International Journal of Energy
ISSN: 2957-9473 | Vol. 6, No. 1, 2025

Prediction Model for Fluid Accumulation of Long distance Pipeline Based on Improved Particle Swarm Optimization Algorithm and Random Forest

Xinhai Zhang

College of Petroleum Engineering, Xi'an Shiyou University, Xi'an Shaanxi, 710065, China

Abstract: At present, as most gas wells gradually enter the middle and late stages of production, the proportion of wet gas has greatly increased, leading to increasingly serious problems of pipeline fluid accumulation. In order to ensure the efficiency and economy of pipeline transportation, this paper optimizes the number and depth of trees in the Random Forest (RF) model for predicting volume of fluid accumulation of pipelines based on an improved Particle Swarm Optimization (PSO) algorithm. Firstly, 45330 sets of on-site data collected from a certain gas gathering station were used as the dataset. Among them, 45231 sets of data were used as the training set for model training, and the remaining 99 sets of data were used as the testing set for validation. At the same time, quantitative analysis was conducted on the influencing factors of pipeline volume of fluid accumulation based on SHAP (Shapley Additive ex Plans) to determine the five characteristics of oil pressure, daily gas production, casing pressure, cumulative gas production, and gas water ratio. Finally, the optimized random forest model is applied to predict the volume of fluid accumulation of the gathering and transportation pipeline. The prediction results show that the MAE of the optimized random forest model based on the improved PSO algorithm reached 0.003168, the MSE reached 0.0004089, and the MAPE reached 1.045%. The accuracy, universality, and feasibility of the optimized model in this paper have been verified, and it can provide some reference for the design and maintenance of natural gas pipelines.

Keywords: Random Forest; Particle Swarm Optimization Algorithm; Volume of Fluid Accumulation; Long Distance Pipeline.

1. Introduction

Pipeline fluid accumulation is an important research topic in the petroleum and natural gas industry. The production of natural gas is always accompanied by liquid hydrocarbons, free water, and compressed steam. Generally, the gas phase flows in a single-phase manner at the inlet of the pipeline, but due to pressure drop, temperature difference, and other reasons, liquid may appear in the pipeline due to retrograde condensation, resulting in pipeline accumulation. This phenomenon of fluid accumulation will have a significant impact on the operational efficiency, safety, and equipment lifespan of pipelines [1-4].

As a significant portion of gas well production enters the mid to late stages, it becomes increasingly prone to liquid accumulation in pipelines. The presence of fluid accumulation can alter the flow characteristics of pipelines, leading to a decrease in airflow velocity and a decrease in transport efficiency. In addition, fluid accumulation can increase the operating pressure of the pipeline, thereby increasing the load and energy consumption of the pipeline. If the accumulated fluid cannot be treated in a timely manner, it may also cause safety accidents such as pipeline corrosion and rupture. This article provides a certain reference for the judgment of pipeline liquid accumulation through the prediction of volume of fluid accumulation, and can also provide theoretical support for the optimization of pipeline cleaning plans, reducing unnecessary cleaning operations and thus lowering operating costs. At the same time, it can also avoid the wear of the inner wall of the pipeline caused by improper cleaning, and extend the service life of the pipeline. Finally, accurate volume of fluid accumulation forecasting also helps with pipeline layout and design. During the pipeline design and construction phase, by predicting the volume of fluid accumulation, the pipeline elevation and slope can be reasonably arranged to avoid or reduce the volume of fluid accumulation phenomenon in low-lying areas and improve the overall design rationality of the pipeline. Considering that gas-liquid two-phase flow is a complex system, most current mechanistic models propose empirical or semi empirical relationships. However, this method is only applicable under specific operating conditions and has significant limitations [5-6]. Therefore, we consider using a random forest network for predicting the volume of fluid accumulation of pipelines, which has strong noise resistance and can adapt to the vast majority of situations. However, traditional random forests are a black box model with many adjustable parameters [7-9].

Therefore, this article adopts the SHAP method to break the "black box" problem and explain the model decisions. To the best of the author's knowledge, there is currently no random forest network model to predict the volume of fluid accumulation of pipelines. Therefore, this article considers using a random forest optimized by improved particle swarm optimization (IPSO) to predict the volume of fluid accumulation in the gathering and transportation pipeline.

2. Algorithm and Principle

2.1. Principle of Random Forest Algorithm

In 2001, American scientists Breiman and Cutler proposed a revolutionary decision tree-based ensemble learning algorithm. It improves the accuracy and stability of predictions by building multiple decision trees and voting or averaging them [10-11]. It combines decision trees, random subspaces, sampling aggregation, and pruning techniques together. Not only does introducing randomness during the training process enhance the model's generalization ability

and reduce the risk of overfitting, but it also has good tolerance for data noise, missing values, and outliers [12]. And for high-dimensional problems, it also has good accuracy and scalability. Therefore, random forests have been widely applied in many fields.

Random forest integrates multiple decision trees, each of which is a learner, and then combines their prediction results to improve the performance of individual trees. It can also be said that the random forest algorithm matches the best result by voting on each decision tree. The algorithm steps for random forest are as follows:

(1) Randomly extract multiple subsets of samples with replacement from the original training dataset D={(x_i, y_i)} (i=1,2,..., n)., Each subset of samples will be used to train a decision tree.

(2) The decision tree trained on each individual subset of training samples first selects a certain number of features from all features at each node, and then selects the corresponding features based on the Gini coefficient for branching and non-pruning growth until it cannot grow, thus avoiding overfitting.

(3) Considering the sampling of data that has been put back, there are some data that have not been selected and are called out of bag data. Using out of bag scores to determine relevant parameters for modeling can improve the generalization ability of the model to a certain extent.

(4) The decision tree corresponding to each subset will predict the input data and generate a predicted value. Due to the fact that the decision tree of a random forest is constructed based on the bagging (Bootstrap Aggregating) method, each tree is trained on different subsets of data, and only one random subset of features is considered when selecting segmentation points, resulting in a certain degree of differentiation and independence for each tree. So the random forest takes the average of the predicted values of all trees for the same sample to obtain the final prediction result for that sample.

The random forest model has a fast convergence speed for prediction, but at the same time, the model can adjust many parameters and is highly sensitive to parameters, as shown in Table 1. Especially, the number of decision trees and the maximum depth of decision trees are mostly determined by experience, which can make the model prone to overfitting. Therefore, this article optimizes the selection of the number and maximum depth of decision trees for random forests.

Table 1. Influence parameters of RF model

parameters	Influence	importance
n_estimators	More trees can increase the stability and accuracy of the model, but it can also increase computational costs. If the number of trees is too small, the model may be affected by random fluctuations, while too many may lead to overfitting.	very high
max_depth	Controlling tree depth helps prevent overfitting. If the depth is too large, the model may learn noise from the data; If it is too small, it may not be possible to capture the underlying structure of the data.	high
max_features	Choosing the correct number of features is crucial for the performance of the model. Too many features may lead to overly complex and overfitting models, while too few features may lead to underfitting.	high
min_samples_split	It determines the growth rate and complexity of the tree. A higher value can prevent the model from creating overly complex tree structures on the training data.	medium
min_samples_leaf	Higher values can increase the robustness of the model and prevent overfitting.	medium

2.2. IMPSO Algorithm Optimization of Random Forest Model

The PSO algorithm is a biomimetic algorithm proposed to simulate bird swarm predation. The algorithm is simple and efficient, requiring fewer parameters to be adjusted, but with clear meaning and high computational efficiency. However, traditional PSO algorithms are prone to overfitting and local optima when dealing with complex problems with high computational dimensions. Therefore, this article improves the standard PSO algorithm and proposes the Improved Particle Swarm Optimization (IMPSO) algorithm, which introduces adaptive mutation and expands the search range of particles, thereby enabling the algorithm to escape from local optima.

The PSO algorithm imitates bird foraging by first empowering particles, and then iteratively updating particle velocity and position through self-learning and global learning, in order to find individual optimal solutions and then converge towards the global optimal solution position. Firstly, initialize the parameters of the PSO algorithm, including particle population position, velocity, individual optimal solution (pBest), global optimal solution (gBest), inertia weight (w), learning factors (c1, c2), and particle count.

Assuming that the position and velocity of the i-th particle in a D-dimensional space are Xi= (xi, 1, xi, 2,..., xi, D) and Vi=(vi, 1, vi, 2,..., vi, D), respectively, the individual optimal solution (pBest) and global optimal solution (gBest) are updated iteratively. The particle velocity and position at time k+1 after the update are:

$$V_{\mathrm{i}}^{\mathrm{k+1}} = \omega V_{\mathrm{i}}^{\mathrm{k}} + c_1 \times rand \times \left(pBest_i^{\mathrm{k}} - x_i^{\mathrm{k}} \right) + c_2 \times rand \times \left(gBest_i^{\mathrm{k}} - x_i^{\mathrm{k}} \right) \quad (1)$$

$$x_i^{\mathrm{k+1}} = x_i^{\mathrm{k}} + v_i^{\mathrm{k+1}} \quad (2)$$

Considering the characteristics of PSO algorithm, in order to improve the calculation speed of the algorithm, the global search ability and local search ability are adjusted by adjusting the inertia weight w. When the value of inertia weight is small, the local search ability of particles is strong; When the value of inertia weight is large, the global search ability of particles is strong [13]. Therefore, a large step size is used in the early stage of training, and a small step size is used in the later stage, so that the algorithm has greater exploration ability in the early stage, and more emphasis is placed on development in the later stage, as shown in equation (3)

$$\omega = \frac{\left(\omega_{\max} - \omega_{\min} \right) gt}{T} \quad (3)$$

In equation (3): ω is the current inertia weight; ωmax is

the maximum value of inertia weight; ωmin is the minimum value of inertia weight; t is the current iteration count; T is the maximum number of iterations.

Considering that the PSO algorithm is prone to getting stuck in local optima and in order to avoid excessive search and waste of search resources, the timing of mutation operations is usually determined based on the population fitness variance. The current motion state of the particle swarm within the population is determined by studying the changes in the population fitness variance σ^2 [14-15]. The smaller the variance of group fitness squared, the closer the particles in the population are, and the higher the probability of extreme value variation in the population; On the contrary, the larger the value of population fitness variance, the lower the probability of population extreme value variation, indicating that the population is in the random search stage. The definition of σ^2 is:

$$\sigma^2 = \frac{\sum_{n=1}^{N}(f_n - f_{av})^2}{N} \tag{4}$$

In the formula, N is the number of particles; fn is the fitness value of the n th particle; fav is the average fitness value within the population; The function of f is to limit the size of σ 2.

Based on the above, calculate the mutation probability Gm and perform random mutation accordingly. The formula for calculating Gm is as follows:

$$G_m = \begin{cases} 0.4, & \sigma^2 < \sigma_d^2 \text{ and} f_{(gBest)} > f_d \\ 0, & \text{other} \end{cases} \tag{5}$$

In equation (5): f_d is the expected optimal solution; The value of σ_d^2 is related to practical problems and is generally much smaller than the maximum value of σ_d^2.

For the mutation operation of gBest, this algorithm will adopt the method of adding random perturbations. Let gBest k be the k-th dimensional value of gBest, where η is a random variable distributed according to Gauss (0,1)

The specific calculation process is as follows:

(1) Initialize the random forest structure.

(2) Initialize particle swarm parameters, including population size, dimensionality, particle velocity, and position [16].

(3) Determine the fitness function as a guide for the PSO algorithm. The function is as follows:

$$F = \left(\sum_{i=1}^{n} \text{abs}(y_i - x_i) \right) \tag{6}$$

(4) By comparing fitness, determine the individual optimal solution (pBest) and the global optimal solution (gBest).

(5) Calculate the mutation probability Gm and determine whether to perform adaptive mutation by generating a random number, in order to escape from the local optimal solution.

(6) Update inertia weights, particle velocity, and position.

(7) Determine the optimal number and depth of decision trees.

(8) Use random forest for prediction.

3. Model Explanation

In the fields of machine learning and deep learning, model interpretability is an important issue. Although complex models such as deep neural networks and ensemble models (such as XGBoost, LightGBM) perform well in predictive performance, they are often seen as "black boxes" that are difficult to explain their internal decision-making processes. SHAP (Shapley Additive exPlans) is a tool that solves this problem by assigning importance values to features to explain the output of the model. It provides a practical and theoretically solid interpretable method.

It has the following characteristics:

(1) Consistency: If the feature contribution of the model increases, the SHAP value will also increase

(2) Local interpretability: SHAP values can explain individual prediction results, providing explanations for a single data point.

(3) Global interpretability: By summarizing the SHAP values of multiple data points, a global interpretation of the model can be provided.

Considering that the SHAP value increases with the contribution of the model's features, a feature importance map was drawn using SHAP as shown in Figure 1.

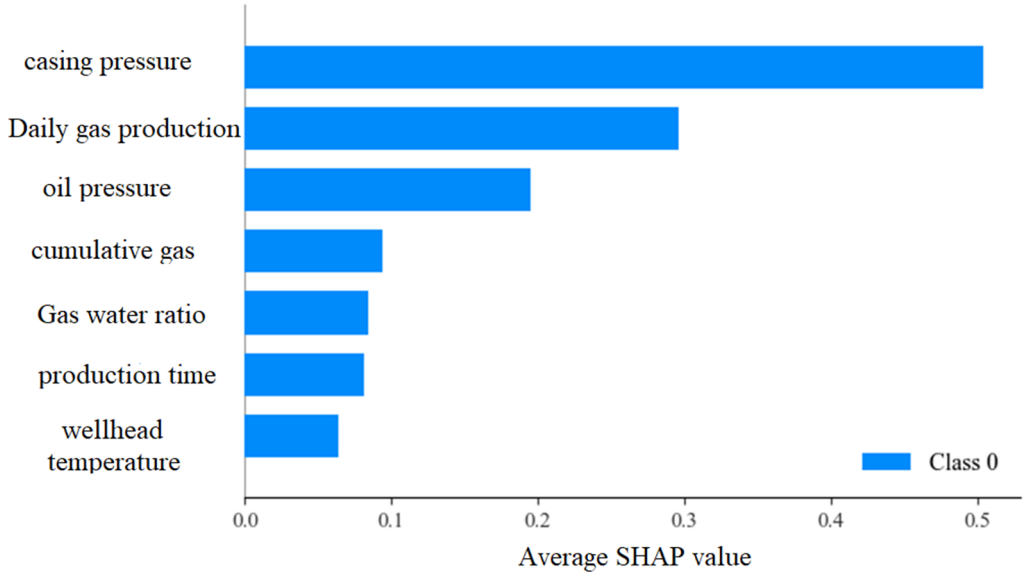


Figure 1. Feature Importance Map

The summary graph shows the distribution of SHAP values for each feature. If the SHAP value is positive, it indicates that the feature value has increased the predicted value of the model; If the SHAP value is negative, it indicates that the feature value has reduced the predicted value of the model. Thus, understanding the contribution of each feature to the model prediction, higher daily gas production and gas water ratio will lead to an increase in volume of fluid accumulation prediction, as shown in Figure 2.

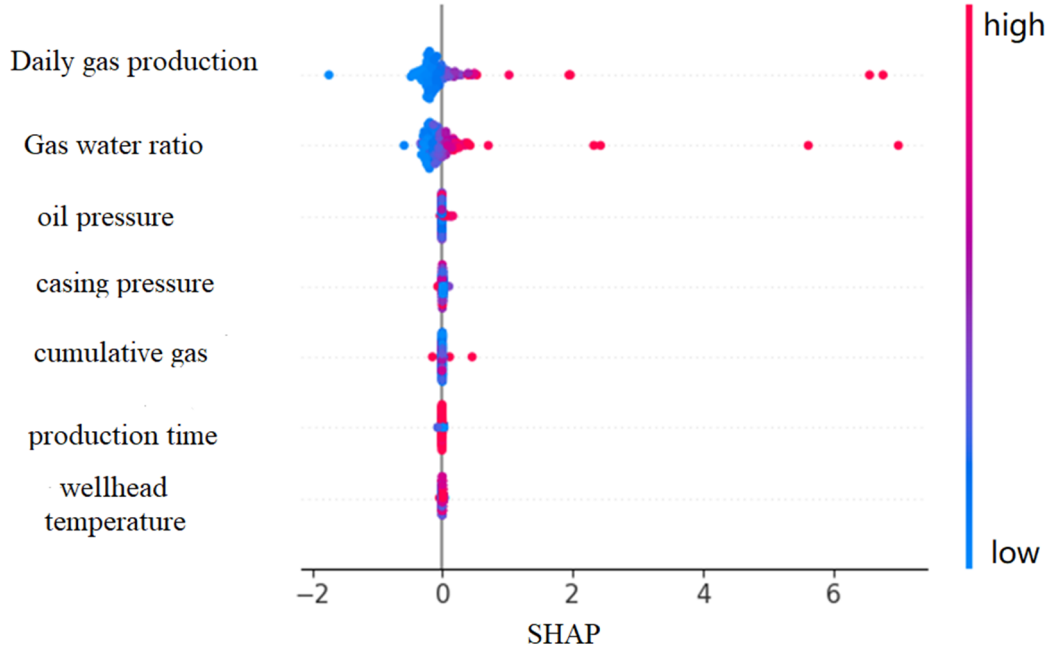


Figure 2. Summary Diagram

As shown in Figure 1 and Figure 2, oil pressure, daily gas production, casing pressure, cumulative gas production, and gas water ratio have a relatively significant impact on volume of fluid accumulation. Therefore, this article considers them as influencing factors for machine learning.

4. Model Training and Validation

4.1. Optimization Objects and Parameters

Given that the characteristics and scale differences of data may have an impact on the analysis and processing results, in order to eliminate dimensional differences, accelerate the learning process, and improve accuracy, this study uses normalization methods to process the data to ensure that all data are on the same scale, and standardizes their range to between 0 and 1, making the data more reliable [17-18]

$$x = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{7}$$

In the formula: X_i is the input data; X_{min} is the minimum value of input data; X_{max} is the maximum value of the input data; X is the output data.

In order to visually observe the comparison results of the models, in this study, the evaluation indicators include mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), and coefficient of determination (R2), which are used to measure the performance of the models:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{8}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{y_i}\right| \tag{9}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \tag{10}$$

In the formula: X_i is the i-th predicted value;y_i is the i-th actual value; N is the number of predicted data.

4.2. Model Establishment and Analysis

With the increasing demand for natural gas, the volume of pipeline transportation in China is growing, but there is still a lack of prediction for the volume of fluid accumulation of pipelines. In order to improve the predictive ability, generalization ability, and escape from local optima of the model, this article adopts the IPSO-RF model. Figures 3-4 shows the prediction of IPSO-RF model for 99 sets of on-site production data.

According to Figure 5, the prediction error of IPSO optimized RF is mostly below 2.5%. This improvement makes the prediction results closer to the actual values, thereby improving the quality of the prediction. Table 2 shows the differences in mean absolute error, squared error, and mean absolute percentage error of IPSO-RF, which better demonstrate the predictive accuracy and efficiency of the model.

The performance of IPSO-RF in model prediction is shown in Table 2. Overall, IPSO-RF has the characteristics of high accuracy and high generalization, and can become an effective method for predicting volume of fluid accumulation in pipelines.

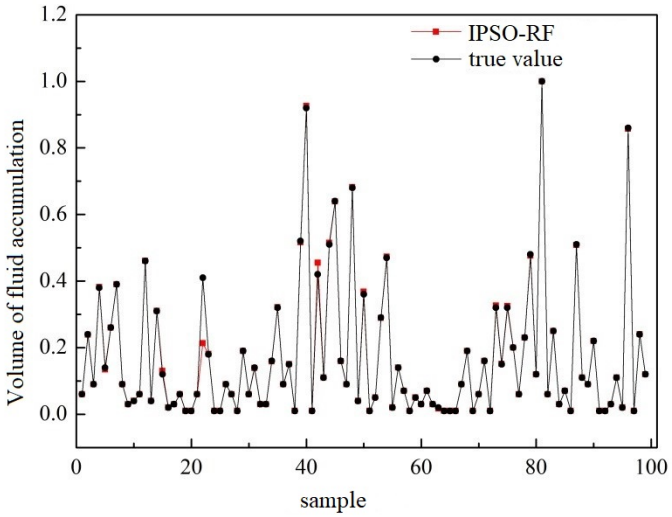


Figure 3. Prediction Results of IPSO-RF Model

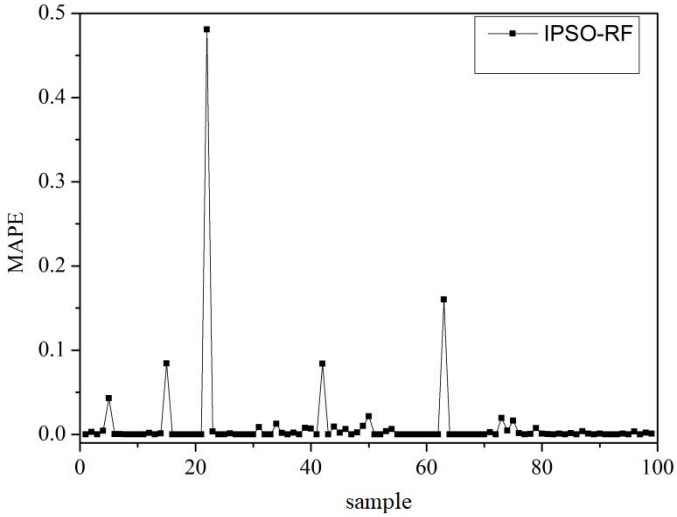


Figure 4. Prediction Error Chart of IPSO-RF Model

Table 2. Comparison of mean absolute error, mean square error, and mean absolute percentage error of IPSO-RF

Model	MAE	MSE	MAPE
IPSO-RF	0.003168	0.0004089	0.01045

5. Conclusion

The accurate prediction of volume of fluid accumulation in pipelines is of great significance for the safe and efficient operation of pipelines in the oil and gas industry. In this study, we propose using an improved particle swarm optimization algorithm to optimize a random forest to solve the problem of predicting volume of fluid accumulation. This solves the problem that the current judgment of pipeline fluid accumulation still mostly relies on manual judgment. And through machine learning prediction of 99 sets of on-site data, we found that the IPSO-RF model has high accuracy in predicting volume of fluid accumulation, with MAE reaching 0.003168, MSE reaching 0.0004089, and MAPE reaching 1.045%, far exceeding the accuracy required by general on-site requirements of 95%. This proves the effectiveness and universality of this study in predicting volume of fluid accumulation in pipelines.

References

[1] Yuan Jingling, Zhong Luo, Li Xiaoyan Research and Development of Grey Neural Networks [J]. Journal of Wuhan University of Technology, 2009, 31 (03): 91-93.

[2] Luan Guohua, He Shunli, Shu Shaoyi, etc Application of artificial neural network method to predict gas well fluid accumulation [J]. Block oil and gas fields,2010, 17(05): 575-578.

[3] CHEN Min, LIU Jun. The improvement of BP neural network and its applications[J] Journal of Hunan University of Arts and Science (Natural Science Edition), 2005, 17(2) : 68-71.

[4] OSMAN E S. Artificial neural network models for identifying flow regimes and predicting liquid holdup in horizontal multiphase flow[J]. SPE Production & Facilities, 2004, 19 (1) : 29-34.

[5] Zhang Yang, Ke Yong, Wang Luxin, etc Multi factor comprehensive evaluation method based on fuzzy mathematics and grey relational analysis for optimizing well completion

methods [J]. Liaoning Chemical Industry, 2012, 41 (10): 1075-1076+1093.

[6] Zhao Binjuan, Liu Qi, Fu Yanxia, etc Prediction of Vortex Flow at the Boundary of Centrifugal Pump Impeller Based on Artificial Neural Network [J]. Journal of Irrigation and Drainage Machinery Engineering, 2020, 38 (02): 127-132.

[7] Fan Zhenyu BP neural network model and learning algorithm [J]. Software Guide, 2011, 10 (7): 66-68. FAN Zhenyu. BP neural network model and learning algorithm [J]. Software Guide, 2011, 10 (7): 66-68.

[8] Chongchong Qi, Xiaolin Tang. Slope stability prediction using integrated metaheuristic and machine learning approaches: A comparative study[J]. Computers & Industrial Engineering, 2018, 118.

[9] SHIPPEN M E, SCOTT S L. A neural network model forprediction of liquid holdup in two-phase horizontal flow [C]. SPE Annual Technical Conference and Exhibition, 2002.

[10] Tian Rui, Meng Haidong, Chen Shijiang, etc RF-AHP Cloud Model for Predicting Rockburst Intensity Grading [J]. Chinese Journal of Safety Science, 2020, 30 (07): 166-172.

[11] Breiman L. Random Forests[J]. Mach Learn, 2001, 45(01): 5-32.

[12] Teacher Dong, Huang Philosophy Analysis of Random Forest Theory [J]. Ensemble Technology, 2013, 2 (01): 1-7.

[13] Guo Chao, Yang Yuxuan, Hu Ronglei, etc WSN Coverage Optimization Based on Particle Swarm Optimization [J]. Research on Computer Applications,2020,37(4):1170-1173+1182.

[14] Zhou Lijun, Peng Wei, Zou Fang, etc Adaptive Mutation Particle Swarm Optimization Algorithm [J]. Computer Engineering and Applications, 2016, 52 (7): 50-55+149.

[15] Cheng Jiawei Prediction Model and Application of Grain Storage Temperature Based on BP Neural Network [D]. An Xue, 2021. DOI: 10.26917/d.cnki.ganhu.2021.000357.

[16] Yan Qunmin, Dong Xinzhou, Mu Jiahao, Ma Yongxiang. 2022. Optimization of Energy Storage Configuration in Active Distribution Networks Based on Improved Multi Objective Particle Swarm Optimization Algorithm [J]. Power System Protection and Control, 50 (10): 11-19.

[17] Chen Xinguo, Leng Xulin, An Yunpeng, etc Hydraulic Prediction Model for Gas Pipeline Based on Deep Learning Structural Network [J]. Oil and Gas Field Surface Engineering, 2018, 37 (08): 52-57.

[18] Zhang Yu Research on wax deposition prediction algorithm for waxy crude oil pipelines based on machine learning [D]. Northeast Petroleum University, 2019: 36-65.