Multi-stage perforation and hydraulic fracture stage selection based on machine learning methods

Yulong Yu and Jiafang Xu*

College of Offshore Oil & Gas Engineering, China University of Petroleum (East China), Qingdao, 266555, China

*Corresponding author. Email: yyl_sn@qq.com

Abstract. In unconventional oil and gas resources, especially shale oil and gas resources with extremely low permeability and porosity, in order to develop effectively, it is necessary to establish perforation and multi-stage fracturing. The machine learning algorithm k-means clustering was used to cluster the five features selected to describe reservoir properties: shale content, porosity, total organic carbon content and those reflecting rock mechanical properties: Young's modulus and Poisson's ratio. The data were classified in a high-dimensional space to determine different perforation fracture stages. The classification algorithm XGboost was then used to predict different perforation stages using four conventional logging curves GR, NPRL, VP and DEN. The K-means clustering algorithm based on Euclidean distance can well classify the selected features in the high-dimensional space. The Hopkins statistics of clustering trend is 0.94, showing a good clustering trend. When the classification algorithm is used for prediction, the average accuracy is 0.92, the average recall rate is 0.90, and the average F1 score is 0.90, which can predict different perforated fracture stages well and optimize the design of the perforated fracture stage.

1. Introduction

In unconventional wells, one of the two largest challenges are low stage and cluster efficiency (Miller et al, 2011) [1]. The change in reservoir rock quality and mechanical properties across the horizontal lateral makes an impact on the completion and the ultimate production of the well and the heterogeneity of reservoirs alone well hole and the differing rock mechanics within a stage result in low completion efficiency. That needs to select perforation clusters and fracture location to optimize well completion of vertical well, especially horizontal wells. That is to say, it is of great importance that optimizing stages selection. There is an evidence that Cipolla et al conducted production tests in more than 100 horizontal wells and found that approximately 40% of the perforation clusters were ineffective because of the differences in reservoir properties and geomechanical parameters within a cluster [2]. It’s important to understand the reservoir and geomechanical properties alone wellbore. To determine the highest quality reservoir and the intervals with the most similar rock mechanics, Byron [3] Cottingham et al uses gamma ray, drill cuttings, rate of penetration (ROP), mud log gas shows, and gas chromatography to correlate each foot of the lateral to understanding of the reservoir and geo-mechanics. Lyle [4] use drilling data to optimize completion efficiency, using neural network to generate geomechanical and reservoir properties from drilling and logging data. Those properties include Young's Modulus, Poisson's Ratio, horizontal stress and total organic carbon content (TOC) which used to optimize stages and perforation clusters selection. Cipolla [2] relies on grouping rocks of similar properties to position stages and
perforations clusters both vertical and horizontal wells, which used effective porosity, permeability and mineralogy, stress, Young’s modulus and Poisson’s ratio.

As development of oil and gas industry, there are “big data” that can be used. Data analytic methods can be used to perforation and fracture treatment. Ishank [5] focus on the use of data analytics to define determine optimal locations to place fractures along the wellbore based on machine learning methods. The petrophysical data of formation, such as velocity, Young’s modulus, Poisson’s ratio, and creep displacements, are used to define different clusters to evaluate which cluster is more suitable for fracture. Ahmed Alzahabi [6] analyse field data to predict fracture stages space, the number of clusters and perforation via machine learning methods, such as Random Forest, GradBoost, Decision Tree and neural network. However, the work just considers the data after production which part of data is reasonable to design clusters and stages for perforation and fracture.

2. Workflow

The objective of this work is to present a workflow to optimize the stage selection and propose a new workflow as Fig. 1 shows.

![Figure 1. Workflow to select stages of perforation and fracturing.](image)

First step, selected feature as input to k-means clustering, and output is different clusters that well segments along wellbore. Because k-means clustering is a cluster algorithm based on Euclidean distance, it will divide the selected data set into different clusters in the space, the distance between the features of each clusters will be the shortest. The uncertainty caused by artificial interval setting is avoided. From a physical point of view, each category has similar properties and each cluster will be in similar rock or the same stratigraphic interval to optimize stages and perforation clusters selection. Second step, training XGBoost model and predict with result of K-means clustering using other four conventional well logs, including VP, GR, DEN and NPRL so that we can use four conventional well logs to predict clusters by trained XGBoost model directly without K-means clustering.

3. Method

As describe above workflow, we use two machine learning method, including unsupervised algorithm K-means cluster and supervised algorithm XGBoost (eXtreme Gradient Boosting).

3.1. K-means cluster

Cluster Analysis is a classical unsupervised learning algorithm. In the case of given samples, clustering analysis automatically divides them into several clusters by means of feature similarity or distance measurement. In this paper, K-mean clustering algorithm [7] is adopted, and Euclidean distance is used as the measurement method of distance between samples. Given m × n sample \( X = \{ x_1, x_2, ..., x_m \} \), the distance between samples \( d(i,j) \) can be defined as follow equation (1):

\[
d(i,j) = \sum_{k=1}^{n}(x_{ik} - x_{jk})^2 = \| x_i - x_j \|^2
\]

The sum of the distance between the sample and its class center is defined as the final loss function as follow equation (2):

\[
L = \sum_{i=1}^{m}d(i, \text{center}_i)
\]
where $\bar{x}_l = (\bar{x}_{1l}, \bar{x}_{2l}, \ldots, \bar{x}_{ml})^T$ is centroid of class $l$. The function $W(C)$ represents the similarity of samples in the same class. Keep iterating until convergence. Any detail can be obtained reference 7.

### 3.2. XGBoost

XGBoost [8] was proposed by Dr. Tianqi Chen of The University of Washington in 2015. It is an additive model composed of multiple base models as follow equation (3):

$$\hat{y}_i = \sum_{t=1}^T f_t(x_i)$$

where $f_k$ donates the $k$th basis model, $\hat{y}_i$ donates $i$th prediction. The loss function can be defined as equation (4):

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i)$$

where $n$ donates the sample size.

### 4. Data description

There are 5 wells and each well contain 9 well logs, including CLA, POR, TOC, E, PR, VP, GR, DEN and NPRL. CLA donates the clay content; POR donates porosity; TOC donates total organic carbon content; E donates the Young's modulus; PR donates the Poisson's ratio; VP donates compressional wave velocity NPRL donates neutron log; GR donates natural gamma log curve; DEN donates density log.

The former five well logs are dataset which are used to cluster with K-means cluster and others are used to predict clusters using XGBoost that the results K-means cluster which are the different stages for perforation and multi-stage hydraulic fracture location.

#### 4.1. Data preprocessing

**4.1.1. Outlier processing.** The dataset contains outliers, and we use boxplot of each features to cleaning outliers as following equations (5), (6):

$$\text{min} = Q1 - 1.5 \times (Q3 - Q1)$$

$$\text{max} = Q3 + 1.5 \times (Q3 - Q1)$$

where $Q1, Q3$ donates first and third quartile of each attribute.

**4.1.2. z-score standardize.** z-score standardize each attribute using equation (7):

$$\hat{x} = \frac{(x - \mu)}{\sigma}$$

where $x$ donates original data; $\hat{x}$ donates standardized data; $\mu$ donates mean value of each attribute.

### 5. K-means clustering

In this section, our dataset will be divided k clusters using K-means clustering.

**5.1. Hopkins Statistic**

Analyse the cluster trend with Hopkins Statistic that the value is near 1, the clustering trend is obvious. The Hopkins Statistic of over data is 0.94, that is to say, the clustering trend is obvious absolutely.

**5.2. Calculate the K of K-means clustering**

The Elbow Method determine the number of clusters, as Figure 2 shows, $K = 4$, that means dataset well be separated 4 clusters.
5.3. Result
Fig. 3 shows the results visualization with different angles using t-SNE dimension reduction method. Dataset is divided four different cluster. Fig. 4 shows the cluster map along borehole of 5 wells.

Figure 2. The Elbow Method determine the number of clusters, and the k = 4 in red circle.

Figure 3. The results visualization with different angles using t-SNE dimension reduction method.

Figure 4. The map of cluster along borehole of 5 wells: A1, A2, A3, A4, A5.
6. Predict on conventional well logs
In this section, we train XGBoost model and predict with result of K-means clustering using other four conventional well logs, including VP, GR, DEN and NPRL.

6.1. Training strategy
Leave-one-out (LOO) method is reserved for training and testing model to improve the utilization rate of samples and better training model, that is to say, training model with four wells’ data and testing model with one well’s data which five experiments totally.

6.2. Evaluation metrics
To evaluation model, there are three evaluation metrics, including precision(weighted), recall(weighted), F1-score (weighted), as following equations (8), (9), (10):

\[ P_{\text{weighted}} = \frac{\sum_{i=0}^{n} P_i}{\sum_{i=0}^{n} \omega_i}, \tag{8} \]
\[ F_{1\text{weighted}} = \frac{\sum_{i=0}^{n} F_i}{\sum_{i=0}^{n} \omega_i}, \tag{9} \]
\[ R_{\text{weighted}} = \frac{\sum_{i=0}^{n} R_i}{\sum_{i=0}^{n} \omega_i}, \tag{10} \]

where \( \omega_i \) donates interclass weights, the number of samples of the class (actual rather than predicted) / total number of samples.

6.3. Results
As Table 1 shows three evaluation metrics of each experiment. The average accuracy is 0.92, the average recall rate is 0.90, and the average F1 score is 0.90. Trained XGBoost model have accurate prediction of classification results.

| Evaluation Metrics     | A1   | A2   | A3   | A4   | A5   | average |
|------------------------|------|------|------|------|------|---------|
| Precision(weighted)    | 0.9023 | 0.9337 | 0.8847 | 0.9230 | 0.9519 | 0.9191  |
| Recall(weighted)       | 0.8090 | 0.9307 | 0.8833 | 0.9196 | 0.9506 | 0.8986  |
| F1(weighted)           | 0.8340 | 0.9311 | 0.8835 | 0.9117 | 0.9503 | 0.9021  |

7. Conclusions
In this paper, we design workflow to select stage of perforation and multi-stage fracturing. Firstly, we select three features that about reservoir quality and two geomechanically property parameters which are important to perforation and multi-stage fracturing designation. K-means clustering based on Euclidean distance can divide dataset into different clusters on high dimension space, which is more reasonable, so each cluster have same reservoir quality and geo-mechanical quality, which is used to optimize the selection of the stage and cluster selection of multi-stage perforation and fracturing. Secondly, five features as inputs, we use machine learning method that K-means clustering based Euclidean distance to separate well into different stages. In this way, the differences in reservoir properties and geomechanical parameters within a section can be reduced. What’s more, we train XGBoost model and predict with result of K-means clustering using other four conventional well logs as mentioned section 3, including VP, GR, DEN and NPRL so that we can use four conventional well logs to predict clusters by trained XGBoost model without K-means clustering.

References
[1] Miller C., Waters G., and Rylander E. Evaluation of Production Log Data from Horizontal Wells Drilled in Organic Shales. Paper SPE 144326 presented at SPE North American Unconventional Gas Conference and Exhibition, The Woodlands, Texas, 14-16 June (2011)
[2] Cipolla, C. L., Weng, X., Onda, H., Nadaraja, T., Ganguly, U. and Malpani, R., 2011. New Algorithms and Integrated Workflow for Tight Gas and Shale Completions, Paper SPE 146872
presented at the SPE Annual Technical Conference and Exhibition, Colorado (2011)

[3] Byron Cottingham, Jerome Truax, Sam Roark, Matt Padgham, Steve Denney, Xiaoxuan An. A Simple and Cost-Effective Workflow for Engineered Perforations. Unconventional Resources Technology Conference, Houston, Texas (2018)

[4] Lyle V. Lehman, Kale Jackson, Bruce Noblett. Big Data Yields Completion Optimization: Using Drilling Data to Optimize Completion Efficiency in a Low Permeability Formation. The SPE Annual Technical Conference, Dubai, UAE (2016)

[5] Shank Gupta, Chandra Rai, Deepak Devegowda, and Carl Sondergeld. Use of Data Analytics to Optimize Hydraulic Fracture Location Along Borehole (2018)

[6] Ahmed Alzahabi, A. Alexandre Trindade, Ahmed Kamel, and Wade Baustian. 2019. Predicting Fracture Clusters, Cluster Spacing and Perforations of Horizontal Wells Using Data Analytics. Paper presented at 53rd US Rock Mechanics/Geomechanics Symposium, NY, USA (2019)

[7] Hartigan, J. A. and M. A. Wong. A k-means clustering algorithm. Applied Statistics 28, 100–108 (1979)

[8] Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. ArXiv e-prints (2016)