Potential Technical Solutions to Mitigate the Effects of Bias on Machine Learning Algorithms

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Eric Armstrong
Spring, 2021

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Eric Armstrong

Approved _ _____ Date 5/4/2021_____
Hannah Rogers, Department of Engineering and Society

Abstract

With machine learning systems starting to dominate society in many ways, these systems should not be allowed to make biased decisions that discriminate against women and non-white people. Actions must be taken to reduce these biases either through fairness metrics, often defined by large companies, to force the algorithms to make fair decisions, or through the use of conditional generative adversarial networks (cGAN's) to generate data in order to unbias the data before training. The analysis will focus on why these approaches may be better or worse than the other in science and technology in society frameworks of wicked problems and technological fix, by analyzing literature about the new approaches. We must ensure that these metrics are "fair" in the way that we intend and that if the data is "unbiased" in the way we see will help, that other biases might arise or still exist that are harder to find.

Potential Technical Solutions to Mitigate the Effects of Bias on Machine Learning Algorithms

Introduction

With machine learning on the rise within businesses, many problems are being solved that were not feasible before; however, these algorithms depend on very large amounts of data that are often unrepresentative of the population (Manyika, 2019). A recent study on some of the leading facial recognition software found that nearly 100% of white males were correctly identified while non-white females were recognized correctly only 65-79% of the time (Mames, 2018). This bias is not something that is coded into the algorithms, but is a result of how the math-centric algorithms in machine learning models used by top companies such as Microsoft and IBM draw conclusions from biased data (Mames, 2018). There are multiple approaches to addressing this problem, but the two that will be the focus of this STS report are using fairness metrics, such as requiring that models have equal predictive value across groups, and using advanced machine learning techniques to generate new "fake" data that is indistinguishable from real data to offset the biases at the data level. While these methods to solve bias in algorithms and data are a solution to mitigate biased predictions, they are an example of a technological fix for a much larger and long-term societal issue that stems from the history of systemic prejudice against women and minority groups that has only been made easier and automated due to the rise of machine learning. Biased models and data can very easily be viewed as a wicked problem showing that there is not a clear or perfect solution. In our current political climate, race and sex have become political issues that is polarizing the nation, leading to lots of heated controversy about what the best solution may be.

We must introduce technical solutions for reducing bias in machine learning models because minority groups are being discriminated against by the models' automated decision making since the data they are learning from is historically biased against them.

Wicked Problem Framing, Technological Fix and Literature Review

The Wicked Problem Framing and Literature Review methodologies will be used to answer the research question. Beginning with background, context will be given for some major examples of how machine learning has shown clear bias in our society even in recent years with our new complex models. Seager's Wicked Problem Framing technique gathers evidence to reveal relationships between actions and consequences by showing how historical actions of social inequality are affecting modern systems (Seager, Selinger, & Wiek, 2012). Newberry's' Technological Fix proposes that just because a problem exists does not mean that we can fix it or should fix it. Following the background will be how the cause of the bias is not in the algorithms themselves, however, but is due to biases that have existed in society that have merely transferred into the models.

The use of literature review will lead the discussion of different researchers' solutions to mitigate the biases in these algorithms. First, the two main approaches will be introduced, the first being a model-based fairness approach and the second being a data generation approach using conditional GAN's to unbias the data in order to unbias the model. Following that, there will be discussion focused on the benefits and drawbacks of both approaches and analyze how effective they would work in practice. There will also be discussion about research on the explainability of models in an attempt to discover where the biases are coming from within the data that might not be obvious from a human perspective. Finally, I will address some of the limitations with the current technology within society followed by a summary of the report and

what needs to happen going forwards. The sources for the review will primarily be recent research papers, published in the last couple years, as well as the fairness metrics that have been published by major companies.

What is Machine learning?

Everyone has, at least, heard of machine learning through some channel by now, be it the news, social media, or in college classes, but many do not understand even the fundamentals of how machine learning works, which is crucial to understanding why machine learning systems make biased decisions. “Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience” (Iriondo, 2020). Machine learning algorithms learn from experience similarly to humans learning from experience, by repeatedly training on data. In each round of training, the algorithms slowly correct themselves from being told if their predictions are either right or wrong. This can occur hundreds of times and runs until the programmer or data scientist thinks that the algorithm has reached its optimal performance. The way it makes predictions in the first place is based on trends and patterns in the data. One common problem and difference is how machines find patterns in data features versus humans. humans have a general idea of which features could be correlated with other features, however, machines will often find hidden correlations that may or may not be relevant, or that humans would not notice. This is partially related to why machine learning algorithms make biased decisions, particularly against black and non-white people who have been discriminated against by humans for generations, creating lots of biased data (Kleinberg et al., 2020). The problem is very real and is not something that we have time to wait to address. The algorithmic bias is found in many hiring algorithms such as the one used by amazon, which discriminated against women because of their choice of words versus men, to criminal justice

algorithms, which have mislabeled African-American defendants as “high risk” at nearly twice the rate it mislabeled white defendants (Manyika, 2019). Both of these areas are very critical to being unbiased as they significantly change the quality of life for the people in either situation. Despite these biases however, it is important to recognize that these algorithms have lots of potential to make a positive difference with making equitable decisions that are easily traceable to understand how the predictions were made, versus humans who can lie about why a certain decision was made or not made (Kleinberg, 2019).

Overview of Potential Solutions

One potential approach to lowering the bias in these algorithms, which I will cover in the following section, is using GAN’s, which are used to generate new synthetic fair data with selective properties from the original data (Abusitta, et al., 2019). In order to understand why this would be a good solution or not it is important to understand how GAN’s work at a high level. Generative modeling is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset (Brownlee, 2019). We know that a GAN is producing realistic new samples by training a discriminator that tries to determine if a sample is real or fake, and when the discriminator is unable to determine which is which, you know the generator is producing plausible examples and, therefore, done training. One important variation of GAN’s are conditional GAN’s, which allow for data to be generated around a certain feature in the data such as race or sex, which could be used to remove specific biases from data.

Another potential approach to lowering bias in algorithms, that is currently used at major companies such as IBM, is by ensuring that models are “fair” according to a technical definition

determined by various researchers. These metrics could be a range of things, but could be having equal predictive value across different groups or requiring that models have equal false positive and false negative rates across groups. I think it is important to note that this technique to mitigate bias does not change the original data, but only looks at sensitive parts of the data to achieve fair results.

Fairness-Driven Approach

Minority groups are being discriminated against by machine learning algorithms due to the data being historically biased against them, therefore, model level solutions that force these algorithms to make an equal amount of fair decisions need to be introduced.

One type of fairness that is commonly seen as a solution in the industry right now is something known as counterfactual fairness. There are other variations of this such as path specific fairness which could potentially be better for some complicated situations, however, I plan to focus on the basic version of counterfactual fairness. The basic idea of Counterfactual fairness is that decisions made in the real world should be the same as decisions made in the counterfactual world, where an individual belongs to different demographic groups. An example of this is with cars and race. Assume that people with cars that are a certain color, red, are more likely to be aggressive drivers and therefore, insurance companies raise rates accordingly to car color. However, there is a particular race that prefers red cars even though individuals of this race are no more likely to be aggressive drivers. Using counterfactual fairness you would be able to see that this is an unfair way to make this prediction. By changing the race variable and keeping the aggressiveness factor constant, there will be a change in the chance to own a red car and consequently the accident rate, showing it is discriminatory on race (Kusner M., et. al).

Data Generation Approach

Introducing new approaches using conditional generative adversarial networks (GAN's) to reduce the bias in the data that machine learning outcomes are based on is necessary, due to the dominance of machine learning in many automated decisions that are often negatively biased towards minority groups.

The cGAN approach is a potential alternative due to the criticism of the model level fairness metrics that take too long to train due to the amount of time it takes to find parameters to make predictions fair, as well as the fact that they largely degrade the accuracy of the underlying machine learning algorithm. The general idea behind this idea is to mitigate the biases similarly to the fairness metrics, while also enhancing the overall accuracy of the model instead of worsening it. The reason this technique is able to achieve both results is because a GAN generates new data, and machine learning algorithms almost always perform better when there is more data. The cGAN will be used to understand how the data is biased by analyzing how different attempts at mitigation work which will ultimately assist a domain expert in inferring the type and quantity of data that needs to be synthetically sampled in order to augment the training data to mitigate the bias.

Wicked Problem

Bias in machine learning algorithms can be viewed as a wicked problem. Wicked problem definitions depend on the solution and the solutions depend on the problem definition and the problems are often viewed differently by different stakeholders. These factors lead to problems that are never solved definitively. While businesses who own the algorithms might worry about the bias itself and potential lawsuits over discrimination, people ultimately affected by the biased decision making want the algorithms to be fair while also wanted to address the

unlerlying problems of why it was biased in the first place, often due to systemic discrimination against certain groups of people over long periods of time. While companies may think they have solved the bias issue, who is the correct person or persons to judge if the metrics in place are actually fair and do not worsen the bias or make it harder to place blame. Additionally, any new real world data will still be biased due to inherent biases in human decision making that has always existed. Researchers studying the long term effects of not understanding black box models are worried that if machine learning algorithms are learning from their own artificially produced data and if we are not able to understand how their decisions are made that we will slowly be unable to remotely determine how decisions are made, making it impossible to determine what the biases may be. Assuming that fair and acceptable algorithms are able to be produced and the algorithmic bias problem solved, influential businesses and people could start to care less about the human biases in society since more and more things are automated and those are fair, causing new problems to stem from the new solution, just as wicked problems are defined. Finally, who gets to decide if these algorithms are fair enough to be introduced into society? If these decisions are being made by only the most influential people in the process, will the rest of society, particularly the non-privileged, believe that the privileged had their best interests in mind, or did they simply want to roll out these new algorithms as quickly as possible. As of right now this decision would probably be made from someone influential inside each organization providing the machine learning service, whether it be a company standard set by the CEO or CTO, or a team leader.

Technical Fix

These potential technical solutions to mitigating bias are clearly seen as a technical fix as well. A technical fix is a solution to a technical problem that fails to address the underlying

problems, and may be addressing a problem that should not be fixed. While this is almost certainly a problem that should be fixed due to its direct effects in sensitive areas of people's lives, both approaches covered in this report fail to address the underlying problems to some extent. The fairness approach is the biggest example of a technological fix. While it may mitigate the biases, it fails to unbias the data the algorithms learn from and does not address the underlying bias in our society. While using GAN's addresses the data itself and can help understand the derivation of the biases in society, it does nothing to mitigate the bias in society. In his definition of the Technological Fix in the *Encyclopedia of Science, Technology, and Ethics*, Byron Newberry writes "It is the aspect of technological fixes – their tendencies to mask the symptoms of complex social problems without addressing their causes or true costs – that generally evokes ethical concern" (Newberry, 1901). While the solution could alleviate some of the effects of the underlying problem by deferring those decisions to machines, it could, however, cause people to forget about the underlying issues or say they are less important at a key time in their progression.

Counter Arguments and Limitations

While these techniques may be able to solve some of the issues regarding bias in machine learning algorithms there are still limitations with these approaches. One issue with defining fairness is that there are different definitions and these fairness definitions usually cannot be satisfied at the same time (Manyika, 2019). Additionally, regarding counterfactual fairness, it has been shown that just hiding sensitive information and features about the data does not necessarily prevent all biases that exist that are harder to see. Due to these reasons, there is no one perfect technical solution, different parts of society must agree on when predictions are fair enough to be released, or if some systems should not be fully automated and a human should be

kept in the loop, which relates back to the wicked problem definition. In practice this would probably have to be done by the end user of the service. If a legal team was outsourcing some companies' machine learning algorithms to classify people's cases as high risk or low risk, they could get a diverse group of colleagues to compare classifications they have made themselves to the predictions made from the algorithm. Assuming companies were able to explain how these algorithms came to a prediction, experts in that area could analyze which factors played the biggest role in how the algorithm made a prediction and decide if it was fair enough or not.

Ultimately, beyond technical solutions, we need to engage in fact-based conversations around potential human biases, particularly potentially unknown biases found by these new technical solutions. "When we do find bias, it is not enough to change an algorithm—business leaders should also improve the human-driven processes underlying it." (Manyika, 2019). Other limiting factors particularly with fairness are related to how long it may take an algorithm to find the fair parameters to make predictions. As data becomes more biased, the time it takes for an algorithm to make a decision increases. While with some fields this may be acceptable, some fields such as medical diagnostics need to be real time and since this data is known to be quite biased, systems may not be able to make decisions fast enough. Finally, one known limitation of the GAN technique is that it fails with unpredictable bias such as bias against men being underpaid in certain positions at Google (Abusitta, et al., 2019).

Conclusion

One of the biggest challenges that both these approaches try to overcome to ensure fairness or to generate data to counter biases is understanding where the bias exists. With many machine learning and deep learning algorithms there is very little explainability with how an algorithm reaches a particular outcome with a given input. Although these solutions can mitigate

the damage done by biases that have existed in our society and, therefore, data, it is important for users and businesses to see where the biases derive from in order to start new conversations around the most biased aspects in data. Being able to see these biases could also unearth biases we may have not known existed or give data to support claims from afflicted groups, to help drive social change more than ever. Therefore, these conversations must include equal input from diverse fields within not just science, but also social fields to ensure the optimal outcome is reached. It is also important to note that even though these algorithms are biased, their predictions are largely more accurate than that of humans, and humans themselves are also prone to exhibiting bias in their decision making, whereas machine learning models can be traced to determine what factors were important in making a prediction. Going forward, companies should be held to a standard to use the most up to date bias mitigation techniques and that more research needs to be focused on detecting and mitigating biases. Additionally, model explainability techniques need to progress in order to build trust with society so that people know a given model is making its predictions for the right reasons.

References

Abusitta A., & Aïmeur E., & Wahab O. A. (2019). Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems. Retrieved September 29th, 2020, from https://deepai.org/publication/generative-adversarial-networks-for-mitigating-biases-in-machine-learning-systems

American Board of Professional Liability Attorneys (2020). What is Medical Malpractice?. Retrieved from https://www.abpla.org/what-is-malpractice

Chiappa S. (2019). Path-Specific Counterfactual Fairness. Retrieved October 15th, 2020, from https://aaai.org/ojs/index.php/AAAI/article/view/4777

Chouldechova A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Retrieved October 15th, 2020, from https://arxiv.org/pdf/1703.00056.pdf

Card D. (2017). The "black box" metaphor in machine learning. Retrieved October 15th, 2020, from https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0#:~:text=The%20black%20box%20metaphor%20dates,Skinner%20conceptualized%20minds%20in%20general

.

Manyika J. (2019, October 25). What Do We Do About the Biases in AI? Retrieved September 29, 2020, from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Brownlee J. (2019). A Gentle Introduction to Generative Adversarial Networks (GANs). Retrieved October 15th, 2020, from https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

Kleinberg J., & Ludwig J., & Mullainathan S., & Sunstein C. (2019). Discrimination in the Age of Algorithms. Retrieved October 15th, 2020, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669#references-widget

Kleinberg J., & Mullainathan S., & Raghavan M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. Retrieved October 15th, 2020, from https://arxiv.org/abs/1609.05807

Mames. (2018, December 11). Impact of Algorithmic Bias on Society. Retrieved September 30, 2020, from https://blogs.ischool.berkeley.edu/w231/2018/12/11/impact-of-algorithmic-bias-on-society/

Guidotti R., & Monreale A., & Ruggieri S., & Turini F., & Giannotti F., & Pedreschi D. (2018). A Survey of Methods for Explaining Black Box Models, Retrieved October 26th, 2020 from https://doi.org/10.1145/3236009

Iriondo R. (2020). What is Machine Learning?. Retrieved October 15th, 2020, from https://medium.com/towards-artificial-intelligence/what-is-machine-learning-ml-b58162f97ec7

Seager, T., Selinger, E., & Wiek, A. (2012). Sustainable Engineering Science for Resolving Wicked Problems. Journal of Agricultural and Environmental Ethics, 25(4), 467–484. https://doi.org/10.1007/s10806-011-9342-2

Karn U. (2016). An Intuitive Explanation of Convolutional Neural Networks. Retrieved from https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/#:~:text=Convolutional%20Neural%20Networks%20(ConvNets%20or,robots%20and%20self%20driving%20cars.

Varshney S. (2018). Introducing AI Fairness 360. Retrieved October 15th, 2020, from https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/