Group Fair Clustering Revisited - Notions and Efficient Algorithm

Extended Abstract

Shivam Gupta
Indian Institute of Technology Ropar, India
shivam.20csz0004@iitrpr.ac.in

Ganesh Ghalme
Indian Institute of Technology Hyderabad, Kandi, India
ganeshghalme@ai.iith.ac.in

Narayanan C. Krishnan
Indian Institute of Technology Palakkad, India
ckn@iitpkd.ac.in

Shweta Jain
Indian Institute of Technology Ropar, India
shwetajain@iitrpr.ac.in

ABSTRACT

This paper considers the problem of group fairness in clustering. We propose a new fairness notion which strictly generalizes existing notions, and we theoretically analyze the relationships between several existing notions. Finally, we propose a simple and efficient greedy round-robin-based algorithm (FRAC$_{OE}$) and extensive experiments to validate its efficacy across multiple datasets.

KEYWORDS

Machine Learning; Unsupervised Learning; Clustering; Fairness

ACM Reference Format:

Shivam Gupta, Ganesh Ghalme, Narayanan C. Krishnan, and Shweta Jain. 2023. Group Fair Clustering Revisited - Notions and Efficient Algorithm: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Fair clustering is an important problem and appears in many situations [8, 10, 11, 22, 26]. Recommender systems cluster their users based on their features and provide recommendations based on the cluster to which a given user is assigned [14]. Suppose the optimal clustering results in a skewed distribution of the users from a given protected group. In that case, the algorithm that provides recommendations, such as job listing based on cluster identity, may give vastly different recommendations across different groups. Clustering is used in many other applications with high societal impact, including facility location[15], job suitability assessments [23], facial recognition [12, 20], and outlier detection [2, 25].

Motivated by such applications, we revisit the first notion of fairness (called `Balance`) introduced by Chierichetti et al. [9] in the clustering setting. When there are only two groups – advantaged and disadvantaged, the `Balance` notion aims to maximize the ratio of people from the disadvantaged and advantaged groups in each cluster. A maximally balanced clustering algorithm tries to achieve ratio same as that present in dataset (called *dataset ratio*). The notion of `Balance` was generalized by Bera et al. [6] using Minority Protection (MP) and Restricted Dominance (RD) that provide lower and upper bounds on data points from each group in every cluster. Along similar lines, Ziko et al. [33] provides a continuous metric called Fairness Error (FE) to enable the use of optimization-based approaches. There are two major drawbacks. First, the resulting clusters can be highly skewed. Secondly, the existing algorithms are computationally complex [6, 7, 24] or require extensive hyper parameter tuning [1, 5, 19, 21, 31, 32].

This paper introduces a new notion of fairness, called as $\boldsymbol{\tau}$`-ratio` fairness and show that satisfying $\boldsymbol{\tau}$`-ratio` fairness also satisfies the τ'`-Balance` property by establishing the relationship between different existing group fairness notions theoretically (See Lemmas 1-4). The paper then proposes a simple and efficient round-robin-based algorithm for the $\boldsymbol{\tau}$`-ratio` that admits $2^{k-1}(\alpha+2)$−approximate solution to fair clustering. Here α is the approximation ratio of vanilla clustering, and k is the desired number of clusters. Finally, through extensive experiments on four datasets, we show the proposed algorithm's efficacy on fairness and objective cost. Further, the cost does not grow exponentially with the k. As a byproduct, our algorithm also solves the capacitated clustering with an ideal cluster size of n/k (See [4, 18]) by setting parameters of $\boldsymbol{\tau}$`-ratio` fairness to satisfy *dataset ratio*.

2 THE MODEL

Let $X \subseteq \mathbb{R}^d$ be a finite set of points that needs to be partitioned into k clusters. A k-clustering algorithm produces a partition $\mathcal{C} = \{C_j\}_{j=1}^k$ of X into k subsets with centers $C = \{c_j\}_{j=1}^k$ using an assignment function $\phi : X \to C$ which maps each point to corresponding cluster center. We consider that each point $x_i \in X$ is associated with a *single* protected attribute ρ_i (say, gender), which takes different group values (like male, female) from the set denoted by $[m]$. Furthermore, let $d : X \times X \to \mathbb{R}_+$ be a distance metric that measures the dissimilarity between features. The vanilla (unconstrained) clustering algorithm minimize the following: $L_p(X, \mathcal{C}, C, \phi)=\left(\sum_{C_j \in \mathcal{C}} \sum_{x_i \in C_j} d(x_i, c_j)^p\right)^{\frac{1}{p}}$. The fairness is measured by a given vector $\boldsymbol{\tau}=\{\tau_a\}_{a=1}^m$ with $0 \le \tau_a \le \frac{1}{k}$ $\forall a \in [m]$. If X_a, n_a represent data points and the number of points having protected attribute value a in X respectively, then $\boldsymbol{\tau}$`-ratio` fairness ensures that each cluster has a predefined fraction of points for every protected group value, i.e. $\sum_{x_i \in C_j} \mathbb{I}(\rho_i = a) \ge \tau_a n_a$, $\forall C_j \in \mathcal{C}$ and $\forall a \in [m]$. Existing discrete group fair notions include τ-Balance i.e. $\left(\min_{a,b \in [m]} \left(\frac{\sum_{x_i \in C_j} \mathbb{I}(\rho_i = a)}{\sum_{x_i \in C_j} \mathbb{I}(\rho_i = b)}\right)\right) \ge \tau$, $\boldsymbol{\tau}$-MP i.e., $\sum_{x_i \in C_j} \mathbb{I}(\rho_i = a) \ge \tau_a |C_j|$ and $\boldsymbol{\tau}$-RD i.e., $\sum_{x_i \in C_j} \mathbb{I}(\rho_i = a) \le \tau_a |C_j|$, $\forall C_j \in \mathcal{C}, a \in [m]$. We now discuss the relationship between group fair notions with a binary protected attribute (i.e., takes only two values $a, b \in [m]$).

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

Lemma 1. *If a cluster* $C_j \in C$ *is* $\boldsymbol{\tau}$-`ratio` *fair, then it also satisfies* $\min_{a,b}\left(\frac{\tau_a}{1-k\tau_b+\tau_b}\frac{n_a}{n_b}\right)$ – *Balance. Further when* τ_a=τ_b=$1/k$ *in* $\boldsymbol{\tau}$-`ratio` *fairness then it is* $\min_{a,b}(n_a/n_b)$-`Balance` *clustering.*

Lemma 2. *A fair clustering instance exists which satisfies* τ'*-Balance with* $\tau' > 0$ *and has arbitrarily low* $\boldsymbol{\tau}$-`ratio`.

Lemma 3. *The cluster satisfying both* $\boldsymbol{\tau}'$*-MP and* $\boldsymbol{\tau}$*-RD ensures* $\min\left(\frac{\tau'_a}{\tau_b}, \frac{\tau'_b}{\tau_a}\right)$*-Balance. Furthermore, satisfying only one of them does not ensure* τ*-Balance.*

Lemma 4. *If a cluster satisfies* τ*-balance then it is also* $\boldsymbol{\tau}$*-MP with* $\boldsymbol{\tau}$=$\{\frac{1}{2}, \frac{\tau}{1+\tau}\}$ *and* $\boldsymbol{\tau}$*-RD with* $\boldsymbol{\tau}$=$\{\frac{1}{1+\tau}, \frac{1}{2}\}$ *for* $\{a, b\}$ *respectively.*

All the above results prove that $\boldsymbol{\tau}$-`ratio` is a generalized notion. Thus, we focus on designing an algorithm satisfying $\boldsymbol{\tau}$-`ratio` fairness while minimizing objective cost irrespective of p.

3 PROPOSED ALGORITHM: FRAC$_{OE}$

We now propose the algorithm that we call **F**air **R**ound-robin **A**lgorithm for **C**lustering **O**ver **E**nd (FRAC$_{OE}$) in Algorithm 1. Our post-processing algorithm derives fair clustering on top of vanilla clustering via a fair assignment procedure described in Algorithm 2. We will now look into the convergence guarantees and objective cost approximation factors in comparison to optimal cost.

Theorem 1. *FRAC$_{OE}$ algorithm results in* $2^{k-1}(\alpha{+}2)$*-approximation to the fair clustering problem for any* k *and* $\boldsymbol{\tau}$.

Proposition 1. *There exists an instance with arbitrary centers and data points on which FRAC$_{OE}$ achieves* 2*-approximation factor compared to optimal assignment.*

Convergence: FRAC$_{OE}$ ensures fairness at the end and makes corrections for every point only once. Thus, given the convergence of the vanilla clustering ([16, 17]), FRAC$_{OE}$ converges in finite time.

Algorithm 1: τ-FRAC$_{OE}$$(X, k, \boldsymbol{\tau}, m, p)$

```
Let (C, ϕ) be solution to vanilla (k, p)-clustering.
if τ-ratio fairness is met then
    return (C, ϕ)
    else
        return FairAssignment(C, X, k, τ, m, p, ϕ)
    end
end
```

4 EXPERIMENTAL RESULT AND DISCUSSION

We compare the performance of FRAC$_{OE}$ against state-of-the-art (SOTA) on different *benchmarking datasets*- Adult (Census) [28], Bank [29], Diabetes [30] and Census-II [27]. The bank dataset has ternary valued protected group, whereas other have binary valued group. However, the datasets differ in sizes and number of features. We show the performance of FRAC$_{OE}$ on metrics, Objective Cost L_p (p=2) and τ-Balance. We take vanilla k-means and k-median as our initial clustering algorithms. Further, we consider Vanilla k-means/k-median, Ziko et al., Backurs et al., Bera et al. as SOTA baselines. In Ziko et al., we consider two variations - tuned (re-tune hyper-parameters) and untuned (hyper-parameter value same as reported in [33]). Results are average and standard deviation over 10 independent trials. The code is available publicly [13].

Algorithm 2: FairAssignment$(C, X, k, \boldsymbol{\tau}, m, p, \phi)$

```
Fix a random center ordering and φ̂(x_i) ← 0 ∀x_i ∈ X.
for ℓ ← 1 to m do
    for t ← 1 to τ_a n_a do
        for j ← 1 to k do
            φ̂(argmin_{x_i ∈ X_a : φ̂(x_i)=0} d(x_i, c_j)) = j
        end
    end
    For all x_i ∈ X_a such that φ̂(x_i) = 0, set φ̂(x_i) = φ(x_i)
end
Recompute centers Ĉ with respect to new allocation φ̂.
return (Ĉ, φ̂).
```

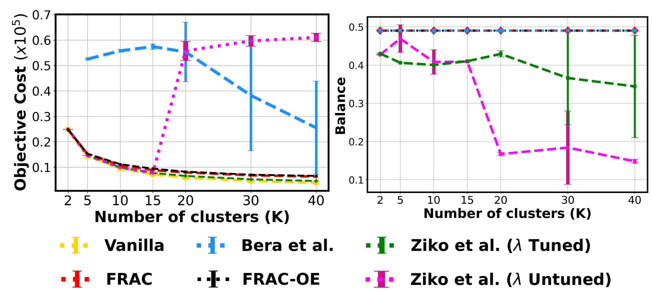

Figure 1: The plot shows evaluation metrics over varying k for k-means setting on adult dataset (*dataset ratio* 0.49).

We analyze different approaches on varying k for τ=$\{1/k\}_{a=1}^{m}$. The results obtained are plotted in Fig. 1 for k=2, 5, 10, 15, 20, 30, and 40 on adult dataset. The complete results on fixed and varying k for k-means/k-median are available in arXiv version [13]. We further check if initial center ordering in Fair Assignment procedure is a critical factor in deciding objective cost. We observe cost over 100 random permutations of k(=10)-means centers that FRAC$_{OE}$ is center invariant. We further report results for FRAC$_{OE}$ on general $\boldsymbol{\tau}$ vector [13]. The runtime of vanilla, FRAC$_{OE}$, Bera et al., Ziko et al. (with tuning) and Ziko et al. (without tuning) are 11.8, 11.55, 188.98, 1310.61 and 15.9 respectively. Thus, FRAC$_{OE}$ can handle fairness with a better cost at considerably less runtime.

5 DISCUSSION

In this paper, a novel $\boldsymbol{\tau}$-`ratio` fairness notion that generalizes existing notion is proposed. We convert fair clustering into a fair assignment problem and propose a simple, efficient round-robin algorithm. We theoretically show cost approximation guarantees. We also provide the relationship between all the discrete group fair notions. Immediate future direction includes tackling multiple protected attributes, and achieving individual and group fairness together.

ACKNOWLEDGEMENT

The authors would like to thank the prime minister research fellowship for generously funding the primary author of this work to travel and present the work at the conference. The research is further supported by the Department of Science & Technology, India, with grant number SRG/2020/001138.

REFERENCES

[1] Savitha Sam Abraham, Deepak Padmanabhan, and Sowmya S Sundaram. 2020. Fairness in Clustering with Multiple Sensitive Attributes. In *EDBT/ICDT 2020 Joint Conference*. 287–298.

[2] Matteo Almanza, Alessandro Epasto, Alessandro Panconesi, and Giuseppe Re. 2022. K-Clustering with Fair Outliers. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 5–15. https://doi.org/10.1145/3488560.3498485

[3] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. In *International Conference on Machine Learning*. PMLR, 405–413.

[4] Arindam Banerjee and Joydeep Ghosh. 2006. Scalable clustering algorithms with balancing constraints. *Data Mining and Knowledge Discovery* 13, 3 (2006), 365–395.

[5] Eustasio Barrio, Hristo Inouzhe, and Jean-Michel Loubes. 2019. Attraction-Repulsion clustering with applications to fairness. *arXiv preprint arXiv:1904.05254* (2019).

[6] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. 2019. Fair Algorithms for Clustering. *Advances in Neural Information Processing Systems* 32 (2019), 4954–4965.

[7] Matteo Böhm, Adriano Fazzone, Stefano Leonardi, and Chris Schwiegelshohn. 2020. Fair clustering with multiple colors. *arXiv:2002.07892* (2020).

[8] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. 2019. Proportionally fair clustering. In *International Conference on Machine Learning*. PMLR, 1032–1041.

[9] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair clustering through fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 5036–5044.

[10] Seyed Esmaeili, Brian Brubach, Aravind Srinivasan, and John Dickerson. 2021. Fair clustering under a bounded cost. *Advances in Neural Information Processing Systems* 34 (2021), 14345–14357.

[11] Seyed A Esmaeili, Sharmila Duppala, John P Dickerson, and Brian Brubach. 2022. Fair Labeled Clustering. *arXiv preprint arXiv:2205.14358* (2022).

[12] Sixue Gong, Xiaoming Liu, and Anil K Jain. 2021. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3414–3424.

[13] Shivam Gupta, Ganesh Ghalme, Narayanan C Krishnan, and Shweta Jain. 2021. Efficient Algorithms For Fair Clustering with a New Fairness Notion. *arXiv preprint arXiv:2109.00708* (2021).

[14] Wenxing Hong, Siting Zheng, Huan Wang, and Jianchao Shi. 2013. A job recommender system based on user clustering. *J. Comput.* 8, 8 (2013), 1960–1967.

[15] Christopher Jung, Sampath Kannan, and Neil Lutz. 2020. Service in your neighborhood: Fairness in center location. *Foundations of Responsible Computing (FORC)* (2020).

[16] Shivaram Kalyanakrishnan. 2016. k-means clustering. https://www.cse.iitb.ac.in/~shivaram/teaching/old/cs344+386-s2017/resources/classnote-2.pdf. [Online; accessed 29-May-2022].

[17] Andreas Krause. 2016. Clustering and k-means. https://las.inf.ethz.ch/courses/lis-s16/hw/hw4_sol.pdf. [Online; accessed 29-May-2022].

[18] Tai Le Quy and Eirini Ntoutsi. 2021. Towards fair, explainable and actionable clustering for learning analytics.. In *EDM*.

[19] Woojin Lee, Hyungjin Ko, Junyoung Byun, Taeho Yoon, and Jaewook Lee. 2021. Fair Clustering with Fair Correspondence Distribution. *Information Sciences* 581 (2021), 155–178. https://doi.org/10.1016/j.ins.2021.09.010

[20] Peizhao Li, Han Zhao, and Hongfu Liu. 2020. Deep Fair Clustering for Visual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Suyun Liu and Luis Nunes Vicente. 2021. A Stochastic Alternating Balance k-Means Algorithm for Fair Clustering. *arXiv:2105.14172* (2021).

[22] Evi Micha and Nisarg Shah. 2020. Proportionally Fair Clustering Revisited. In *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[23] Deepak Padmanabhan. 2020. Whither Fair Clustering?. In *AI for Social Good: Harvard CRCS Workshop*.

[24] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. 2019. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*. Springer, 232–251.

[25] Hanyu Song, Peizhao Li, and Hongfu Liu. 2021. Deep Clustering Based Fair Outlier Detection. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &; Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 1481–1489. https://doi.org/10.1145/3447548.3467225

[26] Suhas Thejaswi, Bruno Ordozgoiti, and Aristides Gionis. 2021. Diversity-aware k-median: Clustering with fair center representation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 765–780.

[27] UCI. 1990. Census-II Dataset. https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29. [Online; accessed 15-August-2021].

[28] UCI. 1994. Adult Dataset (Census). https://archive.ics.uci.edu/ml/datasets/Adult. [Online; accessed 15-August-2021].

[29] UCI. 2014. Bank Dataset. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. [Online; accessed 15-August-2021].

[30] UCI. 2014. Diabetes Dataset. https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008. [Online; accessed 15-August-2021].

[31] Bokun Wang and Ian Davidson. 2019. Towards fair deep clustering with multi-state protected variables. *arXiv preprint arXiv:1901.10053* (2019).

[32] Hongjing Zhang and Ian Davidson. 2021. Deep Fair Discriminative Clustering. *arXiv preprint arXiv:2105.14146* (2021).

[33] Imtiaz Masud Ziko, Jing Yuan, Eric Granger, and Ismail Ben Ayed. 2021. Variational Fair Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11202–11209.