2S-SOM : une méthode de soft-subspace clustering pour données multi-blocs basée sur les cartes topologiques auto-organisées

Mory Ouattara[*,**] Ndèye Niang[*] 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. Nous proposons une méthode de soft subspace clustering basée sur les cartes topologiques pour la classification d'individus décrits par des variables structurées en blocs homogènes. L'algorithme nommé Soft Subspace SOM (2S-SOM) consiste à optimiser la fonction de coût de SOM modifiée en introduisant des poids adaptatifs sur les blocs et sur les variables de chaque bloc. Cette double pondération permet de distinguer les blocs les plus importants prenant ainsi en compte la structuration en blocs, et d'identifier pour chaque bloc les variables les plus informatives pour les classes. La méthode permet alors de déterminer simultanément les groupes d'individus et leurs sous espaces caractéristiques optimaux. La méthode est illustrée sur des données réelles issues des bases de l'UCI repository of machine learning et sur des données simulées.

1 Introduction

Les méthodes de classification non-supervisées (ou clustering) permettent d'explorer des données non-labélisées dans le but de trouver des groupes d'observations homogènes et bien séparés. Les récentes avancées technologiques en capacité de stockage d'informations d'une part, et la multiplication des sources d'informations d'autre part, contribuent à la mise en place de bases de données complexes et de grande dimension. Dans des domaines tels que la génétique, la finance, le traitement de données textuelles et les études environnementales, par exemple, on rencontre des données de grande dimensions. Ce qui conduit à avoir plusieurs blocs de variables caractérisant chacune une vue particulière sur les données, on parle de données multi-vues ou multi-blocs. C'est notamment le cas des études environnementales sur la pollution de l'air intérieur où les vues sont associées à des thématiques précises : concentrations des polluants de l'air intérieur, informations collectées sur la santé des occupants et sur l'aménagement des environnements intérieurs (Kirchner et al., 2011). Par ailleurs, l'usage de capteurs est souvent nécessaire pour mesurer les concentrations des polluants et la possible