[Re] Robust Fair Clustering: A Novel Fairness Attack and Defense Framework

Iason Skylitsis
Zheng Feng
Idries Nasim
Camille Niessink

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Fair clustering algorithms aim to ensure fairness in sensitive applications like healthcare, yet their robustness against adversarial attacks has received little attention. This study reproduces and extends the work of "**Robust Fair Clustering**" by *Chhabra et al.* [1], *which* evaluates fair clustering vulnerabilities via a *black-box adversarial attack* and proposes a robust defense method. Our objective is to reproduce the paper's three main claims:

1. The attack is capable of degrading the fairness performance, by perturbing a small percentage of protected group memberships, in the examined fair clustering models: *Fair K-Center* (**KFC**)[2], *Fair Spectral Clustering* (**FSC**)[3], and *Scalable Fairlet Decomposition* (**SFD**)[4].
2. KFC, FSC, and SFD demonstrate a lack of robustness to adversarial influence, exhibiting significant volatility in terms of fairness utility metrics such as Balance and Entropy.
3. *Consensus Fair Clustering* (**CFC**) exhibits high resilience against the proposed fairness attack, offering a robust solution for achieving fair clustering.

The Attack

In the original paper, the authors propose a novel black-box attack that aims to reduce the fairness utility of fair clustering algorithms by perturbing a small percentage of samples' protected group memberships.

Fairness Attack

The black-box attack perturbs a small subset of protected group memberships (G_A) to reduce fairness utility for the remaining group (G_D).

Threat Model

The adversary controls $G_A \subseteq G$ and observes the clustering outputs of a fair clustering algorithm (F), aiming to compromise fairness for $G_D = G/G_A$.

How it works

The adversary perturbs G_A, modifies the dataset by combining G_A and G_D and iteratively adjusts G_A to find perturbations that degrade fairness the most.

Optimization Approach

Utilizes zeroth-order optimization to identify the best perturbations without requiring access to the internal workings of the clustering algorithm.

Datasets

Following the original methodology, we selected the number of samples used for *MNIST-USPS*, *Office-31*, *DIGITS*, and *Yale* (cropped and uncropped). For *MTFL*, we balanced the dataset by randomly selecting 2,000 images (each with and without glasses)

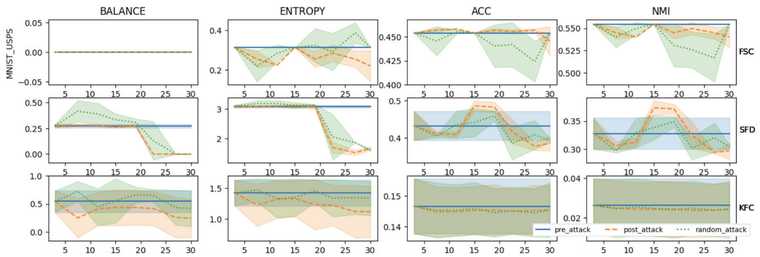
Dataset	Num. samples	Num. categories	Protected attribute	Description
MNIST-USPS	3,800	10	Sample source	Handwritten digits
Office-31	1,293	31	Domain source	Office objects
DIGITS	3,594	10	Source of image	Handwritten digits
Yale	2,414	38	Azimuth and elevation	Frontal-face
uncropped *Yale*	2,414	38	Azimuth and elevation	Full-body & Background
MTFL	2,000	2	Glasses usage	Face

Reproduction - Experiments and results

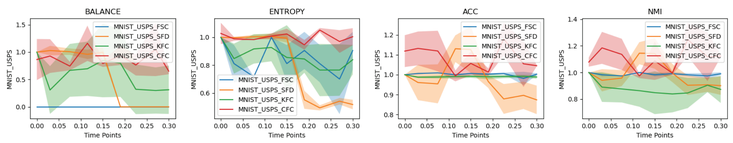
1. Impact on fairness utility (Balance and Entropy) and clustering utility (NMI and ACC) on the MNIST USPS dataset.

Algorithm	Metrics	*MNIST-USPS* Pre-Attack	Post-Attack	Change (%)	Match Original Findings	Random Attack	Change (%)	Match Original Findings
SFD	Balance	0.282 ± 0.001	0.300 ± 0.001	(+)**6.382**		0.330 ± 0.001	(+)17.02	
	Entropy	3.063 ± 0.151	3.104 ± 0.001	(+)**1.339**		3.147 ± 0.000	(+)2.742	
	NMI	0.315 ± 0.000	0.358 ± 0.000	(+)13.65		0.346 ± 0.000	(+)9.841	
	ACC	0.419 ± 0.000	0.473 ± 0.000	(+)12.89		0.456 ± 0.000	(+)8.831	
FSC	Balance	0.000 ± 0.000	0.000 ± 0.000	(-)100.0	✓	0.000 ± 0.000	(-)100.0	✓
	Entropy	0.327 ± 0.000	0.241 ± 0.001	(-)26.30	✓	0.301 ± 0.001	(-)7.951	✓
	NMI	0.549 ± 0.000	0.543 ± 0.000	(-)1.093	✓	0.538 ± 0.000	(-)2.004	✓
	ACC	0.450 ± 0.000	0.454 ± 0.000	(+)0.889	✓	0.443 ± 0.000	(-)1.556	✓
KFC	Balance	0.557 ± 0.324	0.350 ± 0.299	(-)37.16	✓	0.724 ± 0.117	(+)30.20	✓
	Entropy	1.355 ± 0.374	1.202 ± 0.351	(-)11.29	✓	1.417 ± 0.417	(+)4.576	✓
	NMI	0.000 ± 0.000	0.000 ± 0.000	(-)100.0	✓	0.000 ± 0.000	(-)100.0	✓
	ACC	0.147 ± 0.000	0.146 ± 0.000	(-)0.680	✓	0.145 ± 0.000	(-)1.361	✓

2. Pre-attack, post-attack (black-box) and random attack results on fairness utility (Balance and Entropy) and clustering utility (ACC and NMI) for the MNIST-USPS dataset.



3. Pre-attack and post-attack (black-box) ratio trends for FSC, SFD, KFC, and CFC on fairness utility (Balance and Entropy) and clustering utility (ACC and NMI) for the MNIST-USPS dataset.



Summary of the reproducibility: Claim 1 | Claim 2 | Claim 3

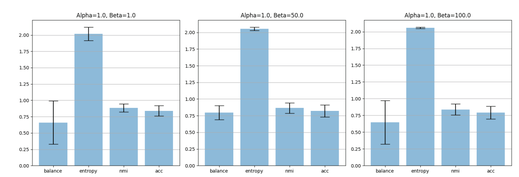
Results beyond the original paper:

Additional Attack Methods

We experimented with various attack strategies, focusing on the *Office-31* dataset using the SFD algorithm, to evaluate its robustness. Despite the marginal improvements, CFC showed resilience to the new attacks.

Metric	Attack Balance	Attack Min. Cluster Ratio	Combined Attack
Balance	0.149 ± 0.004	0.149 ± 0.004	$\mathbf{0.144 \pm 0.011}$
Entropy	9.764 ± 0.037	9.764 ± 0.037	$\mathbf{9.715 \pm 0.089}$
NMI	0.857 ± 0.009	0.857 ± 0.009	$\mathbf{0.857 \pm 0.002}$
ACC	0.757 ± 0.026	0.757 ± 0.026	$\mathbf{0.753 \pm 0.016}$
Min. Cluster Ratio	$0.061, \pm 0.002$	$0.061, \pm 0.002$	$\mathbf{0.059 \pm 0.005}$
Cluster L1	0.178 ± 0.007	0.178 ± 0.007	$\mathbf{0.183 \pm 0.002}$
Cluster KL	0.099 ± 0.009	0.099 ± 0.009	$\mathbf{0.104 \pm 0.006}$
Silhouette diff	-0.009 ± 0.001	-0.008 ± 0.002	-0.005 ± 0.002
Entropy Group A	3.291 ± 0.016	3.291 ± 0.016	$\mathbf{3.287 \pm 0.035}$
Entropy Group B	$\mathbf{3.357 \pm 0.010}$	$\mathbf{3.357 \pm 0.010}$	3.360 ± 0.009
ARI	$\mathbf{0.677 \pm 0.021}$	$\mathbf{0.677 \pm 0.021}$	0.681 ± 0.010
Silhouette Score	$\mathbf{0.153 \pm 0.006}$	$\mathbf{0.153 \pm 0.006}$	0.157 ± 0.003

An ablation study was performed to assess the impact of α, which controls the fair clustering loss, and β, which controls the structural preservation loss, on the CFC model. The results showed minimal influence, emphasizing the robustness of the model's architecture.



Discussion and conclusion

About the methods.

Overall CFC had **superior** performance and was more **robust.**

About the overall reproducibility of the paper.

Most results from the original paper were successfully reproduced.

References

[1] **Chhabra, A. et al.** "Robust Fair Clustering: A Novel Fairness Attack and Defense Framework." *International Conference on Learning Representations, 2022*

[2] **Harb, E. et al.** "KFC: A Scalable Approximation Algorithm for k-center Fair Clustering." *Proceedings of Neural Information Processing Systems (NeurIPS), 2020*

[3] **Kleindessner, M. et al.** "Guarantees for Spectral Clustering with Fairness Constraints." *Proceedings of the 36th International Conference on Machine Learning, 2019.*

[4] **Backurs, A. et al.** "Scalable Fair Clustering." *Proceedings of the 36th International Conference on Machine Learning, 2019*

Acknowledgements

We would like to thank Fernando P. Santos for the excellent organization of the UvA MSc. FACT course, Luca Pantea for his invaluable supervision, and Matteo Tafuro for the help in designing the poster.