Binary Block GTM : Carte auto-organisatrice probabiliste pour les grands tableaux binaires

Rodolphe Priam*, Mohamed Nadif**, Gérard Govaert***

*LMA Poitiers, UMR CNRS 6086, Université de Poitiers,
BP 30179, 86962 Futuroscope Chasseneuil Cedex, France
rpriam@gmail.com
**CRIP5, Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris, France
mohamed.nadif@univ-paris5.fr
***Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne,
BP 20529, 60205 Compiègne Cedex, France
gerard.govaert@utc.fr

Résumé. Ce papier présente un modèle génératif et son estimation permettant la visualisation de données binaires. Notre approche est basée sur un modèle de mélange de lois de Bernoulli par blocs et les cartes de Kohonen probabilistes. La méthode obtenue se montre à la fois parcimonieuse et pertinente en pratique.

1 Introduction

Bien que les méthodes d'analyse factorielle soient très puissantes et contribuent efficacement à la visualisation des données, les grands échantillons nécessitent de nouvelles méthodes mieux adaptées. En effet, les algorithmes de décomposition matricielle rencontrent leurs limites sur les grands tableaux numériques ; en outre, la construction de nombreux plans de projection, du fait des grandes dimensions, rend la tâche d'interprétation difficile pour recouper les informations disséminées sur ces plans. Finalement une grande quantité de données implique une grande quantité d'informations à synthétiser et des relations complexes entre individus et/ou variables étudiés. Il est alors possible, dans ce contexte, d'utiliser les cartes de Kohonen ou cartes auto-organisatrices (SOM) (Kohonen, 1997) qui sont des méthodes de classification automatique utilisant une contrainte de voisinage sur les classes pour conférer un sens topologique aux partitions obtenues. La carte auto-organisatrice originelle peut être vue comme une variante de l'algorithme des *k-means* (MacQueen, 1967) intégrant une contrainte d'ordre topologique sur les centres.

Lorsque la matrice des données $\mathbf{x}$ est définie sur un ensemble I d'objets (lignes, observations) et un ensemble J de variables (colonnes, attributs), différentes approches de classification automatique sont utilisées et la plupart des algorithmes proposés concerne généralement un des deux ensembles. Ces algorithmes peuvent être modélisés par différentes approches. Celle qui a suscité le plus d'intérêt ces dernières années est incontestablement l'approche modèle de mélange (McLachlan et Peel, 2000). Dans ce cadre, il a été proposé diverses versions probabilistes de SOM telles que dans (Lebbah et al., 2007; Verbeek et al., 2005; Luttrell, 1994).