Détection de biais et intégration de connaissances expertes pour l'explicabilité en Intelligence Artificielle

Matthieu Delahaye[1], Lina Fahed[1], Florent Castagnino[2], Philippe Lenca[1]

[1] IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
[2] IMT Atlantique, LEMNA, F-44307 Nantes, France

prénom.nom@imt-atlantique.fr

Le déploiement de modèles d'apprentissage automatique "boîtes noires", dans des secteurs sensibles notamment, a entraîné un fort besoin d'explicabilité adapté au niveau de compréhension des acteurs décideurs. Dans le secteur sensible de la sécurité urbaine, de nombreux modèles de prédiction d'infractions (contraventions, délits, crimes) dans le temps et l'espace ont été proposés. En France, la plupart de ces modèles ne sont plus utilisés à cause, principalement, de l'absence d'informations complémentaires à l'intuition des policiers [2]. Dans l'objectif de comprendre les phénomènes liés à la sécurité urbaine et pour apporter des informations pertinentes aux policiers, nous souhaitons proposer un nouveau modèle d'Intelligence Artificielle eXplicable (XAI) [3] qui relève les deux défis suivants :

Défi 1 : Détection de biais par l'explication

Le biais cognitif présent dans le jugement et la décision humaine est un phénomène naturel [1]. Sachant qu'un modèle d'apprentissage automatique, par définition, extrait des connaissances à partir de données qui sont, elles, souvent générées et collectées par l'humain, cela engendre la présence de biais dans le modèle. La littérature s'est principalement concentrée sur l'atténuation des biais algorithmiques, ce qui peut entraîner une incapacité de justification du résultat du modèle [1]. Or, dans un contexte d'aide à la décision à fort impact social, il est nécessaire de considérer les acteurs impliqués et leurs contextes sociaux tout au long de la modélisation afin de conserver un modèle fiable et non biaisé. Notre intérêt se porte plutôt sur les étapes préliminaires au traitement de biais : la détection et l'évaluation des biais. Pour détecter exhaustivement les biais du modèle, nous cherchons à expliquer son comportement. L'explication des modèles peut se faire aussi bien à l'échelle locale que globale [3]. Localement, générer des contre-factuels, par exemple, permet de modifier légèrement les instances de façon à changer la prédiction. Dans ce cas, la détection de biais peut se faire en analysant des cas précis. À l'échelle globale, l'explication est donnée à propos du comportement entier du modèle ce qui permet de détecter les biais de façon plus complète. Dans notre contexte, les techniques d'explication liées à ces deux échelles sont à explorer.

Défi 2 : Généralisation par intégration de connaissances

Les modèles d'apprentissage automatique fondés exclusivement sur des données, parfois imparfaites, sont limités en terme de généralisation [5]. Dans le but de préciser leur raisonnement, nous proposons d'intégrer des connaissances expertes durant l'apprentissage du modèle afin qu'il assimile le contexte et les contraintes extérieures n'étant pas explicitement dans les données [5]. Dans notre projet, les connaissances expertes sont le résultat d'entretiens réalisés auprès des professionnels de la sécurité, des criminologues et sociologues de la police. La pluralité d'acteurs limite les biais induits par la subjectivité de chaque expert. La piste envisagée est d'incorporer les connaissances dans la fonction de coût du modèle [5] qui se verra ajouter un nouveau terme pouvant être vu comme un terme de régularisation permettant, par exemple, de tempérer la stigmatisation. Afin de vérifier la fiabilité des explications, l'approche d'eXplonatory Interactive Learning (XIL) [4], qui fait intervenir un expert donnant un retour sur la conformité des explications fournies, sera considérée. L'intervention experte contribue à rendre le modèle plus réaliste et renforce la relation de confiance de la police envers l'outil.

Références

[1] Salem Alelyani. Detection and evaluation of machine learning bias. *Applied Sciences*, 11 :6271, 07 2021.

[2] La Quadrature du Net. La police prédictive en France : contre l'opacité et les discriminations, la nécessité d'une interdiction. `https://www.laquadrature.net/`, Janvier 2024.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5) :1–42, 2018.

[4] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat Mach Intell*, 2(8) :476–486, 2020.

[5] L. von Rüden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, et al. Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans. Knowl. Data Eng.*, 35(1) :614–633, 2023.