Bias in Binary Classification: An Argumentation-based Approach to Detection, Explanation and Mitigation

Extended Abstract for ACM WomENcourage 2023, Trondheim, Norway

Madeleine Waller*
madeleine.waller@kcl.ac.uk
King's College London
London, United Kingdom

KEYWORDS

Fairness, Bias, Explainability, Transparency, Argumentation

1 OVERVIEW

Our research investigates the use of computational argumentation as a tool for detecting, explaining and mitigating unwanted bias in tabular data-driven decision-making systems. Specifically we address the following research questions: (1) *Can argumentation be used to model unwanted bias towards an individual compared to similar individuals in a tabular data-driven decision-making system?* (2) *How can potential unwanted bias in an individual's decision compared to similar individuals be explained to a stakeholder of the decision-making system?* (3) *How can feedback from a stakeholder mitigate unwanted bias from an individual's decision?* Our work differs from current research by providing explanations of potential unwanted bias to be able to consider the context and application of a system.

2 INTRODUCTION

Fair artificial intelligence (AI) aims to design AI systems which align with what have previously been defined as fair outcomes, which are outcomes that are impartial and do not discriminate. Designing fair automated decision-making systems is becoming more important as they become widely adopted, therefore having a greater impact on individuals and society as a whole. Decision-making in high impact scenarios, whether by a human or technology, will always have the potential to be discriminatory but the rise of data collection and the reach of technology means these decisions can have a far reaching and substantial effect that can be hard to monitor.

There have been many recent examples of machine learning being used to make impactful decisions that have been unfair including in the domains of criminal justice [6], recruitment [3] and social services [2]. Well known examples are systems used to predict if criminals will re-offend which are widely in use across many US states in particular. An analysis of a popular tool COMPAS [4] discovered that black defendants were identified incorrectly as re-offending at a higher rate than white defendants. In 2018, Amazon had to stop the use of a recruitment tool that was shown to be biased against women [3], reinforcing historical biases due to the small sample size of women who had previously been hired. The potential impact of these systems is immense and ensuring they are fair to individuals and communities is an important cross-disciplinary issue which must consider the context and application of the systems deployed [7].

The overall objective of our research is to contribute a new bias mitigation method for binary classification decision-making systems. Existing methods typically focus on the quantification of bias through the use of fairness metrics without considering the context or application of a system. Instead of using existing metrics to detect unwanted bias or evaluate the effectiveness of a bias mitigation method, we aim to develop a method that prioritises on transparency and explainability. Simply reporting that the fairness of a system has improved by a certain percentage e.g., 20% after the application of a bias mitigation method may not convey they full impact of the method to the individuals affected by it. Therefore, providing explanations for unwanted bias is more informative and useful than simply providing a numerical representation of the bias.

3 METHOD

Our problem is as follows. A black-box machine learning binary classification system has been deployed, and there is no access to the system's training data or training algorithm. A stakeholder of the system wants to ensure a given classification for an individual is fair with respect to similar individuals. Our proposed method to solve this problem is split up into bias detection, explanation and mitigation.

3.1 Bias detection

Can argumentation be used to model unwanted bias towards an individual compared to similar individuals in a tabular data-driven decision-making system?

The overall objective for our research is a novel bias mitigation method for binary classification decision-making systems. The first step towards this goal is a method for representing a queried individual's classification in relation to the classifications of similar individuals, in order to identify attribute values that contribute to the individual's classification.

We propose a quantitative argumentation framework [1] where all the attribute values in the queried and similar individuals are represented by arguments, and the conflicts between attribute values are represented by attacks between those arguments. Specifically, if the classification of the queried individual is negative and the classification of a similar individual is positive, for each attribute that contains different values for the two individuals, there is an

*With support from supervisors Oana Cocarascu and Odinaldo Rodrigues, {oana.cocarascu, odinaldo.rodrigues}@kcl.ac.uk, King's College London

attack relation from each attribute value in the similar individual to that attribute value in the queried individual.

We then apply existing semantics which calculate the final strength of arguments [5] and find the weakest argument which corresponds to the attribute value that contributes most to a negative classification.

Our results of experiments run on available datasets show that our method is effective in identifying unwanted bias with respect to similar individuals.

3.2 Bias explanation

How can potential unwanted bias in an individual's decision compared to similar individuals be explained to a stakeholder of the decision-making system?

Given the graph and the strength of the arguments, the next research question asks how we can best explain this to different potential stakeholders of an AI decision-making system. Key aspects to consider in designing the explanations include:

- How do we extract subgraphs from the graph for meaningful explanations?
- How do we convert the subgraphs to textual explanations?
- Should explanations differ for different stakeholders?
- How should we evaluate the explanations given?

3.3 Bias mitigation

How can feedback from a stakeholder mitigate unwanted bias from an individual's decision?

Our new bias mitigation method will allow a stakeholder to provide feedback based on the explanation they receive and mitigate the potential unwanted bias in the decision queried. The mitigation process will involve adapting weights and relationships of arguments in the graph constructed from the detection method (this is yet to be defined). After mitigation, the classification for the queried individual should either change or a different explanation should be given for the same classification.

Below is an example of how the bias mitigation method would incorporate feedback from a stakeholder.

(1) Explanation given by the proposed bias explanation method: "Queried individual was rejected from the loan compared to similar individuals because... the value of sex is female"
(2) User gives feedback: Does this explanation detect unwanted bias? "Yes"
(3) **Apply proposed bias mitigation method (adapting the relationships between arguments in the framework)**
(4) New explanation given by the proposed bias explanation method: "Queried individual was accepted for the loan compared to similar individuals because... the value of credit score is high" Or "Queried individual was rejected from the loan compared to similar individuals because... the value of age is under 18"

4 RELEVANCE TO WOMENCOURAGE THEME

AI fairness and explainability research are very relevant to the theme of *computing connecting everyone*. As AI becomes increasingly integrated into our daily lives, it is important that it is fair and explainable to all people, regardless of their race, gender, ethnicity, or other personal characteristics.

Explainability refers to the ability to understand how and why an AI system arrived at a particular decision or recommendation. This is important for accountability, transparency, and trust. If people do not understand how an AI system arrived at a decision, they may be less likely to trust it or use it, which could limit its effectiveness in connecting everyone.

5 AUTHOR DETAILS

Madeleine Waller (madeleine.waller@kcl.ac.uk) is a PhD candidate in the UK Research and Innovation Centre for Doctoral Training (CDT) in Safe and Trusted Artificial Intelligence (STAI) at King's College London. Her research and expertise lie in the connection between fairness and explainablity of AI systems, specifically using argumentatation-based approaches. Madeleine has presented her work at various AI events, including at the King's Women in Informatics Conference 2022. Additionally, she was awarded a poster prize at the Safe and Trustworthy AI Workshop 2022 (STAIws22). Her contributions to the academic community include being a reviewer for the Conference on Computational Models of Argument 2022 (COMMA22) and the International Conference on Principles of Knowledge Representation and Reasoning 2023 (KR23). She is also the social secretary of her program (STAI CDT) which involves organising events to enhance collaboration between researchers.

ACKNOWLEDGMENTS

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [grant number EP/S023356/1][1].

REFERENCES

[1] Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument Comput.* 6, 1 (2015), 24–49. https://doi.org/10.1080/19462166.2014.1001791
[2] Philip Gillingham. 2019. Decision support systems, social justice and algorithmic accountability in social work: A new challenge. *Practice* 31, 4 (2019), 277–290.
[3] Jeffrey Dastin. 2018. *Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women.* Technical Report. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[4] Northpointe. 2019. *Practitioner's Guide to COMPAS Core.* Technical Report. https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf
[5] Nico Potyka. 2019. Extending Modular Semantics for Bipolar Weighted Argumentation. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, 1722–1730. http://dl.acm.org/citation.cfm?id=3331903
[6] The Partnership on AI. 2019. Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System. https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/
[7] Madeleine Waller and Paul Waller. 2020. Why Predictive Algorithms are So Risky for Public Sector Bodies. http://dx.doi.org/10.2139/ssrn.3716166

[1] www.safeandtrustedai.org