Improving Fairness Generalization Through a Sample-Robust Optimization Method

Ulrich Aïvodji, Julien Ferry, Sébastien Gambs, Marie-José Huguet and Mohamed Siala
LAAS-CNRS, INSA Toulouse, ETS, UQAM

[1] Context

- Fairness generalization is one of the open challenges in trustworthy Machine Learning
- Current approaches are essentially adhoc with no theoretical framework
- We propose a novel approach to fairness generalization based on distributionally robust optimization
- Our approach can be used with different learning models using different statistical fairness measures
- Empirical results show that our approach outperforms state-of-the-art approaches
- The paper is published in the **Machine Learning** journal in July 2022

[2] Statistical Fairness

The COMPAS Example:

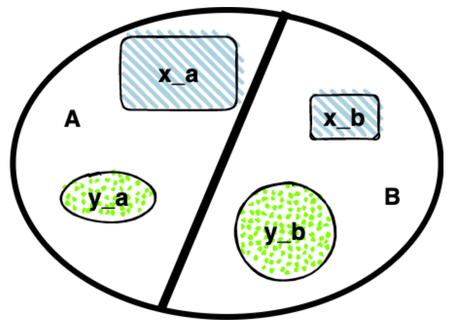
- Binary classification task: Recidivism within two years ?
- **Sensitive attribute:** Ethnicity (African-American/Caucasian)
- **Protected Groups**:
  - A : African-American individuals;
  - B : Caucasian individuals;
- Principle: ensure that some measure $\mathcal{M}$ (e.g., true positive rate) *differs by no more than* ϵ between several *protected* groups
- In the particular case of two protected groups (A) and (B) (binary sensitive attribute), one needs to ensure that $|\mathcal{M}(\mathcal{A}) - \mathcal{M}(\mathcal{B})| \leq \epsilon$

[3] The Sample-Robust Fair Learning Problem

- Let $\mathcal{D}$ be a dataset, ϵ an unfairness tolerence, and h a predictive model
- For a given value $d \in [0, 1]$, we define a perturbation set $\mathcal{B}(\mathcal{D}, d)$ as the set of all subsets of $\mathcal{D}$ whose Jaccard distance from $\mathcal{D}$ is less than or equal to d.
- The robustness of h is the largest distance d such that h is fair on each element in $\mathcal{B}(\mathcal{D}, d)$
- Given a distance d, the sample-robust fair learning problem is:

$$\begin{aligned} \arg\min_{h \in \mathcal{H}} \quad & f_{obj}(h, \mathcal{D}) \\ \text{s.t.} \quad & Robustness(h, \mathcal{D}, \epsilon) > d \end{aligned} \tag{1}$$

[4] Robustness Evaluation: Intuition



- Suppose that $\exists\ x_a$ (x_b) examples in A (B) that satisfy the measure $\mathcal{M}$
- Suppose that $\exists\ y_a$ (y_b) examples in A (B) that do not satisfy measure $\mathcal{M}$
- Let $\mathcal{D}'$ be the subset of $\mathcal{D}$ where the colored sets are removed
- If h is not fair on $\mathcal{D}'$, then h is not robust on the perturbation set defined with the distance between $\mathcal{D}$ and $\mathcal{D}'$

[5] An Integer Programming Approach

$$\min \quad n \tag{2}$$

$$\text{s.t.} \quad n = x_a + x_b + y_a + y_b \tag{3}$$

$$\left|\frac{M_a - x_a}{N_a - x_a - y_a} - \frac{M_b - x_b}{N_b - x_b - y_b}\right| > \epsilon \tag{4}$$

$$0 \leq x_a \leq M_a \qquad 0 \leq x_b \leq M_b$$

$$0 \leq y_a \leq M_a - M_a \qquad 0 \leq y_b \leq N_b - M_b$$

$$x_a + y_a < N_a \qquad x_b + y_b < N_b$$

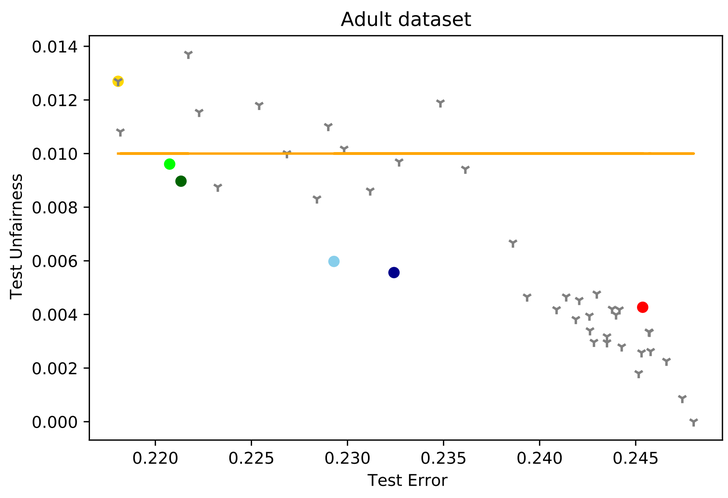
- N_a (respectively N_b) is the number of examples in the group A (respectively B)
- M_a (respectively M_b) is the number of examples that satisfy the measure in the group A (respectively B)
- The optimal value can be used to identify the largest perturbation set where the robustness constraint is satisfied

[6] Experimental Study

- `FairCORELS` and `TensorFlow Constrained Optimization`(`TFCO`)
- Four fairness metrics:
  - Statistical Parity
  - Predictive Equality
  - Equal Opportunity
  - Equalized Odds
- Four biased datasets:
  - Adult Income dataset
  - COMPAS dataset
  - Default Credit dataset
  - Bank Marketing dataset

[7] Some Experimental Results (FairCORELS)

- Test error and unfairness of models generated by `FairCORELS` using our exact and heuristic sample-robust fair methods (Statistical Parity metric, $\epsilon = 0.01$)



[8] Some Experimental Results (TFCO)

- Results of the experimental study of the heuristic approach using `TFCO` (error rates and maximum fairness violations)

	Proxy Lagrangian				Lagrangian			
	Train		Test		Train		Test	
Model	Error	Viol.	Error	Viol.	Error	Viol.	Error	Viol.
Adult Income Dataset								
`unconstrained`	.122	.072	**.144**	.071	.122	.072	**.144**	.071
`baseline`	.141	0	*.154*	.009	.141	0	*.155*	.006
`validation`	.132	-.002	.158	.004	.134	0	.157	*.004*
`dromasks-10`	.14	-.003	.156	*.003*	.143	-.001	*.155*	**-.003**
`dromasks-30`	.14	-.004	.157	**-.001**	.148	-.002	.156	**-.003**
`dromasks-50`	.14	-.003	.157	**-.001**	.151	-.002	.157	**-.003**

[9] Summary

- We propose a novel approach to fairness generalisation inspired by "Distributionally Robust Optimization"
- Our propositions are flexible enough to be used with different models
- We empirically show that our approach is competitive to the state-of-the-art with two learning models on many datasets in the literature using different fairness measures

[10] Future Work

- Can we find an efficient way to approximate the best parameters ?
- How to extend the work with other distance functions ?