Compatibility of Fairness Metrics With EU Non-discrimination Law: A Legal and Technical Case Study

Yasaman Yousefi[1,2,†], Lisa Koutsoviti-Koumeri[3,†], Magali Legast[4,†], Christoph Schommer[2], Koen Vanhoof[3] and Axel Legay[4]

[1]*Università di Bologna*
[2]*University of Luxembourg*
[3]*Hasselt University*
[4]*Université catholique de Louvain*

Abstract

Algorithmic decisions made by Machine Learning (ML) models may pose a threat of discrimination. This research endorses the contextual approach to fairness in the EU non-discrimination legal framework and aims to assess to what extent we can ensure legal fairness using fairness metrics and constraints in ML models. We examine the legal concepts of non-discrimination and differential treatment, using different fairness definitions. In a case study with different scenarios, we train classifiers with bias mitigation methods involving different fairness constraints. Our goal is to determine how effective they are at mitigating prediction bias while respecting the judiciary contextual approach and the substantive notion of equality under EU law.

Keywords

Non-discrimination, Algorithmic decision-making, Fairness, Machine learning, Bias mitigation, Classification

1. Introduction

Algorithmic decisions made by Machine Learning (ML) models affect many sectors of our lives and bring along ethical and legal questions such as that of fairness. It is difficult to define the concept of fairness and guarantee unbiased results, both in human and algorithmic decisions. Fairness is approached mathematically in computer science and multiple fairness definitions and metrics have been proposed for understanding, avoiding, and reducing biases [1]. In the European Union (EU) legal system, fairness is usually achieved through a non-discriminatory framework that includes Article 21 of the European Charter of Fundamental Rights (CFREU) and Article 14 of the European Convention on Human Rights (ECHR). Discrimination based

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
[†] These authors contributed equally.
yasaman.yousefi3@unibo.it (Y. Yousefi); lisa.koutsoviti@uhasselt.be (L. Koutsoviti-Koumeri); magali.legast@uclouvain.be (M. Legast); christoph.schommer@uni.lu (C. Schommer); koen.vanhoof@uhasselt.be (K. Vanhoof); axel.legay@uclouvain.be (A. Legay)
0000-0003-1483-2978 (Y. Yousefi); 0000-0002-9490-6035 (L. Koutsoviti-Koumeri); 0000-0003-4246-115 (M. Legast); 0000-0002-0308-7637 (C. Schommer); 0000-0001-7084-4223 (K. Vanhoof); 0000-0003-2287-8925 (A. Legay)

on protected characteristics such as sex, religion or social origin is prohibited. Legal [2] and technical [3] literature reviews indicate a clear gap between non-discrimination laws and computer science when addressing discrimination. They call for a multi-disciplinary research on the compatibility of the mathematical and legal definitions of fairness.

The main question we address in our research is how adequate are the different fairness definitions in compliance with EU non-discrimination law. We study the suitability of bias mitigation algorithms and fairness metrics in addressing illegal discrimination, both to ensure compliance when developing systems and as tools to detect and prove algorithmic discrimination. This opens a discussion with several sub-questions raising technical and legal points.

To address these questions, we pursue with a practical case study in different scenarios using the ML classification problem. We look at the difference between the discrimination in the data, evaluated through several fairness metrics, and the discrimination in the predictions of models optimized under different fairness constraints. We then discuss those results both from a technical point of view and in the light of EU non-discrimination law. We use legal informatics methodology which interprets the legal concept of fairness and adapts it to emerging technological paradigms and vice versa [4].

2. Methodology

We study the relevant Articles for non-discrimination in EU laws to analyze how their scope of protection can be applied in algorithmic decision-making scenarios. We consider both existing fairness definitions and experimental results of bias mitigation under fairness constraint.

In the experimental setup, we use several publicly available datasets that are widely used in fair ML classification settings, such as COMPAS [5] and Adult [6]. We evaluate the amount of bias they present and use them to train fair classification models, using a learning algorithm with inprocessing bias mitigation. We repeat the process with different choices of fairness constraint and different strength for the constraint, using otherwise the same algorithm. We then evaluate the amount of bias present in the models despite the mitigation, comparing models optimized on different fairness metrics with each other and with unconstrained ones.

We examine different fairness definitions such as Demographic Parity (DP) [7], Conditional Demographic Disparity (CDD) [8] or Disparate Mistreatment [9] to evaluate the level of discrimination in the datasets and predictions. We analyse both direct and indirect discrimination as well as group and individual fairness, using the above-mentioned metrics, and two novel metrics -namely Fuzzy-Rough Uncertainty (FRU) and Fuzzy Cognitive Maps (FCMs)- that consider all features and non-linear relationships [10, 11].

To create models with different fairness constraints, we use the meta-algorithm introduced by Celis et al. [12]. This algorithm trains a classifier while respecting a minimal value allowed for the measure of fairness. This fairness constraint is given as input and can be one out of several metrics, more specifically, any non-convex linear-fractional constraint. This approach allows to use a larger number of existing fairness metrics as constraint as compared to other existing bias mitigation methods, which is ideal to analyze the effect of the fairness constraint in itself as opposed to that of the algorithm. We use the open-source implementation available in AIF360 [13] which uses gradient descent.

3. Expected contribution and results

The innovation of this study is severalfold. First, we incorporate legal considerations into the bias mitigation pipeline. Second, we compare and analyze how different fairness constraints impact bias mitigation, taking other fairness perspectives into account. Further, the current open-source implementation of the meta-algorithm [12] available through AIF360 [13] is only able to handle two existing bias metrics and binary labels and attributes. Therefore, one additional contribution of this work is extending the available code implementation to account for the aforementioned limitations.

In our analysis, we take into account the aim of EU Non-discrimination law to achieve substantive equality, rather than only preventing ongoing discrimination and ensuring formal equality. To achieve substantive equality, treating everyone the same going forward and ignoring past discrimination based on social group attributes is insufficient. True equality involves acknowledging that the status quo is often not neutral [14] because certain groups start from unequal points resulting from historical biases they have experienced.

This perspective is supported by the jurisprudence of the European Court of Justice (ECJ), emphasizing that differences between groups must be recognized in order to achieve substantive equality in practice. This approach to non-discrimination focuses not only on addressing technical biases and discrimination on the surface, but also on tackling the underlying social biases that contribute to inequality.

Considering the above elements, we study in different scenarios how strong the constraint on fairness during training should be to optimize the model, considering both accuracy and the results of fairness metrics. We explore a legal approach based on contextual and substantive equality ideals for the choice of thresholds impacting accuracy and fairness and propose the introduction of a margin for a trade-off between fairness and accuracy in the upcoming Artificial Intelligence Act.

In our preliminary results where DP was used as fairness constraint, we already identified scenarios for which the bias mitigation substantially improved fairness, provided that bias was relatively high in the training data. On the other hand, other scenarios led to different results, sometimes even reducing fairness. We could also observe that the fairness constraint on DP was usually improving other fairness metrics as well, but could also reduce it, which may or not be a problem depending on the situation at hand. Those first results already highlight the importance of taking the context into consideration when taking decisions about AI development. The completion of this research will provide better insight on the impact of the different fairness constraints, including their relationship with other metrics and their compliance with EU legal principles.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning 54 (2021) 115:1–115:35. URL: https://doi.org/10.1145/3457607.

[2] D. Hellman, Measuring algorithmic fairness, Virginia Law Review 106 (2020) 811–866. URL: https://www.jstor.org/stable/27074708.

[3] M. Dolata, S. Feuerriegel, G. Schwabe, A sociotechnical view of algorithmic fairness, Information Systems Journal 32 (2021) 754 – 818. URL: https://doi.org/10.5167/uzh-207228.
[4] G. Sartor, Informatica giuridica, Il diritto nella società dell'informazione (2006).
[5] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism 4 (2018).
[6] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.
[7] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12, Association for Computing Machinery, 2012, pp. 214–226. URL: https://doi.org/10.1145/2090236.2090255.
[8] S. Wachter, B. Mittelstadt, C. Russell, Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai, Computer Law & Security Review 41 (2021) 105567. URL: http://dx.doi.org/10.2139/ssrn.3547922.
[9] M. B. Zafar, I. Valera, M. Gomez Rodriguez, K. P. Gummadi, Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment, in: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2017, pp. 1171–1180. doi:10.1145/3038912.3052660.
[10] G. Nápoles, I. Grau, L. Concepción, L. Koutsoviti Koumeri, J. P. Papa, Modeling implicit bias with fuzzy cognitive maps, Neurocomputing 481 (2022) 33–45. URL: https://www.sciencedirect.com/science/article/pii/S092523122200090X.
[11] G. Nápoles, L. Koutsoviti Koumeri, A fuzzy-rough uncertainty measure to discover bias encoded explicitly or implicitly in features of structured pattern classification datasets, Pattern Recognition Letters 154 (2022) 29–36. URL: https://www.sciencedirect.com/science/article/pii/S0167865522000058.
[12] L. E. Celis, L. Huang, V. Keswani, N. K. Vishnoi, Classification with fairness constraints: A meta-algorithm with provable guarantees, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 319–328. URL: https://doi.org/10.1145/3287560.3287586.
[13] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018. URL: https://arxiv.org/abs/1810.01943.
[14] S. Wachter, B. Mittelstadt, C. Russell, Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law, W. Va. L. Rev. 123 (2020) 735. URL: https://heinonline.org/HOL/Page?collection=journals&handle=hein.journals/wvb123&id=764&men_tab=srchresults.