Argumentation for Bias Detection, Explanation and Mitigation in Decision-Making Systems

Madeleine Waller

madeleine.waller@kcl.ac.uk

SAFE & TRUSTED AI
UKRI CENTRE FOR DOCTORAL TRAINING

UK Research and Innovation

Motivation

- Automated decision-making systems have the potential to have great impact on individuals and society
- Existing systems have been shown to be unfair towards historically disadvantaged groups (e.g. in criminal justice [1], recruitment, social services)
- Individuals impacted by these decisions have a right to contest the decision and be provided with an explanation as to why that decision has been made

What is fairness?

In a binary classification decision-making system, fairness is quantified using metrics which correspond to a notion of *group* or *individual* fairness [2].

- Disparate impact (*group*): difference in positive classifications between protected groups (e.g. Male/Female)
- Consistency (*individual*): count of differences in classifications between similar individuals

XAI for bias detection

Existing bias mitigation methods:

- Focus on the optimisation of some *fairness metrics*
- Are not evaluated on different datasets or with different metrics so it is not clear in what scenarios they can be applied
- Require domain experts to identify *protected* or *proxy* features (ones correlated with protected features)
- Do not take into consideration the **context** or **application** of the system

Detection

Aim: to develop a quantitative abstract argumentation framework [3] to mirror a black-box classifier to be able to detect unwanted bias present in a classifier

Output: a set of feature values that contribute to an individual's classification

Initial ideas:

Similar individuals are ones which are within some threshold of a distance metric e.g. closest five neighbours according to the Manhattan distance [4]

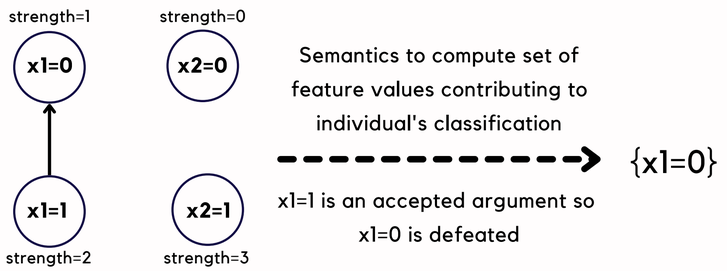
Represent this in a quantitative argumentation framework where:

- the arguments are feature values
- relationship between arguments (attacks) are the features that differ between the individual and *similar individuals* - positive classification attacks negative classification
- the initial strength of arguments is the number of times that feature values appears in the dataset

A very simple dataset

x1	x2	y
0	1	0
1	1	1
1	1	1

Framework



Semantics to compute set of feature values contributing to individual's classification

x1=1 is an accepted argument so x1=0 is defeated

{x1=0}

Explanation

Aim: to develop a method to explain to an individual* the output from the bias detection argumentation framework in natural language

Output: an explanation that allows an individual to understand whether a decision is discriminatory or made on a justifiable basis

*Note an individual could be an individual impacted, but more likely a domain expert such as a judge or social worker

Automated decision-making systems have the potential to be discriminatory. Methods to mitigate unwanted bias in these systems need to be developed that take into consideration the context and application of the system, as well as improving the transparency of the system.

Mitigation

Aim: to develop a method that allows feedback from an individual back into the system to mitigate the bias

Output: a new classification and explanation for that classification that allows for an individual to understand that the new decision is not discriminatory and made on a justifiable basis

Explanation given:

"Rejected from the loan because the value of feature 'sex' is female"

User gives feedback:

"Does this explanation detect unwanted bias in the decision?"

"Yes"

APPLY BIAS MITIGATION

New explanation given:

"Accepted for the loan because the value of feature 'credit score' is high"

Or

"Rejected from the loan because the value of feature 'age' is under 18"

References

[1] J. Larson et al. 2016. 'How We Analyzed the COMPAS Recidivism Algorithm.'
[2] K. Makhlouf et al. 2021. 'On the Applicability of Machine Learning Fairness Notions'
[3] A. Vassiliades et al. 2021. 'Argumentation and Explainable Artificial Intelligence: A Survey'
[4] J. Chakraborty et al. 2020 'Making Fair ML Software Using Trustworthy Explanation'