Supervised learning to detect salt body

Conference Paper · October 2015

5 authors, including:

Pablo Guillen
University of Houston

German Larrazabal
Repsol - USA

Dainis Boumber
University of Houston

Ricardo Vilalta
University of Houston

Supervised learning to detect salt body

Pablo Guillen (University of Houston), German Larrazabal (Repsol USA), Gladys González (Repsol USA)
Dainis Boumber (University of Houston), Ricardo Vilalta (University of Houston)

Summary

In this paper we present a novel approach to detect salt bodies based on seismic attributes and supervised learning. We report on the use of a machine learning algorithm, Extremely Randomized Trees, to automatically identify and classify salt regions. We have worked with a complex synthetic seismic dataset from phase I model of the SEG Advanced Modeling Corporation (SEAM) that corresponds to deep water regions of the Gulf of Mexico. This dataset has very low frequency and contains sediments bearing amplitude values similar to those of salt bodies. In the first step of our methodology, where machine learning is applied directly to the seismic data, we obtained accuracy values of around 80%. A second (post-processing) smoothing step improved accuracy to around 95%. We conclude that machine learning is a promising mechanism to identify salt bodies on seismic data, especially with models that can produce complex decision boundaries, while being able to control the associated variance component of error.

Introduction

Seismic-data interpretation has as its main goal the identification of compartments, faults, fault sealing, and trapping mechanism that hold hydrocarbons; it additionally tries to understand the depositional history of the environment to describe the relationship between seismic data and a priori geological information. Data mining or knowledge discovery in databases (KDD) has become a significant area both in academia and industry. Data mining is the process of extracting novel, useful and understandable patterns from a large collection of data. Automated tools for knowledge discovery are frequently invoked in databases to unveil patterns that show how objects group into some classification scheme; algorithms make use of higher order statistics, feature extraction methods, pattern recognition, clustering methods, and unsupervised and supervised classification. A major strategy in this field is to apply data mining algorithms (Hastie, 2011) to classify points or parts of the 3D seismic data to reinforce correct data interpretations. Multiple studies have shown the benefits of using data mining techniques for seismic-data interpretation. For example, previous work has shown how to generate a set of seismic traces from velocity models containing faults with varying locality, using machine learning to identify the presence of a fault in previously unseen traces (Zhang et. al., 2014). Other techniques segment a seismic image into structural and stratigraphic geologic units (Hale, 2002), which is best done using global optimization methods (Shi et. al., 2000; Hale et. al., 2003). Another solution is to use unsupervised learning techniques (Coléou et. al., 2003), often relying on the application of Self Organizing Maps (Castro de Matos et. al., 2007). Our new approach is essentially a novel salt body detection workflow. The workflow as a whole envisions the creation of a software solution that can automatically identify, classify and delineate salt bodies from seismic data using seismic attributes and supervised learning algorithms. A comparison between the salt body detected and its interpretation from 3D synthetic data set testifies to the effectiveness of our approach.

Method

Automated classification of salt bodies using machine learning

Our approach aims at automatically identifying and delineating geological elements from seismic data. Specifically, we focus on the automatic classification of salt bodies using supervised learning techniques. In supervised learning we assume each element of study is represented as an n-component vector-valued random variable (X_1, X_2,..,X_n), where each X_i represents an attribute or feature; the space of all possible feature vectors is called the input space **X**. We also consider a set {w1, w2,...,wk} corresponding to the possible classes; this forms the output space **W**. A classifier or learning algorithm typically receives as input a set of training examples from a source domain, T = {(xi, wi)}, where x = (x_1, x_2,…,x_n) is a vector in the input space, and w is a value in the (discrete) output space. We assume the training or source sample T consists of independently and identically distributed (i.i.d.) examples obtained according to a fixed but unknown joint probability distribution, P(x,w), in the input-output space. The outcome of the classifier is a hypothesis or function f(x) mapping the input space to the output space, f: $\mathbf{X} \rightarrow \mathbf{W}$. We commonly choose the hypothesis that minimizes the expected value of a loss function (e.g., zero-one loss).

The challenge behind classification of seismic data

Our workflow takes as input a cube of seismic data where each voxel stands as a feature vector (we used three informative features as described below). From the whole cube we take a small fraction of representative voxels to conform a training set T = {(xi, wi)}, where x = (x_1, x_2, x_3); we assume only two classes: w_1 and w_2, corresponding to voxels inside and outside the salt body, respectively. This workflow is challenging because 1) the sheer size of the 3D data cube precludes training predictive models with more than just 1% of the available training data; this implies several regions of the cube may not be fairly represented in the training set; 2) many learning algorithms are unable to cope with millions of training examples; it took days to complete the entire data processing; 3) classification is difficult because many voxels inside and outside salt bodies have very similar appearance. In machine learning terminology this is a problem known as *high Bayes error*. The success of this workflow is clearly contingent on finding useful and informative features to appropriately discriminate among classes.

Informative attributes to generate predictive models in seismic data

A proper characterization of voxels can be attained with useful and informative features. We selected three features for our study exhibiting high correlation with the target class: signal amplitude (directly from seismic data), second derivative, and curve length; the last two derived from amplitude. Second derivative is instrumental to detect edges in images, and curve length capture patterns which characterize different features observed inside a salt structure and in its surroundings.

Supervised learning algorithms

Our data analysis phase receives as input a body of seismic data with the task of automatically identifying salt regions. We randomly sample a small fraction (0.5%) of the total data; the sample is then assigned class labels by an expert (aided by a software tool that simplifies the labeling process). To achieve a class-balanced problem, we made sure exactly one half of the subset corresponded to salt, and the other half as non-salt (the task exhibited equal class priors). The model was built using 2 million training voxels. Accuracy is estimated using 10-fold cross validation (Hastie, 2011). The classification model was subsequently used to automatically label the entire body of seismic data (376,752,501 voxels). Our top performing learning algorithms were the following: Gradient Boosting Trees (Accuracy 80%), Extremely Randomized Trees (Accuracy 80%), and Random Forests (Accuracy 79%). All our learning algorithms are ensemble methods; these techniques have shown remarkable performance due to their ability to attain low bias (using complex decision boundaries), and low variance (achieved by averaging over various models).

Example

We have tested our proposed technique using SEAM I (SEG Advance Modeling Corporation) data. This comes from marine acquisition and represents strong challenges to the geophysical community. Inspiration was deep water (600 – 2000 meters) US GOM Salt Structure and its major structural features are salt body with rugose top and overhangs, twelves radial faults near the root salt, overturned sediment raft proximate to salt root and internal sutures and a heterogeneous salt cap. The migrated seismic volume was obtained with very low frequency, and there are sediments locations with similar amplitude value than salt body. A migrated seismic volume with these kinds of features is very complex for detection of salt body. Mathematical and machine learning algorithms were taken from Python's Numpy and Scikit-learn libraries, respectively. Our final predictive model of choice was Extremely Randomized Trees, which was used to predict the labels of 376,752,501 samples; this resulted in a Boolean mask. The accuracy reported was essentially the same as in cross validation (80%). After that, we have removed outliers and misclassification using mathematical morphological operations and a 3D interactive guided (manual intervention) tool developed in house; finally, we used threshold segmentation using local average threshold to get better detection results.

Results

We describe our results by visually comparing our predictions on a cube of seismic data. Figure 1(a) shows a cross section of the seismic data, figure 1(b) shows the classification obtained with our proposed methodology, and figure 1(c) shows the classification after the post-processing step.

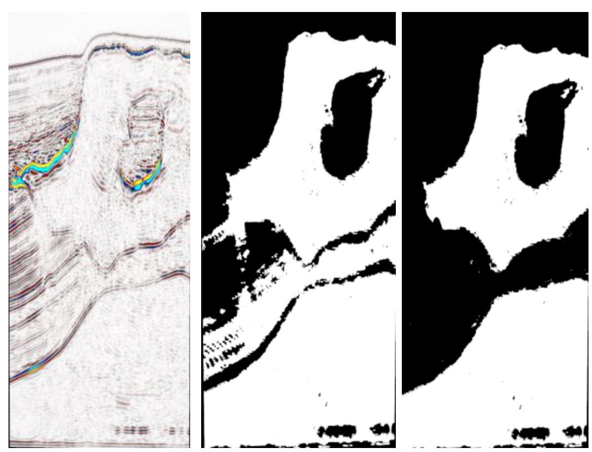

Figure 1: (a) Seismic data, (b) classification using our method, (c) results obtained with a post-processing step.

Figure 2 shows the overlapping between seismic data and salt body (white color) detected on different inline locations. We can observe that seismic attributes used in combination with the machine learning algorithm allows capturing and classifying different patterns and features between sediments and salt body.

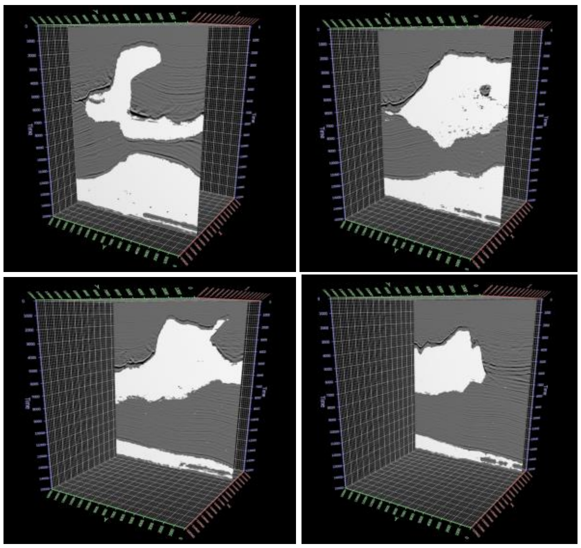

Figure 2: Overlapping between seismic data and salt body detected.

To measure accuracy, we count the number of hits between the detected salt body and the interpretation in the following way: using both volumes, we have counted the number of hits voxel by voxel. We refer to this number as NH. The effectiveness ratio is calculated as: (NH/TS) * 100, where TS is the total number of voxels in the volume. Following this technique, we have obtained an accuracy of 95.22%.

Figure 3 shows a comparison between our salt body detected (white color) and its interpretation (red color). We can see the promising quality of our detection for the synthetic seismic dataset used in this work.

Conclusions

We have shown an efficient approach to classify salt bodies from a very complex synthetic seismic dataset using machine learning techniques. Results show very high accuracy when machine learning algorithms are used to predict class labels of voxels on a seismic cube; this is true even after training with a very small portion of the data (0.5%). After a first step, where machine learning is applied directly to the data, we obtained accuracy values of around 80%. A second (post-processing) step increased accuracy to around 95%. We conclude that machine learning is a promising mechanism to identify geological bodies on seismic data when the selected model has high capacity, and is able to control the variance component of error by model averaging (using ensemble techniques).

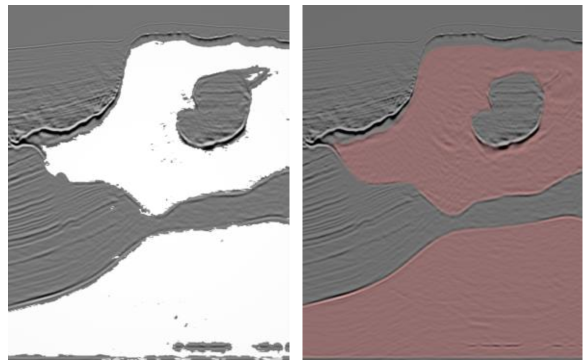

Figure 3: Overlapping between seismic data. (a) salt body detected, and (b) interpretation.

Acknowledgments

The authors thank Repsol for its support and for the authorization to present this work. We would like to acknowledge also SEG Advanced Modeling Corporation (SEAM) for their initiative on creating a realistic salt model and seismic data used for this study.

References

Castro de Matos, M., Manassi, P. L., Osorio, Schroeder, P. R. 2007. Unsupervised Seismic Facies Analysis using Wavelet Transform and Self Organizing Maps. GEOPHYSICS, **72** (1) pp. 9-21.

Coléou, T., Poupon, M., Azbel, K. 2003. Unsupervised Seismic Facies Classification: A Review and Comparison of Techniques and Implementation: The Leading Edge, **22**, pp. 942–953.

Hale, D., Emanuel, J. 2003. Seismic Interpretation using Global Image Segmentation. 73th Annual International Meeting, Society of Exploration Geophysicists.

Hale, D., 2002. Atomic Meshes from Seismic Imaging to Reservoir Simulation. Proceedings of the 8th European Conference on the Mathematics of Oil Recovery, Freiberg, Germany.

Hastie, T., Tibshirani, R., Friedman, J. 2011. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition, Springer.

Shi, J., Malik, J. 2000. Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, **22** (8), pp. 888–905.

Zhang, C., Frogner, C., Araya-Polo, M., Hohl, D. 2014. Machine Learning Based Automated Fault Detection in Seismic Traces. Proceedings of 76th EAGE Conference and Exhibition, Amsterdam.