boosting, des méthodes d'arbres de décision ou bien être intégrés au sein d'un réseau de neurones. Dans ce contexte, disposer de techniques accélérant les temps de calcul par le biais de formules fermées peut être d'un intérêt certain.

Des arbres GLM plus rapides à construire et interprétables

Les arbres de décision sont une technique statistique consistant à réaliser un découpage récursif des données selon des règles afin que les sous-groupes créés respectent un critère d'homogénéité que l'ensemble des données ne respectent pas. Les règles sont définies selon une structure conditionnelle basée sur les variables de partitionnement avec un ordre précis : l'ordre est structuré selon un arbre inversé où la racine (en haut) contient l'intégralité des observations et les différentes « branches » sont composées d'une sous-population de plus en plus petite. Loh (2014) a réalisé l'état de l'art de cinquante ans de ces techniques. La méthode la plus utilisée est l'algorithme CART *(Classification And Regression Trees)*, mais il en existe plusieurs, dont les arbres faisant appel à des modèles GLM décrits ici.

L'algorithme CART peut souffrir d'un biais dans la sélection de variables optimales pour partitionner les données. De nombreux autres algorithmes ont été proposés tels que FACT, QUEST ou encore CTREE. Zeileis, Hothorn et Hornik (2008) ont proposé une méthode appelée MOB *(Model-Based trees)* où la sélection des variables de partitionnement est réalisée selon un test statistique de fluctuation. Cet usage de tests statistiques permet à MOB d'être beaucoup plus robuste que CART dans la construction des arbres sans pour autant changer d'approche comme pour les méthodes ensemblistes *(bootstrap aggregating* ou *boosting).* Au-delà de la robustesse, à chaque nœud terminal de l'arbre, MOB incorpore un modèle paramétrique ou semi-paramétrique, ce qui permet de disposer, non pas d'une prédiction moyenne par nœud, mais bien d'un sous-modèle local dont les paramètres sont spécifiques au nœud considéré.

Dans ce contexte, Dutang et Guibert (2022) utilisent l'algorithme MOB pour les GLM nommé par la suite arbre GLM. Ainsi, un large panel de lois de probabilité pour la variable réponse est possible. Cependant, contrairement à Zeileis, Hothorn et Hornik (2008), nous proposons d'utiliser les estimateurs explicites de la section précédente afin de bénéfier d'une formule fermée pour la log-vraisemblance maximale dans la recherche du découpage optimal de la base selon une variable de partitionnement donnée, qu'elles soient d'ailleurs continues ou catégorielles. Inutile donc d'utiliser l'algorithme IWLS pour déterminer la log-vraisemblance maximale à chaque itération de la procédure de découpage. Cette technique s'adapte aussi à des arbres non binaires et peut-être utilisée dans le cas où le GLM comprend des variables explicatives catégorielles. Nous mettons en avant sur des données simulées et des données réelles l'avantage significatif apporté par cette approche en temps de calcul pour des échantillons de taille supérieure à 1 000.

Dutang et Guibert (2022) s'intéressent aussi aux capacités de prédiction de ces techniques. Une comparaison des arbres GLM, des CART, des CTREE est réalisée, soulignant l'utilité des arbres GLM : les arbres sont peu profonds, et donc plus interprétables, pour une prédiction équivalente. Sur des données réelles de grande dimension, elles offrent des perspectives intéressantes dans un contexte de tarification en assurance non-vie.

Une ouverture vers des forêts de GLM

Enfin, nous proposons dans ces travaux une nouvelle méthode ensembliste fondée sur les arbres GLM, appelée forêt GLM. Elle consiste à reprendre l'algorithme des forêts aléatoires en remplaçant la brique de base (CART) par un arbre GLM. Ceci est rendu possible par le temps raisonnable avec lequel les arbres GLM peuvent être calibrés. La forêt GLM consiste à créer de manière indépendante un ensemble d'arbres GLM établis avec des variables de partitionnement et avec des données tirées aléatoirement. Dutang et Guibert (2022) présentent des situations où les forêts GLM peuvent être plus performantes que les forêts aléatoires usuelles et les CFOREST (construites à partir de l'algorithme CTREE). En permettant de choisir une loi, elle offre néanmoins un degré de liberté supplémentaire qu'il pourrait être intéressant d'exploiter. Cependant, ces méthodes ensemblistes perdent un avantage majeur : celui de l'interprétabilité des arbres de décisions.

Sources : Brouste A., Dutang C. et Rohmer T. (2020), « Closed-form maximum likelihood estimator for generalized linear models in the case of categorical explanatory variables: application to insurance loss modeling », *Computational Statistics*, vol. 35, p. 689-724, https://doi.org/10.1007/s00180-019-00918-7 • Dutang C. et Guibert Q. (2022), « An explicit split point procedure in model-based trees allowing for a quick fitting of GLM trees and GLM forests », *Statistics and Computing*, vol. 32, nº 6, https://doi.org/10.1007/s11222-021-10059-x • Loh W.-Y. (2014), « Fifty years of classification and regression trees », *International Statistical Review*, vol. 82, nº 3, p. 329-348 • OCDE (2020), « The impact of big data and artificial intelligence (AI) in the insurance sector - OECD », https://www.oecd.org/finance/Impact-Big-Data-AI-in-the-Insurance-Sector.htm • Zeileis A., Hothorn T. et Hornik K. (2008), « Model-based recursive partitioning », *Journal of Computational and Graphical Statistics*, vol. 17, nº 2, p. 492-514, https://doi.org/10.1198/106186008X319331.