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

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	Classifier's		
Sensitive	Non-Sensitive		(Has Weapon)	Decision to Stop		
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:

1. 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].

2. 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].
3. 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 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:

$$\hat{y_i} = argmax\hat{P}(Y = y_i) = argmin\hat{P}(Y \neq y_i), Y \in \{0, 1\} \tag{7}$$

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.

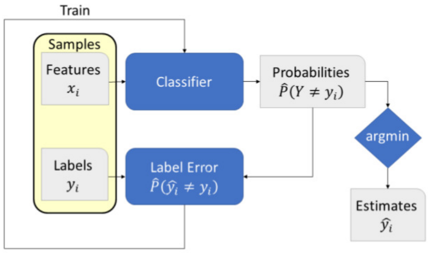

Fig. 1. Probabilistic classifier training

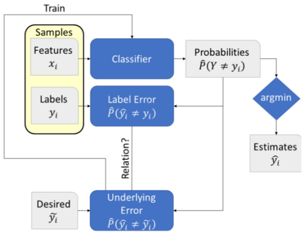

Fig. 2. Directly training on observable desired labels

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.

$$min \sum_{i=1} w_i \hat{P}(\hat{y}_i \neq y_i) \tag{8}$$

$$min \sum_{i=1} \{w_i \hat{P}(\hat{y}_i \neq y_i) - \hat{P}(\hat{y}_i \neq \tilde{y}_i)\}^2 \tag{9}$$

We equate Eq.10 to 0 which yields us:

$$w_i \hat{P}(\hat{y}_i \neq y_i) = \hat{P}(\hat{y}_i \neq \tilde{y}_i) \tag{10}$$

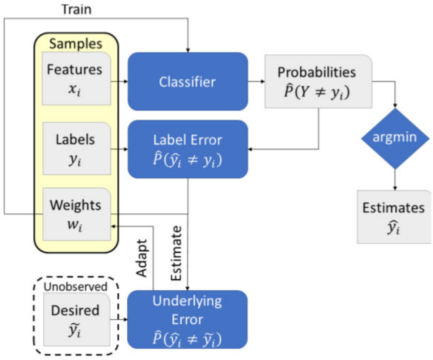

Fig. 3. Training on unobservable desired labels

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:

$$(\hat{P}(\hat{y_i} \neq \tilde{y_i} | y_i = \tilde{y_i}) - \hat{P}(\hat{y_i} \neq y_i))(\hat{P}(\hat{y_i} \neq \tilde{y_i} | y_i \neq \tilde{y_i}) - \hat{P}(\hat{y_i} \neq y_i)) < 0 \tag{11}$$

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. 11 can be represented as:

$$\begin{aligned} \hat{P}(\hat{y_i} \neq \tilde{y_i} | y_i = \tilde{y_i}) &= L_{\beta i}(\pm \hat{P}(\hat{y_i} \neq y_i)) \hat{P}(\hat{y_i} \neq y_i) \\ \hat{P}(\hat{y_i} \neq \tilde{y_i} | y_i = \tilde{y_i}) &= L_{\beta i}(\mp \hat{P}(\hat{y_i} \neq y_i)) \hat{P}(\hat{y_i} \neq y_i) \end{aligned} \tag{12}$$

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

$$\begin{aligned} \beta_i &= \beta_s | i \in S \\ \beta_i &= \beta_{s'} | i \notin S \end{aligned} \tag{13}$$

3.2 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either [17].

Subsequent paragraphs, however, are indented.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Table 2. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 2 gives a summary of all heading levels.

Displayed equations are centered and set on a separate line.

$$x + y = z \tag{14}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 4).

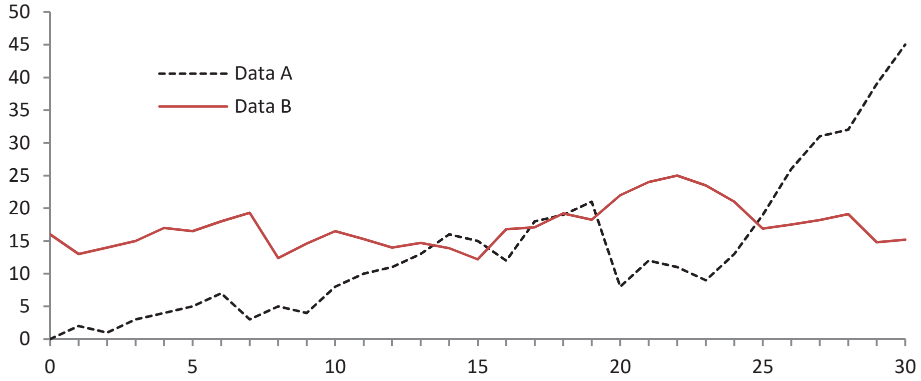

Fig. 4. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [**?**], an LNCS chapter [**?**], a book [**?**], proceedings without editors [**?**], and a homepage [**?**]. Multiple citations are grouped [**?**,**?**,**?**], [**?**,**?**,**?**,**?**].

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)