Unsupervised machine learning applications for seismic facies classification

Satinder Chopra[†], Ritesh Kumar Sharma[†] and James Keay[+]

[+]TGS, Calgary, [+]TGS, Houston

Objectives

The objective of the present study was to apply several unsupervised machine learning (ML) methods such as K-means, principal component analysis (PCA), self-organizing mapping (SOM) and generative topographic mapping (GTM) techniques to seismic data volumes from the Barents Sea, the STACK area in Oklahoma, and the Utica shale play in eastern Ohio, and assess their performance.

Methods

Unsupervised ML uses the attributes themselves as both training data and data to be analyzed. The simplest algorithm is *K-means*, wherein the interpreter defines the number of facies (clusters) to be found. The algorithm then finds means and standard deviations (more generally, covariance matrices) to determine the center and the extent of each cluster in multidimensional attribute space, and thus generates different clusters.

PCA is a useful dimensionality reduction tool. The amount of attribute redundancy is measured by the covariance matrix, which is solved in terms of eigenvalues and eigenvectors. By ordering the eigenvalues from the highest to the lowest, the principal components for the data under investigation are determined. Since seismic attributes are correlated through the underlying geology and the band limitations of the source wavelet, the first two or three principal components will almost always represent most of the data variability.

The Kohonen SOM is a technique that generates a seismic facies map from multiple seismic attributes. In a distribution of *N* attributes lying in an *N*-dimensional data space, the plane that best fits the data is defined by the first two eigenvectors of the covariance matrix. This plane is then iteratively deformed into a 2D surface called a manifold that better fits the data. After convergence, the N-dimensional data are projected onto this 2D surface, which in turn are mapped against a 2D plane or “latent” (hidden) space, onto which the interpreter explicitly defines clusters.

The SOM mapping described above has limitations. There is no theoretical basis for selecting the training radius, neighborhood function and learning rate as these parameters are data dependent. No cost function is defined that could be iteratively minimized and would indicate the convergence of the iterations during the training process, and finally no probability density is defined that could yield a confidence measure in the final clustering results. An alternative approach to SOM mapping that overcomes its limitations is the *generative topographic mapping* (GTM) technique.

Conclusions

Many high amplitude anomalies in the Barents Sea, appear to be associated with high porosity and low Vclay values at different levels. We found that K-means and principal component displays showed the facies distribution on the SOM display, the most detailed distribution of the facies and their distinct definition was seen on the GTM display. In the other datasets at the Woodford and Utica shale level, again the SOM and GTM ML techniques fared better than the other two. Between these two the GTM again takes the lead.

The application of k-means, PCA, and SOM methods have been demonstrated before on different datasets, but as individual applications. We demonstrate for the first time how the comparison of four different ML techniques applied on each of the three data volumes and exhibit their performance.