Real-time monitoring and early warning of well leakage based on big data analysis

Xu Yingzhuo1,a, Wang Liupeng2,b, Zhang Yang3,c, Li MingZhen4,d

1Xi’an Shiyou University, Xi’an Shiyou University, Yanta District, Xi’an City, Shaanxi Province, China, 86-13096949510
2Xi’an Shiyou University, Xi’an Shiyou University, Yanta District, Xi’an City, Shaanxi Province, China, 86-15829913600
3Xi’an Shiyou University, Xi’an Shiyou University, Yanta District, Xi’an City, Shaanxi Province, China, 86-15398010223
4Xi’an Shiyou University, Xi’an Shiyou University, Yanta District, Xi’an City, Shaanxi Province, China, 86-18892089805

a707416631@qq.com
b376338063@qq.com
c957278107@qq.com
d1617676483@qq.com

Abstract—In petroleum drilling engineering operations, complex lost circulation accidents not only affect the efficiency of drilling operations, but serious lost circulation may cause wellbore failure. The occurrence of lost circulation accidents is affected by various factors such as formation conditions, engineering parameters, and operating dynamic parameters. The conventional focus of lost circulation research is to analyze the mechanism of lost circulation, but the data such as engineering parameters and operating dynamic parameters when lost circulation occurs are insufficiently utilized. At present, the field of drilling engineering has accumulated a large amount of historical drilling data, and the complicated occurrence of lost circulation accidents often has regional statistics. How to use big data technology to establish the correlation model between lost circulation and parameters, reason and analyze the main influencing factors of lost circulation, then use the established correlation model comparative analysis. After the data model comparative analysis, it can make different specifications of responses and early warnings for data anomalies with different weights, so as to reduce the impact on drilling operations while avoiding lost circulation accidents.

1. Introduction
Drilling engineering is an extremely complex underground engineering. During the drilling process, the pipe string system, geological stratigraphic system, and drilling fluid system interact with each other in the wellbore. Various factors in the horizontal direction cross-effect, and each point in the vertical direction is different [1- 2]. At present, the monitoring and analysis of drilling technology carried out within the scope of engineering practice is mainly based on ground parameters from the perspective of data sources, and based on mechanism research from the perspective of analysis methods. In this case,
the objects considered and analyzed are generally adopted. The boundary conditions are simplified and idealized, and data such as engineering parameters and operating dynamic parameters when lost circulation occurs are insufficiently utilized. Existing drilling fault identification and early warning are mainly performed by detecting the fluctuations of specific drilling parameters collected in real time, relying on manual analysis methods, and often relying too much on the abnormal fluctuations of a single or several parameters. However, since drilling failures often have complex response relationships on multiple drilling parameters, simple manual analysis alone cannot guarantee the accuracy of drilling failure identification and early warning. In response to this situation, this article uses data mining technology to perform multi-dimensional fitting analysis on these data based on real drilling fault data, block geological data, and mechanical parameter data, so as to identify and early warning through these analysis data, formatter will need to create these components, incorporating the applicable criteria that follow.

2. clustering algorithm and iterative clustering algorithm
Clustering algorithm and iterative clustering algorithm

Clustering is a concept in data mining, which is to divide a data set into different classes or clusters according to a certain standard (such as distance), so that the similarity of data objects in the same cluster is as large as possible, and they are not the same. The diversity of the data objects in the cluster is also as large as possible. That is, after clustering, the data of the same type are gathered together as much as possible, and the data of different types are separated as far as possible [3].

2.1. k-means clustering algorithm
The idea of K-Means algorithm is very simple. For a given sample set, the sample set is divided into K clusters according to the distance between the samples. Make the points in the clusters as close together as possible, and make the distance between the clusters as large as possible.

If expressed by a data expression, assuming that the cluster is divided into (C1, C2,...Ck), then our goal is to minimize the square error:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \| x_i - u_k \|^2$$

Where indicates that the sample is divided into cluster class k, it is 1, otherwise it is 0.

$$r_{ik} = \begin{cases} 1, & x_i \in k \\ 0, & x_i \notin k \end{cases}$$

(2)

Represents the mean vector of cluster class k.

The objective function (1) describes to a certain extent how closely the samples in the cluster surround the cluster mean vector. The smaller the value, the higher the similarity of the samples in the cluster [4]. Minimizing the objective function is an NP (non-deterministic polynomial problem) problem, and k-means clustering uses the EM algorithm (maximum expectation algorithm) to optimize the model.

(1) Initialize the mean vector of K clusters, that is, is a constant, and find when J is minimized. It is not difficult to know that when the data point is divided into the cluster class closest to the data point, the objective function J is the smallest.

(2) It is known that when seeking to minimize J, the corresponding. Let the partial derivative of the objective function J equal to 0:

$$\frac{\partial J}{\partial u_k} = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} (x_i - u_k)$$

$$= \sum_{i=1}^{N} r_{ik} (x_i - u_k) = 0$$

(4)
Nik i

\[ u_k = \sum_{i=1}^{N} \frac{r_{ik} x_i}{\sum_{i=1}^{N} r_{ik}} \]  

(5)

Uk meaning of the expression is that the center of the cluster is equal to the mean of the samples of the cluster.

2.2. Fast iterative clustering algorithm

Before analyzing fast iterative clustering, let's first understand the spectral clustering algorithm. The spectral clustering algorithm is an algorithm based on the spectrogram theory. Compared with the traditional clustering algorithm, it can cluster in a sample space of any shape and converge to the global optimal solution. The main idea of the spectral clustering algorithm is to transform the clustering problem into an undirected graph partition problem.

First, the data point is regarded as the vertex v of a graph, and the similarity of the two data is regarded as the edge of the graph. The set of edges is determined by \( E = Ai \). Indicates that the similarity matrix A of the sample data set is constructed from this, and the Laplacian matrix L is obtained. Secondly, according to the division criteria, make the internal similarity of the subgraphs as large as possible and the similarity between the subgraphs as small as possible, and calculate the eigenvalues and eigenvectors of L.

Finally, select k different feature vectors to cluster the data points.

So how to find the Laplace matrix?

The degree of the vertex can be obtained by adding the elements of each row of the similarity matrix A. We define the diagonal matrix with degree as the diagonal element as the degree matrix D. The Laplacian matrix can be determined by A and D. Laplacian matrices are divided into two types: normative and non-canonical. The normative Laplacian matrix is expressed as L=D-A, and the non-canonical Laplacian matrix is expressed as \( L = I - D^{-1}A \). Compared with traditional clustering methods (such as K-means), the spectral clustering algorithm has many advantages: Similar to K-medoids, spectral clustering only needs the similarity matrix between data, instead of requiring the data to be a vector in N-dimensional Euclidean space like K-means.

Because it captures the main contradiction and ignores the minor things, it is more robust than traditional clustering algorithms, is not so sensitive to irregular error data, and has better performance.

The computational complexity is smaller than that of K-means, especially when running on data with very high dimensions like text data or ordinary image data. Both the fast iterative algorithm and the spectral clustering algorithm embed data points in the low-dimensional subspace derived from the similarity matrix, and then directly or through the k-means algorithm to generate the clustering results, but the fast iterative algorithm has different places. The following focuses on understanding the principles of fast iterative algorithms. In the fast iterative algorithm, we construct another matrix \( w = D^{-1}A \), Compared with the first chapter, we can know that the maximum eigenvector of W is the minimum eigenvector of the Laplacian matrix L. We know that the Laplacian matrix has a characteristic: the second smallest eigenvector (the eigenvector corresponding to the second smallest eigenvalue) defines a solution for the optimal partition of the graph, which can approximate the maximum partition criterion. More generally, the subspace defined by the k smallest eigenvectors is suitable for dividing the graph. Therefore, the eigenvectors of the second smallest, third smallest and k-th smallest Laplacian matrix can well divide the graph W into k parts. Note that the k smallest eigenvectors of the matrix L are also the k largest eigenvectors of the matrix W. Calculating the largest eigenvector of a matrix can be obtained by a simple method, which is fast iteration. PI is an iterative method. It starts with an arbitrary vector and updates it cyclically according to the following formula. \( vt + 1 = C wr^t \). In the above formula, c is a standardized constant to avoid vt Produces too large value, here \( C = ||w^r||_1 \). In most cases, we only care about the k-th feature vector, and not the largest feature vector. This is because the largest feature vector is a constant vector: because the sum of each row of W is 1.
3. CLUSTER PREDICTION MODELING

3.1. Parameter selection of lost circulation characteristics

Lost circulation is an accumulative process. Before a complicated lost circulation accident is discovered, it can be accurately warned. First, it needs to be accurately selected from logging parameters, LWD (logging while drilling) parameters, and MWD (measurement while drilling) parameters. Characteristic parameters for predicting the occurrence of lost circulation. At the same time, the geological parameters around the well are measured to assist in judging the possibility of lost circulation time.

The characteristic parameters should have the following characteristics[6]:

1. Have a stable trend of change before the lost circulation occurs to ensure the accuracy of the prediction results.
2. Continuous changes before the lost circulation occurs to ensure the timeliness of the early warning of lost circulation accidents.

| Characteristic parameters |
|---------------------------|
| Total pool volume         |
| Outlet flow               |
| Inlet flow                |
| Standpipe pressure        |
| Drilling time (drilling speed) |
| Circulating pool lift     |
| Gas measurement total hydrocarbon |
| Hook hanging weight       |
| Geological permeability coefficient |

3.2. Establishment of lost circulation model

Considering the uncertainty of drilling parameter changes when an accident occurs, an early warning model for lost circulation accidents is established based on the clustering algorithm. In order to effectively extract the characteristics of drilling parameter changes from actual drilling data, update the status of the nodes of the clustering algorithm, comprehensively use methods such as normalization and data dimensionality reduction, and formulate rules to judge the status of nodes for the built model Early warning of leakage accidents.

3.3. Data preprocessing

Drilling data monitoring indicators are numerous and highly correlated, which is conducive to correlation analysis and model construction in data analysis. Perform data analysis based on the acquired original drilling lost circulation data. The original data can be divided into two categories, geological data and mechanical data. According to the mechanism of lost circulation, the correlation between lost circulation and mechanical data is relatively small, so the main research is on the influence of geological data on lost circulation.

When lost circulation occurs, the key consideration is the change trend of the lost circulation characteristic parameter rather than the direct value of the characteristic parameter, so the difference in a period of time is used as the training parameter of the lost circulation warning model.

In order to prevent the dimension and value range of different field parameters from affecting the performance of the model results, it is necessary to use a unified unit to convert the data. For example, the units of different characteristic parameters of the drilling lost circulation data are different, and the data has obvious quantities. In order to solve the problem of outline, normalization [7] is used to uniformly transform the data, so that the drilling parameters are mapped to [0,1], as shown in formula (1):

\[ x_{\text{new}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]  

Where: is the normalized data; is the original data; and is the maximum and minimum values of each field in the original data set.
Where: $\bar{d}$ is the normalized data; $d$ is the original data; and $\max$ and $\min$ are the maximum and minimum values of each field in the original data set.

3.4. Removal of data abnormalities
Outliers, also called outliers. As the name implies, it refers to the point far away from the larger sample, the sparser point, generally the maximum or minimum. In the process of solving general problems in machine learning, these abnormal points will affect the fitting effect of the model to normal data, so they should be removed before training the model.

According to the definition of normal distribution, the probability that a data point falls within minus 1 standard deviation of the deviation from the mean is 68.2%; the probability that the data point falls within minus 2 standard deviations of the deviation from the mean is 95.4%; data The probability that a point falls within plus or minus 3 standard deviations from the mean is 99.6%.

Therefore, consider the probability value mentioned above from another perspective. If the probability of a data point falling outside of the mean value by 2 times the standard deviation is less than 5%, it is a small probability event, that is, such a data point is considered an abnormal point. Similarly, if the data points fall outside of the mean value by 3 times the standard deviation, the probability will be smaller, and these data points can be considered as extreme abnormal points. In order to make the reader intuitively understand the probability values mentioned in the article, you can view the probability density diagram of the standard normal distribution, as shown in the following figure:

![Anomalous data distribution](image)

Based on the conclusion of the above figure, the judgment conditions for abnormal points can be obtained, as shown in the following table:

| Judgment criteria                                      | conclusion            |
|-------------------------------------------------------|-----------------------|
| $x > \bar{x} + 2\sigma$ or $x < \bar{x} - 2\sigma$   | Outlier               |
| $x > \bar{x} + 3\sigma$ or $x < \bar{x} - 3\sigma$   | Extreme anomaly       |

In Table 2, $\bar{x}$ is the current data, $\bar{d}$ is the mean of the set of data, and $\sigma$ is the standard deviation of the set of data.

4. REAL-TIME DATA TO MODEL EVALUATION
Perform leakage accident prediction on the trained model, and record the accident probability vector at each time point in the output test set as $P$. Taking into account that lost circulation is a cumulative type of accident, early warning based only on the output of a single point in time will cause a greater false alarm rate. Therefore, in order to reduce the influence weight of the accident probability at a single time point, the cumulative decay probability value $\text{cumP}$ is constructed, which is defined as:

$$\text{cumP}_t = \text{cumP}_{t-1} + P_t - 0.5$$
It can be seen from Figure 2 that 100 sets of data (including 50 sets of lost circulation data and 50 sets of normal drilling data) are divided into two categories through clustering. The blue part indicates lost circulation, and the red part indicates normal drilling.

| project                  | Actual value | Predictive value | Number of false positives |
|--------------------------|--------------|------------------|----------------------------|
| Lost circulation accident| 50           | 56               | 6                          |
| Normal drilling          | 50           | 46               | 6                          |

Table 3  Performance prediction of lost circulation warning

It can be seen from Table 1 that the false alarm rate of accidents reaches 12%. Too many false alarms will affect the normal operation of the driller and reduce the drilling efficiency. This model can achieve effective early warning, but the false alarm rate is still relatively large. Substituting the drilling operation data of lost circulation into the prediction model again, using an iterative clustering algorithm to make the algorithm more data samples, and then iterative training.

It can be seen from Figure 3 that the 100 sets of data (including 50 sets of lost circulation data and 50 sets of normal drilling data) are again divided into two categories through clustering. The blue part indicates lost circulation, and the red part indicates normal drilling.

| project                  | Actual value | Predictive value | Number of false positives |
|--------------------------|--------------|------------------|----------------------------|
| Lost circulation accident| 50           | 53               | 3                          |
| Normal drilling          | 50           | 48               | 2                          |

Table 4  Performance prediction of lost circulation warning

From Table 4, the false alarm rate of accidents reaches 5%. After the iterative clustering algorithm, the false alarm rate of accidents is significantly reduced. The model is more realistic, and the early warning effect of actual drilling operations is more accurate.
5. CONCLUSION
This article analyzes the causes and symptoms of lost circulation, multi-dimensional selection of engineering parameters, geological parameters for correlation analysis, expounds the idea of clustering algorithm, and in-depth explanation of the superiority of iterative clustering algorithm in lost circulation prediction. The iterative clustering algorithm is constantly adding data and constantly making model modifications to make the model more realistic. Finally, a certain lost circulation data set is selected as the experimental object, and the data is analyzed and compared with the common cluster classification results, which fully demonstrates that the iterative clustering algorithm is more reliable in predicting lost circulation.

Acknowledgment
We are grateful to the Natural Science Basic Research Project of Shaanxi Province for their financial supported under Grant No. 2019JM-383 for this paper.

References
[1] He Pengfei, Wang Wei. Research and Application of Drilling Engineering Assistant Decision Method Based on Data Analysis[J]. Technical Supervision in Petroleum Industry, 2018(4): 1-4.
[2] Liu Baosheng, He Pengfei, Yang Baqian, et al. Construction of Drilling Assistant Decision System and Engineering Practice in Bozhong 19-6[J]. Petroleum Drilling and Production Technology, 2018(6): 684-689.
[3] Zhang Yonglai, Zhou Yaojian. Summary of clustering algorithms[J].Computer Applications, 2019,39(0.7) : 1869-1882
[4] Shao Dongheng, Yang Wenyuan, Zhao Hong. Applying k-means algorithm to realize label distribution learning[J].Journal of Intelligent Systems,2017,12(03):325-332.
[5] Quan Yinghui, Gao Xia, Sha Minghui, et al. High-speed multi-target parameter estimation of frequency agile combined orthogonal frequency division multiplexing radar based on expectation maximization algorithm[J].Journal of Electronics & Information Technology, 2020,42(07): 1611-1618.
[6] Yin Hu, Wang Haibiao. Research on Intelligent Early Warning Method of Lost Loss Complicated Accident Based on CBR[J].Bulletin of Science and Technology,2018,34(04):195-199.
[7] Tu Xiyu,Yu Lu,Geng Zichen,Xue Zhi,Zhang Baowen.A method for early warning of leakage accidents based on large-scale time series[J].Information Technology,2018,42(12):1-4.