Active learning algorithms in seismic facies classification

Summary
In this paper we illustrate unsupervised and supervised learning algorithms that accurately classify the lithological variations in the 3D seismic data. We demonstrate blind source separation techniques such as the principal components (PCA) and noise adjusted principal components in conjunction with Kohonen Self organizing maps to produce superior unsupervised classification maps. Further, we utilize the PCA space training in Maximum likelihood (ML) supervised classification. Results demonstrate that the ML supervised classification produces an improved classification of the facies in the 3D seismic dataset from the Anadarko basin in central Oklahoma.

Introduction
3D Seismic data interpretation and processing, unlike most other sensing paradigms, resides at the confluence of several disciplines, including, but not limited to, wave-physics, signal processing, mathematics and statistics. It is not coincidental that algorithms for classification of 3D seismic facies emphasize the strengths and biases of the developers. But, the lack of a common vocabulary has resulted in many duplicate efforts. In this sense, the taxonomies strive to distill the universe of known algorithms to a minimal set. Moreover, we hope the taxonomies provide a framework for future algorithm development, clearly indicating what has and has not been done.

The classification algorithms discussed in our paper is based upon how the learning criterion classifies observed data. Under supervised learning schemes, the classes are predetermined. These classes can be conceived of as a finite set of observations previously arrived at by an expert’s intervention. In practice, a certain segment of data are labeled with these classifications. The machine learner’s task is to search for patterns and construct mathematical models. These models then are evaluated on the basis of their predictive capacity in relation to measures of variance in the data itself.

Unsupervised learning schemes on the other hand are not provided with classifications (labels). In fact, the basic task of unsupervised learning is to develop classification labels automatically. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group. These groups are termed clusters, and there is a plethora of clustering machine learning techniques. In the following section we will provide a set of unsupervised and supervised learning algorithms that accurately classifies lithological structures in 3D seismic data. In the next section we will describe our unsupervised latent space modeling technique, which are based on Kohonen Self Organizing, maps.

Unsupervised latent space modeling
There are many possible techniques for coping with the classification of data with excessive dimensionality. One of the approaches is to reduce the data dimension by combining features. Linear combination to reduce data dimensionality is simple to compute and analytically tractable. One of the classical approaches for effective linear transformation is known as the principal component analysis (PCA) – which seeks the projection that best represents the data in the least square sense. This transform belongs to a small class of image and general signal processing tools, namely, the class of orthogonal transform. Here the multidimensional dataset is projected in a lower 1D or a 2D manifold that approximately contains the majority probability mass of the data. This lower dimensional manifold is known as the latent space.

Kohonen Self-organizing map (SOM) (1982) is one of the most popularly used unsupervised pattern recognition techniques. In the SOM algorithm the neighborhood training of the nodes takes place in a 2D latent space such that the statistical relationship between the multi-dimensional data and the trained nodes are preserved. There are several ways to initialize the SOM training process. The classical way to initialize the SOM training nodes is random initialization of the nodes and arranging them in a predefined 1D or a 2D grid. One of the characteristics of SOM is that the resultant output depends on the initial definition of the latent space. We have compared the SOM outputs from two different ways initialization the 2D latent space with the grid points having the training nodes or the prototype vectors associated with each of the points.

One way to define the initial map of the training nodes (prototype vectors) is to calculate the eigenvectors and the eigenvalues from the covariance matrix created from the input dataset. Then we calculate the eigenvectors and eigenvalues of the covariance matrix. The 2D latent space containing the prototype vectors is defined by considering three standard deviations of the variability (square root of the eigenvalues λ1 and λ2) along the two principal component directions (eigenvectors v(1) and v(2)). Thus this latent space represents approximately 99.7% of the
Active learning algorithms for seismic facies classification

input dataset (Roy et al., 2011). The grid points are uniformly spaced in this latent space. Also these grid points are associated with a prototype vector which we pre-define as the proportional sum of the first two eigenvectors $\mathbf{v}(1)$ and $\mathbf{v}(2)$.

For the second workflow we did iterative PCA analysis. We calculated all the PCA projections of the data. The last PCA component was mostly noise and was subtracted from each of the PCA component of the dataset except to itself. Then we define a new data-vector merging all the modified PCA components. This new noise adjusted dataset is used subsequently to initialize a new latent space for SOM training to take place. We calculate the eigenvalues and the eigenvectors of this new covariance matrix and define the latent space as before.

Unsupervised seismic facies analysis with different initial conditions

We demonstrate the unsupervised SOM workflow on a seismic survey from eastern part of the Anadarko basin in central Oklahoma. Our target zone was the Middle Pennsylvanian Red Fork Formation characterized by three coarsening upwards marine parasequences (Peyton, 1998). The Red Fork Formation has multiple stages of incised valley fills. However our analysis is confined to the Upper Red Fork incised valleys, which can be properly imaged on the 3D seismic. Figure 1 shows the geological interpretation of the coherency slice showing different stages I to V of incised valley fill in the area.

Figure 1: Final interpretation on the coherence slice showing the different stages of the upper Red Fork Formation (Peyton et al., 1998)

We considered a 30ms zone below the flattened skinner horizon as the input data. It has seven samples thus there are 7 PCA components of the data. We subtracted the $7^{th}$ PCA component from all the other PCA components except with itself and did the iterative PCA analysis. The noise adjusted iterative PCA analysis, gave a better interpretation and clarity of different stages of the Red Fork (highlighted in red arrows) than the PCA analysis (Figure 2). This motivated us to initialize our two separate latent space models for SOM training.

Figure 2: Comparison of the 1st component of the PCA analysis (left) and the 1st component of the noise adjusted iterative PCA analysis (right). The output for the noise adjusted PCA is better in interpretation of the different stages of the upper Red Fork Formation.

Figure 3 and figure 4 are the seismic facies maps from SOM training from the set 1 (eigenvalues and eigenvectors from initial PCA analysis) and set 2 (eigenvalues and eigenvectors calculated after the first PCA) respectively. First assigning a gradational 2D HSV colorscale to the trained prototype vectors in the latent space does the coloring. Then different colors are assigned to each $x, y$ location of the dataset according to the similarity of the data vectors with the Prototype Vectors (Roy et al., 2010).

Figure 3: Unsupervised seismic facies analysis after the SOM training in the latent space vectors initialized from the 1st set of eigenvalues and eigenvectors from the dataset. Different arrows show different depositional stages.
Active learning algorithms for seismic facies classification

Figure 4: Unsupervised seismic facies analysis after the SOM training in the latent space vectors initialized from the 2nd set of eigenvalues and eigenvectors from the noise adjusted iterative PCA dataset. The different stages of the Red Fork formation are more clearly visible in this analysis. The channel shows more variation in colors than the previous one. Stage II and stage V are more defined in this analysis (red arrow)

Supervised Maximum Likelihood Classifier

In this section, we shall illustrate the use of the Maximum likelihood supervised learning technique and shall provide classification maps generated due to its training.

This algorithm is based on statistics such as the mean and variance/covariance; a PDF (Bayesian) is calculated from the inputs for classes established from training sites. The ML classifier assumes that the statistics for each class in each time slice are normally distributed and calculates the probability that a given data-vector belongs to a specific class. Unless a probability threshold is selected, all data-vector are classified. Each pixel is assigned to the class that has the highest probability. If the highest probability is smaller than a threshold, the data vector remains unclassified. The following discriminant functions for each data-vector in the time slice are implemented in ML classification:

\[ g(X) = -\frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i) \]

where, \( i \) represents the number of class, \( X \) is a \( n \) dimensional data (where \( n \) is the number of samples present in the seismic data), \( |\Sigma_i| \) is the determinant of the covariance matrix created from the data and \( \Sigma_i^{-1} \) is its inverse.

Application of Supervised ML

The scatter plot in Figure 5 is obtained by projecting the first two principal components of the earlier dataset. The scatter plot essentially depicts the 2D representation of endmember selection. Extreme samples (maximum distance away from the cluster) which ultimately correspond to endmembers and can be determined by rotating this scatter plot in n-dimensions. Thus we are able to partition the feature space based on interclass separation statistics. These color coded samples are then utilized for training. Finally, as shown in Figure 6 we obtain the classification based on Maximum Likelihood.

Figure 5: The scatter plot from projecting the first two principal components. Class 1 (red), class 7 (magenta) and class 9 (purple), which are farthest from the cluster center, represent the endmember classes. These endmember classes represent the purest data-vector of the discrete classes.

Figure 6: Supervised seismic facies analysis from the Maximum Likelihood output. The colors correspond to the classes defined in the features space (Figure 5). The red facies (brown arrow), the magenta facies and the purple facies (cyan arrow) correspond to the discrete classes. Different depositional stages and the geological features are better highlighted in this facies map.
Active learning algorithms for seismic facies classification

Discussions

In both the unsupervised classification techniques, the initial numbers of clusters were over-defined to 256 initial classes. With subsequent iterations the classes clustered to form seismic facies map which we can visually interpret into different geological features. The edges of the different stages of the incised valley filled channels are better visualized in the unsupervised analysis on noise adjusted principal component space.

The supervised classification with Maximum likelihood generates a superior seismic facies map after an accurate and easy training in the feature space. The most isolated points in the scatter plot represent the purest endmember classes (in red, magenta and violet colors) and assists in detecting small variation of the seismic facies from the rest of the interclass variations.

Conclusion

We propose a workflow where we identified different classes in the feature space instead of identifying the facies geologically. The resultant supervised seismic facies map is confirmed with the unsupervised analysis and the depositional history of the area. In a situation where inadequate geological knowledge is an issue this workflow can be recommended for generating seismic facies maps.

Acknowledgements

Thanks to Chesapeake for permission to use and publish their seismic and well data from the Anadarko Basin. We also thank the sponsors of the OU Attribute-Assisted Processing and Interpretation Consortium and their financial support.
EDITED REFERENCES
Note: This reference list is a copy-edited version of the reference list submitted by the author. Reference lists for the 2013 SEG Technical Program Expanded Abstracts have been copy edited so that references provided with the online metadata for each paper will achieve a high degree of linking to cited sources that appear on the Web.

REFERENCES