Cluster Discriminant Prediction of Oil Well Production Based on Mahalanobis Distance

Bo Huang¹, Yunwei Kang² *, Ang Chen¹, Zixi Guo³, Yicheng Sun¹

¹ Engineering Technology Research Institute, Xinjiang Oilfield Company, 834000, China
² School of Sciences, Southwest Petroleum University, Chengdu, Sichuan, 610500, China
³ State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Southwest Petroleum University, Chengdu, Sichuan, 610500, China

*Corresponding author’s e-mail: kangyunweiSWPU@163.com

Abstract. The cost of oil field drilling engineering is very high. Effective well selection and formation selection will avoid waste of energy and resources. Aiming at the problem of well selection and reservoir selection in oilfield, this paper presents a clustering discriminant prediction method for oil well production based on Mahalanobis distance. Based on the field data of Xinjiang oilfield, the main control factors with greater influence are selected after data pretreatment in this paper. Then we randomly divide the data into training data set and prediction data set, and establish a discriminant model of oil well classification based on Mahalanobis distance. Finally, the model is used to predict the production of oil wells in the forecasting data set, and the prediction accuracy reaches 87.5%.

1. Introduction

The development of petroleum industry has promoted the progress of human society and the development of human civilization. Petroleum has been playing an increasingly important role in China’s industry [1] since it was exploited underground in the 19th century. It can be said that oil has become a very important strategic material in the world today. However, oil production is always limited, so how to exploit and utilize oil resources reasonably is a practical problem worth thinking about.

To increase production of petroleum, hydraulic fracturing technology [2] is a common measure at home and abroad. Well selection and reservoir selection is the precondition of hydraulic fracturing, so how to select well and layer greatly affects the production of oil wells. In recent years, due to the development of cloud computing and big data technology, the object studied has a strong complexity. We can not accurately classify and judge by experience and professional knowledge alone. Cluster analysis [3] is thus generated and gradually accepted by people. The parameters obtained by fracturing have strong randomness and fuzziness, which makes the selection of wells and layers more uncertain. We can classify existing wells by clustering discriminant analysis, and classify new wells according to fracturing parameters. The reasonable classification of oil wells by cluster analysis can provide a correct method for well selection and formation selection, thus improving production efficiency.

In this paper, clustering discrimination is applied to oil well production prediction, which improves the prediction accuracy to a certain extent and provides a basis for well selection and reservoir...
selection. Firstly, the field data of oil wells are optimized, and then a clustering discriminant model for low production of oil wells is established. Finally, the model is applied to predict the production of oil wells, which provides guidance for well selection in engineering.

2. 3σ Criterion
The 3σ criterion is also called the Laida criterion [4]. Firstly, it assumes that the data in the sample data have only random errors. After processing, it gets standard deviation and finds an interval according to a certain probability. It holds that as long as the error exceeds the interval, it is not regarded as random error but as wrong data, and that the data beyond the interval should be deleted. 3σ criterion is also applicable when there are many data groups. If the data obey the normal distribution, under the 3σ criterion, the outlier is defined as a set of values whose deviation from the mean exceeds three times the standard deviation. 3σ is a scale, we can also take 4σ, 5σ and so on. Under the assumption of normal distribution, the probability of data appearing outside the mean value of 3σ is \( p(|x - \mu| > 3\sigma) \leq 0.003 \). We can think of these minimal probability events as erroneous data. If the data do not obey the normal distribution, it can also be described by how many times the standard deviation is far from the average value.

3. Correlation Analysis
Correlation analysis [5] refers to the analysis of different correlated parameter statistics, in order to analyze the close impact of the two parameter statistics. There must be some influence or probability between the parameters before correlation analysis can be carried out. Single correlation analysis is to analyze the linear correlation degree of two variables. The statistical measure used in this method is single correlation coefficient, which is called correlation coefficient for short. The general correlation coefficient is a kind of statistic which shows the linear influence between two variables. It is expressed as a constant. The general correlation coefficient is defined as follows:

\[
\gamma = \frac{\text{Cov}(x,y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}
\]

(1)

Where \( \text{Var}(x) \) and \( \text{Var}(y) \) are the variances of parameters \( X \) and \( Y \), \( \text{Cov}(x,y) \) is the covariance of parameters \( X \) and \( Y \).

Since the total values of the total variables \( X \) and \( Y \) can not be observed in practice, the total correlation coefficient is generally unknown. It is usually necessary to randomly extract a certain number of samples from the population and estimate the sample correlation coefficients through the sample observations of \( X \) and \( Y \). The definition formula of sample correlation coefficient is as follows:

\[
\gamma = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}
\]

(2)

Sample correlation coefficient is calculated on the basis of sample statistics. If the sample statistics are different, the calculated values will be different. Sample correlation coefficient is a consistent estimator of population correlation coefficient.

4. Mahalanobis Distance
Mahalanobis distance is proposed by Mahalanobis [6], an Indian statistician. It has statistical significance. Mahalanobis distance is commonly used in distance discriminant analysis. There are several representations as follows:

1. Mahalanobis distance between two vectors of the same population
Suppose there are m-dimensional vectors \( x = (x_1, x_2, \cdots, x_m)^T, y = (y_1, y_2, \cdots, y_m)^T \):

\[
d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}
\]

(3)

The above formula is called the Mahalanobis distance between m-dimensional vectors \( x \) and \( y \). Among them, \( \Sigma \) is the global covariance matrix, which is generally defined as a real symmetric
positive definite matrix. When $\Sigma$ is a unit matrix, the Mahalanobis distance can be regarded as the Euclidean distance.

(2) Mahalanobis distance from one vector to another population
Assuming that the mean of population $G$ is $\mu$ and the covariance matrix is $\Sigma$:

$$d(x, G) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$  \hspace{1cm} (4)

The above formula is called the Mahalanobis distance between the m-dimensional vector $x$ and the population $G$.

(3) Mahalanobis distance between two populations
Assuming that there are two population $G_1$ and $G_2$, their mean vectors are $\mu_1$ and $\mu_2$, and their covariance matrices are equal:

$$d(G_1, G_2) = \sqrt{(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)}$$  \hspace{1cm} (5)

The above formula is called the Mahalanobis distance between two populations $G_1$ and $G_2$.

Because the size of the Euclidean distance is related to the index of the parameter, it is usually used when the component of the vector is different in dimension. We do not use Euclidean distance to do discriminant analysis, and choose Mahalanobis distance which is not required for vector dimension to do discriminant analysis. The specific methods are as follows:

Assuming that there are two population $G_1$ and $G_2$, their mean vectors are $\mu_1$ and $\mu_2$, and their covariance matrices are $\Sigma_1$ and $\Sigma_2$. We take a sample at random and the measured parameter is $x = (x_1, x_2, \cdots, x_q)^T$, the criterion is as follows:

$$d(x, G_i) = \sqrt{(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)}$$  \hspace{1cm} (6)

Since the Mahalanobis distance is related to the covariance matrix of the population, we need to consider whether the covariance matrix of the two populations is equal when we use the Mahalanobis distance for discriminant analysis.

(1) When the covariance matrix of $G_1$ and $G_2$ are equal

$$d^2(x, G_2) - d^2(x, G_1) = (x - \mu_2)^T \Sigma_1^{-1} (x - \mu_2) - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)$$

$$= 2x^T \Sigma_1^{-1} (\mu_1 - \mu_2) - (\mu_1 + \mu_2)^T \Sigma_1^{-1} (\mu_1 - \mu_2)$$

$$= 2 \left[ x - \frac{1}{2} (\mu_1 + \mu_2) \right]^T \Sigma_1^{-1} (\mu_1 - \mu_2)$$

$$= 2 (x - \bar{\mu}) \Sigma_1^{-1} (\mu_1 - \mu_2)$$  \hspace{1cm} (7)

Let $W = (x - \bar{\mu})^T \Sigma_1^{-1} (\mu_1 - \mu_2)$, and the distance criterion is expressed as follows:

$$W(x) > 0, \quad x \in G_1$$

$$W(x) < 0, \quad x \in G_2$$

$$W(x) = 0, \quad \text{undetermined}$$  \hspace{1cm} (8)

(2) When the covariance matrix of $G_1$ and $G_2$ are not equal

The square of the Mahalanobis distances from $x$ to $G_1$ and $G_2$ are respectively:
\[ d^2(x,G_1) = (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) \quad d^2(x,G_2) = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \]  

Let \( W(x) = d^2(x,G_2) - d^2(x,G_1) = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1), \) So the criterion of discrimination is the same as above.

5. Clustering Discrimination of Production of Oil Wells

Aiming at the field data of oil fields in Xinjiang, we first process the abnormal data of 48 wells according to the 3σ criterion. Then we analyze the correlation of 24 factors affecting the production of oil wells and calculate the correlation coefficients of each parameter. The specific results are shown in the table below:

| Rank | Parameter                      | Correlation Coefficient | Rank | Parameter                      | Correlation Coefficient |
|------|--------------------------------|-------------------------|------|--------------------------------|-------------------------|
| 1    | Oil Saturation                 | 0.97471046              | 13   | Percentage of Preflux Displacement Allocation Relation | 0.8699038              |
| 2    | Formation Pressure             | 0.968737075             | 14   | Reservoir Allocation Maximum Displacement Number | 0.8672321              |
| 3    | Reservoir Thickness            | 0.967845103             | 15   | Fracturing Intervals            | 0.8668482              |
| 4    | Total liquid volume            | 0.966730578             | 16   | Formation Temperature          | 0.8666847              |
| 5    | Fracture Pressure              | 0.965928037             | 17   | Casing Pressure                | 0.865624               |
| 6    | Average Displacement           | 0.96540494              | 18   | Reservoir Depth                | 0.8648642              |
| 7    | Formation Permeability         | 0.965205188             | 19   | Average Sand Ratio             | 0.8641495              |
| 8    | Formation Porosity             | 0.964784008             | 20   |                              | 0.8634023              |
| 9    | Maximum Sand Content           | 0.964701519             | 21   | RT                            | 0.8608984              |
| 10   | Sand Allocation Relation       | 0.877000873             | 22   | Pressure Gradient              | 0.8593414              |
| 11   | Half Seam Length               | 0.873563294             | 23   | RI                            | 0.8578376              |
| 12   | Average Sand Content           | 0.870954158             | 24   | Static Filtration Loss Coefficient | 0.8545693          |

As can be seen from the table above, the correlation coefficients after the ninth are at a low level, so we take the first nine factors as the main control factors of oil well production. Then the field data of 48 wells were randomly divided into 40 training data and 8 prediction data by using nine main control factors selected by correlation analysis. Next, we use 40 training data to fit and establish a Mahalanobis distance discriminant model.

- Classification of sample wells.

| Classification Criteria | Well Type      | Group | Well Number |
|-------------------------|----------------|-------|-------------|
| Average oil production ≥2.5t/d | Non-Low yield | \( G_{1} \) | 22           |
Constructing discriminant function.
By calculating the mean vector, covariance matrix, discriminant coefficients and constant terms of samples, we finally get discriminant functions for non-low production wells and low production wells. Discriminant model for non-low yield wells:

\[ W_1(x) = 5.0866x_1 + 0.2548x_2 + 0.0429x_3 + 0.7580x_4 + 125.3366x_5 + 2.1768x_6 - 2.8521x_7 \]

Discriminant model for low yield wells:

\[ W_2(x) = 4.1123x_1 - 0.8028x_2 + 0.0226x_3 + 0.7831x_4 + 130.6708x_5 + 2.2180x_6 - 2.9206x_7 \]

Using the above discriminant model, the fitting accuracy of the model is 92.5%, and the fitting effect is good, which proves that the model can be used in the next stage of prediction. By substituting 8 wells from the predicted data set into the above model, the final prediction and classification results are shown in the following table:

| Well | Discriminant Function | Minimum | Group | Discriminant Type | Real Type |
|------|-----------------------|---------|-------|-------------------|-----------|
| 1    | 0.340780              | 1.298681| 0.340780 | G1                | Non-Low yield | Low yield |
| 2    | 7.043783              | 9.684309| 7.043783 | G1                | Non-Low yield | Non-Low yield |
| 3    | 0.734003              | 2.722185| 0.734003 | G1                | Non-Low yield | Non-Low yield |
| 4    | 3.093142              | 2.545563| 2.545563 | G2                | Low yield     | Low yield |
| 5    | 26.02864              | 22.44502| 22.44502 | G2                | Low yield     | Low yield |
| 6    | 5.492796              | 3.894644| 3.894644 | G2                | Low yield     | Low yield |
| 7    | 2.168031              | 6.602631| 2.168031 | G1                | Non-Low yield | Non-Low yield |
| 8    | 2.060475              | 4.979700| 2.060475 | G1                | Non-Low yield | Non-Low yield |

From the above classification and discrimination results, it can be seen that only the first error is found in the eight prediction data, and the prediction accuracy is 87.5%. The classification prediction effect is good and can meet the actual needs.

6. Conclusion
Clustering discrimination is of great significance for solving many complex engineering problems. The clustering discriminant model is used to optimize the production prediction of oil wells, which improves the prediction accuracy significantly, has good popularization value and can effectively provide guidance for well selection and reservoir selection.

In this paper, we optimize the model and narrow the scope of determining the main control factors in the way of data screening and parameter optimization. We use Mahalanobis distance as discriminant statistics to build clustering model and optimize it by using data features. The fitting accuracy of the discriminant model is 92.5%, and the final prediction accuracy is 87.5%.

Acknowledgment
This study is supported by the National Science and Technology Major Demonstration Project (2017ZX05070).

Reference
[1] Figueroa, H.G., Lagreca, A., Gahan, B.C., et al. (2002) Rock removal using high-power lasers for petroleum exploitation purposes. Proc Spie, 4760:678-690.
[2] Desroches J., Detournay, E, Lenoach, B, et al. (1994) The Crack Tip Region in Hydraulic Fracturing. Proceedings of the Royal Society A, 447(1929):39-48.

[3] Everitt, B. Cluster analysis. (1980) Quality & Quantity, 14(1):75-100.

[4] Cao, H., Zhou, K., Chen, X., et al. (2017) Early chatter detection in end milling based on multi-feature fusion and $3\sigma$ criterion. International Journal of Advanced Manufacturing Technology, 92(9-12):4387-4397.

[5] Cohen, J. (2003) Applied multiple regression/correlation analysis for the behavioral sciences. L. Erlbaum Associates, New Jersey.

[6] Maesschalck, R.D., Jouan-Rimbaud, D., Massart, D.L. (2000) The Mahalanobis distance. Chemometrics & Intelligent Laboratory Systems, 50(1):1-18.