A New Hierarchical Clustering Method using Topological Map

Méziane Yacoub[1], Ndèye Niang[2], 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
[2] CEDRIC, Conservatoire National des Arts et Métiers,
292 rue Saint Martin 75003 Paris France

Abstract. We present a new hierarchical clustering criteria which can be applied to data set. This is done after generating an initial partition by using a Topological Self Organizing Map. This criteria contains two terms which take into account two different errors simultaneously: the square error of the entire clustering (as the Ward criteria) and the topological structure given by the Self Organizing Map. A parameter T allows to control the corresponding influence of these two terms. Results on simulated data are presented which show the effect of this criteria for different values of T.

1 Introduction

The Self-Organizing Map (SOM) as introduced by Kohonen (Kohonen (1984)), has been widely used for quantization and classification. But SOM is also a clustering algorithm which generates some partition of the data. The partition has the particular property of being structured by a topological order represented by a graph. Many authors use SOM as a first stage of a Hierarchical Clustering algorithm (HC) (Thiria et al. (1997), Yacoub et al. (2000) Ambroise et al. (2000)). At each iteration, HC agglomerates two clusters by using a dissimilarity criteria. In this paper, we propose a new parametrized family of criteria for this agglomeration which takes into account both the euclidian distance on the data set and the topological order between the clusters. The main advantage of this family is to provide a flexible way to represent the underlying structure of the data. Depending on the value of the parameter, the criteria is able to identify globular or irregularly shaped clusters.

In the Section 2 we present the batch version of SOM algorithm. In Section 3 we present the new hierarchical clustering criteria. Section 4 is dedicated to preliminary results obtained on simulated data.

2 SOM quantization

The standard Self Organizing Map (SOM) (Kohonen (1984)) 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 ($\lim\limits_{|x| \to \infty} K(x) = 0$) 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:

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

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 (Luttrel (1994), Ritter (1992), Kohonen (1984), Anouar (1997)). It can be expressed as a dynamic cluster method (Diday and Simon (1976)) 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:

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

- **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 $\mathcal{W}$:

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

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 neighbors 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 as 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.

3 A topological hierarchical clustering

The aim of SOM is to provide a "refined" partition of the data space ($\mathcal{D}$) 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 ($\mathcal{D}$) 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 $\mathcal{P}_{K^*}$ of K^* clusters from the initial SOM partition. This can be done by clustering the N_{neuron} referent vectors of SOM using a hierarchical clustering algorithm.

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 J_{som}^T which determines SOM suggests that some new criteria will be able to do it. Rewriting J_{som}^T gives :

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

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

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), 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

$$J_{som}^T = \tfrac{1}{2} \sum_c \sum_{r \neq c} K_T(\delta(c,r)) \left[\sum_{\mathbf{z}_i \in P_r} \|\mathbf{z}_i - \mathbf{w}_c\|^2 + \sum_{\mathbf{z}_i \in P_c} \|\mathbf{z}_i - \mathbf{w}_r\|^2\right]$$
$$+ K_T(\delta(c,c)) \sum_c \sum_{\mathbf{z}_i \in P_c} \|\mathbf{z}_i - \mathbf{w}_c\|^2$$

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

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

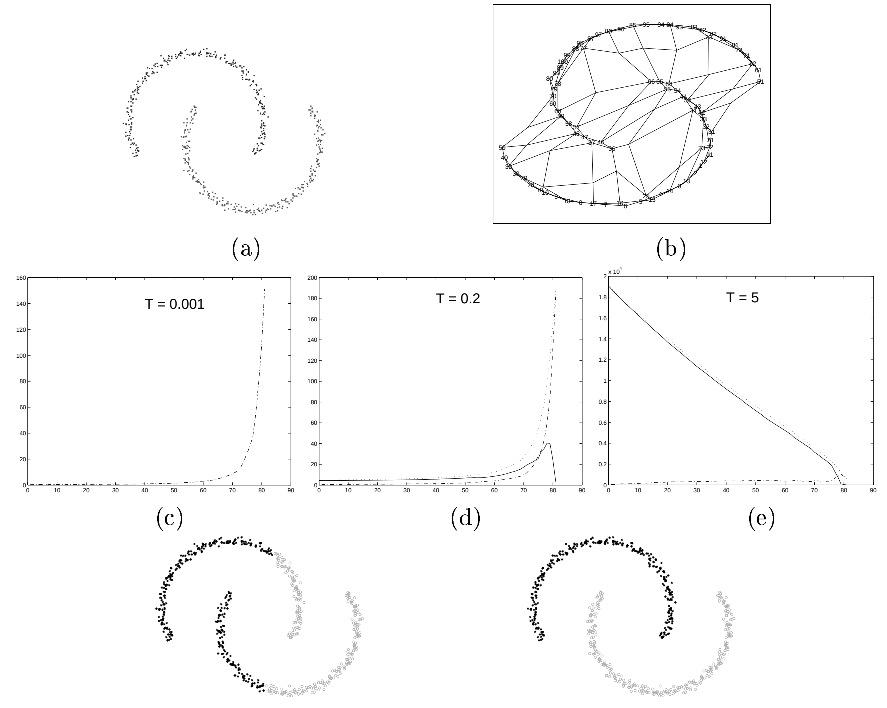
The first term of the decomposition of J^T_{som} takes into account the topological order, the second term corresponds to a weighted square error for the entire clustering and is similar to Ward criteria.

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:

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

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 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 inter-



(a) (b)

(c) (d) (e)

(f) (g)

Fig. 1. (a) : data set (b): the set of referent vectors with the induced topological order. (c, d, e): the evolution of J_{hc}^{T} (dotted ligne), and its two terms a (solid ligne) and b (dashdot ligne) during the aggregation of neurons, for three diffrent temperatures. (f, g) The two clusters obtained for $T = 0.001$ and $T = 0.2$.

mediate 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 use simulated data in order to show some examples of the different partitions we can obtain using HC_{som}.

4 Simulations

HC_{som} described in the previous section was applied on a simulated data set of $\mathcal{R}^2$ shown in figure 1(a). Clearly this data set presents two different clusters which are irregularily shaped. First, we trained a two-dimensional map of size 10×10 with SOM algorithm, the set of referent vectors and the graph induced by $(\mathcal{C})$ is presented in figure 1(b). We used HC_{som} to cluster the set $\mathcal{W}$ of referent vectors given by SOM and select the partition $\mathcal{P}_2$ with 2 clusters. The experiment was repeated, varying the value of the parameter T. For a given value of T, the behavior of the hierarchy can be seen, during the aggregation process, looking simultaneously at the evolution of J_{hc}^T and at those of its first and second terms (a) and (b). The results we obtain for $T = 0.001$, $T = 0.2$ and $T = 5$, are shown in figure 1(c,d,e). It can be seen that for $T = 0.001$, J_{hc}^T is approximatively equal to the second term (b). In this case $HC_{som}^{0.001}$ is similar to the Ward hierarchy. For $T = 5$, J_{hc}^T is approximatively equal to its first term (a), and the topology provided by SOM predominate. For $T = 0.2$, J_{hc}^T can be seen as a Ward criteria regularized by the topological order of the graph, the two terms of the sum have a particular impact. In figure 1 (f,g) we show the two clusters obtained using HC_{som}^T for $T = 0.001$, and for $T = 0.2$. Clearly the intermediate value $T = 0.2$ allows us to find the underlying structure of the data set; the use of the Ward criteria gives more spherical classes and is unable to extract the two circles.

5 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 simulated experiments proves that this hierarchical clustering can be useful for identifying irregularly shaped or nonglobular clusters.

References

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.

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.

DIDAY E. and SIMON J.C. (1976) : Clustering Analysis. In Digittal Pattern Recognition, Edited by K.S.FU. Springer-Verlag

KOHONEN T. (1984) : Self organization and associative memory. Springer Series in Information Sciences, 8, Springer Verlag, Berlin (2^{nd} ed 1988).

LUTTREL S.P. (1994) : A bayesian analysis of self-organizing maps. Neural comput. 6.

RITTER, MARTINEZ and SCHULTEN (1992) Neural computation and self organizing maps. Addison Wesley.

THIRIA S., LECHEVALLIER Y., GASCUEL O. et CANU S., (1997) : Statistique et méthodes neuronales. Dunod

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.