Auditer l'équité : l'union fait-elle la force ?

Martijn de Vos, Akash Dhasade, Jade Garcia Bourrée, Anne- Marie Kermarrec, Erwan Le Merrer, Benoît Rottembourg, Gilles Trédan

Auditer l'équité : l'union fait-elle la force ?

Martijn de Vos[1], et Akash Dhasade[1], et Jade Garcia Bourrée[2], et Anne-Marie Kermarrec[1], et Erwan Le Merrer[2], et Benoît Rottembourg[3], et Gilles Tredan[4]

[1]*EPFL, Lausanne, Switzerland*
[2]*Univ Rennes, Inria, CNRS, IRISA*
[3]*Inria, Paris, France*
[4]*LAAS, CNRS, Toulouse, France*

Usuellement, les agents qui auditent l'équité des algorithmes les étudient de manière indépendante, avec leurs propres données. Dans nos travaux, nous considérons le cas de plusieurs agents auditant un même algorithme pour atteindre différents objectifs. Les agents ont la possibilité d'influencer l'audit à travers deux leviers : une stratégie de collaboration avec les autres agents, avec ou sans coordination préalable, et le choix d'une méthode d'échantillonnage. Nous étudions les interactions possibles qui en résultent. Nous prouvons que, contre-intuitivement, la coordination peut se faire au détriment de la précision de l'audit, alors que la collaboration non coordonnée conduit généralement à de bons résultats. Des expériences sur des jeux de données réels confirment cette observation, lorsque nous observons que la précision de la collaboration non coordonnée atteint celle de l'échantillonnage collaboratif optimal.

Mots-clefs : audit distribué, équité, collaboration, coordination

1 Introduction

Les algorithmes de prise de décision deviennent partie intégrante de nombreux processus médicaux, commerciaux, industriels ou administratifs, ayant un impact croissant sur les vies humaines. L'estimation de l'équité des modèles se fait généralement en évaluant la parité démographique D d'un modèle en boîte noire à laquelle des auditeurs accèdent par un ensemble limité de requêtes-réponses [Eur19].

Actuellement, les tâches d'audit sont effectuées sur chaque attribut d'intérêt de manière séquentielle et indépendamment des autres attributs. C'est pourquoi nous posons la question suivante : *Les auditeurs peuvent-ils bénéficier d'une collaboration entre les différentes tâches d'audit, par exemple en construisant et en partageant stratégiquement les requêtes et les réponses ?*

Nous répondons par l'affirmative en analysant deux formes réalistes de collaboration multi-agents pour les audits d'équité. Dans la collaboration *a-posteriori*, les agents partagent à la fois leurs requêtes et les réponses qu'ils reçoivent. Dans la collaboration *a-priori*, de plus, les agents se coordonnent au préalable sur leurs requêtes afin de maximiser les informations qui peuvent être recueillies.

2 Préliminaires

Contexte. On considère un *algorithme en boîte noire*, $\mathscr{A} : \mathscr{X} \mapsto \{0,1\}$, où l'espace d'entrée $\mathscr{X}$ contient des attributs protégés indépendants $X_1,\ldots,X_m$ et $\mathscr{A}$ donne une réponse binaire Y. m agents différents $A_1,\ldots,A_m$ interrogent $\mathscr{A}$ pour auditer l'équité de chaque attribut avec un budget de R requêtes chacun.

Parité démographique. La parité démographique, D, s'est imposée comme une mesure classique de l'équité. Selon [Eur19], entre autres, si $D_i = 0 \pm 0.2$, alors $\mathscr{A}$ respecte la parité démographique sur l'attribut protégé X_i. Définissons $\hat{D}_i$ un estimateur de D (ce que mesurent les agents) :

$$\hat{D}_i = \hat{\mathbb{P}}(Y=1|X_i=1) - \hat{\mathbb{P}}(Y=1|X_i=0). \tag{1}$$

Chaque attribut X_i induit deux groupes dans $\mathscr{X}$ (le groupe favorisé et le groupe défavorisé). Nous appelons *stratum* chaque intersection de ces $2m$ groupes : il y a 2^m strates. Bien que la parité démographique puisse être estimée relativement à chaque strate (comme dans l'équité intersectionnelle [GC23]), nous traitons la parité démographique au niveau de l'attribut.

Objectif d'un auditeur. L'objectif de chaque auditeur est de réaliser un audit aussi précis que possible en minimisant les erreurs entre les estimations obtenues et la réalité. Nous considérons que toutes les estimations obtenues par équité ont la même valeur pour l'auditeur. Ainsi, *la moyenne des erreurs entre les estimations et les valeurs réelles* est utilisée comme mesure de précision de l'audit, bien que d'autres mesures soient également valables. La formule pour cette mesure est $\varepsilon(I) = \frac{1}{m}\sum_{i\in I}\varepsilon(i)$, où $\varepsilon(i)$ est l'erreur de l'estimateur $\hat{D}_i$ (*i.e* la somme des erreurs d'estimation des deux probabilités de l'équation 1). Cette mesure dépend de la stratégie d'échantillonnage utilisée. Comme $\mathscr{A}$ est étudié en boîte noire, l'échantillonnage uniforme et l'échantillonnage stratifié sont considérés pour leur commodité. L'échantillonnage de Neyman est également utilisé comme point de comparaison, bien qu'il soit infaisable en pratique [MSOA13].

3 Stratégies de collaboration

Les agents sont supposés être homogènes, (*i.e.* utilisant tous la même stratégie). L'analyse de la collaboration entre des agents hétérogènes est laissée à des travaux futurs.

Pas de collaboration (no collab.). Pour quantifier l'efficacité de nos stratégies de collaboration, nous considérons comme point de comparaison un cadre non collaboratif dans lequel chaque agent interroge l'algorithme en boîte noire par lui-même et indépendamment des autres [XWV+23].

Collaboration *a-posteriori*. Le schéma de collaboration le plus naturel implique le partage des requêtes et de leurs réponses entre les agents. Dans cette approche, chaque agent interroge indépendamment $\mathscr{A}$, puis partage les requêtes et les réponses qui en résultent avec les autres agents, de sorte que tous les agents puissent accéder à un ensemble de requêtes mis en commun à la fin de l'échantillonnage. L'échantillonnage sur les attributs autres que X_i pour l'agent i est considéré uniforme par défaut.

L'erreur sur D est définie de la façon suivante pour no collab. ($M = 1$) et a-posteriori ($M = m$) :

$$\varepsilon(i) = \sqrt{\frac{\sigma_i^2}{r_i + (M-1)P_iR}} + \sqrt{\frac{\sigma_{\bar{i}}^2}{r_{\bar{i}} + (M-1)P_{\bar{i}}R}} \tag{2}$$

où σ_i est l'écart-type de Y_i, r_i le nombre d'échantillons fait par A_i tels que $X_i = 1$ et $P_i = P(X_i = 1)$. On définit les mêmes variables pour $\bar{i}$ (situation où $X_i = 0$).

Collaboration *a-priori*. Nous introduisons une deuxième stratégie de collaboration, visant à une coordination préliminaire des agents. Avec la collaboration a-priori, les agents se coordonnent sur la stratégie d'échantillonnage à adopter, en tenant compte des tâches et des objectifs des autres agents. Plus précisément, tous les agents divisent l'espace d'entrée en 2^m strates et se mettent d'accord sur la stratégie d'échantillonnage de chaque strate. Cette coordination permet une approche plus intégrée et stratégique de l'audit, dans le but d'améliorer l'efficacité globale de la collaboration.

En collaboration a-priori avec un échantillonnage stratifié ou Neyman, chaque agent A_i biaise son échantillonnage sur chaque strate. Les erreurs doivent être débiaisées. Chaque estimateur d'attribut est calculé comme la moyenne pondérée des estimateurs de strates par la taille relative de la strate dans l'attribut :

$$\varepsilon^2(i) = 4\sum_{k=1}^{2^m} P_k^2 \frac{\sigma_k^2}{m r_k} [\text{dVDGB}^+24], \tag{3}$$

avec P_k la probabilité d'être dans la k-ième strate, r_k le nombre d'échantillons pris dans cette strate par chaque agent, le reste des notations étant inchangé.

4 Bornes théorique sur l'erreur et validation expérimentale

Nous présentons ici les intuitions des trois principaux résultats sur les interactions entre les stratégies de collaboration et les méthodes d'échantillonnage. Des résultats et démonstrations supplémentaires sont disponibles dans la version préliminaire de nos travaux [dVDGB$^+$24].

Théorème 4.1. ***Les collaborations conduisent généralement à de bons résultats.*** *En dehors de situations pathologiques (voir Résultat 3), la collaboration est toujours bénéfique : la moyenne des erreurs sur D diminue d'un facteur égal à la racine carrée du nombre d'agents qui collaborent.*

Preuve. Donnons ici la preuve de ce résultat pour la collaboration a-posteriori. Puisque $\forall i \in I, (M-1)P_iR \geq 0$ (et $(M-1)P_{\bar{i}}R \geq 0$), L'équation 2 se réécrit :

$$\varepsilon_{a-posteriori}(I) = \sqrt{\frac{\sigma_i^2}{r_i + (M-1)P_iR}} + \sqrt{\frac{\sigma_{\bar{i}}^2}{r_{\bar{i}} + (M-1)P_{\bar{i}}R}} \leq \frac{1}{m}\sum_{i\in I}\sqrt{\frac{\sigma_i^2}{r_i}} + \sqrt{\frac{\sigma_{\bar{i}}^2}{r_{\bar{i}}}} \leq \varepsilon_{nocollab.}(I).$$

Ainsi, il est toujours préférable de collaborer (au moins sans coordination).

Théorème 4.2. ***Les méthodes d'échantillonnage intelligentes sont inutiles avec a-posteriori.*** *Les avantages des stratégies d'échantillonnage disparaissent lorsque la collaboration a-posteriori est utilisée par de nombreux auditeurs.*

Preuve. Comme les r_i ne dépendent pas de m, si $m \to +\infty$ alors $r_i + (m-1)P_iR \sim mP_iR$ (et même chose en remplaçant i par $\bar{i}$). Ainsi :

$$\varepsilon_{a-posteriori}(I) = \sqrt{\frac{\sigma_i^2}{r_i + (M-1)P_iR}} + \sqrt{\frac{\sigma_{\bar{i}}^2}{r_{\bar{i}} + (M-1)P_{\bar{i}}R}} \underset{m\to+\infty}{\sim} \sqrt{\frac{\sigma_i^2}{MP_iR}} + \sqrt{\frac{\sigma_{\bar{i}}^2}{MP_{\bar{i}}R}} \underset{m\to+\infty}{\sim} \varepsilon^{uniforme}_{a-posteriori}(I).$$

En conclusion, lorsque le nombre d'agents est grand, chacun peut échantillonner uniformément son attribut.

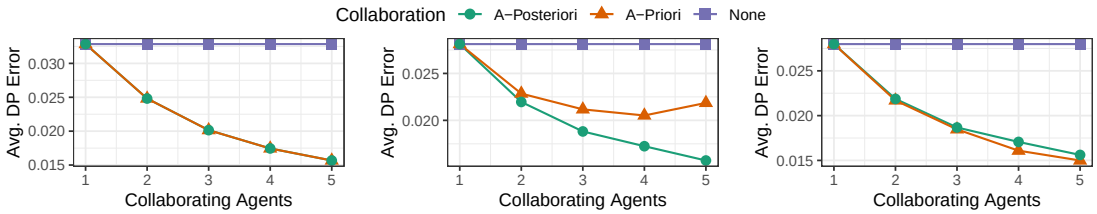
Théorème 4.3. ***La coordination a-priori peut être désavantageuse.*** *L'erreur commise en utilisant stratifié a-priori peut croître avec l'augmentation du nombre d'auditeurs.*

Preuve. Donnons un exemple typique de cas menant à ce résultat. On suppose que les m attributs étudiés sont tous indépendants et très déséquilibrés, *i.e.* $\forall i, P_i \geq 3/4$. Ainsi il existe une des 2^m strates qui est de probabilité $P_* = (3/4)^m$. Dans la majorité des cas, les décisions des $\mathscr{A}$ sur cette strate ne sont pas toujours les mêmes donc $\sigma_* \neq 0$. Par définition, l'erreur moyenne de l'audit est au moins plus grande que l'erreur moyenne faite sur cette strate. Celle-ci est égale à $2P_*\frac{\sigma_*}{\sqrt{mr_*}}$ d'après l'équation (3) (erreur pour la collaboration a-priori) et $r_* = 1/2^m$ (échantillonnage stratifié). On peut donc borner inférieurement l'erreur :

$$\varepsilon^{stratifié}_{a-priori}(i) \geq 2\frac{\sigma_*}{\sqrt{m}}\left(\frac{3\sqrt{2}}{4}\right)^m \underset{m\to+\infty}{\to} +\infty.$$

Ainsi l'union fait la force, mais pas forcément via une coordination préalable.

Validation expérimentale Nous évaluons empiriquement les gains de la collaboration un jeu de données réel (Folktables [DHMS21]) pour voir si les approximations et hypothèses faites concordent avec la réalité expérimentale. Nous considérons les cinq attributs suivants pour l'audit : le sexe, l'état matrimonial, l'âge, la nativité et le statut de mobilité. Pour simuler le modèle en boîte noire, les étiquettes du jeu de données servent de réponses de l'algorithme. Ainsi on élimine le besoin de créer et d'entrainer un modèle d'apprentissage, ce qui biaiserait les résultats. Voir[dVDGB$^+$24] pour plus de détails.
Pour toutes les méthodes d'échantillonnage (Figure 1), les erreurs des collaborations a-posteriori et a-priori sont toujours inférieures à celle de no collab. et que la collaboration a-posteriori produit des résultats presque aussi bons que la stratégie optimale. Ainsi, *il est toujours bénéfique de collaborer* (Théoreme 4.1). La collaboration a-posteriori permet d'obtenir la même erreur pour toutes les méthodes d'échantillonnage, même pour un modeste m = 5 (Théoreme 4.2). Finalement, sous stratification, l'erreur de a-priori diminue ($m = 2,3$), se stabilise ($m = 4$), puis augmente. C'est une illustration du théoreme 4.3.



(a) Uniform Sampling (b) Stratified Sampling (c) Neyman Sampling

FIGURE 1 – Erreur moyenne sur la parité démographique avec l'augmentation du nombre d'agents collaborateurs. Courbes moyennées sur 300 répétitions avec un budget de $R = 500$ requêtes par agent.

5 Travaux et questions connexes

Pour évaluer une collaboration, nous avons utilisé l'erreur moyenne des D. Certains agents vont moins bénéficier de la collaboration que d'autres. Ainsi il peut être intéréssant d'étudier l'équité au sein des agents. C'est ce que font par exemple les auteurs de[XWV+23] en utilisant la théorie des jeux au service de la découverte scientifique (*e.g.* astrophysique). Les objectifs traités par ce papier sont cependant très différents de nos objectifs d'audit. On pourrait aussi définir la métrique de collaboration comme la meilleure des précisions, afin de condamner rapidement pour au moins une infraction.

Nous avons examiné l'équité relative à chaque attribut sans aborder leur intersectionnalité. La collaboration apriori peut se révéler pertinente dans le contexte de telles métriques [GC23]. Ce serait une piste à explorer, bien qu'elle soit hors de notre champ d'étude.

6 Conclusion

Les algorithmes de prise de décision en ligne manquent souvent de transparence, ce qui soulève des inquiétudes concernant leur équité. Les organismes de réglementation cherchent donc à effectuer des contrôles d'équité, mais ils sont limités par le nombre de requêtes qu'ils peuvent émettre. Cet article montre que, si le cadre légal le permet, une collaboration entre des tâches d'audit indépendantes peut améliorer la précision des contrôles tout en respectant les limitations de requêtes. Une étude de cas montre qu'une coordination préalable des requêtes peut être moins efficace qu'une collaboration non coordonnée. Cependant, on constate que la collaboration non coordonnée produit des résultats presque aussi bons que la stratégie optimale (infaisable en pratique). Cela souligne l'importance de la stratégie de collaboration proposée.

Références

[DHMS21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult : New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[dVDGB+24] Martijn de Vos, Akash Dhasade, Jade Garcia Bourree, Anne-Marie Kermarrec, Erwan Le Merrer, Benoit Rottembourg, and Gilles Tredan. Fairness auditing with multi-agent collaboration. *arXiv preprint arXiv :2402.08522*, 2024.

[Eur19] European Parliament and council of the European Union. Regulation 2019/1150 on promoting fairness and transparency for business users of online intermediation services, 2019.

[GC23] Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning : Notions, mitigation, and challenges. *arXiv preprint arXiv :2305.06969*, 2023.

[MSOA13] Olayiwola Olaniyi Mathew, Apantaku Fadeke Sola, Bisira Hammed Oladiran, and Adewara Adedayo Amos. Efficiency of neyman allocation procedure over other allocation procedures in stratified random sampling. *American Journal of Theoretical and Applied Statistics*, 2013.

[XWV+23] Xinyi Xu, Zhaoxuan Wu, Arun Verma, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Fair : Fair collaborative active learning with individual rationality for scientific discovery. *International Conference on Artificial Intelligence and Statistics*, 2023.