A Topological Hierarchical Clustering : Application to Ocean Color Classification

Méziane Yacoub[1], Fouad Badran[2,1], and Sylvie Thiria[1]

[1] Laboratoire d'Océanographie Dynamique et de Climatologie (LODYC)
case 100, Université Paris 6, 4 place Jussieu 75252 Paris cedex 05 France
{yacoub,badran,thiria}@lodyc.jussieu.fr
[2] CEDRIC, Conservatoire National des Arts et Métiers,
292 rue Saint Martin 75003 Paris France
{badran}@cnam.fr

Abstract. We propose a new criteria to cluster the referent vectors of the self-organizing map. This criteria contains two terms which take into account two different errors simultaneously: the square error of the entire clustering and the topological structure given by the Self Organizing Map. A parameter T allows to control the corresponding influence of these two terms. The efficiency of this criteria is illustrated on the problem of top of the atmosphere spectra of satellite ocean color measurements.

1 Introduction

The aim of Self-Organizing Map (SOM) is to provide a "refined" partition of the data space using a huge number of neurons and to induce a topological order between them. The main goal of this partition is to reduce the information provided by the data using a vector quantization method. For practical application, one often looks for a limited number of significant clusters on the data space. Thus the problem is to reduce the number of clusters and to define a new partition of clusters from the initial SOM partition. This can be done by clustering the referent vectors of SOM using a hierarchical clustering algorithm [7, 8, 2].

In the present paper, we look for a new dissimilarity measure which allows us to take into account the two informations provided by SOM: the square error for the entire clustering and the existing topological order on the map. An adequate decomposition of the cost function which determines SOM suggests that some new criteria will be able to do it.

2 A New Hierarchical Clustering criteria

2.1 SOM quantization

The standard Self Organizing Map (SOM) [4] consists of a discrete set $\mathcal{C}$ of neurons called the map. This map has a discrete topology defined by an undirected

graph; usually it is a regular grid in one or two dimensions. We denote N_{neuron} the number of neurons in $\mathcal{C}$. For each pair of neurons (c,r) on the map, the distance $\delta(c,r)$ is defined as the shortest path between c and r on the graph. For each neuron c, this distance allows us to define a neighborhood of order d: $V_c(d) = \{r \in \mathcal{C} / \delta(c,r) \leq d\}$. In the following, in order to control the neighborhood order, we introduce a Kernel positive function K ($\lim_{|x| \to \infty} K(x) = 0$) and its associated family K_T parametrized by T :

$$K_T(\delta) = [1/T]K(\delta/T)$$

Let $\mathcal{D}$ be the data space ($\mathcal{D} \subset \mathcal{R}^n$) and $\mathcal{A} = \{\mathbf{z}_i; i = 1, \ldots, N\}$ the training data set ($\mathcal{A} \subset \mathcal{D}$). The standard SOM algorithm defines a mapping from $\mathcal{C}$ to $\mathcal{D}$ where each neuron c is associated to its referent vector $\mathbf{w}_c$ in $\mathcal{D}$. The set of parameters $\mathcal{W} = \{\mathbf{w}_c; c \in \mathcal{C}\}$, which fully determines the SOM, have to be estimated from $\mathcal{A}$. This is done iteratively by minimizing a cost function:

$$J^T_{som}(\chi, \mathcal{W}) = \sum_{c \in C} \sum_{\mathbf{z}_i \in \mathcal{A}} K_T(\delta(c, \chi^T(\mathbf{z}_i)) \|\mathbf{z}_i - \mathbf{w}_c\|^2 \qquad (1)$$

Where $\chi^T(\mathbf{z}_i)$ represents a particular neuron of $\mathcal{C}$ assigned to $\mathbf{z}_i$. This minimization can be done using a "batch" version of the standard SOM algorithm ([5], [6], [4], [1]). It can be expressed as a dynamic cluster method [3] operating in two steps:

- **The assignment step** assigns each observation $\mathbf{z}_i$ to one neuron c of $\mathcal{C}$ using the assignment function χ^T (relation 2). This step gives a partition of the data space $\mathcal{D}$ in N_{neuron} subsets, each observation $\mathbf{z}_i$ being assigned to its nearest neuron $\chi^T(\mathbf{z_i})$ according to a weighted sum of the euclidian distances:

$$\chi^T(\mathbf{z}) = \arg\min_{r \in C} \sum_{c \in C} K_T(\delta(c,r) \|\mathbf{z} - \mathbf{w}_c\|^2 \qquad (2)$$

- **The minimization step** minimizes the cost function (relation 1) with respect to the set of parameters $\mathcal{W}$ giving rise to the updated values of $\mathcal{W}$:

$$\mathbf{w}^T_c = \frac{\sum_{r \in C} (K(\delta(c,r)) \sum_{z_i \in P_r} z_i)}{\sum_{r \in C} K(\delta(c,r)) n_r} \qquad \text{where } P_r = \{\mathbf{z}_i \in \mathcal{A} / \chi^T(\mathbf{z}_i) = r\} \qquad (3)$$

For a given value of T, the batch algorithm minimizes (1) and leads to a local minima of this cost function with respect to both χ^T and $\mathcal{W}$. Using the batch version iteratively, with decreasing values of T, provides the standard SOM model. The nature of the SOM model reached at the end of the algorithm, the quality of the clustering (or quantization) and those of the topological order induced by the graph depend on the first value of T (T^{max}), its final value (T^{min}) and the number of iterations (N_{iter}) of the batch algorithm.

Formula (3) shows that SOM uses the neighborhood system, whose size is controlled by T, in order to introduce the topological order. When the value of T is large, an observation $\mathbf{z}_i$ will modify a large number of referent vectors $\mathbf{w}_c$, in opposite to small values of T allowing few changes. At the end of the learning algorithm (when T^{min} is reached), two neighbor neurons on the map have close referent vectors in the euclidian space ($\mathcal{R}^n$). In that sense, the map provides a topological order; the clustering associated to this topological order is defined in (2) by taking $T = T^{min}$. If T^{min} is such that the neighborhood of a neuron is reduced to itself for any distance d ($V_c(d) = \{c\}$) the cost function $J_{som}^{T^{min}}$ minimized at the end of the learning phase is the k-means distortion function. So, the successive iterations allow to reach a k-means solution which takes into account the topological constraint. In this case equation (3) shows that, for each neuron c, the referent vectors $\mathbf{w}_c$ is just the mean vector $\mathbf{g}_c$ of $P_c = \{\mathbf{z}_i \in \mathcal{A}/\chi^{T^{min}}(\mathbf{z}_i) = c\}$, in the following we denote by n_c the cardinality of P_c.

2.2 A topological hierarchical clustering

Rewriting J_{som}^T gives : $J_{som}^T = \sum_c \sum_r \sum_{\mathbf{z}_i \in P_r} K_T(\delta(c,r)) \|\mathbf{z}_i - \mathbf{w}_c\|^2$

$$= \left[\sum_c \sum_{r \neq c} \sum_{\mathbf{z}_i \in P_c} K_T(\delta(c,r)) \|\mathbf{z}_i - \mathbf{w}_r\|^2 \right] + \left[K_T(\delta(c,c)) \sum_c \sum_{\mathbf{z}_i \in P_c} \|\mathbf{z}_i - \mathbf{w}_c\|^2 \right] \quad (4)$$

Since usually at the end of the learning phase, $\mathbf{w}_c$ is no more that the mean vector of P_c (see section 2.1), we can decompose J_{som}^T using the square error of each individual cluster (or neuron): $I_c = \sum_{\mathbf{z}_i \in P_c} \|\mathbf{z}_i - \mathbf{w}_c\|^2$ ($I_c = 0$ for $P_c = \emptyset$), and (4) gives

$$\begin{aligned} J_{som}^T &= \tfrac{1}{2} \sum_c \sum_{r \neq c} K_T(\delta(c,r)) \left[\sum_{\mathbf{z}_i \in P_r} \|\mathbf{z}_i - \mathbf{w}_c\|^2 + \sum_{\mathbf{z}_i \in P_c} \|\mathbf{z}_i - \mathbf{w}_r\|^2 \right] \\ &\quad + K_T(\delta(c,c)) \sum_c \sum_{\mathbf{z}_i \in P_c} \|\mathbf{z}_i - \mathbf{w}_c\|^2 \\ &= \frac{1}{2} \sum_c \sum_{r \neq c} K_T(\delta(c,r)) \left[n_r * \|\mathbf{w}_r - \mathbf{w}_c\|^2 + I_r + n_c * \|\mathbf{w}_r - \mathbf{w}_c\|^2 + I_c \right] \\ &\quad + K_T(\delta(c,c)) \sum_c I_c = \end{aligned}$$

$$\left[\frac{1}{2} \sum_c \sum_{r \neq c} K_T(\delta(c,r)) (n_c + n_r) * \|\mathbf{w}_r - \mathbf{w}_c\|^2 \right] + \left[\sum_c \left(\sum_r K_T(\delta(c,r)) \right) I_c \right] \quad (5)$$

The first term of the decomposition of J_{som}^T takes into account the topological order, the second term corresponds to a weighted square error for the entire clustering and is similar to Ward criteria, which minimizes the intra-class inertia.

The hierarchical clustering, presented in this paper and denoted HC_{som}, proceeds by successive aggregations of neurons reducing by one, at each time, the cardinality of the preceding partition. At each iteration a new partition is defined. We denote by $\mathcal{P}_K$, such a partition made of K clusters, each cluster being denoted by an index c. The partition $\mathcal{P}_K = \{P_c / c \in \mathcal{C}_K\}$ is such that the set of index $\mathcal{C}_K$ has a graph structure which induce a discrete topology between the different clusters. For every c in $\mathcal{C}_K$, the cluster P_c is represented by its mean vector $\mathbf{g}_c$, its cardinality n_c and its square error I_c. We use J^T_{som} as a measure of the "quality" of the partition $\mathcal{P}_K$. Using $\mathcal{C}_K$, the dedicated measure becomes a sum of two terms:

$$J^T_{hc} = \left[\frac{1}{2}\sum_c \sum_{r \neq c} K_T(\delta(c,r))(n_c + n_r) * \|\mathbf{g}_r - \mathbf{g}_c\|^2\right] + \left[\sum_c \left(\sum_r K_T(\delta(c,r))\right) I_c\right] \quad (6)$$

Where c and r belong to $\mathcal{C}_K$ and $\delta(c,r)$ represents the distance on the graph $\mathcal{C}_K$ which will be defined below, as in (5) the first term of (6) (a) involves the topological order of the graph $\mathcal{C}_K$ and the second term (b) is similar to Ward criteria.

The initial partition $\mathcal{P}_{K_0}$ is given by the SOM map at the end of the learning algorithm. The graph $\mathcal{C}_{K_0}$ is the sub-graph of the map, where all the neurons such that $n_c = 0$ are removed. The initial distance $\delta(c,r)$ on $\mathcal{C}_{K_0}$ is defined as in section 2.1 by the length of the shortest path on the map. In general, the hierarchical clustering reduce $\mathcal{P}_K$ to $\mathcal{P}_{K-1}$ aggregating two vertices of $\mathcal{C}_K$ which allows us to determine the graph $\mathcal{C}_{K-1}$ of $\mathcal{P}_{K-1}$. If we denoted by $\{c_1, c_2\}$ the new index which aggregate c_1 and c_2 and $P_{\{c_1,c_2\}}$ its related cluster, $P_{\{c_1,c_2\}}$ is represented by its mean and its cardinality on the map :

$$\mathbf{g}_{\{c_1,c_2\}} = \frac{(n_{c_1} * g_{c_1}) + (n_{c_2} * g_{c_2})}{n_{c_1} + n_{c_2}},$$

$$n_{\{c_1,c_2\}} = n_{c_1} + n_{c_2}$$

and its individual square error

$$I_{\{c_1,c_2\}} = n_{c_1} * \|g_{c_1} - g_{\{c_1,c_2\}}\|^2 + n_{c_2} * \|g_{c_2} - g_{\{c_1,c_2\}}\|^2 + I_{c_1} + I_{c_2}.$$

The new distances δ on the graph $\mathcal{C}_{K-1}$ is defined by:

$$\delta(c, \{c_1, c_2\}) = min\{\delta(c, c_1), \delta(c, c_2)\}.$$

HC_{som} is looking for the best aggregation; as we compute the criteria J^T_{hc}, among all the possible pairs of $\mathcal{C}_K$ and the possible resulting partitions, we select the pair for which the value of J^T_{hc} is minimal. This pair gives rise to the new partition $\mathcal{P}_{K-1} = \{P_c / c \in \mathcal{C}_{K-1}\}$. Doing so, the parameter T defines a family of criteria whose characteristics are related to its value. Taking T small (as $T = T_{min}$), cancels the first term (a) of (6); in this case HC_{som} is the Ward criteria. Using a large value of T (as $T = T_{max}$), cancels the term (b); the method classify using only the topological order given by SOM. In this later case, HC_{som} becomes similar to the single link criteria. The intermediate values of T represent

a compromise between these two alternatives. The 'best' value of T has to be specified, as any hyper-parameter.

In the following, we show the behavior of HC_{som} when applied to a real application.

3 Classification of ocean color remote sensing measurements

Satellite ocean color sensors which are now operational or under preparation, measure ocean reflectance, allowing us to a quantitative assessment of geophysical parameters (e.g., chlorophyll concentration).

The major problem for ocean color remote sensing processing is that interactions of the solar light with aerosols in the atmosphere and water constituents are responsible for the observed spectrum of Top Of the Atmosphere (TOA) reflectances for a cloud-free pixel. The critical issue is thus to remove the aerosol contribution to the measured signals in order to retrieve the actual spectrum of marine reflectance which is directly related to the chlorophyl and water constituent content. Ocean color is determined by the interactions of the solar light with substances or particles present in the water, and provides useful indications on the composition of water.

In the following we used the SeaWiFS data products. The SeaWiFS[1] on board the SeaStar satellite is a color-sensitive optical sensor used to observe color variations of the ocean. It contains 8 spectral bands in the visible and near-infrared wavelengths [2]. SeaWiFS ocean color data is available in several different types. We used level 1 GAC data : it consists of raw radiances measured at the top of the atmosphere.

We studied the region shown in figure 1(a). It represents the Atlantic sea in the west north African coast. The image ($536 \star 199 = 106664$ pixels) was taken on January 1999. We removed the land pixels and some other erroneous pixels detected by the SeaWiFS product (the black region in figure 1(b)), and used our method to classify the remaining pixels (the white region in figure 1(b)).

First, we trained a two-dimensional map of size 10×10 with SOM algorithm. Then, we used HC_{som} to cluster the 100 referent vectors given by SOM and select the partition with 3 clusters. The experiment was repeated, varying the value of the parameter T. In figure 2, we show the three areas obtained using HC^T_{som} for $T = 0.0001$ (which correspond to the Ward criteria), $T = 0.1$, $T = 0.3$, $T = 0.5$, and $T = 2$. For technical reasons, we used three colors to show these partitions (white color, grey color, and black color). Thus one of the three partitions will have the same color than the removed region. Hereafter, by black region we means the black region without the removed pixels. First it is clear that the 5 different classifications proposed when using 5 different values of T correspond to different partitions which give rise to different possible interpretations. The

[1] SeaWiFS Project Home Page http://seawifs.gsfc.nasa.gov/SEAWIFS.html
[2] 412nm, 443nm, 490nm, 510nm, 555nm, 670nm, 765nm and 865nm

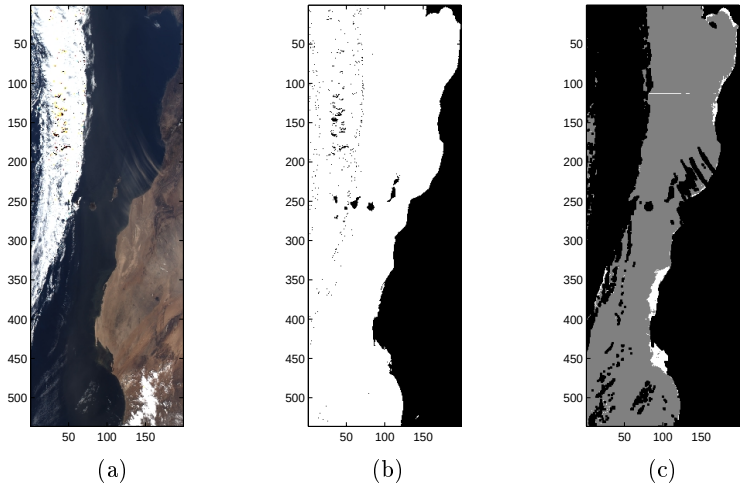

Fig. 1. (a) the studied region. (b) removed pixels according to the classification proposed by SeaWiFS (the black region) and the region to be classified using HC_{som} (the white region) (c) The three classes given by an expert on the studied region without removed pixels.

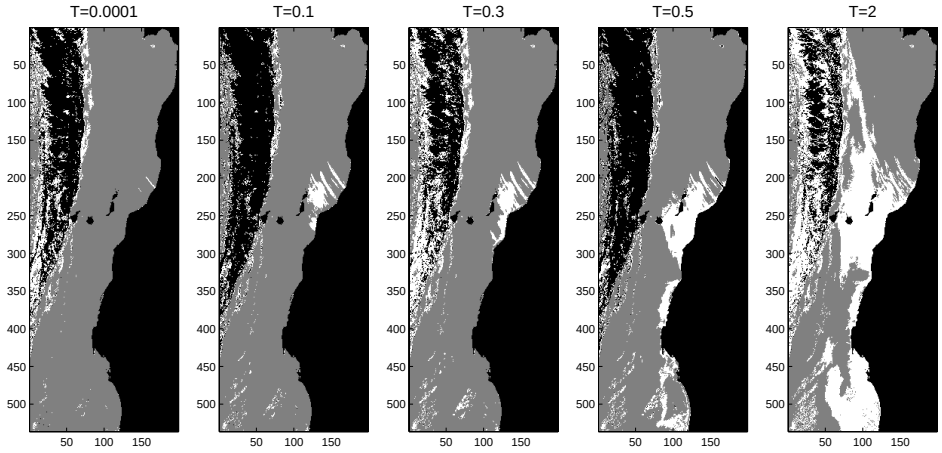

Fig. 2. The three classes obtained using HC_{som} for different value of T

expert gives a physical meaning to each classification. For $T = 0.0001$, the thick clouds became visible bringing together the black and white aereas. For $T = 0.1$ and $T = 0.3$ the black and white aereas represent thick clouds and desertic aerosols. For $T = 0.5$ the black and white areas represent thick clouds, desertic aerosols, and case 2 waters. The expert noticed that desertic aerosols have similar spectrum than case 2 waters. He provides a labeled map where he labeled pixel by pixel the image using physical models of aerosols. The labeled image is shown in figure 1(c). The three proposed classes are: the grey area representing the sea under a clear sky, the case 2 waters (white areas), and the cloud pixels (black areas). Clearly T=0.5 provides a classification similar to the one given by the expert. As case 2 waters and desertic aerosols have similar signature, HC_{som} put them into the same class. So the expert choose the case where $T = 0.5$ as being the most significant classification.

4 Conclusion

In this paper, we introduce a family of new criteria to perform hierarchical clustering. This family presents the new properties to mix two different criteria: the square error of the entire clustering and a graph approach which allows us to take into account the structure of the data set. This approach greatly takes advantage of the neural approach, the Self organizing Map provided an ordered codebook of the initial data and suggest a particular criteria in order to cluster this codebook. Some experiments on the problem of satellite ocean color classification shows that this hierarchical clustering can be useful for identifying different coherent regions.

References

1. Anouar F., Badran F. and Thiria S., (1997) : Self Organized Map, A Probabilistic Approach. proceedings of the Workshop on Self-Organized Maps. Helsinki University of Technology, Espoo, Finland, June 4–6.
2. Ambroise C., Seze G., Badran F. and Thiria S. (2000) : Hierarchical clustering of self-organizing maps for cloud classification. Neurocomputing, vol 30, number 1-4, January 2000. 47–52.
3. Diday E. and Simon J.C. (1976) : Clustering Analysis. In Digittal Pattern Recognition, Edited by K.S.FU. Springer-Verlag
4. Kohonen T. (1984) : Self organization and associative memory. Springer Series in Information Sciences, 8, Springer Verlag, Berlin (2^{nd} ed 1988).
5. Luttrel S.P. (1994) : A bayesian analysis of self-organizing maps. Neural comput. 6.
6. Ritter, Martinez and Schulten (1992) Neural computation and self organizing maps. Addison Wesley.
7. Thiria S., Lechevallier Y., Gascuel O. et Canu S., (1997) : Statistique et méthodes neuronales. Dunod
8. Yacoub M., D. Frayssinet, F. Badran and S. Thiria, (2000) : Clustering and Classification Based on Expert Knowledge Propagation Using a Probablistic Self-Organizing Map: Application to Geophysics. Data Analysis, edyted by W.GAUL, O.OPITZ and M.SCHADER, Springer.