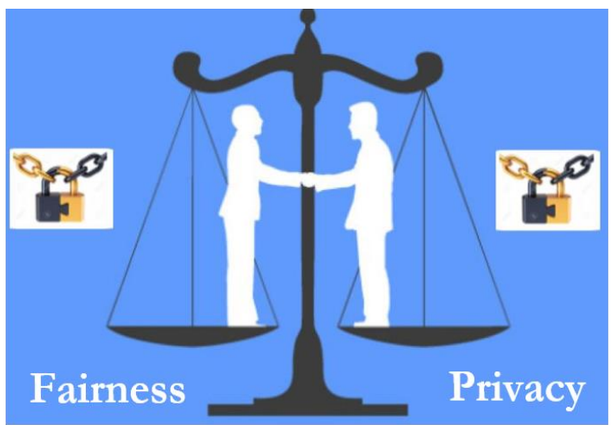
Privacy-Preserving Fair Machine Learning by Utilizing Proxy Data

Hui Hu, Mike Borowczak

University of Wyoming, Department of Computer Science, CEDAR Lab

Introduction

In fair machine learning, predictive models are expected to be non-discriminatory with respect to certain sensitive features. However, most fair learning techniques require full access to sensitive personal data. For example, to learn an auto-hiring model with no discrimination against HIV carriers, most learners may need the HIV records of all individuals in the training set. In practice, it may be impossible to use sensitive data for decision-making due to privacy regulations [1].



To address the dilemma between model fairness and privacy protection, we propose a novel privacy-preserving fair learning mechanism utilizing proxy data. We assume the data holder and modeler are separated, and the original data is privately held by the data holder. To avoid revealing the original sensitive data, the data holder first learns two decoupled dictionaries via dictionary learning technique; then, reconstructs different data representations based on the dictionaries and send them to modeler for model training. We exemplify two non-private fair learners into private ones under the proposed mechanism.

Goal

- To address the dilemma between model fairness and privacy protection.
- To explore model fairness and privacy-preserving performance with different data representations based on dictionary learning technique.

Methodology

The proposed mechanism includes **three** modules: dictionary decoupling module, data reconstruction module and prediction module. Figure 1 shows an overview of the mechanism.

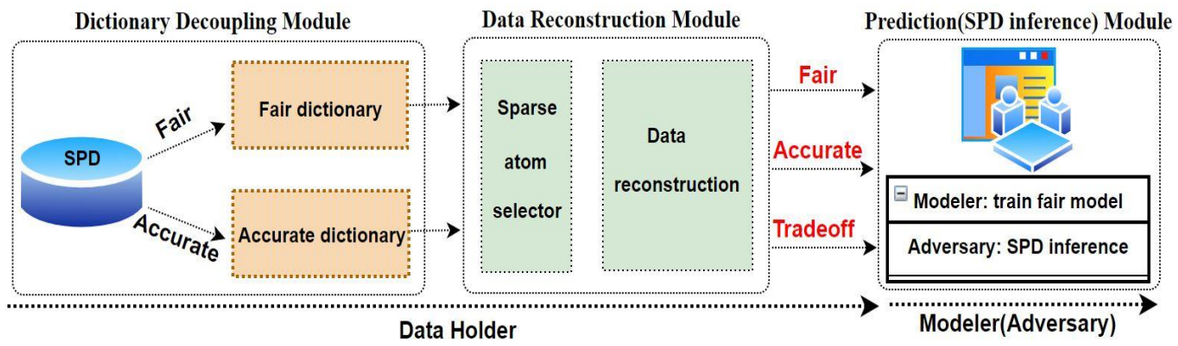


Figure 1. An overview of privacy-preserving fair learning framework (SPD denotes sensitive personal data. Red words denote three different proxy data representations).

- **Dictionary decoupling module**: Learn two decoupled dictionaries – an accurate dictionary D_0 and a fair dictionary D_1.
- **Data reconstruction module**: Learn a sparse selector A_t to reconstruct different representations of data. The objective function is designed as:

$$min_{A_t} ||X - DA_t||_2^2 + \gamma||A_t||_1,$$

where X is the original data matrix, γ is a hyperparameter and $D = [D_1, D_2]$.

- **Prediction module**: Modeler learns a fair model based on different data representations.

Initial Results

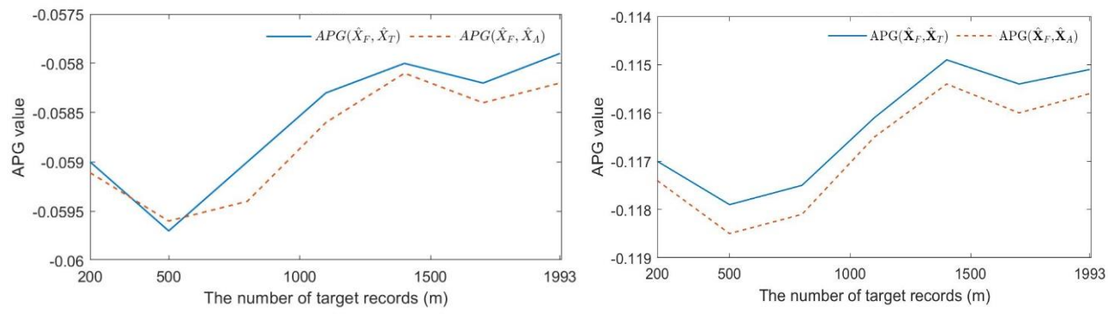
We exemplified two non-private fair learners (fair ridge regression and fair logistic regression) under the proposed mechanism. We evaluated the performance of the two private fair learners on three benchmark datasets.

Evaluation metrics: We used statistical parity[2](SP) to measure model fairness and classifier error was used to measure model error; A smaller SP implies a fairer model, while a smaller classifier error implies a more accurate model. Average privacy gain [3] (APG) was used to measure privacy-preserving capability of different data representations. $APG(X_1, X_2) > 0$ indicates that X_1 leaks more sensitive information than X_2, vice versa.

Table 1. Classification Performance on Crime Data Set

Methods		Statistical Parity	Classifier Error
PFRR	Accurate	.1162±.0113	**.2203±.0067**
	Tradeoff	.0496 ± .0177	.2474 ± .0111
	Fair	**.0303±.0293**	.2467 ± .0171
PFGR	Accurate	.0863±.0198	**.2332±.0127**
	Tradeoff	.0463±.0193	.2416±.0113
	Fair	**.0387±.0179**	.2498±.0165

Figure 2. Comparison of Privacy-preserving Performance of Two Data Representations



Discussion

- Table 1 shows that the fairness and accuracy performance of the proposed fair learners is flexible based on different data representations. Model fairness is often the best on a fair data representation (more fair atoms).
- Figure 2 shows that the privacy-preserving performance of fair data representation outperforms the accurate data representation and trade-off form in most cases.
- This study focuses on qualitative analysis of the privacy-preserving performance of proxy data. It is necessary to do more quantitative analysis. Moreover, it is our future work to design non-linear models under the proposed mechanism.

Conclusion

In this work, we addressed an important problem of fair machine learning. We explored a new direction and proposed a novel privacy-preserving mechanism for fair learning based on dictionary learning technique. Experimental results show that the proposed private fair learners have flexible fairness and privacy-preserving performance based on different data representations.

References

1. Hui Hu, Yijun Liu, Zhen Wang, and Chao Lan. A distributed fair machine learning framework with private demographic data protection. arXiv preprint arXiv:1909.08081, 2019.
2. Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. Provably fair representations. CoRR, 2017
3. Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–a privacy mirage. arXiv preprint arXiv:2011.07018, 2020.

Contact email:
hhu1@uwyo.edu