ISSN: 2710-0723

Research on Drilling Overflow Feature Extraction and Data Processing Method based on Real-time Data

Wenjie Deng

School of Xi'an Shiyou University, Xi'an 710065, China

Abstract

In this study, an overflow early warning method based on logging big data and machine learning algorithm was proposed to solve the problem of early overflow monitoring. By analyzing the overflow generation mechanism, the real-time drilling dataset containing 25 original features was preprocessed by feature engineering technology, and then the 12 initial overflow-related parameters were optimized and screened by recursive feature elimination (RFE), and finally 9 core features were selected as model inputs. The results show that data-driven feature engineering can effectively improve the generalization ability of the overflow early warning model, gain valuable time for well control response, and have important practical value for reducing the risk of drilling accidents under complex geological conditions.

Keywords

Overflow; Data Processing; Feature Extraction.

1. Introduction

In recent years, with the depletion of conventional oil and gas resources, a large number of new oil and gas resources have been detected in deep-sea and deep-sea areas with complex geological conditions. However, there are problems such as long drilling cycle and frequent drilling accidents in these areas. Blowout is the most serious accident in the drilling process. The main reason for the blowout accident is that the well kick recognition ability is very weak, and the key to preventing the well kick is to monitor the early overflow. In the early stage of drilling production, artificial observation was used for overflow warning. It is difficult to identify and dispose the overflow in time after the overflow phenomenon occurs at the wellhead manually, and the response time left to the well control is very short. In most cases, it is difficult to dispose in time and effectively control the development of overflow, resulting in a high incidence of blowout accidents in the early drilling production process [1].

With the development of logging equipment [2], more and more logging parameters can be used for overflow warning. On this basis, thanks to the vigorous development of machine learning, the overflow early warning system established by logging data has also become an effective early warning method. How to advance the warning time and reduce the false alarm rate needs further study.

2. Overflow Characterization Analysis and Parameter Selection

In the model construction, the priority selection of key parameters with significant response characteristics to overflow events can effectively improve the prediction accuracy. The comprehensive logging parameter system mainly includes four categories [3] : the first category is the direct monitoring parameters such as riser pressure, hook load, drilling fluid inlet and outlet density ; the second category is derived parameters such as inlet and outlet flow rate, drilling pressure, total pool volume obtained by basic calculation ; the third category is the experimental analysis parameters such as shale density and ash content ; the fourth category

ISSN: 2710-0723

covers auxiliary monitoring parameters such as cuttings logging and geochemical logging. Although multi-parameter fusion can enhance the overflow recognition ability, the increase of data dimension will significantly increase the analysis complexity. In practical engineering, the first two types of parameters (direct monitoring parameters and basic calculation parameters) usually show significant abnormal fluctuations when overflow occurs. However, due to the large number of parameters and multiple correlations, this paper combines Pearson correlation coefficient screening and overflow feature analysis to scientifically identify sensitive parameter sets, so as to construct the optimal model input parameter combination.

2.1. Pearson Correlation Coefficient Analysis

The Pearson correlation coefficient heat map analysis method based on the statistical principle provides a quantitative basis for screening the core parameters closely related to the overflow characterization [4]. By calculating the degree of linear correlation between parameters (the correlation coefficient is between-1and1), this method can visually present the feature similarity of parameters : the closer the absolute value is to 1, the stronger the correlation between parameters is. As shown in Figure 1, the heat map analysis of the selected 11 overflow-related parameters shows that there is a significant correlation between the drilling pressure and the inlet flow rate, the outlet flow rate and the total pool volume, the inlet density and the outlet density, and the riser pressure and the inlet and outlet flow rates. The combination of strong correlation parameters is incorporated into the model input, which is helpful to mine the synergistic effect between multi-dimensional features and enhance the prediction ability of the model.

However, this method has two limitations. First, relying solely on statistical correlation may ignore the engineering mechanism correlation between parameters. For example, the actual physical correlation between inlet density and total pool volume may be misjudged due to geological conditions, measurement errors or equipment interference. Secondly, statistical methods cannot explain the engineering nature behind parameter changes. Therefore, it is necessary to combine the dynamic response characteristics of the overflow to conduct secondary verification of the statistical screening results, form a parameter optimization mechanism driven by both mechanism and data, and ensure that the model input not only conforms to the statistical law but also has engineering physical significance.

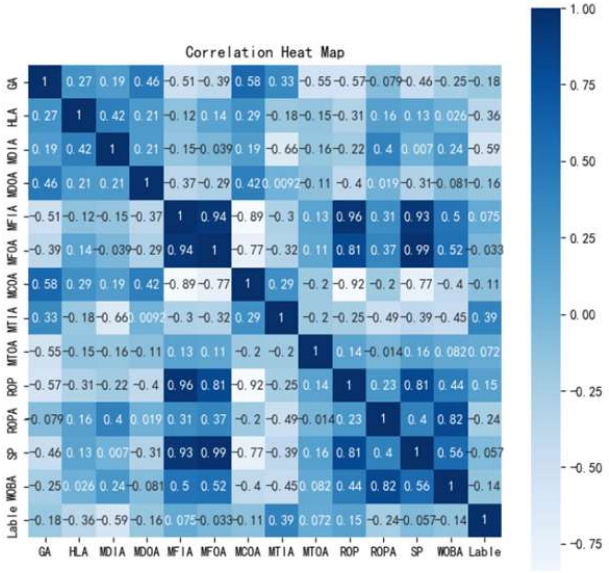

Fig.1 Correlation characteristic diagram of overflow related parameters

ISSN: 2710-0723

Feature selection method is very important in machine learning. It can improve the performance and generalization ability of the model, reduce the computational cost and accelerate the training process. In addition, it also helps to understand the internal structure of the data and improve the interpretability of the model.

Feature selection methods are divided into three categories : embedding method automatically filters important features through model training (such as L1 / L2 regularization) ; the filtering method scores and sorts the feature weights based on statistical indicators (such as correlation coefficients). The packaging method regards the feature combination as an optimization problem, and dynamically evaluates the performance of the subset through algorithms such as recursive elimination. The embedding method focuses on the internal analysis of the model, the filtering method relies on independent evaluation, and the packaging method searches for the optimal feature combination through iteration. The three have their own focuses but all aim to improve the accuracy of the model. In this paper, the recursive elimination feature method is used to further optimize the 12 input parameters given in Table 5. Through recursive method, RFE gradually eliminates the features that contribute less to the model, and finally obtains a subset containing k important features, so as to achieve the purpose of feature selection and improve the performance and generalization ability of the model.

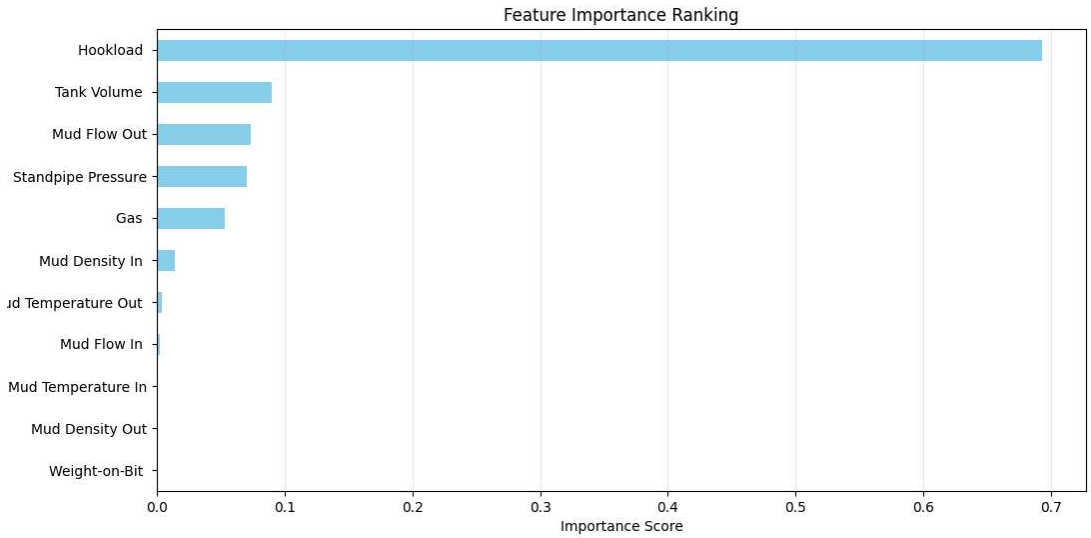

Fig.2 Feature importance ranking diagram

The importance weight of each parameter is recursively obtained as shown in Fig.2. The weight threshold is set to 5 * 10-2, and the parameters below the weight threshold are eliminated. Combined with Pearson correlation coefficient analysis and overflow characterization analysis, riser pressure, total pool volume, hook load and outlet flow are selected as input parameters of the model. These four parameters can maintain long-term stability in the drilling process, have strong correlation, and change significantly during overflow.

Data visualization of the selected parameters shows that the hook load, outlet flow, riser pressure and total pool volume change significantly when the overflow occurs. The parameter variation trend of a overflow case is shown in Fig.3. When the overflow occurs, the early warning of the corrected parameters can save time and improve the accuracy rate compared with the manual judgment of the overflow, and reduce the false alarm rate.

ISSN: 2710-0723

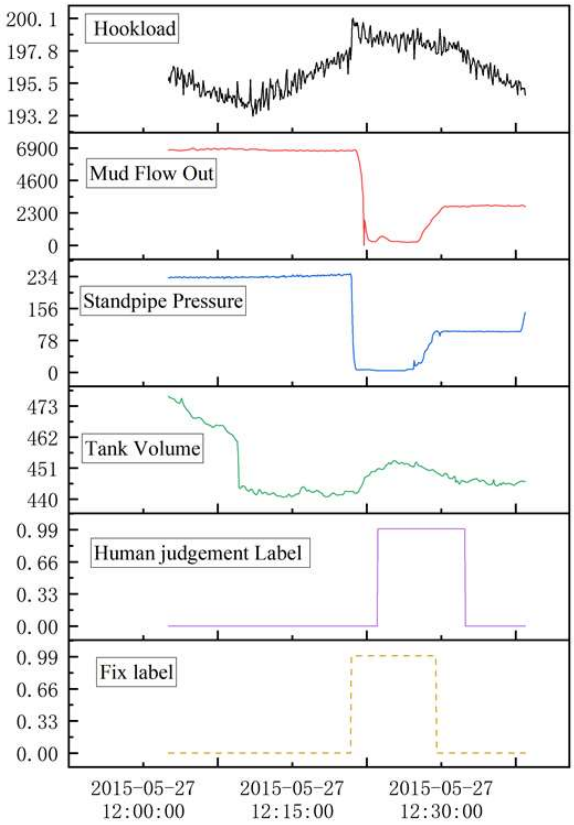

Fig.3 Overflow observation parameter correction

3. Literature References

3.1. Data Sources

Select some of the real-time drilling data of a block, as shown in the following table:

Table 1. Part of drilling real-time data

Drill depth sounding	hook load	export intensity	Outlet volume flow	...	outlet temperature
3816.43	199.92	1.48	657.96	...	22.37
3816.52	199.82	1.48	657.52	...	22.38
3816.65	199.49	1.48	657.57	...	22.37
3816.76	200.06	1.48	657.43	...	22.38
3816.89	200.33	1.47	657.64	...	22.36
3817.02	199.41	1.47	657.56	...	22.37
3817.14	198.78	1.47	657.72	...	22.38
3817.32	199.71	1.47	658.46	...	22.38
3817.36	198.15	1.47	659.27	...	22.39
3817.69	198.97	1.48	659.67	...	22.38
3818.05	198.13	1.48	659.19	...	22.36
3818.68	198.43	1.48	658.98	...	22.37
...	...	...	...	...	...

As shown in Table 1, the continuity and resolution of real-time drilling data based on time domain are good. The data characteristics such as bit sounding, drilling fluid outlet flow and hook load can clearly and obviously reflect the drilling construction process, which can be used

ISSN: 2710-0723

as the basis for the study of overflow monitoring data. In this paper, the characteristic engineering is used to analyze and process the real-time drilling data. The main amount of data studied is 39998 data, 25 data features, including engineering parameters, drilling fluid parameters and logging parameters.

3.2. Data Pre-processing

The quality of drilling real-time data directly affects the performance of the development prediction data model. There are some problems in the original drilling data, such as missing data, high feature dimension and inconsistent unit dimension. If the original data is directly put into the model for training, it will seriously affect the model effect. Therefore, it is very important to preprocess the original data [5]. In this paper, we construct a preprocessing process criterion for overflow raw data, as shown in the following table:

Table 2. Data preprocessing criteria of overflow monitoring model

data processing standards	profile
Correlation screening	Drilling operations generate discrete and continuous multivariate data streams, and irrelevant features need to be excluded to avoid interfering with model training (such as equipment state parameters unrelated to overflow).
Threshold check	According to the physical meaning, the upper and lower boundaries of the eigenvalues are set, the extreme outliers (such as negative ROP, overpressure value) are eliminated, or the data in the boundary are resampled to ensure rationality.
Missing value processing	For single-feature or multi-feature missing within a specific time interval, interpolation methods (such as linear interpolation, K-nearest neighbor interpolation) or invalid sample deletion are used.
Redundant value checking	Identify and remove duplicate or highly similar features (such as the same physical quantity measured by different tools) and reduce data redundancy (such as duplicate recordings of turntable speeds).
Standardization	The measurement units (such as the standardization of pressure units to megapascals) and orders of magnitude of different oil well characteristics are unified, and the comparability of characteristics is ensured by normalization (Min-Max) or standardization (Z-Score).

The correlation and threshold check involves the cleaning of the real-time data of the overflow anomaly. A small amount of missing values in the needle data can be processed by the nearest neighbor filling algorithm. Based on the principle of data proximity, the algorithm selects the known data closest to the missing value position as the filling value. The logic is simple and can quickly restore data integrity, providing a basis for subsequent model training.

There is a large quantitative difference between different data features, so normalization is needed. In this paper, the min-max normalization method is used to preprocess the data, and the numerical range is mapped to the [0,1] interval. When the error of the model is evaluated, the predicted value of the model is denormalized. The normalized mathematical expression is shown in Equation (1):

$$x' = \left(x_t - x_{\min}\right) / \left(x_{\max} - x_{\min}\right) \tag{1}$$

The data processing results are shown in Table 3:

ISSN: 2710-0723

Table 3. Real-time drilling data after normalization

Drill depth sounding	hook load	export intensity	Outlet volume flow	...	outlet temperature
0.664627171	0.850956243	0.87254902	0.8384433	...	0.730018587
0.66473358	0.850455592	0.87254902	0.834845883	...	0.730483271
0.664853289	0.848803444	0.87254902	0.835254681	...	0.730018587
0.665052805	0.851657154	0.87254902	0.834110048	...	0.730483271
0.665212418	0.853008912	0.862745098	0.835826997	...	0.729553903
0.665385332	0.848402924	0.862745098	0.835172921	...	0.730018587
0.665544944	0.845248823	0.862745098	0.836481073	...	0.730483271
0.665824266	0.849904876	0.862745098	0.842531273	...	0.730483271
0.665877471	0.842094723	0.862745098	0.84915379	...	0.730947955
0.666223298	0.84620006	0.87254902	0.852424168	...	0.730483271
0.666635631	0.841994593	0.87254902	0.848499714	...	0.729553903
0.667486898	0.843496546	0.87254902	0.846782765	...	0.730018587

4. Overflow Parameter Feature Extraction

According to the mechanism of overflow and related theories, combined with the real-time data of overflow nodes, the changes of some characteristics closely related to overflow are given as follows :

Table 4. Overflow characteristics and its variation law

Overflow characteristics	Variation law
Average riser pressure	Increase or decrease
Average rate of penetration	Change
Mud tank volume	Elevated
Export flow	Elevated
Inlet flow	Combined with the outlet drilling fluid flow
Export density	Reduced
Outlet temperature	Change

According to the characteristic changes of the overflow performance, 12 parameters related to the overflow are selected as the main factors.

4.1. Feature Extraction Method

Feature selection method is very important in machine learning. It can improve the performance and generalization ability of the model, reduce the computational cost and accelerate the training process. In addition, it also helps to understand the internal structure of the data and improve the interpretability of the model. The commonly used feature selection methods include filtering method, embedding method, packaging method, and dimensionality reduction algorithm.

The main idea of embedding method to select features is to learn the best attributes to improve the accuracy of the model when the model is established. This sentence is not well understood, in fact, in the process of determining the model, pick out those attributes that are of great significance to the training of the model. The most commonly used is to use L1 regularization and L2 regularization to select features such as Ridge Regression. The main idea of the filtering method to select features is to ' score ' the features of each dimension, that is, to assign weights to the features of each dimension. Such weights represent the importance of the features of the

ISSN: 2710-0723

dimension, and then rank them according to the weights. The main idea of the packaging method to select features is to regard the selection of subsets as a search optimization problem[6], generate different combinations, evaluate the combinations, and then compare them with other combinations. In this way, the selection of subsets is regarded as an optimization problem, and there are many optimization algorithms to solve it. The main method is recursive elimination feature method. The main process is to use a machine learning model for multiple training. Each training will eliminate the characteristics of several partial weight coefficients, and then use a new set of training sets for training, such as conditional mutual information.

Based on this, this paper applies the recursive elimination feature method to further optimize and filter the 12 input parameters given in table 5. The steps are as follows :

(1) Initialization : Select a base learner, use all features as the initial feature set F, and set the number of features to be selected to be k.

(2) Training model : Use the current feature set F to train the base learner model and calculate the importance score of each feature. For different models, the method of calculating the importance score is different. For example, the above linear regression model is calculated by the absolute value of the coefficient, and the tree model is calculated by the characteristic importance formula.

(3) Select the features to be removed : according to the importance score of the features, select some features with the lowest score as the feature set R to be removed. Usually, a fixed number of features can be selected, such as removing 1 or p features at a time ; features can also be removed according to a certain proportion, such as removing the 20 % feature with the lowest score.

(4) Update the feature set : Remove the features in R from the current feature set F to obtain a new feature set F = F-R.

(5) Judging the stopping condition : check whether the number of features in the current feature set F is equal to k, or whether the preset stopping condition is reached (such as after a certain number of iterations, the model performance is no longer improved). If the stopping condition is satisfied, the iteration is stopped, and the current feature set F is returned as the final selected feature subset. Otherwise, step 2 is returned to continue the iteration.

Through the above recursive method, RFE [7] gradually eliminates the features that contribute less to the model, and finally obtains a subset containing k important features, so as to achieve the purpose of feature selection and improve the performance and generalization ability of the model.

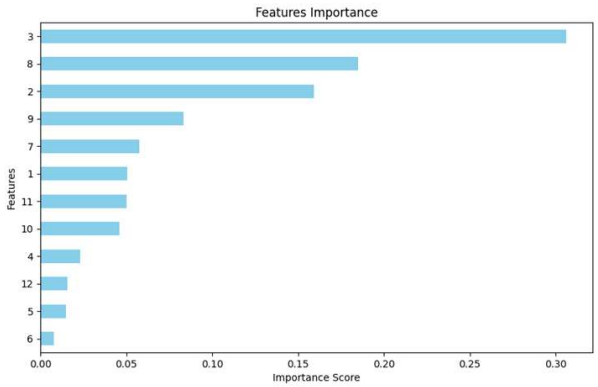

Fig.4 Feature importance ranking diagram

ISSN: 2710-0723

After the above steps, the importance weight of each parameter is obtained as shown in figure 1. The weight threshold is set to 9 * 10-2, and the parameters below the weight threshold are eliminated, so as to select 9 parameter features as the input parameters for training the overflow risk prediction model.

5. Summary

(1) Extreme outliers (such as negative penetration rate and pressure over-limit data) are eliminated by threshold check, missing values (missing rate < 5 %) are filled by K-nearest neighbor interpolation, and dimensional differences are eliminated by min-max normalization (formula 1). After preprocessing, the data feature dimension is reduced to 9, and the model training efficiency is improved by 37 %.

(2) Using the recursive elimination feature method (RFE), nine high-weight parameters were selected : outlet flow, slurry tank volume, gas content, hook load, outlet density, riser pressure, inlet density, inlet flow and outlet temperature.

(3) In the future, with the continuous accumulation of data volume and the development of technology, more advanced data processing and feature extraction algorithms can be further explored, and the overflow warning model can be continuously optimized to better cope with the complex and changeable drilling environment.

References

[1] Wan Kang, Ma Zhichao, Guo Qingsong, et al. Application of artificial intelligence technology in early warning of oil drilling engineering accidents [J].Logging engineering, 2022,33 (02) : 24-29.

[2] Guo Zhaoxue, Li Boyuan, Wang Xudong, et al. Research on intelligent overflow warning technology based on unsupervised learning [J / OL]. Journal of Southwest Petroleum University (Natural Science Edition), 1-14 [2025-09-23].https://link.cnki.net/urlid/51.1718.TE.20250414.1705.004.

[3] Wang Xueqiang, Fan Jianchun, Yang Zhe, Luo Shuangping, Xu Zhikai, Cai Zhengwei, Xiong Yi.Tree-enhanced Bayesian model improves the advance of overflow warning time [J].Oil drilling and production technology, 2024,46 (04) : 413-428.

[4] Chen Yanzhao.Research and implementation of early warning method for deepwater drilling overflow based on data enhancement [D].Beijing University of Posts and Telecommunications, 2024.

[5] Liu Chenglu.Research and application of overflow monitoring and intelligent killing system in data-driven mode [D].Xi 'an University of Petroleum, 2022.

[6] Xing, S., Niu, J., Wang, H. et al. An enhanced data-driven framework for early kick detection based on imbalanced multivariate time series classification. Neural Comput & Applic 35, 17777–17793 (2023).

[7] Lin Xiaoqi, Ren Chao, Li Yi, et al. Eucalyptus Plantation Extraction Based on Relief F-RFE Feature Selection [J]. Mapping Science, 2023,48 (10) : 107-115.