Vision Manuscript

skbh77
Student – Computer Science
Durham University
skbh77@durham.ac.uk

I. Introduction

For this vision manuscript we will discuss and give opinion on a paper titled "Formalizing Fairness in Prediction with Machine Learning" that looks at formalising what fairness in machine learning means [1]. With predictions using machine learning having a growing presence in today's society it is important we understand how to measure and evaluate the fairness of these models, and the paper looks to formalise and critique these methods. We will then go further and discuss the future of bias within AI and highlight any improvements that need to be made.

II. Paper Review.

A. *Importance*

The paper is important because it is crucial to have formalised and strong metrics for measuring fairness. Without these being well defined there will be no ability to evaluate the fairness of machine learning models and correct any bias found within them. Additionally, if bias is incorrectly identified, due to inaccurate notions of fairness, then models could be needlessly made more inaccurate and bias unintentionally added which would be counterproductive. Although having a core set of fairness notions is important, it is also important to continue to evaluate them and improve upon them. The paper takes important steps in suggesting additional fairness formalisations to improve upon the critiques mentioned regarding existing ones. The paper seeks to start a discussion about fairness formalisations, which is important for improving the methods as well as making sure fairness is similarly quantified between different research papers.

B. *Contribution*

The paper makes an interesting contribution by taking existing fairness formalisations and linking them with current fairness notations in social science papers. This contribution is important because it is crucial to make sure that the fairness is not just defined in a machine learning sense but also in a way that fits the complex nature of society. By defining and evaluating seven different fairness formalisations the paper contributes a broad look on the topic whilst also being specific enough each to have a meaningful evaluation. Additionally, by mentioning a multitude of fairness formalisations found in different papers, it serves as a paper for those unsure as to which definition of fairness best suits their project to evaluate all their options. By evaluating each fairness definition using social science literature the paper gives a real-world set of strengths and weaknesses for each that go beyond the scope of many other evaluations that may focus more on accuracy vs fairness, keeping the discussion solely within a computer science framing. Furthermore, the paper goes even further by contributing two notions of fairness found in social science literature that had not yet been discussed in machine learning literature and evaluating if they could be used to improve fairness, these being 'equality of resources' and 'equality of capability of functioning'.

C. *Impact*

The paper does appear to have had a significant impact. According to Google Scholar the paper has been cited 81 times [2]. These papers range from papers discussing the difference between mathematical and human fairness [3], to looking at how fairness impacts people with disabilities [4]. From this we can see that the impact of the paper is far reaching and has helped to further the discussion about fairness metrics. Additionally, the paper is often cited when newer papers talk about short comings of fairness formalisations the paper discusses, so it seems the impact in terms of providing strengths and weaknesses of formalisations is significant.

III. Future of Bias in AI

It can be argued that the identification and mitigation of bias within artificial intelligence has improved dramatically in recent years, with current solutions having a positive impact on the field. However, improvement is still possible and most certainly required. This is because, with increasing use of artificial intelligence and machine learning to make prediction there are an increasing number of bias problems being found. Take for example, COMPAS, a system used to predict the likelihood of an offender reoffending. The system was less likely to mark a white male as a reoffender compared to a black male [5]. This is not the only example of bias within artificial intelligence used in legal context. In 2016, the same year as COMPAS, the United Kingdom deployed a passport photo checking ai despite being aware of the presence of bias [6]. Examples like this are not rare and highlight an issue within the industry, that although methods for mitigating bias are present more work needs to be done in making sure they're used. The fact that bias is involved in machine learning models that have legal consequences for people shows that bias is not yet being taken as seriously as required given its impact. The paper by Gajane and Pechenizkiy [1] goes some length to identifying the strength of current solutions with regards to fairness formalisations but also highlights that there is still improvement to be made, as many definitions of fairness have large weaknesses, such as how a race-blind approach is weak in the long run as opposed to a race conscious approach. When bias is identified there is fair argument to be made that current solutions have a significant impact. Additionally, as there are methods present for pre-processing, such as the one suggested by Feldman et al [7], and for non-bias models, such as Three Naive Bayes [8], there are multiple ways to handle bias depending on the approach required for the project which create suitable improvements. Because of this, I would mainly argue that it is the identification of bias and the definitions of fairness that require the most work, and this improvement can only be made with the combination of computer scientists and social scientists, to be able to create definitions and models which work within the complex nature of society, as backed up in multiple papers [7][9]. As it's with these definitions we can best determine the presence and root cause of bias within models and tackle it.

REFERENCES

[1] Pratik Gajane, Mykola Pechenizkiy. (2021) [1710.03184] On Formalizing Fairness in Prediction with Machine Learning. Retrieved April 14, 2021, from https://arxiv.org/abs/1710.03184

[2] Google Scholar. n.d., Gajane: On formalizing fairness in prediction with... - Google Scholar. Available from: <https://scholar.google.com/scholar?cites=1342272171805998900&as_sdt=2005&sciodt=0,5&hl=en>. [April 15, 2021].

[3] Megha Srivastava, Hoda Heidari, Andreas Krause. n.d., Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning - 3292500.3330664. Available from: <https://dl.acm.org/doi/pdf/10.1145/3292500.3330664?casa_token=XhqpZpfAi3oAAAAA:-Nn3OwoOJsGJU7lfHFbc1tfoj4xoGUHWOGZasfy2IfmWVYX9MEcTtYPQ8_jBmyjcOE1dMn2K_R2hIw>. [April 15, 2021].

[4] Julia Angwin. n.d., Machine Bias — ProPublica. Available from: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. [April 15, 2021].

[5] Maryam Ahmed - BBC News. n.d., UK passport photo checker shows bias against dark-skinned women - BBC News. Available from: <https://www.bbc.co.uk/news/technology-54349538>. [April 15, 2021].

[6] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. (2021) 1412.3756.pdf. Retrieved April 15, 2021, from https://arxiv.org/pdf/1412.3756.pdf

[7] Toon Calders, Sicco Verwer. n.d., Three naive Bayes approaches for discrimination-free classification. Available from: <https://www.researchgate.net/publication/220451718_Three_naive_Bayes_approaches_for_discrimination-free_classification>. [April 15, 2021].

[8] Geoffrey Irving, Amanda Askell. n.d., AI Safety Needs Social Scientists. Available from: <https://distill.pub/2019/safety-needs-social-scientists/>. [April 15, 2021].