Adversarial Fairness Network

Taeuk Jang[1], **Xiaoqian Wang**[1], **Heng Huang**[2]

[1]Elmore Family School of Electrical and Computer Engineering, Purdue University, USA
[2]Department of Computer Science, University of Maryland College Park, USA
{jang141, joywang}@purdue.edu, heng@umd.edu

Abstract

Fairness is becoming a rising concern in machine learning. Recent research has discovered that state-of-the-art models are amplifying social bias by making biased prediction towards some population groups (characterized by sensitive features like race or gender). Such unfair prediction among groups renders trust issues and ethical concerns in machine learning, especially for sensitive fields such as employment, criminal justice, and trust score assessment. In this paper, we introduce a new framework to improve machine learning fairness. The goal of our model is to minimize the influence of sensitive feature from the perspectives of both data input and predictive model. To achieve this goal, we reformulate the data input by eliminating the sensitive information and strengthen model fairness by minimizing the marginal contribution of the sensitive feature. We propose to learn the sensitive-irrelevant input via sampling among features and design an adversarial network to minimize the dependence between the reformulated input and the sensitive information. Empirical results validate that our model achieves comparable or better results than related state-of-the-art methods w.r.t. both fairness metrics and prediction performance.

Introduction

In recent years, machine learning has achieved unparalleled success in various fields, from image classification, speech recognition, to autonomous driving. Despite the rapid development, the discrimination and bias that exist in machine learning models are attracting increasing attention. Recent models have been found to be biased towards some population groups. Hendricks et al. (Hendricks et al. 2018) identified prediction bias towards gender in image captioning model, where the generation of caption is actually based on contextual information (e.g., location and scenes) but not the visual evidence related with the person in the image. In addition, ProPublica (J. Angwin and Kirchner 2016) analyzed a widely used criminal risk assessment tool for future crime prediction and discovered discrimination among different races. For defendants that do not commit a future crime, Black people are more likely to be mistaken by the model as potential future criminals than white people (i.e., a higher false positive rate in Black people than white people).

A model with merely good prediction performance is not convincing enough when we harness the power of machine learning. It is critical to guarantee that the prediction is based on appropriate information, and is not biased towards certain groups of the population characterized by sensitive features like race and gender. To improve model fairness, recent works propose strategies from different perspectives. For example, there are efforts on eliminating data bias by reweighing the samples (Kamiran and Calders 2012; Nam et al. 2020; Jang and Wang 2023; Chai and Wang 2022), generating fair data (Jang, Zheng, and Wang 2021; Sattigeri et al. 2019), or removing the disparity among groups (Feldman et al. 2015). While in-processing methods train a fair model by constraining the prediction not to base on sensitive information (Zhang, Lemoine, and Mitchell 2018; Mary, Calauzenes, and El Karoui 2019; Baharlouei et al. 2019; Cho, Hwang, and Suh 2020; Wang, Wang, and Liu 2022; Balunović, Ruoss, and Vechev 2022). Adel et al. (Adel et al. 2019) also propose an adversarial network that minimizes the influence of sensitive features on the prediction by characterizing the relevance between the latent data representation and the sensitive feature.

Fairness in machine learning is categorized based on different perspectives: group fairness and individual fairness. Group fairness (Li et al. 2021; Celis et al. 2021) guarantees that different groups of the population have equalized opportunity of achieving a favorable prediction result. Whereas for individual fairness (Dwork et al. 2012; Friedler, Scheidegger, and Venkatasubramanian 2016), the goal is to guarantee that similar individuals get similar prediction output. Following the mainstream of the literature, we here focus on mitigating bias in terms of group fairness in this work.

To improve fairness, previous works usually take either the data perspective or model perspective, i.e., modifying input to reduce data bias or optimizing model to reduce prediction bias. These strategies may not guarantee the learned input to be optimal for the model or the designed model to be optimal for the data, such that a fairness constraint usually introduces deterioration in prediction performance.

In contrast, we propose a novel adversarial network to reduce the bias simultaneously from both the *data* and the *model* perspective to improve fairness while maintaining the predictive performance. Specifically, we train a selector module to sample input features that do not propagate bias

while preserving predictive information. The selector reformulates the input with features that contain only sensitive-irrelevant information. To further strengthen the robustness towards the sensitive feature, we minimize the marginal contribution of the sensitive feature so that adding sensitive information will not affect the prediction results. The coupled optimization strategy from both the data and the model aspects improves fairness as well as prediction performance.

To the best of our knowledge, we are the first to introduce an end-to-end method that fuses fair feature selection and fair representation learning. Our model is different from existing fair pre/in-processing methods in three major perspectives. Firstly, our model eliminates data bias in the original data space, which preserves the natural meaning of features. This makes the reformulated fair data easier to be understood and interpreted by practitioners or end users – which is important in real-world applications. Secondly, our model requires a single classifier to address both fairness and performance objectives. Unlike many fair representation learning methods that require an auxiliary adversary classifier to predict the sensitive attribute, we do not need a separate sensitive attribute predictor, which can benefit in training efficiency and model complexity. Lastly, we focus on improving fairness in both data and model aspects. Specifically, we screen the data with the features that get the most impacted by the addition of the sensitive feature and build the predictive model to be the least affected by the sensitive feature.

Problem Definition

For a given dataset $[\mathbf{x}^{(1)},\ \mathbf{x}^{(2)},\ \ldots,\ \mathbf{x}^{(n)}]$ consisting of n samples from the input space $\mathcal{X} \subset \mathbb{R}^d$, each sample $\mathbf{x}^{(i)} = [x_1^{(i)},\ x_2^{(i)},\ \ldots,\ x_d^{(i)}]^\top$ is characterized by d features. The *sensitive feature* characterizes the *groups of population* for which predictions should remain unbiased. Common examples include race, gender, and age. The choice of such features depends on the specific prediction problem. Meanwhile, *sensitive-relevant features* are those that are not regarded as sensitive features, but carry information pertinent to the sensitive feature.

In terms of *prediction bias* in classification tasks, disparate treatment (Barocas and Selbst 2016) occurs when the model makes different predictions when merely the sensitive feature is altered while all other features were held consistent. Moreover, disparate impact arises when seemingly neutral decisions result in different merits to different demographics possibly by inference through sensitive-relevant features. One straightforward idea to improve fairness is *fairness through blindness*, i.e., simply exclude the sensitive feature from the input data. However, this cannot eliminate the prediction bias, as the sensitive-relevant features still provide sensitive information in the input data.

To address the problem, we propose to reduce the prediction bias from two aspects: reformulating the *input data* and strengthening the *model* fairness. We achieve the goal by simultaneously learning selective features $\tilde{\mathbf{x}}$ from the original data $\mathbf{x}$ and training $f^\phi : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is the output space, such that 1) the dependency between $\tilde{\mathbf{x}}$ and the sensitive information is minimized; 2) the influence of the sensitive information to the prediction of f^ϕ is minimized. By improving bias from both directions, the model prediction is based on the sensitive-irrelevant information and earns enhanced robustness towards the sensitive feature.

Adversarial Fairness Network

As discussed above, the simple strategy of fairness through blindness cannot take the existence of sensitive-relevant features into account. In order to reduce the prediction bias, we need to guarantee the prediction is not dependent on either the sensitive feature or the sensitive-relevant features. However, this is challenging since we usually do not have prior knowledge of what are the sensitive-relevant features. In this section, we propose a new FAIrness through AdverSarial network (FAIAS) model to efficiently filter out sensitive-relevant features while maintaining predictive performance.

Proposed FAIAS Model Design

The goal of reducing the prediction bias from both the input and model aspects can be formulated as two folds: 1) from the perspective of input, we propose to learn the new input $\tilde{\mathbf{x}}$ based on the original data $\mathbf{x}$ such that $\tilde{\mathbf{x}}$ contains only sensitive-irrelevant information; 2) for the prediction model, we minimize the marginal contribution of the sensitive feature such that adding the sensitive feature does not change the model prediction too much.

We propose to learn the new input $\tilde{\mathbf{x}}$ by sampling the features in the original data $\mathbf{x}$, i.e., selecting features with a selection function $S : \mathcal{X} \to \{0, 1\}^d$, such that the selected features contain only sensitive-irrelevant information.

Given a data sample $\mathbf{x} = [x_1,\ \ldots,\ x_d]^\top \in \mathcal{X}$, denote corresponding label $\mathbf{y} = [y_1,\ \ldots,\ y_c]^\top \in \mathcal{Y}$, and a selection set of dimension index $\mathbf{s} = \{i \mid \mathbb{1}[S_i(\mathbf{x}) = 1]\} \subset \{1,\ 2,\ \ldots,\ d\}$, where S_i indicates the output of i-th dimension the selection function S. Also, we denote $f^\phi(\mathbf{x}, \mathbf{s}) = f^\phi([x_{s_1},\ x_{s_2},\ \ldots,\ x_{s_m}])$ as the output of function f^ϕ when only the selected features designated by index vector $\mathbf{s} = \{s_1, s_2, \cdots, s_m\}$ are utilized among the entire input features space (the values of not selected features are filtered out by masking with 0). For $t \notin \mathbf{s}$, the marginal contribution to the prediction f^ϕ of the t-th feature of the sample $\mathbf{x}$ can be denoted as $\mathcal{L}\big(f^\phi(\mathbf{x}, \mathbf{s}), f^\phi(\mathbf{x}, \mathbf{s} \cup \{t\})\big)$, i.e., the change in the output when adding the t-th feature, where $\mathcal{L}$ is a distance to describe the difference between $f^\phi(\mathbf{x}, \mathbf{s})$ and $f^\phi(\mathbf{x}, \mathbf{s} \cup \{t\})$.

Let us denote the sensitive feature as x_k[1] for $k \in \{1,\ \ldots,\ d\}$, the goal of FAIAS is to minimize the distance between the distribution $p_\phi(\hat{y}|\mathbf{x} \oplus \mathbf{s})$ and $p_\phi(\hat{y}|\mathbf{x} \oplus \mathbf{s} \cup \{k\})$ for the label $\hat{y}$ predicted by f^ϕ, where the operator $A \oplus B$ is defined as the selection of the features of A per the index vector B. In order to achieve this goal, we propose to minimize $\mathcal{L}\big(f^\phi(\mathbf{x}, \mathbf{s}), f^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})\big)$, where $\mathbf{s}$ represents the selection set produced by the selection function S , which only selects features containing sensitive-irrelevant information. It is no-

[1]For simplicity, here we only consider one sensitive feature. Our FAIAS model can easily extend to the case involving multiple sensitive features.

table that reformulating the input with the selection function S has several advantages:

- Compared with learning a non-interpretable representation, the selection of features maintains *interpretation* of the input, since the natural meaning of features is kept;
- The selection function can be data-dependent, which maintains the *flexibility* such that we learn different sensitive-relevant features for different samples;
- Removing the sensitive-relevant features in the original data space is *theoretically supported* (Kusner et al. 2017), such that learning the observable non-descendants of sensitive feature (i.e., sensitive-irrelevant features in our paper) only needs partial causal ordering without further causal assumptions.

We introduce a probabilistic *selector* function $g^\theta : \mathcal{X} \to [0,1]^d$ with parameter θ which approximates the discrete selection function S. Given the feature vector as the input, the selector function outputs a continuous probability vector $\mathbf{p} = [p_1, p_2, \ldots, p_d] \in \mathbb{R}^d$, which represents the probability of sampling each feature to formulate the input. The probability of getting a joint selection vector $\mathbf{s} \in \{0,1\}^d$ is determined by the individual feature probabilities $\mathbf{p}$ with approximation by Bernoulli sampling as:

$$\pi_\theta(\mathbf{x}, \mathbf{s}) = \Pi_{j=1}^d \big(g_j^\theta(\mathbf{x})\big)^{s_j} \big(1 - g_j^\theta(\mathbf{x})\big)^{(1-s_j)}.$$

From here, the selection set $\mathbf{s}$ refers to the approximated selection set sampled by π_θ unless otherwise specified.

Objective Functions of FAIAS

To quantify the influence of the sensitive feature in the prediction, we formulate the sensitivity loss $l_{sens}(\theta, \phi)$ as:

$$l_{sens}(\theta, \phi) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}) \sim \mathcal{D}_a} \mathop{\mathbb{E}}_{\mathbf{s} \sim \pi_\theta(\mathbf{x}, \cdot)} \Big[JS\big(f^\phi(\mathbf{x}, \mathbf{s}) || h^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})\big)\Big], \tag{1}$$

where $JS(\cdot||\cdot)$ denotes Jensen-Shannon divergence that measures the similarity between two distributions $f^\phi(\mathbf{x}, \mathbf{s})$ and $h^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})$. Here, $\mathcal{D}_a$ denotes a sample distribution with sensitive feature a for $a \in \mathcal{A}$, and $\mathcal{A}$ is a set of all possible sensitive attributes. $h^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})$ is a strengthened predicted confidence by sharpening the probability distribution $f^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})$ as

$$h_l^\phi(\mathbf{x}, \mathbf{s} \cup \{k\}) = \frac{\exp\left(\gamma \cdot \mathbf{z}_l^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})\right)}{\sum_j \exp\left(\gamma \cdot \mathbf{z}_j^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})\right)}, \tag{2}$$

where γ is a sharpening hyperparameter and $\mathbf{z}_j^\phi$ denotes the output of j-th class of the classifier before softmax, i.e., $f^\phi(\mathbf{x}, \mathbf{s} \cup \{k\}) = softmax(\mathbf{z}^\phi(\mathbf{x}, \mathbf{s} \cup \{k\}))$. The function $h^\phi(\cdot)$ is employed to approximate the second input of JS as one-hot vector. Here, we empirically set $\gamma = 10$, which is a design choice by the domain.

The sensitive loss $l_{sens}(\theta, \phi)$ characterizes the marginal contribution of sensitive feature x_k to model prediction given features selected by $\mathbf{s}$. To optimize g^θ to approximate the selection function S and assign higher probability to only sensitive-irrelevant features, we propose an adversarial game between the selector g^θ and the predictor f^ϕ.

The goal of the prediction function f^ϕ is to minimize the sensitivity loss in (1) to ensure that adding the sensitive feature does not influence the prediction. In contrast, we optimize the selector function g^θ to maximize the sensitivity loss in (1), so as to select the subset of features that can be influenced the most by adding the sensitive feature. This allows the selector function g^θ can find the features that are not intrinsically relevant to the sensitive feature. If the selected subset includes the sensitive-relevant features, adding the sensitive feature will not bring significant change since the sensitive information is already inferred by the sensitive-relevant features. When updating the selector function g^θ to maximize the sensitivity loss, g^θ learns to exclude the sensitive information by assigning lower sampling probability to sensitive-relevant features and capturing the input on the basis of only sensitive-irrelevant information.

This setting enjoys the theoretical properties in Theorem 1 as follows. The theorem shows that minimizing the predictor f^ϕ w.r.t. $l_{sens}(\theta, \phi)$ provides a guarantee on the fairness of f^ϕ since l_{sens} upper-bounds the fairness violation. The proof of Theorem 1 is shown in supplementary material.

Theorem 1. *Consider a predictor $f : \mathbb{R}^d \to \mathbb{R}^c$ and a selected feature index $\boldsymbol{s} \in [d]$. Denote $f(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{s})$, $h^+(\boldsymbol{x}) = h(\boldsymbol{x}, \boldsymbol{s} \cup \{k\})$, and the sensitive loss $l_{sens} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \sim \mathcal{D}_a} \Big[\mathcal{D}_{JS}\big(f(\boldsymbol{x}) || h^+(\boldsymbol{x})\big)\Big]$, where $h(\cdot)$ is the sharpened predicted probability of $f(\cdot)$ as in* (2)*. The fairness violation of the predictor f measured by equalized odds difference (Hardt, Price, and Srebro 2016) expressed as*

$$\sum_{y \in \mathcal{Y}} \Big| \mathbb{E}\left[P(f(\boldsymbol{x}) = y | Y = y, A = 0)\right] - \mathbb{E}\left[P(f(\boldsymbol{x}) = y | Y = y, A = 1)\right] \Big|$$

is upper bounded by l_{sens}.

Moreover, we optimize the predictor f^ϕ and g^θ to maximize the utility to maintain the performance. We adopt generalized cross entropy (GCE) (Zhang and Sabuncu 2018) and cross entropy (CE) as the classification losses. Specifically, we minimize the following loss, $l_{gce}(\theta, \phi)$ and $l_{ce}(\theta, \phi)$:

$$l_{gce}(\theta, \phi) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}) \sim \mathcal{X} \times \mathcal{Y}} \mathop{\mathbb{E}}_{\mathbf{s} \sim \pi_\theta(\mathbf{x}, \cdot)} \Big[\frac{1 - f_y^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})^q}{q}\Big],$$

$$l_{ce}(\theta, \phi) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}) \sim \mathcal{X} \times \mathcal{Y}} \mathop{\mathbb{E}}_{\mathbf{s} \sim \pi_\theta(\mathbf{x}, \cdot)} \Big[- \log f_y^\phi(\mathbf{x}, \mathbf{s})\Big],$$

which both measure the performance of the prediction given the features selected by $\mathbf{s}$. Here, f_y^ϕ denotes the probability assigned to correct label $y \in \{1, \cdots, c\}$ and $q \in (0, 1]$ is a hyperparameter.

We train f^ϕ to minimize GCE loss w.r.t samples with sensitive attribute, i.e., $(\mathbf{x}, \mathbf{s} \cup \{k\})$ and Jensen-Shannon divergence between their counterpart without sensitive attribute,

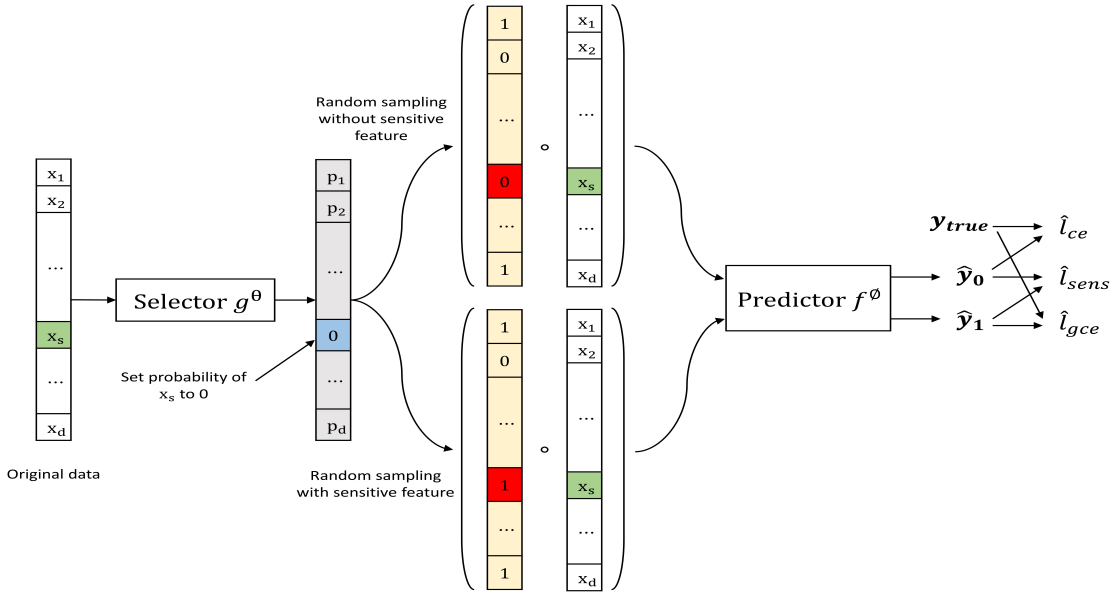


Figure 1: Illustration of the FAIAS model. FAIAS consists of a selector g^θ and a predictor f^ϕ. The selector g^θ takes the feature vector as an input and predict the probability for each feature to be selected, based on which we randomly sample the features. The predictor f^ϕ gets two inputs, one (shown in the upper dot product) is the reformulated input using the sampled features, the other (shown in the bottom dot product) is by adding the sensitive feature to the sampled features. The difference between the output of f^ϕ w.r.t. the two inputs is the sensitivity loss $\hat{l}_{sens}$, which shows the marginal contribution of the sensitive feature to the input. Two classification loss $\hat{l}_{ce}$ and $\hat{l}_{gce}$ takes samples features $\hat{y}_0$ and $\hat{y}_1$ respectively to maximize the utility.

i.e., $(\mathbf{x}, \mathbf{s})$. This enforces the classifier f^ϕ to focus more on high confidence samples while minimizing the effect of the sensitive attribute. On the other hand, the selector g^θ tries to maximize the difference of the prediction by whether the sensitive attribute is included or not. To ensure the selector retains the discriminative power, we minimize the cross entropy loss w.r.t. features without sensitive attribute, i.e., $(\mathbf{x}, \mathbf{s})$. Eventually, the selector would explore features that are useful for the task while the prediction is not vulnerable to the sensitive attribute, i.e., sensitive-irrelevant features. To summarize, our objective can be written as the following:

$$\min_{\theta} \quad l_{ce}(\theta, \phi) - l_{sens}(\theta, \phi),$$
$$\min_{\phi} \quad l_{gce}(\theta, \phi) + \lambda_{sens} l_{sens}(\theta, \phi),$$

where λ_{sens} is a hypterparameter to weight the sensitive loss. We illustrate the overview of FAIAS model in Figure 1, where $\hat{l}$ indicates the empirical loss.

Note that GCE was mainly employed to train a biased model as a reference to training a fair model (Nam et al. 2020; Liu et al. 2021) due to the nature of its focus on easy samples. It has been reported that GCE loss amplifies the bias by weighing more on the samples with high confidence as its gradient is the same as cross entropy except scaled with confidence. Unlike the convention that the literature had, interestingly, we found that GCE can be used to train a fair model. We discuss the contribution of GCE to train fair model later in the experiments.

In Algorithm 1, we summarize the optimization steps of FAIAS model. According to the update rules w.r.t. the gradients, the time complexity of our FAIAS model is linear w.r.t. the number of samples n, the number of parameters in θ and ϕ, as well as the number of iterations T.

Experiments

In this section, we conduct extensive experiments to validate the performance of our FAIAS model. The experiments evaluate: 1) whether FAIAS improves the prediction fairness among different groups w.r.t. sensitive features; 2) how will the prediction performance get affected by including fairness constraints in the FAIAS model.

Experimental Setup

Notably, our FAIAS model is proposed for group fairness in both the pre-processing and in-processing steps. Thus, we compare our model with recent methods for group fairness in pre-processing, in-processing, and post-processing approaches including AdvDeb (Zhang, Lemoine, and Mitchell 2018), CEOP (Pleiss et al. 2017), LAFTR (Madras et al. 2018), LfF (Nam et al. 2020), and baseline with the same structure as f^ϕ of FAIAS, but takes entire features.

To evaluate the models, we employ three fairness benchmark datasets. Adult (Kohavi 1996); COMPAS[2]; CelebA (Quadrianto, Sharmanska, and Thomas 2019). To evaluate fairness, we adopt equalized odds difference (EOD) (Hardt, Price, and Srebro 2016). This metric considers both TPR

[2]https://github.com/propublica/compas-analysis

Algorithm 1: Optimization Algorithm of FAIAS Model

Input dataset $\mathcal{Z} = (\mathcal{X} \times \mathcal{Y}) = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, where $\mathbf{x} \in \mathbb{R}^d$, k-th index of i-th sample $\mathbf{x}_{i,k}$ indicates sensitive information, and $\mathbf{y}$ is one-hot vector. Learning rate α_θ and α_ϕ, batch size n_b, and sensitive attribute $\mathcal{A} = \{2, 1\}$.
Output selector g^θ and predictor f^ϕ.
Initialize parameter θ to one vector and ϕ randomly.
while not converge **do**
for $(\mathbf{x}_{t_i}, \mathbf{y}_{t_i})$ in the t-th mini-batch $\mathcal{Z}_t$ **do**
1. Calculate the selection probability vector $g^\theta(\mathbf{x}_{t_i}) = [p_{t_i}^1, p_{t_i}^2, \ldots, p_{t_i}^d]$.
2. Sample the selection vector $\mathbf{s}_{t_i} \in \mathbb{R}^d$ with

$$\mathbf{s}_{t_i}^j \sim Bernoulli(p_{t_i}^j), \quad \text{for } j = 1, ,2, \ldots, d.$$

3. Calculate

$$\hat{l}_{ce}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}, \mathbf{y}_{t_i}) = -\sum_{l=1}^{c} (y_{t_i})_l \log f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}),$$

$$\hat{l}_{sens}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) = JS\big(f^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) || f^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i} \cup \{k\})\big).$$

$$n_a^k = \sum_i \mathbb{1}(x_{t_i,k} = a), \quad \text{for } a \in \mathcal{A}.$$

end for
4. Update the parameter θ with gradient ascent

$$\theta \leftarrow \theta + \alpha_\theta \cdot \Bigg(\sum_{a \in \mathcal{A}} \frac{1}{n_a^k} \sum_i \hat{l}_{sens}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) \mathbb{1}(x_{t_i,k} = a) - \frac{1}{n_b} \sum_i \hat{l}_{ce}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}, \mathbf{y}_{t_i}) \Bigg) \nabla_\theta \log \pi_\theta(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}).$$

5. Update the parameter ϕ with gradient descent

$$\phi \leftarrow \phi - \alpha_\phi \nabla_\phi \Bigg(\frac{1}{n_b} \sum_i \hat{l}_{gce}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i} \cup \{k\}, \mathbf{y}_{t_i}) + \lambda_{sens} \sum_{a \in \mathcal{A}} \frac{1}{n_a^k} \sum_i \hat{l}_{sens}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) \mathbb{1}(x_{t_i,k} = a) \Bigg).$$

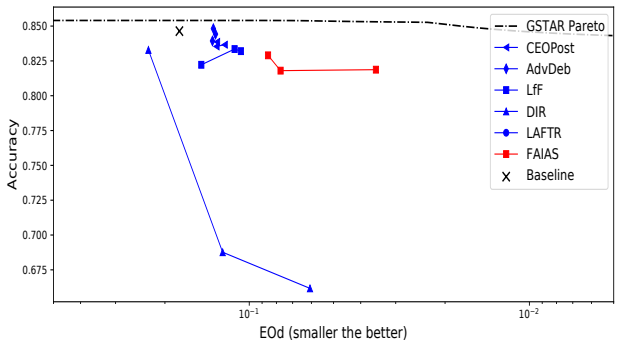
end while

(equal opportunity) difference and FPR (False Positive Rate) difference. Refer to supplementary material for the details of the experimental setup.

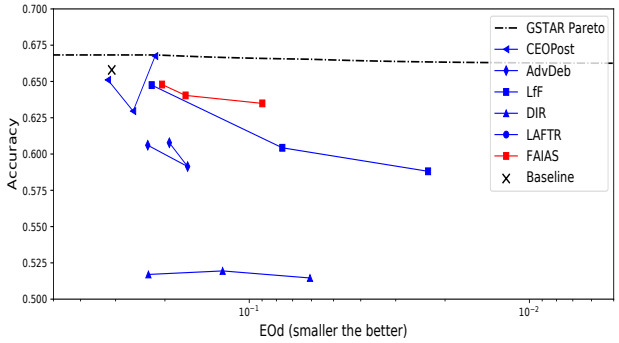
Quantitative Comparison on Tabular Benchmarks

We compare the model performance and summarize the results in Figure 2. We plot the *Pareto frontier* of each method to evaluate the accuracy-fairness trade-off by varying the hyperparameter for fair regularizer for the methods. GSTAR *Pareto frontier* (Jang, Shi, and Wang 2022) depicts the best achievable trade-offs in a model-specific manner, i.e., post-processing of the outcome of the baseline.

The experimental results demonstrate that our approach,



(a) Adult Dataset (Sex)

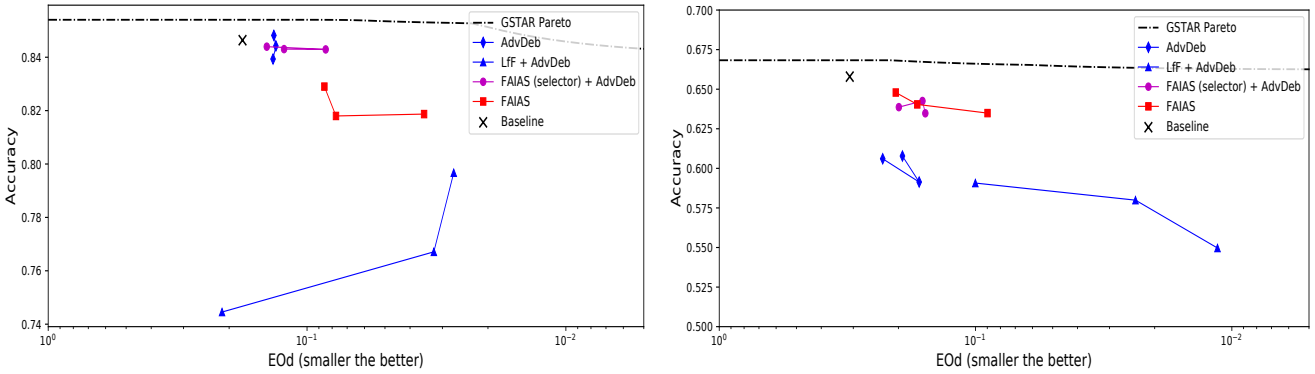


(b) COMPAS Dataset (Race)

Figure 2: The fairness-accuracy trade-off comparison. GSTAR Pareto illustrates the best achievable trade-off in a model-specific manner and upper right region is desired. The result shows that FAIAS achieves the best trade-off.

which incorporates both data and model fairness effectively mitigates the fairness violation while preserving comparable accuracy. Specifically, FAIAS achieves the least fairness violation at a similar accuracy level among the comparing methods. Notably, when compared to the baseline that has the same structure with f^ϕ but utilizes the entire input space, we achieve significant improvement in the fairness violation. This validates the effectiveness of fair feature selection of FAIAS that the adversarial network for feature sampling conducted by selector g^θ. By successfully eliminating the sensitive information, FAIAS ensures the prediction performance is equalized across different groups of the population.

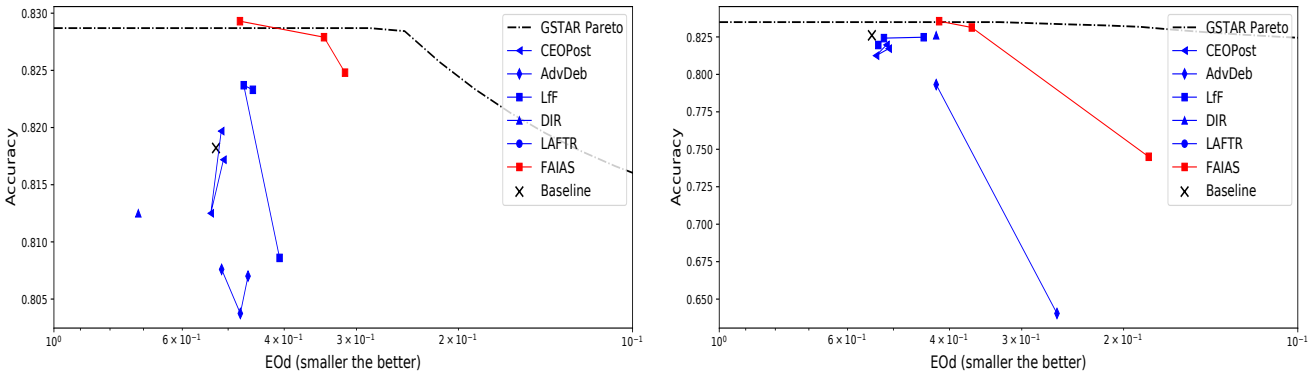
It is notable that even though the removal of sensitive-relevant features sometimes harms the performance because sensitive-relevant features can also be target-relevant, it is beneficial for fairness as demonstrated by the significant improvement in fairness achieved by FAIAS compared to the Baseline in Figure 2. Besides, FAIAS is designed to be able to select a set of features to improve fairness (by eliminating the sensitive-relevant features) while minimizing compromise in discriminative power (by optimizing the predictor using the sensitive-irrelevant information). We show more results in the supplementary material.



(a) Adult Dataset (Sex)

(b) COMPAS Dataset (Race)

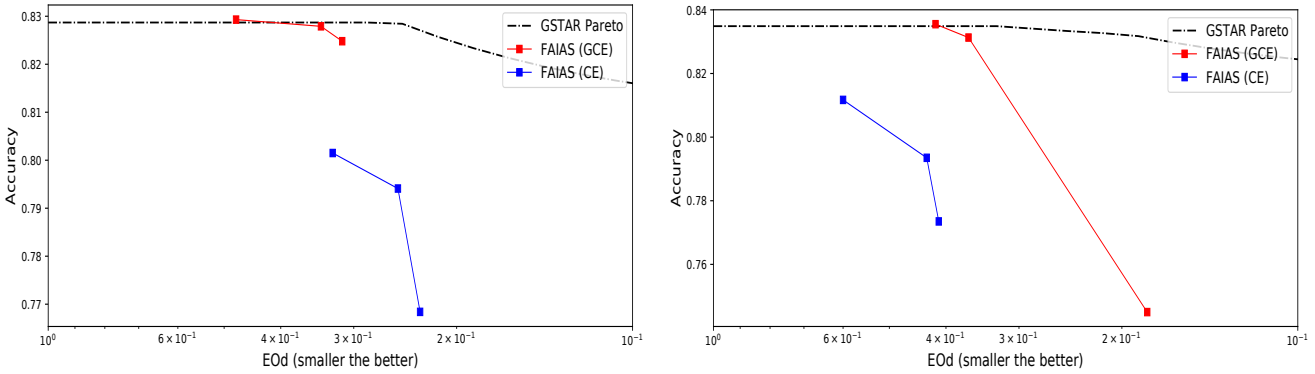
Figure 3: Comparison of different combination of pre-processing (LfF) and in-processing (AdvDeb) methods with FAIAS on adult and COMPAS datasets. We observe that combining the selector of FAIAS, g^{θ}, with AdvDeb (FAIAS (selector) + AdvDeb) outperforms a naive combination of state-of-the-art methods (LfF + AdvDeb).



(a) VGG19

(b) ResNet50

Figure 4: Comparison of the fairness-accuracy trade-off on CelebA dataset. We use the pre-trained models (VGG19 and ResNet50) to extract 1024 latent features. The sensitive feature is *sex*. GSTAR Pareto illustrates the best achievable trade-off in a model-specific manner and upper right region is desired. The result shows that FAIAS achieves the best trade-off.



(a) VGG19

(b) ResNet50

Figure 5: Comparison of employing GCE or CE as a classification loss for f^{ϕ}. The result shows that GCE helps to train a fair model, which contradicts the previous belief.

Ablation Studies

Since FAIAS can be viewed as a two-step method: 1) filter out sensitive-relevant features; 2) learn a fair classifier based on the input, one might think that the combination of other pre-processing and in-processing methods can act similarly to the combination of our selector-predictor structure. Thus, we explore the contribution of each module of FAIAS, i.e., selector g^θ and predictor f^ϕ.

To this end, we compare FAIAS with state-of-the-art pre-processing and in-processing methods, LfF and AdvDeb, respectively. LfF (Nam et al. 2020) reweighs sample loss of fair classifier based on the prediction of a biased vanilla classifier. AdvDeb (Zhang, Lemoine, and Mitchell 2018) debiases the gradient update by removing the conflicting direction of the performance loss to the fairness loss. For a fair comparison, we use the same structure as predictor f^ϕ for the classifier of AdvDeb.

In Figure 3, we depict the fairness-accuracy trade-offs of the methods. There are three combinations of the two-step methods: 1) LfF + AdvDeb; 2) FAIAS (selector) + AdvDeb; 3) FAIAS. In the case of FAIAS (selector), the FAIAS model is pre-trained, and it is not trained when delivering filtered features to train AdvDeb.

In the experiment, we observe that naively combining two fair methods (LfF + AdvDeb, triangle) harms performance significantly despite improved fairness compared with AdvDeb (diamond). However, interestingly, when filtering sensitive-relevant features with FAIAS (selector), not only is fairness improved but sometimes even the accuracy is improved (circle). For example in COMPAS dataset, FAIAS (selector) + AdvDeb improves both fairness and accuracy compared to AdvDeb which takes the entire features of the data. This validates that the selector successfully removes features that could harm fairness while preserving target-related information. With the final structure of FAIAS (square), i.e., the combination of our selector g^θ and predictor f^ϕ, we achieve the best fairness at the comparable accuracy level and vice-versa, since they were trained in an end-to-end fashion.

Another advantage of FAIAS is that a pre-trained selector is applicable to any method because it adopts the original feature space. Given the selector, we can efficiently exclude the sensitive-relevant features that can infer sensitive attributes while maintaining the target-related features to minimize performance degradation.

Image Classification with FAIAS

We further investigate the performance of FAIAS for the image classification task on CelebA dataset. Here, we conduct attractiveness classification and consider *sex* as the sensitive attribute. To evaluate the performance, we extract 1024 features from pre-trained vanilla models (VGG19 and ResNet50). To explicitly provide the sensitive information, we concatenate the sensitive feature to the latent space and train all methods on the 1025 features.

In Figure 4, we compare the fairness-accuracy trade-off of the methods on CelebA dataset. The results demonstrate that FAIAS achieves the best trade-off similar to the tabular benchmarks. It is interesting to note that we could significantly improve the fairness violation from the baseline while improving the performance. For example in VGG19, we could reduce the fairness violation by almost 50%, while improving the accuracy from the baseline. This reveals that FAIAS is able to enhance both fairness and prediction performance in vision tasks, the attractiveness classification.

Validation of GCE as Fair Loss

To optimize f^ϕ, we study the role of GCE loss in our approach. In the previous works, GCE (Zhang and Sabuncu 2018) was commonly considered as a loss function to train a biased model as it concentrates more on the samples with stronger agreement (Nam et al. 2020; Liu et al. 2021; Roh et al. 2020). However, interestingly, we empirically found that GCE could help to train a fair model instead. In Figure 5, we compare FAIAS model with different classification loss for training f^ϕ. As we proposed, the model that adopts GCE (red points) achieves both better fairness and accuracy.

We believe that this result is due to the two-step structure of our model. To begin with, our approach takes the min-max optimization on Jensen-Shannon divergence, which ensures that the selector filters out the biased features and the classifier learns fair prediction. Additionally, even though we focus on easier samples with GCE, we minimize the Jensen-Shannon divergence on the cross entropy. This indicates that instead of the behavior of amplifying the bias, GCE cooperates to focus on simpler features from the output of the selector that could help the generalization and performance boost. These two factors worked synergistically to yield improved performance with less fairness violation.

Conclusion

In this paper, we propose FAIAS, a novel adversarial network approach for fairness that combines both the data and model perspectives to achieve fair feature selection and classification. Our model comprises two primary components: a selector function and a prediction function. The selector function is optimized from the data perspective to select only those features that contain sensitive-irrelevant information. The prediction function, on the other hand, is optimized from the model perspective to minimize the marginal contribution of the sensitive feature and improve prediction performance.

Our experiments demonstrate that the FAIAS model achieves comparable or superior results to existing methods for both prediction performance and fairness metrics in various datasets. Furthermore, the fair feature selection procedure provides valuable insights into the original feature space with regard to sensitive information and target labels. By accurately filtering out sensitive-relevant features, we obtain a better understanding of the feature space and eliminate sources of bias in classification.

While our FAIAS model is proposed for the supervised learning scenario, our future work will explore the extension to unsupervised learning. Specifically, we aim to learn a set of meaningful and interpretable features that preserve data structure for unsupervised learning tasks such as clustering while eliminating bias in the selected features.

Acknowledgements

This work was partially supported by Purdue's Elmore ECE Emerging Frontiers Center, and NSF IIS 1955890, IIS 2146091. H.H. was supported by NSF IIS 2347592, 2348169, 2348159, 2347604, CNS 2347617, CCF 2348306, DBI 2405416.

References

Adel, T.; Valera, I.; Ghahramani, Z.; and Weller, A. 2019. One-network adversarial fairness. In *AAAI*, volume 33, 2412–2420.

Baharlouei, S.; Nouiehed, M.; Beirami, A.; and Razaviyayn, M. 2019. Rényi Fair Inference. In *ICLR*.

Balunović, M.; Ruoss, A.; and Vechev, M. 2022. Fair Normalizing Flows. In *International Conference on Learning Representations*.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California law review*, 671–732.

Celis, L. E.; Huang, L.; Keswani, V.; and Vishnoi, N. K. 2021. Fair classification with noisy protected attributes: A framework with provable guarantees. In *ICML*, 1349–1361. PMLR.

Chai, J.; and Wang, X. 2022. Fairness with adaptive weights. In *ICML*, 2853–2866. PMLR.

Cho, J.; Hwang, G.; and Suh, C. 2020. A fair classifier using kernel density estimation. *NeurIPS*, 33: 15088–15099.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *KDD*, 259–268.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *NIPS*, 29.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 793–811.

J. Angwin, S. M., J. Larson; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Jang, T.; Shi, P.; and Wang, X. 2022. Group-aware threshold adaptation for fair classification. In *AAAI*, volume 36, 6988–6995.

Jang, T.; and Wang, X. 2023. Difficulty-based Sampling for Debiased Contrastive Representation Learning. In *CVPR*, 24039–24048.

Jang, T.; Zheng, F.; and Wang, X. 2021. Constructing a Fair Classifier with Generated Fair Data. In *AAAI*, volume 35, 7908–7916.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *KAIS*, 33(1): 1–33.

Kohavi, R. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, 202–207.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *NIPS*, 4066–4076.

Li, B.; Li, L.; Sun, A.; Wang, C.; and Wang, Y. 2021. Approximate Group Fairness for Clustering. In *ICML*, 6381–6391. PMLR.

Liu, E. Z.; Haghgoo, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021. Just train twice: Improving group robustness without training group information. In *ICML*, 6781–6792. PMLR.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Mary, J.; Calauzenes, C.; and El Karoui, N. 2019. Fairness-aware learning for continuous attributes and treatments. In *ICML*, 4382–4391. PMLR.

Nam, J. H.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from Failure: De-biasing Classifier from Biased Classifier. In *NeurIPS*.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *NIPS*, 5680–5689.

Quadrianto, N.; Sharmanska, V.; and Thomas, O. 2019. Discovering fair representations in the data domain. In *CVPR*, 8227–8236.

Roh, Y.; Lee, K.; Whang, S.; and Suh, C. 2020. Fr-train: A mutual information-based approach to fair and robust training. In *ICML*, 8147–8157. PMLR.

Sattigeri, P.; Hoffman, S. C.; Chenthamarakshan, V.; and Varshney, K. R. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM J Res Dev* ., 63(4/5): 3–1.

Wang, J.; Wang, X. E.; and Liu, Y. 2022. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, 23114–23130. PMLR.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*, 335–340.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *NeurIPS*, 31.