Apprentissage de structure des réseaux bayésiens et données incomplètes

Olivier François et Philippe Leray

INSA Rouen - Laboratoire PSI - FRE CNRS 2645
BP 08 - Av. de l'Université, 76801 St-Etienne du Rouvray Cedex
{Olivier.Francois, Philippe.Leray}@insa-rouen.fr
http://bnt.insa-rouen.fr

Résumé. Le formalisme des modèles graphiques connait actuellement un essor dans les domaines du *machine learning*. En particulier, les réseaux bayésiens sont capables d'effectuer des raisonnements probabilistes à partir de données incomplètes alors que peu de méthodes sont actuellement capables d'utiliser les bases d'exemples incomplètes pour leur apprentissage. En s'inpirant du principe de AMS-EM proposé par (Friedman, 1997) et des travaux de(Chow & Liu, 1968), nous proposons une méthode permettant de faire l'apprentissage de réseaux bayésiens particuliers, de structure arborescente, à partir de données incomplètes. Une étude expérimentale expose ensuite des résultats préliminaires qu'il est possible d'attendre d'une telle méthode, puis montre le gain potentiel apporté lorsque nous utilisons les arbres obtenus comme initialisation d'une méthode de recherche gloutonne comme AMS-EM.

1 Introduction

La détermination d'un réseau bayésien $\mathcal{B} = (\mathcal{G}, \theta)$ nécessite la définition d'un graphe acyclique dirigé (DAG) $\mathcal{G}$ dont les sommets représentent un ensemble de variables aléatoires $X = \{X_1, \cdots, X_n\}$ (la structure), et de matrices de probabilités conditionnelles du nœud i connaissant l'état de ses parents $Pa(X_i)$ dans $\mathcal{G}$, $\theta_i = [\mathbb{P}(X_i/X_{Pa(X_i)})]$ (les paramètres).

De nombreuses méthodes d'apprentissage de structure de réseaux bayésiens ont vu le jour ces dernières années. Alors qu'il est possible de faire de l'apprentissage de paramètres de réseaux bayésiens à partir de données incomplètes et que l'inférence dans les réseaux bayésiens est possible même lorsque peu d'attributs sont observés (Jensen, 1996, Pearl, 1998, Naïm *et al.*, 2004), les algorithmes d'apprentissage de structure avec des données incomplètes restent rares.

Il est possible de différencier trois types de données manquantes selon le mécanisme qui les a générées. Le premier type représente les données manquantes au hasard (MAR, *missing at random*). Dans ce cas, la probabilité qu'une variable ne soit pas mesurée ne dépend que de l'état de certaines autres variables observées. Lorsque cette probabilité ne dépend plus des variables observées, les données manquantes sont dites MCAR (*missing completely at random*). Par contre lorsque la probabilité qu'une variable soit manquante dépend à la fois de l'état de certaines autres variables observées mais également de phénomènes extérieurs, les données sont dites NMAR.

Par la suite, nous supposerons que nous sommes en présence d'une base de données incomplètes suivant un mécanisme MAR ou MCAR. Ainsi, nous possédons toute l'information nécessaire pour estimer la distribution des données manquantes dans la base d'exemples.

Lorsque les données sont incomplètes, il est possible de déterminer les paramètres et la structure du réseau bayésien à partir des entrées complètes de la base. Comme les données manquantes sont supposées l'être aléatoirement, nous construisons ainsi un estimateur sans biais. Néanmoins, dans l'exemple d'une base de 2000 cas sur 20 attributs, avec une probabilité de 20% qu'une mesure soit manquante, nous ne disposerons en moyenne que de 23 cas complets. Les autres données à notre disposition ne sont donc pas négligeables et il serait donc préférable de faire l'apprentissage en utilisant toute l'information à laquelle nous avons accès.

Un avantage des réseaux bayésiens est qu'il suffit que seules les variables X_i et $Pa(X_i)$ soient observées pour estimer la table de probabilité conditionnelle correspondante. Dans ce cas, il est alors possible d'utiliser tous les exemples (même incomplets) où ces variables sont observées (dans