Optimisation Distributionnellement Robuste pour améliorer la généralisation de l'équité en apprentissage

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

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

Mots-clés : *Apprentissage supervisé, Equité, Généralisation, Optimisation Distributionnellement Robuste*

1 Introduction et description du problème

Face au phénomène du biais en apprentissage, différentes notions d'équité ont été proposées, parmi lesquelles les métriques d'équité statistique ont été largement étudiées. Elles reposent sur un principe commun : égaliser la valeur d'une certaine mesure (par exemple le taux de prédictions positives) entre des groupes d'instances différant par la valeur d'un ou plusieurs attribut(s) sensible(s). En effet, ces attributs ne devraient pas (pour des raisons éthiques et légales) influer sur la décision. Pour aborder ce problème bi-objectif, plusieurs méthodes ont été proposées pour obtenir des modèles d'apprentissage respectant ces contraintes d'équité sur leur ensemble d'entraînement. Cependant, la généralisation de l'équité de ces modèles sur de nouvelles données n'est souvent pas au rendez-vous.

Face à ce problème, plusieurs approches formulent le problème d'apprentissage sous contraintes (d'équité) comme un jeu à deux joueurs, où l'un optimise les paramètres du modèle pour une fonction objectif f_{obj} tandis que l'autre cherche à approximer la relaxation Lagrangienne la plus difficile en agissant sur les coefficients de f_{obj} [2, 3]. Dans [2], le second joueur mesure la violation des contraintes d'équité sur un ensemble de validation séparé, ce qui permet d'éviter l'*overfitting* de ces contraintes. Dans [3], le second joueur mesure la violation de l'équité "pire cas" en pondérant les instances d'entraînement. Cette dernière approche est inspirée de l'Optimisation Distributionnellement Robuste, qui consiste à optimiser une fonction f sur le "pire cas", parmi un ensemble de perturbations d'un ensemble $\mathcal{D}$ plutôt que sur $\mathcal{D}$ lui-même. Dans [4], un modèle est construit en minimisant l'erreur maximale sur un ensemble de groupes définis par la valeur d'attributs biaisés.

2 Contribution, résultats et perspectives

Nous proposons une méthode inspirée de l'Optimisation Distributionnellement Robuste (ODR), pouvant être intégrée dans des algorithmes d'apprentissage équitables, et visant à générer des modèles dont l'équité généralise mieux sur de nouvelles données. Notre intuition est qu'assurer l'équité sur plusieurs sous-ensembles de l'ensemble d'entraînement peut mener à la construction d'un modèle plus robuste. Chaque sous-ensemble aléatoire de taille suffisamment importante présente une distribution voisine de celle de l'ensemble global. Pour cette raison, notre méthode est directement inspirée de l'ODR, qui vise à optimiser une métrique (ici l'équité) sur un ensemble donné (ici l'ensemble d'entraînement $\mathcal{D}$), mais également sur un ensemble de distributions voisines (ici les sous-ensembles de $\mathcal{D}$). Formellement, notre méthode consiste à générer des modèles en optimisant une fonction objectif donnée sur un ensemble

d'entraînement $\mathcal{D}_T$, tout en respectant une contrainte d'équité sur $\mathcal{D}_T$, mais également sur n sous-ensembles aléatoires de $\mathcal{D}_T$, définis par n masques binaires.

Afin d'évaluer notre méthode, nous l'avons intégrée dans un algorithme d'apprentissage équitable de la littérature : `FairCORELS` [1]. Nous calculons alors un ensemble de solutions non-dominées (en faisant varier la contrainte d'équité) pour trois valeurs de n (nombre de masques) : 0 (cas où notre méthode n'est pas utilisée), 10 et 30. Cette évaluation a été réalisée pour six métriques statistiques d'équité (*Statistical Parity*, *Predictive Parity*, *Predictive Equality*, *Equal Opportunity*, *Equalized Odds* et *Conditional Use Accuracy Equality*), sur quatre ensembles de données historiquement biaisés (*Adult Income*, *COMPAS*, *Default of Credit Card* et *Marketing*). Tous les résultats obtenus confirment l'intérêt de l'approche proposée. La Figure 1 illustre les tendances constatées. On observe que notre technique permet d'améliorer la généralisation de l'équité au prix d'une dégradation des performances en entraînement. La meilleure généralisation de l'équité permet ainsi l'obtention de fronts de Pareto précision/équité en test plus garnis et présentant de meilleurs compromis (notamment pour de fortes contraintes d'équité).

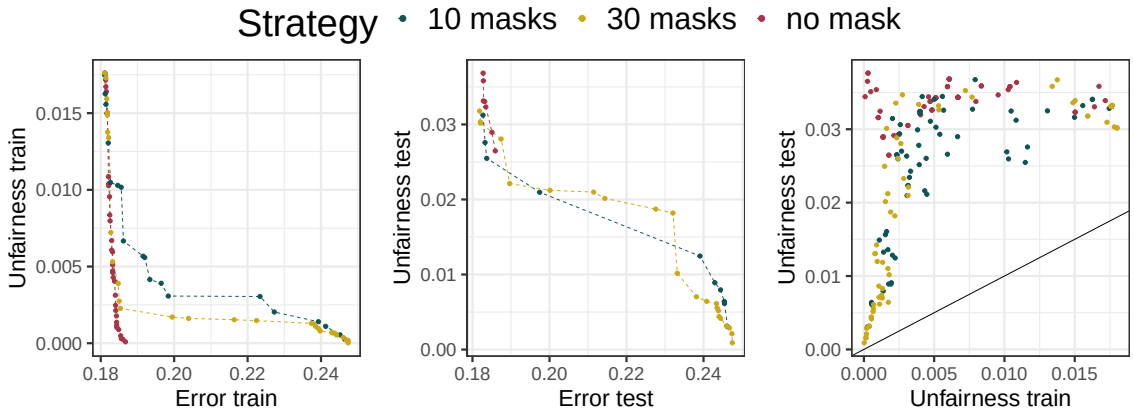

FIG. 1 – Résultats obtenus sur le jeu de données Adult Income pour l'Equal Opportunity.

Ainsi, une évaluation expérimentale extensive de notre méthode en démontre l'intérêt, bien qu'elle ne s'accompagne pas de garanties théoriques. Les perspectives futures portent sur l'étude de l'impact du nombre de masques sur la généralisation ainsi que sur l'extension de la méthode à d'autres modèles d'apprentissage équitable de la littérature.

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] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pages 1397–1405. PMLR, 2019.

[3] Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. Ensuring fairness beyond the training data. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts : On the importance of regularization for worst-case generalization. *arXiv preprint arXiv :1911.08731*, 2019.