CS 689 Project Report

Repairing without Retraining:
Avoiding Disparate Impact with Counterfactual Distributions

https://arxiv.org/abs/1901.10501

Anuj Srivastava

November 2021

1 Introduction

A predictive model is said to have disparate impact when there is a difference in its performance across population groups defined by sensitive attributes such as ethnicity or gender. This work aims to minimize certain kinds of disparities, which can be expressed as difference in probability distributions of output labels and predictions between the aforementioned groups.

Some approaches to solve this problem include integrating the minimization of the disparity metric in the training objective, introducing bias in the output of the model for the disadvantaged groups. *Repairing without Retraining*, as the name suggests, repairs a potentially unfair model without altering the model itself, and no assumptions are made about its structure, treating it as a black box. This is especially useful in applications which use third party models for predictions, where outputs are obtained with APIs or training data is not accessible.

This paper minimizes some classes of disparity metrics by perturbing the probability distribution of input variables for the disadvantaged groups. The perturbed distribution is referred to as the *counterfactual distribution*. A descent algorithm is introduced to estimate a counterfactual distribution from data. A data preprocessor is then built using the estimate, and application of this preprocessor before feeding the input to same model achieves a reduction in the disparity.

2 Framework

For the scope of the paper, we consider a standard binary classification task. Input variables are expressed as a vector $X = (X_1, \ldots, X_d) \in \mathcal{X}$ drawn from

Performance Metric	Disparity Metric
Statistical Parity (SP)	$\Pr(\hat{Y}=0\|S=0) - \Pr(\hat{Y}=0\|S=1)$
False Negative Rate (FNR)	$\Pr(\hat{Y}=0\|Y=1,S=0) - \Pr(\hat{Y}=0\|Y=1,S=1)$
False Positive Rate (FPR)	$\Pr(\hat{Y}=1\|Y=0,S=0) - \Pr(\hat{Y}=1\|Y=0,S=1)$

Table 1: Disparity Metrics $M(P_0)$

the probability distribution P_X, and output variable $Y \in \{0,1\}$. The black-box classifier $h : \mathcal{X} \to [0,1]$ may directly predict an outcome ($h(X) \in \{0,1\}$) such as in SVMs or a predicted probability ($h(X) \in [0,1]$) as in logistic regression.

The sensitive attribute $S \in \{0,1\}$ follows distribution P_S, with $S=0$ and $S=1$ denoting the *target* (disadvantaged) and *baseline* (privileged) groups respectively. The input distributions of the groups are denoted as $P_0 \triangleq P_{X|S=0}$, and $P_1 \triangleq P_{X|S=1}$.

A disparity metric is defined as a mapping $M : \mathcal{P} \to \mathbb{R}$ where $\mathcal{P}$ is the set of probability distributions over $\mathcal{X}$. The metrics implemented in this project are defined in Table 1.

A counterfactual distribution Q_X is a hypothetical distribution of input variables for the target group such that $Q_X \in \underset{Q'_X \in \mathcal{P}}{\operatorname{argmin}} |M(Q'_X)$.

3 Methodology

3.1 Learning Counterfactual Distributions

First let us define the perturbed distributions $\tilde{P}_0$ over the target group ($S=0$)

$$\tilde{P}_0(X) \triangleq P_0(X)(1+\epsilon f(X)), \quad \forall X \in \mathcal{X} \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is a perturbation function with zero mean and unit variance w.r.t. P_0, and $\epsilon > 0$ is a positive scaling constant chosen so that $\tilde{P}_0$ is a valid probability distribution.

Next we define for a given disparity metric M the influence function $\psi : \mathcal{X} \to \mathbb{R}$

$$\psi(X) \triangleq \lim_{\epsilon \to 0} \frac{M((1-\epsilon)P_0 + \epsilon\delta_X) - M(P_0)}{\epsilon} \tag{2}$$

where $\delta_X(z) = \mathbb{I}[z = X]$ is the delta function at X.
Intuitively ψ approximates the change in M when a sample $X \in \mathcal{X}$ is added to the dataset in the target group.

Now, given a disparity metric M, it can be shown that

$$\underset{f(X)}{\operatorname{argmin}} \lim_{\epsilon \to 0} \frac{M(\tilde{P}_0) - M(P_0)}{\epsilon} = \frac{-\psi(X)}{\sqrt{\mathbb{E}[\psi(X)^2 | S = 0]}} \tag{3}$$

if $\mathbb{E}[\psi(X)^2 | S = 0] \neq 0$. This indicates that $-\psi(X)$ reflects the direction of steepest descent in disparate impact.

The influence functions for the disparity metrics in Table 1 are expressed as

$$\begin{aligned}
\psi^{SP}(X) &= -h(X) + \hat{\mu}_0 \\
\psi^{FNR}(X) &= (1 - h(X))\hat{y}_0(X) - \gamma_{0,1}\hat{y}_0(X) \\
\psi^{FPR}(X) &= h(X)(1 - \hat{y}_0(X)) - \gamma_{0,1}(1 - \hat{y}_0(X))
\end{aligned}$$

where $h(X)$ is the black-box classifier to be repaired and $\hat{y}_0(X)$ is a classifier that aims to predict the same outcome but only for individuals in the target group, $P_{Y|X,S=0}(1|X)$. The constants $\hat{\mu}_0$ and $\gamma_{a,b}$ are

$$\begin{aligned}
\hat{\mu}_0 &\triangleq \Pr(Y = 1 | S = 0) \\
\gamma_{a,b} &\triangleq \Pr(\hat{Y} = a | Y = b, S = 0)
\end{aligned}$$

The input distributions for the target group are modelled as a vector of weights $[w_i]_{i \in I_0}$. Let the deployment dataset be $D = \{(x_i, y_i, s_i)\}_{i=1}^n$ and $I_0 = \{\imath = 1, \ldots, n | s_i = 0\}$ is the set of indices of individuals in the target group. Thus the original distribution has $w_i^0 = 1, \forall i \in I_0$. These weights are updated in a loop with the step $w_i^{t+1} \leftarrow w_i^t(1 - \epsilon\psi(x_i)), \forall i \in I_0$, until $M^{t+1} \geq M^t$. Here M^t is the disparity metric computed on D with weights w^t. (For clarity, with weights w_i, $\Pr(\hat{Y} = 0 | S = 0) = \frac{\sum_{i \in I_0} w_i \mathbb{I}[\hat{Y}=0]}{\sum_{i \in I_0} w_i}$.) $\epsilon > 0$ is a suitable step size.

3.2 Model Repair

To mitigate the disparate impact, we construct a preprocessor $T : \mathcal{X} \to \mathcal{X}$ that alters the input variable values of the target group. The repaired classifier $\tilde{h}$ is:

$$\tilde{h}(X) = \begin{cases} h(T(X)) & \text{if } s = 0, \\ h(X) & \text{otherwise} \end{cases} \tag{4}$$

We model T as a probabilistic mapping from $D_0 = \{x_i | i \in I_0\}$ to itself, and $\Pr(T(x_i) = x_j) = \frac{\gamma_{ij}*}{p_i}, \forall i, j = 1, \ldots, m$, where $m = |D_0|$, the number of unique data points in the target group. $\gamma_{ij}*$ is the solution to the following optimal

transport problem:

$$\begin{aligned}
&\min_{\gamma_{ij} \in \mathbb{R}^+} \sum_{i=1}^{m} \sum_{j=1}^{m} C_{ij} \gamma_{ij} \\
&\text{s.t.} \sum_{j=1}^{m} \gamma_{ij} = p_i, \quad i = 1, \ldots, m \\
&\sum_{i=1}^{m} \gamma_{ij} = q_j, \quad j = 1, \ldots, m
\end{aligned}$$

Here C_{ij} represents the cost of altering x_i to x_j, for example L_2-norm: $C_{ij} = ||x_i - x_j||_2$, and p, q are empirical estimates of P_0 and Q_X respectively.

$$\begin{aligned}
p_i &= \frac{1}{|I_0|} \sum_{x \in D_0} \delta_{x_i}(x) \\
q_j &= \frac{w_j}{|I_0|} \sum_{x \in D_0} \delta_{x_j}(x)
\end{aligned}$$

4 Experiments

My implementation of the project is at https://github.com/anuj27596/CS689-Project

The datasets used to test the introduced method are the **adult** dataset and the ProPublica **compas** dataset. For each dataset, 30% of samples are used to train the classifier h, 50% to recover the counterfactual distribution, and 20% to evaluate the performance after repairing.

The classifiers h and y_0 are l_2-regularized logistic regression models.

Dataset	Metric	Target Group	Original AUC	Original Disparity	Repaired AUC	Repaired Disparity
adult	SP	Female	0.8634	**0.1825**	0.8241	**0.0132**
adult	FPR	Male	0.8634	**0.0916**	0.7731	**0.0032**
adult	FNR	Female	0.8634	**0.1952**	0.8559	**0.0568**
compas	SP	White	0.7139	**0.1648**	0.6982	**0.0277**
compas	FPR	Non-white	0.7139	**0.1001**	0.6964	**0.0425**
compas	FNR	White	0.7139	**0.18**	0.6991	**0.0311**

Table 2: Change in disparate impact on repairing the classification model. Values are averaged over 10 runs with different random seeds

5 Conclusion

A new distributional paradigm to understand and mitigate disparate impact has been proposed by the paper. The tools introduced in the paper apply to discrete distributions and binary classification models, but can be extended to continuous distributions and other supervised learning models.
Limitations of this work are as follows: the preprocessor is randomized, which may result in different outcomes in applications such as loan approvals, by simply applying multiple times. The preprocessor also does not handle unseen data points in the discrete setting