A Fair Selective Classifier to Put Humans in the Loop

Daphne Lenders[1,2], Andrea Pugnana[3], Roberto Pellungrini[4], Toon Calders[1,2], Fosca Gianotti[4] and Dino Pedreschi[3]

[1]*Adrem Data Lab, University of Antwerp, Antwerp, Belgium*
[2]*DigiTax, University of Antwerp, Antwerp, Belgium*
[3]*KDD Lab, University of Pisa, Pisa, Italy*
[4]*KDD Lab, Scuola Normale Superiore, Pisa, Italy*

Abstract

In this paper we propose a practical human-in-the-loop approach for algorithmic fairness, utilizing the selective classification framework. We describe a classification model that abstains from making predictions in cases of unfairness or uncertainty. Any rejected predictions can be passed on to a human expert, to review the possible unfairness issues and make the decisions more just.

Keywords

Fair Classification, Selective Classification, Human in the loop

1. Introduction

Fairness in automated decision-making tasks has been an ongoing research area for the last 15 years. While so far fairness has often been treated as a mathematical notation to be optimized, recently more attention has been paid to its highly context-dependent nature. Computer Science and legal scholars have argued that the fairness of an entire system cannot be expressed through a single number, but that instead a system's fairness should be assessed by studying where unfairness occurs, which subgroups are affected by it and in which cases any disparate treatment might be justifiable [1, 2, 3]. Hence, also to improve the fairness of a decision-making model, optimizing for one single fairness notion is not sufficient and instead, one should take a context-dependent approach and fix unfairness where it occurs [1, 3]. Since it is difficult to automatize such nuanced considerations the call for having a human expert, with sufficient knowledge about a domain, is growing. This call is backed by AI legislation, with the recently passed EU AI Act stating in its Article 14 that any "high-risk" AI system should be overseen and adaptable by a human, to minimise any risks that might otherwise be posed by the system [4]. While the necessity for having a human-in-the-loop is clear, no practical guidelines are given about how a human could oversee a decision-making process, especially if a system makes decisions for thousands of individuals that cannot all be manually reviewed.

In our paper, we propose to utilize the framework of selective classification [5]. Selective classification allows building classifiers that can refrain from predicting when not confident enough. This allows one to trade off predictive performance for coverage, i.e. the percentage of

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany
daphne.lenders@uantwerpen.be (D. Lenders)

instances for which a prediction is provided. In our paper, we extend the selective classification framework to take into account not only the uncertainty around a prediction but also its unfairness. Possibly unfair instances can then be passed on to human experts for review. Our selective classifier also provides explanations for why predictions are perceived as unfair, which can further help experts in making more well-informed decisions.

2. Methodology

In this section, we illustrate our selective classifier on the running example of the “income dataset" [6]. This is a dataset, consisting of information about individuals' working life, their education, and their demographics. The associated classification task is to predict peoples' income level, specifically, if it is above 50K a year. To have a clearer idea of how this prediction task may be relevant in real life, imagine a bank using the “high income" prediction as a proxy for whether an individual has the financial means to pay back a loan.

In this task, we consider the attributes “sex" and “race" as sensitive information, that may serve as grounds for illegal discrimination. In this data the possible values for sex are “male" and “female" and the possible values for race are “white", “black" and “other". A standard classification model, that we have trained on this data, performs best on the reference group of white men while other groups are at higher risk of unfair treatment. This unfair treatment holds in terms of the model's predictions and its errors. In other words, the ratio of positive decision outcomes (high income) is considerably lower for non-white-men than for white men. Also, the “False Negative" errors for these groups are severe, meaning that even if individuals from these groups have a high income, the classifier is likely to predict a low income for them.

We have created a selective classification model that can increase the fairness and the accuracy of such a base classifier, by not making predictions for a) individuals that it is not certain about and b) for individuals that it is certain about, but where the prediction is biased.

In Figure 1 we describe the basic intuition of how this selective classification model works. It consists of a base classifier, that makes an initial prediction for an instance and a rejector, that decides whether to keep, reject or intervene on this prediction. To do so, it receives an instance along with the associated prediction label and prediction probability and first analyses the label's fairness on a global and local level. For the first, it checks if the instance and prediction fall under any global patterns of unfairness that have been established in the base classifier, using the methodology of *possibly-discriminated subgroups* by [7]. For the local fairness check, the *Situation Testing* algorithm is executed, and the prediction label for the instance in question is compared to the labels of similar instances in the dataset [8]. If both the global and local fairness checks fail, we consider the prediction label to be unfair. The rejector then takes the prediction probability of the base classifier as a proxy for its certainty, and depending on that performs a fairness intervention or abstains from making a decision. If the certainty of the prediction is below a certain threshold, a fairness intervention is performed; otherwise, it rejects the original prediction. The reasoning behind the fairness intervention is that an uncertain and unfair prediction is likely to be inaccurate, and it is safe to alter it. In the case in which the rejector has not deemed a predicted label as unfair, it may still abstain from predicting in case the prediction probability falls under a dedicated threshold. Thus, the rejector rejects fair but

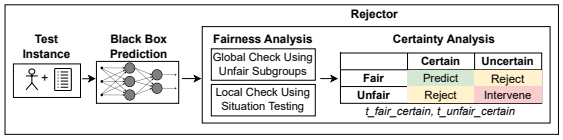

Figure 1: Basic illustration of our selective classification framework.

uncertain predictions and only keeps the original prediction if it is both fair and certain.

Illustrative Example To better illustrate the idea behind our selective classifier, we will go through an illustrative example, showing how our selective classifier rejects the base prediction for a woman, at risk of discrimination. In Figure 2 we see that a baseline classifier needs to make a prediction for a married woman, aged between 60-69 years and working in the Sales sector. The baseline classifier predicts that she has a low income with a probability of 74.17%. To decide whether to keep this prediction our rejector first assesses if this prediction falls under any global patterns of unfairness. To do so, it has a list of subgroups that the classifier is known to behave unfairly on [7]. In this case, our instance falls under the subgroup of women, aged between 60 and 69, working in the Sales sector. This subgroup is deemed to be at risk of discrimination because, the classifier is known to predict a low income for them 90% of the times, compared to only 40% of times for the same subgroup who are *not* female. Because of the high difference, the prediction has failed the first global fairness check, and a local fairness check is performed. In this case, the 3 most similar instances from the reference group of white men, and the non-reference group are selected and their positive label ratios are compared to each other. This allows the rejector to not just focus on the classifier's behaviour on individuals working in Sales and aged between 60 and 69; but also take other relevant characteristics, like their education and their amount of working hours into account. Because even on this fine-grained analysis, the reference group receives more favourable treatment (2/3 positive labels compared to 0/3), the individual fairness check fails. The overall prediction is therefore deemed unfair, and the rejector needs to decide whether to perform a fairness intervention or reject the prediction. To do so, it checks if the prediction probability of 74.17% falls above some threshold (which is learned in a separate step not described in this methodology) of certainty. Since in this case, it does, we fall into the case of an unfair but certain prediction, and the rejector rejects the originally predicted label.

In a next step, the instance could be passed on to a human expert who can review the decision in more detail. The rejector's global and local fairness analysis can help in making the rejection process more transparent and let human experts make more well-informed decisions.

3. Preliminary Results & Discussion

In Table 1 we show some preliminary results of applying our fair selective classifier on the income dataset. We compare the performance of a full coverage classifier (BC), with the performance over all non-rejected instances of a regular uncertainty-based classifier (USC) and our fair selective classifier (FSC). Both selective classifiers had a coverage of 80%, meaning they could reject 20% of the instances. All performances are averaged over 10 test sets. Regarding

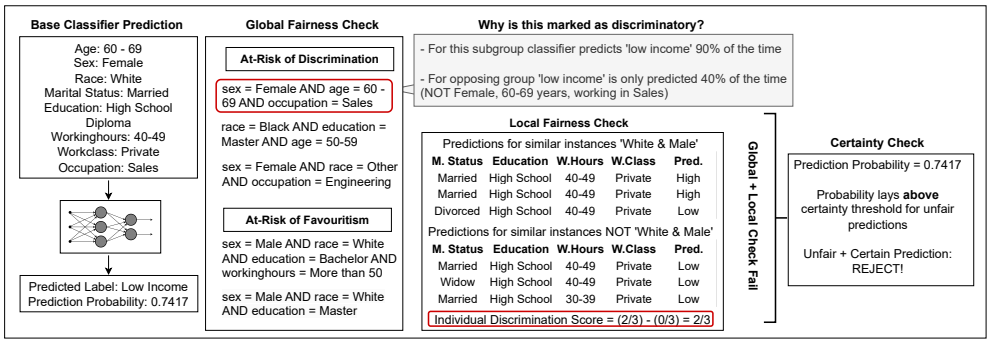

Figure 2: An example of how a prediction of a base-classifier is rejected, because of unfairness concerns.

Table 1
Performances of a baseline classifier, a regular selective classifier and our fair selective classifier

		All			M. Wh.	F. Wh.	M. Bl.	F. Bl.	M. Oth.	F. Oth.
Accuracy	BC	.78 ± .01	**FNR**	BC	.33±.03	.57±.03	.57±.09	.60±.11	.44±.18	.59±.22
	USC	.83 ± .01		CSC	.26±.03	.54±.04	.61±.11	.67±.10	.30±.18	.54±.26
	FSC	.80 ± .01		FSC	.37±.04	.44±.06	.57±.08	.49±.11	.41±.17	.52±.25
Precision	BC	.65 ± .03	**FPR**	BC	.24±.03	.10±.01	.12±.04	.05±.01	.08±.07	.05±.05
	USC	.69 ± .03		USC	.20±.03	.06±.01	.07±.03	.02±.01	.07±.08	.03±.04
	FSC	.64 ± .03		FSC	.18±.03	.11±.01	.10±.04	.04±.02	.08±.07	.05±.05
Recall	BC	.57 ± .02	**Pos. Ratio**	BC	.43±.02	.17±.01	.17±.03	.09±.01	.18±.07	.13±.07
	USC	.62 ± .02		USC	.43±.03	.13±.01	.12±.03	.05±.02	.16±.07	.10±.07
	FSC	.59 ± .04		FSC	.36±.02	.20±.01	.16±.03	.09±.02	.17±.08	.15±.07

overall performance, we see that both selective classifiers manage to increase the performance of the base classifier, by abstaining from some of its decisions. The performance increase of the uncertainty-based classifier is higher, but when focussing on the False Positive-, False Negative and Positive Decision Ratios over different demographics, we see that this comes at the cost of fairness. Regarding all measures, a non-selective baseline has high differences between the group of white men and other groups. In many cases, this difference is increased with the uncertainty-based classifier: it mostly improves the performance for the reference group, but decreases it for others. Our selective classifier manages to make the performance measures over all groups more equal, decreasing the models errors in regards to its False Positive Rates for white men, and its False Negative Rates for other groups. This also results in less differences in Positive Decision Ratios across demographics. While the results of our method are not perfect, and the error rates on e.g. the group of black men are still quite high, we believe that a human-in-the-loop can further enhance the fairness of the system: for instance, they can equalize the positive decision ratios by giving more positive decision labels over all rejected instances of minority groups. Further, they could improve the system, by embedding their domain knowledge in it; e.g. specifying known subgroups at risk of discrimination (used for the global fairness check), that might have been missed by our system.

Acknowledgments D. Lenders and T. Calders were funded by Digitax Centre of Excellence UAntwerp and by Research Foundation Flanders under FWO file number: V467123N A. Pugnana and R. Pellungrini and D. Pedreschi and F. Giannotti have received funding by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme, ERC-2018-ADG G.A. 834756 "XAI: Science and technology for the eXplanation of AI decision making" and Prot. IR0000013. This work was also funded by the European Union under Grant Agreement no. 101120763 - TANGO. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The work has also been realised thanks to NextGenerationEU - National Recovery and Resilience Plan, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR000001 3 - Notice n. 3264 of 12/28/2021.

References

[1] D. Lenders, T. Calders, Users' needs in interactive bias auditing tools introducing a requirement checklist and evaluating existing tools, AI and Ethics (2023) 1–29.

[2] S. Costanza-Chock, I. D. Raji, J. Buolamwini, Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1571–1583.

[3] 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.

[4] The European Commission, The EU Artificial Intelligence Act - Article 14, 2023. https://artificialintelligenceact.com/title-iii/chapter-2/article-14/.

[5] K. Hendrickx, L. Perini, D. V. der Plas, W. Meert, J. Davis, Machine learning with a reject option: A survey, ArXiv abs/2107.11277 (2021). URL: https://api.semanticscholar.org/CorpusID:236318084.

[6] F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning, Advances in neural information processing systems 34 (2021) 6478–6490.

[7] D. Pedreschi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 560–568.

[8] B. T. Luong, S. Ruggieri, F. Turini, k-nn as an implementation of situation testing for discrimination discovery and prevention, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 502–510.