Une approche basée sur STATIS pour la fusion de cartes topologiques auto-organisées

Mory Ouattara*,**, Ndèye Niang*, Rania Gasri ***, Fouad Badran*, Corinne Mandin**

*Statistique Appliquée, CNAM 292, rue Saint Martin, 75141 Paris Cedex 03, France,
n-deye.niang_keita@cnam.fr, fouad.badran@cnam.fr,
**Centre Scientifique et Technique du Bâtiment,
84 Avenue Jean Jaurès, 77420 Champs-sur-Marne
mory.ouattara@cstb.fr,corinne.mandin@cstb.fr,

Résumé. Dans le cadre des cartes topologiques, nous proposons une nouvelle approche d'ensemble clusters basée sur la méthode STATIS. Les méthodes d'ensemble clusters visent à améliorer la qualité de la partition d'un jeu de données à travers la combinaison de plusieurs partitions.
Les différentes partitions peuvent être obtenues en faisant varier les paramètres d'un algorithme (choix des centres initiaux, du voisinage initial et final des cellules dans le cas des cartes topologiques auto-organisée SOM, etc). L'approche présentée dans cette communication repose sur la méthode d'analyse de données multi-tableaux STATIS pour déterminer une matrice compromis représentant au mieux la similarité entre les partitions issues des cartes topologiques. La fusion des cartes topologiques est alors obtenue à travers une classification basée sur cette matrice compromis. La méthode proposée est illustrée sur des données réelles issues de l'UCI et sur des données simulées.

1 Introduction

En apprentissage non-supervisé, la plupart des méthodes de partitionnement souffrent d'une part d'un problème commun de stabilité des résultats par rapport aux paramètres d'initialisations des algorithmes. En effet, les partitions fournies par les algorithmes des K-moyennes ou des cartes topologiques auto-organisées (SOM), par exemple, dépendent du choix des centres de classes initiaux, du voisinage initial et final des cellules de la carte topologique,etc. D'autre part, en fonction de la méthode de classification utilisée, les partitions peuvent être différentes. Ainsi, en classification ascendante hiérarchique la partition obtenue dépend de la stratégie d'agrégation utilisée (critère de ward, lien moyen, lien complet, etc). Récemment, Strehl et Ghosh (2002); Fred et Jain (2003) ont proposé alors d'agréger les différentes partitions afin d'accroître significativement les performances de la partition finale. Ce concept connu sous le terme "d'ensemble clusters" reprend les concepts plus anciens de recherche de consensus de partitions proposés par Régnier (1983) et Gordon et Vichi (1998).
Dans cette communication, on intéresse aux méthodes d'ensemble clusters dédiées aux cartes topologiques. Comme les méthodes de partitionnement classiques (K moyennes, CAH), les

algorithmes d'apprentissage de type auto-organisés SOM (Kohonen, 1998) sont fortement dépendants des paramètres d'initialisation. Nous cherchons donc à améliorer le partitionnement des observations offert par l'algorithme SOM en adaptant les techniques de "cluster ensemble" aux cartes topologiques.
L'approche proposée repose sur la méthode d'analyse de données multi-tableaux STATIS (Lavit et al., 1994) pour déterminer une matrice compromis représentant au mieux la similarité entre les partitions issues des cartes topologiques. La fusion des cartes topologiques est alors obtenue à travers une classification basée sur cette matrice compromis.
La section 2 suivante présente le problème de fusion de SOM. La section 3.1 présente la méthode STATIS. La section 3.2 présente la méthode proposée et ses applications aux données réelles issues de l'UCI et sur des données simulées.

2 Fusion de SOM

La démarche des méthodes d'"ensemble clusters" se résume en deux étapes : une étape de diversification par la création d'un ensemble de partitions et une étape d'agrégation des partitions. Dans le cadre de la fusion de SOM nous désignons par $\mathbb{C}$ l'ensemble des cartes topologiques. Cet ensemble peut être obtenu de diverses manières. Il peut s'agir : de résultats obtenus par application répétée d'un même algorithme avec différentes initialisations des paramètres (Jiang et Zhou, 2004; Georgakis et al., 2005). Dans le domaine des réseaux de neurones, le problème de recherche de consensus consiste d'une part à apprendre indépendamment B cartes SOM, d'autre part à synthétiser les résultats de ces SOM en regroupant les neurones similaires des différentes cartes. Soit $\mathbb{C} = \{\mathcal{C}^1, \ldots, \mathcal{C}^b, \ldots, \mathcal{C}^B\}$ l'ensemble des cartes topologiques, $\mathcal{C}^b = (\pi^b, W^b)$ avec π^b et $W^b = \{w_1^b, \ldots, w_{n^b}^b\}$ définissent respectivement une partition des observations et l'ensemble des vecteurs référents w_c^b associés aux cellules de la carte $\mathcal{C}^b$. Chaque carte $\mathcal{C}^b$ contient n^b cellules. Soit $\mathcal{C}^* = (\pi^*, W^*)$ la carte représentant la fusion des B cartes $\mathcal{C}^b$. Le problème revient alors à définir les quantités π^* et $W^* = \{w_1^*, \ldots, w_{nf}^*\}$ représentant respectivement la partition finale des observations et les référents résultant de la fusion des référents w_c^b associés aux cellules.
Dans la littérature, différentes approches de fusion de SOM ont été proposées. Georgakis et al. (2005) effectuent la fusion des cartes à travers un processus itératif en regroupant leurs cellules les plus proches au sens d'une distance euclidienne définie sur les vecteurs référents de ces cartes. Saavedra et al. (2007) fusionnent deux cellules en se basant sur la proportion d'individus qu'elles ont simultanément captés.

La méthode proposée dans cette communication est basée sur la classification d'une matrice de compromis $\tilde{Co}$ représentant au mieux la relation entre les partitions et qui s'obtient en utilisant la méthode STATIS que nous présentons dans la section suivante.

3 Consensus basé sur STATIS

3.1 STATIS

STATIS est une méthode exploratoire d'analyse de données multi-vues (plusieurs tableaux décrivant les mêmes individus à l'aide de variables pouvant être différentes). On dispose de H^b

$(b = 1, \ldots, B)$ matrices contenant N observations et k variables. Préalablement à l'analyse, les données sont centrées réduites. STATIS associe à chaque matrice H^b, la matrice $X^b = H^b H^{b'}$ $(N \times N)$, où $H^{b'}$ est la matrice transposée de H^b. C'est un objet représentatif de H^b contenant tous les liens inter-individus. La méthode STATIS utilise ensuite le produit scalaire de Hilbert-Schmidt pour induire une distance entre les matrices X^b et les comparer. Ce produit scalaire, pour deux matrices X^a et X^b est représenté par :

$$HS(X^a, X^b) = trace(DX^a DX^b)$$

où D est la matrice des poids associés aux individus, en général ils sont choisis uniformément égaux à 1/N. On définit par S la matrice des produits scalaires entre les tableaux et dont les entrées $S(a, b)$ valent $HS(X^a, X^b)$. Il est habituel de normaliser les matrices X^b. Les entrées de la matrice S deviennent alors des coefficients de corrélation vectorielle R_V entre les objets X^b tel que :

$$R_V(X^a, X^b) = \frac{trace(DX^a DX^b)}{\sqrt{trace(DX^a)^2 trace(DX^b)^2}} \tag{1}$$

Comme en Analyse en Composantes Principales, la diagonalisation de la matrice S fournit une représentation euclidienne des tables X^b dans un espace de dimension réduit permettant de visualiser les différences et les ressemblances entre les matrices X^b. Ce qui constitue l'étude de l'inter-structure dans STATIS. Une étude plus fine au niveau des individus permettant de comprendre la relation entre les tables est réalisée à travers l'étude de intra-structure.

L'intra-structure repose sur la détermination d'une matrice compromis $\tilde{Co}$ de même nature que les matrices H^b telle que $\tilde{Co}$ soit le plus corrélé possible au sens du produit scalaire HS avec les matrices X^b. La recherche du compromis est formalisée comme un problème de recherche de la meilleure combinaison linéaire des matrices X^b, celle qui maximise la corrélation vectorielle avec les matrices X^b :

$$\tilde{Co} = \max_{Co} \sum_{b=1}^{B} R_V(Co, X^b) \tag{2}$$

Comme en ACP, la solution est le premier facteur principal μ défini comme le vecteur propre associé à la plus grande valeur propre de la matrice des R_V :

$$\tilde{Co} = \sum_{b=1}^{B} \mu^b X^b$$

3.2 Approche de Fusion de SOM basée sur STATIS

On s'intéresse au consensus des partitions de l'ensemble $\Pi = \{\pi^1, \ldots, \pi^B\}$ issues de SOM. Notre approche de recherche de consensus est basée sur la matrice compromis $\tilde{Co}$ fournie par STATIS. Les matrices X^b sont les matrices d'adjacence associées aux partitions π^b avec :

$$X^b(i, j) = \begin{cases} 1 & si \quad \pi^b(i) = \pi^b(j) \\ 0 & sinon \end{cases}$$

Pour déterminer le consensus on utilise usuellement la matrice $Co = \frac{1}{B}\sum_{b=1}^{B} X^b$ dont les entrées sont égales au nombre de fois où deux individus ont été regroupés ensemble dans une classe pour déterminer le consensus. Or, les partitions formant l'ensemble Π n'ont généralement pas la même pertinence. Notre approche va alors définir à travers STATIS un système de poids sur les partitions π^b selon le principe suivant : si deux partitions π^a et π^b sont similaires, les poids μ^a et μ^b de ces partitions sont identiques. Par ailleurs, deux partitions différentes auront des poids différents. Par conséquent, une partition totalement différente des autres aura un poids μ faible. Les poids sont donc liés au niveau de ressemblance entre les partitions. La matrice $\tilde{Co}$ définie dans l'équation (2) peut alors être vue comme la matrice consensus des matrices d'adjacence X^b des partitions de l'ensemble Π. L'application d'une classification hiérarchique ascendante sur la matrice $\tilde{Co}$ permet d'obtenir le consensus π^* des partitions. Dans les expérimentations, nous avons utilisé la stratégie d'agrégation de Ward (mais dans le cas général d'autres stratégies pourront être utilisées) et la partition obtenue par coupure du dendrogramme a été consolidée par une méthode des K-means. Les poids μ sont liés au niveau de ressemblance entre partitions.
En considérant les tables Z^b où chaque observation z_i^b est représentée par son vecteur référent w_{ci}^b, il est possible de définir une pseudo matrice compromis (individus $\times$ variables) $Z^* = \sum_{b=1}^{B} \mu^b Z^b$ à travers la moyenne des matrices associées aux tables, pondérée par les poids μ^b définis par STATIS.

L'application de SOM sur cette matrice fournit une visualisation compromis. La figure 1 en est une illustration.

Data	Obs	Dim	Cl	Data	Obs	Dim	Cl
IS	2310	18	7	D1	200	50	4
Glass	214	9	6	D2	200	50	4
Ionosphere	351	34	2	Wine	178	13	3
Iris	150	4	3				

TAB. 1 – *Description des tables ; Obs, Dim et Cl définissent respectivement le nombre d'observations, le nombre de variables et le nombre de classes dans les tables*

3.3 Évaluation

L'ensemble Π des partitions est obtenu à travers 30 applications de SOM en faisant varier des paramètres d'initialisations. Nous calculons ensuite le consensus des 30 partitions à l'aide de la méthode STATIS. Cet expérience est répété 25 fois pour les données IRIS, WINE, GLASS, IONOSPHERE et "Image segmentation (IS)" issues de UCI. Sur les données simulées D1 et D2 qui sont de types multi-vues, chaque vue est une table composée de 5 variables, le consensus est défini sur 10 partitions obtenues sur les vues. Cet expérience est aussi répété 25 fois. Afin de positionner la méthode proposée, que nous appelons, CSTATIS par rapport à quelques méthodes de consensus présentées dans la littérature, nous réalisons la même expérience avec des algorithmes de recherche de consensus basés sur la factorisation de matrice non-négative (NMF, Weighted NMF) présentés par Ding et al. (2006), la méthode d'ensemble cluster (CSPA) présentée par Strehl et Ghosh (2002). Le tableau 1 présente les caractéristiques des différentes tables. Nous utilisons l'indice de pureté suivant pour évaluer la similarité entre

deux partitions :

$$Pureté = \max_{c_i, c_j} \sum_{c_i^1, c_j^2} \frac{M(c_i^1, c_j^2)}{N} \tag{3}$$

où N désigne le nombre d'observations, c_k^1 une classe de partition 1 et c_k^1 une classe de la partition 2 et $M(c_i^1, c_j^2)$ est le nombre d'observations de la classe c_i^1 appartenant à la classe c_j^2. Le tableau 2 présente, par algorithme, la moyenne et l'écart type de l'indice de pureté pour les 25 expériences. CSTATIS donne, sauf pour la base IS, des résultats semblables à la méthode NMF et de performances meilleures que les méthodes W-NMF et CSPA présentées dans la littérature. La figure 1 visualise les classes

Data	SOM	CSTATIS	NMF	WNMF	CSPA
IS	0.59(0.04)	0.61(0.02)	0.59(0.05)	0.64(0.05)	0.61(0.03)
Glass	0.44(0.02)	0.43(0.01)	0.42(0.01)	0.40(0.01)	0.37(0.03
Ionosphere	0.69(0.01)	0.70(0.02)	0.70(0.01)	0.70(0.02)	0.68(0.01)
Iris	0.93(0.07)	0.98 (0.01)	0.98(0.007)	0.98(0.01)	0.90(0.12)
Wine	0.91(0.02)	0.95 (0.02)	0.95(0.008)	0.94(0.009)	0.92(0.01)
D1	0.67(0.08)	0.87(0.02)	0.88(0.01)	0.85(0.06)	0.57(0.08)
D2	0.72(0.10)	0.88(0.02)	0.88(0.06)	0.84(0.04)	0.51(0.06)

TAB. 2 – *Résultats du consensus de partition, on observe, la moyenne des puretés des algorithmes sur 25 expériences. CSTATIS est le résultat du consensus obtenu à l'aide de STATIS, NMF est le résultat du consensus obtenu à l'aide de l'algorithme de factorisation de matrice non-négative NMF, WNMF est le résultat de la version pondérée de NMF et CSPA est le résultat de l'algorithme d'ensemble clusters.*

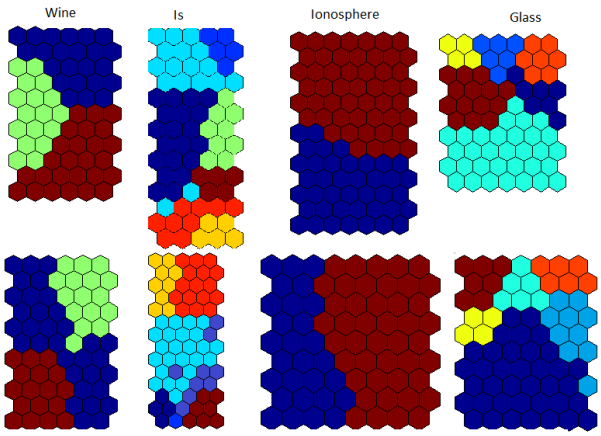

FIG. 1 – *Réprésentation de la carte consensus ; Les figures en haut correspondent à la carte consensus. En bas, une carte de l'ensemble de diversification pour chaque table. On observe une bonne conservation de la topologie des observations sauf pour la table IS*

4 Conclusion

La méthode proposée, basée sur STATIS, détermine des poids sur chaque partition qui sont ensuite utilisés pour déterminer la matrice des compromis des partitions. Une méthode de partitionnement appliquée à la matrice des compromis fournit le consensus recherché. Son évaluation sur des données réelles et simulées montre que la méthode améliore à travers son système de poids des performances de classification en terme de pureté par rapport à la méthode CSPA et W-NMF et des performances similaires à la méthode NMF.

Références

Ding, C., T. Li, W. Peng, et H. Park (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 126–135. ACM.

Fred, A. L. et A. K. Jain (2003). Robust data clustering. *2012 IEEE Conference on Computer Vision and Pattern Recognition 2*, 128.

Georgakis, A., H. Li, et M. Gordan (2005). An ensemble of som networks for document organization and retrieval. In *Int. Conf. on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pp. 6.
English

Gordon, A. et M. Vichi (1998). Partitions of partitions. *Journal of Classification 15*, 265–285.

Jiang, Y. et Z.-H. Zhou (2004). Som ensemble-based image segmentation. *Neural Processing Letters 20*(3), 171–178.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing 21(1-3)*.

Lavit, C., Y. Escoufier, R. Sabatier, et P. Traissac (1994). The act (statis method). *Computational Statistics & Data Analysis 18*(1), 97–119.

Régnier, S. (1983). Sur quelques aspects mathématiques des problèmes de classification automatique. *Mathématiques et Sciences humaines 82*, 13–29.

Saavedra, C., R. Salas, S. Moreno, et H. Allende (2007). Fusion of self organizing maps. In *Computational and Ambient Intelligence*, pp. 227–234. Springer.

Strehl, A. et J. Ghosh (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research 3*, 583–617.

Summary

In the context of the Self -Organizing Maps algorithm, we propose a new approach of "ensemble clusters" based on the STATIS method. This approach find a compromise matrix which synthesizes at best the similarity betwen various clusterings or maps. The maps fusion is then obtained by apply clustering algorithm on this compromise matrix. The method is illustrated on the real data from UCI and simulated data.