A Survey on Intersectional Fairness in Machine Learning- IJCAI 2023

Shin Yun Seop

April 1, 2025

Seoul national university, statistics, IDEA LAB

Contents

Overview

- When machine learning systems are employed for decision-making, issues of discrimination against specific groups (e.g., women, black individuals) arise.
- Existing notions of independent group fairness, which consider only a single sensitive attribute (e.g., gender or race) independently, fail to adequately capture discrimination against groups defined by multiple intersecting attributes.
- To address these limitations, this paper surveys methodologies related to intersectional fairness and provides a discussion on their implications.

Notation

- $x \in \mathcal{X}$: protected attributes.
- $x' \in \mathcal{X}'$: unprotected attributes.
- $X = (x, x') \in \mathcal{X} \times \mathcal{X}' = \mathcal{X}^*$: attributes(features).
- $y \in \mathcal{Y}$: (binary) output (Although the paper does not explicitly specify that the attributes are binary, the context suggests that they are assumed to be binary.)
- $(X, y) \sim \mathcal{P}$: Data distribution.
- $f : \mathcal{X}^* \to \mathcal{Y}$: a predictor, and $f(X)$: predictor output.
- $\mathcal{C} = \{c : \mathcal{X} \to \{0, 1\}\}$: collection of characteristic functions where $c(x) = 1$ indicates that an individual with protected attribute x is in subgroup c.

Contents

Subgroup Fairness

- Traditional notions of group fairness primarily assess disparities between groups defined by independent protected attributes such as gender or race.
- When fairness is assessed solely based on groups defined by independent attributes, there is a significant risk of overlooking unfairness experienced by finer-grained subgroups formed through intersectionality.
- Subgroup fairness is an attempt to evaluate fairness across more fine-grained groups that are defined by combinations of multiple protected attributes.

Subgroup Fairness

Definition 1 (Subgroup Fairness(Kearns et al., 2018))

A classifier $f(x)$ is said to be γ-SP subgroup fair if for all $c \in \mathcal{C}$,

$$|\mathbb{P}(f(X) = 1) - \mathbb{P}(f(X) = 1 \mid c(x) = 1)| \times \mathbb{P}(c(x) = 1) \leq \gamma \quad (1)$$

- The above condition imposes a constraint to ensure that the prediction outcomes for each subgroup do not significantly deviate from those of the overall population.

Subgroup Fairness

- By measuring disparities across more detailed and diverse subgroups rather than using simple group fairness, we can ensure a more accurate and reliable notion of fairness.
- As the size of a subgroup decreases, its corresponding weight in the fairness evaluation tends to diminish.
- Consequently, smaller subgroups may be considered less significant in the overall fairness assessment, which could lead to insufficient protection against discrimination for minority groups.

Calibration-based Fairness

- Calibration-based fairness is an approach that evaluates fairness by assessing the alignment between predicted values (i.e., model confidence) and actual outcomes.
- Fundamentally, it requires that the probabilistic predictions made by the model for a specific subgroup are well-calibrated, meaning they closely reflect the true outcome probabilities.

Calibration-based Fairness

Definition 2 (Multicalibration(Hebert et al., 2018))

Given a parameter $\alpha \in [0, 1]$, a predictor $f(x)$ is said to be $(\mathcal{C}, \alpha)$-multicalibrated if for all predicted values $v \in [0, 1]$ and for all $c \in \mathcal{C}$,

$$|\mathbb{E}[c(x) \cdot (y - v) \mid f(X) = v]| \leq \alpha. \tag{2}$$

- The left-hand side of the condition represents the bias between the actual outcomes and the predicted values. A smaller value indicates that the model is better calibrated.
- The parameter α defines the maximum allowable bias, meaning that a smaller α enforces a stricter calibration requirement.

Calibration-based Fairness

- Since computing the above condition requires high computational cost due to the need for conditional expectations, a relaxed version of the condition has been proposed.

Definition 3 (Weighted multicalibration(Gopalan et al., 2022))

Given a collection of subgroups $\mathcal{C}$ and a weight class $\mathcal{W}$, a predictor $f(x)$ is said to be $(\mathcal{C}, \mathcal{W}, \alpha)$-multicalibrated if for all $c \in \mathcal{C}$ and for all $w \in \mathcal{W}$,

$$|\mathbb{E}\left[c(x) \cdot w(f(X)) \cdot (y - f(X))\right]| \leq \alpha. \tag{3}$$

Calibration-based Fairness

- As the degree of the polynomial weights increases, the condition gradually converges to multicalibration.
- Calibration-based fairness enforces alignment between predicted probabilities and actual outcomes within each subgroup, thereby ensuring fairness for more fine-grained intersectional subgroups.
- However, strict fairness criteria such as multicalibration can incur high computational costs.
- Moreover, when the data for certain subgroups is sparse, it may be difficult or even impossible to satisfy the calibration condition.

Metric-based Fairness

- Metric-based fairness is an approach that extends the notion of individual fairness to protect intersectional groups.
- Individual fairness is grounded in the principle that ”similar individuals should receive similar predictions,” and it evaluates fairness by defining a distance metric over individuals' attributes and assessing prediction consistency with respect to this metric.
- To relax this situation, metric-multifairness, which requires that similar subgroups are treated similarly, is introduced.

Metric-based Fairness

Definition 4

For a small constant $\gamma > 0$ and an unknown similarity metric d, a predictor $f(x)$ is said to be $(\mathcal{C}, d, \tau)$-metric multifair if

$$\mathbb{E}_{(x,x')\sim\mathcal{A}}\left[|f(x) - f(x')|\right] \leq \mathbb{E}_{(x,x')\sim\mathcal{A}}\left[d(x, x')\right] + \gamma. \quad (4)$$

- This definition is the one adopted in the survey paper; however, in my opinion, the notation used is quite unconventional and potentially confusing.
- I dont't understand why using $\mathcal{C}$ and what is the $\mathcal{A}$.
- So I check the original paper(Kim et al., 2018).

Metric-based Fairness

Definition 5 (Metric multifairness(Kim et al., 2018))

Let $\mathcal{C} \subseteq 2^{\mathcal{X}^* \times \mathcal{X}^*}$ be a collection of comparisons and let $d : \mathcal{X}^* \times \mathcal{X}^* \to [0,2]$ be a metric. For some constant $\tau \geq 0$, a hypothesis f is said to be $(\mathcal{C}, d, \tau)$-metric multifair if for all $S \in \mathcal{C}$,

$$\mathbb{E}_{(X,X') \sim S}\left[\left|f(X) - f(X')\right|\right] \leq \mathbb{E}_{(X,X') \sim S}\left[d(X,X')\right] + \tau. \quad (5)$$

- This implies that for each subgroup S, if two instances X and X' have similar feature values (i.e., are close under the metric d), then their predictions $f(X)$ and $f(X')$ should also be similar.

Differential Fairness

Definition 6 (Differential Fairness(Foulds et al., 2020))

A predictor $f(x)$ is said to be ϵ-differentially fair if

$$e^{-\epsilon} \leq \frac{P(f(X) = y \mid x_i)}{P(f(X) = y \mid x_j)} \leq e^{\epsilon}, \tag{6}$$

holds for all tuples $(x_i, x_j) \in \mathcal{X} \times \mathcal{X}$ where $0 \leq P(x_j) \leq 1$.

- This is an intuitive intersectional definition of fairness: regardless of the combination of protected attributes, the probabilities of the out comes will be similar.

Differential Fairness

- This definition does not require the prediction outcomes to be exactly the same across groups, but it enforces that the ratio of outcomes between any two groups must remain within a bounded range.
- A smaller value of ϵ indicates lower disparity between groups, with $\epsilon = 0$ representing perfect fairness.
- This concept offers a stricter and more comprehensive notion of fairness compared to other definitions, as it protects against discrimination across all possible subgroup combinations.

Max-Min Fairness

- The idea essentially is to measure the value of the given fairness metric for every subgroup.
- Then, take the ratio of the minimum and maximum values from this given list.
- The further this ratio is from 1, the greater the disparity is between subgroups.
- Examples of fairness metric are Demographic parity, Conditional statistical parity, Equal opportunity and Group Benefit Equality etc(Ghosh et al., 2021).

Probabilistic Fairness

- Probabilistic Fairness relaxes the requirement of guaranteeing fairness for all subgroups using a probabilistic approach.

Definition 7

For $\epsilon \geq 0$ and $\delta \in [0, 1]$, a predictor is said to be (ϵ, δ)-probably intersectionally fair if

$$\mathbb{P}(U \geq \epsilon) \leq \delta, \tag{7}$$

where $U = u(f(X), s, s')$ measures unfairness for a randomly chosen prediction and two protected groups $s \neq s'$ to compare them.

Probabilistic Fairness

- This means that the probability of the unfairness exceeding ϵ is bounded above by δ, indicating that such violations are rare.
- Here, ϵ represents the allowable threshold for fairness violations, while δ denotes the tolerated probability of exceptions that exceed this threshold.
- In this setting, fairness violations exceeding ϵ are effectively tolerated for a proportion of groups or instances up to δ, which implies that severe discrimination may still persist for a small minority.

Discussions

- Intersectional fairness frameworks evaluate fairness not only across independent groups but also across more fine-grained subgroups formed by the intersection of those attributes.
- However, some approaches apply weights based on subgroup sizes, which can lead to relatively weaker protection for certain minority groups.
- Therefore, approaches like Max-Min and Differential Fairness aim to explicitly protect all possible subgroups.
- However, as the number of intersectional groups increases, these methods also suffer from data sparsity issues, making accurate fairness evaluation more challenging.

Contents

Challenge

- Data Sparsity
  1. As intersectional groups become more fine-grained, the problem of extreme data sparsity arises.
  2. For certain groups, the lack—or complete absence—of data makes fairness evaluation either infeasible or statistically unreliable.
- Selecting subgroups
  1. It is practically infeasible to consider all possible intersectional subgroups.
  2. Moreover, there is no clear criterion for determining which subgroups should be prioritized.
  3. Therefore, it is necessary to develop methods that can automatically identify meaningful subgroups or efficiently detect critical intersectional groups.

Contents

Reference

1. [Main] Gohar, U., Cheng, L. (2023). A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. arXiv preprint arXiv:2305.06969.
2. [Kearns et al., 2018] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In ICML. PMLR, 2018.
3. [Hebert-Johnson et al., 2018] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In ICML, volume 80, pages 1939–1948. PMLR, 2018.

Reference

4. [Gopalan et al., 2022] Parikshit Gopalan, Michael P Kim, Mihir A Singhal, and Shengjia Zhao. Low-degree multicalibration. In COLT, volume 178 of PMLR, pages 3193–3234, 2022.
5. [Kim et al., 2018] Michael P. Kim, Omer Reingold, and Guy N.Rothblum. Fairness through computationally-bounded awareness.In NeurIPS, page 4847–4857, 2018.
6. [Foulds et al., 2020] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In ICDE, pages 1918–1921, 2020.

Reference

7. [Ghosh et al., 2021] A. Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In AIDBEI, 2021.
8. [Molina and Loiseau, 2022] Mathieu Molina and Patrick Loiseau. Bounding and approximating intersectional fairness through marginal fairness. arXiv preprint arXiv:2206.05828, 2022.