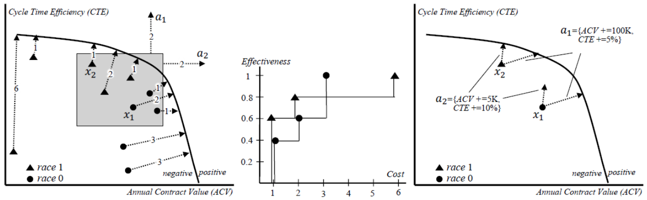
Fairness Aware Counterfactuals for Subgroups

March 21, 2024

Seoul National University

Introduction

- fairness of predction : captures the **explicit bias** reflected in the model's predictions
- an **implicit form of bias** is the difficulty for an individual (or a group thereof) to achieve recourse
- fairness of recourse : captures the notion that the protected subgroups should bear equal burden (mean cost of recourse)



Cycle Time Efficiency (CTE)
a_1
a_2
x_2
x_1
1
2
6
3
race 1
race 0
negative
positive
Annual Contract Value (ACV)
Effectiveness
1
0.8
0.6
0.4
0.2
Cost
1
2
3
4
5
6
Cycle Time Efficiency (CTE)
a_1={ACV +=100K, CTE +=5%}
a_2={ACV +=5K, CTE +=10%}
x_2
x_1
race 1
race 0
negative
positive
Annual Contract Value (ACV)

(a) (b) (c)

the arrow depicts the best action to achieve recourse
and the number indicates the cost of the actions

Fairness of Recourse for Subgroups

Preliminaries

- Feature Space

$$X = X_1 \times ... \times X_n,\ where\ X_n : protected\ \ feature \in \{0, 1\}$$

- Binary Classifier

$$h : X \rightarrow \{-1, 1\},\ 1 :\ \ favorable$$

- A : the set of possible actions
- a (action) : a set of changes to feature values

  e.g., $a = \{country\ \ \rightarrow\ \ US,\ education\ num\ \ \rightarrow\ \ 12\}$
- Counterfactual instance $x' = a(x)$ for a factual instance x
- If $h(x) = -1\ and\ h(a(x)) = 1$, then we say action a offers recourse to the individual x and is thus effective.

Preliminaries

- Recourse cost

$$\mathrm{rc}(A, x) = \begin{cases} \min\{\mathrm{cost}(a, x) \mid a \in A : h(a(x)) = 1\}, & \textit{if } \exists\ a \in A : h(a(x)) = 1 \\ c_\infty, & \textit{otherwise}. \end{cases}$$

- Subspace using predicate p

$$X_p \subseteq X \ , which\ is\ a\ conjunction\ of\ feature - level\ predicates$$

$$e.g.,\ p = (country = US)\ \land\ (education\ num > 9)$$

- Subpopulation group

$$G_p \subseteq D, as\ the\ set\ of\ affected\ individuals\ that\ satisfy\ p$$

$$G_p = \{x \in D \mid p(x)\}$$

- Protected subgroups

$$G_{p,1} = \{x \in D \mid p(x) \land x.X_n = 1\} \text{ and } G_{p,0} = \{x \in D \mid p(x) \land x.X_n = 0\}\,.$$

Effectiveness-Cost Trade-Off

- For a specific action a, we naturally define its **effectiveness** (eff) for a group G, as the proportion of individuals from G that achieve recourse through a:

$$\mathrm{eff}(a, G) = \frac{1}{|G|}|\{x \in G \mid h(a(x)) = 1\}|$$

- We want to examine how recourse is achieved for the group G through a set of possible actions A. We define the **aggregate effectiveness** (aeff) of A for G in two distinct ways.
  (micro viewpoint, macro viewpoint)

Effectiveness-Cost Trade-Off

- Define **micro-effectiveness** of set of actions A for group G as the proportion of individuals in G that can achieve recourse through some action in A

$$\text{aeff}_{\mu}(A, G) = \frac{1}{|G|}|\{x \in G \mid \exists a \in A, \text{eff}(a, x) = 1\}|$$

- Define **macro-effectiveness** of set of actions A for group G as the largest proportion of individuals in G that can achieve recourse through the same action in A,

$$\text{aeff}_{\mathrm{M}}(A, G) = \max_{a \in A} \frac{1}{|G|}|\{x \in G \mid \text{eff}(a, x) = 1\}|$$

Effectiveness-Cost Trade-Off

- Define the **in-budget actions** as the set of actions that cost at most c for any individual in G:

$$A_c = \{a \in A \mid \forall x \in G, cost(a, x) \leq c\}$$

- Define the **effectiveness-cost distribution** (ecd) as the function that for a cost budget c returns the aggregate effectiveness possible with in-budget actions:

$$\mathrm{ecd}(c; A, G) = \mathrm{aeff}(A_c, G)$$

ecd_μ, ecd_M micro, macro viewpoints of aggregate effectiveness.

- The **inverse effectiveness-cost distribution** function $\mathrm{ecd}^{-1}(\phi; A, G)$ takes as input an effectiveness level $\phi \in [0, 1]$ and returns the minimum cost required so that $\phi|G|$ individuals achieve recourse.

Definitions of Subgroup Recourse Fairness

Equal Effectiveness

$$\mathrm{aeff}(A, G_0) = \mathrm{aeff}(A, G_1)$$

Equal Choice for Recourse

$$|\{a \in A | \mathit{eff}(a, G_0) \leq \phi\}| = |\{a \in A | \mathit{eff}(a, G_1) \leq \phi\}|$$

Equal Effectiveness within Budget

$$\mathit{ecd}(c; A, G_0) = \mathit{ecd}(c; A, G_1)$$

Equal Cost of Effectiveness

$$\mathrm{ecd}^{-1}(\phi; A, G_0) = \mathrm{ecd}^{-1}(\phi; A, G_1)$$

Fair Effectiveness-Cost Trade-Off

$$\max_c |\mathrm{ecd}\,(c; A, G_0) - \mathrm{ecd}\,(c; A, G_1)| = 0$$

Equal Mean Recourse

$$\overline{\mathrm{rc}}(A, G_0) = \overline{\mathrm{rc}}(A, G_1)$$

$$, \mathit{where}\ \overline{\mathrm{rc}}(A, G) = \frac{1}{|G|} \sum_{x \in G} \mathrm{rc}(A, x)$$

Fairness-aware Counterfactuals for Subgroups

Fairness-aware Counterfactuals for Subgroups

Method overview

(a) **Subgroup and action space generation**: Used fp-growth algorithm to get subgroups with relatively frequent predicate and effective actions

(b) **Counterfactual summaries generation** : For each subgroup $G_p \in \mathcal{G}$, find set of valid actions and then extracts a subset V_p with each action having exactly the same cost for all individuals of G_p

(c) **CSC construction and fairness ranking** : evaluates all definitions on all subgroups , producing an unfairness score per definition, per subgroup

The outcome of this process is the generation, for each fairness definition, of a ranked list of CSC representations, in decreasing order of their unfairness score.

Experiments

Experimental Settings

- model : logistic regression model, trained on the training set
- subgroup and action generation : used fp-growth algorithm with threshold 1%
- cost functions : implement according to which, the cost of a change of a feature value v to the value v' is defined as follows:
  1. Numerical features: $|norm(v) - norm(v')|$, where norm is a function that normalizes values to [0, 1].
  2. Categorical features: 1 if $v \neq v'$, and 0 otherwise.
  3. Ordinal features: $|pos(v) - pos(v')|$, where pos is a function that provides the order for each value.

Experiments

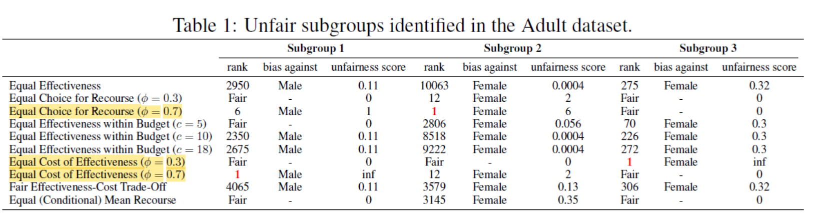
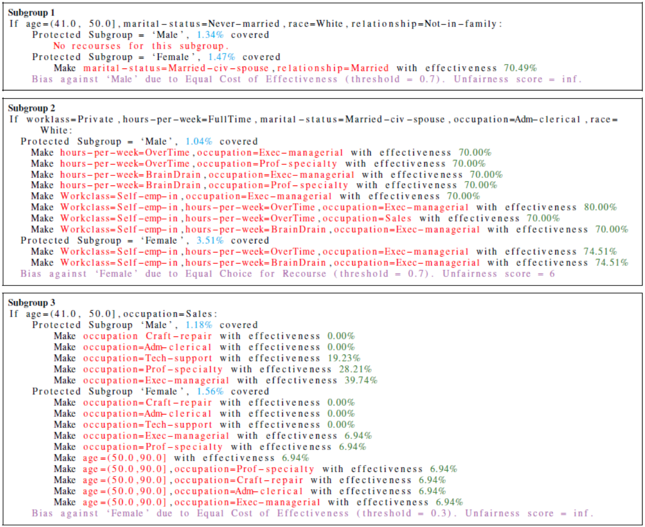


Table 1: Unfair subgroups identified in the Adult dataset.

	Subgroup 1			Subgroup 2			Subgroup 3		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	2950	Male	0.11	10063	Female	0.0004	275	Female	0.32
Equal Choice for Recourse ($\phi = 0.3$)	Fair	-	0	12	Female	2	Fair	-	0
Equal Choice for Recourse ($\phi = 0.7$)	6	Male	1	1	Female	6	Fair	-	0
Equal Effectiveness within Budget ($c = 5$)	Fair	-	0	2806	Female	0.056	70	Female	0.3
Equal Effectiveness within Budget ($c = 10$)	2350	Male	0.11	8518	Female	0.0004	226	Female	0.3
Equal Effectiveness within Budget ($c = 18$)	2675	Male	0.11	9222	Female	0.0004	272	Female	0.3
Equal Cost of Effectiveness ($\phi = 0.3$)	Fair	-	0	Fair	-	0	1	Female	inf
Equal Cost of Effectiveness ($\phi = 0.7$)	1	Male	inf	12	Female	2	Fair	-	0
Fair Effectiveness-Cost Trade-Off	4065	Male	0.11	3579	Female	0.13	306	Female	0.32
Equal (Conditional) Mean Recourse	Fair	-	0	3145	Female	0.35	Fair	-	0

Table 1 presents three subgroups which were ranked at position 1 according to three different definitions: Equal Cost of Effectiveness ($\phi = 0.7$), Equal Choice for Recourse (ϕ= 0.7) and Equal Cost of Effectiveness ($\phi = 0.3$), meaning that these subgroups were detected to have the highest unfairness according to the respective definitions. For each subgroup, its rank, bias against, and unfairness score are provided for all definitions presented in the left-most column. When the unfairness score is 0, we display the value "Fair" in the rank column. Note that subgroups with exactly the same score w.r.t. a definition will receive the same rank.

Experiments



```
Subgroup 1
If age=(41.0, 50.0], marital-status=Never-married, race=White, relationship=Not-in-family:
    Protected Subgroup = 'Male', 1.34% covered
        No recourses for this subgroup.
    Protected Subgroup = 'Female', 1.47% covered
        Make marital-status=Married-civ-spouse, relationship=Married with effectiveness 70.49%
    Bias against 'Male' due to Equal Cost of Effectiveness (threshold = 0.7). Unfairness score = inf.
```

```
Subgroup 2
If worklass=Private, hours-per-week=FullTime, marital-status=Married-civ-spouse, occupation=Adm-clerical, race=
     White:
  Protected Subgroup = 'Male', 1.04% covered
    Make hours-per-week=OverTime, occupation=Exec-managerial with effectiveness 70.00%
    Make hours-per-week=OverTime, occupation=Prof-specialty with effectiveness 70.00%
    Make hours-per-week=BrainDrain, occupation=Exec-managerial with effectiveness 70.00%
    Make hours-per-week=BrainDrain, occupation=Prof-specialty with effectiveness 70.00%
    Make Workclass=Self-emp-in, occupation=Exec-managerial with effectiveness 70.00%
    Make Workclass=Self-emp-in, hours-per-week=OverTime, occupation=Exec-managerial with effectiveness 80.00%
    Make Workclass=Self-emp-in, hours-per-week=OverTime, occupation=Sales with effectiveness 70.00%
    Make Workclass=Self-emp-in, hours-per-week=BrainDrain, occupation=Exec-managerial with effectiveness 70.00%
  Protected Subgroup = 'Female', 3.51% covered
    Make Workclass=Self-emp-in, hours-per-week=OverTime, occupation=Exec-managerial with effectiveness 74.51%
    Make Workclass=Self-emp-in, hours-per-week=BrainDrain, occupation=Exec-managerial with effectiveness 74.51%
  Bias against 'Female' due to Equal Choice for Recourse (threshold = 0.7). Unfairness score = 6
```

```
Subgroup 3
If age=(41.0, 50.0], occupation=Sales:
    Protected Subgroup 'Male', 1.18% covered
        Make occupation Craft-repair with effectiveness 0.00%
        Make occupation=Adm-clerical with effectiveness 0.00%
        Make occupation=Tech-support with effectiveness 19.23%
        Make occupation=Prof-specialty with effectiveness 28.21%
        Make occupation=Exec-managerial with effectiveness 39.74%
    Protected Subgroup 'Female', 1.56% covered
        Make occupation=Craft-repair with effectiveness 0.00%
        Make occupation=Adm-clerical with effectiveness 0.00%
        Make occupation=Tech-support with effectiveness 0.00%
        Make occupation=Exec-managerial with effectiveness 6.94%
        Make occupation=Prof-specialty with effectiveness 6.94%
        Make age=(50.0,90.0] with effectiveness 6.94%
        Make age=(50.0,90.0], occupation=Prof-specialty with effectiveness 6.94%
        Make age=(50.0,90.0], occupation=Craft-repair with effectiveness 6.94%
        Make age=(50.0,90.0], occupation=Adm-clerical with effectiveness 6.94%
        Make age=(50.0,90.0], occupation=Exec-managerial with effectiveness 6.94%
    Bias against 'Female' due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf.
```

Figure 3: Comparative Subgroup Counterfactuals for the subgroups of Table 1.

Conclusion

Conclusion

- delve deeper into the difficulty (or burden) of achieving recourse, an implicit and less studied type of bias
- an efficient implementation that allows the detection and ranking of subgroups according to the introduced fairness definitions and produces intuitive, explainable subgroup representations in the form of counterfactual summaries