A New Benchmarking Dataset for Fair ML

Daphne Lenders[1] [0000−0002−9839−9077] and Toon Calders[1] [0000−0002−4943−6978]

University of Antwerp, Campus Middelheim, Belgium
{daphne.lenders,toon.calders}@uantwerpen.be

Abstract. In our demonstration we present a new benchmarking dataset, that can be used to evaluate the effectiveness of fair Machine Learning algorithms. This dataset contains both a fair and a biased version of its decision labels, where the former one present the labels individuals actually deserve, while the latter ones are the labels as obtained through a biased decision process, where the decision makers were biased against men. Through this dataset it is possible to test the real-life performance of fair algorithms, by seeing how well they can infer the fair labels after being trained on the biased ones. In our demo we describe how we obtained this dataset. Further, we illustrate how the dataset can be used to test fair ML interventions and how this leads to new insights about the effectiveness of such interventions.

Keywords: Fair ML · Fairness Evaluation · Benchmarking Dataset

1 Background

In recent years, interest in the field of fair Machine Learning has risen considerably. One big ground for discussion in this area is how to evaluate the effectiveness of fairness interventions, targeted to make non-discriminatory decisions in areas like hiring or loan allocations. If one fair ML algorithm decides that a job applicant should not be invited to an interview, but the other intervention decides the opposite, there is no clear way to tell which of either decisions is more fair.

To provide an objective way for evaluating fairness interventions, researchers have proposed to assume the existence of a “fair" and “biased" version of decision labels in data [3] [6]. The biased labels show which decision outcome each individual *received*, while the fair labels show which outcomes each individual *deserves* if discrimination would not exist. If information about the biased and fair label is available, it is possible to evaluate fair ML algorithms by training them on the biased labels, and checking how accurately they can predict the fair ones. Because no realistic dataset with a fair and biased version of of its labels is available, researchers so far simulated such data themselves [2] [6] [7]. However, this approach comes with the downside that simulated data is unlikely to mirror the complex dynamics behind real people and the biases they may face. Thus, any experimental results on synthetic data may not generalize to real-life.

In this demo we will present a new realistic dataset, with a fair and biased version of its labels, which overcomes this shortcoming.

2 The Dataset

We collected our dataset, by starting from an existing dataset with information about students, their free time and study behaviour, and their grade for an exam[1]. We made the assumption that the current decision label; i.e., whether students pass or fail the exam, is fair in a sense that every student had the chance to prove their capabilities on it (which is fundamentally different from e.g. hiring decisions, where job applicants who do not get hired, do not get the chance to prove whether they would have performed well in the job). Further, we assumed that we could introduce bias to these decision labels, by setting up an experiment where participants are prompted to make grade predictions for the students, based on limited information about them. In particular, we expected them to be biased against male students in their predictions, as there are many stereotypes about boys being less mature and more lazy throughout high school [1]. After a proof-of-concept study revealed that participants indeed had inherent bias against male students, we set up an experiment where we collected a biased decision label for each student. The complete dataset with both its fair and biased version of its labels is already available on kaggle[2].

3 The Demonstration

After giving some general background information on fair ML and the idea behind our dataset, we will start our demonstration by showing how the dataset can be downloaded and by explaining which information it consists of. Afterwards, we will show some pre-processing steps that can/should be applied on the data. In the main part of the demonstration, we will use our dataset to evaluate the effectiveness of two well-known fair Machine Learning algorithms, namely “Massaging" [4] and “Situation Testing" [5]. Their performance will be compared to two baselines. To end the demonstration, we will highlight some other use cases the dataset can be beneficial for.

Acknowledgements The research surrounding our new dataset was supported by the Flemish Government under the “Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen” programme.

References

1. Brown, C.S., Stone, E.A.: Gender stereotypes and discrimination: How sexism impacts development. Advances in child development and behavior **50**, 105–133 (2016)

[1] This dataset is known as the “Student Alcohol Consumption" dataset and was already publicly available on kaggle: `https://www.kaggle.com/uciml/student-alcohol-consumption`

[2] `https://www.kaggle.com/datasets/daphnelenders/performance-vs-predicted-performance`

2. Fish, B., Kun, J., Lelkes, Á.D.: A confidence-based approach for balancing fairness and accuracy. In: Proceedings of the 2016 SIAM international conference on data mining. pp. 144–152. SIAM (2016)
3. Jiang, H., Nachum, O.: Identifying and correcting label bias in machine learning. In: International Conference on Artificial Intelligence and Statistics. pp. 702–712. PMLR (2020)
4. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd international conference on computer, control and communication. pp. 1–6. IEEE (2009)
5. Luong, B.T., Ruggieri, S., Turini, F.: 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. pp. 502–510 (2011)
6. Wick, M., Panda, S., Tristan, J.B.: Unlocking fairness: a trade-off revisited. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf
7. Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: A causal inference approach. In: IJCAI. vol. 16, pp. 2718–2724 (2016)