Paper review:
Domain Adaptation meets individual fairness.
And they get along.

@ NeurIPS 2022

Kunwoong Kim

2023.1.30

Department of Statistics, Seoul National University

Table of Contents

Motivation

- Is it possible to overcome distribution shifts with algorithmic fairness interventions?
  = From Fairness to Domain Adaptation
- Is it possible to mitigate fairness biases with domain adaptation methods?
  = From Domain Adaptation to Fairness

Contributions

- Methods for individual fairness can help ML models adapt to new domains.
- Domain adaptation algorithms, whose methods are based on aligning representation distributions in the source/target domains, can be used to achieve individual fairness.

Table of Contents

Connection between DA and IF

- Definition of Individual Fairness: L-Lipschitz property

$$d_{\mathcal{Y}}(f(x), f(x')) \leq L d_{\mathcal{X}}(x, x') \tag{1}$$

  for all $x, x' \in \mathcal{X}$. Here, $d_{\mathcal{Y}}$ and $d_{\mathcal{X}}$ are similarity metrics on output and input space, respectively.

- **Goals of IF (Individual Fairness) and DA (Domain Adaptation) coincide.**
  - To ignore uninformative dissimilarity: enforcing invariance/smoothness of the ML model.
  - IF: to ignore variation among inputs that are attributed to variation of the sensitive attribute.
  - DA: to ignore variation among inputs that are attributed to the domains.

Table of Contents

Achieving DA via IF

- Goal of DA: **maximize the accuracy on the unlabeled samples from the target domain** in the training data.
- Assume **covariate shift** exists; the marginal distributions of features from the target domain and the source domain differ.
  - Denote X_s and X_t as the random vector from the source and target domain, respectively.
  - Covariate shift: $Y = f_0(X) + \epsilon$ for $X = X_s, X_t$, i.e., the regression function f_0 in the source and target domains are identical.

Achieving DA via IF

- Classic approach: add regularization $\mathcal{R}_n$ for leveraging the source and target samples:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left(\frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}(y_i, f(x_i)) + \lambda \mathcal{R}_n(f(X)) \right). \quad (2)$$

- Population version:

$$\tilde{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\mathcal{L}(Y_s, f(X_s)) + \lambda R(f, f) \quad (3)$$

where R is the population version of $\mathcal{R}_n$.

Achieving DA via IF

- Example of such $\mathcal{R}_n$: graph Laplacian regularizer.
  - based on a similarity symmetric kernel K on the input space $\mathcal{X}$.
  - The regularizer based on this kernel is defined as

$$\mathcal{R}_n(f(X)) = \frac{1}{n^2} \sum_{i \neq j \in n} K(X_i, X_j)(f(X_i) - f(X_j))^2. \qquad (4)$$

    It means if $K(X_i, X_j)$ is large, then $f(X_i)$ must be close to $f(X_j)$.
  - Example: $K(\cdot, \cdot)$ is a decreasing function of $d_{\mathcal{X}}(\cdot, \cdot)$.

Achieving DA via IF

Assumptions

- $\mathcal{R}_n(f_0(X)) \leq \delta$ for some small $\delta > 0$.
- $\mathcal{R}$ is strongly convex and smooth.
- Classification loss function $\mathcal{L}$ is strongly convex and smooth.

Theoretical results

- Transductive (focusing on training target risk)

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}\left(\hat{f}(x_{t,i}), f_0(x_{t,i})\right) \leq \alpha_n\left[\frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{L}\left(\hat{f}(x_{s,i}), f_0(x_{s,i})\right) + \lambda\mathcal{R}_n(\hat{f}(X))\right] + \beta_n\mathcal{R}_n(f_0(X))\,.$$

- Inductive (focusing on test target risk)

$$\mathbb{E}_Q[\mathcal{L}(\widetilde{f}(x), f_0(x)] \leq C_1\left[\mathbb{E}_P[\mathcal{L}(\widetilde{f}(x), f_0(x)] + \lambda\mathcal{R}(\tilde{f}, \tilde{f})\right] + C_2\mathcal{R}(f_0, f_0)\,.$$

Achieving DG via IF

- In the domain generalization problem, we have no target domain samples in training data, which we have in DA problem.
- Concept
  - An example of regularizer for domain generalization is $R(f, f)$ where

$$R(f, g) = \max_{T} \mathbb{E}\left(f(X) - g \circ T(X)\right)^2 \text{ s.t. } \mathbb{E}\|X - T(X)\| \leq \epsilon. \tag{5}$$

  - T produces the adversarial target domain sample.
  - SenSeI (a method for individual fairness) uses this regularizer, and it is similar to DRO. It finds the worst distribution in the ϵ-ball of P that maximizes the (expectation of) difference of prediction, i.e., $\mathbb{E}(f(X) - f \circ T(X))^2$.

Achieving DG via IF

Theoretical results

$$\sup_{Q \in \mathcal{Q}_\epsilon} \mathbb{E}_{x \sim Q} \left(\tilde{f}(x) - f_0(x) \right)^2 \leq 4 \left[R(\tilde{f}, \tilde{f}) + R(f_0, f_0) + \mathbb{E}_{x \sim P} \left(\tilde{f}(x) - f_0(x) \right)^2 \right]$$

Achieving DA via IF

Experiments

- Datasets: Bios and Toxicity datasets (goal: identifying toxic comments, NLP dataset)
- Measures: balanced accuracy (BA), worst group accuracy, and TNR.
- Methods: algorithms for individual fairness (GLIF, SenSeI, SenSR, CLP)
- Results

Table 1: Enforcing domain generalization using individual fairness methods. Means and stds over 10 runs.

	Bios		Toxicity		
	BA	Worst p. gender	BA	TNR (Annot.)	TNR (Id. tokens)
Baseline	$84.2\% \pm 0.2\%$	$77.9\% \pm 0.4\%$	$\mathbf{80.7}\% \pm 0.2\%$	$79.4\% \pm 2.2\%$	$75.0\% \pm 2.3\%$
GLIF	$\mathbf{84.6}\% \pm 0.3\%$	$77.6\% \pm 1.0\%$	$70.5\% \pm 7.1\%$	$\mathbf{87.0}\% \pm 9.8\%$	$\mathbf{84.5}\% \pm 9.8\%$
SenSeI	$84.3\% \pm 0.3\%$	$\mathbf{80.2}\% \pm 0.4\%$	$79.1\% \pm 0.5\%$	$83.5\% \pm 1.7\%$	$79.4\% \pm 1.5\%$
SenSR	$84.2\% \pm 0.3\%$	$\mathbf{80.2}\% \pm 0.4\%$	$79.4\% \pm 0.3\%$	$81.5\% \pm 1.1\%$	$77.2\% \pm 0.9\%$
CLP	$84.1\% \pm 0.3\%$	$79.9\% \pm 0.3\%$	$79.5\% \pm 0.6\%$	$81.6\% \pm 1.7\%$	$78.0\% \pm 1.8\%$

Table of Contents

Achieving IF via DA

- Are the techniques employed for DA can be applied to IF?
- Many DA methods are to find a representation $\Phi(X)$ of X, such that the source and target distributions of $\Phi(X)$ are aligned.
- Common approaches: minimizing the dissimilarity between the source representation distribution and the target representation distribution.
  - [1, 2] learn transporting map of representations such that the first two moments of the transformed representations are identical in source/target distributions.
  - [3] learns domain-invariant representations by minimizing Wasserstein distance between source/target representations $\Phi(X)$.

Achieving IF via DA

Assume a factor model

- Let U is the random relevant attribute, $Z \in \{0, 1\}$ is the random protected attributes, and ϵ is the random noise.
- We assume the factor model as

$$\begin{aligned} X &= AU + bZ + \epsilon \\ X_S &= AU + b + \epsilon, X_T = AU + \epsilon \end{aligned} \tag{6}$$

for some A (subscripts S and T indicate source and target).

Achieving IF via DA

- If we can estimate a linear transformation $\Phi \in \mathbb{R}^{q \times p}$ of X s.t. ΦX_S and ΦX_T follow an identical distribution, then $\Phi b = 0$.
- This means any classifier built on top of the linear representation Φx will be individually fair because $\Phi x - \Phi x'$ for all x, x' that share U.
- Thus, DA methods can be applied to the problems for achieving individual fairness (when the covariates follow a factor model).

Achieving IF via DA

Experiments

- Datasets: Bios and Toxicity datasets.
- Measures: balanced accuracy (BA) and prediction consistency (PC).
- Methods: algorithms for domain adaptation (DANN, VADA, WDA) - methods for aligning representations adversarially.
- Results

Table 2: Enforcing individual fairness using domain adaptation methods. Means and standard deviations over 10 runs.

	Bios		Toxicity	
	BA	PC	BA	PC
Baseline	**84.2**% ± 0.2%	94.2% ± 0.1%	80.7% ± 0.2%	62.1% ± 1.4%
DANN	84.0% ± 0.3%	94.8% ± 0.3%	**80.8**% ± 0.2%	62.8% ± 1.1%
VADA	84.0% ± 0.3%	94.8% ± 0.3%	**80.8**% ± 0.2%	62.0% ± 1.4%
WDA	83.3% ± 0.3%	**95.5**% ± 0.3%	80.5% ± 0.3%	**65.4**% ± 1.3%
SenSeI	84.3% ± 0.3%	97.7% ± 0.1%	79.1% ± 0.5%	77.3% ± 4.3%

Table of Contents

References

[1] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2058–2065. AAAI Press, 2016.

[2] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation, 2016.

[3] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.


Thanks!