Fairness-Aware Learning with Prejudice Free Representations

Ramanujam Madhavan
Artificial Intelligence, LinkedIn
rmadhavan@linkedin.com

Mohit Wadhwa
Artificial Intelligence, LinkedIn
mwadhwa@linkedin.com

ABSTRACT

Machine learning models are extensively being used to make decisions that have a significant impact on human life. These models are trained over historical data that may contain information about sensitive attributes such as race, sex, religion, etc. The presence of such sensitive attributes can impact certain population subgroups unfairly. It is straightforward to remove sensitive features from the data; however, a model could pick up prejudice from latent sensitive attributes that may exist in the training data. This has led to the growing apprehension about the fairness of the employed models. In this paper, we propose a novel algorithm that can effectively identify and treat latent discriminating features. The approach is agnostic of the learning algorithm and generalizes well for classification as well as regression tasks. It can also be used as a key aid in proving that the model is free of discrimination towards regulatory compliance if the need arises. The approach helps to collect discrimination-free features that would improve the model performance while ensuring the fairness of the model. The experimental results from our evaluations on publicly available real-world datasets show a near-ideal fairness measurement in comparison to other methods.

ACM Reference Format:

Ramanujam Madhavan and Mohit Wadhwa. 2020. Fairness-Aware Learning with Prejudice Free Representations. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, October 19–23, 2020, Virtual Event, Ireland. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3340531.3412150>

1 INTRODUCTION

Machine learning models are increasingly being employed to aid decision-makers in all kinds of consequential decision settings such as healthcare, education, employment, criminal justice, etc. This amplifying use of model-based decision-making process has raised concerns from policymakers, legal experts, philosophers, and civil organizations about the prejudice (also called as bias) in the employed decision models that rely on the statistical patterns learned from the real-world data [3, 7, 9, 10].

A machine learning algorithm learns from historical instances (i.e., training data) to produce decisions for future instances. The learning algorithms are designed to learn statistical patterns in the training data, and since these decision systems use real-world data, the unintended consequence is that it can also potentially learn any unfair prejudice that may exist in the real-world data. One

well-known example is of ProPublica COMPAS's risk assessments – COMPAS algorithm was being used to forecast which criminals are most likely to re-offend – where it was found that the results were displayed differently for black and white offenders [1].

As the application of machine learning and artificial intelligence is becoming more prevalent, there is an increasing concern to make the decision process socially and legally fair i.e. non-discriminatory and fair in sensitive traits such as race, sex, religion, etc. Several techniques have been proposed to reduce or eliminate prejudice in machine learning models. A naïve approach is to avoid using sensitive features such as race, sex, religion, etc. while learning the model. However, machine learning algorithms are capable of learning complex relations present in the data, and there could be certain indirect data points that may serve as the proxy for these sensitive features. For example, the mention of the phrase “Convener of Grace Hopper Summit” in the resume may possibly indicate that the candidate is a female though there is no explicit mention of the sex. Hence, we need a more holistic mechanism to ensure that the data is free of such potential unfair prejudice inducing features.

To address these challenges, we propose an algorithm that automatically detects and treats the prejudice inducing features from the dataset and then learns a fair model. The process to detect and treat prejudice inducing features here is interpretable and easily verifiable. Our approach processes the dataset in an iterative way and removes information about the features which are latently related to the known sensitive attributes. The processed prejudice-free features can be used by the downstream tasks to build fair models. The information about the detected prejudice inducing features can also be used to understand the discrimination in the dataset and the dataset can thus be altered accordingly. By providing interpretable information about the prejudice inducing features in the dataset, our technique allows to collect better prejudice-free features that are more directly related to the model target variable than sensitive attributes. To validate our approach, we do a detailed experimental evaluation on real-world datasets that show significant improvement in fairness measures over prevailing methods. We have shown the comparison of our technique against other well-known methods [2, 4, 6, 8].

The remainder of this paper is organized as follows. We present the problem formulation and general framework in Section 2, the experiments evaluation dataset and results are presented in Section 3, followed by conclusions and future work in Section 4.

2 PREJUDICE FREE REPRESENTATIONS

As in the case of any supervised learning setting, we assume that we have access to a labeled dataset (X, S, Y) where X denotes feature vectors, S denotes known sensitive features, and Y represents the target variable. We also assume that the domain of each of the columns in S is limited to discrete values, which is a common

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6859-9/20/10.

<https://doi.org/10.1145/3340531.3412150>

scenario in most of the real-world tasks, for example, sex, race as sensitive features. We also assume that known sensitive attributes are already removed from X , thereby avoiding disparate treatment i.e. features in S and X are mutually exclusive. The technique we propose here effectively identifies and removes latent discriminating features present in the dataset (X, S, Y) , while learning the model to predict target variable Y .

2.1 Problem Formulation

For the dataset represented by (X, S, Y) , let $\mathcal{M}(Y|X, S)$ be a prediction model that models the conditional distribution of Y given X and S . Our objective is to ensure that model $\mathcal{M}$ is free of prejudice from sensitive features S and latent sensitive features among X . In other words, the conditions $Y \perp\!\!\!\perp S \mid X$ and $S \perp\!\!\!\perp X$ should be satisfied to make the model prejudice free. The first condition is easy to achieve by simply eliminating the sensitive features S from prediction model $\mathcal{M}$, however $Y \perp\!\!\!\perp S \mid X$ doesn't guarantee $Y \perp\!\!\!\perp S$: for example one or more sensitive features are linear of combinations of known non-sensitive features: $S = X + \epsilon_s$. Hence, we need a method to remove prejudice by making X and Y independent from S simultaneously [6].

Before defining a general framework, we pick logistic regression as an example classification algorithm to discuss the approach that ensures $Y \perp\!\!\!\perp S$ and $S \perp\!\!\!\perp X$ conditions. Let $D = \{(x, s, y)\}$ be the dataset comprising of the instances of random variables X , S and Y . We model $\mathcal{M}(Y|X, S; \theta)$ to predict the conditional probability of target variable Y where θ represents the set of model parameter that are learnt. In a typical case, the model parameters θ , are tuned to maximize the log-likelihood given by equation 1:

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} [\ln \mathcal{M}(Y|X, S; \theta)] \\ &= \arg \max_{\theta} \left[\sum_{(x, y) \in D} \left(y \ln \frac{1}{1 + e^{-x^T \theta}} + (1 - y) \ln \frac{1}{1 + e^{x^T \theta}} \right) \right] \quad (1) \end{aligned}$$

In our modified setting, we learn the prejudice free model in a two-step process through equation 2 & 3 where $D' = \{(x, s, y)\}$ be the dataset comprising of the instances of random variables X' , S and Y :

$$\hat{X}' = \arg \min_{X' \subset X} \left[\arg \max_{\omega} \left[\sum_{(x, s) \in D'} \left(s \ln \frac{1}{1 + e^{-x^T \omega}} + (1 - s) \ln \frac{1}{1 + e^{x^T \omega}} \right) \right] \right] \quad (2)$$

$$\hat{\theta}' = \arg \max_{\theta'} \left[\sum_{(x, y) \in D'} \left(y \ln \frac{1}{1 + e^{-x^T \theta'}} + (1 - y) \ln \frac{1}{1 + e^{x^T \theta'}} \right) \right] \quad (3)$$

Equation 2 minimizes the maximum likelihood of the model that learns ω to predict the sensitive variables S from X' where X' is a subset of features from X after eliminating the most predictive features in an iterative manner. Equation 3 learns the maximum likelihood parameter to predict Y based on prejudice free features $\hat{X}'$ which is the residual set of X , and therefore ensures the condition $Y \perp\!\!\!\perp S$. The Prejudice Free Representations (PFR) algorithm in the next section provides a general framework for this approach that is agnostic of the learning algorithm.

2.2 General Framework

In supervised learning, models evolve over iterations of feature engineering and hyperparameter tuning to optimize for one or more key performance metrics such as precision, recall, ROC, AUC,

Algorithm 1 Prejudice Free Representations

Notations:

ζ - function to train a classifier

ψ - function to get AUC score

η - function to get most important feature

Input:

X - feature vectors

S - known sensitive features

τ - auc thresholds for sensitive features

Initiaze: $X' = \frac{X - \min(X)}{\max(X) - \min(X)}$ // min-max scaling X

Output: prejudice free dataset

Process:

```

1:  $\forall (s_i, \tau_i) \in (S, \tau)$ 
2:    $\mathcal{M} := \zeta(X', s_i)$ 
3:    $auc := \psi(\mathcal{M}, X', s_i)$ 
4:   while  $auc > \tau_i$  do
5:      $f := \eta(\mathcal{M}, X')$ 
6:      $X' := X' \setminus f$ 
7:      $\mathcal{M} := \zeta(X', s_i)$ 
8:      $auc := \psi(\mathcal{M}, X', s_i)$ 
9:   end while
10: return  $X'$ 
```

etc. In feature engineering, we tend to add more features to enhance the model's predictive ability. Our approach here applies the inverse of feature engineering where we remove features iteratively so that the AUC of the model that predicts the sensitive variable falls until the point when the model has no capacity to distinguish between the classes of the sensitive variable.

Prejudice Free Representations (PFR) algorithm identifies the latent discriminating features and removes them from the dataset. To do so, it iteratively trains a classifier for each of the features in S as target variable and removes the most important feature from the feature set, X , until the AUC of the sensitive variable predicting model drops to the acceptable fairness measure threshold represented by τ . The process is repeated for each of the features in S . While the underlying learning algorithm attempts to maximize AUC, the PFR algorithm drives the AUC down by removing most predictive feature in every iteration. In this process, latent features that have a correlation to the sensitive features are removed in a thorough manner. The output of the algorithm can be fed to the learning algorithm to learn prejudice free prediction model for target variable Y . Taking an example Logistic Regression (LR) classifier, the PFR algorithm outputs a prejudice free dataset as per our modified setting in equation 2, and the output dataset is then used to train a LR classifier as per equation 3.

Algorithm 1 above provides a pseudocode implementation of the PFR algorithm. The algorithm provides a (best-first) greedy feature selection and removal mechanism to eliminate the prejudice inducing features, but the formulation and approach could be generalized to other feature selection methods as well. A few things to take note of are that the feature vectors of X are scaled before applying the algorithm as it requires relative feature importance values which can be known only when the features are normalized. The other thing is about the AUC threshold value τ . The value τ for a binary classifier is calculated as a function of target class imbalance. For

Figure 1: Accuracy vs. Disparate Impact evaluation on Adult Income Dataset

example, if the domain of target variable is $\{0, 1\}$, then the τ_i for a sensitive feature s_i is calculated as: $\tau_i = \frac{\max(|s_i \setminus \{0\}|, |s_i \setminus \{1\}|)}{|s_i|}$

The parameter τ could be varied to handle Accuracy vs. Fairness-Measure trade-off as discussed more in the experimental section below.

The algorithm generalizes well for any classification algorithm where the feature importance score could be known, such as logistic regression, decision tree-based classifiers, neural networks, etc. As we use the same algorithm that is used to train the target model to identify and remove the prejudice inducing features, the approach provides a verifiable guarantee that the target model is free of prejudice. The dataset processed by the algorithm can also be fed to downstream learning tasks such as regression and classification to build prejudice free models.

3 EXPERIMENTS

We evaluate our approach on a couple of publicly available real-world datasets. The first goal of our experiments is to show the trade-off between accuracy and fairness measure. To this end, we evaluate our approach with different τ parameter values and with multiple sensitive attributes. We demonstrate our approach with the Logistic Regression (LR) classifier and compare it with the standard LR classifier. Second goal of our experiments is to compare the performance metrics of our approach to other known fairness methods. We leverage the fairness-comparison open source library [5] to access datasets and implementations of other known methods.

3.1 Datasets

Adult Income This dataset contains information about individuals from the 1994 U.S. census. It consists of 32,561 instances and 14 attributes, including sensitive attributes race and sex. The prediction task is determining whether an individual makes more or less than \$50,000 per year [5].

ProPublica Recidivism This dataset includes features such as the number of juvenile felonies and the charge degree of the current arrest for 6,167 individuals with 15 attributes, including sensitive attributes race and sex. This dataset was collected by the use of the COMPAS risk assessment tool in Broward County, Florida [Angwin et al. 2016]. Each individual has a binary “recidivism” outcome, that is the prediction task, indicating whether they were rearrested

Attribute	Values
marital-status	Married-AF-spouse, Separated
native-country	Italy, Haiti, Portugal, Trinidad&Tobago, Greece, Yugoslavia, Laos, Scotland, Cambodia, Hungary, China, Canada, South, Poland, Vietnam, Ireland, El-Salvador, Taiwan, Guatemala, Philippines, Honduras, Peru, Hong, Cuba, India, Mexico, Jamaica, Thailand, France, Nicaragua, Columbia, Iran, Germany, United-States

Table 1: PFR Detected Prejudice Inducing Features on Adult Income Dataset with Race Sensitive Attribute at $\tau = 0.70$.

within two years after the first arrest [5].

For both the datasets, we do minority target classes oversampling since the target classes are skewed in terms of sensitive attributes as target variable. Categorical features in both the datasets are converted to one-hot encoded vectors to make them numerical before passing it to the learning algorithm. We consider *White* and *Male* as privileged classes for Race and Sex sensitive attribute in Adult Income Dataset, and *Caucasian* and *Male* as privileged classes for Race and Sex sensitive attribute in ProPublica Recidivism Dataset respectively.

3.2 Accuracy vs. Disparate Impact

Disparate Impact is one of the metrics to evaluate fairness under anti-discrimination laws. Disparate impact occurs when decision system outcomes disproportionately impact certain sub-populations. While there is no mathematical formula for quantifying disparate impact, we use 80% rule (or more generally p% rule) advocated by the US Equal Employment Opportunity Commission (EEOC) [4]. Disparate impact is calculated as the proportion of the unprivileged group that received the positive outcome by the proportion of the privileged group that received the positive outcome i.e. $\text{Disparate Impact} = \frac{\Pr(\hat{Y}=1|S=\text{unprivileged})}{\Pr(\hat{Y}=1|S=\text{privileged})}$, where $\hat{Y}$ represent the predicted variable and S represent the sensitive attribute class.

Figure 1 shows accuracy vs. disparate impact trade-off results on Adult Income Dataset. The results show this trade-off on three different set of sensitive attributes i.e. Race, Sex and Race & Sex, with different τ parameter values which influences the degree of prejudice removal. As expected, PFR iteratively removes prejudice

Attribute	Values
age	continous
workclass	Self-emp-inc, Without-pay, Local-gov
marital-status	Married-civ-spouse, Divorced, Separated, Widowed, Married-spouse-absent
hours-per-week	continuous
native-country	Cambodia, Thailand, Honduras, India, Hungary, Scotland, Dominican-Republic, Outlying-US(Guam-USVI-etc), China

Table 2: PFR Detected Prejudice Inducing Features on Adult Income Dataset with Sex Sensitive Attribute at $\tau = 0.70$.

inducing features from the dataset, as we lower the value of τ , thereby consistently pushing the disparate impact value close to 1.0, which is the ideal value for a fair model. The figures also show that the trade-off between accuracy and fairness measure can be decided based on different τ values.

Since PFR iteratively removes the prejudice inducing features, the drop in accuracy of the target model as we lower the value of τ could be compensated by adding more features that are non-discriminating in nature. Table 1 and 2 show the detected prejudice inducing features at $\tau = 0.70$ for Adult Income Dataset with race and sex as sensitive attributes. We can see that with race attribute, the detected prejudice inducing features are mostly about ‘native-country’ which is as expected. In the case of the sex attribute, what is interesting is that ‘hours-per-week’ is detected as prejudice inducing feature which would not very obvious for human eyes as it takes a continous value and would not have been trivially detected by just looking at the dataset.

With our approach, we advocate that the disparate impact of near one should be achieved at any cost, and accuracy should be improved only through non-prejudice inducing features.

3.3 Algorithms Comparison

We benchmark our algorithm to existing well-known fairness methods; these implementations are leveraged from [5].

In Figure 2a and 2b, we have plotted the output of PFR algorithms at different τ (0.52, 0.55, 0.6, 0.65, 0.85) values and connected the dots to approximately show the trade-off between disparate impact and accuracy as a continuous function though it is not practical to measure the disparate impact metric at all accuracy values. The slopes of the approximate red line plotted in the two figures are opposite as the range on x-axis are different. In addition, the output of other well-known algorithms supported by [5] are also plotted for comparison. In Figure 2a, we see that PFR beats all except Feldman in terms of disparate impact while giving comparable accuracy. However, in Figure 2b, other algorithms perform marginally better than PFR in terms of accuracy. In both the figures, we could see that PFR outperforms other algorithms in terms of disparate impact metric alone as it could push disparate impact closer to 1.0 at lower τ value. As previously called out, the accuracy could be improved by adding more features that are not correlated to the sensitive variables.

4 CONCLUSION & FUTURE WORK

We proposed an efficient approach that effectively identifies and treats the latent prejudice inducing features that (may) exist in

Figure 2: Accuracy vs. Disparate Impact comparison of different algorithms with Race Sensitive Attribute

training data. The approach provides a verifiable guarantee that the model is free of prejudice as we use the same algorithm that is used to train the model also to identify and eliminate the prejudice inducing features. Since the approach eliminates prejudice inducing features in an iterative fashion, it provides information about which features are linked to sensitive attributes. The information about the removed features complements the feature engineering process to add more features that are free of correlation with the sensitive attributes, therefore improving accuracy while maintaining fairness measure. Our experiments support that the approach helps to achieve a near-ideal fairness metric.

The approach can be used to identify and remove latent sensitive features and then train a fair model on the residual features using other learning algorithms, both supervised and unsupervised. Also, we could use our approach with in-processing fairness methods by feeding them with information about the detected prejudice inducing features as input. As an extension to the approach, we plan to conduct experiment by penalizing the prejudiced features instead of removing them completely, by dynamically learning the appropriate penalties for each such feature.

REFERENCES

- [1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica*, May 23 (2016), 2016.
- [2] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.
- [3] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies* 2015, 1 (2015), 92–112.
- [4] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.
- [5] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 329–338.
- [6] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.
- [7] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Queue* 11, 3 (2013), 10.
- [8] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. *arXiv preprint arXiv:1507.05259* (2017).
- [9] Wenbin Zhang and Eirini Ntoutsi. 2019. FAHT: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 1480–1486.
- [10] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *CoRR abs/1511.00148* (2015). *arxiv.org/abs/1511.00148* (2015).