Evaluation of Topographic Clustering and its Kernelization

Marie-Jeanne Lesot, Florence d'Alché-Buc, and Georges Siolas

Laboratoire d'Informatique de Paris VI,
8, rue du capitaine Scott,
F-75 015 Paris, France
Marie-Jeanne.Lesot@lip6.fr

Abstract. We consider the topographic clustering task and focus on the problem of its evaluation, which enables to perform model selection: topographic clustering algorithms, from the original Self Organizing Map to its extension based on kernel (STMK), can be viewed in the unified framework of constrained clustering. Exploiting this point of view, we discuss existing quality measures and we propose a new criterion based on an F-measure, which combines a compacity with an organization criteria and extend it to their kernel-based version.

1 Introduction

Since their definition by Kohonen [1], Self Organizing Maps have been applied in various domains (see [2]), such as speech recognition, image analysis, robotics or organization of large databases. They solve a topographic clustering task, i.e. pursue a double objective: as any clustering method, they aim at determining significant subgroups within the whole dataset; simultaneously they aim at providing information about the data topology through an organized representation of the extracted clusters, such that their relative distance reflects the dissimilarity of the data they contain.

We consider the problem of the results evaluation, which is an important step in a learning process. Defining a quality measure of the obtained model enables to perform model comparison and thus model selection. Many criteria have been proposed but most of them fail to take into account the double objective of clustering and organization.

To address the evaluation question, we show that the various topographic clustering approaches can be viewed as solving a constrained clustering problem, the difference lying in the expression of the constraint which conveys the organization demand. We propose a new criterion which estimates the map's quality by combining, through an F-measure [3], an evaluation of its clustering quality with a measure of its organization quality; we apply it to classic and kernel-based maps to perform hyperparameter selection and data encoding comparison.

2 Constrained clustering

We divide the various formalizations proposed for topographic clustering and summarized in table 1, p. 5. in four categories and highlight the way they express the constraint.

2.1 Neural Networks

A first category of topographic algorithms is based on a neural network representation. Each neuron is associated with a position z_r and a weight vector w_r which represents a cluster center and has the same dimension as the input data. A topology is defined on this neurons set, through a $K \times K$ neighborhood matrix, where K is the number of nodes. It is defined as a decreasing function of the distance between positions, e.g.

$$h_{rs} = \exp\left(-\frac{\|z_r - z_s\|^2}{2\sigma_h^2}\right) \quad . \tag{1}$$

The *Self Organizing Maps* (SOM) algorithm introduced by Kohonen [1] is defined by the following iterative learning rule:

$$\begin{gathered} w_r(t+1) = w_r(t) + \alpha_t h_{rg(x_t)}(t)(x_t - w_r(t)) \\ \text{with } g(x_t) = \arg\min_s \|x_t - w_s\|^2 \quad , \end{gathered} \tag{2}$$

where x_t is a data point; the neighborhood term h_{rs} and the learning rate α_t decrease during the learning procedure; $g(x_t)$ denotes the winning node, i.e. x_t nearest neuron in terms of weight. At each step, the similarity between a data point and its winning node is increased; as a result, similar data are assigned to the same node, whose weight vector corresponds to an average representant of its associated data. Moreover, through the coefficient $h_{rg(x_t)}$, a data point modifies the weights of its node's neighbors: the organization constraint is expressed as an influence sphere around the winning node, which affects its neighbors. The parameter $\sigma_h(t)$ monitors the width of the influence areas and thus the distance at which the organization constraint is still to be felt.

Heskes [4] showed that the learning rule (2) cannot be derived from an energy function when the data follow a continuous distribution, and therefore lacks theoretical properties (see also [5]). Thus, he proposes to train the network by optimizing an energy function that leads to a slightly different map, which still fulfills the topographic clustering aims. In the case of a finite dataset $X = \{x_i, i = 1..N\}$, it is defined as

$$E = \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{K} h_{rg(x_i)}\|x_i - w_r\|^2 \text{ with } g(x_i) = \arg\min_s \sum_{t=1}^{K} h_{st}\|x_i - w_t\|^2 \quad . \tag{3}$$

Thus, the winning node is not only x_i nearest neighbor as in SOM, but takes into account the resemblance to neighbor nodes. If h is defined as in (1), $\forall r,\ h_{rr} = 1$, we propose to write $E = E_1 + E_2$ with

$$E_1 = \frac{1}{N}\sum_{i=1}^{N}\|x_i - w_{g(x_i)}\|^2 \quad \text{and} \quad E_2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{r\neq g(x_i)} h_{rg(x_i)}\|x_i - w_r\|^2, \tag{4}$$

and thus to interpret E in a constrained clustering context: E_1 equals the cost function optimized by the k-means algorithm. E_2 imposes the organization: when it is minimum, neighbor cells, corresponding to high $h_{rg(x_i)}$ values, have similar weight vectors. Indeed, $\|w_{g(x_i)} - w_r\|^2 \leq \|w_{g(x_i)} - x_i\|^2 + \|x_i - w_r\|^2$ where the first term is low because of E_1 minimization and the second because of the term $h_{rg(x_i)}\|x_i - w_r\|^2$ in E_2.

The parameter σ_h can be interpreted in a regularization framework: it monitors the relative importance of the main goal, clustering, and the imposed constraint, and thus the number of free parameters. When it is low, most h_{rs} terms are zero and the dominant term in E is E_1; when E_2 becomes prevalent (for high σ_h values), the map falls in a degenerate state where only the furthest nodes on the grid are non empty. Indeed, for such neurons, the weight is exclusively determined by the constraint and not by a tradeoff taking into account its assigned data, which enables to maximize organization.

Graepel, Burger and Obermayer [6] propose a deterministic annealing scheme to optimize this cost function, which leads to global and stable minima; beside the weights w_r, it provides assignment probabilities $p(x_i \in C_r) \in [0, 1]$, where $C_r = \{x_i / g(x_i) = r\}$. The associated algorithm is called *Soft Topographic Vector Quantization* (STVQ).

2.2 Markov Chains

Luttrell [7] considers topographic clustering as a noisy coding-transmission-decoding process, which he models by a specific Markov chain, called Folded Markov Chain (FMC): it consists in a chain of probabilistic transformations followed by the chain of the inverse (in a Bayes' sense) transformations.

He shows that SOM are a specific case of a two-level FMC. The first level corresponds to the coding step, which is equivalent to a clustering phase: it assigns data according to rule defined in equation (3) and codes them by the corresponding vector. The second level represents the transition probability to other clusters, it is fixed *a priori* and enables to express the constraint in the same way as a normalized neighborhood matrix (see [6]). The optimized cost function is defined as the reconstruction cost; it is equivalent to the function (3) if the neighborhood matrix is normalized.

2.3 Probability Distribution Modelling

Other formalizations aim at explicitely modelling the data probability distribution.

Utsugi [8] considers a bayesian framework to learn a gaussian mixture, constrained through a smoothing prior on the set $\mathcal{W} = \{w_r, 1 \leq r \leq K\}$. The prior is based on a discretized differential operator D:

$$p(\mathcal{W}/\alpha) = \prod_{j=1}^{d} C \exp\left(-\frac{\alpha}{2}\|Dw_{(j)}\|^2\right) \quad \text{with} \quad C = \left(\frac{\alpha}{2\pi}\right)^{l/2} (\det{}^{+} D^T D)^{\frac{1}{2}} \quad , \tag{5}$$

where $w_{(j)}$ is the vector of the jth components of the centers, $l = \text{rank}\,(D^T D)$, and $\det^{+} D^T D$ denotes the product of the positive eigenvalues of $D^T D$. Thus, a weights set is *a priori* all the more probable as its components have a low amplitude evolution, as expressed by the differential operator D. The centers w_r are learnt by maximizing the penalized data likelihood, computed as a gaussian mixture with this prior on centers; α monitors the importance of the constraint.

Bishop, Svensén and Williams [5] also consider a gaussian mixture, based on a latent variable representation: a data $x \in \mathcal{R}^d$ is generated by a latent variable $z \in \mathcal{L}$ of lower dimension $l < d$, through a function ψ of parameters $\mathcal{A}$: $x = \psi(z; \mathcal{A})$. Denoting

by β the variance of a gaussian noise process, and defining $p(z)$ as the sum of functions centered at nodes of a grid in $\mathcal{L}$, $p(z) = 1/K \sum_{r=1}^{K} \delta(z - z_r)$, $p(x)$ is defined as

$$p(x/\mathcal{A}, \beta) = \frac{1}{K} \sum_{r=1}^{K} \left(\frac{\beta}{2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{\beta}{2}\|\psi(z_r; \mathcal{A}) - x\|^2\right) \quad , \tag{6}$$

which corresponds to a constrained gaussian mixture: the centers $\psi(z_r; \mathcal{A})$ cannot evolve independently, as they are linked through the function ψ, whose parameters $\mathcal{A}$ are to be learnt. The continuity of $\psi(\cdot; \mathcal{A})$ imposes the organization constraint: two neighbor points z_A and z_B are associated with two neighbor images $\psi(z_A; \mathcal{A})$ and $\psi(z_B; \mathcal{A})$.

Heskes [9] shows that the energy function (3) can be interpreted as a regularized gaussian mixture: in a probabilistic context, it can be written as the data likelihood plus a penalization term, defined as a deviation of the learnt center w_r from the value imposed by organization $\tilde{w}_r = \sum_s h_{rs} w_s$. The solution must find a tradeoff between adaptating to the data and abiding by a low deviation, thus it solves a constrained clustering task.

2.4 Kernel Topographic Clustering

Graepel and Obermayer [10] propose an extension of topographic clustering, called Soft Topographic Mapping with Kernels (STMK), using the kernel trick: it is based on a non-linear transformation $\phi : \mathcal{R}^d \rightarrow \mathcal{F}$ to a high, or infinite, dimensional space, called the feature space; it must enable to highlight relevant correlations which may remain unnoticed in the input space. STMK transposes the cost function (3) in $\mathcal{F}$, by appplying STVQ to $\phi(x_i)$; the centers, denoted w_r^ϕ, then belong to $\mathcal{F}$. The cost function becomes

$$E^\phi = \frac{1}{N} \sum_{i=1}^{N} \sum_{r=1}^{K} h_{r g(x_i)} \|\phi(x_i) - w_r^\phi\|^2 \text{ with } g(x_i) = \arg\min_s \sum_{t=1}^{K} h_{st} \|\phi(x_i) - w_t^\phi\|^2.$$

Provided w_r^ϕ is searched as a linear combination of $\phi(x_i)$, as $w_r^\phi = \sum_i a_{ir} \phi(x_i)$, the computations are expressed solely in terms of dot products $< \phi(x_i), \phi(x_j) >$ [10]. Thus, defining a kernel function k such that $< \phi(x_i), \phi(x_j) >= k(x_i, x_j)$, it is possible to optimize E^ϕ without doing costly calculations in the high dimensional space $\mathcal{F}$. This algorithm being a direct transposition of STVQ to $\phi(x_i)$, it has the same interpretation in terms of constrained clustering, in the feature space.

3 Topographic Clustering Evaluation

Table 1 summarizes the previous algorithms. Whichever choice is made, the result map must be evaluated, to determine its validity and possibly to perform *a posteriori* model selection: in topographic clustering, it implies choosing the appropriate neighborhood parameter and the adequate size of the grid[1], plus the kernel parameter in the kernelized approach. According to the previous constrained clustering framework, maps must be assessed along two lines: their clustering capacity and their respect of the constraint, i.e. their organization. Yet most existing measures only take into account one aspect; using the notations of section 2.1, we discuss some of the existing criteria.

[1] assuming the dimension of the visualization space is 2.

Table 1. Summary of some caracteristics of topographic clustering algorithms (see section 2).

Designation	Principle	Learning algorithm	Probabilistic modelling	Constraint expression	Associated references
SOM	neural net	iterative rule	no	influence areas	[2]
STVQ STMK	neural net	deterministic annealing	possible	influence area	[6, 4, 9] [10]
FMC	probabilistic transformation	EM	yes	probabilistic influence	[7]
Utsugi	gaussian + prior	EM	yes	smooth weight differential	[8]
Bishop *et al.*	latent variable	EM	yes	continous gene--ration process	[5]

3.1 Clustering Quality

Kohonen [2] proposes to use the classic criterion called quantization error, which is the cost function of the k-means algorithm and is defined as the cost of representing a data x by the center of the cluster it is assigned to:

$$qC_1 = \frac{1}{N}\sum_{i=1}^{N}\|x_i - w_{g(x_i)}\|^2 = \frac{1}{N}\sum_{r=1}^{K}\sum_{i/x_i\in C_r}\|x_i - w_r\|^2 \quad . \tag{7}$$

For topographic clustering, contrary to clustering, the center of a cluster and its mean are distinct, as centers are influenced by their neighbors due to the organization constraint. Thus, computing the distance to centers introduces a bias in the homogeneity measure, and under-estimates the clustering quality. We propose to measure the cost obtained when representing a data by the mean of the cluster $\bar{x}_r$:

$$qM_1 = \frac{1}{N}\sum_{i=1}^{N}\|x_i - \bar{x}_{g(x_i)}\|^2 = \frac{1}{N}\sum_{r=1}^{K}\sum_{i/x_i\in C_r}\|x_i - \bar{x}_r\|^2 \quad . \tag{8}$$

Only the identified subgroups intervene in the measure, which makes it a justified clustering quality criterion.

Compacity can also be measured by the average variance of clusters [11] :

$$qM_2 = \frac{1}{K^*}\sum_{r=1}^{K}\frac{1}{|C_r|}\sum_{i/x_i\in C_r}\|x_i - \bar{x}_r\|^2 \quad K^* = \text{ number of non empty clusters.} \tag{9}$$

One can notice that qM_1 is also a weighted average variance, whose weighting coefficients are $|C_r|K^*/N$ i.e. the quotient between the cluster cardinal and an average cluster cardinal, under an equi-distribution assumption.

Some learning algorithms, like STVQ, provide assignment probabilities $p(x_i \in C_r)$ which are normalized so that $\forall i$, $\sum_r p(x_i \in C_r) = 1$ and equal the conditional probabilities $p(C_r/x_i)$. They lead to a probabilistic quantization error, qM_1^p computed by averaging the individual probabilistic errors, and a probabilistic variance mean qM_2^p:

$$qM_1^p = \frac{1}{N}\sum_{i=1}^{N}\gamma(x_i) \quad \text{with} \quad \gamma(x_i) = \sum_{r=1}^{K} p(C_r/x_i)\|x_i - \bar{x}_r\|^2 \quad , \tag{10}$$

$$qM_2^p = \frac{1}{K^*}\sum_{r=1}^{K}\sigma^2(C_r) \quad \text{with} \quad \sigma^2(C_r) = \sum_{i=1}^{N} p(x_i/C_r)\|x_i - \bar{x}_r\|^2 \quad , \tag{11}$$

$$\text{where} \quad \bar{x}_r = \frac{1}{\sum_j p(x_j \in C_r)}\sum_{i=1}^{N} p(x_i \in C_r)x_i \quad . \tag{12}$$

Likewise, one can define a probabilistic equivalent to qC_1. As previously, the differences between qM_1^p and qM_2^p come from normalization coefficients: considering equiprobable data, $p(x_i/C_r) = p(C_r/x_i)/\sum_{j=1}^{N} p(C_r/x_j)$.

3.2 Organization Quality

The organization criteria can be divided in three groups. The first measure was proposed by Cottrell and Fort [12] for one dimensional maps, as the number of inversions, i.e. the number of direction changes, which evaluates the line organization. It was generalized to higher dimensions by Zrehen and Blayo [13].

A second category is based on the data themselves and uses the winning nodes: if the map is well organized, then for each data the two best matching units must be adjacent on the grid. This principle has inspired measures such as the topographic error [14], Kaski and Lagus criterion [15], or the Hebbian measure [16].

Some organization measures are computed using only the neurons, without the data, which leads to an important computational saving and is more independent from the learning dataset. They evaluate the correlation between the distance in terms of weights and the distance imposed by the grid, that is $dW_{rs} = \|w_r - w_s\|^2$ and $dG_{rs} = \|z_r - z_s\|^2$. Indeed, the aim of organization is that the nearer the nodes the higher their similarity in terms of weight vectors. Bauer and Pawelzik [17] evaluate the conservation of ordering between nodes sorted by dW or dG. Flexer [18] evaluates the organization by a measure of correlation on the distance matrices. It only considers the map itself, without taking into account the data whose role is reduced to the training phase. Denoting for any KxK matrix A, $\Sigma A = \sum_{i,j} A_{ij}$, and $N_A = \left(\Sigma A^2 - (\Sigma A/K)^2\right)$, he uses the Pearson correlation:

$$\rho = \frac{\sum dGdW - \frac{\sum dG \sum dW}{K^2}}{\sqrt{N_G N_W}} \quad \in [-1, 1] \quad . \tag{13}$$

3.3 Combination

The previous measures do not consider the double objective of topographic clustering, but only one of its aspects; only two measures evaluate the compromise quality.

In the case of probabilistic formalizations, the result can be evaluated by the penalized likelihood of validation data. This measure evaluates the two objectives as it integrates the probability distribution on the weights, which expresses the constraint.

Independently of the learning algorithm, one can use as criterion a weighted quantization error $q_w = E$, where E is the function (3) whose decomposition (4) shows it considers both clustering and organization. Yet, it does not enable to select an optimal σ_h value if the h matrix is not normalized: when σ_h is small, most $h_{rg(x_i)}$ terms are low; it appears that when σ_h increases, the augmentation of the number of summing terms entails a more important increase than the decrease of cost due to a better organization. Thus, q_w augments, without its reflecting a real deterioration of the map's quality.

4 Proposed Criterion

To evaluate globally a topographic map, we propose a criterion combining a clustering quality measure with an organization measure, which we extend to kernel-based maps.

4.1 Classic Topographic Clustering

To measure the clustering quality, we choose a normalized expression $\tilde{q}_p = q/\eta$ of the criteria presented in section 3.1, $q = qC_1^p$, qM_1^p, or qM_2^p. The normalization aims at making the measure independent of the data norm scale. We propose to define:

$$\eta = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \bar{x}\|^2 \quad \text{with } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad . \tag{14}$$

If $q = qM_1^p$ or qC_1^p, η is interpreted as an *a priori* quantization error, obtained when all data are coded by the mean of the dataset. If $q = qM_2^p$, η is seen as the variance of the dataset before any subdivision. The criterion $\tilde{q}_p = q/\eta$ constitutes a clustering quality measure which varies in the interval $[0, 1]$ and must be minimized.

As organization measure, we choose a criterion derived from the Pearson correlation

$$c = \frac{1+\rho}{2} \in [0, 1] \quad . \tag{15}$$

$\tilde{q}_p$ only depends on the data, c on the weight vectors, which makes them independent. We combine them through the F-measure defined by Van Rijsbergen [3] and classically used in the Information Retrieval field to combine recall and precision. We apply it to c and $1 - \tilde{q}_p$ which are both to be maximized and define the global criterion $\mathcal{Q}_b$

$$\mathcal{Q}_b = \frac{(1+b^2)(1-\tilde{q}_p)c}{b^2(1-\tilde{q}_p) + c} \quad , \tag{16}$$

which must be maximized too. b is a weighting parameter controlling the relative importance of the two aims in the evaluation: if $b = 2$ for instance, $\mathcal{Q}_b$ rewards a high organization four times more than a good clustering. Thus this classic measure offers a mean to agregate in a single quantity the two criteria, and provides a numerical value, which always belong to the interval $[0, 1]$, to compare different maps; its advantage comes from the flexibility provided by the b agregation weight which allows the user to define numerically a tradeoff level between the two objectives.

4.2 Kernel-Based Topographic Clustering

The evaluation of the kernel-based topographic map requires us to compute the previous measure without computations in the feature space. $\tilde{q}_p^\phi$ imposes to evaluate $\|\phi(x_i) - \bar{x}_r^\phi\|$ and η^ϕ; both can be expressed solely with the kernel matrix $k_{ij} = k(x_i, x_j)$: denoting $p_{ir} = p(C_r/x_i)$ et $\alpha_{ir} = p_{ir} / \sum_j p_{jr}$, we have

$$\|\phi(x_i) - \bar{x}_r^\phi\|^2 = k_{ii} - 2\sum_{j=1}^{N} \alpha_{jr} k_{ij} + \sum_{j,l=1}^{N} \alpha_{jr}\alpha_{lr} k_{jl}$$
$$\eta^\phi = \frac{1}{N}\sum_{i=1}^{N} \|\phi(x_i) - \bar{x}^\phi\|^2 = \frac{1}{N}\left(\sum_{i=1}^{N} k_{ii} - \frac{1}{N}\sum_{i,j=1}^{N} k_{ij}\right) \quad .$$

Thanks to the normalisation η^ϕ, a small $\tilde{q}_p^\phi$ value is not due to the kernel itself, but indicates that the corresponding feature space defines a data encoding which highlights the presence of homogenous subgroups in the dataset.

The adaptation of c^ϕ requires to compute $dW_{rs} = \|w_r^\phi - w_s^\phi\|^2$. Using the decomposition $w_r = \sum_i a_{ir}\phi(x_i)$, we have $dW_{rs}^\phi = \sum_{i,l=1}^{N} k_{il}(a_{ir}a_{lr} - 2a_{ir}a_{ls} + a_{is}a_{ls})$.

The global quality of the map is then computed without too important additional costs as the F-measure between $1 - \tilde{q}_p^\phi$ and c^ϕ :

$$\mathcal{Q}_b^\phi = \frac{(1+b^2)(1-\tilde{q}^\phi)c^\phi}{b^2(1-\tilde{q}^\phi) + c^\phi} \quad . \tag{17}$$

5 Numerical Experiments

The numerical experiments highlight the relevance of the proposed criterion for map evaluation, and model selection including the algorithm hyperparameters (grid size K and neighborhood parameter σ_h), the kernel hyperparameters (type and parameter) and the data encoding. They are based on the STMK algorithm applied to a 2D square map. Indeed, STMK contains the classic maps as a special case, using the linear kernel $k(x, y) = (x \cdot y)/d$ which is equivalent to the scalar product in the input space.

5.1 Criterion Validation and Hyperparameter Selection

We study the behavior of the proposed criterion on an artificial 2D database, varying the hyperparameters. The base is generated by two distributions: a gaussian centered along a parabolic curve, and an isotropic gaussian (see fig. 1). As the data belong to $\mathcal{R}^2$, the resulting clusters can be visually displayed, by representing data belonging to a same node with a same symbol; the organization is represented by joining the means (computed in $\mathcal{R}^2$) of non empty clusters corresponding to adjacent nodes (in the kernel-based case, the centers w_r^ϕ belong to the feature space $\mathcal{F}$, and cannot be represented).

Figure 2 represents, for the linear kernel, the evolution of the clustering criteria qC_1^p, and qM_1^p (left), and qM_2^p (right), as functions of σ_h for various K values. All are monotonous functions of K and σ_h: the clustering quality is higher if the number of clusters is high and the organization constraint is low; thus they are not sufficient to

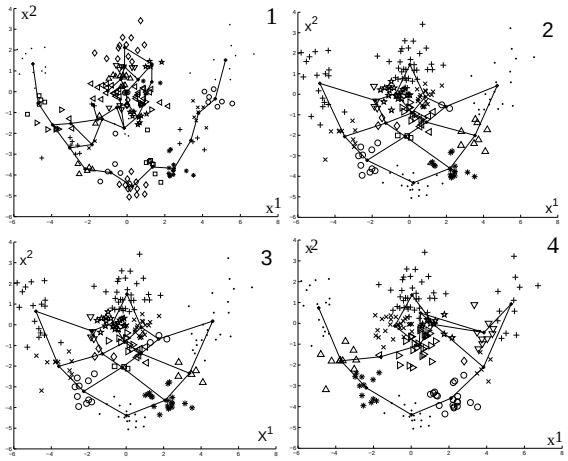

Fig. 1. Maps obtained with the optimal hyperparameter values. 1. Best linear map for $\mathcal{Q}_{0.5}$, $(K, \sigma_h) = (49, 0.30)$. 2. Best linear map for $\mathcal{Q}_2$, $(K, \sigma_h) = (16, 0.28)$. 3. Best 4x4 gaussian map for $\mathcal{Q}_2^\phi$, $(\sigma_h, \sigma_k) = (0.28, 1)$. 4. Best 4x4 polynomial map for $\mathcal{Q}_2^\phi$, $(\sigma_h, m_k) = (0.28, 2)$.

select optimal parameter values. One can notice that the difference between qC_1^p and qM_1^p remains low; yet, as qM_1^p evaluates the clustering quality using only the identified data subgroups, it appears as more satisfying on an interpretation level. Lastly the range of qM_2^p (right graph) is smaller than that of qC_1^p and qM_1^p: it is less discriminant and thus less useful to compare maps. In the following, we shall keep the qM_1^p measure.

Figure 3 represents the evolution of $\mathcal{Q}_{0.5}$ (left) and $\mathcal{Q}_2$ (middle), for a linear kernel. They are not monotonous and indicate optimal values, which depend on b. We chose to test $b = 0.5$ and $b = 2$ to favor respectively each objective: for $b = 0.5$, clustering is considered as the main task, thus the optimal K value is the highest tested value, $K = 49$; if $b = 2$, the organization demand is stronger, large grids which are difficult to organize obtain a lower score, and the optimum is $K = 16$.

The right graph of fig. 3 shows the "trajectory" of the map in the $(1 - \tilde{q}, c)$ plane, when σ_h varies, for different grid sizes, and highlights σ_h influence : small σ_h lead to a high quality clustering, but a poor organization. The parameter b, which expresses the relative importance of the two objectives in the evaluation phase enables the user to define the tradeoff level he desires: the points denoted A and B correspond to the optima associated with $\mathcal{Q}_{0.5}$ and $\mathcal{Q}_2$ and represent two different compromises. From our experiments, it seems that $b = 2$ is a good compromise in a visualisation framework.

Graphs 1 and 2 of fig. 1 show the maps obtained with the optimal (K, σ_h) values for $\mathcal{Q}_{0.5}$ and $\mathcal{Q}_2$ respectively. For $b = 0.5$, the clusters have low variances, but their organization is not satisfying: the chain of clusters associated with the parabolic data is too sensitive to data. For $b = 2$, it is more regular, and the map distinguishes between the two generative sources; it reflects the intern structure of data, in particular their symmetry. In the following, we conserve the value $b = 2$, which better reflects the visualisation goal, and consider 4x4 maps.

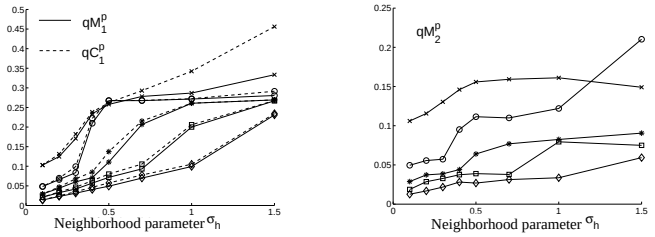

Fig. 2. Variation of the clustering criteria qC_1^p, et qM_1^p on the left, qM_2^p on the right, as functions of the neighborhood parameter σ_h for various grid sizes $K = \kappa^2$. Caption: x $\Leftrightarrow$ $\kappa = 3$, $\circ \Leftrightarrow \kappa = 4$,$* \Leftrightarrow \kappa = 5$, $\Box \Leftrightarrow \kappa = 6$, $\Diamond \Leftrightarrow \kappa = 7$.

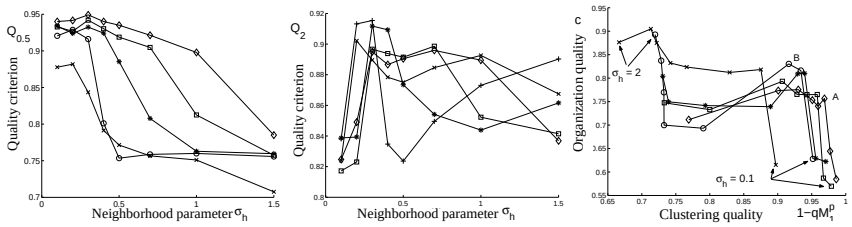

Fig. 3. Variations of $\mathcal{Q}_{0.5}$ on the left, $\mathcal{Q}_2$ in the middle, as functions of σ_h ; on the right, "trajectory" of the map in the $(1-\tilde{q}, c)$ plane when σ_h varies ; for different grid sizes $K = \kappa^2$. Caption : x $\Leftrightarrow$ $\kappa = 3$, $\circ \Leftrightarrow \kappa = 4$,$* \Leftrightarrow \kappa = 5$, $\Box \Leftrightarrow \kappa = 6$, $\Diamond \Leftrightarrow \kappa = 7$.

We tested STMK with the polynomial kernel k_p and the gaussian kernel k_g

$$k_p(x,y) = \left(\frac{x \cdot y}{d} + 1\right)^{m_k} \quad k_g(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma_k^2 d}\right) \quad . \tag{18}$$

Figure 4 represents $\mathcal{Q}_2^\phi$ as a function of σ_h for various values of σ_k (resp. m_k), for k_p and k_g. It shows that a large σ_h range leads to similar results. It also indicates that the gaussian kernel outperforms the linear one: $\mathcal{Q}_2^\phi$ optimal value is 0.941 in the gaussian case, and 0.917 in the linear case. The associated graphs (2 et 3, fig. 1) are very similar, the evaluation difference has a double reason: the slight assignment differences are in favor of the gaussian kernel; moreover, even identical clusters appear as more compact in the feature space than in the input space and lead to a better score. This is justified as the higher compacity leads to a faster convergence (5.3 times faster with these parameter values). According to $\mathcal{Q}_2^\phi$, the polynomial kernel gives poor results which is confirmed by the graphical representation (graph 4, fig. 1): the optimal polynomial map enables to distinguish the two sources but lacks organization.

This artificial base highlights the fact that the quality criterion based on the F-measure enables to select the hyperparameters values that indeed correspond to optimal maps, by both rewarding good maps and penalizing bad ones. It also enables to highlight the relevance of kernels in solving a topographic clustering problem.

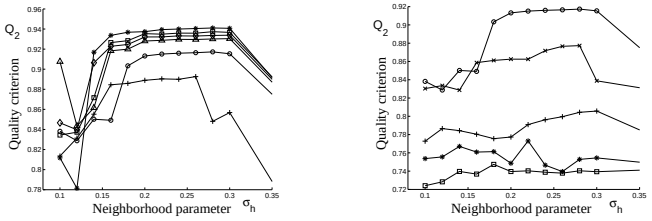

Fig. 4. Variations of $\mathcal{Q}_2^\phi$, for gaussian (left) and polynomial (right) kernels, as function of σ_h. $\circ$ corresponds to the linear map ; right, x to $\sigma_k = 0.1$, $+$ to $\sigma_k = 0.5$,$*$ to $\sigma_k = 1$, $\square$ to $\sigma_k = 1.5$, $\Diamond$ to $\sigma_k = 1.7$, $\triangle$ to $\sigma_k = 2$; left x to $m_k = 2$, $+$ to $m_k = 3$,$*$ to $m_k = 4$, $\square$ to $m_k = 5$.

Table 2. Best gaussian parameter combinations for various databases. The correspondance with the newsgroups is: 1 = alt.atheism, 2 = comp.graphics, 3 = rec.autos, 4 = rec.sport.hockey, 5 = sci.crypt, 6 = sci.electronics, 7 = soc.religion.christian, 8 = talk.politics.guns.

Dataset content	*tfidf* encoding (500 attributes)					*mppca* encoding (20 attributes)				
	σ_h	σ_k	$1-\tilde{q}^\phi$	c^ϕ	$\mathcal{Q}_2^\phi$	σ_h	σ_k	$1-\tilde{q}^\phi$	c^ϕ	$\mathcal{Q}_2^\phi$
$\mathcal{D}_1$: 2, 3, 5, 8	0.14	2	0.43	0.73	0.645	0.18	0.5	0.66	0.79	0.761
$\mathcal{D}_2$: 1, 2, 6, 8	0.14	1.5	0.36	0.72	0.601	0.22	1	0.69	0.78	0.762
$\mathcal{D}_3$: 3, 4, 6, 7	0.14	1.7	0.32	0.74	0.582	0.24	1.5	0.69	0.79	0.769

5.2 Data Encoding Comparison

We applied the proposed criterion to compare two document encodings: the *tfidf* method and a semantic based representation, called *mppca*, proposed by Siolas and d'Alché-Buc [19]. The latter exploits, through Fisher score extraction, a generative document model combined with a generative word model which captures semantic relationships between words thanks to a mixture of probabilistic PCAs. The dataset is built from the 20 newsgroup database[2] by selecting 100 texts of four different newsgroups. These 400 documents are encoded either by a 20 PCA mixture learnt on a 4x200-text set, or by the *tfidf* also learnt on this set, with a 500-word vocabulary. Table 2 presents the caracteristics of the best 7x7 gaussian maps. It shows the relevance of the semantic based model for the topographic clustering task: it leads to far better results, both globally and individually on clustering and organization. These tests on an unsupervised learning task confirm the results obtained in a supervised framework [19].

6 Conclusion

We have presented topographic clustering algorithms, from the original formulation by Kohonen of Self Organizing Maps to the Soft Topographic Mapping with Kernel extension which enables to use the kernel functions, in the same context of constrained

[2] `http://www.ai.mit.edu/people/jrennie/20Newsgroups/`

clustering, and considered the map evaluation problematic. We defined a new criterion which flexibly combines by an F-measure a clustering quality criterion with an organization criterion. The numerical experiments show it constitutes an efficient map comparison tool and enables to perform hyperparameter selection. Its main advantage lies in its flexibity which makes it possible for the user to explicitely define the tradeoff level between the two contradictory objectives of self organizing maps; thus it adapts itsef to the user's demands.

The next step of our work consists in applying bootstrap or other robust statistical method of estimmation to the proposed evaluation measure. The perspectives also include the application of the criterion to micro-array data where visualization is at the heart of the problematic and where such a criterion would enable to objectively select the best maps.

References

1. Kohonen, T.: Analysis of a simple self-organizing process. Biol. Cybern. **44** (1982) 135–140
2. Kohonen, T.: Self Organizing Maps. Springer (2001)
3. Van Rijsbergen, C.J.: Information Retrieval. Butterworth, London (1979)
4. Heskes, T.: Energy functions for self organizing maps. In Oya, S., Kaski, E., eds.: Kohonen Maps. Elsevier, Amsterdam (1999) 303–316
5. Bishop, C., Svensén, M., Williams, C.: GTM: The generative topographic mapping. Neural Computation **10** (1998) 215–234
6. Graepel, T., Burger, M., Obermayer, K.: Phase transitions in stochastic self-organizing maps. Physical Review E **56** (1997) 3876–3890
7. Luttrell, S.: A Bayesian analysis of self-organizing maps. Neural Computation **6** (1994) 767–794
8. Utsugi, A.: Hyperparameter selection for self organizing maps. Neural Computation **9** (1997) 623–635
9. Heskes, T.: Self-organizing maps, vector quantization, and mixture modeling. IEEE TNN **12** (2001) 1299–1305
10. Graepel, T., Obermayer, K.: Fuzzy topographic kernel clustering. In: Proc. of the 5th GI Workshop Fuzzy Neuro Systems, W. Brauer (1998) 90–97
11. Rezaee, M., Lelieveldt, B., Reiber, J.: A new cluster validity index for the fuzzy *c*-means. Pattern Recognition Letters **19** (1998) 237–246
12. Cottrell, M., Fort, J.: Etude d'un processus d'auto-organisation. Annales de l'Institut Poincaré **23** (1987) 1–20
13. Zrehen, S., Blayo, F.: A geometric organization measure for Kohonen's map. In: Proc. of Neuro-Nîmes. (1992) 603–610
14. Kiviluoto, K.: Topology preservation in Self Organizing Maps. In: Proc. of Int. Conf. on Neural Networks. Volume 1., IEEE Neural Networks Council (1996) 294–299
15. Kaski, S., Lagus, K.: Comparing self-organizing maps. In: Proc. of ICANN. (1996) 809–814
16. Polani, D., Gutenberg, J.: Organization measures for self-organizing maps. In: Proc. of the Workshop on Self-Organizing Maps, HUT (1997) 280–285
17. Bauer, H., Pawelzik, K.: Quantifying the neighborhood preservation of self-organizing feature maps. IEEE TNN **3** (1992) 570–579
18. Flexer, A.: On the use of self organizing maps for clustering and visualization. Intelligent Data Analysis **5** (2001) 373–384
19. Siolas, G., d'Alché Buc, F.: Mixtures of probabilistic PCAs and Fisher kernels for word and document modeling. In: Proc. of ICANN. (2002) 769–776