Application of GBDT for division of petroleum reservoirs

Yaqiong Qin, Zhaohui Ye and Conghui Zhang

1 Department of Automation, Tsinghua University, Beijing 100084, China
2 China Oilfield Services Limited, Sanhe, Hebei 065201, China
*E-mail: qyq16@mails.tsinghua.edu.cn

Abstract. Traditional methods of dividing petroleum reservoirs are inefficient. The machine learning method has been applied in the two-class and three-class tasks of the reservoir, but there is no research on the classification of all classes. In this paper, the GBDT (Gradient Boosting Decision Tree) model is proposed for the reservoir classification problem in argillaceous sandstone areas, and the model is implemented by XGBoost algorithm. The classification effect of this model is better than that of multiple-hidden-layer neural networks, and the macro-average AUC is 0.89.

1. Introduction
In the field of petroleum well-logging, reservoirs are those that have interconnected pores and are capable of storing fluids such as water and oil [1]. The task of reservoir division is to determine the layered interface and the thickness of different types of reservoirs. In the process of petroleum exploration and development, people use specialized detection instruments and utilize physical measurement methods such as sound, electricity, magnetism and radioactivity to obtain various logging curves. At present, the artificial method of dividing reservoirs is mainly observing the characteristics of the logging curves to qualitatively identify the reservoir type and the layered interface.

The artificial method of dividing reservoirs is time-consuming and laborious, what's more, its stratification results are greatly influenced by the professional experience of the logging analysts. Therefore, it is meaningful to explore the automatic approaches to dividing reservoirs. In the early days, people's idea was to integrate expert experience into computers and build an expert system for well-logging [2]. This method has high requirements for researchers' expertise and is not good for atypical reservoirs. After that, researchers used mathematical methods to study the division of reservoirs. For example, [3] proposed a stratification method based on logging curves' activity. The activity of the curve implies the degree of curve' saltation, which can help identify the interface of different layers, but due to the complex variation of logging curves, the accuracy of this method is not ideal.

With the development of artificial intelligence technology, people began to use neural networks to study the problem of reservoir partitioning. [4-5] used BP neural network with one hidden layer to divide reservoirs; [6] introduced the coordinate-method of solving the optimization problem into BP neural network, but the network still had only one hidden layer. These attempts did not significantly increase the accuracy of automatic stratification. [7-8] utilized one-hidden-layer BP neural network to deal with the oil-gas-water reservoir classification problem. The classification accuracy is about 88%. However, they make judgments for a given reservoir and cannot automatically determine the layered interface.
The GBDT (Gradient Boosting Decision Tree) method has a strong fitting ability in classification tasks in a densely distributed data set, and the calculation speed is fast. In the field of oil logging, Han Qidi et al. [9] used the GBDT method to study lithology identification, inputing grade data of chemical elements, and outputing lithology categories, and the test accuracy is optimally 92.7%. In the reservoir partition problem, no researchers have used the GBDT method. Therefore, this paper introduces the GBDT method into the reservoir partition problem, and the comparison algorithm uses multiple hidden layer BP neural network. Experiments show that the layering effect of GBDT method is better than other existing methods such as multi-hidden-layer BP neural network.

The rest of the paper is organized as follows: Section 2 will introduce the preprocessing method of logging data, the structure and optimization strategy of GBDT models. Section 3 gives the results of the experiment. Section 4 summarizes the content of the article.

2. GBDT models for reservoir division

2.1. Features and samples

2.1.1. Feature selection. There are many kinds of logging curves. However, not every oil-well has a full range of logging curves, and the conventional 9 curves are available in most wells. There are 9 main types of conventional logging curves: natural gamma ray (GR), spontaneous potential (SP), caliper (CAL), deep lateral resistivity (RD), shallow lateral resistivity (RS), and micro lateral resistivity (RMLL), acoustic time difference (DT), neutron (CNCF), density (ZDEN). From the perspective of artificial method, the conventional 9 curves are extremely important for dividing reservoirs. Therefore, for practical considerations, the conventional 9 curves are chosen as input features.

2.1.2. Sample selection. In this paper, the oil wells of argillaceous sandstones in the marine area are taken as research objects. The logging data is selected from four oil wells in a certain sea area. The depth ranges of these four wells are Well1: 950m-3376.2m, Well2: 2778.9m-3883.1m, Well3: 2768.6m - 3392.3m, Well4: 843.8m - 3873.8m. The artificial interpretation results table shows the type of reservoir at each depth point of each well, including Oil Layer, Poor-oil layer, Oil-water Layer, Water-with-oil Layer, Water Layer, Dry Layer and Shale Layer. Therefore, this topic requires division tasks in seven types of reservoirs.

Each depth corresponds to a sample, with a label which is one of the seven types of reservoirs. This sample is a 9-dimensional data, and each dimension corresponds to a logging variable. Part of the real data is shown in Figure 1.

![Figure 1. Part of the real logging data](image)

2.1.3. Processing of unbalanced dataset. The number of the seven types of reservoirs in this subject is extremely uneven, and there are few samples in the Oil layer and the Poor-oil layer. If left unprocessed, the classifier will prefer to predict the test sample class as a majority class, so the unbalanced dataset needs to be processed.
The common processing methods of unbalanced datasets are undersampling and oversampling. These two methods have obvious shortcomings, so the method adopted in this paper is SMOTE. The SMOTE algorithm is as follows: for each minority class sample, find its K-nearest neighbor where the distance used as the Euclidean distance. According to the proportion of data that needs to be amplified, m is randomly selected among K neighbors (m=n-1 if n is to be amplified), and interpolation is performed between its m neighbors to generate a new sample. As shown below

\[
x_{\text{new}} = x_i + (\hat{x}_i - x_i) \times \delta
\]

(1)

Where \(x_{\text{new}}\) represents the new sample generated, \(\hat{x}_i\) represents the neighbor of \(x_i\), and \(\delta\) represents the random number between 0 and 1.

The Borderline SMOTE algorithm [10] divides the minority class samples into three types:
1) Noise samples: All K neighbors of candidate samples belong to the majority class.
2) Dangerous samples: More than half of the K-nearest neighbors of the candidate sample belong to the majority sample.
3) Safety samples: More than half of the K-nearest neighbors of the candidate samples belong to the minority class.

Selecting candidate samples only from dangerous samples allows the synthesized new sample to contain enough useful information without introducing significant noise. In this paper, the Borderline SMOTE algorithm is adopted when dealing with multi-class unbalanced datasets, and the following improvements are made: since the number of the minority class samples is too small, the candidate samples are chosen not only from dangerous samples but also from a number of safety samples. Table 1 shows the data of well1 before and after SMOTE processing.

|      | Shale | Water | Water-with-oil | Oil-wate | Dry | Poor-oil | Oil |
|------|-------|-------|---------------|----------|-----|----------|-----|
| Raw data | 7920  | 1195  | 474           | 137      | 437 | 175      | 643 |
| After SMOTE | 7920  | 3000  | 2000          | 2000     | 2000| 2000     | 2000 |

2.2. GBDT model

GBDT (Gradient Boosting Decision Tree) [11] is a boosting method whose base classifier is decision tree. The training method of GBDT is gradient boosting. We use XGBoost [12] as an implementation tool for GBDT, because of its good performance and ease of customization.

2.2.1. Model design. Using the GBDT model to deal with classification problems, there is no need to do missing value filling and normalization of the data. For the multiple classification problem in this paper, because the datasets are very unbalanced, the minority classes tend to be ignored in the prediction. However, these minority classes are often oil-layer reservoir categories, which is the class that log interpreters are most concerned about, and it is hoped that these minority classes will be correctly predicted as much as possible. Therefore, this paper increases the importance of the minority classes in the design of the loss function. The basic loss function still selects the default Softmax-cross entropy of XGBoost. As an improvement, different classes are given different weights on the loss, as follows:

\[
\text{loss} = - \sum_{j=1}^{N} \sum_{k=1}^{K} w_{t_{jk}} \log(y_{jk})
\]

(2)

\[
w_k = \log \left( \frac{\sum_{k=1}^{K} m_k}{m_k} \right)
\]

(3)
Where \( t_{ik} \) is the probability that sample \( x_i \) belongs to class \( k \), \( y_{ik} \) is the probability that the model predicts sample \( x_i \) as class \( k \), \( w_k \) is the class weighting coefficient of class \( k \), and \( m_k \) is the total number of samples of class \( k \) after SMOTE synthesis.

For the minority classes, the value of \( w_k \) is larger, so the penalty from the loss function for the error of this class will be greater. Such a design is equivalent to adding a priori of the data distribution to the optimization goal, and weighting the loss function with the class proportion of the training set. When the test set is consistent with the training set, this assumption will bring good results, otherwise it will bring bad effects.

3. Experiments

In the four oil wells in the experiment, for the datasets of well1, well2 and well3, 80% of the total data of each well was selected and mixed together as a training set, and the remaining 20% of the data was used as the verification set. The whole well4 data is used as a test set. According to the prediction effect of the model on the verification set, the parameter setting is determined as shown in Table 2 after fully adjusting the parameters.

| Parameters                          | Values |
|-------------------------------------|--------|
| Learning rate                       | 0.05   |
| Iterations                          | 80     |
| Max-depth of the tree                | 10     |
| Min leaf node                       | 6      |
| Feature sampling ratio              | 0.8    |
| Min loss required for node splitting| 0.01   |
| L2 regularization coefficient       | 0.6    |

In this paper, the neural network with multiple hidden layers is selected as the contrast algorithm. The network model uses the conventional 9 lines as the input, so the input layer has 9 neurons. The number of neurons in the output layer is 7. The setting of the hidden layer is to be tried through experiments. After trying, the number of hidden layers in the neural network is determined to be 3, and the number of neurons in each hidden layer is 20, 15, and 9, respectively.

The test accuracy and macro-average AUC results of the four wells are shown in Table 3. Take the well4 as an example to observe the classification effect, the classification ROC curve is shown in Figure 2 and Figure 3.

| Number | Accuracy of GBDT | Accuracy of Neural Network | AUC of GBDT | AUC of Neural Network |
|--------|------------------|----------------------------|-------------|-----------------------|
| Well1  | 0.9508           | 0.8352                     | 0.97        | 0.77                  |
| Well2  | 0.9477           | 0.8701                     | 0.96        | 0.81                  |
| Well3  | 0.9301           | 0.8124                     | 0.96        | 0.74                  |
| Well4  | 0.8908           | 0.8002                     | 0.82        | 0.80                  |
It can be seen that the classification accuracy and AUC of the GBDT model exceed the neural network model. GBDT is superior to neural networks in its ability to express reservoir classification problems. However, in the test set, the classification effect of GBDT is worse than that of the verification set, and this is the direction of further research.

4. Conclusions
This paper uses nine conventional logging curves to classify petroleum reservoirs. Since the dataset is unbalanced, we applied the Borderline SMOTE algorithm to increase the number of the minority classes samples.

Since the logging data is a densely distributed numerical data, it is suitable to use the GBDT model. In this paper, the GBDT method is introduced into the reservoir classification problem, and the class weighted loss function is designed for the imbalance of the dataset. We choose multiple hidden layer neural networks as comparison algorithms, and experiments show that the classification accuracy and AUC of the GBDT model exceed the neural network model.
For the generalization effect, further research is still needed to promote the practicality of reservoir classification by GBDT.

Acknowledgments
This work was supported by the China Oilfield Services Limited.

References
[1] Hong Y. Logging Principles and Comprehensive Interpretation[M]. China university of petroleum press, 1993. (in Chinese)
[2] Yi G, Lv J, et al. Expert system of petroleum reservoir evaluation [J]. Well Logging Technology, 1986, 10(1):1-8. (in Chinese)
[3] Du W, Activity division method of well-logging curves[J]. Coal Geology of China, 1991, 3(3):83-88. (in Chinese)
[4] Chen Z. Application of Artificial Neural Network for Oil-layer Recognition[J]. Petroleum Geology & Oilfield Development in Daqing, 1994(3). (in Chinese)
[5] Zhang L, Wang L, Dividing fractured carbonate reservoirs with neural networks[J]. Compound oil and gas field, 1995(2):43-45. (in Chinese)
[6] Sun K, Zhou Y. Improved neural network algorithm and its application in oil layer recognition[J]. China Petroleum Machinery, 2004, 32(3):28-29. (in Chinese)
[7] Zhu K, Song H, Gao J, et al. Application of Evolutionary Neural Networks for Well-logging Recognition in Petroleum Reservoir[C]// Seventh International Conference on Computational Intelligence and Security. IEEE Computer Society, 2011:362-366.
[8] Li M, Zhang J. Application of neural network technique for logging fluid identification in low resistance reservoir[C]// Sixth International Conference on Natural Computation. IEEE, 2010:163-166.
[9] Han Qidi, Zhang Xiaotong, Shen Wei. Lithology identification technology based on gradient lifting decision tree (GBDT) algorithm[J]. Bulletin of Mineralogy, Geochemistry, 2018(6): 1173-1180. (in Chinese)
[10] Huang P J. Classification of Imbalanced Data Using Synthetic Over-Sampling Techniques[J]. Dissertations & Theses - Gradworks, 2015.
[11] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5):1189-1232.
[12] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2016.