SOMbrero : Cartes auto-organisatrices stochastiques pour l'intégration de données décrites par des tableaux de dissimilarités

Laura Bendhaïba[a], **Madalina Olteanu**[a] et **Nathalie Villa-Vialaneix**[a,b]

[a]SAMM, Université Paris 1
F-75634 Paris - France
laurabendhaiba@gmail.com
{madalina.olteanu,nathalie.villa}@univ-paris1.fr

[b]INRA, UR875, MIAT
F-31326 Castanet Tolosan - France

Mots clefs : cartes auto-organisatrices, dissimilarités, graphes, classification, visualisation

Dans de nombreuses situations réelles, les individus sont décrits par des jeux de données multiples qui ne sont pas nécessairement de simples tableaux numériques mais peuvent être des données complexes (graphes, variables qualitatives, texte...). Un cas typique est celui des graphes étiquetés dans lequel les individus (les sommets du graphe) sont décrits à la fois par leurs relations les uns aux autres mais aussi par des attributs de natures diverses. Dans [5, 2], nous avons proposé d'utiliser des cartes auto-organisatrices [1] pour combiner classification et visualisation en projetant les individus étudiés sur une grille de faible dimension. Notre approche permet de traiter des données non numériques par le biais de noyaux ou de dissimilarités, et est basée sur une version stochastique de l'apprentissage de cartes auto-organisées, comme décrit dans [4, 3]. Les différentes dissimilarités sont combinées et la combinaison est optimisée au cours de l'apprentissage de la carte.
Nous avons testé notre approche sur un jeu de données simulé : dans celui-ci, les observations sont décrites par un graphe séparé en deux groupes denses de sommets (figure 1, en haut à gauche), les sommets étant étiquetés par des valeurs numériques de $\mathbb{R}^2$ tirées selon deux Gaussiennes (figure 1) ainsi que par un facteur à deux niveaux. Seules les trois informations permettent de retrouver les 8 groupes de sommets, représentés par 8 couleurs différentes sur la figure 1. La combinaison des trois informations sous la forme de trois tableaux de dissimilarités (longueur du plus court chemin entre deux sommets pour le graphe, distance euclidienne pour les étiquettes numériques et distance de Dice pour les facteurs) permet de retrouver les huit groupes initiaux avec une bonne précision et de bien les organiser sur la carte (figure 1, en bas à droite). L'apprentissage adaptatif des distances donne un poids prépondérant à la dissimilarité basée sur la valeur du facteur qui est la seule valeur non bruitée (figure 1).
La méthodologie proposée est en voie d'implémentation dans un package R appelé **SOMbrero**. La version 0.1 du package, disponible depuis mars 2013 propose l'implémentation de l'algorithme de cartes auto-organisatrices pour des données numériques simples ainsi que diverses fonctionnalités permettant l'interprétation (fonctionnalité graphique pour visualiser les niveaux des diverses variables, les valeurs des prototypes de la carte...). Le package n'est pas encore disponible sur le CRAN mais peut être téléchargé à `http://tuxette.nathalievilla.org/?p=1099&lang=en` (sources et compilation windows).

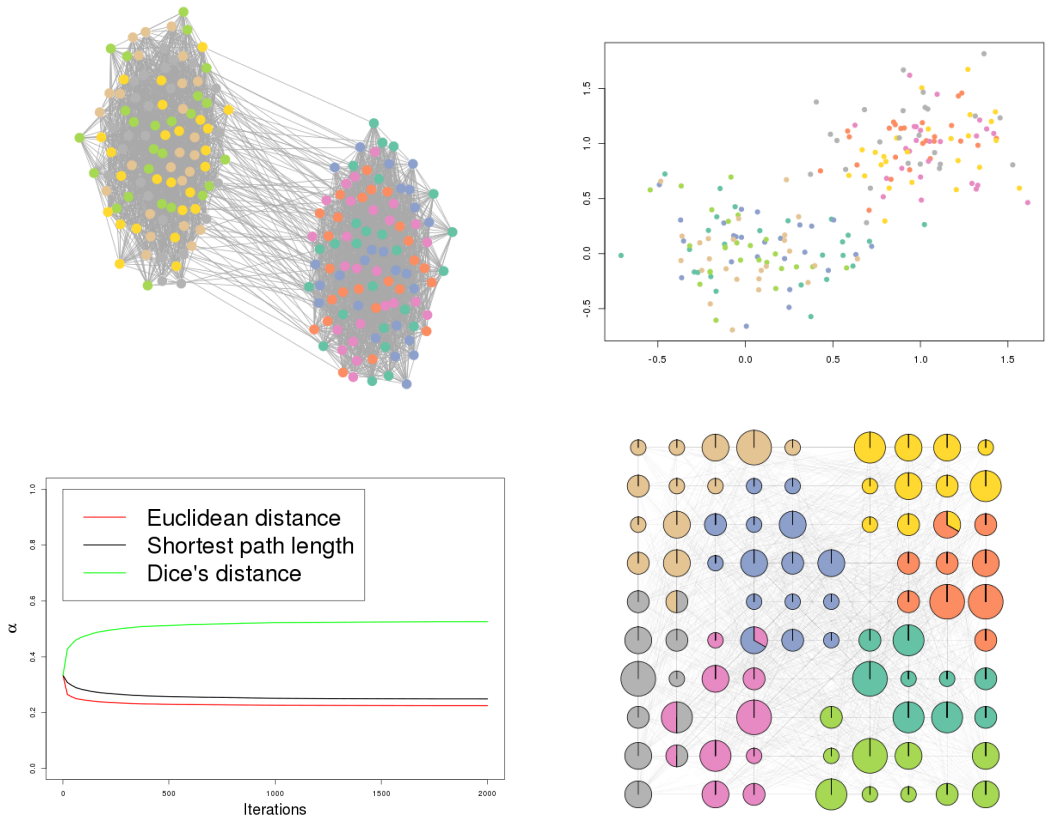

Figure 1: Données simulées : graphes et valeurs des étiquettes numériques des sommets (en haut à gauche et à droite). Évolution des poids des diverses dissimilarités (en bas à gauche). Carte finale obtenue (en bas à droite, les couleurs représentent les classes initiales, les aires des disques sont proportionnelles au nombre d'observations de la classe)

References

[1] T. Kohonen. *Self-Organizing Maps, 3rd Edition*, volume 30. Springer, Berlin, Heidelberg, New York, 2001.

[2] M. Olteanu, N. Villa-Vialaneix, and C. Cierco-Ayrolles. Multiple kernel self-organizing maps. In M. Verleysen, editor, *XXIst European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, 2013. d-side publications. Forthcoming.

[3] M. Olteanu, N. Villa-Vialaneix, and M. Cottrell. On-line relational som for dissimilarity data. In P.A. Estevez, J. Principe, P. Zegers, and G. Barreto, editors, *Advances in Self-Organizing Maps (Proceedings of WSOM 2012)*, volume 198 of *AISC (Advances in Intelligent Systems and Computing)*, pages 13–22, Santiago, Chile, 2012. Springer Verlag, Berlin, Heidelberg.

[4] N. Villa and F. Rossi. A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph. In *6th International Workshop on Self-Organizing Maps (WSOM)*, Bielefeld, Germany, 2007. Neuroinformatics Group, Bielefeld University.

[5] N. Villa-Vialaneix, M. Olteanu, and C. Cierco-Ayrolles. Carte auto-organisatrice pour graphes étiquetés. In *Actes des Ateliers FGG (Fouille de Grands Graphes), colloque EGC (Extraction et Gestion de Connaissances)*, Toulouse, France, 2013.