Incorporating the Concept of Bias and Fairness in Cybersecurity Curricular Module

Sheikh Rabiul Islam
University of Hartford
West Hartford, CT, USA
shislam@hartford.edu

Ingrid Russell
University of Hartford
West Hartford, CT, USA
irussell@hartford.edu

Maanak Gupta
Tennessee Tech University
Cookeville, TN, USA
mgupta@tntech.edu

ABSTRACT

Although Artificial Intelligence has become an integral part of modern cybersecurity solutions, data bias and algorithmic bias have made it vulnerable to many cyberattacks, in particular, adversarial attacks where the attacker crafts input of the AI system to exploit the existence of possible bias in the data or algorithms. In this paper, we share our experiences with ongoing work to develop and evalu-ate a cybersecurity curricular module that demonstrates (a) data bias detection, (b) data bias mitigation, (c) algorithmic bias detec-tion, and (d) algorithmic bias mitigation, using a network intrusion detection problem on real-world data. The module includes lectures and hands-on exercises, using state-of-the-art and open-source bias detection and mitigation software on a real-world dataset. The goal is to identify and mitigate the prevailing conscious/unconscious bias in data and algorithms that the attacker might exploit.

1 INTRODUCTION

Skewed data during the training phase of AI-based security systems, incorrect data sampling, and embedded assumptions as logic are the main sources of biased decisions. This is an underexplored area of research, specifically from the cybersecurity perspective. Therefore, there is a need for a curricular module in this area to make students (i.e., future workforce) aware of this serious pitfall that attackers can easily exploit. Our curricular module on bias and fairness in AI-enabled cybersecurity solutions (i.e., network intrusion detection) consists of (1) lecture and demonstration, and (2) hands-on exercises on a real-world dataset. The module is for computer science course on 'Intrusion Detection and Security' that covers techniques and algorithms for detecting unusual activities, in networks or devices that typically signal a break-in or breach.

2 DESCRIPTION

The lecture component of the module contains some of the prominent cases of bias and discrimination with automated decision-making with the demonstration of bias detection and mitigation tools. For bias detection, we use the open source tool Aequitas[1], and for bias mitigation, we use IBM AI Fairness 360[2]. In the hands-on component of the module, students are given four sequential lab exercises (see Table 1). In exercise 1, students detect data bias (perhaps stemming from sampling, data poisoning, or imbalanced data) in the data using Aequitas, and thus, detect possible scopes of adversarial attack in the network intrusion detection scenario. In exercise 2, students mitigate data bias, using a reweighing technique of IBM AI 360, resulting in a more balanced (i.e., bias-free) dataset, to mitigate possible adversarial attack scenarios. In exercise 3, students proceed with the bias-free dataset, and apply both customized algorithms (e.g., human bias embedded as rules) and standard algorithms to identify the difference stemming from the infused bias (e.g., embedding rules that consider network trafic from a part of the world safe). Exercise 2 gives a comparatively less biased dataset, and exercise 3 gives a comparatively less biased algorithm. Finally, in exercise 4, students mitigate algorithmic bias using a bias mitigated algorithm on bias mitigated data and neutralize the scope of adversarial attacks that might exploit the presence of bias in the data and/or algorithm for a successful attack.

Table 1: Exercise, description, and relevant issues with Cybersecurity and Artificial Intelligence

Exercise	Description	Cybersecurity Issues	AI/ML issues
1. Data bias detection	Use of a bias detection tool (e.g., Aequitas) to detect bias in the data	Detect adversarial attacks in network intrusion detection	Sampling, poisoned data, data imbalance
2. Data bias mitigation	Use of a bias mitigation tool (e.g., IBM AI 360) to mitigate bias	Mitigate possible adversarial attack scenarios	Reweighing, corrected and balanced data
3. Algorithmic bias detection	Use of corrected data on a customized algorithm and in other well-known algorithms	Detect human infused bias from the difference in customized algorithms	Algorithmic bias
4. Algorithmic bias mitigation	Use of curated data (i.e., bias-free data) and chosen algorithm (e.g., less bias prone one)	Neutralize the scope of adversarial attacks that stems from data/algorithmic bias	Algorithmic bias mitigation

3 CONCLUSION

Pre- and post-survey results of prior work [1, 2] on data bias detection and mitigation, in a data mining course, suggest that the introduced module helped to instill the consciousness of fairness and bias in students. We plan to conduct similar surveys, for our module, in the offering of the Intrusion Detection and Security course, and expect to see similar student experiences.

REFERENCES

[1] Sheikh Rabiul Islam, Ingrid Russell, William Eberle, and Darina Dicheva. 2022. Incorporating the Concepts of Fairness and Bias into an Undergraduate Computer Science Course to Promote Fair Automated Decision Systems. In Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 2. 1075–1075.
[2] Sheikh Rabiul Islam, Ingrid Russell, William Eberle, and Darina Dicheva. 2022. Instilling conscience about bias and fairness in automated decisions. Journal of Computing Sciences in Colleges 37, 8 (2022), 22–31.

[1] http://aequitas.dssg.io
[2] https://aif360.mybluemix.net

SIGCSE 2023, March 15–18, 2023, Toronto, Canada

ACM ISBN 978-1-4503-9433-8/23/03.
https://doi.org/10.1145/3545947.3576302