SimFair: A Unified Framework for Fairness-Aware Multi-Label Classification (AAAI 2023)

Liu et. al.

Reviewer: Jihu Lee, Jinwon Park

IDEA lab
Department of Statistics
Seoul National University

March, 21, 2024

Table of Contents

Multi-Label Classification

Examples

- An applicant may apply for multiple positions
- Undergraduates submit applications to multiple programs when applying to graduate schools
- **Target label**: Admission decision of each position

Simple Approach

- Decompose into multiple binary classification
- Ignores the correlations among labels
  - Applicants usually apply for positions with similar requirements of skill sets and experiences

Fairness in Multi-Label Classification

- No existing work to define fairness directly in the context of multi-label classification
- Proposed framework s_γ-**SimFair** unifies DP and EOp
- It works even when imbalanced label distributions exist

Table of Contents

Notations

- Samples: $\mathcal{D} = \left\{ \left(\boldsymbol{x}^{(i)}, a^{(i)}, \boldsymbol{y}^{(i)} \right) \right\}_{i=1}^{N}$
- $\boldsymbol{x}^{(i)} \in \mathcal{X} = \mathbb{R}^M$, $a^{(i)} \in \mathcal{A} = \{1, \ldots, K\}$, $\boldsymbol{y}^{(i)} \in \mathcal{Y} = \{0, 1\}^L$
- $(\boldsymbol{x}, a, \boldsymbol{y}) \sim p$
- Classifier: $h = f \circ g : \mathcal{X} \rightarrow [0, 1]^L \rightarrow \mathcal{Y}$

- **Advantaged Label**: $\boldsymbol{y}_{\text{adv}}$
  - Only favorable outcomes present
  - Ex) *received offers of postion A and position B* from job screening example

DP and EOp

- **DP**: $\hat{\boldsymbol{y}} \perp a$
- **EOp**: $\hat{\boldsymbol{y}} \perp a | \boldsymbol{y}_{\text{adv}}$

Proposition (0.1)

For a multi-label classifier that takes the form $h = f \circ g$, where $\tilde{\boldsymbol{y}} = g(\boldsymbol{x})$ is the predicted probability and $\hat{\boldsymbol{y}} = f(\tilde{\boldsymbol{y}})$ is computed elementwisely, DP and EOp hold if for any $k \in \mathcal{A}$

$$\begin{aligned} \textit{DP} &: \mathbb{E}\left[\tilde{\boldsymbol{y}} | a = k\right] = \mathbb{E}\left[\tilde{\boldsymbol{y}}\right] \\ \textit{EOp} &: \mathbb{E}\left[\tilde{\boldsymbol{y}} | a = k, \boldsymbol{y} = \boldsymbol{y}_{\textit{adv}}\right] = \mathbb{E}\left[\tilde{\boldsymbol{y}} | \boldsymbol{y} = \boldsymbol{y}_{\textit{adv}}\right] \end{aligned} \quad (1)$$

DP and EOp

Estimation

$$\mathbb{E}[\tilde{\boldsymbol{y}}|a=k] \approx \frac{\sum_{i=1}^{N} \tilde{\boldsymbol{y}}^{(i)} \mathbf{1}\left(a^{(i)}=k\right)}{\sum_{i=1}^{N} \mathbf{1}\left(a^{(i)}=k\right)} \tag{2}$$

$$\mathbb{E}[\tilde{\boldsymbol{y}}|a=k, \boldsymbol{y}=\boldsymbol{y}_{\text{adv}}] \approx \frac{\sum_{i=1}^{N} \tilde{\boldsymbol{y}}^{(i)} \mathbf{1}\left(a^{(i)}=k\right) \mathbf{1}\left(\boldsymbol{y}=\boldsymbol{y}_{\text{adv}}\right)}{\sum_{i=1}^{N} \mathbf{1}\left(a^{(i)}=k\right) \mathbf{1}\left(\boldsymbol{y}=\boldsymbol{y}_{\text{adv}}\right)} \tag{3}$$

- most labels only associate with few samples $\rightarrow$ additional challenges for EOp estimation

s_γ-SimFair

EOp (Eq. (1)) is equivalent with

$$\frac{\mathbb{E}\left[\tilde{\boldsymbol{y}}\mathbf{1}\left(\boldsymbol{y}=\boldsymbol{y}_{\text{adv}}\right)\right]}{\mathbb{E}\left[\mathbf{1}\left(\boldsymbol{y}=\boldsymbol{y}_{\text{adv}}\right)\right]}=\frac{\mathbb{E}\left[\tilde{\boldsymbol{y}}\mathbf{1}\left(a=k\right)\mathbf{1}\left(\boldsymbol{y}=\boldsymbol{y}_{\text{adv}}\right)\right]}{\mathbb{E}\left[\mathbf{1}\left(a=k\right)\mathbf{1}\left(\boldsymbol{y}=\boldsymbol{y}_{\text{adv}}\right)\right]} \tag{4}$$

Definition 1 (s-SimFair)

Given a similarity function $s:\mathcal{Y}\times\mathcal{Y}\rightarrow[0,1]$, a multi-label classifier h satisfiew Similarity s-induced Fairness if for $\forall k\in\mathcal{A}$,

$$\frac{\mathbb{E}\left[\tilde{\boldsymbol{y}}s\left(\boldsymbol{y},\boldsymbol{y}_{\text{adv}}\right)\right]}{\mathbb{E}\left[s\left(\boldsymbol{y},\boldsymbol{y}_{\text{adv}}\right)\right]}=\frac{\mathbb{E}\left[\tilde{\boldsymbol{y}}\mathbf{1}\left(a=k\right)s\left(\boldsymbol{y},\boldsymbol{y}_{\text{adv}}\right)\right]}{\mathbb{E}\left[\mathbf{1}\left(a=k\right)s\left(\boldsymbol{y},\boldsymbol{y}_{\text{adv}}\right)\right]} \tag{5}$$

$$\text{Jac}(\boldsymbol{y},\boldsymbol{y}_{\text{adv}})=\frac{|\text{cate}(\boldsymbol{y})\cap\text{cate}(\boldsymbol{y}_{\text{adv}})|}{|\text{cate}(\boldsymbol{y})\cup\text{cate}(\boldsymbol{y}_{\text{adv}})|}$$

$$s_\gamma(\boldsymbol{y},\boldsymbol{y}_{\text{adv}})=\exp\left(\gamma\left(\text{Jac}\left(\boldsymbol{y},\boldsymbol{y}_{\text{adv}}\right)-1\right)\right)$$

s_γ-SimFair Unifies DP and EOp

1. DP and EOP are special cases of s_γ-SimFair
   - $s \equiv c \to$ DP
   - $s(\boldsymbol{y}, \boldsymbol{y}') = \mathbf{1}\,(\boldsymbol{y}, \boldsymbol{y}') \to$ EOp
2. s_γ-SimFair helps achieve DP and EOp
   - γ sufficiently small $\to$ small DP violation
   - γ sufficiently large $\to$ small EOp violation

Training

Violation of s_γ-SimFair

$$l_{s_\gamma(\boldsymbol{y},\boldsymbol{y}_{\text{adv}})}(h) = \sum_{k=1}^{K} \left\| \frac{\mathbb{E}\left[\tilde{\boldsymbol{y}} s_\gamma(\boldsymbol{y},\boldsymbol{y}_{\text{adv}})\right]}{\mathbb{E}\left[s_\gamma(\boldsymbol{y},\boldsymbol{y}_{\text{adv}})\right]} - \frac{\mathbb{E}\left[\tilde{\boldsymbol{y}}\mathbf{1}\left(a=k\right) s_\gamma(\boldsymbol{y},\boldsymbol{y}_{\text{adv}})\right]}{\mathbb{E}\left[\mathbf{1}\left(a=k\right) s_\gamma(\boldsymbol{y},\boldsymbol{y}_{\text{adv}})\right]} \right\| \tag{6}$$

Objective

$$\min_h l_{\text{mlc}}(h) + \lambda l_{s_\gamma(\boldsymbol{y},\boldsymbol{y}_{\text{adv}})}(h) \tag{7}$$

Combination with a Backbone Model (MPVAE)

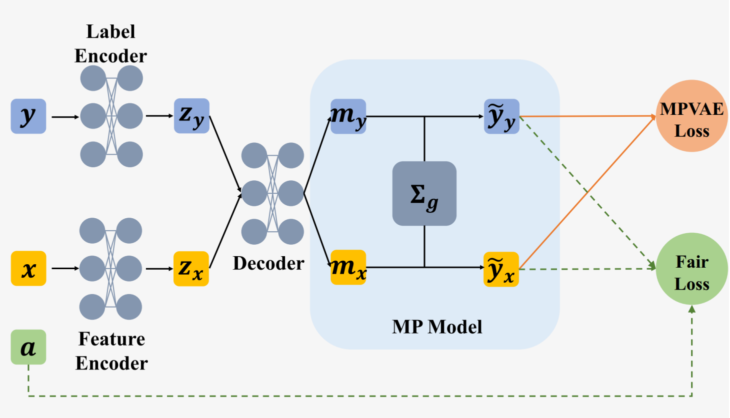

Figure 1: Framework of training MPVAE with fairness regulrization. Both probability vectors $\tilde{y}$ on two brancehs are used to construct the s_γ-SimFair regularizer

Table of Contents

Datasets

- Adult
  - Label: income, workclass, occupation
  - Sensitive: binarized age
- Credit
  - Label: default payments, education level
  - Sensitive: gender

Settings

Baselines

- MPVAE with No regularizer
- MPVAE with DP regularizer
- MPVAE with EOp regularizer

Evaluation metrics

- micro-F1, macro-F1, example-F1

Estimate DP and EOp with s_γ-SimFair

	y_{adv}	DP	$s_{0.1}$-SF	$s_{0.5}$-SF	s_1-SF	s_5-SF	s_{10}-SF	EOp
Adult	100%	0.11*	0.12	0.12	0.13	0.18	0.17	0.17*
	70%	0.11	0.11	0.12	0.13	0.18	0.18	0.17
	30%	0.10	0.10	0.11	0.12	0.17	0.18	0.17
	10%	0.10	0.10	0.10	0.11	0.14	0.16	0.17
	5%	0.10	0.10	0.10	0.11	0.15	0.23	0.27
Credit	100%	0.03*	0.03	0.03	0.03	0.03	0.03	0.03*
	70%	0.03	0.03	0.03	0.03	0.03	0.04	0.04
	30%	0.02	0.02	0.02	0.02	0.03	0.03	0.03
	10%	0.02	0.02	0.02	0.02	0.03	0.04	0.04
	5%	0.02	0.02	0.02	0.02	0.03	0.04	0.05

Figure 2: Varying the numbers of samples in the advantaged group to different levels. Ground truth is marked with asterisk. More stable EOp estimates when EOp estimator fails.

Performance of Regularization

	$\lvert \boldsymbol{y}_{adv} \rvert$	Metric	Regularzier					
			DP	s_1-SF	s_5-SF	s_{10}-SF	EOp	None
Adult	No.1	DP	0.038	**0.031**	0.038	0.043	0.045	0.111
		EOp	0.051	0.042	**0.030**	0.034	0.035	0.161
	No.18	DP	**0.038**	**0.038**	0.043	0.045	0.094	0.111
		EOp	0.076	0.072	0.037	**0.027**	0.066	0.095
Credit	No.1	DP	0.018	0.018	**0.017**	0.018	0.018	0.029
		EOp	0.026	0.026	**0.025**	**0.025**	0.026	0.038
	No.9	DP	**0.018**	**0.018**	0.019	0.019	0.030	0.030
		EOp	0.202	**0.192**	0.193	0.197	0.241	0.241

Figure 3: DP and EOp violations of MPVAE trained with such regularizers. Best results are in bold.

Fairness-Accuracy Tradeoff

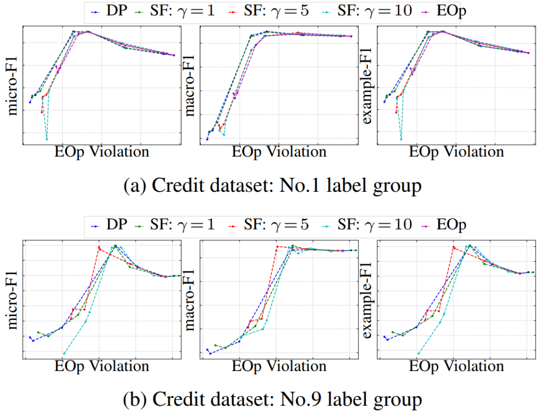

Figure 4: EOp-accuracy tradeoffs on Credit dataset. EOp regularizer is unstable and ineffective when the advantaged group is small.


THANK
you!