Investiguer la notion d'équité algorithmique dans les environnements informatiques pour l'apprentissage humain

Mélina Verger[0000−0002−5839−882X], 1ère année de thèse

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
melina.verger@lip6.fr

Résumé L'utilisation croissante de systèmes de décision fondés sur l'analyse de données dans l'éducation suscite des inquiétudes quant à leur équité envers certains apprenants ou groupes d'apprenants. Dans cet article, qui s'inscrit dans le cadre d'une problématique plus générale d'évaluation de l'équité des systèmes algorithmiques, nous nous intéressons ici à la définition de l'équité ainsi qu'à ce qui l'affecte dans l'utilisation des systèmes algorithmiques. Par conséquent, nous mettons en évidence les multiples sens de la notion d'équité et son lien avec les biais algorithmiques pour en pointer quelques sources. Nous présentons ensuite des travaux relatifs à l'évaluation de l'équité et concluons sur les objectifs de recherche de la thèse en cours.

Mots-clés : Equité · Biais algorithmiques · Educational data mining.

1 Introduction

La fouille de données éducatives (*Educational Data Mining* - EDM) est un domaine de recherche qui vise à comprendre et améliorer l'apprentissage humain à partir des données [11]. Cependant, les données comportent des biais historiques, souvent illustrés par des inégalités de genre dans certains cursus académiques ou par la diversité des profils d'apprenants [2]. De fait, des *patterns* de discrimination sont observés, appris et éventuellement reproduits par les modélisations en EDM. Au-delà des données elles-mêmes, d'autres biais peuvent s'introduire dans la manière dont elles sont traitées (e.g. suppression des données anormalement éloignées de la moyenne, au détriment d'élèves au comportement atypique), ainsi que dans le jugement humain qui a lieu dans l'analyse des résultats obtenus (e.g. simplification dans l'interprétation de regroupements d'élèves comme dépendant uniquement de leur niveau même si l'analyse présentée est multidimensionnelle).

Par conséquent, les recherches sur l'évaluation de l'équité des *systèmes algorithmiques*, visant à ne pas perpétuer les biais dans les prises de décision (en éducation ou ailleurs), se sont accélérées ces dernières années [10]. Nous emploierons les termes *systèmes algorithmiques* pour se référer aux systèmes fondés sur une analyse de données (*data-driven*), comprenant à la fois le jeu de données, le modèle (algorithme entraîné sur les données) et les résultats obtenus guidant la prise de décision. Ces recherches sont motivées par une demande sociétale forte

de transparence et une pression des instances de régulation pour l'utilisation éthique des données et des algorithmes, notamment au niveau européen.

Les travaux actuels en EDM se concentrent en majorité sur la considération d'un biais, d'origine démographique, comme le genre ou l'ethnie dans le milieu anglo-saxon [1]. L'équité est ensuite évaluée par différentes mesures selon les travaux [1], donnant lieu à des interprétations diverses de l'équité.

Dans cet article, il s'agit donc d'investiguer la notion d'équité dans les environnements informatiques pour l'apprentissage humain (EIAH), pour préparer une réponse future à la problématique d'évaluation de l'équité des systèmes algorithmiques soulevée précédemment. Plus précisément, nous énonçons des objectifs intermédiaires à cette problématique en section 5 qui seront ceux abordés dans la suite de la thèse.

Ainsi, nous étudions la notion d'équité en section 2, puis nous distinguons équité et biais algorithmiques en section 3 pour déterminer les sources courantes d'inéquité dans les systèmes algorithmiques. Nous présentons ensuite des exemples d'évaluation de l'équité en EDM en section 4 pour, une fois posé le cadre, qui est l'objet de cet article, énoncer en section 5 les objectifs de recherche envisagés dans ce travail de thèse. Enfin, l'article conclut sur une synthèse en section 6.

2 Définition de l'équité et implications

Dans le sens commun, l'équité est décrite comme l'absence de discrimination, c'est-à-dire l'absence de distinction entre deux ou plusieurs personnes à partir de certains critères. Toutefois, un traitement non-discriminant comme un traitement égal pour tous n'est pas systématiquement équitable.

Pour dépasser ce conflit entre équité et non-discrimination, la philosophie introduit un sens moral de l'équité qui consiste à questionner les distinctions qui sont conformément acceptables ou non d'opérer. Ainsi, l'équité est définie comme la capacité d'adapter ce qui s'applique à tous à la singularité des situations [4]. Par conséquent, plutôt qu'éliminer les discriminations, on préférera prendre en compte les inégalités de faits pour rendre possible l'égalité de résultats, ou ce qui est communément appelé l'égalité des chances en éducation.

En conclusion, cette définition de l'équité implique de devoir définir, pour chaque situation, ce pour quoi l'égalité de résultats est attendue (e.g. la maîtrise d'une connaissance) et quelles différences considérer comme inégalités de faits (e.g. la vitesse d'apprentissage, les capacités métacognitives, le milieu socio-économique...). De fait, la multiplicité des sens possibles de l'équité, dépendants du contexte d'application, se traduit par une absence de consensus sur la manière de l'introduire et de l'évaluer dans les travaux de recherche en EDM, ce qui sera montré en section 4.

3 Sources d'inéquité algorithmique

En plus des enjeux que pose la définition de l'équité, les biais algorithmiques inhérents à l'utilisation de systèmes algorithmiques peuvent eux aussi affecter l'équité des résultats. Plus précisément, nous verrons dans cette section 1) où peuvent apparaître les biais algorithmiques pour 2) en identifier certains via une cartographie et 3) pourquoi ils peuvent être source d'inéquité algorithmique, c'est-à-dire d'inéquité engendrée par l'utilisation des systèmes algorithmiques.

Premièrement, plusieurs classifications des biais algorithmiques ont été proposées selon les différentes phases de développement des systèmes algorithmiques. [8] distingue les phases de (a1) mesure, (a2) d'apprentissage du modèle, et (a3) d'action. La mesure est la phase de collecte des données. L'apprentissage du modèle est la phase qui utilise les données collectées – les données d'entraînement – pour développer une représentation de l'environnement. L'action est l'utilisation des prédictions du modèle pour de nouveaux cas de jugement et de prise de décision. L'étude [10], elle, différencie les phases de (b1) pré-traitement, (b2) traitement, et (b3) post-traitement. Les mécanismes de pré-traitement consistent à modifier les données d'apprentissage avant de les introduire dans un algorithme d'apprentissage automatique ; ceux en cours de traitement consistent à modifier les algorithmes d'apprentissage automatique pour tenir compte de l'équité pendant la période d'entraînement ; et ceux en post-traitement effectuent un traitement des résultats en sortie du modèle pour rendre les décisions plus équitables.

Deuxièmement, nous nous sommes appuyés sur ces classifications pour cartographier en Figure 1 un ensemble de biais et leurs sources principales dans les systèmes algorithmiques. Sans être exhaustive, cette cartographie permet d'exposer des points de vigilance dans le développement d'un tel système. Dans la première source principale, les données, nous retrouvons la phase de mesure ou collecte identifiée par [8], puis nous avons distingué les sous-groupes "valeurs" et "attributs", correspondant au pré-traitement de [10], ainsi que "variable cible". Dans le sous-groupe "collecte", le biais d'échantillonnage concerne la sélection d'un échantillon partiellement ou pas représentatif de la population étudiée ; de plus, une documentation insuffisante dissimule des biais issus de la manière dont la collecte a été effectuée (e.g. choix des participants, acquisition et format des données). Dans le sous-groupe "attributs", la sélection de ceux-ci pour expliquer la variable cible et la présence d'attributs "proxy", c'est-à-dire dont les valeurs permettent de déduire un autre attribut ou même la variable cible, engendrent des biais. Dans le sous-groupe "variable cible", la définition de celle-ci présente en général un écart avec le concept abstrait et non mesurable à représenter (e.g. prédire une note à un exercice – la variable cible – pour juger la maîtrise d'une connaissance – le concept abstrait) ; par ailleurs, les erreurs de labellisation sont courantes, qu'elles soient dues à une mauvaise mesure ou à une saisie manuelle. Dans le sous-groupe "valeurs", le biais de déséquilibre fait référence au problème de sous-représentation des minorités sur lesquelles le modèle apprend moins bien et des biais sont induits par la manière dont sont gérées les valeurs inconsistantes.

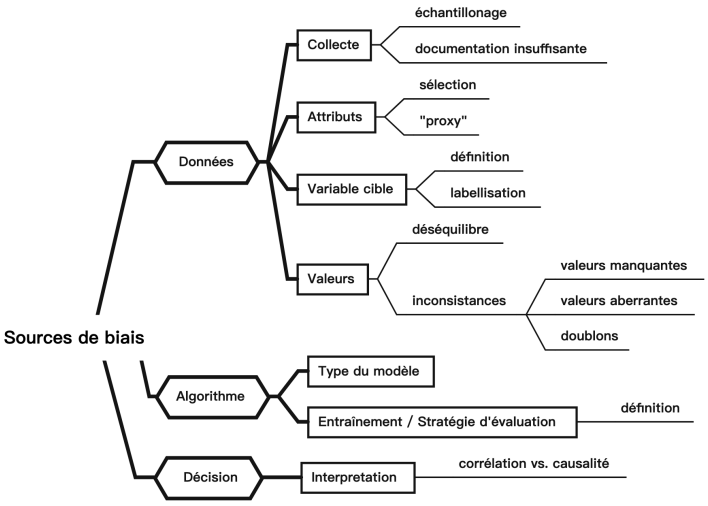

Fig. 1. Cartographie de biais et leurs sources dans les systèmes algorithmiques.

Dans la deuxième source principale, l'algorithme, des biais peuvent également provenir du type de modèle utilisé (e.g. biais de conception) et de la manière dont le modèle est entraîné (phases (a2) et (b2)).

Dans la dernière source, nous avons considéré les biais humains dans l'interprétation des résultats du modèle et par conséquent dans les décisions, conduisant aux phases (b3) et (a3).

Pour conclure, l'ensemble des biais présents dans un système algorithmique, mais également les choix faits pour les traiter, modifie la façon dont le système produit des résultats. Les biais impactent donc l'équité algorithmique des systèmes, pouvant avoir des répercussions sur des groupes distincts de personnes présentes dans les données. Dans le cadre des EIAH, bien que les biais d'origine humaine peuvent influer sur les systèmes algorithmiques, nous ne considérons que les biais algorithmiques, induits par l'utilisation de tels systèmes.

4 Travaux associés à l'évaluation de l'équité

Suite aux conclusions des deux sections précédentes, nous nous demandons comment mesurer l'impact des biais sur les systèmes, autrement dit comment évaluer l'équité de leurs résultats. Nous avons identifié deux travaux en EDM traitant directement cette question. D'une part, [7] a comparé la performance de plusieurs modèles de prédiction du décrochage dans les MOOCs envers les femmes ou les hommes. Pour cela, l'étude propose une nouvelle mesure d'équité, ABROCA (*Absolute Between-ROC Area*), et utilise une méthode d'analyse par tranches pour tester l'équité de leurs modèles dans différents sous-groupes d'apprenants. Cette approche sous-tend que l'équité est satisfaite lorsqu'un modèle produit des

résultats équivalents pour chaque genre, c'est-à-dire quand l'ABROCA, l'aire entre les courbes ROC (*receiver operating characteristic*) associées aux genres à travers chaque groupe, est moindre.

D'autre part, [9] a pris en compte deux variables de comparaison, le genre, constitué de femmes et d'hommes, et l'origine ethnique, distinguant les ethnies majoritaires et les ethnies minoritaires aux États-Unis. L'étude compare la performance d'un modèle de prédiction du succès dans un cours envers ces quatre groupes avec quatre mesures différentes : la précision du modèle, l'égalité des chances, la parité démographique et la parité des prédictions correctes. Entre les groupes, des écarts plus importants étaient constatés selon les mesures, suggérant une inéquité du modèle envers les hommes d'ethnies minoritaires en termes de parité démographique et d'égalité des chances.

Nous constatons qu'aucun consensus n'apparaît autour du choix des variables de comparaison, aussi appelées *attributs protégés*, et que l'équité peut s'étudier de manière multi-attribut. Il n'y pas non plus de consensus sur le choix des mesures à employer. En revanche, l'équité est à chaque fois adressée par la recherche d'équivalence des performances des modèles à travers les groupes. Seulement, cette approche réductrice pose problème quand les groupes de comparaison comportent effectivement des inégalités et qu'une équivalence de performances engendrerait des résultats altérés pour certains apprenants [6].

5 Objectifs de recherche

Dans cette section, nous présentons les différents objectifs de recherche (OR) envisagées pendant la thèse. Nous précisons que, dans le cadre des EIAH, nous nous intéressons uniquement à l'équité du point de vue algorithmique.

5.1 OR 1 : Choix de la formalisation de l'équité

Comme vu en section 3, il existe de nombreux biais algorithmiques. Il existe aussi de nombreuses définitions formelles de l'équité, représentées par des *fairness metrics* [12,3,10,1]. Par exemple, si l'on pose S l'attribut protégé (e.g. le genre) et $S = 1$ le genre pour lequel les biais sont favorables (groupe privilégié), $P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S \neq 1)$ définit la mesure de parité démographique citée en section 4, avec $\hat{Y}$ la valeur de la prédiction arbitrairement posée à 1.

Par conséquent, nous allons déterminer un contexte de travail pour choisir une formalisation adéquate. En effet, le contexte de travail permettra de poser l'égalité de résultats attendue et les différences à considérer, comme soulevé en section 2. Ainsi, nous comparerons les formalisations existantes, identifiées notamment par un travail théorique de revue systématique.

5.2 OR 2 : Identification des biais algorithmiques dans les données éducatives

Nous souhaitons ensuite nous concentrer sur une des principales sources de biais algorithmiques, les données (section 3). Plus précisément, dans le contexte de

travail défini en OR 1, nous étudierons l'impact sur les biais des spécificités des données éducatives, telles que : leur granularité, leur multi-modalité (données de logs, de comportements, d'état émotionnel, académiques...), le caractère latent de certaines variables dans la modélisation des connaissances (e.g. dans le *knowledge tracing* [5]), leur non-généralisabilité (e.g. données spécifiques à un seul cours), etc. Nous prendrons également en compte l'effet de phénomènes couramment étudiés dans la communauté EDM sur les données (e.g. *wheel spinning*, *mind wandering*, *gaming the system*). Ce travail permettra d'identifier la présence systématique de certains biais dans des familles de jeux de données éducatives, ainsi que leur impact sur certains algorithmes en particulier, en donnant des indications sur l'aspect multi-attributs de l'équité.

5.3 OR 3 : Mitigation automatique de biais algorithmiques

À l'aide des résultats des OR 1 et 2, nous souhaitons déterminer automatiquement le risque de biais d'équité dans un jeu de données en fonction de divers critères afin de recommander des stratégies de mitigation locales adaptées (e.g. collecter des données supplémentaires auprès d'une population particulière, éviter l'utilisation de tel attribut ou de telle famille d'algorithmes).

De plus, en se concentrant sur un ou plusieurs biais algorithmiques éducatifs fréquents clairement identifiés dans l'OR 2, nous souhaitons concevoir et implémenter une méthode de mitigation globale pour voir s'il est possible de combiner différents algorithmes équitables selon des critères différents pour obtenir une décision globalement plus équitable.

6 Conclusion

Dans cet article, nous avons mis en évidence les multiples sens de la notion d'équité, distingué équité et biais algorithmiques pour en déterminer les principales sources dans les systèmes algorithmiques, présenté des travaux relatifs à l'évaluation de l'équité et enfin décrit les objectifs de recherche envisagés dans ce travail de thèse. L'évaluation de l'équité en EDM et dans les EIAH contribue à renforcer la prise de conscience des biais liés aux systèmes algorithmiques et à outiller la communauté pour mieux traiter ce problème à l'avenir.

Références

1. Baker, R.S., Hawn, A. : Algorithmic Bias in Education. International Journal of Artificial Intelligence in Education (2021). https://doi.org/10.1007/s40593-021-00285-9
2. Bickel, P., Hammel, E., O'connell, J. : Sex Bias in Graduate Admissions : Data from Berkeley. Science (1975). https://doi.org/10.1126/science.187.4175.398
3. Castelnovo, A., Crupi, R., Greco, G., Regoli, D. : The zoo of Fairness metrics in Machine Learning. arXiv :2106.00467 [cs, stat] (2021), `http://arxiv.org/abs/2106.00467`

4. Comte-Sponville, A. : Dictionnaire Philosophique. Presses Universitaires de France - PUF (2001)
5. Corbett, A.T., Anderson, J.R. : Knowledge tracing : Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction **4**, 253–278 (1995)
6. Corbett-Davies, S., Goel, S. : The Measure and Mismeasure of Fairness : A Critical Review of Fair Machine Learning. arXiv :1808.00023 [cs] (2018), http://arxiv.org/abs/1808.00023
7. Gardner, J., Brooks, C., Baker, R. : Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In : Proceedings of the 9th International Conference on Learning Analytics & Knowledge. pp. 225–234. ACM, Tempe AZ USA (2019). https://doi.org/10.1145/3303772.3303791
8. Kizilcec, R.F., Lee, H. : Algorithmic Fairness in Education. arXiv :2007.05443 [cs] (2021), http://arxiv.org/abs/2007.05443
9. Lee, H., Kizilcec, R.F. : Evaluation of Fairness Trade-offs in Predicting Student Success. arXiv :2007.00088 [cs] (2020), http://arxiv.org/abs/2007.00088
10. Pessach, D., Shmueli, E. : Algorithmic Fairness. arXiv :2001.09784 [cs, stat] (2020), http://arxiv.org/abs/2001.09784
11. Romero, C., Ventura, S. : Educational data mining and learning analytics : An updated survey. WIREs Data Mining and Knowledge Discovery **10**(3), e1355 (2020). https://doi.org/10.1002/widm.1355
12. Verma, S., Rubin, J. : Fairness definitions explained. In : Proceedings of the International Workshop on Software Fairness. pp. 1–7. ACM, Gothenburg Sweden (2018). https://doi.org/10.1145/3194770.3194776, https://dl.acm.org/doi/10.1145/3194770.3194776