Artificial Intelligence Technologies and Applications
C. Chen (Ed.)

doi:10.3233/FAIA231342

Study on PM 2.5 Concentration Prediction Model Based on Bagging-Gradient Boosting Decision Tree

Xueka HU [a1], Jiahao SHI [a2], and Riliu LIANG[b3]
[a] *Guangxi Science and Technology Normal University, Laibin 546100, Guangxi, China*
[b] *Guangxi Power Grid Corporation Laibin Power Supply Bureau, Laibin 546100, Guangxi, China*

Abstract. To accurately predict PM 2.5 concentration and provide a reference for early warning and control of PM 2.5 pollution, the PM 2.5 concentration prediction model based on the Bagging-Gradient Boosting Decision Tree (GBDT) is proposed. Pearson correlation coefficient and distance correlation coefficient are used for feature selection by analyzing the linear and nonlinear correlations between PM 2.5 and relevant initial feature variables. The bagging algorithm is used to construct multiple new samples with bootstrap sampling and GBDT is used to train the new samples to obtain weak models. The mean value method is used to assemble the weak models into the final strong model and obtain the final prediction results. The results show that for the Bagging-GBDT model, the *RMSE* is 4.83, the goodness of fit R^2 is 0.94, the predicted value can fit the actual value well, and about 80% prediction error is between [-5, +5]. It can be concluded that the Bagging-GBDT model can predict PM 2.5 concentration accurately and the error between the predicted value and the actual value is acceptable. Otherwise, compared with some machine learning models, this model has better performance, higher accuracy, better stability, and better fitting effects.

Keywords. PM 2.5 Concentration Prediction Bagging Gradient Boosting Decision Tree

1. Introduction

Global Burden of Disease studied 87 disease risk factors in 204 countries and territories from 1990 to 2019. The result shows that 6.67 million global deaths were attributable to air pollution in 2019. Ambient particulate matter pollution was a risk factor that accounted for more than 1% of Disease-Adjusted Life Years (DALYs) and was increasing in exposure by more than 1% per year [1]. In China, PM 2.5 related to 1.06 million deaths and resulted in a 705.9 billion yuan economic loss in 2016 [2]. Air pollution, especially PM 2.5 pollution, has been a threat to global human health[3-4]. It is

[1] Xueka HU, Guangxi Science and Technology Normal University, Laibin 546100, Guangxi, China; Email: huxueka513@163.com
[2] Jiahao SHI, Guangxi Science and Technology Normal University, Laibin 546100, Guangxi, China; Email: 415937208@qq.com
[3] Corresponding author: Riliu LIANG, Guangxi Power Grid Corporation Laibin Power Supply Bureau, Laibin 546100, Guangxi, China; Email: lrl407627336@163.com

significant to predict PM 2.5 concentration to provide a reference for early warning and control of PM 2.5 pollution.

For nearly two decades, soft sensors have played an important role in monitoring, controlling, and optimizing industrial processes. Machine learning is an approach to soft sensor modeling [5]. For example, Lu et al. studied soft sensing modeling of metatitanic acid particle size based on machine learning [6]. Similarly, for the monitoring and controlling of PM 2.5, a soft sensor based on machine learning is also an efficient method. For PM 2.5 prediction, numerical models and statistical models are widely used. The numerical models, which fully consider the formation and transport mechanism between atmospheric state and PM 2.5, mainly include the CMAQ, WRF-Chem [7-8], and so on. The numerical model requires a full understanding of the physicochemical processes of the source, transport, and settlement of PM 2.5, but the relevant parameters of these processes are highly uncertain, which results in uncertainty of prediction results. Statistical models make predictions by obtaining potential relationships between large amounts of data, mainly including multiple linear regression and machine learning models such as Neural Networks, ARIMA, and LSTM modules [9-10]. Comparatively, statistical models, especially machine learning models, are easier and more efficient to predict.

Gradient Boosting Decision Tree (GBDT) is a machine learning algorithm that can be used to solve regression and classification problems, which has been widely used in prediction. Zha et al. built a prediction model of end-point manganese content based on GBDT [11], Gong et al. applied GBDT to the prediction of blood glucose [12], and the results show that GBDT has a strong learning ability and good prediction effect. Liang et al. built a wind power prediction model based on a Bagging-Neural Network [13], and Qiu et al. built a load forecasting model based on a Bagging-combined Kernel Function Relevance Vector Machine [14]. The results show that bagging combined with machine learning can improve prediction performance. Based on the above ideas, the PM 2.5 concentration prediction model based on Bagging-GBDT is proposed in this paper.

2. PM 2.5 concentration prediction model based on Bagging-GBDT

The PM 2.5 concentration prediction model based on Bagging-GBDT mainly includes three modules: feature selection, sampling with the bagging algorithm, and training and prediction with GBDT.

2.1. Feature selection

The initial feature vectors are air quality indicators related to PM 2.5, such as AQI, PM 10, SO_2, CO, NO_2, and O_3. The features highly related to PM 2.5 are selected as input vectors through feature selection, and features with less impact on the prediction results are eliminated, so that the model can quickly establish the input-output relationship. To fully analyze the linear and nonlinear correlation between feature vectors and PM 2.5, the Pearson correlation coefficient and distance correlation coefficient are comprehensively used here.

Pearson correlation coefficient R is used to study the linear correlation between variables, whose value is the quotient of the covariance and standard deviation of two variables X and Y as Formula (1).

$$R = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

Distance correlation coefficient d_R is used to study the nonlinear correlation between variables. For two variables (X,Y), the algorithm is as follows:

- We calculate the norm distance between the data in each row.

$$a_{jk} = \left\| X_j - X_k \right\|, b_{jk} = \left\| Y_j - Y_k \right\| \tag{2}$$

- We centralize all pairwise distances. For Variable *X*, $\bar{a}_{j.}$ is the average value of row *j*, $\bar{a}_{.k}$ is the average value of Column *k*, and $\bar{a}$ is the average value of the distance matrix. For Variable *Y*, the implication of $\bar{b}_{j.}$, $\bar{b}_{.k}$, and $\bar{b}$ are the same.

$$A_{jk} = a_{jk} - \bar{a}_{j.} - \bar{a}_{.k} + \bar{a}, B_{jk} = b_{jk} - \bar{b}_{j.} - \bar{b}_{.k} + \bar{b} \tag{3}$$

- We calculate the arithmetic mean of the square covariance of the distance between two variables and the distance variance of each variable.

$$dCov_n^2(X,Y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{jk}B_{jk}, dVav_n^2(X) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{jk}^2, dVav_n^2(Y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} B_{jk}^2 \tag{4}$$

- We calculate the distance correlation coefficient between the two variables.

$$d_R = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}} \tag{5}$$

2.2. Bagging

Bagging is an algorithm that integrates weak learners into strong learners. The weak learners can be fitted in parallel because they are independent from each other. For PM 2.5 prediction, given the original sample data, the Bagging algorithm conducts *N* times bootstrap sample and reconstructs *N* training sets with the same size as the original samples. The weak prediction model is obtained by training the *N* training sets, and the final strong prediction model is obtained by assembling the *N* weak prediction models with certain rules. The Bagging algorithm flow is shown in Figure 1. Through Bagging, the accuracy and stability of the prediction model can be improved, and overfitting can be reduced.

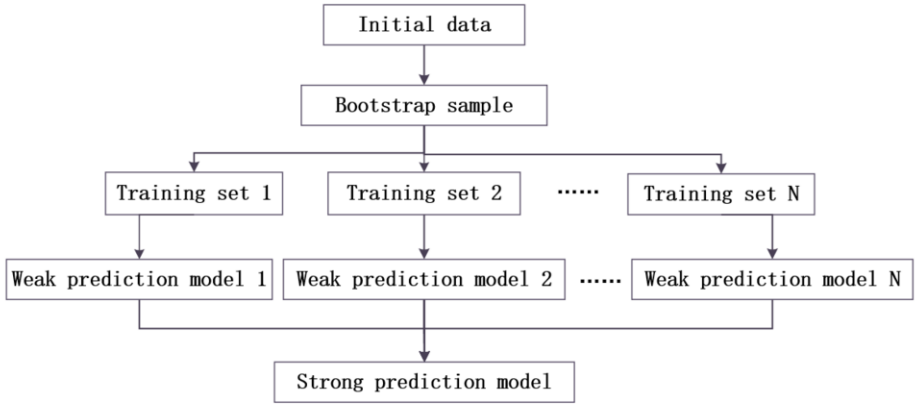

Figure 1. Bagging algorithm flow

2.3. GBDT

GBDT is an algorithm based on Decision Tree and Gradient Boosting Machine. Its base learning algorithm is the Classification and Regression Tree (CART). GBDT is used to solve the regression problem of PM 2.5 prediction here. The Decision Tree gradually finds the optimal classification rule with a series of decision conditions from top to leaf, and finally gets the predicted value at the leaf node. GBDT uses the negative gradient of the loss function to fit the specific value of the current model. The model output gradually approximates the real value with each iteration. The goal of each iteration is to find a weak evaluator to minimize the sum and loss function. The final output is the combination of the results of these weak evaluators. Compared with the general boosting tree, GBDT has higher applicability because its loss function is not limited to the square loss function, which can also optimize the model quickly. The process for solving regression problems with GBDT is as follows:

- We input a data set $D=\{(x_1,y_1),(x_2,y_2)...,(x_n,y_n)\}$ and initialize the regression tree $f_0(x)$, where L is the loss function, y_i is the true value corresponding to x_i, and c is the constant that minimizes the loss function.

$$f_0(x)=\arg\min_c \sum_{i=0}^{N} L(y_i,c) \tag{6}$$

- We perform M iterations. For the m iteration, we calculate the value of the negative gradient in the current model for each sample (x_i,y_i):

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

We use $\left(x_i, r_{m,i}\right)_{i=1,2,\ldots n}$ to train a new decision tree $T\left(x, \theta_m\right)$ and then update $f_m\left(x\right)$, where θ_m is the parameters of the decision tree.

$$f_m\left(x\right) = f_{m-1}\left(x\right) + T\left(x, \theta_m\right) \tag{8}$$

- The final GBDT model is the integration of *M* decision trees obtained by *M* iterations.

$$f_M\left(x\right) = \sum_{m=1}^{M} T\left(x, \theta_m\right) \tag{9}$$

2.4. Prediction model based on Bagging-GBDT

The algorithm flow of the PM 2.5 concentration prediction model based on Bagging-GBDT is shown in Figure 2:

- Feature selection. Feature selection is conducted by the Pearson correlation coefficient and distance correlation coefficient to obtain the features strongly related to PM 2.5 for training and modeling.
- Bagging. We conduct *N* times bootstrap sampling and reconstruct *N* training sets.
- GBDT. We train the *N* training sets with GBDT, and *N* weak models and *N* prediction results are obtained. The mean value method is used to assemble the *N* weak models into the final model and obtain the final prediction results.
- Model performance evaluation. *RMSE* and R^2 are used to evaluate prediction performance, in which *RMSE* is the root mean squared error and R^2 is the goodness of fit.
- Parameter determination. The key parameters of the model are determined by the grid search method to optimize the performance of the model.

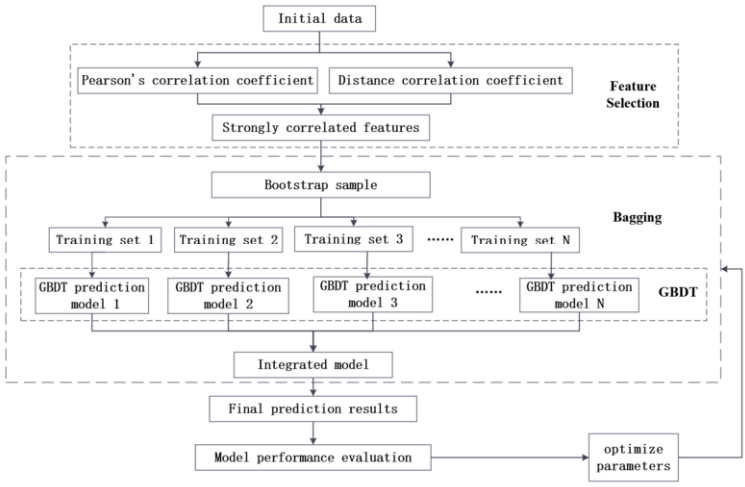

Figure 2. Algorithm flow of the Bagging-GBDT model

3. Experiments and results analysis

3.1. DataSet and feature selection

The dataset is PM 2.5, AQI, PM 10, SO_2, CO, NO_2, and O_3, which is the daily data of Liuzhou from January 2016 to May 2023, with a total of 2, 708 groups, of which 2, 165 groups (about 80%) are used as training sets and the remaining 543 groups are used as test sets. According to the theory and algorithm in 2.1, the cor () and decor () functions in R language are used to calculate the Pearson correlation coefficient and distance correlation coefficient between variables respectively. The greater the absolute values of both coefficients are, the stronger the correlation is. If both of them are greater than 0.5, the feature vector will be considered to be strongly correlated with PM 2.5. The correlation coefficient calculation results are shown in Table 1. It can be seen that the Pearson correlation coefficient and distance correlation coefficient between PM 2.5, AQI, PM 10, SO_2, and NO_2 are greater than 0.5, so they are selected as feature vectors for subsequent model training.

Table 1. The correlation coefficient between PM 2.5 and relevant feature vectors

Relevant feature vectors	AQI	PM10	SO_2	CO	NO_2	O_3
Pearson correlation coefficient	0.90	0.95	0.55	0.07	0.72	0.22
Distance correlation coefficient	0.89	0.95	0.51	0.24	0.72	0.31

3.2. Model specification and experiment results

For optimizing the performance of the PM 2.5 concentration prediction model based on Bagging-GBDT, the key parameters need to be determined. For Bagging, the key parameter is sampling numbers. The more the sampling numbers are, the larger the calculation amount is, and the longer the calculation time is, which is more likely to cause trade-off errors. However, if the number of samples is too small, the effect on improving stability and accuracy will not be obvious. For GBDT, the key parameters are the loss function, maximum number of iterations, learning rate, and maximum depth of the decision tree. For regression models, the loss function can be Mean Square Error, Absolute, and Huber loss. In general, when the data quality is good, the Mean Square Error is better, if not, the anti-noise loss function Huber is recommended. The data is reliable in this paper, so the loss function of Mean Square Error is selected. The above parameters are determined by the grid parameter optimization method. First, the default parameters are used, and then the optimal parameters are gradually determined according to the *RMSE* and R^2 values in the order of maximum number of iterations, learning rate, maximum depth, and sampling numbers. The smaller the *RMSE* is, the higher the prediction accuracy is. The closer R^2 is to 1, the better the fit is. The results of grid parameter optimization are shown in Tables 2-5.

Table 2. RMSE and R^2 values with different maximum number of iterations

Maximum iterations	20	30	40	50	60	70	80	90	100
RMSE	6.18	5.52	5.35	5.30	5.34	5.38	5.41	5.42	5.43
R^2	0.90	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93

Table 3. RMSE and R^2 values with different learning rates

Learning rate	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
RMSE	13.34	5.75	5.30	5.51	5.50	5.31	5.36	5.44	5.56	5.45	5.46	6.04
R^2	0.55	0.92	0.93	0.92	0.92	0.93	0.93	0.92	0.92	0.92	0.92	0.91

Table 4. RMSE and R^2 values with a different maximum depth

Maximum depth	1	2	3	4	5	6	7	8	9	10
RMSE	7.13	6.05	5.30	5.00	4.87	4.93	4.93	5.00	5.05	5.08
R^2	0.87	0.91	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.93

Table 5. RMSE and R^2 values with different sampling numbers

Sampling numbers	5	10	15	20	25	30
RMSE	5.01	4.94	4.87	4.83	4.87	4.87
R^2	0.94	0.94	0.94	0.94	0.94	0.94

It can be seen from the above tables that the prediction performance of the model is best when the maximum number of iterations is 50, the learning rate is 0.1, the maximum depth is 5, and the sampling number is 20. With the optimized parameters, *RMSE* is 4.83, and R^2 is 0.94. The prediction effect is shown in Figures 3 and 4. They show that the model can predict PM 2.5 concentration accurately, and the predicted value can fit the actual value well. The error between the predicted value and the actual value is acceptable, and about 80% error is between [-5, +5].

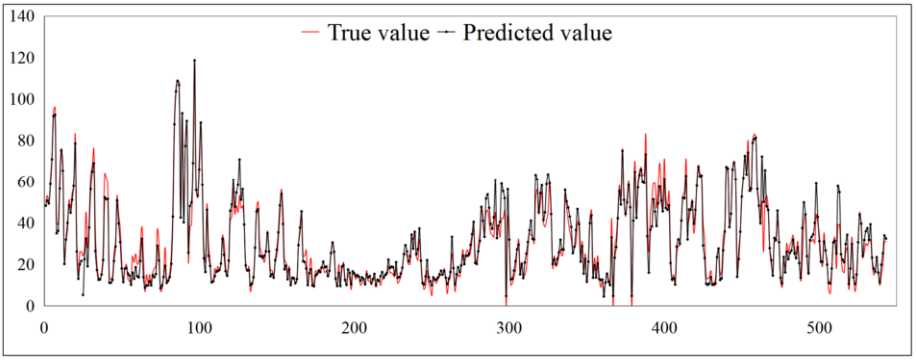

Figure 3. Prediction performance

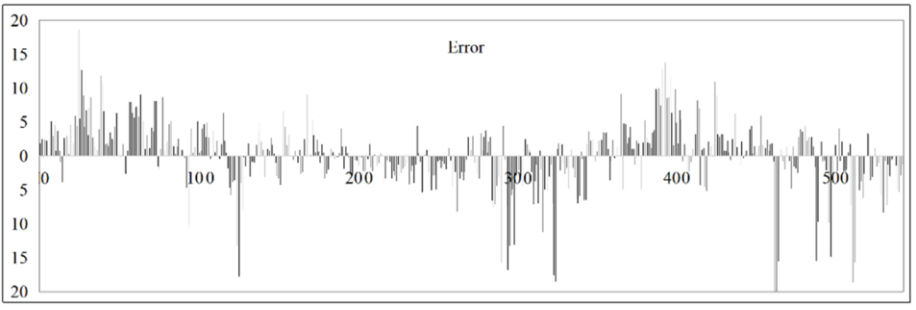

Figure 4. The error between the predicted value and the true value

3.3. Model comparison

To further evaluate the performance of the Bagging-GBDT model in predicting PM 2.5 concentration, a separate GBDT model, BP Neural Network, and Random Forest are selected as comparison models. The key parameters of the comparison models are also adjusted to make the prediction performance best. The parameters of the separate GBDT and Random Forest are the same as the relevant parameters of the GBDT in the Bagging-GBDT model. The epochs of BP Neural Network are 200 and the error threshold is 1-6. *RMSE* and R^2 are evaluation indexes. The comparison results are shown in Table 6, which shows that the Bagging-GBDT model has better prediction performance. Compared with the separate GBDT model, BP Neural Network, Random Forest, and the Bagging-GBDT model decreased by 0.41, 5.06, and 0.14 in *RMSE* and increased by 0.01, 0.19, and 0 in R^2, respectively.

Table 6. Performance of different models

Module	RMSE	R^2
Bagging-GBDT	4.83	0.94
GBDT	5.24	0.93
BP Neural Network	9.89	0.75
Random Forest	4.97	0.94

4. Conclusions

The PM 2.5 concentration prediction model based on Bagging-GBDT is proposed in this paper. That is, the Pearson correlation coefficient and distance correlation coefficient are used for feature selection by analyzing the linear and nonlinear correlation between PM 2.5 and initial feature vectors, such as AQI, PM 10, SO_2, CO, NO_2, and O_3. The bagging algorithm is used to construct multiple new samples with bootstrap sampling and GBDT is used to train the new samples to obtain weak models. The mean value method is used to assemble the weak models into the final strong model. The results show that for the Bagging-GBDT model, the *RMSE* is 4.83, the goodness of fit R^2 is 0.94, the predicted value can fit the actual value well, and about 80% prediction error is between [-5, +5]. It can be concluded that the Bagging-GBDT model can predict PM 2.5 concentration accurately and the error between the predicted value and the actual value is acceptable. Otherwise, compared with a separate GBDT model, BP Neural Network, and Random Forest, Bagging-GBDT has better performance, higher accuracy, better stability, and better fitting effect in predicting PM 2.5 concentration. The study of this model is helpful for more efficient and accurate prediction of PM 2.5, and it can provide a reference for related industrial production and residents' lives. In future studies, more variables related to PM 2.5, such as seasons, meteorological factors, and geographical conditions, can be considered for analysis, and deep learning methods can be combined to build better models for air quality prediction.

Acknowledgments

This paper is supported by Guangxi University Young and Middle-aged Teachers' Basic Ability Improvement Project (Prediction Model of PM 2.5 Concentration in

Central Guangxi Based on Deep Learning, 2020KY23019).

References

[1] Christopher J. L., Murray, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019 [J]. *Lancet*, **2020 (396)**: 1, 223-49. https://www.thelancet.com/journals/lancet/issue/vol396no10258/ PIIS0140-6736(20).

[2] Li Y., Liao Q., et al. Influence of PM2. 5 Pollution on Health Burden and Economic Loss in China [J]. *Environmental Science*, **2021, 42 (4)**: 1, 688-1, 695. https://doi.org/10.13227/j. hjkx.202008313.

[3] Mohsen Abbasi-Kangevari, Mohammad-Reza Malekpour, Masoud Masinaei, et al. Effect of air pollution on disease burden, mortality, and life expectancy in North Africa and the Middle East: a systematic analysis for the Global Burden of Disease Study 2019 [J]. *Lancet Planet Health,* **2023,7(5)**: e358–69. https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(23)00053-0/fulltext.

[4] Katrin Burkart, Kate Causey, Aaron J Cohen, et al. Estimates, trends, and drivers of the global burden of type 2 diabetes attributable to PM2·5 air pollution, 1990–2019: an analysis of data from the Global Burden of Disease Study 2019 [J]. *Lancet Planet Health,* **2022,6(7)**: e586–600.https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(22)00122-X/fulltext#%20.

[5] Jiang Y. C., Yin S., et al. A Review on Soft Sensors for Monitoring, Control, and Optimization of Industrial Processes [J]. *IEEE Sensors Journal*, 2021, **21 (11):** 12, 868-12, 881. https://www.nstl.gov.cn/paper_detail.html?id=51a9eba0a1eab19fe62b59ff44caacba.

[6] Lu R. F., Liu C., et al. Soft sensing modeling of metatitanic acid particle size based on machine learning [J]. *Iron Steel Vanadium Titanium*, 2021, **42 (02)**: 36-42. https://doi.org/10.7513/j. issn.1004-7638.2021.02.007.

[7] Chen M., Hao J., et al. Test analysis of air quality forecast effect in Yinchuan City based on WRF-CMAQ model [J]. Meteorological and Environmental Sciences, 2023, **46 (02)**: 83-91. https://doi.org/10.16765/j.cnki.1673-7148.2023.02.011.

[8] Chen J., Li Z. Q., et al. Impact Evaluation of Aerosol Variational Assimilation Based on Improved GSI System on WRF-Chem PM 2.5 Analysis and Forecast [J]. *Journal of Atmospheric and Environmental Optics,* 2020, **15 (15)**: 321-333. https://kns.cnki.net/kcms2/article/abstract?v =3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7i8oRR1PAr7RxjuAJk4dHXoikAN4NOAy1j4cNMs0nlN96JXDzY_0OUKxdJPRdDF0zr&uniplatform=NZKPT.

[9] Gu K., Jiao R. L., et al. PM 2. 5 Concentration Prediction Based on the Composite LSTM Model [J]. *Environmental Monitoring in China,* 2023, **39 (1)**: 170-180. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7ioT0BO4yQ4m_mOgeS2ml3UIRrANpPHVSzoVXAL3izA483a_3ov-Ld7VY2IrWZYdae&uniplatform=NZKPT.

[10] Liu Y. M., Luo H. Y., et al. Research and application of PM 2.5 mass concentration prediction model based on XGBoost ARIMA method [J]. *Journal of Safety and Environment*, 2023, **23 (1)**: 211-221. https://doi.org/10.13637/j.issn.1009-6094.2021.1849.

[11] Zha W., Dong Y. W., et al. Prediction model of end-point manganese content in vacuum consumable ingot based on GBDT algorithm [J]. *China Metallurgy*, 2023, **33 (07)**: 107-114. https://doi.org/10.13228/j.boyuan.issn1006-9356.20230098.

[12] Gong Y. C., Du C. H., et al. Prediction of blood glucose value based on principal component and GBDT [J]. *Mathematics in practice and theory*, 2019, **49 (14)**:116-122. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iLik5jEcCI09uHa3oBxtWoHyKrXYgZsH1edc8jQZtU1Oxn9sshvS-yAGREtRQyZfY&uniplatform=NZKPT.

[13] Liang T., Shi H., et al. Wind power prediction based on Bagging neural network integration [J]. *water resource and power*, 2020, **38 (04)**: 205-208. https://kns.cnki.net/kcms2/article/ abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7i8oRR1PAr7RxjuAJk4dHXorGYjoqCfEq3L4ijGjGgULM2HVd29L_4KJwQ7qS-ZoNX&uniplatform=NZKPT.

[14] Qiu S., Gong W. J., et al. Research on Short-term Load Forecasting Model Based on Bagging-combined Kernel Function Relevance Vector Machine [J]. *Journal of Electrical Engineering*, 2023, **18 (02)**: 142-148. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44 YLTlOAiTRKu87-SJxoEJu6LL9TJzd50nhwF9M7cHVN5lSADpbUCeDLUte1Ndmj_hyh PNOzjTxEygHyVL0vp_D&uniplatform=NZKPT.