UNSUPERVISED SEGMENTATION IN SEISMIC DATA ANALYSIS

B. Grennberg Fismen, S. Clausen,
L. Yang, M. Carlin, T. Kavli

SINTEF electronics and cybernetics
Pb 124 Blindern
NO-0314 Oslo, Norway

G. Wansink

de Groot-Bril Earth Sciences BV
Boulevard 1945-24
7511 AE Enschede, The Netherlands

ABSTRACT

Segmentation of seismic data is an important visualisation tool which helps geologists to interpret data. Important issues for the design of a reliable seismic data segmentation system are selection of features, choice of segmentation methods, and criteria for evaluating the results. We present some results addressing the above issues.

Real seismic data were used in our experiments. Seismic features were calculated from the seismic traces and used as input for segmentation. The size and content of the feature set were determined from a correlation analysis. Two segmentation methods were tested and evaluated; the hard c-means clustering (HCM) algorithm and the complete linkage (CL) clustering algorithm. The number of clusters was suggested by using Dunn's index for cluster validation. The HCM algorithm is most effective, but the CL algorithm is deterministic and allows hierarchical access to the data. The algorithms give visually comparable results.

1. INTRODUCTION

Interpretation of seismic data has played an important role in oil exploration. Seismic data provides an image of the sub-surface by recording at the surface the reflections of acoustic waves that were emitted into the subsurface by exciting a high energy source such as dynamite (onshore) or airgun (offshore). Huge amounts of 2D and 3D seismic data are recorded by using modern acquisition techniques. Usually there exists no a priori information about the geological structure, and performing a systematic analysis of all the seismic data is time consuming.

By applying clustering methods to a seismic data set it will be partitioned into g groups or clusters, with similar data placed in the same group. Ideally, each cluster should correspond to areas with similar geological structure. The number of clusters usually needs to be predefined by the user. However, for some algorithms the number of clusters is suggested by the algorithm itself.

Limited work is published in this field. Shen et al [5] segments seismic images by the single linkage hierarhical clustering method. The single linkage method tends to form long and loosely connected clusters, which join each other easily. Köster et al [6] extracts features both by using Gabor filter banks and by computing the instantaneous frequency. A bottom-up region-growing method is used to cluster the features. This method is compared to the single linkage algorithm, and the region-growing method performs better.

We study two clustering algorithms applied to seismic data, the hard c-means (HCM) and the complete linkage (CL) algorithm [4]. The c-means algorithm place most of the clusters in high-density areas, whereas the CL algorithm spreads the clusters more evenly in data space. The HCM algorithm is non-deterministic, and the resulting partitioning will vary with initialisation. The CL algorithm is deterministic and builds a hierarchy of clusters starting on a local level by merging nearby points. In order to validate the different partitions we calculate a generalised Dunn's index [1] to estimate the number of clusters.

We show that the CL algorithm is promising for analysing continuous seismic data without any well-separated clusters. It is robust, deterministic and allows hierarchial access to the data.

The rest of this paper is outlined as follows: In section 2 we describe our data sets in more detail. In section 3 the HCM and CL clustering algorithms are described together with Dunn's validity index. In section 4 we present our results and finally, in section 5 we make some concluding remarks.

This work was conducted in the context of the SMART-3D project that is funded in part by the EC under the THERMIE program (OG/ 183/98). SMART-3D is a partnership between Dutch-based dGB, Norwegian Sintef and IES from Germany. The project is further supported by RWE-DEA from Germany.

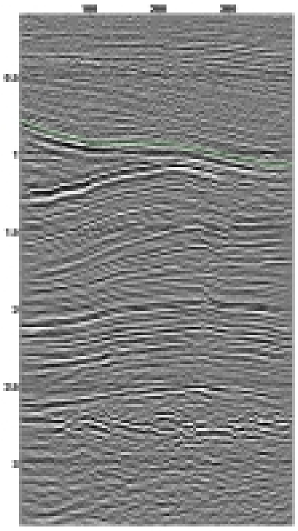


Fig. 1. Seismic amplitude for line A.

2. DATA ACQUISITION

The seismic data are from onshore Germany. We study two lines from a 3D volume (hereinafter referred to as lines A and B). Each seismic trace consists of 851 data points. The resolution is 4 ms, giving us a total time span in depth of 3.5 s. Seismic features based on the amplitude, frequency, phase information of the seismic traces, and on the similiarity between neighboring traces, were extracted from the raw data by applying a sliding window technique along each trace [2]. 17 different features were calculated in windows of varying size centered around each point.

3. METHODS

3.1. The hard c-means clustering algorithm

The HCM algorithm [4] partitions the data set by minimising the following squared-error criterion:

$$J(U,V:X)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}\|\mathbf{x}_k-\mathbf{v}_i\|^2 \qquad (1)$$

where $X=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n)$ is the set of n sample points and
$V=(\mathbf{v}_1,\mathbf{v}_2,...,\mathbf{v}_c)$ denotes the c unknown cluster centres. U is a crisp $c \times n$ partition matrix consisting of zeros and ones. We use the L_1 norm throughout this paper. Minimisation of Equation 1 is obtained if

$$u_{ik}=\begin{cases}1; & \|\mathbf{x}_k-\mathbf{v}_i\|\leq\|\mathbf{x}_k-\mathbf{v}_j\|, j=1,...,c \\ & 1\leq i\leq c; 1\leq k\leq n \\ 0; & otherwise\end{cases} \qquad (2)$$

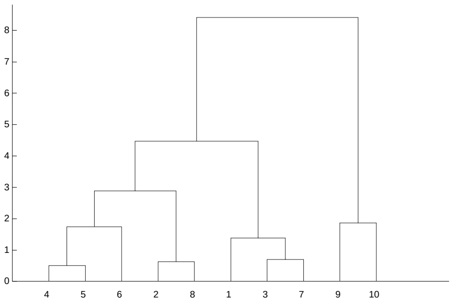


Fig. 2. Dendrogram plot showing a hierarchy of 10 clusters.

and

$$\mathbf{v_i}=\frac{\sum_{k=1}^{n}u_{ik}\mathbf{x}_k}{\sum_{k=1}^{n}u_{ik}}, 1\leq i\leq c. \qquad (3)$$

The HCM algorithm is based on iteration through Equation 2 and 3. One starts the algorithm by choosing c cluster centres randomly. In our case we randomly pick c distinct points from X as our initial cluster centres. The resulting partition is calculated by 2 and the cluster centres are updated according to 3. The algorithm is iterated until the cluster centres stop moving, i.e., until $\|V_t - V_{t-1}\| \leq \epsilon$, where V_t denotes the position of the c cluster centres after t iterations.

3.2. The complete linkage clustering algorithm

The complete linkage algorithm (CL) builds a hierarchy of clusters by merging samples and clusters, thus forming clusters of clusters [4]. The result can be visualized in a dendrogram plot; see Figure 2.

The merging decision is based on calculation of the intercluster distances. The single linkage defines the distance between the clusters as the closest pair of sample points in each cluster (also called the nearest neighbour algorithm) [3], [5]. It tends to produce long and loosely connected clusters, and the algorithm is sensitive to outliers. The class of averaging algorithms, where intercluster distances are defined from some average measure of the cluster, are less sensitive to outliers than the single linkage algorithm. The complete linkage (CL) algorithm defines in contrast to the single linkage the furthest distance between samples in each cluster as the intercluster distance (4):

$$\delta(X_i,X_j)=max(d(\mathbf{x}_i,\mathbf{x}_j)),\mathbf{x}_i\in X_i,\mathbf{x}_j\in X_j \qquad (4)$$

where X_i denotes the set of feature vectors belonging to cluster i. The complete linkage algorithm generally finds tight, hyperspherical clusters that join others only with difficulty. These qualities makes the complete linkage distance measure well suited for our application.

By cutting the resulting dendrogram tree at a specified horizontal level, and follow each branch to its leaf (i.e. the original samples), all samples are assigned to a cluster. The cluster centres $V = (\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_n})$ are defined as the centroid of all its sample members:

$$\mathbf{v_i} = \frac{1}{n_i} \sum_{\mathbf{x} \in X_i} \mathbf{x} \tag{5}$$

3.3. Cluster validation

In order to validate a resulting crisp partition U there are several different validity indexes that can be calculated. All these indexes try in one way or other to maximise the ratio between intercluster distance δ and cluster diameter Δ. Dunn's index (DI) is based on geometrical considerations and is designed to identify sets of clusters that are compact and well separated [1]. Simply stated it evaluates the ratio between the shortest intercluster distance and the largest cluster diameter and is defined by:

$$\nu_D(U) = \underbrace{min}_{1 \leq i \leq c} \left\{ \underbrace{min}_{1 \leq j \leq c} \left\{ \frac{\delta(X_i, X_j)}{\underbrace{max}_{1 \leq k \leq c} \Delta(X_k)} \right\} \right\} \tag{6}$$

Dunn used the minimum distance between points in a pair of set as a measure of the intercluster separation δ. This measure of interset distance is sensitive to outliers in the clusters and in [3] the Dunn's index is generalised by using more robust estimates of both cluster diameter and intercluster distance.

The cluster diameter Δ of cluster i is defined to be the average Euclidean distance between all points X_i belonging to the cluster and the cluster centre $\mathbf{v}_i$:

$$\Delta(X_i) = \frac{2}{n_i} \sum_{\mathbf{x} \in X_i} \|\mathbf{x} - \mathbf{v}_i\| \tag{7}$$

where n_i is the number of samples belonging to cluster i and the multiplier of 2 converts a radius to a diameter.

4. RESULTS

4.1. Feature selection

In general, clustering of data gives bad results if the clustering features are not selected with care. We have used a rough feature selection method based on a correlation analysis of the data. Our goal is to remove redundant features, and features without structure.

In order to find redundant features we first calculate the $n \times n$ correlation coefficient matrix C. Here C_{ij} is the cross correlation between traces of feature i and feature j evaluated at several different cross-line positions. The matrix reveals that the different energy features are redundant. This is also the case for the three different similarity features. In addition, the two amplitude features and the two gradient features are strongly correlated with the energy features. Therefore it is enough to choose one of these features for the clustering. Noisy features (phase and count zero crossing) were disregarded. We chose energy and similarity in a $[-40, 40]$ window as our two final features.

Fig. 3. Dunn's index for 2-15 clusters with the CL algorithm (highest Dunn's index for $c < 7$) and the HCM algorithm.

The data set was normalized to zero mean and unit variance.

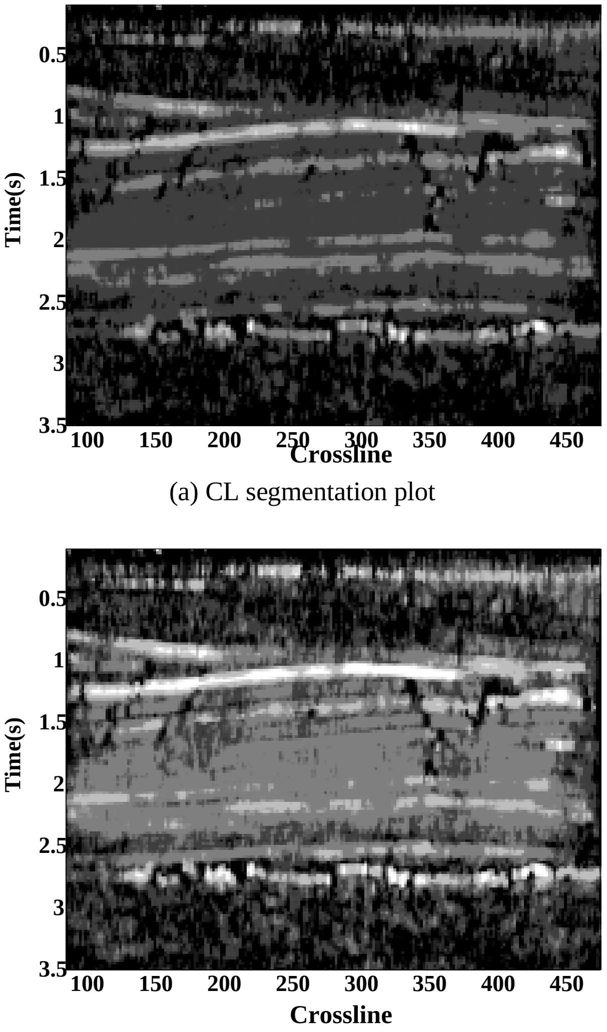
4.2. Training data set

The data sets we consider in this paper consist of $390 \times 851 = 331890$ sample points. In order to speed up the execution time, the clustering algorithms are run on subsampled parts of the complete data sets. Such an approach is quite similar to what is done in supervised learning or classification, where the algorithms learn by adapting to a limited set of training points. For the algorithms to perform well it is then important that the training set is representative to the whole data set. In the examples shown below every 20th point in the horizontal and vertical directions is subjected to the clustering algorithms, i.e., a total of 830 points. By doing so, the computing time is reduced drastically. After clustering the training data, all the other sample points in the data set are presented to the resulting partition in turn, and placed into the most nearby cluster according to a Euclidean distance measure. Increasing the density of training samples does not alter the clustering results by much.

The final segmentation plots are obtained by transforming back to real space. The sample points are coloured according to which cluster they belong to.

4.3. Clustering results

Figure 3 shows a plot of Dunn's validity index (DI) for $c = 2, 3, ..., 15$ clusters for both the HCM algorithm and



(a) CL segmentation plot

(b) HCM segmentation plot

Fig. 4. Segmentation results of line A for $c = 6$.

the CL algorithm. The HCM values are obtained by averaging over 50 different initialisations for all values of c. According to the DI values, the different partitions do not vary much. The largest variation is found for $c = 8$, where the standard deviation $s(DI) = 0.05$ for the 50 initialisations. Visual inspection of the $c = 8$ segmentation plots for the best and worst partition reveals very little difference. According to Dunn's index, the best partition of the data set is obtained with $c = 2$ clusters for both algorithms. However, geologists are interested in finer details; typically they cluster 6-10 clusters. Dunn's index at $c = 6$ indicates that the CL clustering is better than the HCM clustering. This can be seen in the segmentation results in Figure 4, where the CL plot is smoother with more well-defined layers than the HCM plot.

The approach presented in this paper was also applied to line B. The segmentation results were similar.

5. CONCLUDING REMARKS

We have studied two different clustering algorithms applied to seismic data: the hard c-means clustering algorithm and the complete linkage algorithm. The performance of the algorithms was evaluated by calculating Dunn's validity index. According to this index, the best partition of the data sets is obtained with $c = 2$ clusters. This coincides with a visual inspection of the scatter plots.

The CL algorithm spreads the cluster centres more evenly, whereas the HCM algorithm places most of the cluster centres in high density areas. According to Dunn's validity index, the CL algorithm performs better when the number of clusters is small (in our example less than 7). This observation agrees qualitatively with the resulting segmentation plots: for small values of c the CL algorithm seems to reveal more of the geological structure than the HCM algorithm, whereas for high values of c the HCM algorithm gives better segmentation plots. The HCM algorithm is more efficient, but as the CL algorithm gives hierarchial access to the data it is suitable when both an overview of the data as well as detail inspection of segments is requested.

6. REFERENCES

[1] J. C. Dunn,*"A fuzzy relative to the ISODATA process and its use in detecting compact and well-separated clusters"*, J. Cybern., 3(3):32-57, 1973.

[2] J. H. Justice, D. J. Hawkins and G. Wong, *"Multidimensional attribute analysis and pattern recognition for seismic interpretation"*, Pattern Recognition 18(6):391-407, 1985.

[3] J.C. Bezdek and N.R. Pal, *"Some new indexes of cluster validity"*, IEEE Trans. Syst., Man, Cybern. - part B: Cybernetics, 28(3):301-315, 1998.

[4] B.D. Ripley, *"Pattern recognition and neural networks"*,Cambridge university press, 1997.

[5] X. Shen, M. Spann, P. Nacken, *"Segmentation of 2D and 3D images through a hierarchiacal clustering based on region modelling"*, Pattern Recognition, 31(9):1295-1309, 1998.

[6] K. Köster and M. Spann, *"Unsupervised segmentation of 3D and 2D seismic reflection data"*, Int J. Pattern Recogn. Artif. Intell., 13(5):643-663, 1999.