Improve Fairness of Deep Clustering to Prevent Misuse in Segregation

Anonymous Authors[1]

Abstract

Deep clustering scales well due to mini-batch training and usually achieves better clustering performance as joint nonlinear embedding and clustering has been shown beneficial (Yang et al., 2017). However, deep clustering approaches may generate "monochromatic" cluster that all members in that cluster are from the same demographic group, even if their moral status information is not directly used in the algorithm. This might be misused to segregate people by sensitive attributes like gender, race etc, which is usually not desirable. To this end, we propose a differentiable approach, which learns a discriminative but less biased cluster assignment function. By improving the fairness degree, our model can be utilized to redress the algorithmic bias in deep clustering to prevent potential misuse in segregation.

1. Introduction

The purpose of clustering is often given as trying to find groups of instances that are similar to each other but dissimilar to others. Common clustering methods include the spectral clustering (Ng et al., 2001; von Luxburg, 2007) that tries to effectively maximize the edges within a subgraph (clusters), the Louvain method (Blondel et al., 2008) that tries to maximize the number of edges per cluster, and k-means style algorithms (Lloyd, 1982; Ostrovsky et al., 2006) that attempt to minimize the distortion. Due to limitations of clustering in high-dimensional space or linear subspace, deep clustering approaches (Yang et al., 2016; Dizaji et al., 2017; Law et al., 2017; Xie et al., 2016; Yang et al., 2017) are proposed to simultaneously perform nonlinear embedding and clustering and achieve superior performance.

However, when the clusters are of people or other entities deserving moral status (Warren, 1997) then the issue of fairness of the clustering must be considered. For example, we may build special interest groups by clustering the authors' publication records and maybe 90% of the members in some group are male researchers. It is worth noting that such clustering result could be sound according to classic evaluation metrics and may truly reflect the current state in this area based on the publication records. However, such result will aggravate the imbalance and finally lead to sex segregation in the future, which definitely contradicts the intention of building special interest groups. In other words, clustering without fairness consideration may lead to misuses of this technique.

The standard way of ensuring fairness is to declare some attributes as *sensitive* and these attributes are *not* used by the algorithm when forming the clusters but rather are given to the algorithm to ensure the clustering is fair. The first paper on this topic (Chierichetti et al., 2017) shows a clever way of pre-processing data set into chunks. This step guarantees that when k-center and k-median algorithms are applied to the chunks they produce fair clusters. Besides, there are several recent papers adding fairness constraints to k-means (Schmidt et al., 2018) and spectral clustering (Kleindessner et al., 2019) or devise scalable variants of previous algorithm (Backurs et al., 2019). However, there still are some limitations in existing work:

First, those algorithms provide fair variants of classic clustering approaches but not applicable to mini-batch training based deep clustering approaches, although the unfairness still exists in deep ones.

More importantly, existing work (Chierichetti et al., 2017; Backurs et al., 2019; Kleindessner et al., 2019) treats sensitive attributes like gender, ethnicity as **Boolean variables** and attempts to balance the populations of two demographic groups within every cluster so that each cluster has equal (within mathematical limits) numbers of males and females, non-whites and whites etc. However, most of sensitive attributes are not intrinsically binary. Consider the classic eight tenants of demographic status: sex, race, age, disability, color, creed, national origin and religion. None are now considered binary with each having more than two responses in most forms used to collect data. For example, nowadays gender contains not only male and female but also other states like transgender, queer, etc. However, the area of fair deep clustering with multi-state sensitive attribute has not been studied to our knowledge.

To address the limitations above, we first define a general fairness measure that works for both binary and multi-state sensitive attribute, which is intuitive and efficient to compute. Based on the fairness measure, we explore the potential of incorporating fairness consideration in deep clus-

tering models. The core idea of our method is to learn a "neural" assignment function that assigns a label to each instance and achieves two aims: i) separating the instances into high-quality clusters and ii) making the composition of those clusters be equalized with respect to the sensitive attribute. The second one is achieved by exploiting the idea that the instances with the same demographic status form a demographic group and inherently a "fairoid" (fairness centroid). That is if the demographic status is race, there would be a fairoid for each of white, black, hispanic etc. Then we wish the cluster centroids to be **equi-distant** from each and every fairoid to ensure fairness. In Section 3.2 we discuss why this is reasonable to ensure fairness.

The contributions of this paper are summarized as follows:

(1) We define fair clustering problem for multi-state sensitive attributes. This has significant practical importance as most sensitive attributes are not inherently binary.

(2) We propose a novel mechanism to incorporate notion of fairness into deep clustering. All instances for a particular state of sensitive attribute defines a fairoid and finding centroids equi-distant to the fairoids produces fair clustering.

2. Measure of Fairness

Notations. We first introduced the notations needed to define the problem of fair clustering with multi-state sensitive attribute: Let $X \in \mathbb{R}^{N\times D}$ denote N data points with D-dimension features. The aim of clustering C is to learn an assignment function $\alpha : \mathbf{x} \to \{1, ..., K\}$, which allocates N data points into K disjoint clusters $C_1, C_2, ..., C_K$ ($\bigcap_{k=1}^{K} C_k = \emptyset$). In addition, let the sensitive attribute A be available for fairness consideration. If there are T unique values in A, the data points X can be naturally partitioned into T disjoint demographic groups $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_T$ ($\bigcap_{t=1}^{T} \mathcal{G}_t = \emptyset$) by another function $\chi : \mathbf{x} \to \mathcal{A}$. Here $\mathcal{A}$ is a finite set of unique values in sensitive attribute, $\mathcal{A} = \{a_1, a_2, ..., a_T\}$. When considering fairness notion for the clustering problem, the aim is to get a fair partition *w.r.t.* specific sensitive attribute A but still a useful clustering *w.r.t.* topology of original data.

Chierichetti et al. first define the measure of fairness for a cluster when sensitive attribute is binary (*e.g.*, male/female), based on disparate impact (Feldman et al., 2015) and $p\%$-rule (Saini, 2006), which is called "balance":

$$\text{balance}(C_k) = \min\left[\frac{N_k^1}{N_k^2}, \frac{N_k^2}{N_k^1}\right] \in [0, 1] \qquad (1)$$

Here C_k denotes the k-th cluster and N_k^1 and N_k^2 represent the populations of the first and second demographic groups in C_k. $\text{balance}(C_k) = 1$ means the k-th cluster is perfectly balanced and $\text{balance}(C_k) = 0$ means the k-th cluster only contains members from one demographic group. The balance of the whole clustering C (all clusters) is defined as $\text{balance}(C) = \min_k \text{balance}(C_k)$. Such a measure of fairness has clear meaning. For example, if there are 60 males and 40 females in a cluster, the fairness of the cluster is defined as the population of minor group (female) over the population of major group (male), *i.e.*, the disparity between two demographic groups. Considering $P(\mathbf{x} \in \mathcal{G}_i \big| \mathbf{x} \in C_k) = \frac{N_k^i}{|C_k|}, i = 1, 2$, we can re-write Formula 1 more formally as follows:

$$\min\left[\frac{P(\mathbf{x} \in \mathcal{G}_1 \big| \mathbf{x} \in C_k)}{P(\mathbf{x} \in \mathcal{G}_2 \big| \mathbf{x} \in C_k)}, \frac{P(\mathbf{x} \in \mathcal{G}_2 \big| \mathbf{x} \in C_k)}{P(\mathbf{x} \in \mathcal{G}_1 \big| \mathbf{x} \in C_k)}\right] \in [0, 1] \qquad (2)$$

We can further understand the fairness measure above and extend it to multi-state sensitive attribute by introducing the "demographic histogram".

Definition 2.1 *Demographic histogram h_k in the k-th cluster is the histogram whose bins are the unique states of sensitive attribute and the value on each bin is the proportion of that group in that cluster, which can be shown as:*

$$h_k = \Big[P(\mathbf{x} \in \mathcal{G}_1 \big| \mathbf{x} \in C_k), ..., P(\mathbf{x} \in \mathcal{G}_T \big| \mathbf{x} \in C_k)\Big] \qquad (3)$$

When sensitive attribute is binary, we can understand the "balance" via the demographic histogram with #bins = 2: Here the bins are "male" and "female" and the values on two bins are the proportions of male/female in the k-th cluster. Thus, "balance" measures the disparity between values on two bins by ratio.

When the sensitive attribute is multi-state, the demographic histogram has more than 2 bins and value on each bin represents the proportion of the demographic group in this cluster. To derive a measure of fairness for this case, a straightforward approach is to compute all balances scores between every possible pair of bins in the demographic histogram. For example, if the sensitive attribute race has 4 states: {White, Black, Asian, Other}, we can compute $\binom{4}{2}$ balances between every possible pair: *e.g.* White *v.s.* Black, White *v.s.* Asian, Black *v.s.* Asian etc. There are two drawbacks of such treatment: a) Unlike the measure defined in Formula 2, we have several scores instead of a single score that reflects the overall fairness degree of a cluster; b) When number of states (T) is huge, we need to calculate $\binom{T}{2}$ such scores, which are neither insightful nor convenient.

A New Fairness Measure. Considering drawbacks of the simple treatment above, we propose a new measure of fairness for clustering with multi-state sensitive attributes, which is convenient and insightful. We generalize the idea

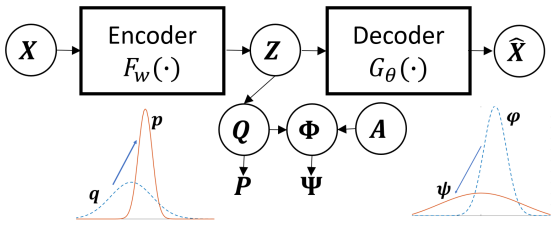

Figure 1. Proposed model for fair deep clustering. X is the original data and encoder $F_{\mathcal{W}}$ maps X into latent representations Z. Decoder G_θ reconstructs $\hat{X}$ from Z. A denotes given sensitive attribute. Meanings of Q, P, Φ, Ψ are explained in Section 3.

of balance and Calders-Verwer score that measures the disparity between values on 2 bins into measuring the disparity among values on $T \geq 2$ bins in demographic histogram. Intuitively, the larger disparity is, the more efforts we need to make it a perfectly fair cluster (uniform demographic histogram), which resembles the process of moving the earth.

Definition 2.2 ***Earth-mover Distance as a Fairness Measure***. *Let the demographic histogram for the k^{th} cluster be $h_k \sim \mathcal{H}_k$ and the histogram of perfectly fair cluster be $u_k \sim \mathcal{U}\big(\inf \mathcal{A}, \sup \mathcal{A}\big)$. Then a fairness measure is simply the distance between them. We utilize the Wasserstein distance W_p as $\mathcal{D}(\cdot,\cdot)$:*

$$\mathcal{D}(\mathcal{H}_k, \mathcal{U}) = \Big(\inf_{J_k \in \mathcal{J}_k} \int \|h_k - u_k\|^p dJ_k \Big)^{1/p} \quad (4)$$

Here $\mathcal{J}_k(\mathcal{H}_k, \mathcal{U})$ is the joint distribution of $\mathcal{H}_k$ and $\mathcal{U}$.

3. Deep Learning Formulation

In this section we outline our formulation for deep fair clustering. We first begin by explaining why finding centroids that are equi-distant from the fairoids is desirable.

3.1. Overview

In deep clustering formulation, a latent representation $\mathbf{z} = F_{\mathcal{W}}(\mathbf{x})$ and cluster assignment function $\alpha : \mathbf{z} \to \{1, ..., K\}$ are simultaneously learned. Thus the K centroids $\mu_1, \mu_2, ..., \mu_K$ of clusters $C_1, C_2, ..., C_K$ are simply $\mu_k = \mathbb{E}_{\alpha(\mathbf{z}_i)=k}[\mathbf{z}] = \frac{1}{|C_k|} \sum_{i=1}^{N} \mathbf{z}_i 1[\alpha(\mathbf{z}_i) = k]$. Each fairoids $\pi_1, \pi_2, ..., \pi_T$ of demographic groups $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_T$ is represented as $\pi_t = \mathbb{E}_{\chi(i)=a_t}[\mathbf{z}] = \frac{1}{|\mathcal{G}_t|} \sum_{i=1}^{N} \mathbf{z}_i 1[\chi(i) = a_t]$, given the sensitive annotations A. We propose an inductive model (Figure 1) to directly learn the mapping function α that transforms X into soft assignment $Q \in \mathbb{R}^{N \times K}$, which is fairer *w.r.t.* protected attribute $A \in \mathbb{R}^N$ but still discriminative enough. At the beginning, cluster centroids $M = \{\mu_k\}_{k=1}^K$ and fairoids $\Pi = \{\pi_t\}_{t=1}^T$ can be calculated from input data X, given sensitive attribute A, as well as the initial encoding function $f_{\mathcal{W}}$. During the training process, clustering objective and fairness objective are jointly optimized until convergence to refine the network parameter $\mathcal{W}$, cluster centroids M, and fairoids Π for improving clustering performance as well as fairness.

3.2. Differentiable Demographic Histogram

Consider the measure of fairness in Definition 2.2, which requires that the demographic histogram should be closer to uniform histogram. However, it is challenging to build a trainable model because calculating the demographic histogram is a non-differentiable operation. Instead, we can approximate the value h_k^t on the t-th bin in demographic histogram h_k during mini-batch training. Intuitively, bigger value on t-th bin in demographic histogram h_k means that the k-th cluster is more monochromatic towards demographic group $\mathcal{G}_t$. In other words, there are more overlapping instances between cluster C_k and group $\mathcal{G}_t$. In deep clustering, the centroids and cluster assignment function are based on the lower-dimensional representation $Z = F_{\mathcal{W}}(x)$. Suppose the lower-dimensional representations of instances in C_k are $Z_k \sim U_k$ and those of instances in $\mathcal{G}_t$ are $Z_t \sim V_t$. The disparity between C_k and $\mathcal{G}_t$ could be computed by Maximum Mean Discrepancy (MMD) (Gretton et al., 2006) between $Z_k \sim U_k$ and $Z_t \sim V_t$:

$$\mathrm{MMD}(U_k, V_t) = \Big\| \mathbb{E}_{U_k}\big[F_{\mathcal{W}}(X_k)\big] - \mathbb{E}_{V_t}\big[F_{\mathcal{W}}(X_t)\big] \Big\| \quad (5)$$

The bigger the disparity above is, the fewer common instances between C_k and $\mathcal{G}_t$ are so that the value h_k^t on the t-th bin in (normalized) demographic histogram h_k should be smaller. Following the idea that uses student's t-distribution for soft assignment (Xie et al., 2016), we can approximate h_k by ϕ_k:

$$h_k^t \approx \phi_k^t = \frac{\Big(1 + \frac{\|\mu_{\alpha(\mathbf{z}_i)} - \pi_t\|^2}{\alpha}\Big)^{-\frac{\alpha+1}{2}}}{\sum_{t'} \Big(1 + \frac{\|\mu_{\alpha(\mathbf{z}_i)} - \pi_{t'}\|^2}{\alpha}\Big)^{-\frac{\alpha+1}{2}}}, \forall \alpha(\mathbf{z}_i) = k \quad (6)$$

3.3. Objective Function

To make the demographic histogram closer to uniform histogram, we can minimize the KL divergence between them. However, this objective may be difficult to optimize. Instead, we utilize the self-training strategy (Xie et al., 2016) to minimize the KL divergence between the differentiable demographic histogram ϕ_k and its self-training target ψ_k, which is smoothed ϕ_k:

$$\psi_{kt} = \frac{\hat{\phi}_{kt} / f_t}{\sum_t' \hat{\phi}_{kt'} / f_t'} \quad (7)$$

Methods	ACC ↑	EMD (gender) ↓	Balance (gender) ↑	EMD (race) ↓
AE + KM (Vincent et al., 2008)	**0.758**	0.438	0.299	0.493
DCN (Yang et al., 2017)	**0.758**	0.441	0.295	0.496
DEC (Xie et al., 2016)	0.744	0.396	0.363	0.500
Ours	0.735	**0.368**	**0.430**	**0.477**

Table 1. Experimental results (over 10 trials) on Adult dataset with gender/race as sensitive attributes. "Balance" metric is defined in (Chierichetti et al., 2017). Best performance is in **bold black**. "↑" and "↓" mean that larger value indicates better or worse performance.

Here $\hat{\phi}_{kt} = \sqrt[\beta]{\phi_{kt} + \epsilon}$, $\beta \geq 2$ and ϵ is a small number for numerical stability. $f_t = \sum_t \phi_{kt}$ is the frequency over histogram bins $a_1, a_2, ...a_T$. Let $\Phi = [\phi_1; ...; \phi_K]$ and $\Psi = [\psi_1; ...; \psi_K]$. Our objective function for fairness consideration can be written as:

$$\mathcal{L}_{fr} = \mathrm{KL}(\Psi||\Phi) = \sum_k \sum_t \psi_{kt} \log \frac{\psi_{kt}}{\phi_{kt}} \tag{8}$$

The overall objective $\mathcal{L}$ combines $\mathcal{L}_{fr}$ with the clustering loss $\mathcal{L}_{cl}$ defined in (Xie et al., 2016) (to improve cluster purity):

$$\mathcal{L} = \mathcal{L}_{cl} + \gamma \mathcal{L}_{fr} = \mathrm{KL}(P||Q) + \gamma \mathrm{KL}(\Psi||\Phi) \tag{9}$$

P denotes cluster assignment matrix and Q is its self-training target defined in (Xie et al., 2016).

4. Experiments

We evaluate the performance of our proposed model on Adult dataset[1] with two types of sensitive attributes: gender (Female/Male) and race (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black). We compare our method, in terms of both clustering quality and fairness, with representative deep clustering baselines including DEC (Xie et al., 2016), Autoencoder + k-means (AE+KM) and Deep Cluster Network (DCN) (Yang et al., 2017) to show that our approach can improve the fairness of deep clustering while only having minor loss in terms of clustering quality.

To evaluate the clustering quality of our approach and compared methods mentioned above, we use the widely used clustering accuracy (ACC) metric (higher values indicates better performance). To evaluate fairness of clustering, we use the protocol defined in Definition. 2.2. We use Earth-Mover distance between demographic histogram and uniform histogram as metric named "EMD", where smaller value denotes better performance. Because "gender" is a binary sensitive attribute, we also evaluate the methods in terms of balance score as described in Equation 1 from previous work (Chierichetti et al., 2017).

[1]https://archive.ics.uci.edu/ml/datasets/adult

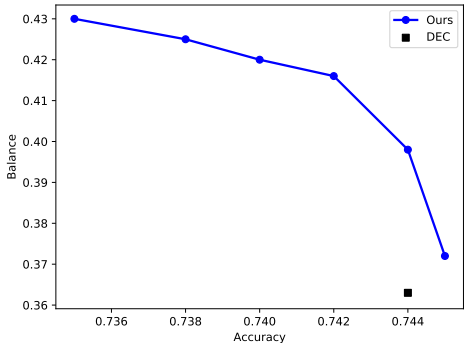

Figure 2. Trade-off between the clustering accuracy and balance of clusters on Adult dataset when hyper-parameter γ varies. The black point shows clustering accuracy and FWD of base model DEC and the blue line shows the performance of ours.

As shown in Table 1, ACC of our model is mildly lower than the base model DEC. On the other hand, the fairness measured by EMD or Balance is remarkably improved.

Because the fairness loss could be seen as a "fairness" regularization term for clustering loss, we are also interested in how the hyper-parameter γ controls the trade-off between clustering accuracy and fairness. Thus we testify our model under different values of γ ranging from 10^{-2} to 10^3 when sensitive attribute is gender, the results can be seen in Figure 2. With larger γ our model is fairer but less accurate.

5. Conclusion

In this work, we propose a differentiable model to improve fairness of deep clustering approaches, when the sensitive attribute is binary or multi-state. We compare our approach with representative deep clustering approaches on a real-world with different types of sensitive attributes. The experimental results demonstrate our approach can improve fairness while only having minor loss in terms of clustering quality. By controlling one hyper-parameter, our approach can provide flexible trade-off between clustering accuracy and fairness. This model can be used to redress the algorithmic bias in deep clustering to prevent potential misuse in segregation.

References

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., and Wagner, T. Scalable fair clustering. *arXiv preprint arXiv:1902.03519*, 2019.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pp. 5029–5037, 2017.

Dizaji, K. G., Herandi, A., Deng, C., Cai, W., and Huang, H. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 5747–5756. IEEE, 2017.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pp. 259–268, 2015.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *NIPS*, pp. 513–520. MIT Press, 2006.

Kleindessner, M., Samadi, S., Awasthi, P., and Morgenstern, J. Guarantees for spectral clustering with fairness constraints. *ICML*, 2019.

Law, M. T., Urtasun, R., and Zemel, R. S. Deep spectral clustering learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1985–1994, 2017.

Lloyd, S. P. Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28(2):129–136, 1982. doi: 10.1109/TIT.1982.1056489.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *NIPS*, pp. 849–856. MIT Press, 2001.

Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. The effectiveness of lloyd-type methods for the k-means problem. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pp. 165–176, 2006.

Saini, D. Book review: Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing, by dan biddle (gower publishing limited, alderhshot, 2005). *Vision—The Journal of Business Perspective*, 10:113–114., 01 2006.

Schmidt, M., Schwiegelshohn, C., and Sohler, C. Fair coresets and streaming algorithms for fair k-means clustering. *arXiv preprint arXiv:1812.10854*, 2018.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pp. 1096–1103, 2008. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Warren, M. A. *Moral status: Obligations to persons and other living things*. Clarendon Press, 1997.

Xie, J., Girshick, R. B., and Farhadi, A. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 478–487, 2016.

Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 3861–3870, 2017.

Yang, J., Parikh, D., and Batra, D. Joint unsupervised learning of deep representations and image clusters. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5147–5156, 2016.