Apprentissage de structure à partir de données incomplètes et application à la classification

Olivier François et Philippe Leray

Laboratoire LITIS (Informatique, Traitement de l'Information et Systèmes)
INSA et Université de Rouen, BP08, 76801 Saint-Etienne-Du-Rouvray
Francois.Olivier.C.H@gmail.com, Philippe.Leray@insa-rouen.fr

Résumé. Les réseaux bayésiens sont des modèles probabilistes de plus en plus utilisés pour la modélisation de systèmes complexes, l'aide à la décision et la classification. De nombreuses heuristiques ont alors été proposées pour déterminer la structure d'un réseau bayésien à partir de bases d'exemples complètement observées. Néanmoins, les techniques existantes pour trouver cette structure à partir de bases incomplètes restent rares et sont de complexité élevée. En nous inspirant de l'heuristique *Structural-EM* (SEM) introduite par Friedman (1998), nous proposons une méthode *d'identification de réseau bayésien à partir de bases d'exemples incomplètes* rapide et optimale pour un critère de score dans l'espace des structures arborescentes. Nous verrons ensuite comment utiliser cette technique dans le cadre d'une problématique de classification.

Résumé long

Selon Rubin (1976), il est possible de différencier trois types de données manquantes selon le mécanisme qui les a générées. Pour les données *manquantes au hasard* (MAR) 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*). Dans ces deux cas, le mécanisme d'absence des données est dit *ignorable* car il est possible d'inférer les données manquantes à l'aide des données observées. 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 *non ignorables* (NMAR).

À structure fixée, effectuer l'apprentissage des paramètres d'un réseau bayésien est une problématique bien résolue en présence de données MCAR ou MAR en utilisant une technique de type EM (Heckerman (1998)). Friedman (1998) a été l'un des premiers à proposer une méthode déterministe efficace pour rechercher une structure à partir de données incomplètes en se basant sur les principes de l'algorithme EM. La méthode SEM est itérative, de type recherche gloutonne, et propose de choisir la nouvelle solution parmi un ensemble de voisins du graphe courant.

Dans de précédents travaux (François et Leray (2004)), nous avons montré que, dans le cas de données complètes, la méthode MWST (*maximum weight spanning tree*) de recherche

de l'*arbre de recouvrement maximal* adaptée aux modèles graphiques probabilistes (Heckerman (1998); Chow et Liu (1968)) permet d'obtenir 'très' rapidement une structure simple représentant au mieux les données et pouvant également servir d'initialisation pour une méthode de recherche gloutonne. Nous introduirons alors l'algorithme *itératif* MWST-EM Leray et François (2005), adaptant cette méthode aux bases de données incomplètes en recherchant le meilleur graphe dans l'ensemble des arbres plutôt que le meilleur dans un voisinage. L'algorithme MWST-EM, tout comme l'algorithme SEM, est une méthode itérative. Nous proposons ensuite d'initialiser SEM avec l'arbre rendu par MWST-EM pour étudier les apports éventuels d'une telle initialisation.

Nous observerons alors comment l'obtention d'un modèle de dépendances conditionnelles (réseau bayésien) sous forme d'arbre peut donner de bons résultats rapidement. Nous verrons également comment cet arbre peut contribuer à trouver un modèle plus général meilleur, voire de trouver un modèle aussi performant que celui rendu par la méthode SEM mais mis en évidence plus rapidement. à trouver un modèle plus performant que celui obtenu par SEM, ou sinon un modèle de performances équivalentes mais obtenu plus rapidement.

Nous comparerons empiriquement ces différentes approches pour des bases d'exemples jouets ainsi que sur des bases d'exemples incomplètes disponibles en ligne pour une problématique de classification.

Références

Chow, C. et C. Liu (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory 14*(3), 462–467.

François, O. et P. Leray (2004). Evaluation d'algorithmes d'apprentissage de structure dans les réseaux bayésiens. In *14ieme Congrès francophone de Reconnaissance des formes et d'Intelligence artificielle*, pp. 1453–1460.

Friedman, N. (1998). The Bayesian structural EM algorithm. In G. F. Cooper et S. Moral (Eds.), *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, San Francisco, pp. 129–138. Morgan Kaufmann.

Heckerman, D. (1998). A tutorial on learning with bayesian network. In M. I. Jordan (Ed.), *Learning in Graphical Models*. Kluwer Academic Publishers.

Leray, P. et O. François (2005). Bayesian network structural learning and incomplete data. In *in the proccedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 2005), Espoo, Finland*.

Rubin, D. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Summary

Many heuristics have been proposed to learn Bayesian Networks from complete datasets, but only few ones from incomplete datasets exist. Using the Structural-EM heuristics (SEM) introduced by Friedman (1998), we proposed a fast technic to learn Bayesian Networks from incomplete datasets which is optimal in the tree DAGs space and show how to adapt it for classification tasks.