Methods to Reduce Bias in Machine Learning Applications

CS4991 Capstone Report, 2023

Henry Todd
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, VA USA
ht6xd@virginia.edu

ABSTRACT
Machine learning (ML) algorithms inherit characteristics of the dataset from which they are trained, leading to potential for unfairness in classification as a result of bias. To reduce unfairness of ML algorithms this report outlines methods for minimizing the effects of bias. Methods include collecting unbiased data and altering or removing existing data in order to reduce bias in preexisting datasets. Additionally, algorithms should provide context to underlying data alongside outputs to minimize the influence of bias. The proposed methods should result in a black-box framework for reducing bias which can be applied generally to ML. Additional work is needed on in-process bias reduction, tools for identifying bias, and best practices for data collection.

1. INTRODUCTION

Whether it is online shopping, scrolling through Netflix, or your smartphone's facial recognition, artificial intelligence has become a part of our everyday lives. While the technology has allowed for many exciting innovations, for example the possibility of self-driving cars, it comes with inherent limitations that left unchecked can lead to systematic unfairness. One key limitation is how "artificial intelligence algorithms rely on big data… [which] is incomplete in some levels" [1]. When artificial intelligence uses incomplete or biased information unchecked, the issues are "strengthened and amplified" [1].

The amplification of bias in artificial intelligence can lead to misclassifications that disproportionately affect certain groups of people [1,2]. As researcher Zhang (2020) describes, "the needs and wishes of [some] disadvantaged groups [may not be] reflected in the final calculation results" if "data shielding" occurs during data collection [1].

Although a bad movie recommendation does not result in any severe consequences, the seriousness of misclassifications become clear when considering other applications of artificial intelligence. One such example is the use of facial recognition in automated detection of criminality [3]. The software proposed by Wu and Zhang attempts to identify the likelihood of committing crime, determining IQ, and other traits simply from one's face. The importance of accurate classification in such an example is paramount considering how the result of biased classification towards any minority group would be socially catastrophic.

Although the need for minimizing bias in artificial intelligence is clear, the means by which bias can be eliminated are not. The issue is with the sheer magnitude of data necessary to train machine learning algorithms. Dealing with 'big data' comes with inherent challenges, some of which Gudivada et al. [2015] list "how to capture… clean, analyze, filter" the data. Two means of minimizing bias are: 1) by initially constructing unbiased datasets; and 2) minimizing bias by removing data from preexisting datasets. Both of these methods are nontrivial, as they both require either capture or cleaning, analyzing, and filtering of data. The handling of big data "require[s] making several tradeoffs among desired scalability, availability, performance, and security" [4]. The methods and results sections of this report will detail proposed means of handling big data such that bias might be minimized.

2. RELATED WORKS

Benjamin [2019] discusses how racial inequities can be perpetuated by artificial intelligence algorithms, citing "BeautyAI," an algorithm which selected what it deemed the most beautiful faces out of all submitted in a competition. The outcome saw a majority of light-skinned faces deemed most beautiful, with only 6 out of the 44 winners being non-white. This case study informed the need for research to minimize unfairness in machine learning. She describes how "social bias embedded in technological artifacts" can have "the allure of objectivity without public accountability," thus perpetuating unfair treatment of minorities without restriction. This report seeks to outline methods for reducing the potential mistreatment of minorities through algorithmic unfairness.

This report proposes to improve the collection and adjustment of datasets using methods from statistics, the use of which for artificial intelligence are introduced by Friedrich et al. [2022]. They discuss the relationship between artificial intelligence and statistics and indicate the necessity of statistical methodology for "assessment of data quality and data collection," as well as "assessment of uncertainty in results." I propose the use of statistical methods as described to collect unbiased data or improve preexisting biased data, in addition to using assessments of uncertainty to provide context to the output of AI algorithms where removal of bias is not possible.

Mehrabi et al. [2021] define bias and algorithmic unfairness, as well as outline methods for developing fair machine learning algorithms. These methods consist of "pre-processing, in-processing, and post-processing" of datasets. The methods outlined provide specific context towards machine learning applications where Friedrich et al. and others remain more general. This paper seeks to combine statistical methodology with methods proposed here for mitigating bias to provide additional context to ML classification.

3. PROPOSED METHODS

The methods proposed in this report synthesize current research for increasing fairness in machine learning algorithms. Two key improvements to machine learning fairness are proposed: 1) the development of better datasets specific to machine learning; and 2) increasing the amount of contextual information provided alongside output to give better understanding of output origins.

The former can be achieved by building new datasets following an "interventionist" policy [7], or by the improvement of existing datasets. The latter requires record-keeping through the data collection and maintenance processes, will possible use of statistics for uncertainty estimation.

3.1 Bias in Machine Learning

Unfairness in ML applications can arise from two areas: 1) through biased underlying datasets; and 2) bias inherent in the algorithm itself. As Mehrabi et al. [2021] describe, "biases in data skew what is learned by machine learning algorithms." ML trained on skewed data "can even amplify and perpetuate existing biases" [2]. This amplification leads to worsened effects of bias in society through ML applications.

3.2 Improved Datasets for ML

Due to the dependence of machine learning on its training data, if left unchecked the datasets used can be a major source of bias [2,4,6,7]. The development of unbiased datasets is generally called "pre-processing" [2]. This indicates the reduction of bias prior to its use in ML. As such, pre-processing methods can be applied generally to any ML algorithm. Pre-processing can be conducted via the creation of new datasets following an "interventionist" policy [7], or by the removal of bias-inducing data.

3.2.1 Constructing New Datasets

Friedrich et al. [2022] describe how "AI applications often use data that were collected for a different purpose." These datasets, with what they term "secondary data," as a result may have underlying biases related to their initial purpose. Additionally, when data is collected for AI it is often done so in a "laissez-faire" method, as described by Jo and Gebru [2020], in which mass amounts of data are collected without oversight.

The collection of mass data for AI with little oversight is done both out of convenience as well as "the misconception that 'Big Data'

automatically leads to exact results" [6]. It is easy to mistakenly assume that collecting "enough" data is sufficient to avoid bias. However, "datasets composed without an adequate degree of intervention with replicate biases" [7].

Thus, it is imperative to perform data collection with an interventionist policy. This involves "critical investigation of the motivations and purpose of the data" [7]. Especially when data used for AI is drawn from sources originally uses for other purposes, it is paramount to analyze the biases of those origins, and actively only use data meeting certain criteria.

Data collection for machine learning should also be performed by following a specific policy. To avoid bias, datasets must cover all demographics of a certain population such that no groups are underrepresented [2,6,7]. As Ding et al. [2021] describe, datasets should be "larger" and "more representative [and] reflective of social progress" in order to minimize bias.

Following a policy for data collection can "address gaps in sociocultural diversity and inclusivity" [7]. This helps increase the representivity of datasets required for use in AI. Additionally, in order to avoid biases arising from Simpson's paradox (a form of aggregation bias), data collected should be disaggregated [7]. Friedrich et al. [2022] introduce additional means for identifying if data is "good" for use in AI.

3.2.2 Improving Preexisting Datasets

Two methods have been proposed for removing bias from preexisting datasets. Both methods are considered pre-processing, and as a result are generally applicable to all ML algorithms by treating the implementation as a black-box, in which only the input and output are considered.

The first method, proposed by Li and Vasconcelos [2019], formulates bias removal as an optimization problem. Specifically, dataset resampling is performed in which the ML algorithm "obtain[s] sample points with different frequencies than those of the original distribution" [8]. Resampling can help address underrepresented groups within the dataset. This works by "oversampling minority classes and undersampling majority ones" [8].

By treating the problem of resampling as an optimization problem, using a minimax algorithm following stochastic gradient descent, a common optimization technique, the optimal resampling can be obtained [8].

A second method for improving preexisting datasets involves analyzing and finding which data points are most responsible for causing bias and removing them [9]. This follows the assumption that "some training data is more biased than the rest" [9]. The method defines discriminatory pairs as "similar individuals who receive dissimilar predictions" [9]. By defining an "influence function," which determines the influence of data points on a particular decision, bias can be removed from the dataset by deleting those data most responsible for discriminatory pairs.

3.3 Providing Context to Machine Learning Outputs

While methods for reducing the amount of bias inherent in the underlying dataset are described above, it is impossible to eliminate all bias from the outputs of machine learning algorithms. As a result, it is also important to provide context regarding the dataset with all outputs, such that end-users can gain a better understanding of their results.

3.3.1 Providing Auxiliary Information

In order to continue increasing algorithmic fairness in artificial intelligence, I propose methods for recording and providing auxiliary information as context to machine learning outputs. As discussed in the previous section on constructing new unbiased datasets, following a specific policy for collection is important to reduce certain biases [7]. This is important for contextualizing output as well, as following a policy allows for better documentation of sources of data (and therefore bias) [7,10].

Hutchinson et al. [2021] describe all documentation of data necessary for providing adequate context, including "making apparent what is valued in the data," accounts of dataset design and "justify[ing] the design decisions made," as well as accounts of testing and maintenance of the dataset. This documentation is imperative to provide all necessary context to ML outputs concerning the origins of data and its properties. This is echoed by Jo and Gebru [2020], who describe how the "origin,

motivation, platform, and potential impact" of data should be recorded as it is collected.

The collection of auxiliary information on the data collection and implementation process allows be the production of a "datasheet," which provides information on the potential skews of the data used [2,6,7].

3.3.2 Using Statistics for Uncertainty Approximation

In addition to providing auxiliary data in a datasheet alongside outputs, Friedrich et al. [2022] outline methods from statistics which can be used to provide uncertainty approximations to ML. These estimations provide a level of confidence associated with the classifications of ML algorithms.

With the datasheet described above and the certainty level provided alongside outputs, end-users can take into account for themselves the weight with which they hold outputs.

Methods for uncertainty approximation include "Bayesian approximations, bootstrapping… cross-validation techniques" and others [6]. Additionally, the use of auxiliary models, or "comparatively simple statistical models which… describe the most important patterns" could potentially be used to quantify uncertainty [6].

4. ANTICIPATED RESULTS

The results anticipated from the methods proposed are a general framework for increasing fairness of any machine learning algorithm. The methods accomplish this in two ways. The first is that bias as a whole can be reduced in ML by creating better datasets, either from scratch or by improving previous datasets. The second is to provide contextual information about the underlying dataset to outputs, to give the end-user a better understanding of the potential biases which influence the decision made.

I anticipate that the combination of creating new datasets with the explicit purpose of minimizing bias towards use in machine learning, following an interventionist policy, in addition to performing bias minimizing techniques such as resampling through optimization and elimination of bias inducing data points will cause a non-trivial reduction in bias of the underlying dataset.

However, while bias can be reduced as described above, its effects cannot be eliminated completely. To further improve the fairness of machine learning algorithms, I anticipate that the additional context provided with outputs will allow users to have a better understanding of the confidence with which they can trust decisions made. This can improve fairness by allowing users to refute a decision made, citing potential sources of bias. However, this leaves a substantial burden of avoiding unfair decision making on the user, rather than the designer of the algorithm. Thus, the methods proposed in this paper are not a full solution, but rather an improvement which requires additional consideration.

5. CONCLUSION

As artificial intelligence becomes increasingly present in technology, ensuring fair decision-making by machine learning algorithms is paramount. Failing to account for bias in ML will undoubtedly have consequences on society, especially in cases of protected characteristics such as race or gender. I propose methods for reducing the effect of biased data, as well as means for end-users to take the bias of underlying data into consideration using additional contextual information. These methods in conjunction provide a black-box framework which can be applied to any ML algorithm.

6. FUTURE WORKS

Additional efforts to improve upon the black-box methods proposed include bias reduction at the post-processing level, in which only the outputs are considered for bias.

The best methods for data collection is currently an active research topic. In addition to using an active, interventionist policy for using unbiased data are methods of statistical sampling for disaggregated data such as stratified random sampling, as well as identifying the best specific policies for accepting or rejecting data to be used in the dataset.

Finally, bias arises both from the dataset as well as from the algorithm. In addition to the black-box approaches to reducing bias for all machine learning algorithms described in this report, research is needed for in-processing to reduce the bias developed inherent within the

algorithm for implementation-specific algorithms.

REFERENCES

[1] Kefei Zhang. 2020. The data limitations of artificial intelligence algorithms and the political ethics problems caused by it. *Advances in Intelligent Systems and Computing* 1303, 1 (2020), 438-443. DOI: https://doi.org/10.1007/978-981-33-4572-0_64

[2] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristena Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (Jul. 2021), 1-35. DOI: https://doi.org/10.1145/3457607

[3] Xiaolin Wu, Xi Zhang. 2017. Responses to critiques on machine learning of criminality perceptions. arXiv:1611.04135. Retrieved from https://arxiv.org/abs/1611.04135

[4] Venkat N. Gudivada, Ricardo Baeza-Yates, Vijay V. Raghavan. 2015. Big data: Promises and peril. *Computer* 48, 3 (Mar. 2015), 20-23. DOI: 10.1109/MC.2015.62

[5] Ruha Benjamin. 2019. Race after technology, 49-76. Polity, Cambridge, UK.

[6] S. Friedrich, G. Antes, S. Behr, H. Behr, H. Binder, W. Brannath, F. Dumpert, K. Ickstadt, H. A. Kestler, J. Lederer, H. Leitgöb, M. Pauly, A. Steland, A. Wilhelm, and T. Friede. 2022. Is there a role for statistics in artificial intelligence? *Advances in Data Analysis and Classification* 16 (2022), 823-846. DOI: https://doi.org/10.1007/s11634-021-00455-6

[7] Eun S. Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3351095.3372829

[8] Yi Li and Nuno Vasconcelos. 2019. REPAIR: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 16 - 20, 2019, Long Beach, CA. IEEE, Piscataway, NJ, USA, 9572-9581. https://doi.org/10.48550/arXiv.1904.07911

[9] Sahil Verma, Michael Ernst, and Rene Just. 2021. Removing biased data to improve fairness and accuracy. arXiv:2102.03054. Retrieved from https://arxiv.org/abs/2102.03054

[10] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 560-575. https://doi.org/10.1145/3442188.3445918

[11] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2022. Retiring adult: New datasets for fair machine learning. In *Proceedings of the 2021 Advances in Neural Information Processing System Conference*. https://doi.org/10.48550/arXiv.2108.04884

[12] Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. Mitigating gender bias in machine learning data sets. *Communications in Computer and Information Science*, CCIS 1254. https://doi.org/10.1007/978-3-030-52485-2_2

[13] Tal Feldman and Ashley Peake. 2021. End-to-end bias mitigation: Removing gender bias in deep learning. arXiv:2104.02532. Retrieved from https://arxiv.org/abs/2104.02532