Comparison of Post-processing Bias Mitigation Strategies to Accelerate the Creation and Adoption of Fairer Machine Learning Algorithms

Nathaniel Valdez-Cagua[1], Diego Aguirre[2], Martine Ceberio[2], and Patricia Morreale[1]

Kean University[1], School of Computer Science and Technology, Union, NJ 07083

The University of Texas at El Paso[2], Computer Science Department, El Paso, TX 79968

Keywords: Fairness, Machine Learning, Bias Mitigation, Post-processing Mitigation Strategies, Fairness Metrics.

Many algorithms are currently using machine learning. Machine learning algorithms create mathematical models to understand data. They learn by tuning parameters in the models that adapt to observed data. After being "fit" to previously seen data, they can be used to predict and understand newly observed data. Some of these algorithms can make predictions that have a disproportionately negative impact on unprivileged groups in protected classes such as race and sex. One example is machine learning algorithms using race as a "risk factor" to predict success in STEM majors, a strategy that disproportionately predicts that Black and Latinx students would fail in STEM majors, which could cause advisors to steer Black and Latinx students away from becoming STEM majors. To prevent this from occurring, many bias mitigation strategies have been created to increase fairness for unprivileged groups.

There are three types of bias mitigation strategies: pre-processing, in-processing, and post-processing.

- Pre-processing strategies mitigate bias by modifying a dataset before a machine learning algorithm is trained with it.
- In-processing strategies mitigate bias by changing the algorithm's classifier to take fairness into consideration.
- Post-processing strategies mitigate bias by modifying the results of an algorithm to increase fairness.

For bias mitigation strategies to increase fairness, fairness must first be defined so that it may be measurable. There is no universal definition of fairness, so researchers have created many definitions of fairness. This led to the creation of fairness metrics. Each fairness metric measures different things. *Statistical Parity Difference* measures how each protected attribute, such as race and sex, affects the predictions. *Equal Opportunity Difference* measures the difference in positive outcomes for each group. Each bias mitigation strategy works differently, so the results can differ greatly between them. Currently, there is a lack of standards regarding which bias mitigation strategy should be used in certain situations. Some bias mitigation strategies work better for certain kinds of datasets or improving certain fairness metrics. This prevents the adoption of bias-mitigation strategies since it is difficult for practitioners interested in using them to know which one to use in their situation.

The purpose of the research project was to test post-processing strategies in various scenarios and compare the results to find out which post-processing strategy is best for specific scenarios.

To test post-processing strategies, I created a machine learning algorithm. The algorithm I created uses logistic regression as its machine learning model. I used the algorithm on three

different datasets to predict the value of a variable. One variable I tried to predict was whether someone was a good or bad credit risk using a German credit dataset. This kind of prediction may lead to discrimination if the algorithm decided that certain races were risky to lend money to, causing those races to have fewer opportunities to borrow money. The fairness metrics I measured were balanced accuracy, average odds difference, disparate impact, statistical parity difference, equal opportunity difference, and Theil index. These metrics cover a wide range of fairness definitions. The post-processing strategies I have tested so far are Equalized Odds Post-processing and Calibrated Equalized Odds Post-processing. I chose them because of their popularity and performance. The machine learning algorithm I created split the dataset into three parts for training, validating, and testing. I ran many experiments where different combinations of datasets and mitigation strategies were tested to analyze how they perform in different contexts across the previously mentioned fairness metrics.

Using the results of my experiments, I have made a few interesting observations. Equalized Odds Post-processing (EOP) was generally 32 times better than Calibrated Equalized Odds Post-processing (CEOP) and 16 times better than no post-processing at reducing equal opportunity difference. EOP was generally 1.6 times better than CEOP and 1.5 times better than no post-processing at improving disparate impact.

The intellectual merit of the work was that the mitigation strategies were able to reduce bias while maintaining acceptable performance. These results could be used by a practitioner to decide on what strategy to use. Someone interested in reducing equal opportunity differences should use EOP according to this information.

The broader impacts of this work will help reduce bias in machine learning by allowing practitioners to compare and select bias mitigation strategies more easily, leading to the creation and use of fairer algorithms. In the future, I will test more strategies and test them on other kinds of datasets to get a clearer picture of their performance.