Fairness without Demographics through Knowledge Distillation

SeongSik Choi

January 30, 2023

Seoul National University

Introduction

Consider the classification problem without causing disadvantage for each sensitive class.

Most fairness strategies in machine learning models have focused on achieving fairness objectives when the sensitive information is available.

However, in practice, due to legal or privacy concerns, when sensitive information is not available, it is crucial to find alternative objectives to ensure fairness.

Existing methods on fairness without demographics can be divided into two categories: **Max-Min fairness** and **fairness with proxy sensitive attribute.**

ARL : Max-Min Fairness Approach

We are given non-protected features $x_i = (x_i^1, \cdots, x_i^K)$, protected features s_i, and class labels $y_i \in \{0, 1\}$, $1 \leqslant i \leqslant n$.

Definition (Rawlsian Max-Min Fairness) Suppose H is a set of classifiers, and $U_{\mathcal{D}_s}(h)$ is the expected utility of the classifier h for the individuals in group s, then a classifier h^* is said to satisfy Rawlsian Max-Min fairness principle if it maximizes the utility of the group with the lowest utility.

$$h^* = \arg \max_{h \in H} \min_{s \in S} U_{\mathcal{D}_s}(h)$$

where $\mathcal{D}_s = \{(x_i, y_i) : s_i = s\}_{i=1}^n$

ARL : Max-Min Fairness Approach

Replacing the expected utility with an appropriate loss function $L_{\mathcal{D}_s}(h)$ over the set of individuals in group s, we can formulate our fairness objective as:

$$\min_{h \in H} \max_{s \in S} L_{\mathcal{D}_s}(h) = \min_{\theta} \max_{\lambda : \|\lambda\|_1 = 1} \sum_{s \in S} \lambda_s L_{\mathcal{D}_s}(h_\theta)$$

$$= \min_{\theta} \max_{\lambda : \|\lambda\|_1 = 1} \sum_{i=0}^{n} \lambda_{s_i} \ell\left(h_\theta\left(x_i\right), y_i\right)$$

However, we cannot optimize this objective directly because we do not observe s_i.

ARL : Max-Min Fairness Approach

In adversarial reweighted learning(ARL), we optimize this objective instead:

$$\min_{\theta} \max_{\phi} \sum_{i=1}^{n} \lambda_{\phi}\left(x_i, y_i\right) \cdot \ell\left(h_{\theta}\left(x_i\right), y_i\right) \left(=: J(\theta, \phi)\right)$$

$$\lambda_{\phi}\left(x_i, y_i\right) = 1 + n \cdot \frac{f_{\phi}\left(x_i, y_i\right)}{\sum_{i=1}^{n} f_{\phi}\left(x_i, y_i\right)} \text{ (ARL)}$$

Note that this optimization includes maximization for $\left\{\lambda_{\phi}\left(x_i, y_i\right), 1 \leqslant i \leqslant n\right\}$, which is different from maximization for $\left\{\lambda_{\phi}\left(x_{s_i}, y_{s_i}\right), 1 \leqslant i \leqslant n\right\}$. It could lead to a great decrease in accuracy.

FairRF : Fairness with Proxy Sensitive Attribute

For $\hat{y}_i = h_\theta(x_i)$, $\mathcal{L}_{cls}(\hat{y}_i, y_i) = -y_i \log \hat{y}_i - (1 - y_i) \log (1 - \hat{y}_i)$, if we know s_i, then we can consider penalty term with respect to dependency between prediction and sensitive attribute as:

$$\mathcal{R}(\mathbf{s}, \hat{\mathbf{y}}) = \left| \sum_{i=1}^{n} (s_i - \mu_s)(\hat{y}_i - \mu_{\hat{y}}) \right|$$

where μ_s and $\mu_{\hat{y}}$ are the mean of $\mathbf{s}$ and $\hat{\mathbf{y}}$, respectively.

However, as sensitive attribute is unavailable in our problem, directly adopting the above regularization is impossible.

FairRF : Fairness with Proxy Sensitive Attribute

If non-protected feature $\mathbf{x}^j$ has high correlation with s, reducing the correlation between $\mathbf{x}^j$ and $\hat{\mathbf{y}}$ can indirectly reduce the correlation between $\mathbf{s}$ and $\hat{\mathbf{y}}$. In FairRF, the regularization term is written as

$$\mathcal{R}_{\text{related}} = \sum_{j=1}^{K} \lambda_j \cdot \mathcal{R}\left(\mathbf{x}^j, \hat{\mathbf{y}}\right)$$

The final objective function of FairRF is

$$\min_{\theta,\lambda} \sum_{i=1}^{n} \mathcal{L}_{cls}\left(\hat{y}_i, y_i\right) + \eta \cdot \sum_{j=1}^{K} \lambda_j \cdot \mathcal{R}\left(\mathbf{x}^j, \hat{\mathbf{y}}\right) + \beta \|\lambda\|_2^2.$$

However, this approach requires a assumption that some non-protected features are highly correlated with the sensitive attribute of concern.

Fairness with Knowledge Distillation

Fairness without demographics through knowledge distillation is newly proposed.

The main content of this paper is that by using actual labels and teacher model's output for label smoothing, then training the student model with these smoothed labels leads to improved accuracy and fairness metrics.

Fairness with Knowledge Distillation

Let g be the teacher model, and h be the student model.

The training objective for student model can be formulated as

$$L(h) = \frac{1}{N} \sum_{i=1}^{N} \left[\alpha \mathcal{L}_{cls} \left(h\left(x_i\right), \hat{y}_i^{\mathrm{t}} \right) + (1-\alpha) \mathcal{L}_{cls} \left(h\left(x_i\right), y_i \right) \right]$$

where $\hat{y}_i^{\mathrm{t}} = g\left(x_i\right)$ is the soft label from teacher model, $\mathcal{L}_{cls}$ is the classification loss, and $\alpha \in (0, 1)$ is the trade-off hyper-parameter.

Fairness with Knowledge Distillation

The teacher model can be decomposed into two functions, $g = \phi \circ f$, where f is the nonlinear function that predicts the logit, and ϕ is the mapping function for soft labelling.

Given the logit $z_i = f(x_i)$ calculated by the teacher model,

$$\hat{y}_i^{\mathrm{t}} = \phi(z_i) = \frac{\exp(z_i/T)}{1 + \exp(z_i/T)},$$

where T is the temperature that controls the probability distribution over classes.

Fairness with Knowledge Distillation

Method	Accuracy	Disparate impact	Equalized odds
Teacher	84.41%	20.27%	39.64%
Student (with hard label)	64.13±0.32%	23.27±2.43%	38.34±3.37%
DRO (Hashimoto et al., 2018)	62.67±0.73%	21.41±2.19%	30.43±3.24%
ARL (Lahoti et al., 2020)	63.23±0.47%	21.37±3.46%	29.46±1.74%
FairRF (Zhao et al., 2022)	63.26±0.83%	21.47±1.76%	25.67±2.63%
Student (with softmax label)	63.47±0.44%	19.52±2.46%	21.32±1.97%

Results on COMPAS dataset with sensitive attribute race.

Similarly, the experimental results show student with softmax label outperforms DRO, ARL, and FairRF in all aspects.

Fairness with Knowledge Distillation

The contribution of this paper is to show the improvement of accuracy and fairness metrics compared to other methods through experiments.

However, in the proof of the only theorem that theoretically supports this method, it seems that there are some critical errors.

Theorem 3.1. *Consider a classifier* $f : \mathcal{X} \to [0, 1]$ *for binary classification. Denote the classification loss as* $L_{soft} = -y' \log(f(x)) - (1 - y') \log(1 - f(x))$ *with soft label* $y' = \alpha \hat{y}^t + (1 - \alpha)y$*, where* $\hat{y}^t \in [0, 1]$ *is the predicted label from teacher model,* $y \in \{0, 1\}$ *is the binary label, and* α *is the balance parameter. The equal odds fairness metrics w.r.t. classifier* f *is upper bounded by* L_{soft}*.*

Appendix

Dataset :

New adult(predict whether an individual's income exceeds certain threshold race and gender as sensitive attributes)

COMPAS(predict whether a defendant reoffends within two years with sex and race as sensitive attributes)

CelebA(predicting attractiveness with gender as sensitive attribute, and predicting gender with age as sensitive attribute)

Appendix

In the paper, the definitions of the disparate impact and equalized equalized odds are not stated. One of the possible definitions is as follows:

Disparate impact :

$$\left| \frac{1}{\#(i : s_i = 1)} \sum_{i:s_i=1} I(\hat{y} = 1) - \frac{1}{\#(i : s_i = 0)} \sum_{i:s_i=0} I(\hat{y} = 1) \right|$$

Equalized odds :

$$\frac{1}{2} \sum_{y_0=0}^{1} \left| \frac{1}{\#(i : y_i = y_0, s_i = 1)} \sum_{i:y_i=y_0,s_i=1} I(\hat{y} = 1) - \frac{1}{\#(i : y_i = y_0, s_i = 0)} \sum_{i:y_i=y_0,s_i=0} I(\hat{y} = 1) \right|$$