Lithology prediction from seismic data using Random Forests algorithm

Introduction

Using seismic data to predict lithology away from wells is an important and challenging task in hydrocarbon exploration. The common lithology prediction workflow (e.g. Larsen et al., 2006) involves cross-plot analysis, which is usually restricted by the overlapping of lithology samples. Therefore, alternative method must be developed. Recently, the Random Forests algorithm (Breiman, 2001) shows superior performance in mining high-dimensional data, and has been successfully applied to solve geophysical problems such as seismic-facies classification (Kim et al., 2018) and lithologic mapping (Kuhn et al., 2018). Predicting lithology from seismic data can also be considered as a problem of high-dimensional data mining, which aims to explore complex and non-linear relations between lithology and various seismic derived features like seismic attributes and rock-physics parameters. Random forests algorithm is suitable for lithology prediction because it excels at handling large number of input features (Zhang et al., 2018) and is superior in learning ability and robustness over other machine learning algorithms (Naghibi and Pourghsemi, 2015). However, the performance of Random Forests is usually affected by the class imbalanced problem, where the training data are dominated by the samples belonging to one class. In some real cases, the lithology samples from cores or well logs interpretation can be dominated by a certain class of lithology due to a stable deposition environment. The prediction results could bias toward the majority class of lithology if the Random Forests classifier is trained by such an imbalanced data set.

In this study, we use Random Forests algorithm to predict lithology from seismic data. We use lithology samples interpreted from well logs as training labels while seismic attributes and rock-physics parameters are corresponding features. For class imbalanced problem, we integrate synthetic minority over-sampling technique (Chawla et al., 2002) and NearMiss-1 method (Mani and Zhang, 2003) to balance the sample amount of each lithology class in the training set. The Random Forests classifier is then trained by the balanced training set, and applied to predict lithology by taking seismic attributes and seismic-inverted rock-physics parameters as input. We tested our method on field data and the results showed a promising prediction accuracy of all classes of lithology including the minority classes.

Method

To construct the lithology data set, we used lithology samples of 14 wells in the study area as labels, including salt, mudstone, limestone, dolomites, sandstone and shale. P-wave impedance, S-wave impedance, Young’s modulus, frequency change, average frequency, instantaneous bandwidth, instantaneous frequency and amplitude acceleration were used as features. The lithology samples were first converted from depth domain to time domain through seismic-well tie, then down-sampled by 2ms to match the corresponding seismic attributes and seismic-inverted rock-physics parameters. We randomly selected lithology samples of 13 wells as training set for Random Forests classifier while lithology samples of the remaining well were the test set. The amounts of lithology samples in the training set were severely imbalanced that approximately 77% of the lithology samples belong to mudstone (Figure 1a). The prediction result will bias toward mudstone if the Random Forests classifier is trained by such an imbalanced data set.

**Figure 1** The lithology sample distribution in the original training set (a) and balanced training set (b).
To balance the amounts of lithology samples in the training set, we used synthetic minority over-sampling technique (SMOTE) to over-sample minority class samples while under-sampling majority class samples with NearMiss-1 method. The SMOTE generates new minority samples in feature space by taking each minority class sample and introducing synthetic samples along the line segments joining any of the \( k \) minority class nearest neighbours:

\[
x_{\text{new}} = x + \text{rand}(0,1) \times (x' - x)
\]

where \( x_{\text{new}} \) is the coordinate of the synthetic sample in the feature space, \( x \) is the coordinate of the minority class sample in the feature space, \( x' \) is the coordinate of the minority class nearest neighbour in the feature space and \( \text{rand}(0,1) \) means a random number between 0 and 1. Here, our implementation used three nearest neighbours (\( k=3 \)). Although over-sampling minority class samples can balance class distribution, the class clusters are not well defined since some majority class samples might be invading the minority class space. Training a classifier under such a situation can lead to over-fitting (Batista, et al., 2004). To create better-defined class clusters, we applied the NearMiss-1 method to clean the majority class samples that had invaded the minority class space. The NearMiss-1 method selects the majority class samples whose average distances to the \( N \) closest samples of the minority class are the smallest, and remove them from the feature space. Here, our implementation used three closest samples of the minority class (\( N=3 \)). Through the integration of SMOTE and NearMiss-1, we are able to balance the class distribution and create better-defined class clusters. The demonstration of this method is shown in Figure 2. First, the minority class samples in the original data set is over-sampled with SMOTE, then NearMiss-1 method is implemented to undersample the majority class samples, producing a balanced data set with well-defined class clusters. For the lithology samples, we considered mudstone as majority class and other lithology classes are minority classes. We applied the SMOTE + NearMiss-1 method to balance the lithology samples, and the result is shown in Figure 1b.

![Figure 2](image1.png)

**Figure 2** Balancing a data set with SMOTE + NearMiss-1: (a) original data set; (b) over-sampling with SMOTE; (c) over-sampled data set; (d) under-sampling with NearMiss-1; (e) balanced data set.

The balanced lithology training set was then used to train the Random Forests classifier which consists of 300 decision trees. In Random Forests, each decision tree is built from a bootstrap data set which consists of samples drawn with replacement from the training set. When splitting each node during the construction of a decision tree, the best split is found using the feature and threshold that yield the largest impurity decrease at each node:

\[
\arg \max_{f \in \{f_1, f_2, ..., f_{F}\}} \left\{ I(\theta) - \left[ P_l \times I(\theta_l) + P_r \times I(\theta_r) \right] \right\}
\]

where \( f_i \) represents the \( j \)th randomly selected feature of the \( i \)th sample, \( n \) is the amount of samples in the node, \( F \) is the amount of randomly selected features, \( I(\theta_l) \), \( I(\theta) \) and \( I(\theta_r) \) represent the
impurity of parent node, left and right child node respectively, $P_L$ and $P_R$ are the probability that the samples are split to left and right child node respectively. For classification, the impurity of each node is measured by Gini index, which is given by:

$$I(\theta) = 1 - \sum_{c=0}^{K} [P(c|\theta)]^2$$

(3)

where $P(c|\theta)$ can be seen as the samples’ proportion of class $c$ in node $\theta$, and $K$ is the total amount of classes. The nodes of each decision tree will keep splitting until the Gini impurity can no longer decrease or the node contains less than two samples. Finally, all decision trees vote to decide the class of unknown samples. The random feature selection at each node of Random Forests produces a lower generalization error compared with simple bagging-classification and regression tree (CART) (Wang et al., 2020). Here, we use $F=\log_2M$ as Pavlov (1997) recommended, where $M$ is the total amount of sample features in the training set.

**Results and discussion**

After the Random Forests classifier was trained by the balanced lithology training set, we also trained another Random Forests classifier with the original lithology training set for comparison. These two classifiers were first tested by the test set respectively and the prediction results are demonstrated by confusion matrices (Figure 3). Due to that the original training set is dominated by samples belonging to mudstone, the predicted result of Random Forests classifier trained by the original training set biases toward the mudstone, which leads to low prediction accuracy of other classes of lithology. The Random Forests classifier trained by the balanced training set has a significantly improved prediction accuracy of the minority classes of lithology such as limestone, sand and shale. However, the side effect also arises that the prediction accuracy of mudstone is decreased. Next, the Random Forests classifier trained by the balanced lithology training set was used to predict lithology of the study area by taking the corresponding seismic attributes and seismic-inverted rock-physics parameters as input. The predicted lithology cube (Figure 4b) shows the spatial distribution of six classes of lithology, including thin layers and faults, which are consistent with the corresponding seismic data (Figure 4a).

![Figure 3](image)

**Figure 3** Confusion matrices of lithology prediction results of Random Forests classifier trained by the original imbalanced training set (a) and balanced training set (b).

![Figure 4](image)

**Figure 4** Seismic cube (a) and lithology cube predicted by Random Forests classifier (b).
Conclusions

In this paper, we have applied the Random Forests algorithm to predict lithology from seismic data. The Random Forests classifier can learn inherent relations between seismic attributes, rock-physics parameters and lithology through training, which can be seen as an alternative method to cross-plot analysis. We have also developed a class balancing method that integrates SMOTE and NearMiss-1 to balance the sample amount of each lithology class in the training set. The Random Forests classifier trained by the balanced lithology training set worked well to predict spatial lithology distribution from 3D seismic data, which will help lithology interpretation and reservoir characterization.

Acknowledgements

This research is supported by the National Science Foundation of China (Grant No. 41674125) and the Major Scientific and Technological Projects of CNPC (ZD2019-183-003-001).

References


Kuhn, S., Cracknell, M. J., & Reading, A. M. [2018]. Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. GEOPHYSICS, 83(4), B183-B193.


