Carte auto-organisatrice probabiliste sur données binaires

Rodolphe Priam, Mohamed Nadif

LITA, Université de Metz
Ile du Saulcy, 57045 Metz

Résumé. Les méthodes factorielles d'analyse exploratoire statistique définissent des directions orthogonales informatives à partir d'un ensemble de données. Elles conduisent par exemple à expliquer les proximités entre individus à l'aide d'un groupe de variables caractéristiques. Dans le contexte du datamining lorsque les tableaux de données sont de grande taille, une méthode de cartographie synthétique s'avère intéressante. Ainsi une carte auto-organisatrice (SOM) est une méthode de partitionnement munie d'une structure de graphe de voisinage -sur les classes- le plus souvent planaire. Des travaux récents sont développés pour étendre le SOM probabiliste *Generative Topographic Mapping* (GTM) aux modèles de mélanges classiques pour données discrètes. Dans ce papier nous présentons et étudions un modèle génératif symétrique de carte auto-organisatrice pour données binaires que nous appelons *Bernoulli Aspect Topological Model* (BATM). Nous introduisons un nouveau lissage et accélérons la convergence de l'estimation par une initialisation originale des probabilités en jeu.

1 Introduction

La visualisation des corrélations et similarités principales dans un échantillon de données est l'objectif des méthodes factorielles (Lebart et al., 1984). Ces méthodes cherchent souvent des directions informatives orthogonales dans un nuage de données. Ces directions concentrent l'essentiel de la variance projetée car l'inertie est porteuse de sens. Une décomposition pertinente de l'inertie sur des plans de projection révèle quels individus sont similaires et quelles variables sont dépendantes. Bien que ces méthodes soient très pertinentes, les grands échantillons de données demandent de nouvelles méthodes efficaces pour leur analyse. Dans ce contexte, les cartes de Kohonen (1997) sont connues dans le domaine de l'analyse exploratoire des données pour généraliser les méthodes factorielles telles que la méthode d'Analyse en Composantes Principales ou ACP (Lebart et al., 1984) pour les données continues. Plus généralement, les cartes auto-organisatrices ou SOM (Kohonen, 1997) sont des méthodes de classification avec une contrainte de voisinage sur les classes conférant un sens topologique à la partition finale. Le GTM ou Generative Topographic Mapping (Bishop et al., 1998) est une carte auto-organisatrice probabiliste avec des contraintes sur les moyennes d'un mélange gaussien pour données continues, mais ce modèle est inopérant pour des données catégorielles ou binaires. Des modèles récents (Girolami, 2001; Kabán et Girolami, 2001; Tipping, 1999) ont été proposés pour étendre le GTM aux modèles de mélanges classiques pour données discrètes. Hofmann et Puzicha (1998) ont par contre proposé l'approche du modèle symétrique à *aspects*