Midterm STOR390

Collin Register

2024-03-22

With the increasing prevalence of artificial intelligence in today's decision-making processes, concerns about algorithmic bias have also been on the rise. One common application of artificial intelligence is using algorithms to determine if prospective job candidates are a good fit for employment. Within this classification, algorithms often misclassify groups at different rates, maintaining a level of bias in decisions, which often leads to discrimination. The prevalence of bias within an algorithm is a complex issue that becomes increasingly complicated when the tradeoff between fairness and accuracy is introduced. Often, to maintain a high level of fairness, a certain degree of accuracy must be compromised to ensure that fairness. There is often a struggle to find the right balance of fairness and accuracy, making it a difficult task. Also, the bias within the classification is often hard to detect and can make it difficult to alter the decision-making process fairly.

One study, using an artificial intelligence software toolkit, highlights this tradeoff in terms of accuracy and fairness. In a paper detailing this study, "Bias mitigation with AIF360: A comparative study," the methods and results of how this is explored are explained in detail. The goal of this study is to determine the performance of these bias mitigation algorithms and analyze the fairness and accuracy of each of them in predicting employment status, excluding gender and race as variables.

The dataset used for this analysis is the 2013 U.S. census data, with original records of over 130,000 observations. Upon initial assessment and using a logistic regression classification, it was found that men were associated with 9.5% more favorable labels (employed) than women, and whites were associated with 9.7% more than non-whites. Given this, the privileged groups were set to be whites and males, and the unprivileged groups to be women and non-whites. After finding this initial bias within the logistic regression, the researchers first did pre-processing on the data, where all null values were removed, and they were left with 56827 observations. Then, the data was sorted so that age was binned by decades, and anyone under the age of 14 was removed due to not being of working age. The race attribute was grouped into non-white and white.

After detecting the initial bias, the researchers aim to combat this bias using bias mitigation algorithms. They use Reject Option Based Classification (ROC) and Prejudice Remover (PR), two of the bias mitigation algorithms within the package AIF360. Once the algorithms are applied, they are evaluated and compared to the classification without the fairness constraints, using statistical measures of fairness such as disparate impact, statistical parity difference, average odds difference, and equal opportunity difference. Disparate impact computes the measure of the intentional bias in the label assignment. Statistical parity is the difference in probabilities that the protected and unprotected groups have when being assigned the favorable label. The average odds difference is a combination of the difference

between the true-positive rates of both groups and the false-positive rate of both groups in classification. The equal opportunity difference is the difference in probability of the two groups being wrongly assigned the unfavorable label. In this study, the fairness metrics looked at group fairness as opposed to individual fairness due to the lack of specificity within the dataset and to avoid having to make assumptions regarding individuals.

Following the application of each of the bias mitigation techniques, we are left with the following results. For ROC with gender as the protected attribute, accuracy decreased by about 2 percent compared to the classification without the fairness constraint, while significant improvements were made with regard to the measures of statistical parity difference, disparate impact, average odds, and equal opportunity. When looking at ROC with race as a protected attribute, there was a small decrease in accuracy (roughly half a percent), while significant improvements were made regarding the measures of statistical parity difference, disparate impact, average odds, and equal opportunity.

When comparing the application of PR with gender as the protected attribute to the original, accuracy decreased by roughly four percent and average odds decreased. However, significant improvements were made regarding the measures of statistical parity difference, disparate impact, and equal opportunity. However, when comparing the application of PR with race as the protected attribute to the original, we see an improvement in all the statistical measures of fairness and a slight decrease in accuracy.

An interesting result is that ROC overshoots for both gender and race for the average odds difference. For example, ROC with fairness constraints sways from 13.88% in favor of men to 4.01% in favor of women. All of the other fairness metrics seemed to increase drastically towards removing bias. In this instance, we can see that for most of the bias mitigation techniques, accuracy decreased by a few percent, but the statistical measures of fairness improved greatly.

Sources

Hufthammer, Knut T., et al. “Bias Mitigation with AIF360: A Comparative Study.” Bergen Open Research Archive, Norsk IKT-konferanse for forskning og utdanning, 15 Dec. 2020, bora.uib.no/bora-xmlui/handle/11250/2764230.