Debiasing Algorithms Using Multiple Adversaries

By Kara Davis

Rensselaer

Problem Statement

Can multiple adversaries be combined to create high accuracy algorithms that are fair by eliminating group discrimination even when generated from datasets that exhibit biases?

Why is this an issue?

Often image recognition systems are created from datasets that are biased toward light-skinned people and have poorer accuracy on darker female faces.

The UCI Adult Dataset discriminates against female individuals which is in part due to unequal representation of male and female demographics in the data set.

A fair adversarial debiaser can only do so much to learn a fair classifier if it is trained on unfair datasets.

Fair datasets are shown to exhibit biases when implemented into unfair algorithms.

Goal: create a generative adversarial network that helps to reduce discrimination in datasets by generating data with fairness constraints and combine this with an adversarial debiaser trained on fairness attributes to ensure a fair classifier is created

Related Work

FairGAN: Fairness-aware Generative Adversarial Networks: (Xu et. al. 2018) Uses fairness-aware generative adversarial networks to ensure the classifiers which are trained on generated data can achieve fair classification on real data.

Mitigating Unwanted Biases with Adversarial Learning: (Zhang et. al. 2018) Original paper presented on in class discussing different fairness measures and implementations to reduce an adversary's ability to determine the protected attribute from the predictions.

Estimating and Improving Fairness with Adversarial Learning: (Li et. al. 2021) In this paper they use a GAN to improve the original dataset and then add an additional adversarial network that predicts the fairness scores for the given trained model and the testing data to improve skin lesion detection.

Understanding Unequal Gender Classification Accuracy from Face Images: (Muthukumar et. al. 2018) Explores the root causes of unequal performance of commercial face classification services in the gender classification task across intersectional groups defined by skin type and gender.

Generative Adversarial Network with Fairness Discriminator

Generative Adversarial Network:

- Involves a **generator** that generates fake samples from a prior distribution and learns a generative distribution to match the real data distribution.
- The **discriminative component** is a binary classifier that predicts whether an input is real data or fake data generated from the generator.
- The discriminator **outputs the probability** that a value is from the real data or the generated fake data.
- The generator is then trained to fool the discriminator enough to the point where it cannot distinguish the generated data from the real data (want high probability that generated data is real)

Fairness Component:

For data to be free from discrimination and be considered fair, generated samples must be free from both

- **disparate treatment:** discriminatory outcomes are due to explicitly using the protected attribute to make decisions
- **disparate impact:** discriminatory outcomes are not explicitly from the protected attribute but from the proxy unprotected attributes (harder to eliminate, and focus on removing only disparate treatment in this project, although this was accomplished by the paper on FairGAN)

Generating Fair Data with Adult Dataset

The Adult Income dataset is shown to be biased against the female individuals, so the FairGAN algorithm sets the following parameters:

- **protected attribute**: sex
- **underprivileged group:** female
- **decision label:** income
- **desired value label:** >50k

Goal: to generate data that is less biased than the original dataset

Trained with 200 epochs, batch size of 256, with 20 fair epochs and a lambda parameter of 0.5 to produce 32561 data records

```
$ python TabFairGAN.py --help
usage: TabFairGAN.py [-h]
                     df_name S Y underprivileged_value desirable_value
                     num_epochs batch_size num_fair_epochs lambda_val
                     fake_name size_of_fake_data

positional arguments:
  df_name               Reference dataframe
  S                     Protected attribute
  Y                     Label (decision)
  underprivileged_value
                        Value for underpriviledged group
  desirable_value       Desired label (decision)
  num_epochs            Total number of epochs
  batch_size            the batch size
  num_fair_epochs       number of fair training epochs
  lambda_val            lambda parameter
  fake_name             name of the produced csv file
  size_of_fake_data     how many data records to generate
```

Second Adversarial Network

- Once the new dataset is generated, it is fed into an algorithm with an adversarial debiaser
- This provides further metrics for ensuring that the dataset is fair by learning a classifier to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute (in this case sex and race) from the predictions.
- This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

Original Dataset Results

Plain model - without debiasing - classification metrics

```
Test set: Classification accuracy = 0.805978
Test set: Balanced classification accuracy = 0.659291
Test set: Disparate impact = 0.000000
Test set: Equal opportunity difference = -0.447797
Test set: Average odds difference = -0.274461
Test set: Theil_index = 0.177591
```

Model - with debiasing - classification metrics

```
Test set: Classification accuracy = 0.799154
Test set: Balanced classification accuracy = 0.664351
Test set: Disparate impact = 0.323661
Test set: Equal opportunity difference = -0.236832
Test set: Average odds difference = -0.143563
Test set: Theil_index = 0.174275
```

Fair Dataset Results

Plain model - without debiasing - classification metrics

```
Test set: Classification accuracy = 0.774244
Test set: Balanced classification accuracy = 0.597273
Test set: Disparate impact = 0.209727
Test set: Equal opportunity difference = -0.226973
Test set: Average odds difference = -0.135728
Test set: Theil_index = 0.228893
```

Model - with debiasing - classification metrics

```
Test set: Classification accuracy = 0.765782
Test set: Balanced classification accuracy = 0.612658
Test set: Disparate impact = 1.215056
Test set: Equal opportunity difference = 0.039271
Test set: Average odds difference = 0.043794
Test set: Theil_index = 0.217727
```

Experimental Results on Fair Data

Results with Original Dataset

Plain model - without debiasing - dataset metrics

```
Train set: Difference in mean outcomes between unprivileged and privileged groups = -0.209651
Test set: Difference in mean outcomes between unprivileged and privileged groups = -0.205555
```

Model - with debiasing - dataset metrics

```
Train set: Difference in mean outcomes between unprivileged and privileged groups = -0.139003
Test set: Difference in mean outcomes between unprivileged and privileged groups = -0.137160
```

Results with Fairly Generated Dataset

Plain model - without debiasing - dataset metrics

```
Train set: Difference in mean outcomes between unprivileged and privileged groups = -0.106604
Test set: Difference in mean outcomes between unprivileged and privileged groups = -0.100891
```

Model - with debiasing - dataset metrics

```
Train set: Difference in mean outcomes between unprivileged and privileged groups = 0.022080
Test set: Difference in mean outcomes between unprivileged and privileged groups = 0.026784
```

Conclusion

Key Take-Aways

- Equality of opportunity difference and average odds difference greatly reduced
- Able to significantly reduce the difference between outcomes for privileged and underprivileged groups
- Trade off: noticeable decrease in accuracy
- Addtionial increase in disparate impact as the adversary cannot make its decision from protected attribute when we ultimately want this to be eliminated as well

Future Directions to be Explored

- A next step would be to use this method with algorithms involving models that complete facial recognition. With the ability to generate high quality images with a GAN we would likely see the same results for increasing fairness and mitigating discrimination
- Other implementations go further for mitigating disparate impact which was seen to increase in this experiment
  - Implementations include adding a second discriminator to your GAN that attempts to predict a protected attribute given a generated sample and the generator aims to fool the discriminator
  - Once the generated sample cannot be used to predict the protected attribute the correlation between the two is removed and you can ensure the generated samples do not have disparate impact

Repositories Cited

- GAN with fairness constraints: TabFairGAN
- Adversarial Debiaser: AI Fairness 360
- My repository on combining the two: Debiasing Using Multiple Adversaries