A Bio-Inspired Framework for Machine Bias Interpretation

Jake Robertson
robertsj@cs.uni-freiburg.de
University of Freiburg
Freiburg, BW, Germany

Catherine Stinson
c.stinson@queensu.ca
Queen's University
Kingston, ON, Canada

Ting Hu
ting.hu@queensu.ca
Queen's University
Kingston, ON, Canada

ABSTRACT

Machine learning algorithms use the past and the present to predict the future. But when given biased historical data, these algorithms can quickly become discriminatory. The area of machine learning fairness has emerged to detect and de-bias these algorithms, but has received widespread criticism for its one-size-fits-all approach, which allows certain cases of bias to slip through the cracks. In this study, we take a deeper look at the mechanisms by which machine learning algorithms develop harmful bias. We introduce a new method to interpret discriminatory systems, an Evolutionary algorithm for Feature Interaction (EFI), which we apply to several commonly used machine learning algorithms in two real-world problem instances: violent crime and median house price prediction. In the results, we discover several complex forms of bias including the encoding of race through other seemingly unrelated attributes. Ultimately we suggest that more informative interpretation tools such as EFI can be used to not only explain machine learning outcomes, but supplement and improve existing machine bias detection approaches to provide a more robust and in-depth ethical evaluation of machine learning algorithms.

CCS CONCEPTS

• **Computing methodologies** → Genetic algorithms.

KEYWORDS

Interpretability, fairness, machine bias, feature importance, feature interaction

ACM Reference Format:
Jake Robertson, Catherine Stinson, and Ting Hu. 2022. A Bio-Inspired Framework for Machine Bias Interpretation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3514094.3534126

1 INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have revolutionized the modern world, but also raised a host of difficult questions. This paper focuses on two such questions: 1) how can we understand the decisions made by AI systems, and 2) how can we ensure that these systems are treating people fairly? Understanding AI systems is important both in its own right, and as a step toward evaluating fairness and ethical decision making.

The predictive power of ML algorithms comes from their ability to internalize and represent features and their complex relationships. However, this same complexity poses a challenge when it comes to deciding ethical questions. Interpretable Machine Learning (iML) has the goal of making black-box ML systems understandable to human interpreters [14], but is a challenge in itself due to the high complexity of these systems.

The challenge of fairness has received significant attention due to discoveries of models making skewed decisions with respect to attributes like race and gender [6]. ML algorithms can internalize the complex systems of inequality reflected in our data, and when applied in the wrong setting, reinforce cycles of systemic oppression. There have been many advances in ML fairness, including statistical measures of fairness and bias removal methods which allow us to automatically configure fairer models [17]. Although such approaches have strong statistical roots, many practitioners are unsatisfied with their one-size-fits-all nature [2], which allows special cases of bias to slip through the cracks. In order for a more robust and interactive evaluation of fairness, an interplay between other machine learning objectives like interpretability, transparency, and explainability is required.

In this study, we develop a method to improve interpretability in complex ML models. We introduce an Evolutionary algorithm for Feature Interaction (EFI)[1], whose contributions to interpretability are twofold: 1) a bio-inspired approach to efficiently search for synergistic feature interactions and 2) a novel, permutation based method to approximate feature interaction strength with improved efficiency. In the following sections, we present EFI's methodology and apply it to both synthetic and real world example problems. We show that interpretability tools can supplement fairness metrics by providing an in depth understanding of complex ML biases. EFI could also be used as a robust method of selecting combinations of attributes to examine for intersectional fairness.

2 BACKGROUND

In this section, we introduce and motivate the methodology of our approach with an overview of the current state and challenges in ML interpretability and fairness.

2.1 Interpretability

According to the Predictive Descriptive Relevant (PDR) framework for iML development [14], interpretable models and interpretability tools should maintain a high degree of accuracy (predictive), comprehensively describe the model's predictive process (descriptive), and offer insights that are comprehensible to the human interpreter

[1] https://github.com/jr2021/EFI

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534126