How Much Does Cluster Fairness Cost? A Counterfactual-based Approach

Antonia Karra, Georgios Vardakas, Evaggelia Pitoura, Aristidis Likas
{a.karra, g.vardakas, pitoura, arly}@uoi.gr
Department of Computer Science & Engineering, University of Ioannina, Greece

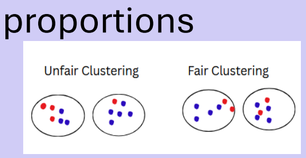
Motivation

- **Clustering is widely used** in sensitive domains (education, healthcare, finance, employment)
- Existing fair clustering methods (balance, social) **change the cluster structure** to protect fairness-but **how sensitive are clusters to these interventions?**
- **Key gap:** *What is the **cost** of enforcing fairness?*
  - At the **structural level** → how much the clustering solution changes
  - At the **individual level** → how much an individual must change to be reassigned fairly.

Fairness Definition

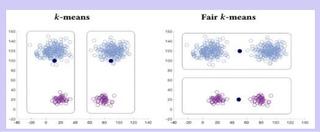
Balance Fairness

- Ensures that each cluster reflects the overall group proportions



Social Fairness

- Focuses on equal clustering cost across groups.



Proposed Cost Measures: Similarity-based Costs

Goal: Compare unfair and fair clustering.

Definitions

$C = \{C_1, \dots, C_k\}$: Original (unfair) clustering
$C^F = \{C_1^F, \dots, C_k^F\}$: Fair clustering
Y: $\emptyset \neq Y \subseteq X$ a protected group(e.g., female, male)
Y' $\subseteq$ Y: the set of misaligned points

Metrics:

- **NMI (Normalized Mutual Information):** Measures structural similarity between unfair and fair clustering.

$$\text{NMI} - \text{cost}(\text{C}, C^F, Y) = 1 - \text{NMI}(\text{C}, C^F, \text{Y})$$

- **Misalignment Cost:** Counts instances assigned to different clusters after alignment.

$$mis - cost(C, C^F, Y) = |\{\text{x} \in \text{Y}: \text{C(x)} \neq C^F(\text{x})\}|$$

Proposed Cost Measures: Counterfactual-based Costs

Goal: Quantify the effort (minimal changes) required to make the clustering **fair** by fixing **misaligned points**.

Counterfactual

- A counterfactual $x'$of a point x is a minimally modified version of x that moves it to a different cluster.

$$x' = \arg\min_{y \in R^d} d(x, y), s.t. d(y, c_\ell) < d(y, c_i), 1 \leq \ell \leq k.$$

- $dist(x, y)$: measures how much x must change to belong to another cluster (e.g., Euclidean).

Counterfactual Cost

- **For a misaligned point x:**

$$cfcost(C, C^{AF}, x) = d_{C_i \to C_j}(x)$$

distance to move x from its unfair cluster C_i to the aligned fair clustering C_j^{AF}

- **For a group Y:**

$$cfcost(C, C^F, Y) = \frac{\sum_{\text{x} \in \text{Y}'} cfcost(C, C^F, \text{x})}{|\text{Y}'|}$$

Feature-level Contribution

- **For a feature m in instance x:** Measures how much each feature contributes to unfairness.

$$r_m(x) = \frac{(x_m - x'_m)^2}{\sum_{l=1}^{d} (x_l - x'_l)^2}$$

- **Average over group Y:**

$$r_m(Y) = \frac{\sum_{x \in Y'} r_m(x)}{|Y'|}$$

Experimental Setup

•Datasets: **Adult** sensitive = *sex* ,**Credit Card** sensitive = marriage, **Bank** sensitive = *marital status,* **Student** sensitive = *sex* (600 samples)

Methods compared:

- k-means (unfair baseline)
- Balance fairness
- Social fairness

•10 runs for each k with different initial centers

Results and Insights

RQ1: Social vs Balance Fairness

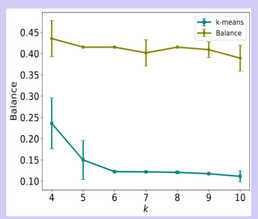
***Social Fairness*:**

- Lower NMI-cost, fewer structural changes **than *Balance fairness***.
- Keeps clusters closer to the original, more stable **than *Balance fairness***.
- Fewer reassignments **than *Balance fairness***.
- Higher counterfactual cost → individuals must change more **than *Balance fairness***.

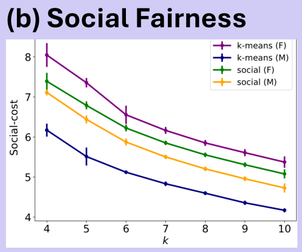
RQ2: Relation between unfairness, similarity-based costs, and counterfactual cost

- NMI & misalignment follow the same trend: more misalignment → higher costs.
- Counterfactual cost shows **how much a data point must change** to reach fairness.
- Fairness and cost are not directly correlated: a group with higher unfairness does not necessarily face higher similarity-based or counterfactual costs.

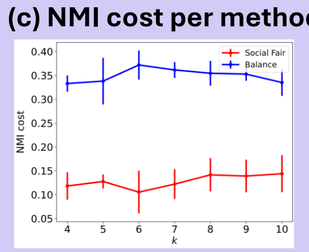
(a) Balance Fairness



(b) Social Fairness

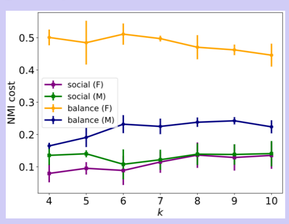


(c) NMI cost per method

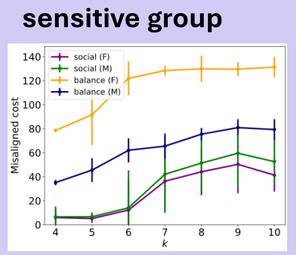


Adult dataset: Fairness in original and fair k-means

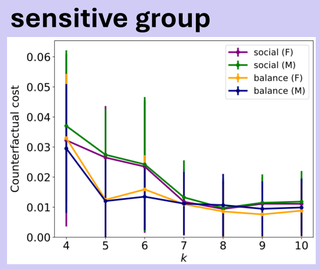
(a) NMI cost per sensitive group



(b) Misaligned cost per sensitive group

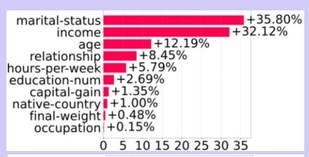


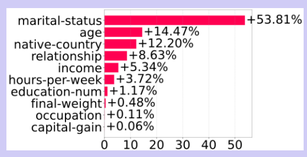
(c) Counterfactual cost per sensitive group

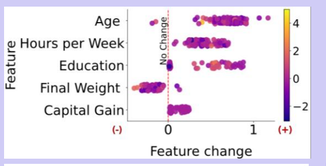


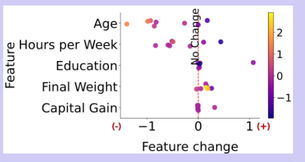
RQ3: What features contribute most to the cost?

- We analyze which features drive counterfactual changes (Adult dataset, k = 5)
- The model mainly uses **marital status** to adjust fairness, even though **sex** is not used directly.
- The plots show that different features matter more or less, depending on the fairness method and the group

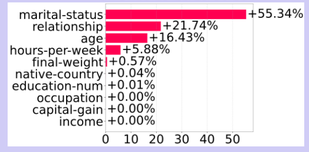


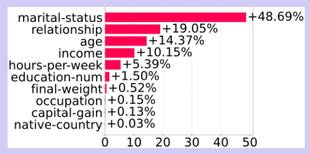




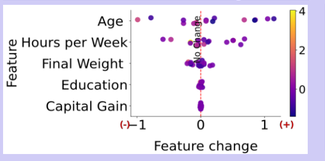


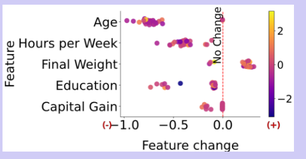
Adult dataset: Feature changes for both fairness objectives.





Adult dataset: Feature changes for both fairness objectives.





Summary & Future work

Summary

- Our work **fills the gap** between clustering fairness cost, and counterfactual explanations.
- Proposed cost measures to evaluate fairness in clustering.

Future Work

- Extend the analysis to other clustering algorithms, such as deep clustering and Gaussian mixture models.
- Explore alternative definitions of fairness and study their impact on clustering and counterfactuals.

Counterfactuals for Clustering and Deep Clustering: Explainability, Fairness and Quality
Project Number: 15940
Hellenic Foundation for Research & Innovation (H.F.R.I)