Discrimination of AVO Types in Shallow Reservoirs Based on Random Forest Algorithm

Liu Haonan1, Wen Xiaotao1, He Jian1, Chen Qianshu1, Zhang Xiaoqi1

1Key Laboratory of Earth Exploration and Information Technology of Ministry of Education, Chengdu University of Technology, Chengdu, 610059, China
E-mail: 710490592@qq.com

Abstract-As we all know, AVO technology can be used to identify gas-bearing reservoirs and is of great significance to oil and gas exploration. The manual identification of AVO types of reservoirs has large human interference factors, low identification accuracy and long time-consuming. Therefore, this paper introduces the random forest algorithm, uses bootstrap repeated sampling and branch and leaf node splitting techniques to generate a large number of decision tree classifiers, and realizes the identification of the AVO type of the reservoir by counting the classification results of all decision trees. Firstly, a velocity density model is established based on logging data in the working area. Secondly, use the Shuey approximation formula to calculate the AVO curve and obtain the fitted polynomial corresponding to the curve. Thirdly, the morphological feature parameters are extracted according to the fitting polynomial as the input parameters of the training data set of the random forest algorithm, and the artificial AVO type recognition results are used as the output parameters to train and obtain a decision tree classifier. Finally, using the characteristic parameters of the AVO curve of the actual pre-stack seismic data as input parameters, the classification of the AVO type of the reservoir in the working area is obtained through the classification of the random forest decision tree. By comparing the results with the approximate support vector machine algorithm, it can be seen that the two algorithms approximate the AVO type discrimination results of the reservoir, and both have high accuracy, but the random forest algorithm requires fewer feature attributes, shows stronger generalization and has better universality.

1. Introduction
The AVO analysis technology uses the relationship between the amplitude of seismic waves and the offset to find oil and gas, which is of great significance to the prediction of the oil and gas content of the reservoir. Since Ostrander[1] proposed the AVO analysis technology, there are mainly four types of AVO characteristic curves used to describe gas-bearing sandstones. Type I, II, and III were proposed by Rutherford[2] et al., and Type IV was proposed by Castagna[3] et al. Under normal circumstances, people classify by analyzing the variation of CRP trace concentration amplitude with the angle of incidence, but the manual identification method is time-consuming and laborious. Therefore, in recent years, people have gradually considered using machine learning algorithms to automatically identify the AVO type. Large number of studies have proved that machine learning algorithms are more computationally efficient, can reduce the negative impact of human factors, and can perform multi-attribute fusion analysis, including Random Forest (RF) and Proximal Support Vector Machine (PSVM), etc. In the application of non-artificial recognition of AVO types of reservoirs, only Li Wenxiu [4] et al. implemented the PSVM algorithm in 2018, but the algorithm requires more feature attributes and poor generalization.
Considering that the RF algorithm has the characteristics of less generalization error, less interference, and less prone to overfitting[5], this paper intends to introduce the RF algorithm to distinguish the AVO type of the reservoir in the work area, and test the method's AVO type recognition applicability.

This paper extracts the morphological characteristic parameters from the four types of AVO reflection coefficient curves as the training set, and introduces the random forest algorithm to determine the morphological characteristic parameters extracted from the actual area pre-stack seismic data to determine the AVO at the top of the target layer in the study area. At the same time, in order to reflect the advantages of the random forest algorithm, this paper compares the application effect of the algorithm with the approximate support vector machine algorithm.

2. Theory and Realization Method of Random Forest Algorithm

2.1. Principle of algorithm

Random forest algorithm is a combination classification prediction algorithm proposed by Breiman on the basis of Bagging algorithm and integrated learning idea [6-7]. It uses bootstrap repeated sampling technology to obtain multiple decision tree combinations, and finally aggregates the prediction results of these multiple decision tree combinations and outputs the overall output.

The algorithm makes predictions based on a combination of multiple decision trees, so each decision tree depends on the value of a random vector that is sampled independently and has the same distribution for all trees. If the number of all attributes is N, the optimal attribute \( n_{\text{split}} \) satisfies \( n_{\text{split}} \leq N \), that is, the optimal attribute is the optimal result obtained by comparing the advantages and disadvantages of some attributes randomly selected from all N attributes. During the splitting process, these optimal attributes \( n_{\text{split}} \) will be split in the form of internal nodes. After each split, there will be an attribute feature value \( (Q_{i}, i = 1,2,\cdots,q) \) corresponding to the branch, and the node of each branch leaf is used as the category of our sample, and the cycle is repeated until the decision tree reaches the termination decision condition and exits and outputs result. The decision tree classifier \( \{C(X,\varphi_k), k = 1,2,\cdots,K\} \) is constructed irregularly based on the bootstrap resampling method, where, \( \varphi_k, k = 1,2,\cdots,K \), K and X represent a random vector sampled independently and with the same distribution, the sum of the number of all decision trees in the random forest, and the independent variables that need to be given. In the case where the independent variable X is known, the classification judgment is first performed through the decision tree, and then the classification prediction results of each decision tree are aggregated, and finally the mode category is selected from the aggregated classification prediction results by voting as the optimal sample Classification results.

2.2. Realization Method

1) First, on the basis of the parameter model, the original training set A is made, and then P training set subsets with the same capacity as A are randomly selected through the bootstrap method with replacement, and P decision tree classifiers are constructed from this.

2) When generating a decision tree, first randomly select \( n_{\text{try}} \) variables at the tree's branch and leaf nodes, and then select the variable \( n_{\text{split}} \) with the best classification ability to split from \( n_{\text{try}} \). In this way, the correlation can be reduced while keeping the classification strength unchanged, thereby reducing generalization errors and improving the classification accuracy of the system.

3) The growth of each decision tree in the forest is unrestricted and no cutting is done.

4) A large number of decision trees are formed into an RF classifier, and then the classifier is used to classify multiple feature attributes containing sample features, and then the classification results of all the decision trees are counted to vote, and finally the mode of the classification results is selected as The result of sample classification prediction.
3. AVO discrimination based on random forest

3.1. Preparation of training data set

Since the change of reflection coefficient with the angle of incidence is related to the difference in fluid properties and the degree of compaction of the reservoir overlying strata, etc., we can start with its change characteristics and quantify the four types of AVO curve morphological characteristic parameters to obtain the training set. Each characteristic attribute. First, establish a velocity density model based on the logging data in the work area (Table 1); then use the Shuey approximate formula to calculate the model to obtain the reflection coefficient curve. Because the calculated reflection coefficient is discrete, it is not easy to extract the morphological characteristic parameters, so it is necessary to perform polynomial fitting on the discrete data. Figure 1 is the continuous curve after polynomial fitting; and then express from the fitting of the reflection coefficient curve. In the formula, multiple morphological feature parameters are extracted as the feature attributes of the training set and input into the training set; finally, the training set training and decision tree classification output the AVO type discrimination result.

Table 1  Velocity and density model

| Elastic Parameters | Overlying Medium | Gas-bearing sandstone |
|--------------------|------------------|-----------------------|
|                    | Vp1 (m/s)        | Vs1 (m/s)             | ρ1 (g/cm³) | Vp2 (m/s) | Vs2 (m/s) | ρ2 (g/cm³) |
| Type I             | 4200             | 1810                  | 2.5        | 6000      | 3900      | 2.1        |
| Type II            | 4200             | 1810                  | 2.5        | 5200      | 3300      | 2.1        |
| Type III           | 4200             | 1810                  | 2.5        | 4600      | 3000      | 2.1        |
| Type IV            | 4200             | 1810                  | 2.5        | 3800      | 2600      | 2.1        |

Figure 1 The variation of four types of AVO reflection coefficients varying with incident angles

As can be seen from Figure 1, the four types of AVO reflection coefficient curves have different characteristics. By analyzing the difference of the four types of AVO reflection coefficients with the angle of incidence, mathematical derivation can be used to extract morphological feature parameters from the curve fitting expression as the feature attributes of the training set, that is, the classification feature of the decision tree. The main morphological characteristic parameters include the starting and ending position of the curve, the monotonicity, the number of extreme points and their positions, the convexity, the number of inflection points and their positions, etc. They are the most representative and can describe the characteristic attributes of the AVO curve in detail. And the training set and sample set are formed by extracting this information (Table 2). Because when processing actual pre-stack seismic data, the morphological characteristic parameters are extracted from the variation curve of seismic wave...
amplitude with the angle of incidence, so the velocity density model can be used to calculate the four types of AVO reflection coefficients through the Shu-ey approximation formula and polynomial fitting. Then, the RF algorithm training set is made through the synthetic seismic records obtained by convolution with wavelets. Use the same method to obtain the training data set of the PSVM algorithm (Table 3), but because the PSVM algorithm itself has certain limitations for multi-classification problems, it is necessary to further mathematically derive the curve fitting expression to obtain intermediate variables that strengthen the algorithm’s binding force, and therefore make the number of its characteristic attributes more.

| Number | Starting point | Monotonicity | Number of extreme points | Extreme point location | Convexity | Number of inflection points | Inflection point location | End position | Tag |
|--------|----------------|--------------|--------------------------|-----------------------|-----------|-----------------------------|-------------------------|--------------|-----|
| 1      | 1.1            | 1            | 2                        | 2                     | 1         | 2                          | 2                       | 1            | 1   |
| 2      | 0.4            | 1            | 2                        | 1                     | 1         | 1                          | 0                       | 2            | 2   |
| 3      | -0.3           | 1            | 0                        | 1                     | 1         | -1                         | -1                     | 2            | 2   |
| 4      | -0.8           | 1            | 0                        | 1                     | 2         | -2                         | -2                     | 3            | 3   |
| 5      | -1.3           | 2            | 0                        | 1                     | 2         | 2                          | 0                       | 2            | 4   |

| Number | Starting point | Monotonicity | Number of extreme points | Extreme point location | Convexity | Number of inflection points | Inflection point location | End position | Intermediate Variables | Intermediate Variables | Intermediate Variables | Intermediate Variables | Tag |
|--------|----------------|--------------|--------------------------|-----------------------|-----------|-----------------------------|-------------------------|--------------|------------------------|------------------------|------------------------|------------------------|-----|
| 1      | 1.1            | 1            | 2                        | 1                     | 2         | -1                         | -1                     | 1            | 1                     | -1                     | -1                     | -1                     | 1   |
| 2      | 0.4            | 1            | 2                        | 1                     | 0         | -1                         | -1                     | 1            | 1                     | 1                      | 1                      | 1                      | 2   |
| 3      | -0.3           | 1            | 0                        | 1                     | -1        | -1                         | -1                     | -1           | 1                     | 1                      | -1                     | -1                     | 2   |
| 4      | -0.8           | 1            | 0                        | 1                     | 2         | -2                         | 1                      | -1           | -1                     | -1                     | -1                     | -1                     | 3   |
| 5      | -1.3           | 2            | 0                        | 1                     | 2         | 1                          | 1                      | -1           | -1                     | -1                     | -1                     | -1                     | 4   |

4. Applications

4.1. Working area overview
The working area is located in an offshore oil and gas field in the South China Sea. The overlying rock is mudstone, the reservoir section is sandstone, and the strata is the Zhujiang Formation (indicated by ZJ in the figure). In the seismic cross section of wells 1 and 2 (Figure 2), the positions indicated by the arrows are the reservoir sections of Wells 1 and 2, and the logging curve inserted in the figure is a gamma curve. Log interpretation (Figure 3) shows that Well 1 encountered a gas layer at the top of the target layer, and Well 2 encountered a gas-water layer at the top of the target layer. Through synthetic records and seismic data (Figure 4), the top and bottom positions of the sandstone in the reservoir section (shown by the red line in Figure 4) can be calibrated. At the same time, it can be seen from the logging curve that the gamma value of the reservoir section is abnormally low.
4.2. Application effect

This paper uses the AVO type of the synthetic record to test the classification results. The velocity density model of the synthetic record comes from logging data. Figure 5 is the forward gather of well 1. Figure 6 shows the extraction of the target top AVO curve from the actual gather (Take Well 1 as an example), and Figure 7 shows the extraction of the target top AVO curve based on the forward simulation gather (Well 1). Figures 6 and 7 both show that the AVO type at the top of the target layer in Well 1 is Type III.

After extracting the sample set from the curve fitting expression of the original pre-stack seismic data, use the RF algorithm (input 8 feature attributes) and PSVM algorithm (input 8 and 13 feature attributes respectively) to perform AVO on the sample set Type discrimination, the results are shown in Figure 8-10. Among them, the green represents the III AVO, and the purple represents the IV AVO. Comparing the three figures, we can see that: 1) The AVO type discrimination results of the RF algorithm (input 8 feature attributes) and the PSVM algorithm (input 13 feature attributes) are very close, only in the circle in the figure (Figure 8,9) With minor differences, the classification result at the side of the well is also consistent with the AVO type obtained by the numerical simulation, which is a type III AVO. It
shows that the results of using the two algorithms to distinguish the AVO type are accurate and reliable; 2) If we use the same input attributes as the random forest (8 feature attributes, as shown in Figure 10) when using the PSVM algorithm, the well type of AVO is not clear. The left side of the well curve is Type IV AVO, and the right is Type III AVO. It is difficult to accurately determine the AVO type of the reservoir section. In summary, the RF algorithm can achieve a very close and accurate recognition effect with the PSVM algorithm with fewer feature attributes when performing AVO type discrimination; the PSVM algorithm has inaccurate discrimination results when there are fewer feature attributes, indicating that the RF algorithm is similar Compared with the PSVM algorithm, it has stronger generalization in the discrimination of AVO types and has better universality.

![Figure 5 Partial near-well seismic trace gathers for Well 1 using forward simulation](image5)

![Figure 6 The variation of after-fitting near-well seismic data in target stratum varying with incident angles(Well 1)](image6)

![Figure 7 AVO response characteristics based on forward modeling of logging data(Well 1)](image7)
5. Conclusions

1) When the random forest algorithm is used to distinguish AVO types, when the number of feature attribute categories is within a certain range, the more types, the higher the accuracy of the discrimination. Therefore, the appropriate number of morphological feature parameters should be selected when performing AVO type discrimination.

2) The random forest algorithm uses the morphological characteristic parameters after fitting multiple original seismic data to distinguish the AVO type, reducing the ambiguity caused by a single attribute, and reducing the types of morphological characteristic parameters as much as possible, thereby reducing operational cost.

3) The application effect of random forest algorithm and approximate support vector machine algorithm in actual areas is very close, but in contrast, the former requires fewer feature attributes in the discrimination process, shorter calculation time, and stronger generalization ability. It has certain advantages in classification.
Acknowledgements
This article is one of the phased achievements of the National Natural Science Foundation of China
Deep Carbonate Reservoir Fluid Seismic Prediction Theory and Method (U1562111) and
Frequency Information Based Fluid Identification and Fluid Mobility Prediction (41774142)

References
[1] Ostrander W J. Plane-wave reflection coefficients for gas sands at non-normal angles of
incidence[J]. Exploration Geophysics, 1984, 15(3):193.
[2] Rutherford S R, Williams R H. Amplitude-versus-offset variations in gas sands[J]. Geophysics,
1989, 54(6):680–688.
[3] Castagna J P, Swan H W. Principles of AVO cross plotting[J]. The Leading Edge, 1997, 16(4):
337-342.
[4] Li Wenxiu, Wen Xiaotao, Li Tian, et al. AVO types discrimination based on a proximal support
vector machine[J]. Oil Geophysical Prospecting, 2018, 53(5):7-8+95-100.
[5] Fang Kuangnan. (2009) Random forest combination forecasting theory and its application in
finance[M]. Xiamen University Press, Xiamen, 216-217.
[6] Breiman L. Bagging Predictors[J]. Machine Learning, 1996, 24(2):123-140.
[7] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.