Reconstruction exacte de réseau bayésien à partir d'observations complètes

Thématique : apprentissage, réseau bayésien, logique propositionnelle, optimisation combinatoire
Équipe d'accueil : Statistique et Algorithmique pour la Biologie
Laboratoire d'accueil : Mathématiques et Informatique Appliquées de Toulouse, Institut National de la Recherche Agronomique
Lieu : Auzeville-Tolosane (près de Toulouse), France
Encadrant : Simon de Givry (degivry@toulouse.inra.fr Tel : 05 61 28 50 74)
Gratification : environ 400 euros / mois

Contexte

Le problème étudié est celui de la reconstruction automatique de la structure d'un réseau bayésien à partir d'observations. On suppose les variables aléatoires entièrement observées pour un ensemble de réalisations et la difficulté est de retrouver la structure qui maximise la vraisemblance pénalisée des observations. Il s'agit d'un problème d'optimisation combinatoire NP-dur avec un espace de recherche constitué d'un nombre exponentiel de graphes dirigés sans circuit. Récemment d'important gains de performance ont été obtenus en limitant le nombre de parents potentiels et en exploitant la taille de l'échantillon pour *pré*-évaluer un ensemble de configurations possibles de parents pour chaque variable aléatoire qui soit de taille raisonnable. Cette technique dite de *cache* a ouvert la voie à plusieurs approches d'optimisation dont la programmation dynamique (Silander et al, 2006), la programmation linéaire en nombre entiers (Barlett et Cussens, 2013) et la logique propositionnelle (Berg et al, 2014). L'équipe d'accueil mène des travaux sur la reconstruction de réseau de régulation de gènes à l'aide de différentes techniques (Vignes et al, 2011), dont les réseaux bayésiens (Vandel et al, 2012). De plus, elle a une activité en optimisation combinatoire dans les sciences du vivant et les modèles graphiques probabilistes pour lesquels elle développe un outil ayant remporté plusieurs compétitions (UAI Evaluation 2014 `http://www.hlt.utdallas.edu/~vgogate/uai14-competition/leaders.html` Proteus&Robin utilisant toulbar2 `https://mulcyber.toulouse.inra.fr/projects/toulbar2/`).

Sujet

L'objectif du stage est dans un premier temps de faire un état des lieux des différentes approches existantes pour la reconstruction exacte de réseaux bayésiens. Une comparaison des méthodes sur des données *benchmarks* de la communauté UAI et aussi sur des données simulées et des données réelles de réseaux de gènes disponibles à l'INRA (arabette, tournesol). Le travail portera ensuite sur l'amélioration d'une méthode capable de borner la complexité du réseau appris en terme de requêtes d'inférence (Berg et al, 2014). Il s'agira de revoir la modélisation du problème éventuellement dans d'autres cadres que la logique propositionnelle et d'étudier des algorithmes efficaces pour propager la contrainte d'acyclicité.

Possibilité de poursuite en thèse (financement INRA) sur les thématiques apprentissage/optimisation/BigData avec application à des données biologiques de grande taille pour l'étude de la résistance au stress chez le tournesol.

Bibliographie

M. Barlett, J. Cussens,
Advances in Bayesian Network Learning using Integer Programming.
In Proc. of UAI, 2013

J. Berg, M. Jarvisalo, B. Malone,
Learning Optimal Bounded Treewidth Bayesian Networks via Maximum Satisfiability
In Proc. of AISTATS, 2014

T. Silander, P. Myllymäki,
A simple approach for finding the globally optimal Bayesian network structure.
In Proc. of UAI, 2006

M. Vignes, J. Vandel, D. Allouche, N. Ramadan-Alban, C. Cierco-Ayrolles, T. Schiex, B. Mangin, S. de Givry,
Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso
and their meta-analysis. PLoS ONE, 6(12), 2011

J Vandel, B Mangin, and S de Givry
New Local Move Operators for Bayesian Network Structure Learning.
In Proc. of PGM-12, Granada, Spain, 2012