Defining depositional environments by using neural networks

Janina Horváth

Department of Geology and Palaeontology, University of Szeged, H-6722, Szeged, Egyetem u. 2–6, Hungary; (th.janina@geo.u-szeged.hu)

doi: 104154/gc.2011.21

1. INTRODUCTION

The neural network approach is a well-known development tool, which became popular within the last couple of decades. Supervised and unsupervised trainable networks are used in many different fields of geology, especially petroleum geology or facies analysis e.g. the unsupervised network as a tool for lithofacies identification (CHANG et al., 2002), application of a supervised neural network for predicting permeability from porosity (ROGERS et al., 1995), using a supervised neural network tool in reservoir characterization (AHMED et al., 1997). However, supervised networks are more frequently used, because the unsupervised learning networks can solve specific problems like clustering or, pattern recognition problems as a way of defining spatial patterns. Application of a particular unsupervised neural network called Kohonen’s network (in other words Self Organized Map, abbreviated SOM) in the identification spatial pattern of some delta-plain sub-environments is demonstrated here. Kohonen’s neural network has been successfully applied to studies with different representation methods or as a tool to define clusters e.g. a well log interpretation model for the determination of reservoir facies and fluid contents (AKINYOKUN et al., 2009), lithofacies identification (CHANG et al., 2002), and classification of biogenic sedimentation (ULTSCH et al., 1995).
2D or 3D geometry of the depositional facies identified. This is the sedimentary environment according to PETTIJOHN and POTTER (1972).

In this study, the main emphasis is given to the following issues: (1) Demonstration of how a known depositional geometry outlined by sand-content contours can be honored by using SOM analysis of points of three petrophysical grids. This analysis relies on good average porosity, net pore volume and hydraulic conductivity grids; (2) Description of an example of the application of this robust approach in re-recognizing the known distributary mouth bar shape in a particular Pannonian reservoir.

2. APPLIED METHOD

2.1. Self organized maps (SOM)

Kohonen’s neural network (or Self Organized Map (SOM)) means that the neurons are organized in a grid (Fig. 1) as a ‘map’. However it is not a real map since it does not assign any spatial coordinates to the samples. So, self organizing gives a ‘map’ where the nearby locations represent inputs with similar properties.

SOM is a type of competitive and unsupervised network. The competitive and unsupervised learning algorithm implies that the network has to have the ability to recognize the structure of a multidimensional basic data set, by the method of dimension reduction (KOHONEN, 1982; KOHONEN 2001; HAYKIN, 1999; PATTERSON, 1996; FAUSETT, 1994). This reduction is only a “queasy” one, since each neuron is an n-dimensional weight-vector, where n is equal to the dimension of the input vectors (Fig. 1). The main goal of SOM is to represent data points with fewer representatives preserving the original topology (JOHANNSON, 2003).

Even the unsupervised learning character implies that only one data set is available for analysis. So the data structure is explored within the input set without using any reference dataset. In the case of the supervised neural network, the training processes depend on training examples and each example is a pair of an input objects. In contrast with the unsupervised network, the learning processes are based on unlabelled examples. The framework (Fig. 2) of Kohonen’s Neural Network also demonstrates this difference between the supervised and unsupervised neural networks. There are no real output layers, but neurons are arranged into a grid of dots. In other words, the grid is the output layer. This is the so-called Kohonen’s layer. It contains m neurons; each neuron includes one output cluster. Figure 1 shows that in this layer neurons give a two dimensional map (usually we can use one or two dimensional Kohonen layer), however the neurons saved and represent the original data dimension because each neuron is represented as an n dimensional vector, where n is the dimension of the input, training, vectors.

Weighted $x_i$ inputs are processed by all three neurons (Fig. 2). Each neuron computes its weighted input, but only the neuron with the largest excitation is the winner. These units are adjusted to cluster the training data using the activation values of winning neurons. In Kohonen’s layer the other units are reserved by this active element through the lateral connections during training shown in Fig. 2.

An advantage of the SOM network is that it attempts to learn the structure of the original data. It means that the training data set gives the contact between the original data points and identified classes. If new data, unlike previous cases, is encountered, the network fails to recognize it which indicates novelty. Therefore, in the next learning step, the grid is updated based on the novelty. The updating process is quite similar to that of the K-means algorithm, where the arrangements of cluster-centroids change in the regular low-dimensional grid. However, in case of SOM, changing the position of neurons in the data space influences the positions of its neighbors through the lateral connections within a predetermined distance.

Since SOM consists of neurons organized on a regular low-dimensional grid, these neurons control the classification of data using a suitable distance. The input vectors are compared step by step with the reference vectors using a neighboring function and the grid is revised continuously using a winning neuron algorithm. During the comparison
process, each sample input vector is randomly drawn from the input data set, and their similarities to the ‘codebook’ vectors are computed. This process is based on Euclidean distance that measures distance from the data sample to each neuron to obtain the winning neuron (LAMPINEN et al., 2005). However better selections use either Mahalanobis distance, (developed by MAHALANOBIS in 1936), or cosine distance as alternative measures. The latter is derived from correlations between variables in the property space.

Kohonen’s network learns iteratively and uses the modified Hebb-rule as the winning neuron algorithm (LAMPINEN et al., 2005).

The modified law is \( w(t+1) = w(t) + \rho(t) \times (x - w(t)) \). Where \( x \) is the training case, \( w(t) \) is weight, and \( \rho(t) \) is the learning rate.

The application of neural networks is relatively simple. An artificial network studies the data structure, which means that it necessarily must have some preliminary empirical knowledge e.g. on appropriate size of network, rate of learning, number of learning steps. In this case, the size of the network corresponds to the number of neurons representing the clusters of which numbers must be given previously. This setting depends on preliminary expertise.

2.2.1. Pattern recognition as a clustering method

SOM is a type of clustering method. It differs from any traditional clustering method in its robustness. However it needs, like any ‘traditional k-mean algorithm’, to input the number of groups to be revealed.

SOM, (as any clustering method), provides an opportunity for deriving solutions for geological problems which can be related to different clustering problems, such as pattern recognition (ULTSCH, 1995). For implementation, Kohonen’s network has been used as it is possible to recognize spatial patterns being defined as spatial groups of points bounded by a polygon. It has potential for applying pattern recognition to problems such as identifying the spatial pattern of a sub-environment on a delta plain, which is resolvable by applying the SOM method based on a clustering algorithm.

Shapes shall be construed as clusters of data points that are bounded by a polygon which implies that pattern recognition is a clustering problem. However, this approach raises the question of the definition of patterns as statistical groups. Within a pattern, there are points with the same or similar properties. In practice, this means that all points in a sub-set represent a certain cluster which is bounded by polygons. A polygon represents the shape and reflects the borders of sub-environments, where the properties change between separate clusters. Figure 3 demonstrates an example for mapping clusters using the original coordinates. SOM creates \( c_m \) clusters (\( m \) is the number of clusters) based on the property values of data points without original spatial coordinates. After the clustering process we map the resulting clusters as patterns using the original coordinates. In this manner the different properties clusters are mapped and the pattern is visualized on the map. Thus after the clustering process, the resulting clusters are patterns in a map using the original coordinates.

2.2.2. Recognisable patterns in the map

The mapping processes compose the closed patterns by the arrangement of the points in space through the original \((x, y)\) coordinates. In this case it is necessary to give some conditions and definitions for recognizing patterns for mathematical reasons.

(1) A Closed pattern means that there is at least one inner point of the group.

(2) A point is an inner point if all of its neighboring points belong to the same cluster, i.e. there are at least four edge-neighbours to it.

(3) The border of a shape is defined by a point which has neighbours from another cluster. Of course a pattern might also have more neighbours.

(4) Within a shape we can also define another shape if it has inner points. Otherwise we should disregard these points as a pattern. However in this case, the question is how many inner points are enough to define a pattern. The answer depends on the grid dissolution. Since a 100 x 100 m grid size was applied over the study area, we can define the inner pattern which has at least one inner point. (This consideration gives 5 inner points in case of 50 x 50 m grid size and 25 inner points in case of 25 x 25 m grid size and so on.)

(5) The ‘inner shape’, should be covered by border points of only one other cluster.

3. STUDY AREA

Data originated from the Szöreg-1 reservoir (Algyő Field) that was used as a test area for the application of pattern recognition of palaeoenvironments in a clastic reservoir using the Kohonen’s neural network. The Algyő Field is the largest Hungarian hydrocarbon accumulation consisting of several oil and gas bearing reservoirs (Fig. 4). The upper mem-
bers of these reservoirs developed in a Pannonian delta system as the consequences of complex lateral delta-shifting and prograding phases. They can be subdivided into delta slope and delta plain rock bodies. The sills of the individual reservoirs were formed during delta abandonment phases. Below this series, the lower reservoirs are regarded as turbidity rock bodies of partly pro-delta fans and partly deep basin origin (RÉVÉSZ, 1982).

The Szőreg-1 oil reservoir, with a large gas cap is one of the largest rock bodies within the delta plain record (Fig. 4). Its average gross thickness is about 35 m, but locally it can be up to a maximum of 50 m thick. Earlier works by RÉVÉSZ (1982), GEIGER et al., (1998) and GEIGER (2005) have proved its delta plain origin with a significant amount of fluvial channel sedimentation. Figure 5 shows the original well density, the dataset is from 506 well on Szőreg-1 study area.

4. DATA PRE-PROCESSING AND ANALYSIS BY SOM

The network depends on the quality and quantity of training data, so data pre-processing is very important. This is as follows: 1) definition of the data type (discrete or constant), 2) calculation of statistics: mean, standard deviation, and 3) selection of outlier data. No 1 is critical because the outlier data reduces the efficiency of the network. In the case of category data, it is necessary to encode or scale the data. In this study we used continuous variable data.

It is imperative to set the initial values for the training rate and the neighborhood radius. The Kohonen-learning rate is altered linearly from the first to the last training cycle. We specified 0.3 for the start value and 0.02 for the end value. The neighborhood radius designates the adjacent area centered on the winning unit; in this case the size was 1 and specified a 3X3 square. Normal randomization of weights was used for the training; the mean and variance are specified, and are used to draw the initial weight values.

The second important step is to define the training, test and validation sets for training of SOM. The original database comprises 7500 points (grid points). This dataset is deconstructed into three sub-sets (training-, test- and validation set) using a random number generator to avoid bias. 60% of this database is used for building a neural network by SOM as a training set, 20% as a test, and 20% as a validation set.

Training set: this set is important to learn to fit the parameters of the clusters (in this case clusters are the neurons that characterized the clusters). SOM is applied to find the optimal weights to update the Kohonen’s layer and to position the neurons of clusters.

Validation set: this set is applied to tune the parameters of a classifier, and determine the stopping point of learning process.

Test set: this set lends itself to assess the performance of the trained clusters.

After pre-processing, the first step only sampled distributary mouth bars developed in Szőreg-1 to analyze how the SOM method can recognize the pattern for this part. The pattern of distributary mouth bars was given by sand contours that can reflect the geometry (Figs. 6a,b). It was an impor-
tant step since it can show how well the SOM-method reflects patterns jointly through three other variables. These variables were porosity, net pore volume and hydraulic conductivity data (Figs. 7a,b). The input data set comprises 2200 grid points (with 100X100 m dissolution), randomly divided into three (training, test and validation) subsets. The predefined number of clusters was tuned into 3, 4 and higher. In the case of more than 4 clusters, the algorithm became too confused for interpretation. Even, after mapping these results it did not constitute disjunctive clusters. In the first run, the minimum number of clusters is easier to tune in 3 because distributary mouth bars can be divided into three sub-environments as channel, outer bar, and bar crest.

After the clustering, we mapped the patterns. In the maps these patterns are the separated clusters of the analyzed area (distributary mouth bars) and this patterns mean the blocks of similar properties. Figure 8a,b represents the mapped result of the first run of SOM. The pattern within this depositional environment can affect the spatial arrangement of reservoir properties. Usually the sand contour used to reflect the spatial patterns but in the case of the clustering method and after that the mapping of clusters, similar patterns are discernible like the maps of three original properties (Figs. 7a,c) or map of sand contour (Fig. 6b). In this case the result-maps can demonstrate that it is also possible to apply clustering methods for outline the depositional geometry which are based on other three petrophysical parameters jointly.
In the second run, we have involved each of the available 7500 grid points of the ‘basic’ grid system (with 100X100 m resolution, derived from well 506, Fig. 5). In addition to recognizing the patterns again, the goal was also to identify the contours of the distributary bars. In this case, the number of clusters was tuned in 4 and 6, because we assumed at least one, but possibly two additional groups compared to the first run.

Figs. 9a,b represents the mapped result of the second run of SOM. It expresses again, that the boundary of this depositional environment can also affect the spatial arrangement of reservoir properties. However, if we compare the mapped clusters to the original contours of the distributary mouth bars, Figure 10 shows the similarities and differences. It can be seen that joint use of three variables in the SOM method produces a similar pattern within the distributary mouth bars as using one variable (i.e. sand content), but the contours of the shapes are different. So, the mapped results show the clusters which contain the points with similar properties in the present stage of diagenesis.

In both cases, the run time of the training algorithm depends primarily on the size of the network and only secondarily on the number of training steps. The presented mapped results (Figs. 8,9) are composed by setting 1000 learning steps in the SOM process. For this setting, (above mentioned numbers of clusters and number of learning steps) the run of SOM gives similar values for the error rates. Table 1 shows the rate of errors for each set.
Besides the error rates of sets, we can determine the goodness of clustering of SOM. Discriminant analysis can show how well the group a particular case belongs to can be predicted (see Table 2 a-d). Since the size of each cluster is not the same (the clusters contain different amounts of points), the priori probabilities \( p \) are not equal to 1/3 for each subset. The first column of the spreadsheet indicates the percentage of cases that are correctly classified in each cluster. Other columns show the number of misclassifications. According to discriminant analysis, the results of SOM are 95% adequate on average.

**Table 1: Error rates of running sets in case of running a-b-c-d.**

<table>
<thead>
<tr>
<th>errors rate of sets</th>
<th>case a)</th>
<th>case b)</th>
<th>case c)</th>
<th>case d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>training set</td>
<td>0.02786</td>
<td>0.028644</td>
<td>0.01515</td>
<td>0.01515</td>
</tr>
<tr>
<td>test set</td>
<td>0.020765</td>
<td>0.028264</td>
<td>0.031158</td>
<td>0.031158</td>
</tr>
<tr>
<td>validation set</td>
<td>0.019769</td>
<td>0.028178</td>
<td>0.0319757</td>
<td>0.031957</td>
</tr>
</tbody>
</table>

**Table 2 A Classification Matrix (distributary mouth bars – 4 clusters)**

<table>
<thead>
<tr>
<th>C_i</th>
<th>Percent Correct</th>
<th>C_i</th>
<th>p=0.39853</th>
<th>C_i</th>
<th>p=0.18582</th>
<th>C_i</th>
<th>p=0.24830</th>
<th>C_i</th>
<th>p=0.16789</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_1</td>
<td>88.02589</td>
<td>544</td>
<td>56</td>
<td>0</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_2</td>
<td>99.38651</td>
<td>0</td>
<td>1458</td>
<td>9</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_3</td>
<td>84.31085</td>
<td>0</td>
<td>107</td>
<td>575</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_4</td>
<td>92.88840</td>
<td>42</td>
<td>23</td>
<td>0</td>
<td>849</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>93.07253</td>
<td>586</td>
<td>1644</td>
<td>584</td>
<td>867</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2 B Classification Matrix (distributary mouth bars – 3 clusters)**

<table>
<thead>
<tr>
<th>C_i</th>
<th>Percent Correct</th>
<th>C_i</th>
<th>p=0.45314</th>
<th>C_i</th>
<th>p=0.31459</th>
<th>C_i</th>
<th>p=0.23327</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_1</td>
<td>98.32134</td>
<td>1640</td>
<td>28</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_2</td>
<td>96.80483</td>
<td>30</td>
<td>1121</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_3</td>
<td>96.49123</td>
<td>0</td>
<td>30</td>
<td>825</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>97.41918</td>
<td>1670</td>
<td>1179</td>
<td>832</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2 C Classification Matrix (entire dataset – 6 clusters)**

<table>
<thead>
<tr>
<th>C_i</th>
<th>Percent Correct</th>
<th>C_i</th>
<th>p=0.23560</th>
<th>C_i</th>
<th>p=0.14747</th>
<th>C_i</th>
<th>p=0.25400</th>
<th>C_i</th>
<th>p=0.13680</th>
<th>C_i</th>
<th>p=0.11200</th>
<th>C_i</th>
<th>p=0.11413</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_1</td>
<td>94.68025</td>
<td>1673</td>
<td>0</td>
<td>72</td>
<td>4</td>
<td>18</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_2</td>
<td>99.18626</td>
<td>0</td>
<td>1097</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_3</td>
<td>97.42782</td>
<td>43</td>
<td>0</td>
<td>1856</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_4</td>
<td>94.15205</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>966</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_5</td>
<td>88.33334</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>742</td>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_6</td>
<td>92.87383</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>795</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>95.05334</td>
<td>1776</td>
<td>1115</td>
<td>1928</td>
<td>1000</td>
<td>803</td>
<td>878</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2 D Classification Matrix (entire dataset – 4 clusters)**

<table>
<thead>
<tr>
<th>C_i</th>
<th>Percent Correct</th>
<th>C_i</th>
<th>p=0.35520</th>
<th>C_i</th>
<th>p=0.15800</th>
<th>C_i</th>
<th>p=0.32853</th>
<th>C_i</th>
<th>p=0.15827</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_1</td>
<td>94.68025</td>
<td>1673</td>
<td>0</td>
<td>72</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_2</td>
<td>99.18626</td>
<td>0</td>
<td>1097</td>
<td>0</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_3</td>
<td>97.42782</td>
<td>43</td>
<td>0</td>
<td>1856</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_4</td>
<td>94.15205</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>966</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>95.05334</td>
<td>1776</td>
<td>1115</td>
<td>1928</td>
<td>1000</td>
<td>803</td>
<td>878</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. CONCLUSIONS

Usually Kohonen’s network is applied to solve classical pattern recognition problems such as the recognition of a writing character or pictures. Frequently nonetheless, it can be applied to identifying a pattern as a classification problem, since in the background of pattern recognition there are clustering problems. This paper demonstrated the approach of using an unsupervised neural network to distinguish parts of the depositional sub-environment as clusters which contain points with similar properties in the present stage of diagenesis. Thus the SOM is an adequate tool to help identify genetic geological environments, sub-environments supplemented with the mapping process. The first test using the SOM-method reflected considerably well the patterns of the distributary mouth bars using three variables (Figs. 7 a,b). In the case of the second run of SOM, the mapped results cluster does not show exactly the same boundaries of the distributary mouth bars, but a similar pattern within the area can be recognized.

This paper focused on joint handling of more property to characterize parts of the sub-environment; and then mapping these as arranged points in space using some defined rules for geometry.

In fact, the neural network answers do not differ significantly from the original geological model obtained. However, differences are visible, especially at the contour of distributary mouth bars. This raises the question of whether the differences show greater accuracy or greater uncertainty. Nevertheless the discriminant analyses show that the clusters are well-defined at least on average for 95 % of the time.

REFERENCES


Manuscript received December 03, 2010
Revised manuscript accepted July 11, 2011
Available online October 26, 2011