# Accélération des cartes auto-organisatrices sur tableau de dissimilarités par séparation et évaluation

Brieuc Conan-Guez*, Fabrice Rossi**

*LITA EA3097, Université de Metz, Ile du Saulcy, F-57045 Metz
Brieuc.Conan-Guez@univ-metz.fr
**Projet AxIS, INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105,
78153 Le Chesnay Cedex
Fabrice.Rossi@inria.fr

**Résumé.** Dans cet article, nous proposons une nouvelle implémentation d'une adaptation des cartes auto-organisatrices de Kohonen (SOM) aux tableaux de dissimilarités. Cette implémentation s'appuie sur le principe de séparation et évaluation afin de réduire le temps de calcul global de l'algorithme. Une propriété importante de ce nouvel algorithme tient au fait que les résultats produits sont strictement identiques à ceux de l'algorithme original.

## 1 Introduction

Dans beaucoup d'applications réelles, les individus étudiés ne peuvent pas être décrits efficacement par des vecteurs numériques : on pense par exemple à des données de tailles variables, comme les séquences d'acides aminés constituant des protéines, ou bien à des données (semi-)structurées (par exemple des documents XML). Une solution pour traiter de telles données est de s'appuyer sur une mesure de dissimilarité permettant de comparer les individus deux à deux.

Nous nous intéressons dans cet article à une adaptation des cartes auto-organisatrices de Kohonen (SOM pour *Self-Organizing Map*, (Kohonen, 1995)) aux données décrites seulement par un tableau de dissimilarités, proposée dans (Kohonen et Somervuo, 1998, 2002). Nous désignons cette adaptation par le sigle DSOM (pour *Dissimilarity SOM*). On trouve aussi dans la littérature l'appellation *Median SOM*. Le DSOM et ses variantes ont été appliqués avec succès à des problèmes réels d'analyse exploratoire portant sur des protéines, des données météorologiques (El Golli et al., 2004a), spectrométriques (El Golli et al., 2004b) ou encore provenant de l'usage d'un site web (Rossi et al., 2005; El Golli et al., 2006). Comme dans le cas classique, les résultats obtenus par le DSOM en terme de la qualité de la classification sont comparables à ceux obtenus avec d'autres méthodes applicables à des tableaux de dissimilarités (comme PAM (Kaufman et Rousseeuw, 1987) ou les algorithmes de type nuées dynamiques (Celeux et al., 1989)). L'avantage du DSOM sur ces algorithmes réside dans prise en compte d'une structure *a priori* qui permet la représentation graphique des classes et prototypes obtenus, ce qui facilite l'analyse exploratoire des données étudiées.

Le problème majeur du DSOM réside dans son temps de calcul : pour $N$ observations et $M$ classes, le coût algorithmique du DSOM est de $O(N^2M + NM^2)$ par itération. A titre de