Arbres de régression et de classification (CART)

Olivier Lopez, Xavier Milhaud, Pierre-Emmanuel Thérond

Arbres de régression et de classification (CART)

Ces outils fournissent de nouvelles méthodes pour répondre aux **problématiques des assureurs**, notamment pour l'analyse des comportements des assurés et des prospects. Ils permettent de construire des classes de risques.

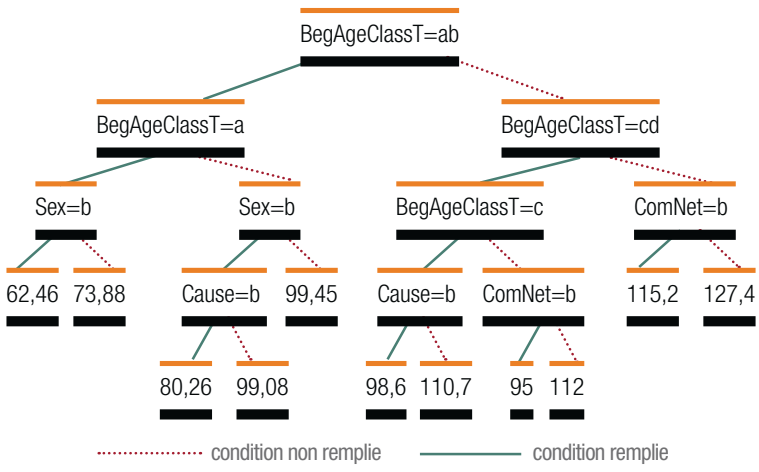
Les problématiques associées aux mégadonnées ont donné un éclairage particulier aux méthodes non paramétriques de classification et de régression. Au premier rang desquelles les arbres de régression et de classification (CART en anglais). Outre l'aspect intelligible de leur restitution (sous forme d'arbre, voir figure ci-dessous), ces techniques présentent des intérêts actuariels évidents dès lors qu'il s'agit de construire des classes de risques. Bien évidemment, la tarification est une application immédiate mais cette problématique se fait de plus en plus ressentir également en termes de suivi de portefeuille, d'identification de comportements client (déclaration de sinistre en auto, résiliation/rachat ou renouvellement d'une police d'assurance, abandon d'une souscription web d'un nouveau contrat, etc.).

Comment cela fonctionne-t-il ?

Un arbre de régression et de classification se construit de manière itérative, en découpant à chaque étape la population en deux sous-ensembles. Le découpage s'effectue suivant des règles simples portant sur les variables explicatives, en déterminant la règle optimale qui permet de construire deux populations les plus différenciées en termes de valeurs de la variable à expliquer. Les variables à expliquer[1] et explicatives peuvent être quantitatives ou pas.

Arbre de régression

La figure ci-contre présente le résultat d'un arbre de régression obtenu sur du risque de maintien en arrêt de travail à partir de données en partie censurées à droite. La variable expliquée est la durée moyenne dans l'arrêt (en jours). Les variables explicatives sont la classe d'âge à l'entrée, le sexe de l'assuré, le réseau commercial dont est issu le contrat concerné et la cause de l'arrêt.



Ces techniques nécessitent donc la définition d'un critère de discrimination[2], permettant de procéder à chaque étape au découpage, et d'un critère d'arrêt (élagage). Des algorithmes permettant leur mise en œuvre sont disponibles dans la plupart des outils statistiques (R, SAS, etc.).

Intérêts et limites

À la différence d'approches paramétriques plus classiques (ex. : modèles linéaires généralisés), ces techniques permettent de ne pas introduire de structure a priori du lien de dépendance entre variable à expliquer et covariables. Elles ne sont pas conditionnées non plus par des types ou structures de données particulières (ce qui fait leur succès dans l'exploitation de mégadonnées). Elles conduisent à des segmentations invariantes aux transformations de variables pour lesquelles on dispose d'estimateurs du taux de mauvais classement. Enfin (et surtout ?) elles produisent des résultats simples à interpréter et à utiliser en cela qu'ils sont obtenus sur la base d'une succession de problèmes appréhendables par l'esprit humain (sachant le sous-ensemble de départ, quel nouveau découpage discrimine le mieux la variable à expliquer).

Ces méthodes commencent à se diffuser dans la sphère actuarielle

Ces approches ont néanmoins des limites qu'il faut garder à l'esprit :
- la segmentation finale obtenue n'est pas nécessairement optimale au global[3] (l'optimalité n'est assurée qu'à chaque nœud de l'arbre) ;
- chaque segmentation n'est opérée que sur une des variables (pas de combinaison de variables explicatives à un nœud donné) ;
- une variable qui n'apparaît pas dans l'arbre n'est pas forcément non explicative ;
- de légères variations de données peuvent

•••/•••

conduire à des arbres sensiblement différents (une segmentation optimale à un nœud prenant le dessus sur une autre).
Néanmoins, cette instabilité peut être corrigée par la combinaison avec des méthodes d'apprentissage (forêts aléatoires), au prix d'une perte sensible d'interprétation.

Perspectives

Au-delà des aspects mégadonnées, ces méthodes commencent à se diffuser dans la sphère actuarielle notamment du fait de leur simplicité de mise en œuvre et surtout de leur caractère tout-terrain en cela qu'elles ne nécessitent pas d'a priori et sont donc particulièrement adaptées aux nouvelles problématiques des assureurs. C'est particulièrement le cas dans l'analyse des comportements des assurés ou prospects.

Olivier Lopez, ENSAE Paris-Tech, Xavier Milhaud, ENSAE Paris-Tech, Pierre Thérond, Galea & Associés, ISFA – Université Lyon-1

Notes :
1. On parle de régression lorsque la variable à expliquer est quantitative, de classification sinon.
2. Par exemple la déviance ou la décomposition de variance pour une variable explicative quantitative, l'entropie ou l'indice de Gini pour une variable qualitative.
3. Les techniques de forêts viennent compléter ces approches pour obtenir des arbres optimaux au global.

Références :

L. Breiman, J. Friedman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*, Chapman and Hall, 1984.

Lopez O., Milhaud X., Thérond P., *Consistency of tree-based estimators in censored regression with applications in insurance*, *working paper*, 2014.

Walter Olbricht, *« Tree-based methods: a useful tool for life insurance »*, *European Actuarial Journal*, 2(1): 129-47, 2012.

Imad Bou-Hamad, Denis Larocque, Hatem Ben-Ameur, *« A review of survival trees »*, *Statistics Surveys*, 5:44-71, 2011.