Mitigation of Bias in Data to Achieve Fairness-Aware Classification

Pooja Hanavadi Balaraju[1]

RWTH Aachen University pooja.balaraju@rwth-aachen.de

Abstract. Machine learning has seen a rapid growth in recent years in a variety of sectors and services. Automated decision making uses machine learning techniques to remove human interference . It predicts the decision based on past experience of similar situation but makes it challenging because the predictions need to match with accuracy. However, one its major concerns is misclassification errors caused due to bias in the training data. These errors put certain sensitive groups at an unfair advantage creating different types of unfairness: disparate impact, disparate treatment and disparate mistreatment. Therefore fairness concerns have become increasingly important. In this paper, we discuss some of the current methodologies developed recently to mitigate bias and achieve fairness like massaging of the labels, group thresholding, covariance and adaptive sensitive reweighting. We learn that it is unsufficient to remove the sensitive information for eliminating biases because it has an indirect correlation. We analyse and compare the performance of these methods based on different fairness metrics for three datasets. We observe that adaptive sensitive reweighting model achieves better or similar trade-offs between accuracy and unfairness mitigation when compared to other fairness-aware approaches.

Keywords: Machine Learning · Classifier · Unfairness · Bias in data

1 Introduction

Machine learning is a brach of artificial intelligence in which systems can learn from data or experience to make decisions with minimal human intervention. It is growing across many fields because it can analyze large amount of data to identify patterns in a short interval of time. Some of the major appplications are: speech recognition, email spam and malware filtering, search engine result refininng and automated decisition making process and so on.
With the increase in shift towards automated decision making process of machine learning in various services that affects people's lives, there is an immense need for fairness concerns. The decisions should be unbiased and nondiscriminatory in relation to the sensitive features such as gender, sex, religion, race and so on. These sensitive features should be carefully treated, and if not constrained well, it leads to bias in the decision making process which gives an unfair treatment to certain people based on sensitive attributes. For example, training a logistic

regression classifier on the ProPublica COMPAS dataset of crime recidivism [12] yields differences between black and white defendants that amount to 17% for false positives and 25% for false negative rates [19].
Researchers have previously recognized that classification is often caused by data rather than classifier [10] [18]. Elimination of sensitive features from data is insufficient for avoiding misclassification, due to the indirect influencec of the sensitive information. For example, when determining credit scoring, let us remove the sensitive feature race. People of a specific race live in a specific area and address is used as a feature for training the prediction model, then we can expect unfair determinations even though race is not considered. This is called a red-lining effect [4] or indirect discrimination [20]. This happens due to inherent inability to treat datasets and we discuss some of the reasons in the section 5.
In this paper, we discuss the following methods to mitigate bias in the training data:

- Group Thresholding: The goal of this method is to predict a true outcome Y from features X based on labeled training data, while ensuring the prediction is "non-discriminatory" with respect to a specified protected attribute A [11].
- Regularizer: Adjusting regularizers to reduce indirect prejudice (statistical dependence between sensitive information) to restrict learner's beahviour [16].
- Covariance : Convex concave programming is used to reduce the different unfairness measures discussed in the next section [22] [23].
- Adaptive-Sensitive Reweighting : Assumes that there exists an underlying set of class labels corresponding to training samples, that if, predicted would yield unbiased classfification with respect to a fairness objective. Weights are obtained using the CULEP model that stands for Convex Underlying Label Error Pertubation [19].

The structure of the paper is as follows: Section 2 discusses the background and related work. Section 3 constitutes the different methodolgies explained in detail. Section 4 provides information about dataset editing deficiencies. Section 5 provides information on the experiments conducted along with the results depicting the performance analysis of each of the methods. Section 6 concludes the paper summarizing the results and the future work.

2 Background and Related Work

In this section we first elaborate on the different types of unfairness and their corresponding metrics used to measure them in automated decision making process. Throughout this paper, we consider binary classifiers that produce label estimations $\tilde{y_i} \in \{0, 1\}$ for samples i of features x_i and labels $y_i \in \{0, 1\}$. If a certain group of sample is associated with the sensitive attributes then they are considered as sensitive samples S and the non-sensitive complement S'.

2.1 Types of Unfairness

Classification unfairness is often expressed through the notions of disparate treatment, disparate impact and disparate mistreatment [23]. Let us take an example of [22] to illustrate the three types of unfairness. The classifier needs to decide whether or not to stop a person on suspicion of having an illegal weapon based on set of features like bulge in clothing and proximity to a crime scene. The ground truth tells whether a person actually possesses an illegal weapon or not.

User Attributes			Ground Truth (Has Weapon)	Classifier's Decision to Stop		
Sensitive	Non-Sensitive					
Gender	Clothing Bulge	Prox Crime		C_1	C_2	C_3
Male1	1	0	y	1	1	1
Male2	1	0	y	1	1	0
Male3	0	1	n	1	0	1
Female1	1	1	y	1	0	1
Female2	1	0	n	1	1	1
Female3	0	0	y	0	1	0

Table 1. Decision of three classifiers (C_1, C_2 and C_3) on whether (1) or not (0) to stop a pedestrian on the suspicion of possessing an illegal weapon

1. Disparate treatment elimination: ability of a trained classifier to yield the same output $\hat{y_i}$ for features x_i irrespective of the sample belonging to the sensitive group S or the non-sensitive group S' [19].

$$P(\hat{y_i}|x_i, i \in S) = P(\hat{y_i}|x_i) \tag{1}$$

   As seen in 1, C_2 and C_3 are unfair due to disparate treatment since C_2's and C_3's decisions for Male1 and Female1 are different even though they have the same values of non-sensitive attributes [22].
2. Disparate impact elimination: ability of a classifier to achieve statistical parity —citekamiran2012data [15] [16] i.e, assigns the same portion of the users to a class for sensitive and non/sensitive groups [19].

$$P(\hat{y_i} = 1|i \in S) = P(\hat{y_i} = 1|i \notin S) \tag{2}$$

   As depicted in Fig. 1, C_1 is unfair due to disparate impact because the fraction of males and females that were stopped are different (1.0 and 0.66 respectively) [22].
3. Disparate mistreatment elimination: ability of a classifier to achieve equal misclassification rates across sound ground truth tables(i.e. not suffering from dataset construction problems, such as historical biases) [23] [22] [19]. For example, if the race is a sensitive attribute for prediction of criminal

behaviour [12], disparate mistreatment elimination would ensure the same error rate between white and non-white defendants [19]. The most common mistreatment constraint is equal number of false positive rates(FPR) and false negative rates (FNR).

$$\begin{aligned} P(\hat{y_i} \neq y_i | y_i = 1, i \in S) = P(\hat{y_i} = 1 | i \neq y_i | y_i = 1, i \notin S) \\ P(\hat{y_i} \neq y_i | y_i = 0, i \in S) = P(\hat{y_i} = 1 | i \neq y_i | y_i = 0, i \notin S) \end{aligned} \tag{3}$$

Fig. 1 shows that C_1 and C_2 are unfair due to disparate mistreatment because their rate of erroneous decisions for males and females are different: C_1 has different false negatives for males and females (0.0 and 0.5 respectively) [22].

2.2 Metrics

Disparate treatment and disparate impact is measured using $p\%$ rule. The $p\%$ [1] rule is an empirical rule which does not allow sensitive group identification to be lower than a set percentage of non-sensitive group identification:

$$pRule = min\{\frac{P(\hat{y_1}|i \in S)}{P(\hat{y_1}|i \notin S)}, \frac{P(\hat{y_1}|i \notin S)}{P(\hat{y_1}|i \in S)}\} \tag{4}$$

Let us consider a specific instantiation supported by the U.S Equal Employment Opportunity Commission: the "80%-rule". The $p\%$ rule states that the ratio between the percentage of subjects having a certain sensitive attribute value assigned the positive decision outcome and the percentage of subjects not having the value also assigned the positive outcome should be non less than p:100 [23]. On the other hand, the disparate mistreatment elimination conditions in Eq. 3, it is measured using the following measures [19]:

$$\begin{aligned} D_{FPR} = P(\hat{y_i}|y_i = 1, i \in S) - P(\hat{y_i}|y_i = 1, i \notin S) \\ D_{FNR} = P(\hat{y_i}|y_i = 0, i \in S) - P(\hat{y_i}|y_i = 0, i \notin S) \end{aligned} \tag{5}$$

The overall disparate mistreatment is a combination of the above two metrics:

$$|D_{FPR}| + |D_{FNR}| \tag{6}$$

2.3 Related Work

Methodologies to reduce bias can be boradly classified into the following groups: **Preprocessing the training data**: this is based on the assumption that disparate impact of the trained classifier is due to the disparate impact on the training data. In order to avoid the disparate impact, the sensitive attributes are ignored. However, ignoring the sensitive attribute information may still lead to disparate impact in outcomes: since automated decision making systems are often trained on historical data , if a group with certain sensitive attribute was unfairly treated in the past, this unfairness may persist in future predictions

through indirect discrimination, leading to disparate impact [23]. These approaches include massaging the dataset [2] [13] [14] [24] by changing class labels that are identified as mislabeled due to bias and reweighting training samples so that more importance is places on sensitive sttributes [2] [14].
Training under fairness constraints: solves disparate impact by adjusting the training rules by editing the rules themselves [3] [23] or by introducing linear programs constraints [5] [8] [21] [23].
Edit posteriors: these methods attempt to edit posteriors to satisfy fairness coonstrains but requires information about the sensitive group to make appropriate decisions [6] [7] [9] [11]

3 Methodology

In this section, we will discuss some of the methods that have been proposed to mitigate bias in the traning data to achieve accuracy.

3.1 Covariance-Based Models

In a binary classification task, we need to find the mapping function $f(x)$ between user feature vectors $x \in \mathbb{R}^d$ and class labels $y \in \{-1, 1\}$. The task is achieved by utilizing a training set, $(x_i, y_i)_{i=1}^N$ to construct a mapping that works on unseen dataset. We need to look for a decision boundary defined by a set of parameters θ^*, that achieves the greatest classification accuracy in a test set by minimizing a loss function over a training set $L(\theta)$ i.e $\theta^* = argmin_\theta L(\theta)$ [23]. Given an unseen feature vector x_i from the test set, the classifer predicts $f_\theta(x_i) = 1$ if $d_\theta^*(x_i) \geq 0$ and $f_\theta(x_i) = -1$ otherwise, where $d_\theta^*(x)$ denotes the signed distance from the feature vector x to the decision boundary [23].
Decision boundary covariance: Decision boundary unfairness is defined as the covariance between the users' sensitive attributes, $\{z_i\}_{i=1}^N$ and the signed distance from the users' feature vectors to the decision boundary $\{d_\theta(x_i)\}_{i=1}^N$ which is represented as:

$$\begin{aligned} Cov(z, d_\theta(x)) &= \mathbb{E}[(z - \bar{z})d_\theta(x)] - \mathbb{E}[(z - \bar{z})\bar{d_\theta}(x)] \\ &= \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \end{aligned} \tag{7}$$

where $\mathbb{E}[(z - \bar{z})d_\theta(x)]$ cancels out since $\mathbb{E}[(z - \bar{z})] = 0$. In linear models for classification like logistic regression or Support Vector Machines (SVM), the decision boundary is just the hyperplane defined by $\theta^T x = 0$ which reduces the Eq. 7 to $\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})\theta^T(x_i)$ [23]. Eq. 7 represents a convex function with respect to the decision boundary parameters θ, since $d_\theta(x_i)$ is convex with respect to θ for all linear, convex margin-based classifiers [23].
Maximizing accuracy under fairness constraints: Maximize an accuracy object for example: pRule by finding the decision boundary parameters θ by

minimizing the corresponding loss function over the training set under fairness constraints. This is done by minimizing $L(\theta)$ as follows :

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \leq c \\ \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \geq -c \end{aligned} \tag{8}$$

where c is the covariance threshold which specifies an upper bound on the covariance between each sensitive attribute and the signed distance from the feature vector to the decision boundary [23]. As c decreases to 0, then it increases the pRule but will potentially suffer from a larger trade-off in accuracy.
Maximizing fairness under accuracy constraints: Underlying correlation between the class labels and the sensitive attributes in the training data is high. Even though it is very difficult to completely eliminate disparate impact, we need to make sure that decision making causes least possible disparate impact. In order to achieve this, the authors in [23] find the decision boundary parameters θ by minimizing the corresponding decision boundary covariance over the training set under constraints on the classifier loss function:

$$minimize|\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)| \tag{9}$$

where $L(\theta) \leq (1+\gamma)L(\theta)^*$. Here $L(\theta)^*$ denotes the optimal loss over the training set provided by the unconstrained classifier and $\gamma \geq 0$ speicfies the maximal additional loss with respect to the loss provided by the unconstrained classifier. When $\gamma = 0$, we can achieve maximum fairness with no loss in accuracy.

3.2 Reguralizer Method

This method is implemented as a regularizer to reduce indirect prejudice(statistical dependence between sensitive features). Index to quantify the degree of indirect prejudice is called prejudice index (PI). Y, X and S are random variables corresponding to a class, non-sensitive feature, respectively. A training data set consists of the instances of these random variables,i.e: $D = (y, x, s)$ [16]. The conditional probability of a class given non-sensitive and sensitive features is modeled by $M[Y|X, S; \odot]$, where $\odot$ is the set of modeled parameters which are estimated based on the maximum likelihood principle.

$$L(D; \odot) = \sum_{(y_i, x_i, s_i) \in D} lnM[y_i|x_i, s_i; \odot] \tag{10}$$

Two types of regularizers were adopted. The first is a standard $L_2|| \odot ||_2^2$ to avoid over-fitting. The second regularizer, $R(D, \odot)$ enforces fair classification.

Addition of the two regularizers to Eq.10, the objective funtion to minimize is obtained

$$L(D;\odot) + \eta R(D,\odot) + \frac{\lambda}{2}||\odot||_2^2 \tag{11}$$

where λ and η are positive regularization parameters.
A prejudice remover regularizer directly reduces the prejudice index which is denoted by R_{PR} [16]. The prejudice index is computed as follows:

$$PI = \sum_{Y,S} \hat{Pr}[Y,S] ln \frac{\hat{Pr}[Y,S]}{\hat{Pr}[S]\hat{Pr}[Y]} = \sum_{X,S} \hat{Pr}[X,S] \sum_{y} M[Y|X,S;\odot] ln \frac{\hat{Pr}[Y,S]}{\hat{Pr}[S]\hat{Pr}[Y]} \tag{12}$$

Eq 12 can be rewritten by replacing $\sum_{X,S} \hat{Pr}[X,S]$ by $({}^{1}/_{|D|}) \sum_{x_i,s_i \in D}$ and removing the scaling factor $({}^{1}/_{|D|})$, we get:

$$\sum_{x_i,s_i \in D} \sum_{y \in 0,1} M[y|x_i,s_i;\odot] ln \frac{\hat{Pr}[y,S]}{\hat{Pr}[y]} \tag{13}$$

$\hat{Pr}[y,S]$ is obtained is marginalizing $M[y|X,s_i;\odot]Pr^*[X|s]$ and then applying the sample mean to get the following equation:

$$\hat{Pr}[y,S] \simeq \frac{\sum_{x_i,s_i \in Ds.ts_i \in s} M[y|x_i,s;\odot]}{|(x_i,s_i) \in Ds.ts_i = s|} \tag{14}$$

$$\hat{Pr}[y] \simeq \frac{\sum_{x_i,s_i \in D} M[y|x_i,s_i;\odot]}{|D|} \tag{15}$$

From Eq.14 and Eq.15, the prejudice remover regularizer is computed as:

$$\sum_{x_i,s_i \in D} \sum_{y \in 0,1} M[y|x_i,s_i;\odot] ln \frac{\hat{Pr}[y|s_i]}{\hat{Pr}[y]} \tag{16}$$

This regularizer becomes large when a class is determined mainly based on sensitive features;thus, sensitive features becomes less influential in the final determination [16].

3.3 Group Thresholding

The author's goal in [11] is to predict a true outcome Y from features X based on labeled training data, while ensuring that the prediction is non-discriminatory with respect to a specific protected attribute A. It can be assumed that supervised learning setting has labeled data that can be used to construct a predictor $\tilde{Y}(X)$ or $\tilde{Y}(X,A)$.

Definition 1. *Equalized odds We say that a predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y [11].*

$$Pr\{\hat{Y} = 1|A = 0, Y = y\} = Pr\{\hat{Y} = 1|A = 1, Y = y\}, yin0, 1 \tag{17}$$

Definition 2. *Equal Opportunity We say that a predictor $\hat{Y}$ satisfies equal opportunity with respect to protected attribute A and outcome Y, if $Pr\{\hat{Y} = 1|A = 0, Y = 1\} = Pr\{\hat{Y} = 1|A = 1, Y = 1\}$*

Definition 3. *Oblivious We say that a predictor $\hat{Y}$ or score R is oblivious if it only depends on the joint distribution of (Y, A, Y) or (Y, A, R) respectively [11].*

The goal is to obtain equalized odds or equal opportunity predictor $\hat{Y}$ from a possibly discriminatory learned binary predictor $\hat{Y}$ or score R. The model does not require changing the training process as it might introduce additional complexity but it will construct a non-discriminatory predictor which is derived from $\hat{Y}$ or R.

Definition 4. *Derived Predictor A predictor $\tilde{Y}$ is derived from a random variable R and the protected attribute A if it is a possible randomized function of the random variables (R,A) alone. In particular, $\tilde{Y}$ is independent of X conditional on (R,A) [11].*

A protected attribute way of deriving a binary predictor from a score R would be to threshold it using $\hat{Y} = \mathbb{I}\{R > t\}$ [11]. If R satisfies the equalized odds, even the predictor will and the optimal threshold would be chosen to balance false and true positive rates so as to minimize the expected loss.

3.4 Adaptive Sensitive Reweighting + CULEP

Analysis is conducted on binary probabilistic classifier, which produces probability estimates $\hat{P}(Y = y_i) = 1 - \hat{P}(Y \neq y_i)$ for samples i with features x_i and each class label $Y \in \{0, 1\}$. This classifier estimates class labels as:

$$\hat{y_i} = argmax\hat{P}(Y = y_i) = argmin\hat{P}(Y \neq y_i), Y \in \{0, 1\} \tag{18}$$

A well-calibrated classifier has the misclassification error $P(Y \neq y_i)$ reaching a minimization target in the learning process. For the training sample i, features x_i and class labels y_i, there exists underlying class labels $\tilde{y_i}$ (unobservable) that yields estimated labels $\hat{y_i}$ which conforms to designated fairness and accuracy trade-offs. Training goals are two-fold: a) make the calssifier yield accurate predictions, minimize $\hat{P}(\hat{y_i} \neq y_i)$ and b) make classifier predictions approach the underlying labels, minimize $\hat{P}(\hat{y_i} \neq y_i)^2$. It is difficult to attain both of them when the original labels do not coincide with the underlying labels.

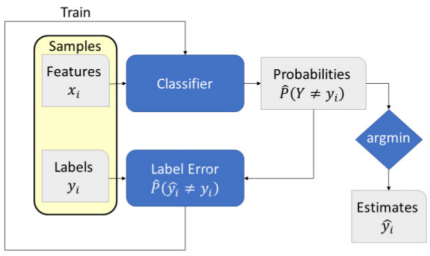

Fig. 1. Probabilistic classifier training

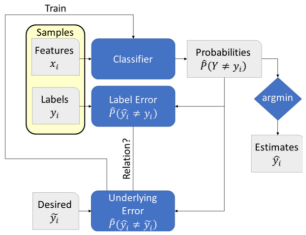

Fig. 2. Directly training on observable desired labels

Training the data labels could be achieved using the mechanism in Fig 1and training towards underlying labels could be done using the mechanism in Fig 2.However, estimating the underlying labels and directly using them for training is considered under data falsification under legal constraints. In order to solve this contradiction, weights w_i are trained on the data labels which makes them equivalent to the unweighted training on the underlying labels. The goal is now is to minimize the weighted error and also bridge the gap between the weighted data labels and unweighted underlying labels.

$$min \sum_{i=1} w_i \hat{P}(\hat{y}_i \neq y_i) \tag{19}$$

$$min \sum_{i=1} \{w_i \hat{P}(\hat{y}_i \neq y_i) - \hat{P}(\hat{y}_i \neq \tilde{y}_i)\}^2 \tag{20}$$

We equate Eq.20 to 0 which yields us:

$$w_i \hat{P}(\hat{y}_i \neq y_i) = \hat{P}(\hat{y}_i \neq \tilde{y}_i) \tag{21}$$

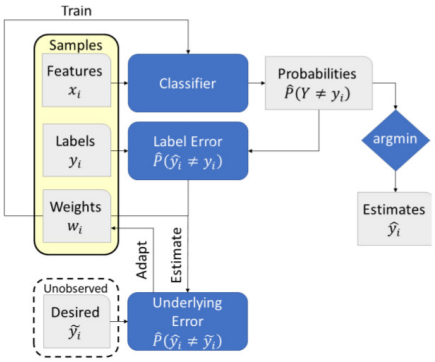

Fig. 3. Training on unobservable desired labels

The authors in [19] have proposed a model which employs conditional probabilities to make informed estimations based on $\hat{P}(\hat{y_i} \neq y_i)$. This model achieves the goals by estimating the underlying labels while training on weighted original labels as seen in Fig.3 .This process shifts the focus from the training scheme to discovering probability estimation model that can train towards the goal than searching explicitly searching for underlying labels. The advantages of using this model are three fold: a) the classifier is not accuse of being trained on falsified data, b) selection of estimation models that trains towards objectives that could not be formulated as deficiencies in training data and c)there is no need to introduce massaging heuristics to distribute the relabeling.

Algorithm 1 Adaptive Sensitive Reweighting

function REWEIGHT(Classifier C, Data D, Sensitive Group S)
  $w_i \leftarrow \forall i \in D$
  $w_i, prev \leftarrow 1 + \sqrt{e} \forall i \in D$
  while $\sum_{i \in D} \{w_i - w_i, prev\}^2 \geq e$ **do**
    Train C samples: $i = (x_i, y_i) \in D$ and weights $\frac{w_j}{\sum_{j \in D} w_j}$.
    Use C to Obtain $\hat{P}(\hat{y_i} \notin y_i)$.
    Estimate $\hat{P}(\hat{y_i} \notin \tilde{y_i})$ using $\hat{P}(\hat{y_i} \notin y_i) \forall i \in D$.
    $w_i, prev \leftarrow w_i \forall i \in D$
    $w_i \leftarrow {}^{P(\hat{y_i} \neq \tilde{y_i})}/_{P(\hat{y_i} \neq y_i)} \forall \in D$
  return trained classifier $C, \{w_i\}$

The weights are determined using the Convex Underlying Label Error Pertubation (CULEP) model. When the original labels coincide with the underlying labels we expect overfitting and similarly when they do not coincide, we expect underfitting. This is represented as follows:

$$(\hat{P}(\hat{y_i} \neq \tilde{y_i} | y_i = \tilde{y_i}) - \hat{P}(\hat{y_i} \neq y_i))(\hat{P}(\hat{y_i} \neq \tilde{y_i} | y_i \neq \tilde{y_i}) - \hat{P}(\hat{y_i} \neq y_i)) < 0 \quad (22)$$

To satisfy this property the authors in [19] propose conditional probabilities by pertubating classifier error of training samples i. To achieve this, they multiply it with values of a non-decreasing convex function $L_{\beta i}(p_i) \geq 0$, $L_{\beta i}(0) = 1$ of pertubation paramters $p \in [-1, 1]$ whose Lipschitz constant is proportional to β_i^3. Depending on overestimatio(+) or underestimation Eq. 22 can be represented as:

$$\begin{aligned} \hat{P}(\hat{y_i} \neq \tilde{y_i} | y_i = \tilde{y_i}) &= L_{\beta i}(\pm \hat{P}(\hat{y_i} \neq y_i)) \hat{P}(\hat{y_i} \neq y_i) \\ \hat{P}(\hat{y_i} \neq \tilde{y_i} | y_i = \tilde{y_i}) &= L_{\beta i}(\mp \hat{P}(\hat{y_i} \neq y_i)) \hat{P}(\hat{y_i} \neq y_i) \end{aligned} \quad (23)$$

The authors have selected different Lipschitz constants for the senstive group S and the non-sensitive group S' as follows:

$$\begin{aligned} \beta_i &= \beta_s | i \in S \\ \beta_i &= \beta_{s'} | i \notin S \end{aligned} \quad (24)$$

Conditional probability needs to consider the probability bias for inadequate labeling. Data mislabeling would occur with a fixed probability, depending on whether samples belong to the sensitive group which can be modeled as two Bernoulli processes, one for the sensitive group S and the other for the non-sensitive group S' as follows:

$$\begin{aligned} q_i &= q_s | i \in S \\ q_i &= q_{s'} | i \notin S \end{aligned} \tag{25}$$

Substituting the above in Eq.21, we get:

$$w_i \hat{P}(\hat{y_i} \neq y_i) = L_{\beta i}(\pm \hat{P}(\hat{y_i} \neq y_i)) \hat{P}(\hat{y_i} \neq y_i) q_i + L_{\beta i}(\mp \hat{P}(\hat{y_i} \neq y_i)) \hat{P}(\hat{y_i} \neq y_i)(1 - q_i) \tag{26}$$

Convex Unedrlying Label Error Pertubation (CULEP) model can be obtained through previous propositions as:

$$w_1 = \alpha_i L_{\beta i}(\hat{P}(\hat{y_i} \neq y_i)) + (1 - \alpha_i) L_{\beta i}(\hat{P}(-\hat{y_i} \neq y_i)) \tag{27}$$

where $\alpha_i = q_i$ or $(1 - q_i)$ depending on the sign of $\pm$.

4 Dataset Editing Deficiencies

In this section, we discuss some of the common shortcomings of dataset-editing fairness aware mechanisms discussed in the previous sections:

Limitations of Preprocessing: They are dependent on the types of bias in the training data and use statistical models to eliminate them. This is a good approach for simple dataset-related biases but fails to take into consideration more intricate source of awareness. For example, there may exist weaker feature correlations [3] that cause the bias against only a subset of the sensitive group. Moreover, certain types of biases are specific to certain classifiers. For example: linear classifiers have difficulty non-linear types of biases [19].

Heuristic Statistical Models: Although, classifier bias has a high correlation with the training bias, their structural difficulties may cause inadequate bias elimination. The statistical models employed arrive at a minimum condition that guarantees correct but not necessarily full treatment [19].

Inability to justify disparate mistreatment eliminaton: The relation between dataset bias and disparate mistreatment remains unclear till date [19]. This is because: a) we cannot attribute the similar misclassification rates on the biased data and b) since disparate mistreatment is not caused by disparate impact, methods to remove sensitive attributes from training data (like massaging approaches, dataset editing labels) i.e constructing datasets unbiased does not solve the disparate mistreatment problem.

5 Experimental Setup

5.1 Datasets

The experiment is performed with two synthetic datasets suffering from disparate mistreatment previously proposed in [22] along with three datasets: the *Adult*

income dataset [**?**], the *Bank* market dataset [**?**] and the ProPublica *COMPAS* dataset [12]. The two synthetic datasets suffering from disparate mistreatment comprises of 10,000 samples wuth 2 features, a binary sensitive label and a binary classification label. Their features are obtained through bivariate normal distributions, chosen so that their sensitive labels yield $D_{FPR}D_{FNR} < 0$ and opposite sign $D_{FPR}D_{FNR} > 0$ [19]. *SynthOpp* dataset is formed by opposite signs of disparate mistreatment between FPR and FNR whereas, *SynthSame* dataset is formed bz the same signs of FPR and FNR.
The Adult dataset comprises of 48,842 test samples with 14 features and a binary label indicating whether income is above 50K [19]. Gender is a sensitive feature in this dataset. The Bank dataset consists of 41,188 samples with 20 features and a binary label, indicating whether the client has subscribed to a term deposit [19]. In this dataset, age less than 25 and greater than 60 is considered to be sensitive attribute. The COMPAS dataset used in [23] consists of 6,150 samples with 5 features (age category, gender, race, priors and charge degree)and a binary label indicating whether the defendant reoffended wthin two years [19]. Here, race is considered as a sensitive attribute. Logistic regression without regularizer is employed as the base classifier.

5.2 Compared Methods

The following methods are compared against each other based on their performance to achieve the fairness constraints (accuracy, FPR, FNR) on the two synthetic and three datasets discussed previously.

1. **ASR+CULEP**: Adaptive Sensitive Reweighting along with the CULEP model to mitiagte disparate impact and mistreatment.
2. **Covariance**: Models proposed in [23] [22] which employs covariance to approximate linear program constraints to mitigate disparate impact and disparate mistreatment.
3. **Group Thresholding**: Model proposed in [11] to mitigate disparate mistreatment.
4. **Regularizer**:Method proposed in [16] to remove prejudice-related disparate impact but, it still suffers from disparate treatment.

6 Results

6.1 Results for Disparate Mistreatment

Experiments for disparate mistreatment is explored both with and without the disparate treatment. In the first case, sensitive information is not included in the training and validation datasets, where as in the second case it is. Removing the sensitive group feature yields inadequate levels of prediction for the dataset. Table 2 shows that when the sensitive information is available, ASR+CULEP outperforms covariance-based contraints in eliminating disparate mistreatments= and provides equally favourable results to group thresholding [19]. It reduces

the mistreatment by 12% on the COMPAS dataset in exchange for 1% accuracy compared to covariance based methods and yields identical results to the group thresholding. It performs well on all respects on SynthOpp dataset and in Syth-Same it reduces the overall mistreatment by 4% when compared to covariance-based methods.

	Disparate Treatment								
	COMPAS			SynthOpp			SynthSame		
Fairness Approach	acc	D_{FPR}	D_{FNR}	acc	D_{FPR}	D_{FNR}	acc	D_{FPR}	D_{FNR}
None	66%	17%	−25%	78%	−16%	19%	80%	25%	14%
ASR+CULEP	65%	−1%	−1%	81%	0%	0%	77%	0%	−16%
Covariance	66%	3%	−11%	80%	1%	2%	77%	14%	6%
Group Thresholding	65%	−1%	−1%	79%	0%	−1%	67%	2%	0%

Table 2. Disparate mistreatment elimination for logistic regression on both $|D_{FPR}|$ and $|D_{FNR}|$ constraints for disparate treatment

Table 3 depicts that when the disparate treatment is avoided, ASR+CULEP produces better accuracy vs. overall mistreatment elimination trade-offs compared to covariance based linear constraints [19]. It yields (1%) overall disparate mistreatment elimination while preserving accuracy on the SynthOpp dataset and trades 6% overall disparate mistreatment to gain 6% accuracy on the SynthSame dataset.

	Avoiding Disparate Treatment					
	SynthOpp			SynthSame		
Fairness Approach	acc	D_{FPR}	D_{FNR}	acc	D_{FPR}	D_{FNR}
None	78%	−16%	19%	80%	25%	14%
ASR+CULEP	77%	0%	−1%	75%	0%	−13%
Covariance	75%	−1%	1%	69%	−1%	6%
Group Thresholding	–	–	–	–	–	–

Table 3. Disparate mistreatment elimination for logistic regression on both $|D_{FPR}|$ and $|D_{FNR}|$ constraints for avoiding disparate treatment

6.2 Results for Disparate Impact

Table 4 shows that ASR+CULEP has the ability to peform better to remove disparate impact when compared to the state-of-the-art solutions because it achieves higher pRule for smaller accuracy trade-offs. As seen in Table 4, ASR+CULEP attains 6% pRule gain for the same accuracy on the Adult dataset and 16% pRule gain for 2% accuracy loss on the Bank dataset.

	Adult		Bank	
Fairness Approach	pRule	acc	pRule	acc
None	27%	85%	31%	91%
ASR+CULEP	100%	82%	99%	89%
Covariance	94%	82%	83%	91%
Group Thresholding	85%	83%	100%	91%

Table 4. Adult dataset disparate impact elimination for logistic regression

7 Conclusion

In this paper we have discussed one of the major problems of automated decision making process which causes misclassification error leading to an unfair advantage to a section of people bounded by sensitive attributes like gender, race, financial status and so on. We have learnt the different types of unfairness caused due to the misclassification problem and the corresponding metrics to measure them. Following that, we discussed some of the recent developments and state-of-the art solution to mitigate bias in the data along with the limitations of some of these methods on dataset editing deficiencies which hinders the methods to completely mitigate the unfairness constraints (disparate impact, disparate treatment and disparate mistreatment). Lastly, we performed an experiment on two synthetic (SynthOpp and SynthSame) and three real-world datasets (Adult, Bank and COMPAS) to evaluate the performance of four approaches (ASR+CULEP, Covariance, Group Thresholding and Regularizer) to mitigate disparate mistreatment and disparate impact. We observed that ASR+CULEP achieves better or similar trade-offs between accuracy and unfairness mitigation on the all the datasets.
Results indicate that there is merit in further developing non-heuristic dataset editing mechanisms as competent alternatives to existing fairness-aware approaches [19]. Further development can be done to adjust the CULEP parameters that guarantees convergence by training towards optimal sample weights rather than analytical derivation. As a next logical step, the performance of ASR+CULEP model can be tested on more datasets and with different base classifiers to identify limits and potential developmental areas to make it perform better than the state-of-the-art solution across many other fairness constraints.

References

1. Biddle, D.: Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing. Gower Publishing, Ltd. (2006)
2. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: 2009 IEEE International Conference on Data Mining Workshops. pp. 13–18. IEEE (2009)
3. Calders, T., Karim, A., Kamiran, F., Ali, W., Zhang, X.: Controlling attribute effect in linear regression. In: 2013 IEEE 13th International Conference on Data Mining. pp. 71–80. IEEE (2013)

4. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery **21**(2), 277–292 (2010)
5. Celis, L.E., Straszak, D., Vishnoi, N.K.: Ranking with fairness constraints. arXiv preprint arXiv:1704.06840 (2017)
6. Dionne, G., Rothschild, C.: Economic effects of risk classification bans. The Geneva Risk and Insurance Review **39**(2), 184–221 (2014)
7. Doherty, N.A., Kartasheva, A.V., Phillips, R.D.: Information effect of entry into credit ratings market: The case of insurers' ratings. Journal of Financial Economics **106**(2), 308–330 (2012)
8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226. ACM (2012)
9. Feldman, M.: Computational fairness: Preventing machine-learned discrimination (2015)
10. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 259–268. ACM (2015)
11. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems. pp. 3315–3323 (2016)
12. J. Larson, S. Mattu, L.K.J.A.: COMPAS dataset. https://github.com/propublica/compas-analysis (2017), [COMPAS dataset (2017)]
13. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication. pp. 1–6. IEEE (2009)
14. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012)
15. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: 2010 IEEE International Conference on Data Mining. pp. 869–874. IEEE (2010)
16. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 35–50. Springer (2012)
17. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 35–50. Springer (2012)
18. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)
19. Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, Y.: Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web. pp. 853–862. International World Wide Web Conferences Steering Committee (2018)
20. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568. ACM (2008)
21. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review **29**(5), 582–638 (2014)

22. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1171–1180. International World Wide Web Conferences Steering Committee (2017)
23. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. arXiv preprint arXiv:1507.05259 (2015)
24. Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: 2011 IEEE 11th International Conference on Data Mining. pp. 992–1001. IEEE (2011)