

Intégrer le savoir de l'expert à l'apprentissage pour identifier le « vrai » DAG causal

V.Asvatourian^{1,2}, S.Michiels^{1,2}, E.Lanoy^{1,2}

¹Université Paris-Saclay, Univ. Paris-Sud, UVSQ, CESP, INSERM, Villejuif, France;

²Service de Biostatistique et d'Epidémiologie, Gustave-Roussy, Villejuif, France;

Thème : Approche Statistique de la causalité

La méthode IDA (intervention calculus when the DAG is absent) a été introduite pour identifier des effets causaux à partir de données multidimensionnelles. La première étape de l'IDA est d'identifier les potentiels DAG causaux qui contiennent tous les effets causaux des marqueurs entre eux et sur le critère de jugement d'intérêt. Ces DAGs causaux sont encodés dans un super graph contenant le vrai DAG : le CPDAG (Completed Partially Directed Acyclic Graph). Nous avons montré qu'il était possible de réduire le nombre de DAGs causaux potentiels dans le cas de données multidimensionnelles longitudinales en intégrant l'ordre chronologique à la première étape de l'IDA, permettant ainsi de mieux identifier les effets causaux en limitant le nombre d'associations à tester [1].

Les DAGs ou autres modèles graphiques décrivent les relations entre variables d'un domaine par des liens d'indépendances conditionnelles. Ces modèles graphiques peuvent être modélisés par des algorithmes d'apprentissage (méthodes de score, contraintes ou hybrides), ou par le savoir de l'expert si celui-ci comporte peu de variables. Récemment une méthode améliorant l'apprentissage grâce au savoir de l'expert a été développée pour les méthodes de score [2]. Pour les méthodes dites « contraintes », plus robustes en situation multidimensionnelle et utilisées dans l'IDA, cette intégration n'a pas encore été développée.

L'objectif de ce travail est donc d'intégrer le savoir de l'expert dans l'algorithme d'apprentissage dit de « contrainte » pour la recherche d'indépendances (conditionnelles) des paires de variables afin de réduire le nombre de DAGs causaux potentiels et d'avoir un CPDAG qui se rapproche du vrai DAG.

L'avis d'expert sera recueilli pour N paires de variables sous la forme d'un vecteur de probabilités $P = \{p_{\rightarrow} + p_{\leftarrow} + p_{\emptyset}\}$ avec $P = p_{\rightarrow} + p_{\leftarrow} + p_{\emptyset} = 1$ s'il existe une information a priori et $P = \emptyset$ s'il n'y a pas d'information a priori.

Une étude de simulation sera réalisée pour l'impact de l'ajout de l'information a priori sur la structure du CPDAG trouvée grâce à l'algorithme d'apprentissage.

Références

1. Asvatourian V, Coutzac C, Chaput N, Robert C, Michiels S, Lanoy E: **Estimating causal effects of repeated multi-dimensional biomarkers in an observational setting**. *Under Review*
2. Amirkhani H, Rahmati M, Lucas P, Hommersom A: **Exploiting Experts' Knowledge for Structure Learning of Bayesian Networks**. *IEEE Trans Pattern Anal Mach Intell* 2016, 8828(c):1-1.