

Segmentation hiérarchique des cartes topologiques

Mustapha Lebbah^{*,**}, Hanane Azzag^{**}

^{*} LIM&BIO - UFR (SMBH)- Université Paris 13,
74, rue Marcel Cachin 93017 Bobigny Cedex France

^{**}LIPN - UMR 7030

Université Paris 13 - CNRS

99, av. J-B Clément - F-93430 Villetaneuse

{hanane.azzag, mustapha.lebbah}@lipn.univ-paris13.fr

Résumé. Dans ce papier, nous présentons une nouvelle mesure de similarité pour la classification des référents de la carte auto-organisatrice qui sera réalisée à l'aide d'une nouvelle approche de classification hiérarchique. (1) La mesure de similarité est composée de deux termes : la distance de Ward pondérée et la distance euclidienne pondérée par la fonction de voisinage sur la carte topologique. (2) Un algorithme à base de fourmis artificielles nommé AntTree sera utilisé pour segmenter la carte auto-organisatrice. Cet algorithme a l'avantage de prendre en compte le voisinage entre les référents et de fournir une hiérarchie des référents avec une complexité proche du $n\log(n)$. La segmentation incluant la nouvelle mesure est validée sur plusieurs bases de données publiques.

1 Introduction

Le problème de la classification de données est identifié comme une des problématiques majeures en extraction des connaissances à partir de données. Depuis des décennies, de nombreux sous-problèmes ont été identifiés, comme par exemple la sélection des données ou des variables, la variété des espaces de représentation (numérique, symbolique, etc), l'incrémentalité, la nécessité de découvrir des concepts, ou d'obtenir une hiérarchie, etc. La popularité, la complexité et toutes ces variantes du problème de la classification de données, (Jain et al. (1999)), ont donné naissance à une multitude de méthodes de résolution. Ces méthodes peuvent faire appel à des principes heuristiques ou encore mathématiques.

Les méthodes qui nous intéressent dans ce travail, sont celles qui permettent de faire de la classification non supervisée de données en utilisant les cartes topologiques (appelées aussi SOM :Self-organizing Map). Celles-ci sont souvent utilisées parce qu'elles sont considérées à la fois comme outils de visualisation et de partitionnement non supervisé de différents types de données (quantitatives et qualitatives). Elles permettent de projeter les données sur des espaces discrets qui sont généralement en deux dimensions. Le modèle de base, proposé par Kohonen (Kohonen (2001)), est uniquement dédié aux données numériques. Des extensions et des reformulations du modèle de Kohonen ont été proposées dans la littérature, (Bishop et al. (1998);