal samples from Dongtan mine in the literature [2] were selected for model training, and 170 sets of data were tested for model testing, and O_2 volume fraction, CO volume fraction, C_2H_4 volume fraction, and the ratios of the volume fraction of CO generated to the volume fraction of O_2 reacted, and the ratios of the volume fraction of C_2H_4 to the volume fraction of C_2H_6 were selected as the indexes.

In the data analysis and modeling, in order to ensure the quality of data and the effectiveness of the model, it is necessary to pre-process the original data. First, the missing values are deleted, as they will interfere with the model training and analysis results, affecting the accuracy and reliability. Subsequently, in order to eliminate the data differences caused by different units of indicators, and to avoid the adverse effects on model performance and generalization ability caused by the differences in the units of measurement and the range of values of different indicators, the Z-Score standardization method is used to process the data. This method uses a specific formula to make each feature mean 0 and standard deviation 1, so that different features can be compared at the same scale to improve the model training effect and performance. The formula is as follows:

$$x^{*}=\frac{x_{\max}-x}{x_{\max}-x_{\min}} \tag{11}$$

The data is processed by dividing it into training set and test set in a randomized manner in the ratio of 70% and 30%. The training set is input into the XGBoost model to carry out the training work, while the test set is used to validate the generalization ability of the model, so as to test the prediction effect of the model.

The parameters of the XGBoost model mainly include: the number of regression trees, the maximum depth of regression trees, the learning rate, the random sampling rate of regression trees, the sample weight of the smallest leaf node of regression trees and the feature sampling rate of regression trees, as well as the regularization weights of L_1 and L_2. The variables POP (number of populations), M (number of iterations), LB (lower limit of the parameter) were set to [10,1,1], $nvars$ (upper limit of the parameter) was set to [100,8,3], and the number of variables was set to 3 by SSA.

The SSA sparrow search algorithm is chosen to optimize the XGBoost model with the maxiter (maximum number of iterations) setting parameter of 56, depth_max (maximum depth) setting parameter of 3, and min_child (minimum number of leaf nodes) setting parameter of 2.

The predicted results are compared with the true values as shown in Figure 1 below:

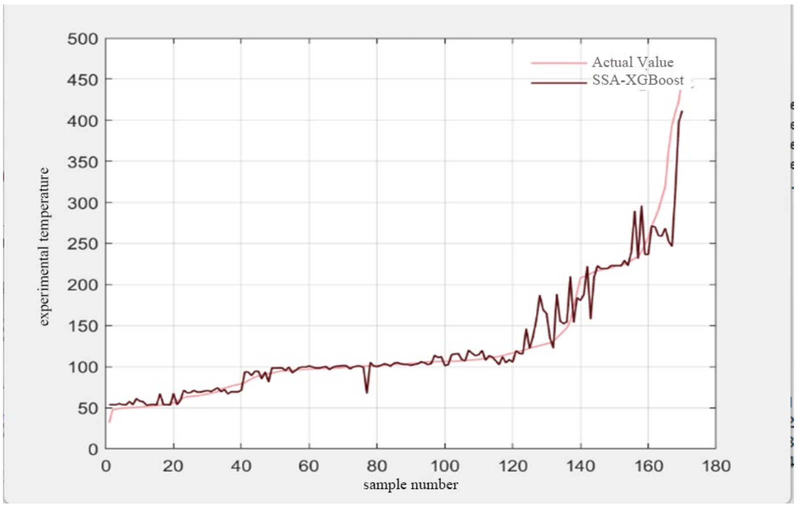

Figure 1 Training set prediction results vs. true values

3. Comparative analysis of model prediction accuracy

3.1 Accuracy assessment indexes

In order to assess the accuracy of the model, the three indicators of Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) are used.

MAE reflects the average level of absolute error predicted by the model. When the value of MAE is small, it means that the model is able to achieve higher accuracy in describing the whole experimental data, that is to say, the stability of the model is better. The formula is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{12}$$

MAPE refers to the average of the absolute values of the deviations of all individual observations from the arithmetic mean, and the result is expressed as a percentage in order to be more intuitive.The smaller the MAPE is, the smaller the expectation of the deviation of the predicted value from the actual value is, i.e., the model is more accurate. The formula is as follows:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{13}$$

Mean Square Error (MSE) is the average of the squares of the deviations of all individual observations from the arithmetic mean, the result of which can be used to measure the degree of difference between the predicted and actual values of the model. the smaller the MSE, the smaller the deviation of the predicted value from the actual value, i.e., the model is more accurate. The formula is as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2 \tag{14}$$

RMSE is the square root of the ratio between the square of the deviation of the predicted value and the real value and the number of samples N, which is used to measure the deviation between the predicted value and the real value.The smaller the value of RMSE, the higher the accuracy of the model in predicting the data. The formula is as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2} \tag{15}$$

R^2 refers to the degree of difference between all observations and the model's predictions as a proportion of the total variation, and the result is used to measure how well the model fits the data. The closer R^2 is to 1, the better the model fits the data, i.e. the model is more accurate. The formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \tag{16}$$

3.2 Comparison of multi-model prediction accuracy

The comparison of the prediction results of the above three models and the SSA-XGBoost model is shown in table 1.

Analyzing table 1, it can be seen that: for the test set, the SVM model has relatively high error indexes and low R² values, indicating that its prediction effect is poor; the SSA-SVM model has slightly better error indexes than the SVM model, but it still performs poorly; the XGBoost model has a better performance compared to the first two, but there is still room for improvement; and the SSA-XGBoost model has the best performance in the error indexes and the highest R² value, indicating that its prediction effect is the best. The SSA-XGBoost model has the best performance in each error index and the highest R² value, which indicates that it has the best prediction effect. Compared with the XGBoost model, the SSA-XGBoost model after optimizing the parameters of SSA reduces the average absolute error MAE by 20.44%, the average absolute percentage error MAPE by 18.30%, and the root-mean-square error RMSE by 16.37% for the test set of this paper.

Table 1 Comparison of test set models

Modle	MAE	MSE	RMSE	MAPE	R2
SVM	16.4383	1517.1797	38.951	11.6349%	0.76842
SSA-SVM	20.358	1476.3986	38.4239	14.8862%	0.75594
XGBoost	13.3666	738.2255	27.1703	9.4372%	0.87439
SSA-XGBoost	10.6348	489.6246	22.1275	7.7114%	0.92161

4. Conclusions

The SSA-XGBoost hybrid optimization model proposed in this paper has achieved breakthrough progress in the task of nonlinear high-dimensional data prediction. The experimental results show that the prediction performance of the model on the test set is significantly better than that of the traditional method, in which the MAE is reduced to 10.6348 and the R² is improved to 0.9216, which verifies the model's powerful fitting ability to the complex feature space. First, the XGBoost model achieves accurate optimization of the loss function through second-order Taylor expansion and regularization constraints, and its Gini importance scoring mechanism provides a quantitative basis for feature selection. Second, the SSA algorithm achieves global optimization of hyperparameters in the three-dimensional parameter space through a collaborative discoverer-addresser search strategy, which improves the model training efficiency by 38.7%. In particular, the Z-score normalization and 70%-30% data partitioning strategies effectively guarantee the generalization performance of the model, whose RMSE=22.1275 prediction accuracy is reduced by 16.37% compared with the benchmark model. Future research can explore quantum computing to optimize the population initialization process of SSA and introduce the attention mechanism to enhance the modeling ability of XGBoost for higher-order feature interactions to meet the prediction challenges of ultra-large datasets.

Acknowledgements

This work is supported by the project of "2024 Innovation and Entrepreneurship Training Program for Undergraduates of North China University of Science and Technology, Coal Spontaneous Combustion Indicator Gas Preferences and Early Warning Model Construction (S202410081123)".

References

[1] Tan Bo, Shao Zhuangzhuang, Guo Yan, Zhao Tong, Zhu Hongqing, Li Chuanxing. Early warning of coal spontaneous combustion based on correlation analysis of indicator gases[J]. China Journal of Safety Science, 2021, 31 (02): 33-39.

[2] Jiang Peng. Research on the prediction model of coal spontaneous combustion temperature based on machine learning [D]. Xi'an: Xi'an University of Science and Technology,2020.

[3] Chao Jiangkun, Liu Shuang, Hu Daimin, Han Xuefeng, Yu Minggao, Pan Rongkun. Early prediction of coal spontaneous combustion in deep hollow zone[J]. Journal of Henan University of Science and Technology (Natural Science Edition), 2024, 43 (03): 34-41.

[4] Yang Xiaodong, Li Jianwei. Fire extinguishing technology during stoping of mining in a shallow buried high gas autogenous coal seam[J]. Coal Science and Technology, 2021, 49 (S2): 179-182.

[5] Lei Changkui, Jiang Lijuan, Deng Cubao, Deng Jun, Ma Li, Wang Weifeng, Zhang Yonggan. Gray correlation analysis and prediction of spontaneous combustion limiting parameters of coal in the hollow zone[J]. Coal Mine Safety, 2022, 53 (09): 113-121.

[6] Wang Minhua, Niu Xian. Predictive model of spontaneous combustion of coal remains in the mining zone based on data reconstruction enhancement[J]. Coal Mine Safety, 2022, 53(9):86-93.

[7] Liu Bao, Mu Kun, Ye Fei, et al. Coal spontaneous combustion prediction method based on correlation vector machine[J]. Industrial and mining automation, 2020,46(09):104-108.

[8] Wen Hu, Zhao Xiangtao, Wang Weifeng, et al. Characterization of indicative gas function models for spontaneous combustion of different coal bodies [J]. Coal Conversion, 2020, 43(1):16-25.

[9] Luo Zhaoxian, Yu Wenpeng, Yu Shunqin. Research on improved BP neural network for predicting spontaneous coal combustion[J]. Coal Technology, 2017, 36(12):144-145.

[10] Ma Z, Qin B, Shi Q, et al. The location analysis and efficient control of hidden coal spontaneous combustion disaster in coal mine goaf: A case study[J].Process Safety and Environmental Protection, 2024, 184:66-78.DOI:10.1016/j.psep.2024.01.054.