Algorithmic Fairness Final Report

MORPH 2020

Borovkova Kseniia

Contents

1 Introduction

In the recent years, there have arisen major concerns about the fairness of the widely used Machine Learning algorithms. Bias was discovered in real-world applications that affected thousands of human lives. Algorithmic fairness has gained recognition and attention from the leading companies and researchers, and numerous advances have been made.[2] Unbiased and fair algorithms are extremely important since the use of technology is rapidly growing, more decisions are being made with its help, and therefore there are more people suffering from unjust and biased Machine Learning algorithms. A major breakthrough was made in 2011, when "Fairness through Awareness" came out, resulting in a flow of further, massive research. Natural language processing, classification, computer vision have all been found to have significant fairness issues present. In this report I aim to give an overview of current bias-detection and bias-mitigation pratices in these areas.

2 Aware and Unaware Approaches

There are two distinct approaches for managing bias: unaware approach and aware approach.

The Unaware approach is based on the concept of a "sensitive attribute/feature", a feature in the training dataset that might be discriminated against, for example gender or race. In this approach, the sensitive attribute is completely ignored, deleted or not included in the dataset, which is insufficient for various reasons. First, there can be proxy features in the dataset that give indirect information about the sensitive attribute. For example, if a candidate for some job attended Barnard College, odds are that this candidate is not male.

On the other hand, the aware approach takes sensitive attributes into account and works with them. It was first introduced in the "Fairness through Awareness" paper by Cynthia Dwork et al.[4] The paper provides fairness definitions and bias mitigation methods, specifically, introduces Individual and Group Fairness.

3 Individual and Group Fairness

There are different definitions of fairness, as well as approaches to ensuring it. For classification problems, two of such approaches are Individual and Group Fairness.

3.1 Individual Fairness

The key idea of this approach, given a binary classification problem, is to treat similar individuals similarly. Therefore, there is a need to act within the particular task at hand.

Definition 3.1. Any two individuals who are similar according to a specific task at hand **should be classified similarly**.

In order to understand whether and how similar are two individuals, we need to use a distance metric. First, we have to find the similarity/distance between the individuals,$d(x, y)$, where d is the distance metric and x, y are the characteristics of the two individuals. That can be done through

graphical distance measurement, features in the dataset, and other techniques However, once we have this information, we need to define how fair the algorithm's output for these individuals is. Similarity is measured between distributions of the outcomes, $D(M(x), M(y))$, where $M(x), M(y)$ are mappings from individuals over outcomes. Then fairness in a certain algorithms is defined as follows:

Definition 3.2. Any two individuals x, y that are at distance $d(x, y)[0, 1]$ map to distributions $M(x)$ and $M(y)$, respectively, such that the statistical distance between $M(x)$ and $M(y)$ is at most $d(x, y)$

Which in other words means that the difference of the outcomes that were returned for the two individuals cannot exceed the initial difference between them. Lipschitz condition:

$$D(M(x), M(y)) \leq d(x, y) \tag{1}$$

However, it is also possible to relax the condition, for example, by setting a threshold:

$$D(M(x), M(y)) - d(x, y) \leq e \tag{2}$$

Where e is the fairness threshold.

3.2 Group Fairness

Group fairness aims to treat and perform actions on groups of individuals, and ensure fairness according to the group. There exist several approaches to that: Statistical Parity, Predictive Rate Parity and Equalized Odds/Equality of Opportunity.

3.2.1 Predictive Rate Parity

To be considered fair according to Predictive Rate Parity, individuals from both groups should have an equal chance for success given the outcomes. For example, if a binary classification algorithms decided whether to accept/reject a student, applicants from both groups (the privileged and discriminated) ideally must have the same probability of getting accepted, or rejected. If C is whether the applicant is actually qualified enough ($c = 1$) or not ($c = 0$), which is determined through past experience/interviews/test, etc., and Y is the decision to accept ($y = 1$) or reject ($y = 0$), $A = a_1$ if the individual is in the privileged group, and $A = a_2$ if the individual is in the discriminated group, then:

$$P_{a_1}[C = c|Y = y] = P_{a_2}[C = c|Y = y] \tag{3}$$

3.2.2 Equalized Odds/Equality of Opportunity

To satisfy Equalized Odds, individuals from both groups should have an equal chance for getting hired/rejected, if they are equally qualified. Therefore, the main idea of the Equalized Odds approach is to accept the same percentage of individuals from the qualified subsets of both groups. Given the same notation as before:

$$P_{a_1}[Y = y|C = c] = P_{a_2}[Y = y|C = c] \tag{4}$$

3.2.3 Statistical Parity

The idea of Statistical Parity is to accept/reject a certain percentage of people in both groups that matches their percentage in the overall demographic. One way to do it is to accept the same percentage of people from both groups:

$$P[Y = y|A = a_1] = P[Y = y|A = a_2] \tag{5}$$

However, there are several issues with this approach which are discussed in detail in [4]. Briefly put, the outcome might be considered "fair" but individuals will still be discriminated. For example, in a set of individuals S, most skilled students study Engineering. In a different set T, however, most skilled students study Marketing. A hypothetical company wants to hire students interested in Marketing and wishes to satisfy Statistical Parity. It would hire the not-the-most-skilled subset of S, and the most-skilled subset of T, which will in turn result in self-fulfilling prophecy, when unqualified members of S could be hired to justify future discrimination against S. This arises further, more ethical, concerns: is it moral to determine how "bright" and "talented" students are, and hire only the "brightest" ones, not taking into account special circumstances/their background? And if so, how can then the "not the brightest" students reach their potential, improve and gain experience? Such questions cannot be answered with Mathematics, so I would encourage Philosophers, Ethicists, Educators, and other qualified professionals to try to answer them.

4 Counterfactual Fairness

Counterfactual Fairness [6] is a completely different approach to defining fairness. It is based on the concept of a "counterfactual", which is an event that would have happened under different, often reversed, circumstances. Usually the sensitive feature is reversed, and the definition of fairness becomes the following:

Definition 4.1. A classifier is counterfactually fair if it returns the same decision with the sensitive attribute reversed.

If Y is an output, is the counter-output, A is the sensitive feature,X are all other features then:

$$P(\hat{Y}_{A\leftarrow a} = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a_0} = y|X = x, A = a) \tag{6}$$

For example, given the same problem as before, if a classifier returned 0 (reject) for a candidate with sensitive feature $A = 1$, then in order to be considered counterfactually fair, it too has be return 0 for a candidate with the same set of other features (GPA, interview score, etc.) but this time with the sensitive attribute A (A is now 0). This approach can also be applied to detect bias of crowd-workers [5] which can often lead to constructing discriminating datasets, and result in spread of the bias even further.

5 Fairness in AI

During the past several years, numerous discoveries have been made in the field. In this section, I would like to briefly describe two breakthroughs that stood out to me personally: "Gender Shades" [3] and "Man is to Computer Programmer is What Woman is to Homemaker?" [2].

5.1 Gender Shades

In their innovative work, Joy Buolamwini and Timnit Gebru draw attention to unfairness in Computer Vision, discuss possible reasons for that. One reason they highlight is biased/unbalanced image datasets, where the majority of images contain males and white people, which results in the image recognition algorithms performing poorly on colored and female individuals. Buolamwini and Gebru introduce a dataset that has an equal number of images of darker females, darker males, as well as lighter females and lighter males. They then compare the accuracy of classifiers trained on popular unbalanced datasets to the same classifiers trained on their dataset, and the accuracy is certainly higher. This research paper played an important role in encouraging and drawing the industry's attention to biases in Computer Vision. These problems are not yet solved, however, they are undoubtedly closer to being resolved than ever before.

5.2 Man is to Computer Programmer is What a Woman is to Homemaker?

This paper was an alarm-raiser in Natural Language Processing. Issues that arose in various NLP-driven applications were because of the biased word-datasets. In order to measure and mitigate bias in word-embeddings, they define the "gender-subspace", which is the difference between two historically female/male words("she" and "he" etc.), and gender direction, and then compute the distance between a given word's vector and the gender direction. That would be the "gender score". In order to mitigate the existing bias, they adjust and re-embed the vectors so that they are either completely neutral and have no gender meaning, or remain with the desired amount of gender context to them. This paper offered a groundbreaking method for mitigating and measuring bias in word-embeddings. Thanks to it, now there is much more awareness of the issues in NLP, and more research aimed at fighting bias.

6 Bias-detection and bias-mitigation toolkits

6.1 AI Fairness 360

An IBM-produced toolkit, AI Fairness 360 [1] provides numerous methods both for bias-detection and bias-mitigation. For assessing existing bias in the dataset, there are several options:

1. Statistical Parity Difference, which is computed as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.
2. Equal Opportunity Difference, which is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group.
3. Average Odds Difference, computed as average difference of false positive rate (false positives / negatives) and true positive rate (true positives / positives) between unprivileged and privileged groups.
4. Disparate Impact, computed as the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.

5. Theil Index, computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 1. It measures the inequality in benefit allocation for individuals.

There are methods to mitigate bias during diiferent stages of implementing the model:

1. Pre-processing: distributes and adjusts the weights after training, but before classification, in order to ensure fairness and give more weight to specific attributes.
2. In-processing: learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions.
3. Post-processing - uses a linear program to decide whether to change the classification labels for an instance, provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups to equalize them.

6.2 Fairlearn

Fairlearn [7] is a fairness toolkit developed by Microsoft which focuses on quality-of-service and allocative harms. A major advantage of Fairlearn is its interactive visualization dashboard, which makes understanding the algorithm's performance easier. Its features include:

- Showing how a group can be negatively effected by an algorithm.
- Comparing the fairness of multiple models.
- Evaluating either classification or regression models.
- Using fairness metrics such as demographic parity, equalized odds, and worst-case accuracy rate.

It also has algorithms that can mitigate unfairness in a model. Fairlearn has one post-processing model, and two reduction models. Their post-processing algorithms will alter the output in regards to certain fairness metrics, and their reduction algorithm retrains the model with the constraints implied by the fairness requirements.

6.3 What-If

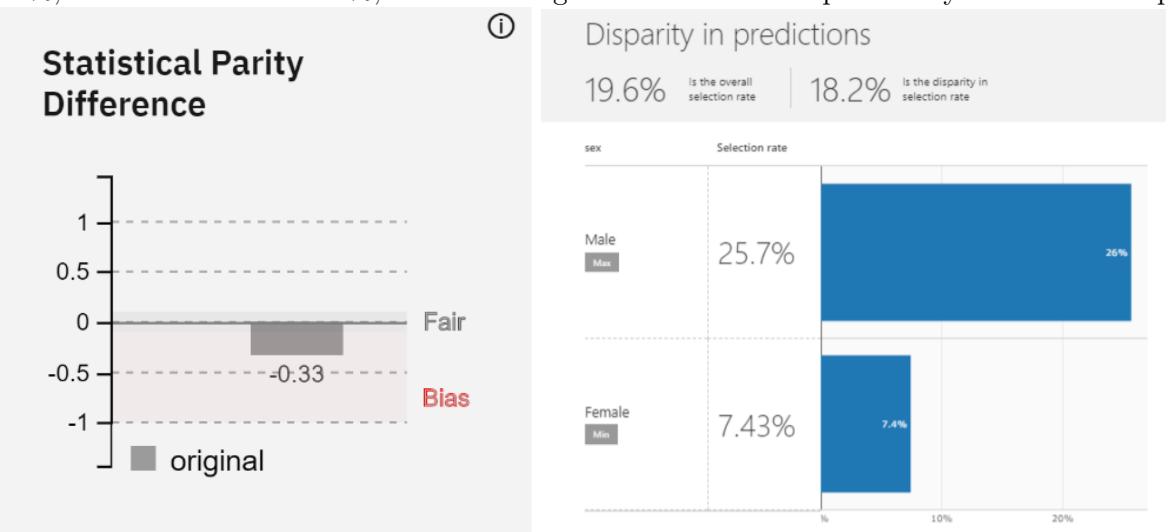
Another toolkit presented by Google, What-If-Tool, takes a completely different approach and suggests to mitigate bias using Counterfactuals, and conducts intersectional analysis of subgroups. It is possible to edit individual data points (individuals/instances) to see how it affects the model at large. The main idea is to invert the counter factual examples, and manually analyze the results. The main disadvantage of this toolkit is that the bias-assessing methods are not extensive and automated, and a developer would need to examine the results on their own, which does not seem to add accessibility to the toolkit. However, the main approach is certainly innovative, since Counterfactual Fairness is an effective technique that, unfortunately, has not yet gained wide recognition in the industry.

6.4 Fairlearn VS. AI Fairness 360

Since both Fairlearn and AI360 toolkits have several bias-assessment methods in common, it seems reasonable to compare them to each other. We chose the Statistical Parity Difference, where the

sensitive attribute with respect to which the parity was assessed was gender. The dataset chosen to be evaluated was Adult Data Set, also known as "Census Income" dataset, and the task would be to predict whether income exceeds 50K dollars/yr based on census data.

For AI360, the parity spectrum goes from -1 to 1 (rather than 0 to 1), so the magnitude of difference of 0.33 should be divided by the total range of 2 to get the percentage, which is 16.5%. As for Fairlearn, we calculate the difference as 25.7% - 7.43%, which is 18.27%. Interestingly, 16.5% and 18.27% are rather far apart, although the bias-assessment method was the same, as well as the data. This can be explained by inner differences in the bias-detection algorithms, and although it is not huge, in some cases it might affect the outcome. For example, if a certain company decides to asses bias present in the dataset using AI360 toolkit, and chooses to set a threshold of 15% (if the bias rate is higher than 15%, then work on mitigating it, otherwise - leave it as it is), AI360 might output 14%, but Fairlearn - 15.77%, then this slight difference could potentially influence real people.



7 Conclusions and Personal Opinion

All approaches have their own advantages and disadvantages. However, the one that is closest to my idea of ensuring fairness is Counterfactual Fairness, since it acknowledges the uniqueness of each individual, and acts in a somewhat personalized manner. On the other hand, however, it might be hard and time-consuming to implement in real life. As technology influences our world more and more, it is becoming our major priority to ensure its fairness towards people. Although this topic is not widely discussed in the media, it is steadily gaining recognition and interest from all fields. There is much left to uncover about the data, the algorithms, the ways of discrimination, even with the existing research. Now that the world is becoming as connected as ever, it is getting harder to divide arts and humanities from more technical fields, and vise versa. That is why we should raise awareness and spread knowledge about bias in algorithms: to encourage interest and interdisciplinary thinking of young professionals. It also works in the opposite direction: if we teach future generations of developers and engineers about ethics and history, they might become more concerned about aspects of Machine Learning algorithms other than mere accuracy and speed.

References

[1] Rachel K. E. Bellamy et al. *AI FAIRNESS 360: AN EXTENSIBLE TOOLKIT FOR DETECTING, UNDERSTANDING, AND MITIGATING UNWANTED ALGORITHMIC BIAS*. 2018.

[2] T. Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *ArXiv* (2016).

[3] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *ArXiv* (2018).

[4] Cynthia Dwork et al. "Fairness through Awareness". In: *ArXiv* (2011).

[5] Bhavya Ghai et al. "Measuring Social Biases of Crowd Workers using Counterfactual Queries". In: *ArXiv* (2020).

[6] Matt J. Kusner et al. "Counterfactual Fairness". In: *ArXiv* (2017).

[7] Microsoft. "Fairlearn: A toolkit for assessing and improving fairness in AI". In: (2020).