Apprentissage de la structure des réseaux bayésiens à partir des motifs fréquents corrélés : application à l'identification des facteurs environnementaux du cancer du Nasopharynx

Alexandre Aussem*, Zahra Kebaili*, Marilys Corbex**, Fabien De Marchi***

*Equipe COMAD, Lab. PRISMa, Université Lyon 1,
alexandre.aussem@univ-lyon1.fr,
**Unité d'épidémiologie génétique,
Centre International de Recherche sur le Cancer (CIRC), Lyon,
corbex@iarc.fr,
***LIRIS UMR CNRS 5205, Université Lyon 1,
fabien.demarchi@liris.cnrs.fr

Résumé. L'apprentissage de structure des réseaux bayésien à partir de données est un problème NP-difficile pour lequel de nombreuses heuristiques ont été proposées. Dans cet article, nous proposons une nouvelle méthode inspirée des travaux sur la recherche de motifs fréquents corrélés pour identifier les causalités entre les variables. L'algorithme opère en quatre temps : (1) la découverte par niveau des motifs fréquents corrélés minimaux ; (2) la construction d'un graphe non orienté à partir de ces motifs ; (3) la détection des V_structures et l'orientation partielle du graphe ; (4) l'élimination des arêtes superflues par des tests d'indépendance conditionnelle. La méthode, appliquée au réseau *Asia*, permet de retrouver la structure du graphe initial. Nous l'appliquons ensuite aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC). 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

Les réseaux d'inférence bayésiens (RB) sont des outils d'apprentissage numérique qui permettent de rendre compte de relations causales entre des variables aléatoires et de construire un raisonnement probabiliste à partir de connaissances, parfois incertaines et incomplètes, consignées dans les bases de données. L'apprentissage automatique des *valeurs numériques* des probabilités conditionnelles s'opère d'ordinaire à partir d'un ensemble d'apprentissage, même incomplet, si la structure du réseau est *connue*. En revanche, l'apprentissage de la *structure* du RB à partir de données est plus problématique ; la taille de l'espace de recherche est super-exponentielle en fonction du nombre de variables et le problème combinatoire associé est NP-difficile. Deux grandes familles de méthodes existent : celles fondées sur la recherche de causalités via des tests d'indépendance conditionnelle et celles fondées sur la maximisation d'un score. Avec les méthodes à base de score, l'ajout d'un arc repose sur un compromis entre