Application des réseaux bayésiens à l'analyse des facteurs impliqués dans le cancer du Nasopharynx

Alexandre Aussem*, Sergio Rodrigues de Morais*, Marilys Corbex**

*Université de Lyon 1,
EA 2058 PRISMa, F-69622 Villeurbanne
aaussem@univ-lyon1.fr,
**Unité d'épidémiologie génétique,
Centre International de Recherche sur le Cancer (CIRC),
150 cours Albert Thomas - 69280 Lyon Cedex 08
corbex@iarc.fr

Résumé. L'apprentissage de la structure des réseaux bayésien à partir de données est un problème NP-difficile. Une nouvelle heuristique de complexité polynômiale, intitulée Polynomial Max-Min Skeleton (PMMS), a été proposée en 2005 par Tsamardinos et al. et validée avec succès sur de nombreux bancs d'essai. PMMS présente, en outre, l'avantage d'être performant avec des jeux de données réduits. Néanmoins, comme tous les algorithmes sous contraintes, celui-ci échoue lorsque des dépendances fonctionnelles (déterministes) existent entre des groupes de variables. Il ne s'applique, par ailleurs, qu'aux données complètes. Aussi, dans cet article, nous apportons quelques modifications pour remédier à ces deux problèmes. Après validation sur le banc d'essai $Asia$, nous l'appliquons aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC) de 1289 observations, 61 variables et 5% de données manquantes issues d'un questionnaire. L'objectif est de dresser un profil statistique type de la population étudiée et d'apporter un éclairage utile sur les différents facteurs impliqués dans le NPC.

1 Introduction

L'apprentissage de la *structure* des réseaux bayésiens (RB) à partir de données est un problème ardu ; la taille de l'espace des graphes orientés sans circuits (*DAG* en anglais) est super-exponentielle en fonction du nombre de variables et le problème combinatoire associé est NP-difficile (Chickering et al., 2004). Deux grandes familles de méthodes existent pour l'apprentissage de la structure des RB : celles fondées sur la satisfaction de contraintes d'indépendance conditionnelle entre variables et celles à base de score fondées sur la maximisation d'un score (BIC, MDL, BDe, etc.). Les deux méthodes ont leurs avantages et leurs inconvénients. Les méthodes sous contraintes sont déterministes, relativement rapides et bénéficient des critères d'arrêt clairement définis. Les contraintes imposées à la structure du graphe proviennent des informations statistiques sur les dépendances et indépendances conditionnelles observées dans les données. Elles reposent cependant sur un niveau de signification arbitraire