La carte GHSOM comme alternative à la SOM pour l'analyse exploratoire de données

Françoise Fessant*, Fabrice Clérot*
Pascal Gouzien*

* Orange Labs, 2 av. Pierre Marzin, 22307 Lannion, France
francoise.fessant@orange-ftgroup.com

Résumé. L'objecif de cet article est de faire de la carte auto-organisatrice hiérarchique (GHSOM) un outil utilisable dans le cadre d'une démarche d'analyse exploratoire de données. La visualisation globale est un outil indispensable pour rendre les résultats d'une segmentation intelligibles pour un utilisateur. Nous proposons donc différents outils de visualisation pour la GHSOM équivalents à ceux de la SOM.

1 Introduction

Le modèle des cartes auto-organisatrices hiérarchiques (ou GHSOM pour *Growing Hierarchical Self Organizing Map*) est un arbre de cartes SOM qui s'adapte aux données d'apprentissage par expansion ou agrandissement des feuilles SOM. La taille des branches et la configuration des feuilles varient en fonction des données. Ce modèle a été proposé initialement par Rauber et al. (2002) comme une alternative à la carte SOM traditionnelle. La carte SOM suppose de fixer a priori l'architecture initiale (le nombre de prototypes et la topologie du réseau). La GHSOM se contruit sans que l'utilisateur ait à définir la granularité du modèle ni sa profondeur. Seule la forme des feuilles est fixée a priori : les feuilles sont des grilles bidimensionnelles carrées. Le processus d'apprentissage est géré par différents paramètres qui contrôlent l'expansion et l'élargissement des feuilles. Moins contraint que la SOM, il offre de meilleures performances de quantification car ses prototypes se positionnent mieux dans l'espace des données. Dans cet article nous nous intéressons à l'adaptation des outils de visualisation et d'interprétation des classifications de la SOM à la GHSOM. L'objectif est d'en faire un outil utilisable dans le cadre d'une démarche d'analyse exploratoire de données pour laquelle il est nécessaire de disposer de représentations graphiques et de visualisations très parlantes des données aussi bien quantitatives que qualitatives.

2 La carte GHSOM

Le processus d'apprentissage combine une phase d'élargissement et une phase d'expansion qui sont controlées par deux paramètres α et β. Les cartes d'un niveau sont indépendantes les unes des autres. Le modèle est initialisé par la création de deux cartes SOM :

- Une carte avec un seul prototype est créée au niveau 0 de l'arbre. Le prototype est défini par un vecteur m_0 égal à la moyenne de tous les vecteurs d'entrée.
- Une carte SOM de taille 2x2 est ensuite initialisée au niveau 1 avec 4 vecteurs choisis aléatoirement $m_i, i \in \{1, 2, 3, 4\}$. La carte ainsi créée est ensuite entraînée avec l'algorithme standard.

La stratégie d'apprentissage spécifique à l'arbre GHSOM commence ensuite. Plusieurs actions peuvent être effectuées sur la carte :

- Elargissement de la carte par insertion d'une ligne et d'une colonne pour améliorer la qualité de représentation du niveau courant. L'algorithme d'apprentissage standard est réappliqué à la carte.
- Expansion de la carte à partir d'un ou de plusieurs prototypes de la carte courante.
- Arrêt si la carte représente correctement les données.

Lors de l'ajout d'une nouvelle carte le processus se poursuit sans que la hiérarchie précédemment créée ne soit remise en cause et modifiée.

La stratégie d'élargissement des cartes commence avec le calcul de l'erreur moyenne de quantification de la carte du niveau 0 qui sert d'erreur de référence pour toutes les cartes qui vont être créées ensuite dans la hiérarchie. On cherche à élargir la carte d'un niveau jusqu'à ce que les données soient bien représentées. Le critère de représentation est basé sur l'erreur de quantification moyenne de la carte (donnée par la moyenne des erreurs de quantification de l'ensemble des prototypes de la carte).

$$EMQ_l = \frac{1}{u} \sum_{i=1}^{u} emq_i \tag{1}$$

où EMQ_l est l'erreur de quantification de la carte l à un niveau hiérarchique donné, emq_i l'erreur de quantification du prototype i, u le nombre de prototypes de la carte.

L'idée sous-jacente est que chaque couche du GHSOM explique la déviation standard des données projetées dans la couche suivante. Les cartes SOM d'un niveau donné s'élargissent jusqu'à ce que l'erreur de quantification du prototype de la couche précédente, qui a généré l'expansion, soit réduite dans le rapport α.

$$EMQ_l \geq \alpha . emq_u \tag{2}$$

où emq_u est l'erreur moyenne de quantification du prototype u au niveau précédent. Par exemple, $EMQ_1 \geq \alpha . emq_0$ pour la carte de niveau 1.

Aussi longtemps que le critère (2) reste vrai, on ajoute une ligne et une colonne à la carte. Une fois que la condition est atteinte, l'élargissement de la carte est stoppé et les différents prototypes de la carte sont observés pour voir si un niveau supplémentaire est nécessaire. Les prototypes avec une grande erreur de quantification vont être étendus par création d'une nouvelle carte à un niveau supplémentaire. Un prototype i qui vérifie le critère $emq_i > \beta . emq_0$ subira l'expansion. Le processus s'arrête quand plus aucune unité ne doit subir l'expansion.

α et β sont les deux paramètres que l'utilisateur doit régler pour contrôler le modèle.

- α contrôle l'élargissement de la carte et fixe la forme de l'arbre. α permet de gérer l'erreur de reconstruction à chaque étage de la hiérarchie ; plus α est grand, plus le rapport de variance d'un niveau à l'autre est important et donc plus la carte s'élargit pour mieux représenter les données.

– β contrôle la croissance de l'arbre et le niveau de coupure.

Les valeurs généralement choisies pour les deux paramètres sont telles que : $1 > \alpha >> \beta > 0$. Il n'y a pas de règle pour déterminer les bonnes valeurs des paramètres. Ces valeurs et donc la vitesse à laquelle on veut organiser les prototypes dépendent de la structure et du type de représentation que l'on souhaite favoriser et donc du type d'exploration que l'on veut faire :

– Un arbre profond avec de petites cartes sur chaque feuille qui chacune explique une faible partie des caractéristiques des données à chaque niveau, mais privilégie l'aspect hiérarchique. Dans ce cas on prendra une valeur élevée de α.
– Un arbre peu profond avec des cartes plus larges dans les feuilles qui expliquent une plus grande partie des données à chaque niveau au détriment de la structure hiérarchique. Dans ce cas on choisira une valeur plus faible.

3 Application à une base de données artificielles

On observe le comportement de la GHSOM pour l'analyse exploratoire de données à travers un exemple artificiel 2-D. L'exemple consiste en 8 000 points à deux dimensions répartis sur le plan avec différentes zones de densité [1].

Les outils de visualisation de la SOM sont adaptés à la GHSOM : carte des hits et cartes des poids. On propose également une visualisation des distances entre les neurones de la carte concurrente à la U-matrice.

3.1 Mise en oeuvre de la GHSOM

La GHSOM est construite avec le couple de paramètres $\alpha = 0,5$ et $\beta = 0,01$. On privilégie pour cet exemple l'aspect hiérarchique en construisant un arbre profond avec des petites cartes.

Un test du χ^2 a été rajouté lors de l'apprentissage des feuilles SOM. Pour chaque carte nouvellement créée, on compare les comptes des populations par neurone pour la population d'apprentissage (qui a servi à entraîner la carte) et une population de validation. On teste au sens du χ^2 l'hypothèse nulle selon laquelle ces deux vecteurs de compte sont issus d'une même distribution. Si cette hypothèse est rejetée (à un niveau fixé ici à 5%), on considère que les neurones ont sur-appris la population d'apprentissage et on arrête le déploiement de la hiérarchie au niveau supérieur.

On obtient sur cet exemple un arbre à 5 niveaux avec 88 cartes 2x2 dans la hiérarchie. Chacun des 4 neurones du niveau 1 a généré une carte 2x2 au niveau 2 et chacun des neurones du niveau 2 a creusé au niveau 3. Il y a 352 paramètres au total dans l'architecture et 265 poids dans la couche terminale. Les neurones terminaux sont de niveau 3, 4 ou 5.

En terme de performance de quantification, la GHSOM s'est montrée plus performante que la SOM. On a obtenu une erreur de quantification de 6,5 pour la GHSOM et une erreur de 6,7 pour la SOM avec une carte 19x19.

3.2 Outils pour la visualisation

Orientation des cartes

[1] Ce jeu de données provient du cours de Data Mining que Philippe Leray effectue à l'INSA de Rouen http ://moodle.insa-rouen.fr/course/

Dans la carte GHSOM, les cartes d'un niveau hiérarchique n'ont pas de relation entre elles en dehors de leur provenance commune par le biais de la carte du niveau précédent. Elles peuvent donc être orientées sans qu'il y ait une relation de proximité entre elles. Chan et Pampalk (2002) ont proposé une stratégie pour orienter les différentes cartes d'un même niveau à partir de l'organisation de la carte au niveau précédent. La stratégie s'applique avant l'apprentissage, lors de l'initialisation de la nouvelle carte. Les poids du prototype qui va subir l'expansion dans la carte mère et les poids des prototypes voisins sont partiellement recopiés pour l'initialisation de la carte créée au niveau suivant.

Nous proposons une stratégie différente d'orientation des cartes qui consiste à agir après apprentissage. Lorsque qu'une nouvelle carte est créée par l'expansion d'un prototype, elle peut prendre 8 orientations différentes par rotation et retournement horizontal ou vertical. On recherche la meilleure orientation pour la carte nouvellement créée et on fait l'hypothèse que c'est celle qui minimise la distance de ses prototypes aux prototypes voisins du prototype parent dans la carte de niveau précédent.

Cartes des poids

Les cartes de poids sont des visualisations couramment utilisées notamment pour corréler les différentes variables entre elles. La Figure 1 présente la carte des poids de la première variable pour l'exemple analysé. On cherche, comme dans la SOM classique, à visualiser les valeurs des variables. On ne s'intéresse qu'aux valeurs des poids terminaux (niveaux 3, 4 et 5 de l'arbre). Un carré clair correspond à une forte valeur de la variable, un carré sombre à une faible valeur.

On compare, sur la première variable, la carte des poids obtenue après application de la stratégie d'orientation des cartes SOM dans l'arbre et la carte des poids que l'on aurait obtenue si on avait laissé les cartes SOM se positionner librement.

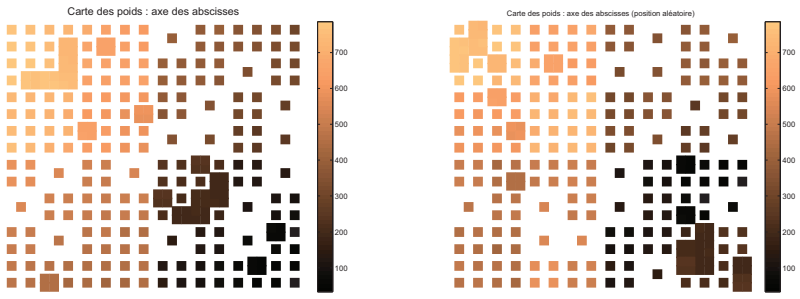

FIG. 1 – *Cartes des poids de la variable axe des abscisses après application de la stratégie d'orientation des cartes SOM (gauche) ou avec une orientation non contrainte (droite).*

La représentation de gauche montre une carte de poids bien structurée avec un gradient de valeurs sur la diagonale SE-NO. La structure d'arbre de la GHSOM, bien qu'elle soit moins contrainte que celle de la SOM et adaptée à partir des données, permet de conserver les propriétés de voisinage entre prototypes comme dans la SOM standard. Les cartes des poids de

la GHSOM sont exploitables de la même manière que le sont les cartes de poids de la SOM. La comparaison des deux représentations Figure 1 montre que le positionnement libre des cartes entraîne une rupture dans l'organisation avec notamment une rupture dans le gradient des valeurs (quart NO et quart SE par exemple).

Cette visualisation reflète également sur un plan l'organisation hiérarchique et les différentes zones qui se dégagent. Les zones les plus denses en prototypes correspondent aux zones ayant été les plus creusées dans la hiérarchie.

Carte des distances entre prototypes

On calcule la distance entre deux prototypes voisins dans la carte (distance euclidienne) et on visualise cette distance à l'aide d'un lien. La carte résultante est donnée Figure 2 (à gauche). Le lien entre les prototypes est d'autant plus large qu'ils sont proches. La couleur sombre code une faible distance entre les prototypes, la couleur claire code une grande distance. Si on se place dans le cadre d'un voisinage diamant, chacun des prototypes possède au minimum 4 voisins à un niveau donné. On balaie les 4 directions (N,S,E,O) et on retient le ou les neurones les plus proches sur la grille dans chaque direction. Pour des cartes de niveaux hiérarchiques différents, un neurone peut avoir plusieurs voisins à la même distance sur la grille dans une direction donnée. On rappelle la position des prototypes et la taille de leur population sur la carte. Une représentation de la position des poids dans l'espace des données est juxtaposée à la carte des distances. L'ensemble des poids de la structure est affiché pour la GHSOM. Chacune des zones de densité est repérée par un symbole. On fait correspondre à chaque prototype sur la carte des distances la zone de densité qu'il représente.

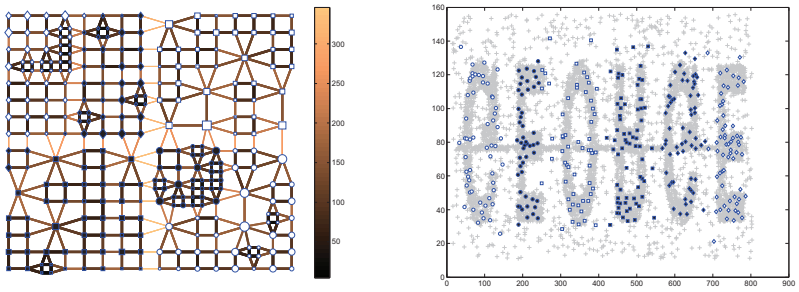

FIG. 2 – *Carte des distances entre prototypes (gauche) et répartition des poids dans l'espace des données (droite) pour la GHSOM.*

On met en évidence différentes zones dans la carte. Les zones de ruptures (traits clairs, fins) coïncident avec les différentes branches de l'arbre et séparent les zones homogènes (traits sombres plus épais) qui correspondent aux feuilles. On peut clairement rattacher les 6 zones homogènes qui se dégagent dans la carte aux 6 zones de densités identifiées dans les données. Les zones les plus denses sont également celles qui sont les plus peuplées par les prototypes. Les zones de rupture correspondent à des zones frontières auxquelles la GHSOM alloue peu de neurones.

4 Conclusion

Dans cet article nous nous sommes intéressés au modèle des cartes auto-organisatrices hiérarchique (GHSOM). La structure GHSOM est un arbre de cartes SOM qui s'adapte à la distribution des données pendant la phase d'apprentissage par un processus d'élargissement ou d'expansion de ses feuilles SOM. Ce modèle nécessite de définir deux paramètres qui vont conditionner la forme de l'arbre et sa profondeur. On ne fait aucune hypothèse a priori sur la forme de l'arbre ni sur le nombre de paramètres des feuilles.

La SOM est très utilisée pour l'analyse exploratoire de données car on dispose d'outils de représentations graphiques et de visualisation très parlantes des données. Nous avons montré que l'on peut transposer les différentes visualisations de la SOM à la GHSOM (carte des poids, carte des observations, carte des distances entre prototypes).

Le comportement de la GHSOM a été comparé à celui d'une SOM classique. Le modèle GHSOM a montré un comportement intéressant avec des prototypes qui se positionnent mieux dans l'espace réel des données que les prototypes de la SOM. La GHSOM ne consacre pas de prototypes aux zones frontières comme c'est le cas dans la SOM traditionnelle (donc à des zones peu peuplées). Les frontières dans la carte GHSOM correspondent aux branches de l'arbre. Il y a globalement une meilleure utilisation des prototypes par le découpage produit par la GHSOM.

Ce comportement s'explique par le fait que les paramètres de la GHSOM ne sont pas contraints à évoluer sur une topologie fixée a priori avant l'apprentissage et avec un nombre de paramètres fixe. Le modèle s'adapte au données au cours de l'apprentissage. En ce sens le modèle parait plus souple que la carte SOM. Cependant la structure de la GHSOM est quand même suffisamment contrainte (les feuilles de l'arbre sont des cartes SOM standards) pour que les propriétés de voisinage qui sont le grand intérêt des modèles SOM puissent être conservées et exploitées.

Plus performant que la SOM en quantification et doté d'outils de visualisation similaires, la GHSOM se révèle un outil très intéressant pour l'analyse exploratoire de données.

Références

Chan, A. et C. Pampalk (2002). Growing hierarchical self organising map (ghsom) toolbox: Visualisations and enhancements. *In the Proceedings of the* 9^{th} *International Conference on Neural Information Processing (ICONIP'02)*, 2537–2541.

Rauber, A., D. Merkl, et M. Dittenbach (2002). The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks 13*, 1331–1341.

Summary

In this paper we show how the visualisation tools of the SOM can be transposed to the Growing Hierarchical Self Organizing Map (GHSOM) model. With these visualisations the GHSOM can be used efficiently for data exploratory analysis.