Évaluation des cartes auto-organisatrices et de leur variante à noyaux

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

Laboratoire d'Informatique de Paris 6
Université Pierre et Marie Curie
8, rue du capitaine Scott, 75 015 Paris
Marie-Jeanne.Lesot@lip6.fr

Résumé : Nous considérons la tâche de clustering topographique traitée par les cartes auto-organisatrices et le problème de leur évaluation : nous replaçons les variantes des algorithmes de clustering topographique, depuis les cartes de Kohonen jusqu'à leur extension basée sur les fonctions noyaux (STMK), dans un cadre commun du clustering contraint. Exploitant ce point de vue, nous discutons les mesures de qualité existantes et nous proposons un nouveau critère basé sur une F-mesure, qui combine une mesure de qualité de clustering avec un critère d'organisation et nous l'étendons aux cartes à noyaux.

Mots-clés : Cartes auto-organisatrices, Clustering topographique, Évaluation, Méthodes à noyaux

1 Introduction

Depuis leur définition par Kohonen en 1982, les cartes auto-organisatrices ont été appliquées dans de multiples domaines (voir (Kohonen, 2001)), tels que la reconnaissance de la parole, le traitement d'images, la robotique, le contrôle de processus ou l'organisation de grandes bases de données, textuelles (Kohonen *et al.*, 2000) ou génomiques (Tamayo *et al.*, 1999) par exemple. Elles poursuivent le double objectif du «clustering topographique» : comme tout algorithme de classification non supervisée, elles permettent de déterminer des sous-groupes pertinents dans l'ensemble des données ; simultanément, elles préservent des informations sur la topologie des données, par le biais d'une représentation organisée des *clusters* extraits, telle que la distance relative entre les groupes reflète la dissimilarité entre les données qui les constituent. De nombreux algorithmes d'apprentissage ont été proposés, basés sur des représentations par réseaux de neurones (Kohonen, 1982; Heskes, 1999) ou par chaînes de Markov (Luttrell, 1994), modélisant la densité de probabilité des données (Utsugi, 1997; Bishop *et al.*, 1998) ou exploitant une transformation non linéaire des données dans un espace de dimension élevée, par le biais des fonctions noyaux (Graepel & Obermayer, 1998).

Comme toute tâche d'apprentissage non supervisé, le clustering topographique pose le problème de l'évaluation des résultats. De nombreux critères ont été proposés, mais

pour la plupart ils ne prennent en compte qu'un seul aspect du double objectif des cartes auto-organisatrices. Pour aborder ce problème d'évaluation, nous montrons que toutes les variantes des cartes auto-organisatrices s'inscrivent dans le cadre commun du clustering contraint : elles effectuent une tâche de regroupement à laquelle est imposée une contrainte qui diffère suivant les algorithmes, mais qui répond à l'objectif de préservation de topologie et peut être vue comme une contrainte d'organisation. Nous proposons un critère qui évalue la qualité d'une carte en combinant par une F-mesure (Van Rijsbergen, 1979), une mesure de la qualité du clustering avec une évaluation de l'organisation et l'appliquons aux cartes classiques et aux cartes à noyaux.

La section 2 replace les algorithmes de clustering topographique dans le cadre du clustering contraint ; ce point de vue conduit, en section 3, à une classification des critères d'évaluation existants. Nous présentons en section 4 le critère proposé, et montrons en particulier comment il peut être calculé dans le cas des cartes auto-organisatrices à noyaux. La partie 5 présente les résultats des expériences numériques, qui montrent la pertinence du critère proposé pour les cartes classiques et les cartes à noyaux dont l'efficacité pour la tâche de clustering topographique est illustrée.

2 Clustering contraint

De nombreux algorithmes ont été proposés pour résoudre le problème du clustering topographique qui consiste à déterminer des sous-groupes pertinents de l'ensemble des données, en conservant une information sur la similarité des *clusters*. Ils peuvent être divisés en quatre catégories principales, et présentés dans un cadre commun de clustering contraint ; ils diffèrent par la formulation de la contrainte qui traduit l'exigence d'organisation. Leurs principales caractéristiques sont résumées dans le tableau 1.

2.1 Réseaux de neurones

Un premier groupe d'algorithmes est associé à un réseau de neurones constitué de K cellules n_r ; chaque neurone est associé à une position[1] z_r et à un vecteur de poids (ou vecteur de référence) w_r qui représente le centre du cluster associé et qui a même dimension (notée d) que les données. Une topologie est définie sur cet ensemble de neurones, par le biais d'une matrice de voisinage, de taille $K \times K$ définie comme une fonction décroissante de la distance entre les cellules, par exemple comme

$$h_{rs} = \begin{cases} 1 \text{ si } \|z_r - z_s\|^2 < \delta \\ 0 \text{ sinon} \end{cases} \qquad \text{ou} \quad h_{rs} = \exp\left(-\frac{\|z_r - z_s\|^2}{2\sigma_h^2}\right) \qquad (1)$$

Dans la suite, nous considérerons un voisinage gaussien, défini par la fonction de droite.

Cartes de Kohonen

L'algorithme SOM (*Self Organizing Maps*) proposé par Kohonen (1982) a été le premier algorithme de clustering topographique. Il est défini par la règle d'apprentissage

[1] Si les neurones sont situés sur une ligne ($z_r \in \mathcal{R}$) ou sur une grille 2D ($z_r \in \mathcal{R}^2$), ils permettent directement une visualisation graphique de la structure des données.

itérative suivante : la donnée x_t produit la mise à jour

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

où le terme de voisinage h_{rs} décroît au cours de l'apprentissage, de même que α_t qui définit le taux d'apprentissage, et où $g(x_t)$ désigne le neurone gagnant, c'est-à-dire le neurone auquel x_t est affectée, défini comme son plus proche voisin en terme de poids.

Cette règle permet, à chaque étape, d'accentuer la ressemblance entre une donnée et le vecteur de poids dont elle est la plus proche ; à la fin du processus d'apprentissage, des données similaires sont donc associées à un même neurone dont le vecteur de référence est un représentant moyen, ce qui correspond à l'objectif de clustering. De plus, à travers le coefficient $h_{rg(x_t)}$, une donnée modifie également les poids des voisins de son neurone gagnant, ce qui permet d'imposer la contrainte de similarité de neurones proches : la contrainte d'organisation est exprimée comme une zone d'influence autour du neurone gagnant, qui affecte et détermine partiellement les poids de ses voisins. Le paramètre $\sigma_h(t)$ (ou le paramètre $\delta(t)$), qui est décroissant, règle le rayon de cette zone et donc la distance à partir de laquelle la contrainte n'est plus sensible.

Définition d'une fonction d'énergie

Heskes (1999) montre qu'il n'existe pas de fonction d'énergie dont l'optimisation conduise à la règle d'apprentissage (2) si les données suivent une distribution continue, ce qui pose des problèmes théoriques, comme le manque de preuve de convergence du processus dans le cas général. Aussi, il introduit une fonction d'énergie, qui correspond à une règle d'apprentissage différente, mais dont le résultat respecte aussi les objectifs du clustering topographique. Dans le cas d'une base de données finie, $X = \{x_i, i = 1..N\}$, elle est définie comme

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

Elle utilise une nouvelle définition du neurone gagnant qui prend en compte la ressemblance aux poids des neurones voisins et fait donc intervenir la notion de voisinage.

Si h est le voisinage gaussien de (1), on a $\forall r, h_{rr}$=1, et l'on peut écrire $E = E_1 + E_2$,

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

ce qui permet d'interpréter E dans un contexte de clustering contraint : E_1 est égale à la fonction de coût de l'algorithme de clustering des k-moyennes, E_2 impose l'organisation. En effet, quand E_2 est minimal, des neurones voisins, correspondant à des valeurs de $h_{rg(x_i)}$ élevées, ont des poids proches : $\|w_{g(x_i)} - w_r\|^2 \leq \|w_{g(x_i)} - x_i\|^2 + \|x_i - w_r\|^2$, où le premier terme est faible en raison de la minimisation de E_1, le second à cause du terme $h_{rg(x_i)}\|x_i - w_r\|^2$ de E_2.

Le paramètre σ_h peut être interprété dans un cadre de régularisation car il contrôle l'importance relative entre le but principal (clustering) et la contrainte imposée, et donc le nombre de paramètres libres du système : quand il est faible, la plupart des termes h_{rs} sont nuls, E_1 est le terme dominant de E. Quand E_2 devient dominant (σ_h élevé), la carte peut se trouver dans un état dégénéré d'organisation extrême, où seuls les nœuds les plus éloignés sont non vides : pour de tels neurones, le poids est exclusivement déterminé par la contrainte et non par un compromis prenant en compte l'adaptation à des données qui lui seraient affectées, ce qui permet de maximiser l'organisation.

Graepel, Burger et Obermayer (1997) proposent une procédure d'optimisation de la fonction (3) basée sur le recuit déterministe, elle permet d'obtenir des minima globaux indépendants de l'initialisation ; en plus des vecteurs w_r, elle fournit des probabilités d'affectation $p(x_i \in C_r) \in [0,1]$, où $C_r = \{x_i / g(x_i) = r\}$. L'algorithme associé est appelé *Soft Topographic Vector Quantization* (STVQ).

2.2 Chaînes de Markov

Luttrell (1994) considère le clustering topographique comme un processus bruité de codage-transmission-décodage, qu'il modélise par une chaîne de Markov particulière, appelée *Folded Markov Chain* (FMC) : elle est constituée d'une chaîne de transformations probabilistes, suivie de la chaîne des transformations inverses, au sens de Bayes.

Luttrell montre que les SOM constituent un cas particulier de FMC à deux niveaux. Le premier correspond à la phase de codage, qui est équivalente à une étape de classification non supervisée : elle affecte (en probabilité) une donnée au vecteur qui lui est le plus similaire et la code par ce vecteur. Le second niveau exprime la probabilité de transition vers d'autres *clusters*, il est fixé *a priori* et n'est pas optimisé au cours de l'apprentissage ; il permet de traduire la contrainte de façon identique à la matrice de voisinage si elle est normalisée (voir (Graepel *et al.*, 1997)).

La fonction de coût optimisée est définie comme le coût de reconstruction ; il est équivalent à la fonction (3) si la matrice de voisinage est normalisée.

2.3 Modélisation de distribution de probabilité

D'autres formalisations des cartes auto-organisatrices ont pour but de modéliser explicitement la distribution de probabilité des données.

Utsugi (1997) se place dans un cadre bayésien pour l'apprentissage d'un mélange de gaussiennes, contraint par le biais d'un *a priori* sur les centres des gaussiennes :

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

où $\mathcal{W} = \{w_r, 1 \leq r \leq K\}$, $w_{(j)}$ est le vecteur des $j$ème composantes des centres, $l = \text{rang}\,(D^T D)$, et $\det^{+} D^T D$ est le produit des valeurs propres positives de $D^T D$, où D est un opérateur différentiel discrétisé qui permet d'exprimer la contrainte d'organisation : un ensemble de poids est d'autant plus probable que les composantes de ses vecteurs ont une évolution de faible amplitude, l'évolution étant traduite par D. Les

centres w_r sont appris en maximisant la vraisemblance pénalisée des données, calculée comme un mélange de gaussiennes avec cette loi *a priori* sur les centres ; α détermine l'importance de la contrainte.

Bishop, Svensén et Williams (1998) définissent le modèle GTM (*Generative Topographic Mapping*) qui considère aussi un mélange de gaussiennes mais utilise une représentation par variables latentes : une donnée $x \in \mathcal{R}^d$ est générée par une variable latente $z \in \mathcal{L}$ de dimension l, $l < d$, par le biais d'une fonction ψ de paramètres $\mathcal{A}$: $x = \psi(z; \mathcal{A})$. En notant β la variance d'un bruit gaussien affectant le processus de génération, et en définissant la probabilité de z comme la somme de fonctions centrées en des nœuds d'une grille de $\mathcal{L}$, $p(z) = 1/K \sum_{r=1}^{K} \delta(z - z_r)$, $p(x)$ est définie par

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

Cette distribution correspond à un mélange de gaussiennes contraint : dans l'espace d'entrée $\mathcal{R}^d$, les centres $w_r = \psi(z_r; \mathcal{A})$ ne peuvent évoluer indépendamment, ils sont liés par la fonction ψ, dont les paramètres $\mathcal{A}$ doivent être appris. Ils vérifient une propriété d'organisation grâce à la continuité de $\psi(\cdot; \mathcal{A})$ car deux points voisins z_A et z_B sont associés à deux points $\psi(z_A; \mathcal{A})$ et $\psi(z_B; \mathcal{A})$ proches.

Enfin, Heskes (2001) montre que la fonction d'énergie (3) peut être associée à un mélange de gaussiennes régularisé : dans un cadre probabiliste, elle s'écrit comme une vraisemblance à laquelle on ajoute un terme de pénalisation, défini comme la déviation des poids w_r par rapport aux valeurs imposées par l'organisation $\tilde{w}_r = \sum_s h_{rs} w_s$. Aussi le résultat de l'apprentissage est un compromis entre l'adaptation aux données, et l'obtention d'une faible déviation, et répond donc à un objectif de clustering contraint.

2.4 Clustering topographique avec noyau

Graepel et Obermayer (1998) proposent une extension du clustering topographique appelée STMK (*Soft Topographic Mapping with Kernel*), utilisant les fonctions noyaux, développées dans le cadre des machines à vecteurs supports par Vapnik (Cortes & Vapnik, 1995) : elle est basée sur une transformation non linéaire des données dans un espace de dimension élevée, voire infinie, appelé espace des caractéristiques, $\mathcal{F}$, qui peut éventuellement mettre en évidence des corrélations qui pourraient n'être pas remarquées dans l'espace initial.

L'algorithme STMK transpose la fonction de coût (3) dans l'espace $\mathcal{F}$, lié à l'espace des entrées $\mathcal{R}^d$ par une fonction non linéaire $\phi : \mathcal{R}^d \to \mathcal{F}$. Il consiste à appliquer STVQ aux $\phi(x_i)$; les centres, notés w_r^ϕ, appartiennent alors à $\mathcal{F}$. La fonction de coût devient :

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

Si les poids sont cherchés sous la forme de combinaisons linéaires des $\phi(x_i)$, $w_r^\phi = \sum_i a_{ir} \phi(x_i)$, les calculs s'expriment uniquement en fonction des produits scalaires $(\phi(x_i), \phi(x_j))$ (Graepel & Obermayer, 1998). Aussi en définissant une fonction noyau

Désignation	Principe	Type de l'apprentissage	Modélisation probabiliste	Expression de la contrainte
SOM	réseau neuronal	règle itérative	non	zone d'influence
STVQ STMK	réseau neuronal	recuit déterministe	possible	zone d'influence
FMC	transformations probabilistes	EM	oui	influence probabiliste
Utsugi	gaussienne + a priori	EM	oui	différentielle des poids
GTM	variables latentes	EM	oui	génération continue

TAB. 1 – Récapitulatif de quelques unes des caractéristiques des algorithmes d'apprentissage des cartes auto-organisatrices (voir le détail dans la partie 2).

k telle que $(\phi(x_i), \phi(x_j)) = k(x_i, x_j)$ l'optimisation de E^ϕ ne nécessite pas de calculs coûteux dans l'espace de grande dimension $\mathcal{F}$. Comme pour STVQ, Graepel et Obermayer proposent d'utiliser une procédure de recuit déterministe pour optimiser E^ϕ.

Cet algorithme est la transposition directe de STVQ aux $\phi(x_i)$, il a la même interprétation en terme de clustering contraint, transposée à l'espace des caractéristiques.

3 Évaluation

Le tableau 1 résume les caractéristiques des algorithmes précédents. Quel que soit l'algorithme utilisé, il est nécessaire d'évaluer le résultat, pour déterminer la validité de la carte obtenue et éventuellement effectuer une sélection *a posteriori* des hyperparamètres. D'après le cadre du clustering contraint par un objectif d'organisation introduit précédemment, les cartes doivent être évaluées suivant deux axes : la qualité des regroupements qu'elles proposent et la qualité d'organisation des *clusters*. Cependant, la plupart des mesures existantes ne prennent en compte qu'un seul de ces deux aspects, sans évaluer globalement la qualité du compromis. En utilisant les notations et le vocabulaire spécifiques à la représentation par réseau de neurones (partie 2.1), nous discutons ces critères qui, pour la plupart, sont applicables aussi aux autres formulations.

3.1 Qualité du clustering

Dans le cas du clustering, les mesures de qualité évaluent le compromis entre la ressemblance entre données affectées à un même groupe (homogénéité), et la différence entre données de *clusters* différents (séparabilité). Pour les cartes auto-organisatrices, la

séparabilité n'est pas un objectif principal : des groupes très similaires sont possibles, s'ils sont associés à des neurones proches au sens du voisinage ; le compromis s'effectue entre l'homogénéité et le respect de la topologie (organisation).

Cas d'affectations binaires

Kohonen (2001) propose d'utiliser la mesure classique d'homogénéité, appelée erreur de quantification, définie comme

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

Elle correspond à la fonction E_1 de la décomposition (4), i.e. à la fonction optimisée par l'algorithme des k-moyennes et est définie comme le coût du codage d'une donnée par le centre du *cluster* auquel elle est affectée, pris comme représentant du groupe.

Dans le cadre de la classification non supervisée, le centre d'un groupe est souvent confondu avec sa moyenne ; dans le cas des cartes auto-organisatrices, ils sont distincts, puisque les centres sont influencés par leurs voisins du fait de la contrainte d'organisation. Aussi, en calculant l'écart au centre, on introduit un biais dans la mesure d'homogénéité et on sous-estime la qualité du regroupement. Pour pallier cet effet, nous proposons de mesurer une erreur de quantification considérant que le codage associe à une donnée la moyenne du *cluster* auquel elle est affectée :

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

$$\text{où} \quad \bar{x}_r = \frac{1}{|C_r|}\sum_{i/x_i \in C_r} x_i \quad \text{avec} \quad |C_r| \text{ le cardinal du cluster } r$$

Cette mesure fait uniquement intervenir les données, c'est-à-dire les regroupements obtenus, ce qui en fait un critère justifié de la qualité du clustering.

D'autres algorithmes de classification non supervisée utilisent comme mesure d'homogénéité la variance moyenne des groupes (Rezaee *et al.*, 1998), c'est-à-dire

$$qM_2 = \frac{1}{K^*}\sum_{r=1}^{K} \frac{1}{|C_r|} \sum_{i/x_i \in C_r} \|x_i - \bar{x}_r\|^2 \quad (9)$$

en notant K^* le nombre de clusters non vides. Elle est proche de qM_1, qui constitue aussi une moyenne des variances des groupes, mais les pondère par $|C_r|K^*/N$, c'est-à-dire le quotient entre le cardinal du *cluster* et le cardinal moyen d'un groupe dans le cas où les données sont équi-réparties dans les *clusters*.

Cas de probabilités d'affectation

Certains algorithmes d'apprentissage, en particulier STVQ, fournissent des probabilité d'appartenance $p(x_i \in C_r)$ et non des affectations binaires. Elles sont normalisées de telle sorte que $\forall i$, $\sum_r p(x_i \in C_r) = 1$, et sont égales aux probabilités conditionnelles $p(C_r/x_i)$. Leur existence conduit à définir des équivalents probabilistes des

mesures précédentes : une erreur de quantification probabiliste qM_1^p, moyennant les erreurs probabilistes individuelles, et une variance moyenne des *clusters* probabiliste qM_2^p

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

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

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

On peut définir de la même façon un équivalent à qC_1. Les différences entre qM_1^p et qM_2^p se situent une nouvelle fois au niveau des normalisations, car, en considérant des données équiprobables,

$$p(x_i/C_r) = \frac{p(C_r/x_i)p(x_i)}{p(C_r)} = \frac{p(C_r/x_i)}{\sum_{j=1}^{N} p(C_r/x_j)}$$

3.2 Qualité de l'organisation

Même si les expressions de la contrainte varient suivant les algorithmes, elle peut être considérée comme une contrainte d'organisation dont on cherche le degré de satisfaction. On peut distinguer principalement trois types de mesures, essentiellement dédiées à la formulation neuronale des cartes auto-organisatrices.

La première mesure, définie comme une mesure d'inversion, a été proposée par Cottrell et Fort (1987) pour des cartes unidimensionnelles et des données unidimensionnelles, comme le nombre de changements de direction des vecteurs de poids. Elle a été généralisée à des cartes de dimension supérieure par Zrehen et Blayo (1992).

Un second type de mesures se base sur les neurones gagnants associés aux données : si la contrainte d'organisation est respectée, le neurone gagnant et le «second meilleur neurone» sont adjacents sur la carte, pour toute donnée. Ce principe a inspiré de nombreux critères (voir (Kiviluoto, 1996; Kaski & Lagus, 1996; Polani & Gutenberg, 1997)).

Enfin, certaines mesures sont calculées uniquement grâce aux neurones, sans utiliser les données, ce qui permet une économie de temps calcul. Elles évaluent la corrélation entre la distance en termes de poids et la distance imposée sur la carte, c'est-à-dire entre $dW_{rs} = \|w_r - w_s\|^2$ et $dG_{rs} = \|z_r - z_s\|^2$ En effet, l'organisation impose que deux neurones soient d'autant plus proches au sens de dG qu'ils ont des vecteurs proches, au sens de dW. Bauer et Pawelzik (1992) évaluent la conservation de l'ordre entre cellules ordonnées suivant dG ou dW. Flexer (2001) propose d'utiliser directement une mesure de corrélation sur les matrices de distance : notant pour toute matrice A de taille $K \times K$, $\Sigma A = \sum_{i,j} A_{ij}$, $N_A = \left(\Sigma A^2 - (\Sigma A/K)^2\right)$, il utilise la corrélation de Pearson

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

3.3 Combinaison

Les mesures précédentes ne prennent pas en compte le double objectif des cartes auto-organisatrices, mais seulement l'un de ses aspects : la qualité du clustering ou le respect de la contrainte d'organisation. Seules deux mesures existantes évaluent simultanément les deux aspects et la qualité du compromis atteint par la carte.

Dans le cas d'algorithmes qui modélisent la distribution de probabilité des données, le résultat de l'apprentissage peut être évalué par la vraisemblance pénalisée d'un ensemble de validation. Cette mesure spécifique évalue simultanément les deux objectifs car elle incorpore la distribution de probabilité sur les poids, qui traduit la contrainte.

Indépendamment de l'algorithme utilisé, on peut évaluer les cartes par l'erreur de quantification pondérée, $q_w = E$, où E est la fonction (3). La décomposition (4) montre qu'elle prend en compte à la fois le clustering(terme E_1) et la qualité de l'organisation (terme E_2). Elle ne permet toutefois pas de sélectionner une valeur optimale de σ_h si la matrice h est non normalisée : quand σ_h est faible, la plupart des termes $h_{rg(x_i)}$ sont petits ; lorsque σ_h augmente, l'augmentation du nombre de termes de la somme induit une élévation de l'erreur supérieure à la diminution due au gain en organisation. Aussi, q_w augmente, sans que cela reflète une réelle détérioration de la qualité de la carte.

4 Critère proposé

Pour évaluer globalement la qualité du clustering topographique, nous proposons un nouveau critère combinant une mesure de qualité du clustering $\tilde{q}$ avec une mesure d'organisation c, que nous étendons ensuite aux cartes auto-organisatrices à noyaux.

4.1 Clustering topographique classique

Pour mesurer la qualité du clustering, nous choisissons une version normalisée $\tilde{q} = q/\eta$ des critères présentés dans la partie 3.1, $q = qC_1^p$, qM_1^p, ou qM_2^p. La normalisation doit rendre la mesure indépendante de l'ordre de grandeur des normes des données. Dans tous les cas, nous proposons de définir

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

Si $q = qM_1^p$ ou qC_1^p, η est vu comme une erreur de quantification *a priori*, qui considère que toutes les données sont codées par la moyenne de l'ensemble des données. Si $q = qM_2^p$, η est vu comme la variance des données considérées avant subdivision.

Le critère $\tilde{q} = q/\eta$ ainsi défini permet de comparer des cartes correspondant à des données de variance totale très différente et représente la variance résiduelle non expliquée. Il constitue bien une mesure de la qualité de clustering, qui varie dans l'intervalle $[0, 1]$ et qui doit être minimisée.

Nous mesurons l'organisation par un critère linéairement lié à la corrélation de Pearson, mais qui varie dans $[0, 1]$, défini comme

$$c = \frac{1 + \rho}{2} \tag{15}$$

Nous combinons ensuite ces deux mesures pour évaluer globalement la qualité de la carte ; l'une ne dépend que des données et l'autre des vecteurs de poids, ce qui les rend indépendantes. Pour les combiner, nous proposons d'utiliser la F-mesure, définie par Van Rijsbergen (Van Rijsbergen, 1979) dans le contexte de la recherche d'information, pour combiner les mesures de rappel et de précision. Nous l'appliquons à $1-\tilde{q}$ et c qui doivent tous deux être maximisés, ce qui conduit au critère $\mathcal{Q}_b$

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

que l'on doit maximiser ; b est un paramètre qui détermine l'importance relative accordée à chacun des deux objectifs lors de l'évaluation ; ainsi, si $b>1$, $\mathcal{Q}_b$ récompense davantage une forte organisation qu'une bonne qualité de clustering. Il permet à l'utilisateur de définir le niveau de compromis qu'il souhaite obtenir et s'adapte aux propriétés qu'il désire, ce qui en fait un critère flexible.

4.2 Clustering topographique avec noyau

L'évaluation des cartes apprises avec des fonctions noyaux est faite en remplaçant x_i par $\phi(x_i)$ et w_r par w_r^ϕ dans les équations précédentes. Le problème des calculs impliquant des vecteurs de $\mathcal{F}$, qui sont coûteux car $\mathcal{F}$ est de dimension élevée voire infinie, peut être évité grâce à la matrice de noyau k, définie par $(\phi(x_i),\phi(x_j)) = k_{ij}$.

$\tilde{q}^\phi$ est défini en transposant qM_1^p et qM_1^p et η, ce qui nécessite l'évaluation de $\|\phi(x_i) - \bar{x}_r^\phi\|$ et η^ϕ : en notant $p_{ir} = p(C_r/x_i)$ et $\alpha_{ir} = p_{ir}/\sum_j p_{jr}$, on a

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

$$\eta^\phi = \frac{1}{N}\sum_{i=1}^{N} \|\phi(x_i) - \bar{x}^\phi\|^2 = \frac{1}{N}\left(\sum_{i=1}^{N} k_{ii} - \frac{1}{N}\sum_{i,j=1}^{N} k_{ij}\right)$$

Le calcul de c^ϕ nécessite de calculer les distances $dW_{rs} = \|w_r^\phi - w_s^\phi\|^2$. En utilisant la décomposition $w_r = \sum_i a_{ir}\phi(x_i)$, elles s'expriment par

$$dW_{rs}^\phi = \sum_{i,l=1}^{N} k_{il}(a_{ir}a_{lr} - 2a_{ir}a_{ls} + a_{is}a_{ls})$$

La qualité globale de la carte est alors calculée sans coût additionnels trop importants comme la F-mesure de $1-\tilde{q}^\phi$ et c^ϕ :

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

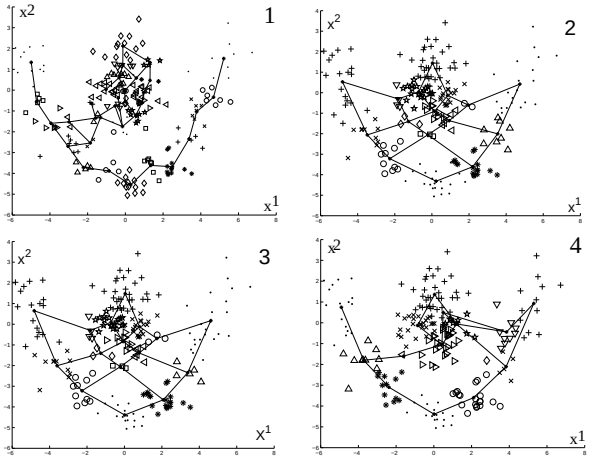

FIG. 1 – Cartes obtenues avec les valeurs optimales des hyperparamètres. 1. Meilleure carte linéaire pour $\mathcal{Q}_{0.5}$, $(K, \sigma_h) = (49, 0.30)$. 2. Meilleure carte linéaire pour $\mathcal{Q}_2$, $(K, \sigma_h) = (16, 0.28)$. 3. Meilleure carte gaussienne 4x4 pour $\mathcal{Q}_2^\phi$, $(\sigma_h, \sigma_k) = (0.28, 1)$. 4. Meilleure carte polynomiale 4x4 pour$\mathcal{Q}_2^\phi$, $(\sigma_h, \sigma_k) = (0.28, 2)$.

5 Expériences numériques

Les expériences réalisées illustrent la pertinence du critère proposé pour l'évaluation de la qualité des cartes, la sélection des valeurs des hyperparamètres et la comparaison de codages de données. Elles sont effectuées avec l'algorithme STMK, pour une carte 2D, comportant $K = \kappa^2$ neurones situés sur une grille carrée $\kappa \times \kappa$. STMK contient en effet les cartes classiques comme cas particulier, si l'on considère le noyau linéaire $k(x, y) = (x \cdot y)/d$, équivalent au produit scalaire dans l'espace des entrées.

5.1 Validation du critère et sélection d'hyperparamètres

Nous étudions le comportement du critère proposé sur une base artificielle 2D, en faisant varier les hyperparamètres, i.e. le nombre de cellules K, le paramètre de voisinage σ_h, et éventuellement le paramètre du noyau. La base est générée par deux distributions : une gaussienne centrée sur une parabole et une gaussienne isotrope (voir fig. 1). Les données appartenant à $\mathcal{R}^2$, les *clusters* peuvent être représentés graphiquement en associant aux points affectés à un même groupe un même symbole ; l'organisation est représentée en joignant les moyennes (calculées dans $\mathcal{R}^2$) des *clusters* non vides correspondant à des neurones adjacents sur la carte (dans le cas de cartes à noyaux, les centres w_r^ϕ appartiennent à l'espace de grande dimension et ne peuvent être représentés).

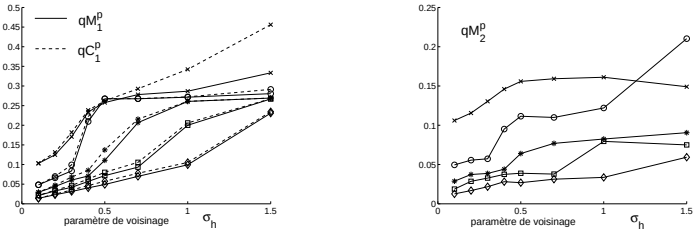

FIG. 2 – Variations des critères de clustering qC_1^p, et qM_1^p à gauche, qM_2^p à droite, en fonction du paramètre de voisinage σ_h pour différentes tailles de grille. Légende : x $\Leftrightarrow \kappa = 3$, $\circ \Leftrightarrow \kappa = 4$, $* \Leftrightarrow \kappa = 5$, $\square \Leftrightarrow \kappa = 6$, $\Diamond \Leftrightarrow \kappa = 7$.

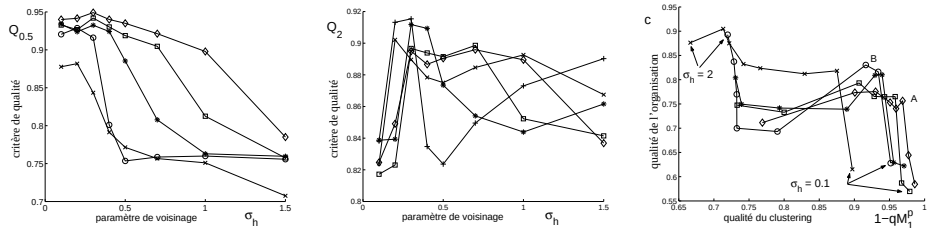

FIG. 3 – Variations de $\mathcal{Q}_{0.5}$ (gauche), $\mathcal{Q}_2$ (centre) en fonction de σ_h ; à droite, «trajectoire» de la carte dans le plan $(1-\tilde{q}, c)$ quand σ_h varie ; pour différents κ. Légende : x $\Leftrightarrow \kappa = 3$, $\circ$ (+ au centre) $\Leftrightarrow \kappa = 4$,$* \Leftrightarrow \kappa = 5$, $\square \Leftrightarrow \kappa = 6$, $\Diamond \Leftrightarrow \kappa = 7$.

Choix du critère de clustering

La figure 2 représente l'évolution des critères de clustering , qC_1^p, et qM_1^p à gauche, qM_2^p à droite, en fonction de σ_h pour différentes valeurs de K. Ces trois critères sont des fonctions monotones de K et σ_h : le clustering est de meilleure qualité si le nombre de clusters est élevé et si la contrainte d'organisation est faible (i.e. σ_h faible) ; ils ne sont donc pas suffisants pour sélectionner les valeurs optimales de ces hyperparamètres.

On constate que la différence entre qC_1^p et qM_1^p reste le plus souvent faible ; qM_1^p paraît toutefois plus satisfaisante sur le plan de l'interprétation, car elle évalue la qualité du regroupement en ne faisant intervenir que les groupes constitués. La plage de valeurs couverte par qM_2^p (graphique de droite) est moins étendue que pour qC_1^p et qM_1^p, ce qui rend cette mesure moins discriminante et donc moins utile pour la comparaison des cartes. Aussi, dans la suite des expériences, nous utiliserons la mesure qM_1^p.

Variation du nombre de cellules et du paramètre de voisinage

La figure 3 représente l'évolution des critères $\mathcal{Q}_{0.5}$ à gauche et $\mathcal{Q}_2$ au centre pour un noyau linéaire. On constate qu'ils ne sont pas monotones et indiquent des valeurs optimales pour les hyperparamètres K et σ_h, qui dépendent de la valeur de b choisie, en

particulier pour K. En effet, pour $b = 0.5$, la tâche considérée comme principale est le clustering, et la valeur optimale de K est égale au plus grand nombre de cellules testé, soit ici $K = 49$. Si $b = 2$, l'exigence d'organisation est prépondérante, ces grandes grilles difficiles à organiser obtiennent un score plus faible, et une carte plus petite ($K = 16$) est considérée comme optimale.

On constate que le critère $\mathcal{Q}_2$ qui accorde une importance plus grande à l'organisation que $\mathcal{Q}_{0.5}$ est plus sensible au paramètre qui contrôle cette contrainte, σ_h. Dans un premier temps, lorsque σ_h augmente, le gain en organisation est plus élevé que la détérioration du clustering, et $\mathcal{Q}_2$ croît, cette répartition s'inverse ensuite. La dernière phase de croissance, observée pour la grille 3×3, correspond au cas dégénéré où où seuls les quatre coins sont occupés : la qualité du clustering ne diminue plus, mais les autres cellules sont progressivement mieux organisées. Cette phase d'augmentation se produit pour les grilles de taille supérieure, pour des valeurs de σ_h plus élevées. Notons que si l'exigence d'organisation devient trop grande (valeurs de b supérieures), de telles cartes «vides» peuvent apparaître comme optimales, ce qui n'est pas souhaitable.

Le graphe de droite de la figure 3 montre la «trajectoire» de la carte dans le plan $(1 - \tilde{q}, c)$ quand σ_h varie, pour différentes valeurs de K. Il confirme que les grandes cartes associées à de petites valeurs de σ_h effectuent un clustering de bonne qualité mais sont mal organisées. Le paramètre b qui exprime l'importance relative des deux objectifs dans la phase d'évaluation permet à l'utilisateur de définir le niveau de compromis qu'il souhaite : les points notés A et B correspondent aux optima de $\mathcal{Q}_{0.5}$ et $\mathcal{Q}_2$, et représentent deux compromis différents.

Les graphes 1 et 2 de la fig. 1 montrent les cartes obtenues avec les valeurs optimales de (K, σ_h) pour $\mathcal{Q}_{0.5}$ et $\mathcal{Q}_2$ respectivement. Pour $K = 49$, les *clusters* sont de variance faible, mais l'organisation est peu satisfaisante : la chaîne des *clusters* associés aux données paraboliques est trop sensible aux données. La carte 4x4 conduit à une chaîne plus régulière, distingue les deux sources génératrices et reflète la structure interne des données, en particulier leur symétrie. Pour la suite, nous conservons une valeur de b supérieure à 1, pour favoriser l'organisation étant donné l'objectif de visualisation ; nous choisissons $b = 2$ et nous considèrons des cartes 4x4 sur une plage de σ_h donnant des cartes non «vides».

Variation de l'hyperparamètre du noyau

Nous avons testé les cartes auto-organisatrices à noyaux, en considérant les noyaux[2] polynomial k_p et gaussien k_g, définis par

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

La figure 4 représente $\mathcal{Q}_2^\phi$ en fonction de σ_h, pour différentes valeurs de σ_k ou m_k, pour les noyaux gaussien à gauche et polynomial à droite, ainsi que les résultats de la carte linéaire 4x4. Elle montre qu'une large plage de valeurs de σ_h conduit à des résultats comparables. Elle indique aussi que le noyau gaussien permet d'améliorer les performances par rapport aux cartes linéaires : la valeur optimale de $\mathcal{Q}_2^\phi$ vaut 0.941 dans le

[2]Le facteur d permet de comparer des valeurs de σ_k ou m_k pour des codages de données différents.

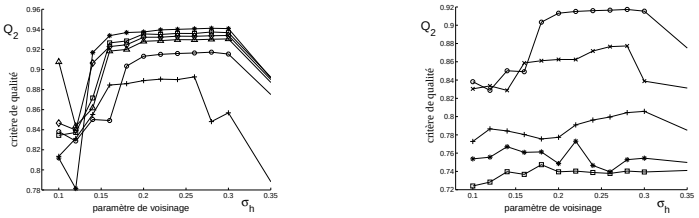

FIG. 4 – Variations de $\mathcal{Q}_2^\phi$, à gauche pour les noyaux gaussiens, à droite pour les noyaux polynomiaux, en fonction du paramètres de voisinage σ_h. Légende : $\circ$ correspond à la carte linéaire 4x4 ; à droite, x $\Leftrightarrow \sigma_k = 0.1$, $+ \Leftrightarrow \sigma_k = 0.5$,$* \Leftrightarrow \sigma_k = 1$, $\square \Leftrightarrow \sigma_k = 1.5$, $\Diamond \Leftrightarrow \sigma_k = 1.7$, $\triangle \Leftrightarrow \sigma_k = 2$; à gauche x $\Leftrightarrow m_k = 2$, $+ \Leftrightarrow m_k = 3$,$* \Leftrightarrow m_k = 4$, $\square \Leftrightarrow m_k = 5$.

cas gaussien, alors qu'elle est de 0.917 pour les cartes linéaires. Les graphes associés (2 et 3 sur la fig. 1) sont cependant très similaires, la différence d'évaluation a une cause double : les légères différences d'affectation apparaissent en faveur de la carte gaussienne ; de plus, des *clusters*, même similaires, sont plus compacts dans l'espace des caractéristiques que dans l'espace des entrées, et sont donc évalués comme de qualité supérieure. Ceci est justifié car cette meilleure compacité induit une convergence beaucoup plus rapide (5.3 fois plus rapide en moyenne pour ces valeurs de paramètres).

Les valeurs de $\mathcal{Q}_2^\phi$ indiquent que les noyaux polynomiaux donnent de mauvais résultats, ce qui est confirmée par la représentation graphique (graphe 4 de la fig. 1) : le noyau polynomial optimal conduit à une carte qui permet de distinguer les deux sources, mais qui manque d'organisation.

Cette base artificielle montre que le critère de qualité basé sur la F-mesure permet de sélectionner des valeurs d'hyperparamètres qui sont effectivement des valeurs optimales, à la fois en récompensant les bonnes cartes et en pénalisant les cartes de mauvaise qualité. Elle permet également de mettre en évidence l'intérêt des noyaux dans l'apprentissage des cartes auto-organisatrices,

5.2 Comparaison de codages de données

Nous avons appliqué le critère proposé à un problème de sélection de représentation, pour comparer deux codages de documents : la méthode *tfidf* et une représentation sémantique, appelée *mppca*, proposée par Siolas et d'Alché-Buc (2002). Celle-ci exploite, par l'extraction de scores de Fisher, un modèle génératif de documents combiné à un modèle génératif de mots, qui capture les relations sémantiques entre les mots grâce à un mélange d'ACP probabilistes.

La base de données considérée est construite à partir de la base des 20 *newsgroups*[3] en sélectionnant 100 textes de 4 groupes de *news* différents. Ces 400 documents sont codés soit par un mélange de 20 ACP appris sur un ensemble de 4x200 textes, soit par

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

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

TAB. 2 – Meilleures combinaisons d'hyperparamètres obtenues pour les deux codages de documents sur différentes bases avec le noyau gaussien. Les *newsgroups* utilisés sont indiqués par la correspondance suivante : 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.

la méthode *tfidf* apprise aussi sur cet ensemble, avec un vocabulaire de 500 mots. Le tableau 2 contient les caractéristiques des meilleures cartes obtenues, pour des cartes de taille 7x7 à noyaux gaussiens. Il montre la pertinence du modèle extrayant des concepts sémantiques, qui obtient de bien meilleurs résultats, aussi bien au niveau de la mesure globale que sur les plans de la qualité du clustering et de l'organisation. Ces tests sur une tâche d'apprentissage non supervisé confirment les résultats obtenus pour des tâches de classification supervisée (Siolas & d'Alché Buc, 2002) .

6 Conclusion

Nous avons présenté les algorithmes de clustering topographique, depuis la formulation originelle de Kohonen des *Self Organizing Maps* jusqu'à l'extension *Soft Topographic Mapping with Kernel* qui permet d'utiliser les fonctions noyaux, dans le même cadre de clustering contraint, et avons abordé la problématique de l'évaluation des cartes produites. Nous avons défini un nouveau critère d'évaluation qui combine de manière souple par une F-mesure un critère de qualité de la classification obtenue avec un critère de satisfaction de la contrainte d'organisation imposée. Les expériences numériques montrent qu'il constitue un outil efficace de comparaison numérique de carte et permet de sélectionner les combinaisons de valeurs optimales d'hyperparamètres. Son avantage principal provient de sa flexibilité, qui permet à l'utilisateur de définir explicitement le degré de compromis qu'il souhaite obtenir entre les deux objectifs contradictoires des cartes auto-organisatrices ; il s'adapte ainsi à ses exigences.

Les perspectives de ce travail incluent l'estimation du critère par des méthodes robustes comme le bootstrap, et son application à des données issues du traitement de *micro-array*, où la visualisation organisée des données se situe au cœur de cette problématique et où un tel critère numérique de comparaison de carte permettra de sélectionner objectivement les meilleures représentations.

Références

BAUER H. & PAWELZIK K. (1992). Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE TNN*, **3**(4), 570–579.

BISHOP C., SVENSÉN M. & WILLIAMS C. (1998). GTM: The generative topographic mapping. *Neural Computation*, **10**(1), 215–234.

CORTES C. & VAPNIK V. (1995). Support vector networks. *Machine learning*, **20**, 273–297.

COTTRELL M. & FORT J. (1987). Etude d'un processus d'auto-organisation. *Annales de l'Institut Poincaré*, **23**(1), 1–20.

FLEXER A. (2001). On the use of self organizing maps for clustering and visualization. *Intelligent Data Analysis*, **5**(5), 373–384.

GRAEPEL T., BURGER M. & OBERMAYER K. (1997). Phase transitions in stochastic self-organizing maps. *Physical Review E*, **56**(4), 3876–3890.

GRAEPEL T. & OBERMAYER K. (1998). Fuzzy topographic kernel clustering. In *Proc. of the 5th GI Workshop Fuzzy Neuro Systems*, p. 90–97: W. Brauer.

HESKES T. (1999). Energy functions for self organizing maps. In S. OYA & E. KASKI, Eds., *Kohonen Maps*, p. 303–316. Amsterdam: Elsevier.

HESKES T. (2001). Self-organizing maps, vector quantization, and mixture modeling. *IEEE TNN*, **12**, 1299–1305.

KASKI S. & LAGUS K. (1996). Comparing self-organizing maps. In *Proc. of ICANN*, p. 809–814: Springer.

KIVILUOTO K. (1996). Topology preservation in Self Organizing Maps. In *Proc. of Int. Conf. on Neural Networks*, volume 1, p. 294–299: IEEE Neural Networks Council.

KOHONEN T. (1982). Analysis of a simple self-organizing process. *Biol. Cybern.*, **44**(2), 135–140.

KOHONEN T. (2001). *Self Organizing Maps*. Springer.

KOHONEN T., KASKI S., LAGUS K., SALOJÄRVI J., HONKELA J., PAATERO V. & SAARELA A. (2000). Self organization of a massive document collection. *IEEE TNN*, **11**(3).

LUTTRELL S. (1994). A Bayesian analysis of self-organizing maps. *Neural Computation*, **6**(5), 767–794.

POLANI D. & GUTENBERG J. (1997). Organization measures for self-organizing maps. In *Proc. of the Workshop on Self-Organizing Maps*, p. 280–285: HUT.

REZAEE M., LELIEVELDT B. & REIBER J. (1998). A new cluster validity index for the fuzzy *c*-means. *Pattern Recognition Letters*, **19**, 237–246.

SIOLAS G. & D'ALCHÉ BUC F. (2002). Mixtures of probabilistic PCAs and Fisher kernels for word and document modeling. In *Proc. of ICANN*, p. 769–776: Springer.

TAMAYO P., SLONIM D., MESIROV J., ZHU Q., KITAREEWAN S., DMITROVSKY E., LANDER E. S. & GOLUB T. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS*, **96**, 2907–2912.

UTSUGI A. (1997). Hyperparameter selection for self organizing maps. *Neural Computation*, **9**, 623–635.

VAN RIJSBERGEN C. J. (1979). *Information Retrieval*. Butterworth, London.

ZREHEN S. & BLAYO F. (1992). A geometric organization measure for Kohonen's map. In *Proc. of Neuro-Nîmes*, p. 603–610.