Utiliser l'équité d'un modèle d'apprentissage pour reconstruire les attributs sensibles de son ensemble d'entraînement

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
[3] UQAM, Montréal, Canada

Mots-clés : *Apprentissage, attaque contre la vie privée, équité, programmation par contraintes, programmation linéaire en nombres entiers.*

1 Introduction et contexte

L'utilisation croissante de modèles d'apprentissage dans les systèmes d'aide à la décision impactant la vie d'individus motive la nécessité de s'assurer de l'équité de ces modèles (*i.e.,* qu'ils ne créent ni ne reproduisent de discriminations). Différentes notions d'équité ont été proposées dans la littérature, parmi lesquelles *l'équité statistique*. Plus précisément, les métriques d'équité statistique mesurent la différence d'une certaine grandeur, qui est fonction de la matrice de confusion d'un modèle (ex : taux de vrais positifs), entre différents *groupes protégés*, définis par la valeur d'un ou plusieurs *attributs sensibles* (genre, origine ethnique, ...).

Un autre enjeu fondamental est le respect de la vie privée. En effet, les modèles sont souvent entraînés sur de grandes quantités de données personnelles. Il est alors important de s'assurer que ces modèles apprennent des motifs génériques utiles sans révéler des informations concernant des individus spécifiques du jeu de données. Dans ce contexte, les *attaques d'inférence* cherchent à exploiter le résultat d'un calcul (ex : un modèle entraîné) pour retrouver des informations sur ses entrées (ex : un ensemble d'entraînement). Notre travail [3] fait partie de la famille des *attaques de reconstruction* [2], dans lesquelles un attaquant essaie de reconstruire tout ou partie (ici, les attributs sensibles) du jeu de données utilisé pour entraîner un modèle.

2 Description du problème et contributions

Soit $D = (X, A, Y) \in \mathcal{X}^n \times \mathcal{A}^n \times \mathcal{Y}^n$ un ensemble de données comportant n exemples, avec $\mathcal{X}$ (resp., $\mathcal{A}$) le domaine des attributs non sensibles (resp., sensibles) et $\mathcal{Y}$ celui des labels. Soit $h : \mathcal{X} \mapsto \mathcal{Y}$ un modèle appris à partir de D, en utilisant un algorithme d'apprentissage supervisé équitable par conception. Ainsi, h respecte une certaine contrainte d'équité vis-à-vis du vecteur des attributs sensibles A, caractérisée par une métrique et une valeur de tolérance ϵ, définissant *l'information de l'équité*. Bien que h n'utilise pas explicitement l'attribut sensible pour ses prédictions (ce qui est en général interdit), A a néanmoins été utilisé pendant l'entraînement. L'objectif de l'attaque de reconstruction [2] proposée [3] est de reconstruire les attributs sensibles A en utilisant les attributs non sensibles X de l'ensemble d'entraînement, les prédictions $\hat{Y} = h(X)$ du modèle équitable ainsi que *l'information de l'équité*.

Dans le pipeline considéré, un attaquant *baseline* propose une première reconstruction des attributs sensibles de l'ensemble d'entraînement. Le vecteur de cette reconstruction, noté $\hat{A} \in \mathcal{A}^n$, s'accompagne éventuellement d'un vecteur $\hat{P} \in [0, 1]^n$ traduisant la confiance de l'attaquant pour chacun des éléments de $\hat{A}$. Un mécanisme de *correction de reconstruction* vient ensuite modifier $\hat{A}$ de sorte à (1) assurer la cohérence de $\hat{A}$ par rapport à l'information de l'équité,

et (2) minimiser les changements (pondérés par $\hat{P}$). La version corrigée de la reconstruction est notée $\hat{A}^*$. La performance de l'attaque *baseline* (resp., *corrigée*) est mesurée comme la proportion d'attributs sensibles correctement reconstruits dans $\hat{A}$ (resp. dans $\hat{A}^*$).

La correction de reconstruction a d'abord été implémentée via un modèle générique de Programmation Linéaire en Nombres Entiers, $\mathcal{RC}(\hat{A}, \hat{P}, \hat{Y}, \epsilon)$. En encodant explicitement la valeur de chaque élément de $\hat{A}^*$, ce modèle permet d'exprimer n'importe quelle contrainte sur ce vecteur, mais comprend un espace de recherche exponentiel (par rapport à n). Pour remédier à cela, nous avons proposé un modèle de programmation par contraintes $\mathcal{RC}_{\mathcal{E}}(\hat{A}, \hat{P}, \hat{Y}, \epsilon)$, comprenant un espace de recherche polynomial (par rapport à n). Nous avons démontré que bien que ce modèle soit moins général que le précédent, dans le cas de métriques d'équité statistique, $\mathcal{RC}_{\mathcal{E}}(\hat{A}, \hat{P}, \hat{Y}, \epsilon)$ et $\mathcal{RC}(\hat{A}, \hat{P}, \hat{Y}, \epsilon)$ partagent le même ensemble de solutions optimales.

En utilisant trois jeux de données de tailles et de caractéristiques différentes, ainsi que quatre métriques d'équité et différentes valeurs de tolérance d'iniquité ϵ, nous avons étudié expérimentalement l'efficacité du mécanisme de correction proposé. La reconstruction *baseline* est d'abord effectuée par un attaquant de la littérature [1] prédisant $(\hat{A}, \hat{P})$ à partir de $(X, \hat{Y})$ et entraîné sur un jeu de données auxiliaire. Cette reconstruction est ensuite corrigée en utilisant $\mathcal{RC}_{\mathcal{E}}(\hat{A}, \hat{P}, \hat{Y}, \epsilon)$. Les résultats présentés dans la figure 1 montrent que l'information de l'équité permet d'améliorer significativement la qualité de la reconstruction des attributs sensibles. Plus précisément, plus la contrainte d'équité imposée est forte (petites valeurs d'ϵ), plus le mécanisme de correction proposé permet d'améliorer la reconstruction des attributs sensibles.

Ces résultats illustrent la tension (déjà observée dans la littérature) entre protection de la vie privée - en particulier des attributs *sensibles* - et apprentissage équitable vis-à-vis de ces mêmes attributs. Des expériences complémentaires montrent par ailleurs que même lorsque l'information de l'équité n'est pas divulguée, un adversaire peut néanmoins l'estimer conduisant à une performance de l'attaque qui reste significative.

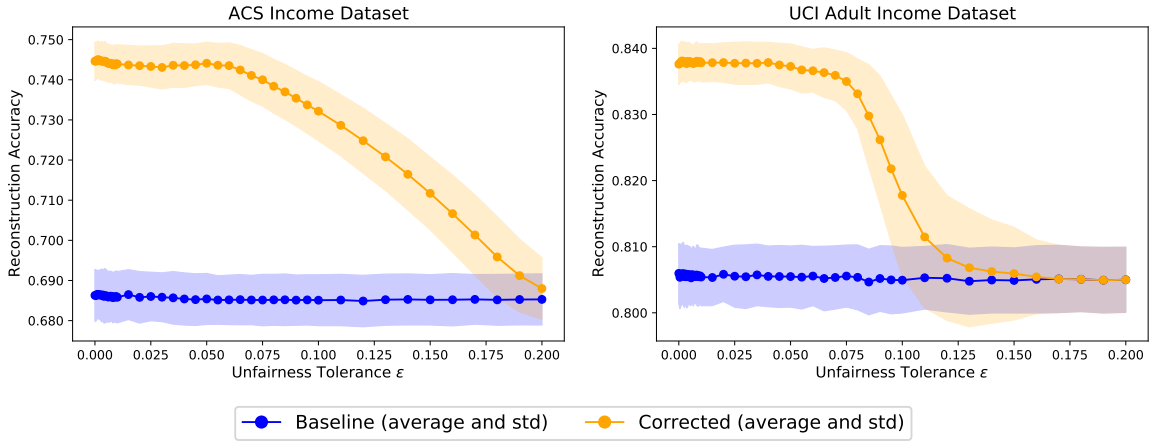

FIG. 1 – Résultats obtenus (moyennés sur 100 séparations entraînement/test différentes) sur deux jeux de données pour la métrique d'équité *Statistical Parity*.

Références

[1] Jan Aalmoes, Vasisht Duddu, and Antoine Boutet. Dikaios : Privacy auditing of algorithmic fairness via attribute inference attacks. *arXiv preprint arXiv :2202.02242*, 2022.

[2] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed ! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 2017.

[3] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Exploiting fairness to enhance sensitive attributes reconstruction. In *The 1st IEEE Conference on Secure and Trustworthy Machine Learning*, SATML, 2023.