A general purpose framework for fairness in job hiring

Sam Vanspringel, Alexandra Cimpean[1], Pieter Libin *[1,2], and Ann Nowé *[1]

[1] Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium,
{sam.vanspringel,ioana.alexandra.cimpean,pieter.libin,ann.nowe}@vub.be
[2] Data Science Institute, UHasselt, Hasselt, Belgium

Abstract. Due to the increasing popularity of machine learning algorithms to automate decision-making, attentiveness with regard to fairness implications is crucial. In this bachelor thesis, a methodology to detect and mitigate bias and unfairness in algorithms is proposed. We surveyed existing fairness notions and their applicability in the context of job hiring scenarios, as well as investigated a selection of pre- and post-processing techniques to mitigate potential bias. A framework capable of dealing with multiple fairness and bias requirements based on distinct problem settings is proposed. Using this framework, we conducted experiments to investigate a series of job hiring scenarios based on realistic populations and highlight unfairness for different biases.

Keywords: Fairness · Supervised machine learning · Bias mitigation

1 Introduction

It becomes increasingly important to ensure that algorithms avoid discriminating against minorities or even specific individuals, as such algorithms play a critical role in various decision-making processes that affect people. To this end, we build a framework that can quantify bias using fairness notions and mitigate said bias [10]. The term fairness cannot be used unambiguously. Different scenarios require different fairness notions, considering distinct criteria [8]. Fairness notions quantify how fair groups or individuals were treated. A criterion for using a certain fairness notion is the availability of the ground truth. As job hiring decisions are inferred from human reasoning, selecting the appropriate fairness notions requires domain expertise to identify if the objective ground truth is (partially) available.

In job hiring, it is important to treat all candidates fairly, when determining who is the best suited for the job. As supervised machine learning-based decision-making systems are increasingly employed in the real world, it is important to assess how fair such systems are. Since such algorithms learn from historical data, the algorithm might track biased signals and may consequently invoke an unfair selection process.

* Supervisor - Equal contribution

2 Fairness and mitigation

We use our framework to investigate the suitability of distinct fairness notions [8], as well as to assess the performance of specific mitigation techniques [5, 7, 9, 3, 6, 2, 11]. If bias is present in the training dataset, pre- or post-processing techniques can be used to mitigate this from the dataset prior to training or by altering the model's prediction afterwards to satisfy fairness constraints [3, 2, 4, 6, 11, 7].

The framework elucidates the shortcomings of certain fairness notions, in a visual manner. For example, the fairness notion *equal opportunity* [8] is unable to detect bias in the training dataset as it relies on the ground truth, which in itself might be biased. Next, our framework illustrates when machine learning algorithms adopt discriminative behaviour in case it is encoded in the dataset they learn from. A pre- or post-processing technique is then recommended to mitigate the bias.

In a scenario that considers gender bias, we observe that machine learning algorithms encode the bias present in the training data (**Fig. 1**). While applying pre-, in- or post-processing techniques to mitigate this bias can produce a more fair prediction, it is noteworthy that there is a trade-off to be made when considering multiple fairness notions.

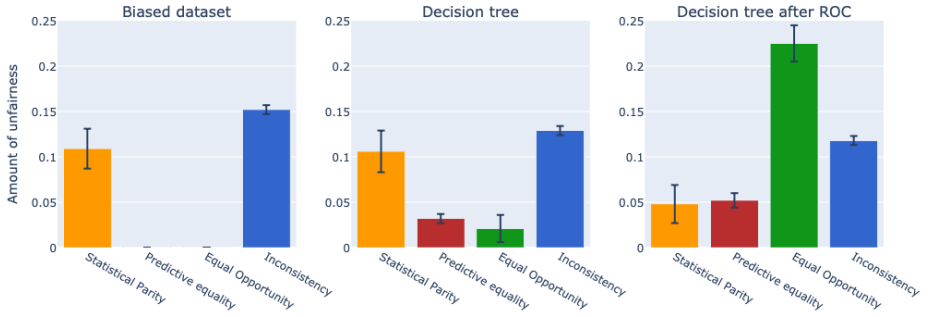

Fig. 1. The fairness notions of a gender-based biased dataset and the decision tree predictions based on the biased dataset before and after mitigation with *reject option classification* [7]. The mean and standard deviations for 20,000 candidates across 100 runs are shown.

The experiments show that distinct scenarios and machine learning models impact the output. We illustrate the need for mitigating techniques that influence the machine learning model on their functionality. Furthermore, we observe the trade-off between satisfying multiple fairness notions, demonstrating the difficulty of achieving total fairness [1]. Therefore, our framework provides tools to detect bias through multiple fairness notions. Additionally, our framework shows an overview of pre- and post-processing techniques that can be applied. Next, we emphasise that trade-offs between the fairness notions need to be considered.

Bibliography

[1] Berk, R.A., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**, 3 – 44 (2018)

[2] Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: 31st International Conference on NIPS. p. 3995–4004 (2017)

[3] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: 21st International Conference on Knowledge Discovery and Data Mining. p. 259–268. KDD '15 (2015). https://doi.org/10.1145/2783258.2783311

[4] Kamiran, F., Calders, T.: Classifying without discriminating. In: 2nd International Conference on Computer, Control and Communication. pp. 1–6 (2009). https://doi.org/10.1109/IC4.2009.4909197

[5] Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012). https://doi.org/10.1007/s10115-011-0463-8, https://doi.org/10.1007/s10115-011-0463-8

[6] Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: 2010 IEEE International Conference on Data Mining. pp. 869–874 (2010). https://doi.org/10.1109/ICDM.2010.50

[7] Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th International Conference on Data Mining. pp. 924–929 (2012). https://doi.org/10.1109/ICDM.2012.45

[8] Makhlouf, K., Zhioua, S., Palamidessi, C.: "On the applicability of ML fairness notions" (2020)

[9] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)

[10] Vanspringel, S., Cimpean, A., Libin, P., Nowé, A.: A general purpose framework for fairness in job hiring. Bachelor thesis, Vrije Universiteit Brussel, Brussels (2023)

[11] Woodworth, B., Gunasekar, S., Ohannessian, M.I., Srebro, N.: Learning non-discriminatory predictors. In: Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 65, pp. 1920–1953. PMLR (2017)