Concilier l'équité statistique et la précision en apprentissage machine interprétable grâce à la PLNE

Julien Ferry, Ulrich Aïvodji, Sebastien Gambs, Marie-José Huguet, Mohamed Siala

Concilier l'équité statistique et la précision en apprentissage machine interprétable grâce à la PLNE

Julien Ferry[1], Ulrich Aïvodji[2], Sébastien Gambs[3] , Marie-José Huguet[1], Mohamed Siala[1]

[1] LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France
{jferry, huguet, msiala}@laas.fr
[2] École de Technologie Supérieure, Montréal, Canada
ulrich.aivodji@etsmtl.ca
[3] UQAM, Montréal, Canada
gambs.sebastien@uqam.ca

Mots-clés : *apprentissage, équité, interprétabilité, recherche opérationnelle, optimisation.*

1 Introduction et description du problème

L'interprétabilité est une propriété de plus en plus recherchée en apprentissage machine [3]. En effet, la construction de modèles interprétables, c'est à dire compréhensibles par un humain, présente d'importants enjeux éthiques, légaux, et sociétaux. Elle est souvent un pré-requis à l'adoption de modèles d'apprentissage, et facilite l'extraction de connaissances et le débogage. L'équité est également un domaine de recherche très actif. En effet, les données à partir desquelles un modèle est construit peuvent comporter des biais indésirables. Différentes notions d'équité ont été proposées pour permettre l'apprentissage de modèles qui ne reproduisent pas ces biais. Ainsi, les métriques d'équité statistique visent à égaliser différentes grandeurs exprimées en fonction de la matrice de confusion d'un classifieur, entre différents groupes protégés. Adresser simultanément ces deux problématiques semble donc crucial, mais présente des difficultés. En effet, l'apprentissage de modèles interprétables optimaux et soumis à des contraintes (par exemple, d'équité statistique) a été identifié comme l'un des grands challenges techniques dans le domaine de l'interprétabilité [3].

Les listes de règles (*rule lists*) de taille raisonnable sont communément admises comme un type de modèle interprétable. `FairCORELS` [1] est un algorithme d'apprentissage produisant des *rule lists* optimales et respectant certaines contraintes d'équité statistique. `FairCORELS` est basé sur `CORELS` [2], un algorithme de branch-and-bound permettant la construction de *rule lists* optimales (la fonction objectif considérée est la somme pondérée de la précision et de la longueur d'une *rule list*). `CORELS` utilise plusieurs bornes et différentes structures de données intelligentes pour explorer efficacement l'espace de recherche. Toutefois, en raison des contraintes d'équité imposées dans `FairCORELS`, certaines structures de données de `CORELS` ne peuvent plus être utilisées, et ses bornes originales présentent une efficacité amoindrie.

2 Contribution et résultats

Nous proposons une approche basée sur la Programmation Linéaire en Nombres Entiers (PLNE), visant à élaguer efficacement l'espace de recherche de `FairCORELS`. Nous étudions également une variante de cette approche, dans laquelle le modèle de PLNE ne se contente pas d'élaguer l'espace de recherche, mais guide son exploration. Tout comme `CORELS`, `FairCORELS` représente son espace de recherche (l'ensemble des *rule lists*) sous la forme d'un arbre des préfixes, dans lequel chaque chemin depuis la racine vers un noeud représente un préfixe (ensemble ordonné de règles). Chaque préfixe p peut lui-même être étendu pour former d'autres préfixes,

plus longs (qui seront ses noeuds enfants dans l'arbre). A chaque noeud de l'arbre des préfixes, nous résolvons un problème de satisfiabilité (encodé en PLNE) afin de savoir s'il est possible qu'une extension de ce préfixe améliore la valeur courante de la fonction objectif, tout en respectant la contrainte d'équité. Si le problème est insatisfiable, alors le sous-arbre issu de ce préfixe peut être élagué sans compromettre l'optimalité de la résolution. Si ce problème est résolu au moment d'extraire un noeud de la frontière d'exploration, on parle de `lazy pruning`. S'il est résolu avant d'insérer un noeud dans la frontière d'exploration, on parle d'`eager pruning`. Ce second cas implique un nombre plus important d'appels au solveur, mais limite au maximum la taille de la frontière d'exploration. Enfin, en résolvant un problème d'optimisation (également basé sur la PLNE), il est aussi possible de mesurer la meilleure valeur de fonction objectif qu'une *rule list* basée sur p peut atteindre tout en respectant la contrainte d'équité. On peut ainsi élaguer l'espace de recherche (de la même manière qu'avec l'`eager pruning`), mais également guider l'exploration de l'arbre, en ordonnant la frontière d'exploration en fonction de la valeur retournée par le problème de PLNE (`MILP Best First Search`).

Notre implémentation utilise le solveur IBM ILOG CPLEX Optimizer, associé à une memoisation efficace tirant profit des nombreuses symétries de l'arbre. Une évaluation expérimentale menée pour quatre métriques d'équité statistique (*Statistical Parity*, *Predictive Equality*, *Equal Opportunity*, et *Equalized Odds*) et deux jeux de données historiquement biaisés de la littérature (*German Credit* et *COMPAS*) a confirmé l'intérêt de la méthode. La Figure 1 illustre les tendances constatées : nos trois approches présentent de meilleures performances que le `BFS original` de `FairCORELS` [1]. Elles permettent en effet d'atteindre de meilleures solutions, et d'en prouver l'optimalité. Le `lazy pruning`, effectuant moins d'appels au solveur, met à jour la fonction objectif plus rapidement au début de l'exploration. Toutefois, il est globalement moins rapide à atteindre les meilleures solutions et à en prouver l'optimalité. Enfin, on note que l'utilisation de notre méthode pour élaguer l'espace de recherche mais aussi guider son exploration (`MILP Best First Search`) accélère la convergence et la preuve d'optimalité.

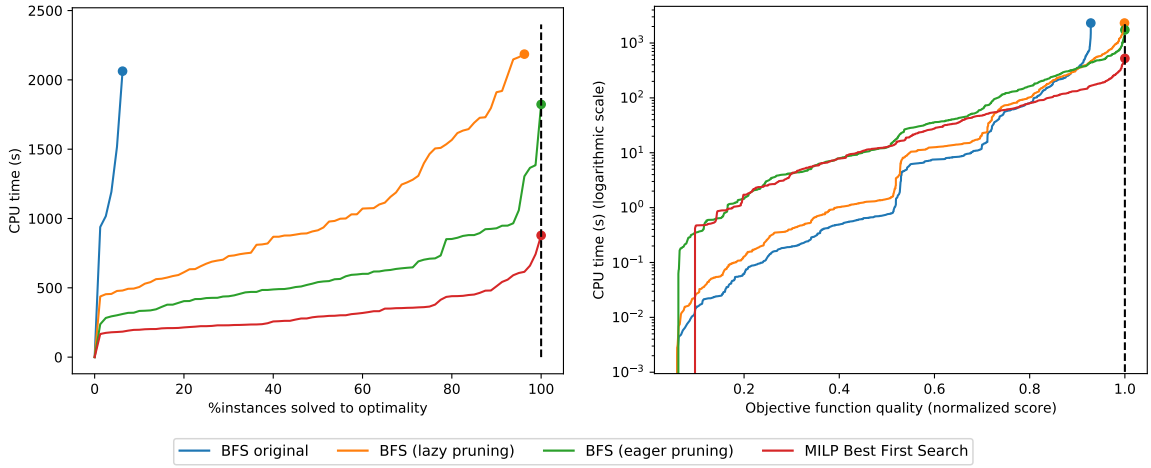

FIG. 1 – Résultats obtenus sur le jeu de données *German Credit*, pour la métrique *Equalized Odds*, pour différentes valeurs d'équité (comprises entre 0.98 et 0.995) et 20 ensembles d'entraînements.

Références

[1] Ulrich Aïvodji, Julien Ferry, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Learning fair rule lists. *arXiv preprint arXiv :1909.03977*, 2019.

[2] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1) :8753–8830, 2017.

[3] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning : Fundamental principles and 10 grand challenges. *arXiv preprint arXiv :2103.11251*, 2021.