

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 s'intéresse aux méthodes d'ensemble clusters dédiées aux cartes topologiques. Comme les méthodes de partitionnement classiques (K moyennes, CAH), les