PONDÉRATION POUR CORRECTION DE LA NON-RÉPONSE TOTALE ET MACHINE LEARNING

Brigitte GELEIN (*), David HAZIZA (**), David CAUSEUR (***)

(*) Ensai
(**) Département de mathématiques et de statistique, Université de Montréal
(***) Agrocampus Ouest

brigitte.gelein@ensai.fr

Mots-clés : apprentissage supervisé, machine learning, non-réponse totale, pondération.

Résumé

Afin de mieux comprendre la démographie, la sociologie et l'économie, les analystes et les chercheurs mettent en œuvre des méthodes statistiques pour analyser les données. Ces dernières peuvent être fournies par des recensements, des enquêtes ou encore des sources administratives. Peu importe l'origine de ces données, elles sont toutes susceptibles de présenter des données manquantes. Le traitement de la non-réponse est d'un intérêt pratique très important étant donnée la baisse constante du taux de réponse aux enquêtes depuis plusieurs décennies.

Nous considérons le problème de l'estimation des probabilités de réponse dans un contexte de pondération pour correction de la non réponse totale. Les probabilités de réponse peuvent être estimées par des méthodes paramétriques ou non paramétriques. La classe des modèles paramétriques inclut la régression logistique comme cas particulier. Les méthodes paramétriques présentent cependant plusieurs inconvénients : (i) elles ne sont pas robustes par rapport à une mauvaise spécification de la forme du modèle, (ii) elles ne sont pas non plus robustes à la non prise en compte d'éventuelles interactions entre prédicteurs ou de termes quadratiques, (iii) elles peuvent conduire à des probabilités estimées très proches de zéro, conduisant à des estimateurs potentiellement instables (Little et Vartivarian, 2005, et Beaumont 2005).

La classe des méthodes non paramétriques comprend la régression par noyaux (Giommi, 1984, Da Silva et Opsomer, 2006), la régression par polynômes locaux (Da Silva et Opsomer, 2009), la pondération de classes formées sur la base d'une estimation préliminaire des probabilités de réponse (Little, 1986, Eltinge et Yansaneh, 1997, Haziza et Beaumont, 2007), l'algorithme CHi square Automatic Interaction Detection (CHAID de Kass, 1980), Classification and Regression Trees (CART Breiman et al., 1984, Phipps et Toth, 2012), Conditional inference trees (Ctree) pour des cibles simples ou multiples (Hothorn et al. 2006).

Nous présentons une vaste étude par simulation pour comparer un grand nombre de méthodes d'estimation des probabilités de réponse par apprentissage supervisé, dans un cadre de population finie. Dans ces simulations, nous couvrons un large champ de méthodes paramétriques ou non, avec des règles de décisions simples ou agrégées telles que Bagging, Random Forests (Breiman, 1996), Boosting (Freund et Shapire, 1996, Friedman et al. 2000); voir également Hastie et al. (2009) pour une revue très complète des méthodes d'apprentissage. Pour chaque méthode, ce sont les performances de l'estimateur par expansion et de l'estimateurs de Hajek d'un total qui sont mesurées en termes de biais relatif et d'efficacité relative.