When Fairness Meets Privacy: Fair Classification with Semi-Private Sensitive Attributes

Thomas Wimmer, Jules Soria, Maximilien Chau

Chen, Canyu, et al. "When Fairness Meets Privacy: Fair Classification with Semi-Private Sensitive Attributes."
Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022.

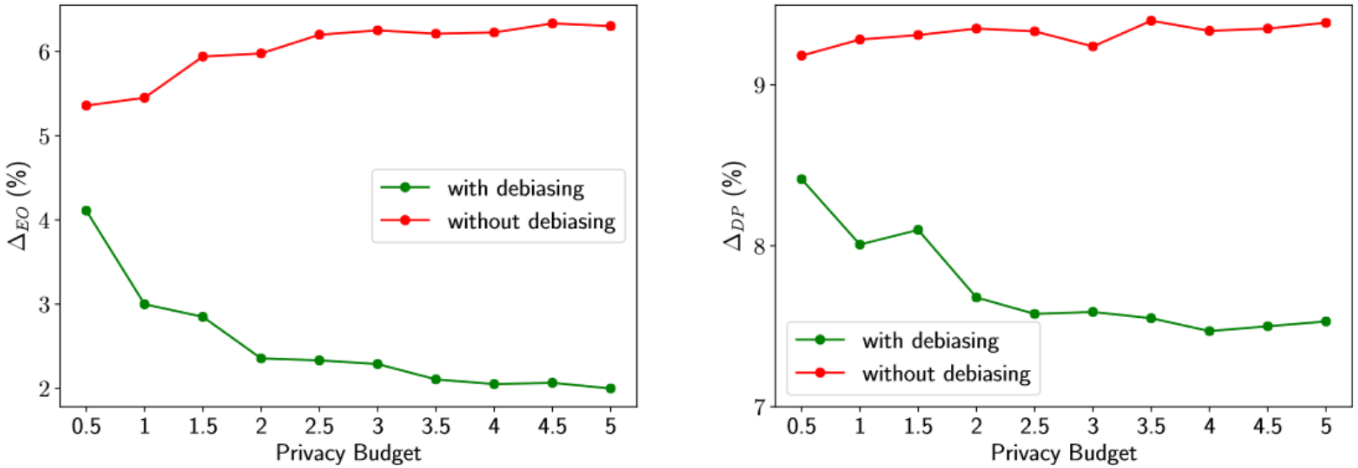
Motivation

- Sensitive information is often protected through law (GDPR, ECPA) and thus in many cases only available as noise (processed values using e.g., Local Differential Privacy (LDP))
- Most fairness-preserving methods require direct access to sensitive attributes
- We can alleviate the little known information on clean sensitive attributes to make educated guesses about the values of noisy sensitive attributes

We can thus create a method better suited to most practical use cases

Preliminary study Results

What impact does privacy (in the form of using LDP) have on fair ML algorithms?



(a) ADULT (b) ADULT

Impact of privacy on fairness performances on the ADULT dataset

Higher privacy budget = lower privacy guarantees (= lower probability of "flipping")

ΔEO: Equal Opportunity

$$\Delta_{EO} = |\mathbb{E}(\hat{Y}|A=1, Y=1) - \mathbb{E}(\hat{Y}|A=0, Y=1)|$$

> Positive instances with arbitrary sensitive attributes are equally likely to be assigned a positive outcome

ΔDP: Demographic Parity

$$\Delta_{DP} = |\mathbb{E}(\hat{Y}|A=1) - \mathbb{E}(\hat{Y}|A=0)|$$

> Positive rate across sensitive attributes is equal

Preliminary study Conclusions

What impact does privacy (in the form of using LDP) have on fair ML algorithms?

- Non-debiasing methods (usual MLPs) improve in fairness when using a stronger privacy guarantee (more noise in sensitive attributes)
- For debiasing methods, stronger privacy guarantees lead to worse fairness performance

Improving the fairness performance of debiasing methods requires (among others) reducing the noise in sensitive attributes

Problem Statement

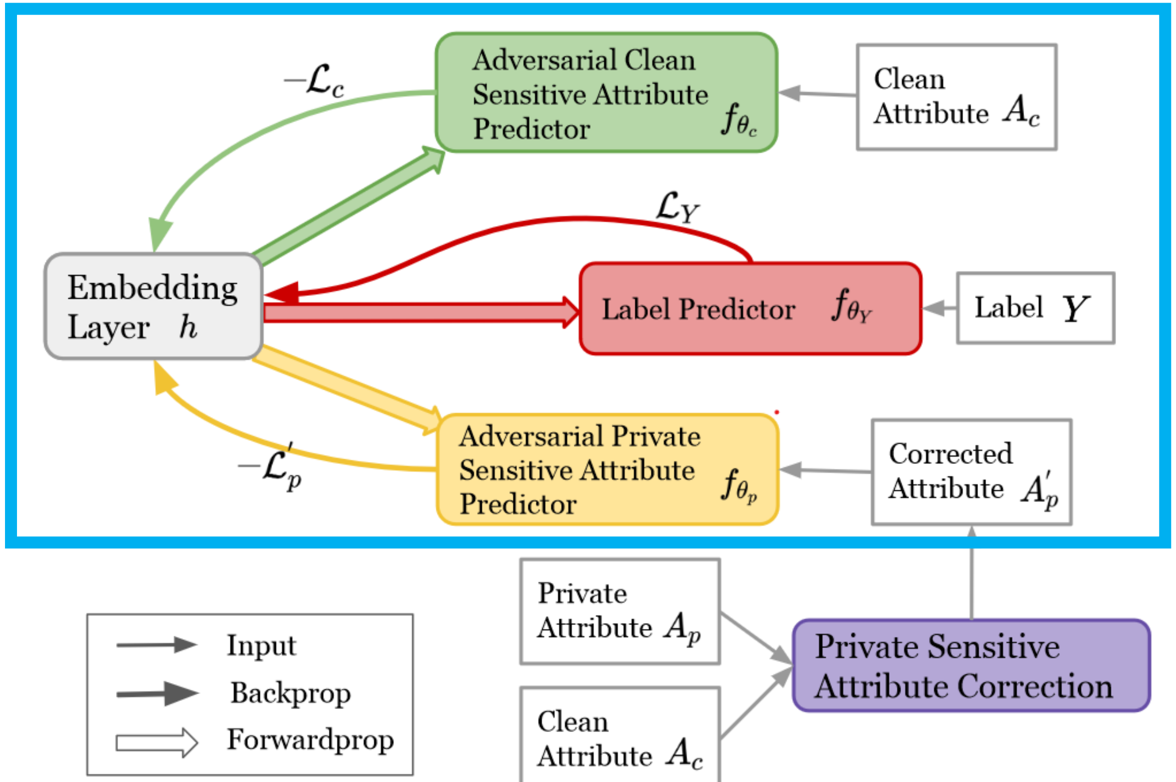
Given the training data D with a limited number of clean sensitive attributes A_c and a large amount of private sensitive attributes A_p, learn an effective classifier that generalises well to unseen instances, while satisfying the fairness criteria such as demographic parity.

	Sex	Marital status	Ethnicity	Income > $50K
Person A	Male	**Never-married**	**Amer-Indian-Eskimo**	Yes
Person B	**Female**	**Divorced**	White	No

Samples from the ADULT dataset in a semi-private setting

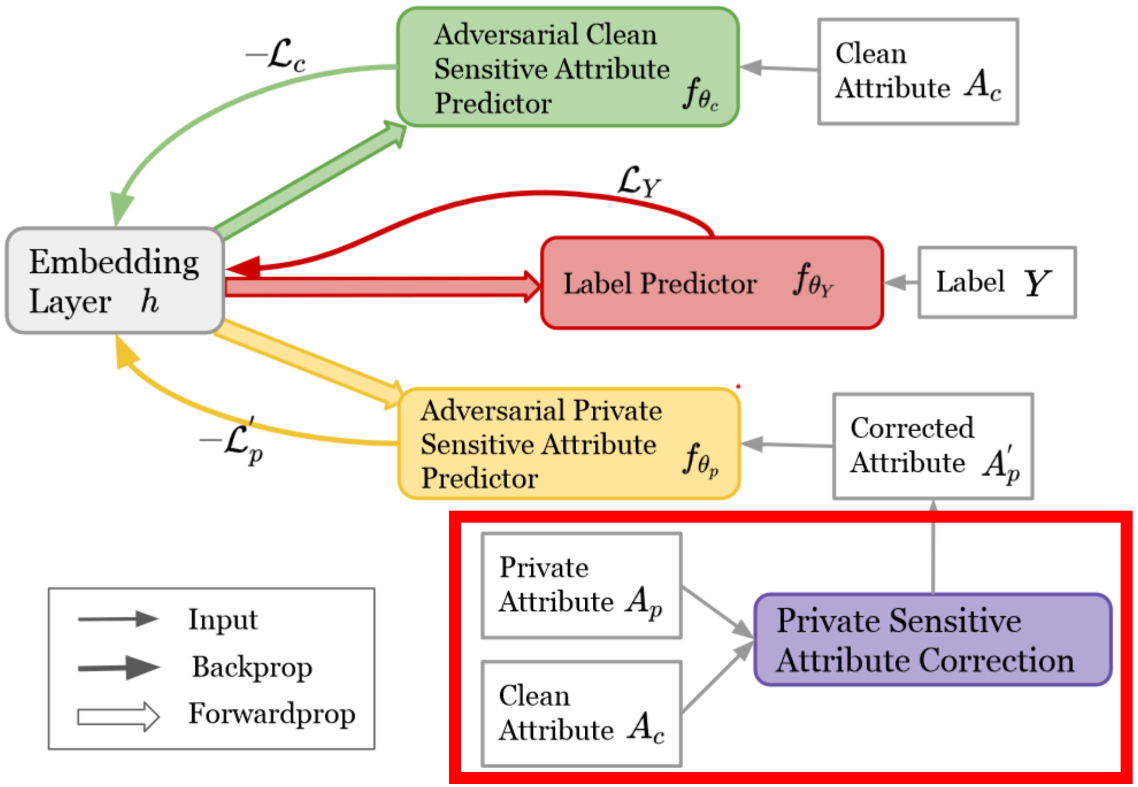
Proposed Method Semi-private Adversarial Debiasing

- Shared Encoder Layer to learn an "anonymous" embedding vector that is fed into the predictor network
- Adversarial Learning: Train *clean sensitive attribute predictor* and *private sensitive attribute predictor*
  - Min-max game between encoding layer and predictors



Proposed Method Private Sensitive Attribute Correction

- Directly applying adversarial debiasing may lead to sub-optimal results
- Before feeding attributes into the network, we try to clean them using a learned correction matrix to estimate the true sensitive attributes from the private ones



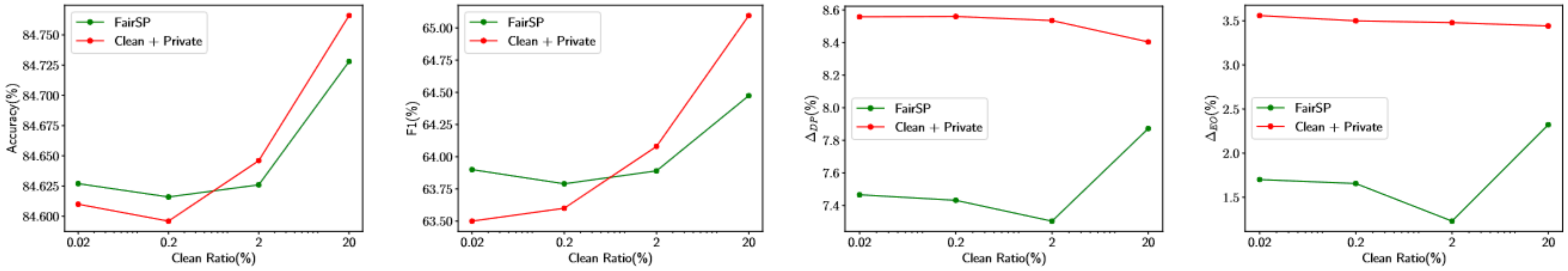
Experiments

Table 1: The performance comparison for fair classification under semi-private setting.

Datasets	Metric	Vanilla	RemoveS	RNF [13]	FariRF [47]	Clean	Private	C+P	FAIRSP
ADULT	Acc.(%)	84.8±0.2	84.9±0.3	83.5±1.2	84.0±0.5	84.9±0.4	84.7±0.3	84.8±0.5	84.7±0.4
	F1(%)	65.4±0.7	64.8±0.8	63.3±0.8	63.5±0.7	64.6±0.7	64.6±0.3	64.8±0.6	64.5±0.7
	Δ_{DP}(%)	9.1±0.4	8.4±0.2	8.3±1.0	8.2±0.3	8.4±0.4	8.4±0.3	8.1±0.2	**7.8±0.3**
	Δ_{EO}(%)	5.3±1.0	4.1±1.1	4.0±0.5	3.5±0.8	4.1±1.0	4.1±1.2	3.4±1.4	**2.3±1.2**
COMPAS	Acc.(%)	67.0±0.6	67.3±0.8	66.9±0.8	66.3±0.7	67.2±0.6	67.1±0.7	67.2±0.6	67.0±0.6
	F1(%)	64.3±0.9	64.2±1.2	63.5±0.9	63.2±0.5	64.8±1.0	64.6±1.1	63.9±1.1	63.8±1.4
	Δ_{DP}(%)	13.8±1.1	13.0±0.4	13.1±0.6	13.8±2.4	13.1±0.5	13.0±0.4	12.9±0.2	**12.7±0.5**
	Δ_{EO}(%)	12.8±1.4	12.2±0.6	12.3±1.3	15.3±1.2	12.3±0.8	12.1±0.7	12.2±0.5	**12.1±0.6**
MEPS	Acc.(%)	86.1±0.1	86.1±0.2	85.8±0.1	85.9±0.2	86.1±0.1	86.0±0.1	86.1±0.1	86.0±0.1
	F1(%)	48.5±2.0	49.9±1.6	49.5±1.5	47.0±1.9	50.6±1.6	50.8±2.3	48.8±1.8	47.3±1.7
	Δ_{DP}(%)	4.5±0.5	4.7±0.5	4.8±0.3	4.9±1.0	4.8±0.6	4.8±0.7	4.4±0.4	**4.1±0.8**
	Δ_{EO}(%)	4.5±1.0	4.6±1.1	4.8±0.9	4.7±1.3	4.4±1.2	4.5±1.1	4.3±0.7	**4.0±1.2**

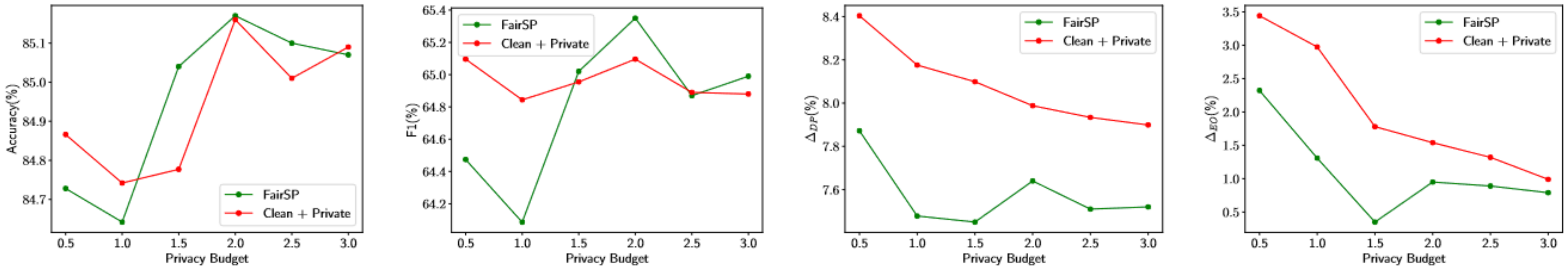
Every metric is important to assess the performance of the model!

Experiments



(a) Accuracy (b) F1 (c) Δ_{DP} (d) Δ_{EO}

The impact of clean data ratio on prediction and debiasing performances on ADULT.



(a) Accuracy (b) F1 (c) Δ_{DP} (d) Δ_{EO}

The impact of privacy budget ϵ on prediction and debiasing performances on ADULT.

Conclusion and Assessment of the Paper

- Working in the semi-private setting is a novel idea
- Preliminary study gives a clear motivation for the work
- Proposed method shows well-balanced results in the experiments

- Paper is not particularly well-written
- Analysis of the „goodness“ of the correction matrix would have been interesting