Approches combinatoires et bayésiennes pour la détection de biais dans les algorithmes en ligne

Jordan Thieyre[1], Benoît Rottembourg[1]
[1] INRIA, France
{jordan.thieyre,benoit.rottembourg}@inria.fr

Mots-clés *: Fairness, biais, optimisation, sequential decision making, active learning*

1 Introduction

Les recommandations algorithmiques sont omniprésentes dans notre quotidien. Qu'il s'agisse de choisir une pizza à se faire livrer, une série à regarder sur son écran ou un partenaire pour la nuit. Les acteurs économiques qui développent ces algorithmes, parfois opaques, et souvent à base de *machine learning*, ne respectent pas toujours la loi et se font fréquemment sanctionner. L'algorithme induit alors des biais et peut ainsi s'avérer trompeur ou discriminatoire.

Les textes de loi européens de Juillet 2022, le DSA et le DMA obligent les grandes plateformes non seulement à une transparence renforcée quant à la description de leurs algorithmes, mais à des audits réguliers, garantissant le respect du droit et la minimisation des risques ([2]). Certaines entreprises elles-mêmes procèdent à des audits pour s'assurer du respect des règles éthiques qu'elles se donnent.

Auditer un algorithme de recommandation ou de *pricing* en ligne, en quête de présence de biais ou de non-respect du droit oblige l'auditeur à recourir à des méthodes de détection de plus en plus sophistiquées afin que les plateformes numériques ne puissent les esquiver ([1]). C'est dans ce cadre que nos travaux visent à développer des méthodes de détection de biais efficaces pour ces algorithmes.

2 Contraintes et objectifs d'un algorithme d'audit en boîte noire

2.1 La recherche d'un « flagrant délit » par *testing* algorithmique

Un biais de genre par exemple, où une femme se verrait refuser un service alors qu'un homme avec les mêmes données contextuelles se le verrait accorder, est interdit par certaines autorités de régulation comme l'ACPR pour les crédits bancaires. Ainsi si X sont les données client, et A(X) = Y est la réponse de l'algorithme à auditer, effectuer un *testing* va constituer à identifier un sous-ensemble des parties de X, S(X) dans lequel le prédicat (« le genre influe sur la réponse de l'algorithme ») est valide. S(X) sera appelée zone d'intérêt dans la suite.

S(X), sur lequel le prédicat est vérifié doit avoir au moins quatre caractéristiques supplémentaires C :

- C1 : S(X) doit être représentatif des usages de l'algorithme (et non atypique)
- C2 : L'échantillonnage de S(X) doit être statistiquement représentatif pour que la satisfaction du prédicat soit fort probablement vraie en cas de reproduction du test
- C3 : S(X) doit pouvoir se décrire de manière compacte pour un auditeur humain
- C4 : Enfin, il doit représenter un niveau de préjudice suffisant pour motiver les autorités de régulation dans leur sanction.

2.2 Efficacité, frugalité et furtivité

Rechercher une zone S(X) vérifiant C1, .., C4, quand X est un espace multi-dimensionnel de variables à valeurs continues, discrètes et catégorielles est évidemment une tâche ardue, même quand C4 impose un préjudice élevé. Pour aider l'auditeur humain, on s'attend toutefois à ce que des procédés statistiques puissent identifier des variables d'intérêt si l'échantillonnage est intense et couvre l'espace densément. Toutefois, le contexte de l'audit impose un budget B en terme de nombre de requêtes à la boîte noire, pour d'une part éviter d'endommager le service (on fausserait l'algorithme avec des requêtes artificielles) et ne pas être facilement identifié par la plateforme.

2.3 Recherche de zones d'intérêt à échantillonnage fixé

Nous montrerons l'intérêt de combiner des méthodes de régression, de clustering et d'optimisation combinatoire pour identifier des zones d'intérêt avec à la fois une probabilité forte de validité (l'échantillonnage peut être insuffisant dans la zone) et une forme exploitable par un régulateur.

2.4 Intérêt de stratégies d'échantillonnages adaptatives

A l'image des problèmes de bandits multi-bras, où l'on chercherait ici à bien dépenser son budget B d'échantillonnage, pour trouver la machine/zone qui maximise les chances d'être une zone d'intérêt, nous montrerons la valeur d'une stratégie d'échantillonnage de type *Optimisation Bayésienne* ou inspirée par des techniques *d'Active Learning*, en les comparant avec des échantillonnages aléatoires et semi-aléatoires.

Nous illustrerons sur la Figure (1) les efficacités comparées de plusieurs stratégies d'échantillonnage portant sur des algorithmes de *pricing* de livraison de repas et d'agences de voyage en ligne.

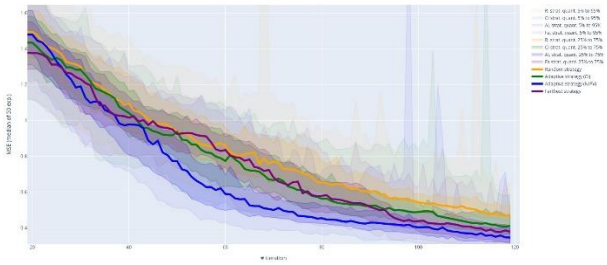

FIG. 1 – Evolution du MSE par itération selon la stratégie d'échantillonnage

3 Conclusions

L'audit en boîte noire est une des manières d'assurer une surveillance des plateformes numériques, mais elle demande un support algorithmique conséquent. Ces algorithmes d'audit peuvent être rendus plus efficaces par des approches combinatoires et statistiques et ouvrent un champ de recherche très riche, tant les algorithmes à auditer sont variés et en perpétuelle évolution.

Références

[1] E. Le Merrer, G. Trédan. Remote explainability faces the bouncer problem. *Nature Machine Intelligence* 2 (9), 529-539

[2] B. Rottembourg. The Need for AI Regulation and Audit Tools for Bias and Disloyalty Detection. *ERCIM News Special Theme: Ethical Software Engineering and Ethically Aligned Design*. 131 (2022), pp. 8–9