Améliorer la généralisation de l'équité en apprentissage grâce à l'Optimisation Distributionnellement Robuste

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

Améliorer la généralisation de l'équité en apprentissage grâce à l'Optimisation Distributionnellement Robuste

J. Ferry[1], U. Aïvodji[2], S. Gambs[2], M-J. Huguet[1], M.Siala[1]

[1] LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France
[2] UQAM, Montréal, Canada

jferry@laas.fr

Résumé

Pour répondre aux enjeux de biais non-désirés en apprentissage machine, de nombreux travaux ont proposé des techniques d'amélioration de l'équité se basant sur des métriques statistiques. Cependant, l'expérience montre que la généralisation sur de nouvelles données n'est pas toujours au rendez-vous. Pour répondre à ce problème, nous proposons une technique d'Optimisation Distributionnellement Robuste permettant de générer des modèles d'apprentissage dont l'équité généralise mieux sur des nouvelles données. L'évaluation expérimentale de cette technique démontre son efficacité.

Mots-clés

Apprentissage supervisé, Equité, Généralisation, Optimisation Distributionnellement Robuste.

Abstract

Recent work have proposed fairness enhancement methods based on statistical metrics to address the problem of negative bias in machine learning. However, it has been shown that these methods do not always generalize well to new data. To address this issue, we propose a Distributionally Robust Optimization technique allowing for the generation of learning models whose fairness generalizes better to new data. Empirical evaluation demonstrates the effectiveness of the approach.

Keywords

Supervised Learning, Fairness, Generalization, Distributionally Robust Optimization

1 Introduction

Le déploiement croissant de systèmes d'intelligence artificielle soulève l'enjeu, de plus en plus étudié, de l'équité de ces systèmes. En particulier, l'objectif est souvent d'assurer que des systèmes, entraînés sur des données historiquement biaisées, ne reproduisent pas ces biais, parfois implicites. Considérant que certains attributs, dits sensibles, ne devraient pas (pour des raisons éthiques et/ou légales) influer sur la décision, différentes métriques d'équité statistique ont été proposées. Elles reposent généralement sur le même principe : égaliser la valeur d'une certaine mesure (par exemple le taux de prédictions positives) entre des groupes différant par la valeur d'un ou plusieurs attribut(s) sensible(s). Plusieurs méthodes ont été proposées pour apprendre des modèles respectant ces contraintes d'équité sur leur ensemble d'entraînement. Ces méthodes peuvent être regroupées en trois grandes familles d'approches. D'une part, les approches de *preprocessing* [8] ou de *postprocessing* [6] interviennent respectivement avant (pour éliminer le biais des données d'entraînement) et après (en modifiant les prédictions pour les rendre équitables) la phase d'apprentissage qui conduit à la construction du modèle. Enfin, les approches *inprocessing* [11] consistent à modifier directement l'algorithme d'apprentissage afin de construire des modèles équitables à partir de données possiblement biaisées. La méthode que nous proposons s'applique à cette dernière famille de techniques. Cependant, quelle que soit l'approche choisie, la généralisation de l'équité de ces modèles sur de nouvelles données n'est souvent pas au rendez-vous.

Pour améliorer cette généralisation, [7] propose l'ajout d'un terme de régularisation (mesurant la stabilité) à la fonction objectif d'un problème d'apprentissage équitable. En s'assurant que les prédictions du modèle ne varient pas trop lorsque l'ensemble d'entraînement est perturbé, cette méthode permet de borner théoriquement l'erreur de généralisation.

Plusieurs approches améliorant la généralisation de métriques statistiques d'équité [3, 4, 9] sont basées sur la méthode de [1], qui formule le problème d'apprentissage équitable comme un jeu à deux joueurs. L'un des joueurs optimise la fonction objectif (incluant un terme d'équité) et l'autre cherche à approximer la relaxation Lagrangienne la plus difficile en modifiant les coefficients Lagrangiens associés à la violation de l'équité. Dans [3, 4], le second joueur met à jour les coefficients Lagrangiens associés aux contraintes d'équité en mesurant la violation de l'équité sur un ensemble de validation séparé (plutôt que sur l'ensemble d'entraînement lui-même), ce qui permet d'améliorer la généralisation de ces contraintes.

Enfin, plutôt que d'optimiser une fonction objectif f sur un ensemble $\mathcal{D}$, l'Optimisation Distributionnellement Robuste (ODR) consiste à optimiser f sur le "pire cas", parmi un ensemble de perturbations de $\mathcal{D}$ [10] (et donc d'optimiser f

pour un ensemble de distributions voisines de $\mathcal{D}$). D'autres travaux récents utilisent le principe de l'ODR pour améliorer la généralisation de l'équité [9, 10]. Dans [10], un modèle est construit en minimisant l'erreur maximale sur un ensemble de groupes définis par la valeur d'attributs biaisés. L'approche d'ODR de [9] reprend la formulation de [1]. Le second joueur mesure alors la violation de l'équité "pire cas" en pondérant les instances d'entraînement (l'équité est ainsi optimisée pour un ensemble de pondérations des instances de l'ensemble d'entraînement).

2 Problématique

Nous considérons un algorithme d'apprentissage supervisé équitable de la littérature, produisant des modèles interprétables de type *rule list* : `FairCORELS` [2] [1]. Cet algorithme se base sur une approche de type branch and bound explorant l'ensemble $\mathcal{R}$ des *rule lists* à l'aide d'un arbre des préfixes. Dans cet arbre, chaque noeud correspond à une règle et chaque chemin (depuis la racine) est une solution possible. Ainsi, une étape préalable à l'utilisation de `FairCORELS` est le minage de règles. Ces dernières peuvent être n'importe quelle combinaison des attributs du jeu de données et doivent prendre une valeur binaire pour toutes les instances du jeu de données d'entraînement. L'objectif de `FairCORELS` est de déterminer la solution minimisant la fonction objectif f_{obj} (somme pondérée de l'erreur de classification et du nombre de règles) tout en respectant une contrainte d'équité ϵ donnée.
Soit $\mathcal{D} = (X, Y, A)$ l'ensemble de données d'entraînement, où X est l'ensemble des attributs non sensibles, Y l'ensemble des étiquettes et A l'attribut (ou l'ensemble des attributs) sensible(s). La fonction `misc(.)` mesure l'erreur de classification alors que `unf(.)` quantifie la violation de l'équité (selon la métrique choisie). La *rule list* recherchée r^* est la solution du problème suivant, où K_r désigne la longueur de r, et λ est un coefficient de régularisation :

$$\begin{aligned} \underset{r \in \mathcal{R}}{\arg\min} \quad & f_{obj} = \texttt{misc}(r, X, Y) + \lambda . K_r \\ \text{s.t.} \quad & \texttt{unf}(r, X, Y, A) \leq \epsilon, \end{aligned}$$

A chaque itération, un préfixe r est évalué. S'il améliore f_{obj} et respecte la contrainte d'équité (sur l'ensemble d'entraînement), la meilleure solution courante est mise à jour. Plusieurs heuristiques peuvent être utilisées pour guider l'exploration, dont l'efficacité est améliorée par l'existence de différentes bornes. `FairCORELS` retourne théoriquement la solution optimale, c'est-à-dire la *rule list* pour laquelle f_{obj} est minimisée, et qui respecte la contrainte d'équité. Toutefois, la taille de l'arbre des préfixes augmentant exponentiellement avec le nombre de règles, un paramètre n_{iter} définit le nombre maximal de noeuds à explorer dans l'arbre des préfixes. Ce paramètre limite l'espace mémoire utilisé par le programme, permettant ainsi d'obtenir de bonnes solutions dans un délai maîtrisé.

1. https ://github.com/ferryjul/fairCORELS

Il est possible d'utiliser `FairCORELS` pour générer un ensemble de solutions, en réalisant des appels successifs pour différentes valeurs d'ϵ. On obtient ainsi un ensemble de compromis précision/équité, qu'on peut représenter par un front de Pareto sur l'ensemble d'entraînement. Cependant, ces solutions définissent souvent un ensemble de compromis moins intéressants sur leur ensemble de test, en raison d'une mauvaise généralisation de l'équité, notamment lorsque les contraintes sont fortes (ϵ faible). Plusieurs travaux récents ont ainsi proposé des méthodes pour palier à ce problème [3, 4, 7, 9, 10].

3 Méthode proposée

Notre objectif est d'obtenir des modèles dont l'équité généralise bien sur l'ensemble de test (c'est-à-dire de nouvelles données non observées pendant l'entraînement de ces modèles). A l'inverse de la précision, l'équité statistique se mesure uniquement sur des ensembles d'instances (et pas sur des instances seules). En outre, il est possible qu'un modèle paraisse équitable sur un ensemble donné, tout en prenant des décisions non équitables localement [5]. Notre intuition est qu'assurer l'équité sur plusieurs sous-ensembles de l'ensemble d'entraînement peut forcer le modèle construit à une meilleure organisation de ses décisions, et par conséquence à une meilleure généralisation de l'équité. 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'Optimisation Distributionnellement Robuste, 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 $\mathcal{D}$).
Dans le contexte de `FairCORELS`, l'approche proposée consiste à s'assurer que la *rule list* générée respecte la contrainte d'équité sur l'ensemble d'entraînement, et sur un certain nombre de sous-ensembles aléatoires de celui-ci, approximant les "distributions voisines" du cadre de l'ODR. Pour cela, n masques binaires aléatoires sont générés et utilisés afin de définir n sous-ensembles de l'ensemble d'entraînement. Lorsque la *rule list* évaluée r permet d'améliorer la fonction objectif, on ne met à jour la meilleure solution courante que si r respecte la contrainte d'équité sur l'ensemble d'entraînement et sur chacun des sous-ensembles définis par les n masques. La *rule list* recherchée r^* est donc la solution du problème suivant, où V_i est le sous-ensemble de V défini par le masque i, pour $V \in \{X, Y, A\}$:

$$\begin{aligned} \underset{r \in \mathcal{R}}{\arg\min} \quad & f_{obj} = \texttt{misc}(r, X, Y) + \lambda . K_r \\ \text{s.t.} \quad & \texttt{unf}(r, X, Y, A) \leq \epsilon, \\ & \underset{\forall i \in [n]}{\texttt{unf}(r, X_i, Y_i, A_i) \leq \epsilon} \end{aligned}$$

4 Evaluation expérimentale

Afin d'évaluer notre méthode, nous calculons un ensemble de solutions non-dominées (en faisant varier la contrainte

Jeu de données	Nombre d'instances	Attribut sensible	Nombre de règles minées	Prédiction
Adult Income	33 917	Homme/Femme	183	Salaire : haut/bas
COMPAS	5 273	African-American/Caucasian	165	Récidive : oui/non
Default of Credit Card	29 986	Homme/Femme	189	Paiement refusé le mois prochain : oui/non
Marketing	41 175	Age : entre 30 et 60 ans ou pas	179	Souscription : oui/non

TABLE 1 – Caractéristiques des différents jeux de données (après notre pré-traitement) utilisés pour évaluer notre méthode

d'équité) pour trois valeurs 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 (*Adult Income*, *COMPAS*, *Default of Credit Card* et *Marketing*). Les caractéristiques (taille, attribut sensible, ...) de ces ensembles de données sont résumés dans le tableau 1.

Les quatre ensembles de données regroupent des données historiquement biaisées (notamment par rapport à l'attribut sensible considéré), et peuvent donc être utilisés pour générer des ensembles de compromis entre la précision et l'équité. Afin de pouvoir utiliser `FairCORELS` sur ces ensembles de données, nous les avons binarisés, avant de calculer, pour chacun d'entre eux, un ensemble de règles. Les règles minées sont des conjonctions d'au plus deux attributs (ou leur négation) capturant plus de $min_{support}\%$ des instances de l'ensemble de données, où $min_{support}$ est choisi de sorte à limiter le nombre de règles minées. Pour toutes nos expérimentations, n_{iter} est fixé à 25.10^5, $\lambda = 10^{-3}$ et l'heuristique d'exploration utilisée est le `BFS obj.-aware` (une recherche en largeur priorisant les solutions présentant une meilleure valeur de f_{obj} parmi celles de même profondeur). Les valeurs reportées (précision et équité) sont obtenues au moyen d'une *validation croisée 5-folds*. Tous les résultats obtenus confortent les observations présentées ci-après sur l'ensemble de données *Default of Credit Card* et pour la métrique d'équité *Predictive Equality*.

La figure 1 présente la généralisation des solutions construites, où les points les plus proches de la diagonale correspondent aux solutions dont l'équité généralise le mieux. Elle illustre l'amélioration apportée pour la généralisation de l'équité, en particulier pour des contraintes fortes d'équité. En effet, on peut voir que lorsque la violation de l'équité sur l'ensemble d'entraînement est faible, les solutions obtenues en utilisant la méthode basée sur l'ODR (`10masks` et `30masks`) présentent une violation de l'équité en test plus faible que celles obtenues sans ODR (`0masks`). La figure 2 suggère qu'un compromis entre performance sur l'ensemble d'entraînement (équité et précision) et généralisation doit être fait. En effet, la méthode basée sur l'ODR permet de générer des solutions moins performantes sur l'ensemble d'entraînement, mais plus robustes (généralisant mieux). La figure 3 démontre qu'une conséquence de l'amélioration de la généralisation

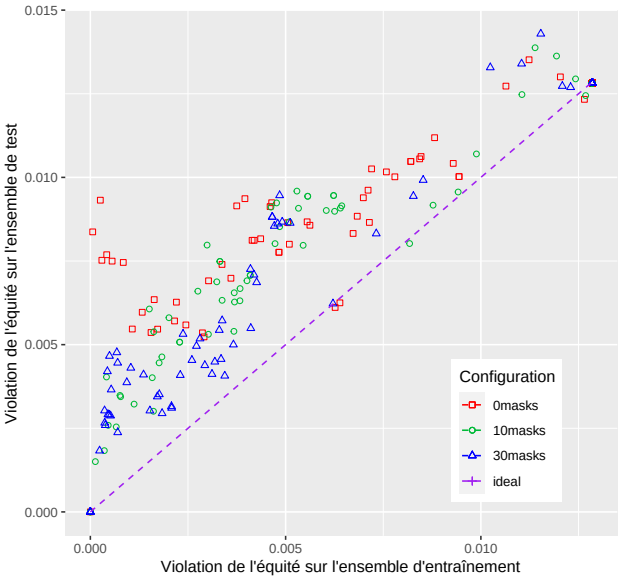

FIGURE 1 – Equité sur l'ensemble de test en fonction de l'équité sur l'ensemble d'entraînement pour la *Predictive Equality* sur l'ensemble de données *Default of Credit Card*

de l'équité est la production d'un front de Pareto sur l'ensemble de test plus garni. L'utilisation de notre méthode permet en effet de générer des solutions avec des violations d'équité très faibles, y compris sur de nouvelles données, sans affecter de manière trop importante la précision. Le nombre de masques conduisant à la meilleure généralisation varie selon la métrique d'équité et le jeu de données considérés. L'ensemble des résultats obtenus démontrent ainsi l'intérêt de l'approche proposée : générer des solutions plus robustes, et présentant de meilleurs compromis équité/précision sur de nouvelles données, au prix d'une performance dégradée sur l'ensemble d'entraînement.

5 Conclusion

Nous proposons une méthode inspirée de l'ODR améliorant la généralisation de l'équité en apprentissage supervisé. Une évaluation expérimentale 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 des extensions pour d'autres modèles d'apprentissage équitable.

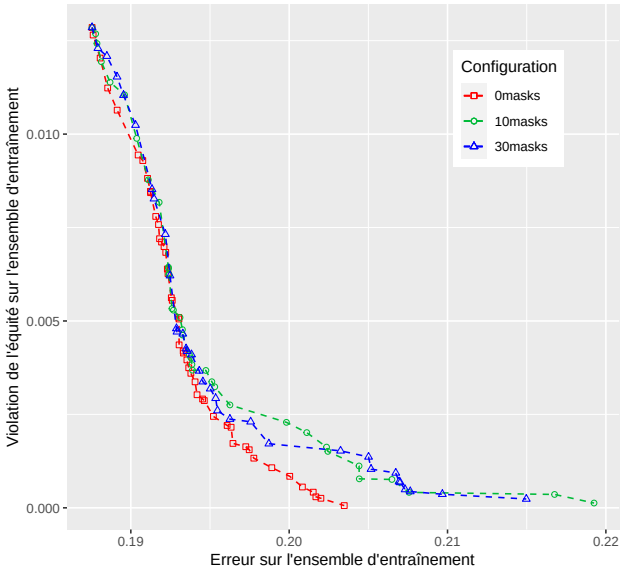

FIGURE 2 – Front de Pareto (violation de l'équité en fonction de l'erreur) sur l'ensemble d'entraînement, pour la *Predictive Equality*, sur l'ensemble de données *Default of Credit Card*

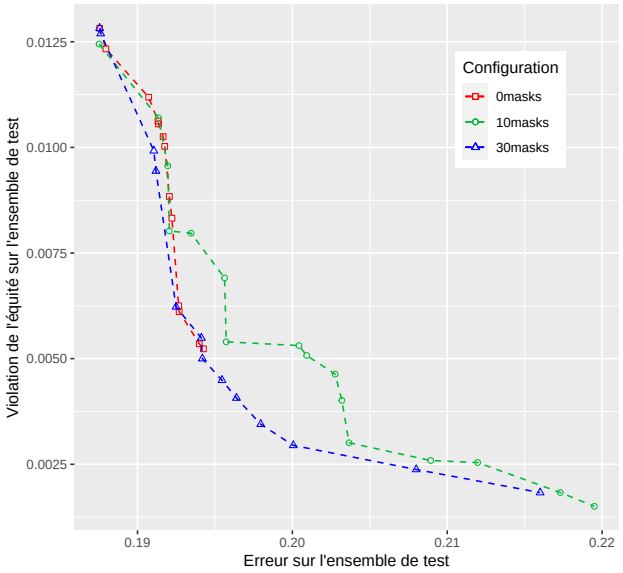

FIGURE 3 – Front de Pareto (violation de l'équité en fonction de l'erreur) sur l'ensemble de test, pour la *Predictive Equality*, sur le jeu de données *Default of Credit Card*

Références

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 2018.

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

[3] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training fairness-constrained classifiers to generalize, 2018.

[4] 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.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[6] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv :1610.02413*, 2016.

[7] Lingxiao Huang and Nisheeth K. Vishnoi. Stable and fair classification. *36th International Conference on Machine Learning, ICML 2019*, 2019-June :5130–5144, 2 2019.

[8] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1) :1–33, 2012.

[9] 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.

[10] 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.

[11] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact : Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.