Co-clustering sous contraintes pour identifier les profils à risque : application à la sinistralité des flottes de véhicules d'entreprises

Romain Lorenzi, Clément Fauré, Dimitri Petroviche

GAC Technology, Lyon, France
rlorenzi, cfaure, dpetroviche@gac-technology.com

1 Contexte

Pour un gestionnaire de flotte, la compréhension du comportement de ses véhicules est le point central d'une gestion optimale. L'identification de profils de véhicules/conducteurs permet notamment de traiter cette problématique. Cette pratique, appelée profiling, consiste à trouver des individus qui partagent plusieurs caractéristiques en commun et se comportent de la même façon vis-à-vis d'une cible d'étude. Afin de tirer le maximum de cette méthode, il est possible de limiter l'extraction des profils à ceux qui sont les plus intéressants. Ces derniers seront appelés "profils extrêmes" ou "profils à risque" et correspondent à des groupes d'individus ayant des habitudes extrêmement différentes des autres.
Nous présentons ici un algorithme permettant de répondre à ces besoins, nommé EPR pour Extreme Profiles Recognition. Son objectif est d'identifier des profils de véhicules ayant une sinistralité extrême (inexistante ou très importante) tout en garantissant des résultats variés et représentatifs des données. Il s'inscrit dans la catégorie des algorithmes de co-clustering sous contraintes. Ces dernières concernent notamment la cible d'étude : seuls les profils avec une sinistralité particulièrement haute ou basse sont gardés.
Nous avons isolé quatre composantes principales dans les profils à retourner :

- Les caractéristiques qu'ont en commun tous les véhicules du profil (parmi les variables de la base de données).
- Le nombre de véhicules composant le profil.
- La valeur de la cible d'étude associée au profil.
- La pureté : parmi les véhicules appartenant au profil, le pourcentage de ceux ayant la même valeur de la cible d'étude que celle du profil.

L'objectif est d'identifier des profils dont la pureté est proche des 100% tout en garantissant un nombre de véhicules suffisamment grand et un nombre de caractéristiques en commun raisonnable pour que l'information rendue soit pertinente et facilement compréhensible.

2 Extreme Profiles Recognition

Notre étude s'appuie sur une base de données orientée sinistralité. La cible d'étude est un score de sinistralité calculé à partir de plusieurs critères (nombre d'accidents corporels et maté-

riels, responsabilité, ...) et le but consiste à extraire des profils possédant un score de sinistralité très bas ou très haut. Pour cela, nous nous sommes inspirés du fonctionnement des arbres de décision tout en essayant de compléter les résultats qu'ils proposent. La correspondance entre un arbre et les profils se fait directement : les variables de la base de données rempliraient les noeuds de l'arbre tandis que les profils correspondraient aux feuilles. En revanche, n'utiliser qu'un seul arbre limiterait fortement la diversité des résultats : un profil ne comprenant pas la variable à la racine de l'arbre ne pourrait ressortir. Ainsi, afin d'élargir le spectre des résultats, l'EPR va créer autant d'arbres qu'il y a de variables dans la base, en forçant chacune des variables à être placée à la racine d'un arbre.

Construction et regroupement des profils La construction des profils se fait alors par niveaux. On appellera un profil de niveau n un profil où les individus le composant possèdent n variables en commun. L'EPR commence à créer des profils de niveau 2 en combinant deux à deux toutes les variables de la base de données. Ensuite, il construit le niveau 3 en ajoutant une nouvelle variable aux combinaisons de niveau 2 et ainsi de suite jusqu'à un niveau n_{max} permettant donc de créer tous les profils possibles. En revanche, le nombre de profils générés est bien trop important et la plupart ne sont pas intéressants. C'est pourquoi des contraintes concernant la pureté et la taille sont rajoutées. Seuls les profils suffisamment purs et ayant une taille suffisamment grande sont retournés par l'algorithme.
Si les contraintes permettent de ne retenir que les profils les plus intéressants, un important travail reste à faire pour avoir une visualisation simple des résultats. C'est là qu'interviennent les groupes. L'idée est de regrouper les profils jugés similaires (au sens des caractéristiques qui les composent) pour faciliter la lecture et l'analyse des résultats.

Performance et résultats L'EPR a pour vocation une utilisation en temps réel : le temps de calcul est donc un point essentiel de son développement. Dans ce sens, nous avons rendu l'algorithme facilement parallélisable. Sur chaque niveau, les profils sont calculés indépendamment les uns des autres. Il est donc possible de paralléliser ces calculs en séparant la base de données entre les différents processus. En revanche, il n'est pas possible de paralléliser les niveaux entre eux car la génération du niveau $n + 1$ dépend des profils retournés au niveau n. Pour repérer des profils de niveau 8 d'une base comportant 25000 lignes et 40 colonnes, l'algorithme tourne en environ 35 secondes sur une machine à 4 coeurs.
Afin de mesurer la précision de l'EPR, il peut être intéressant d'étudier la précision d'algorithmes de Machine Learning pour mieux comprendre la qualité de la base de données et la difficulté d'en sortir des résultats cohérents. En effet, si ces algorithmes parviennent à prédire à 90% des cas la bonne classe de sinistralité, alors trouver des profils purs à 75% n'a pas réellement d'intérêt. En revanche, si la précision se situe vers 50%, l'EPR prend tout son sens. Après un paramétrage optimal de l'algorithme, nous sommes parvenus à extraire des profils purs à 81% quand les algorithmes de classification de Machine Learning (Gradient Boosting et Random Forest) ne prédisaient la bonne classe qu'avec 62% de précision.

L'EPR apporte donc une réelle plus-value et permet l'extraction d'informations difficilement identifiables autrement. De plus, il offre à l'utilisateur des résultats complets et non centrés sur une seule partie de la base de données. Les profils extrêmes sont ainsi mis en lumière et rendent possible une analyse plus fine du thème abordé.