Whither Fair Clustering?

Deepak P
Queen's University Belfast, UK
deepaksp@acm.org

Abstract

Within the relatively busy area of fair machine learning that has been dominated by classification fairness research, fairness in clustering has started to see some recent attention. In this position paper, we assess the existing work in fair clustering and observe that there are several directions that are yet to be explored, and postulate that the state-of-the-art in fair clustering has been quite parochial in outlook. We posit that widening the normative principles to target for, characterizing shortfalls where the target cannot be achieved fully, and making use of knowledge of downstream processes can significantly widen the scope of research in fair clustering research. At a time when clustering and unsupervised learning are being increasingly used to make and influence decisions that matter significantly to human lives, we believe that widening the ambit of fair clustering is of immense significance.

1 Introduction

Fair Machine Learning (Fair ML) is a flourishing discipline of study that has gathered much attention in the last several years, starting from an early pioneering work in [Dwork *et al.*, 2012]. Of late, a newly instituted interdisciplinary conference series, ACM FAT*/FAccT[1] has bolstered further interest. Broadly, there have been two fairness streams explored in Fair ML literature: (i) *individual fairness* that prefers adherence to *treating similar people similarly*, and (ii) *group fairness* which involves ensuring some notion of *'fair'* distribution of analytics results across groups defined on *sensitive attributes* such as gender, race, ethnicity and religion. Over the past years, significant progress has been made in fair classification, with emergence of computational notions such as *independence*, *separation* and *sufficiency* [Barocas *et al.*, 2017]. Supervised learning has the luxury of availability of labelled data that encompasses information of historical decisions. In the case of binary decision making (success/fail), the chasm between the base success rate for each sensitive class (e.g., gender) and their representation within the training data provides a fertile ground for the pursuit of fairer supervised learning. On the other hand, unsupervised or exploratory learning does not assume availability of labels in the data, making fairness within unsupervised learning quite a distinct notion from the former. It may be noted that unsupervised learning is of growing significance in ML, and is often referred to as the *next frontier in AI*[2].

Fairness in unsupervised machine learning may be expected to increase in importance with the broadening scope of unsupervised learning, facilitated by the growth of data volumes far outpacing any attempt at getting them labelled. This data growth has been facilitated in the public sector by an expansion of the methods for 'passive' data collection, where data is collected through safety/surveilance cameras and IoT devices as part of smart city infrastructure. In the private sector, the user's mobility patterns are available to map services (e.g., Google Maps, Bing Maps) and black-box car insurance providers, social interests are available to social media companies (e.g., Facebook, Twitter), and with the advent of PDAs (e.g., Echo, Home), web tech giants can potentially have access to audio conversations within homes.

Clustering, arguably the most popular task in unsupervised learning, has seen much fairness-oriented research attention in the last few years. The pioneering work on this stream was one on data pre-processing to facilitate fair clustering [Chierichetti *et al.*, 2017]. Our analysis of the community's approach to the task across the 15+ papers in literature leads us to an argument that the literature has been quite restricted in scope. This is arguably due to treating it as a well-defined computational task, despite it being much more nuanced due to being situated within the space of a sophisticated landscape of normative principles. We assess fair clustering literature in the backdrop of the political philosophy around fairness and justice, and make the following arguments:

- **Normative Target:** The normative principles, the space of values targeted, that have been used across fair clustering formulations have been quite narrow in scope, and significantly narrower than in the case of fair supervised learning. This is to be seen in the backdrop of the plethora of normative principles available in political philosophy. In particular, we observe that most clustering formulations have relied on *alleviating disparate impact through representational parity*, a pursuit of *group fairness* that relates to

[1] https://facctconference.org/

[2] https://bit.ly/2zWjTEo - Yann LeCun, 2018 Turing Laureate

luck egalitarianism [Lang, 2009] when sensitive attributes are considered as manifestations of *brute luck* choices. It is also noteworthy that the relationship between egalitarianism and discrimination avoidance has been argued to be nuanced [Binns, 2018].

- **Shortfall Characterization:** Clustering, as a dataset-level optimization task, is very well understood to be complex. Given the complexities, most formulations fall short of achieving the representational parity goal that they target for. Techniques have focused on either bounding the shortfall theoretically, or illustrating empirically that the quantum of shortfall is tolerable. The critical missing piece is that the shortfall, while being quantified as above, has been left uncharacterized. It has not been elucidated as to what *what kind of data objects are likely to suffer more or less* from the shortfall. For usage in practical scenarios, especially within public sector, absence of such a characterization of the shortfall could be a potential dealbreaker.
- **Application Space:** Most clustering formulations seek to achieve their fairness goal in each of the clusters in the output. In a way, they are being application-agnostic and try to ensure that *whatever be the downstream application that makes use of the clustering, there is some form of fairness assurance* that the techniques provide. However, typical clustering outputs could be used in order to decide from among a small set of decision choices, which could additionally be placed somewhere in the spectrum of positive or negative. Information about the downstream usage of clustering outputs could both: (i) improve the ability to optimize better for the chosen optimization goal, and (ii) render the formulation more suited to particular domains.

2 Case Study: Clustering for Job Shortlisting

Towards putting forth the arguments raised above, we will use the backdrop of a setting where clustering is used to inform consequential decisions directly. Consider the case of a heavily oversubscribed job vacancy, where manual perusal of each of them is out of question. Such a scenario is routinely encountered in the case of government jobs in populated developing countries[3]. We consider a pipeline of clustering usage for such a scenario. *First*, the received applications would be subject to clustering using a similarity measure that is relevant to assessing the suitability to the job, to generate perhaps hundreds of clusters. *Second*, a representative application from each cluster, perhaps the *medoid*, would be subject to manual assessment for suitability to the job. *Third*, the arrived assessment for the medoid, likely one of *shortlist, reject, scrutinize further* would be applied to all applications in its cluster. *Fourth*, those labelled *scrutinize further* by virtue of enough ambiguity on the suitability assessment, could be subject to further clustering, or if there is enough manual bandwidth available, subject to individual manual assessments. As an illustrative example to appreciate the need for clustering fairness within this pipeline, observe that generating a set of gender-skewed clusters could help reinforce gender stereotypes that play a part in manual perusal. The cluster-level decisions made over such gender-skewed clusters could then become implicitly gender-aligned. Data analytics' role in reinforcing social and economic inequalities has been the topic of several recent books [O'neil, 2016].

3 Normative Target: What to Optimize for

The normative principle used in a number of fair clustering formulations is that of assuming that each attribute be either considered *sensitive* or task-relevant, followed by *targeting to preserve the dataset-wide distribution of objects along sensitive attributes within each cluster*. For example, with gender regarded sensitive, this translates to ensuring that the gender ratio within each cluster be very similar to the gender ratio in the dataset. Different fair clustering formulations differ in the *number* and *kind* of sensitive attributes they admit; such a characterization of literature appears in [Abraham *et al.*, 2020] (Ref. Table 1). The similarities between data objects on task-relevant attributes are deemed to be relevant to the task *in the same manner*; weights may be attached to attributes to differentiate the quantum of influence, but the *nature* of the influence remains similar.

First, the crisp binary distinction between sensitive and non-sensitive attributes begs apparent criticism. There are often attributes on which discrimination could be avoided, but not necessarily as strongly. For example, the *age* or *region/province* attribute could be such; there is typically a higher degree of tolerance towards skew in age and regions (e.g., urban skew), but purely age-homoegeous or region-homogeneous clusters are nevertheless undesirable. *Second*, while sensitive attributes are often outcomes of what are called *brute luck*, there exist other luck-influenced attributes whose placement is not clear in the sensitive/task-relevant dichotomy. These include the likes of *option luck* [Dworkin, 2002] which relate to choices made on the face of considerable uncertainty of how things would turn out. For example, a career-break due to startup failure is unlike *brute luck*, but still not something that the candidate should heavily scored down on. Some addressal of option luck may be achieved by manually engineering covariate features to control. *Third*, there is significant space to expand the normative target, notably the Rawlsian choice [Rawls, 1971] in the fairness-efficiency trade-off. There are other possibilities *outside fairness* within the so-called *patterned notions* (as outlined in [Nozick, 1974]), and prefer to allocate resources in accordance with patterns such as *need* or *moral desert*[4]. This would be especially true of hiring in the public sector where the government could use such patterned allocation in order to associate esteem with certain values. This would require identifying attributes that correlate with *need* and *desert* and treating them specially so that people with similar needs and deserts be clustered together. Desert may often need to be specified through attribute-combinations; a candidate from a backward region who has shown exceptional interest in a trade despite limited access to facilities may be considered as scoring high on moral desert. Similarity search has explored multiple unconventional and complex aggregation operators [Deepak and Deshpande, 2015]. The lack of diversity

[3] https://www.bbc.co.uk/news/world-asia-india-43551719

[4] Desert (in philosophy) $\approx$ quality of being considered deserving.

in normative targets is also true of supervised machine learning, though perhaps only to a lesser extent.

While we started off observing that group fairness on sensitive attributes has been the mainstay in fair clustering, a few deviant formulations are worthy of mention. Proportionally fair clustering [Chen *et al.*, 2019] proposes an ingenious notion of *collective desert*; it requires that a sufficiently large collective of proximal objects would deserve a cluster of their own. *Representativity Fairness* [P and Abraham, 2020], on the other hand, prefers egalitarian distribution of the *cost of abstraction* incurred due to the clustering.

4 Characterization of Residual Unfairness

Once the normative target is decided, fair clustering formulations translate the target to a mathematical optimization formulation. With even simple clustering formulations being computationally hard [Mahajan *et al.*, 2012], fair clustering will also involve approximations. These might be in the form of theoretical approximation bounds [Chierichetti *et al.*, 2017; Bera *et al.*, 2019] or demonstration of empirical effectiveness [Abraham *et al.*, 2020]. While it is eminently desirable that the chosen target be achieved as much as possible, it is also useful to have an understanding of *how* it falls short when it does indeed fall short; this aspect has not been explored at all to our best knowledge. An important question that one may ask is whether the residual unfairness is *Rawlsian* [Rawls, 1971]; whether it is arranged to the greatest benefit of the least advantaged (ref. *difference principle*). Answers to such questions are crucial for uptake in practical applications since some kinds of systematic unfairness may be considered as intolerable, especially within public sector.

Consider the immensely popular K-Means formulation for clustering [MacQueen, 1967], which some fair clustering formulations build upon (e.g., [Abraham *et al.*, 2020; Ziko *et al.*, 2019]). K-Means clusters may be seen as being located within Voronoi cells centered on the cluster means. Since fair clustering algorithms building upon the K-Means framework are intuitively likely to make the *pro-fairness adjustments* through membership re-assignments at the fringes of clusters, fringe objects would likely bear the *cost/benefit of fairness* more than others. For example, in attributes with a bimodal distribution, say, a mixture of people with no career breaks at all, and long career breaks (e.g., maternity etc.), people with mid-sized career breaks may get reassigned, and could benefit or lose out depending on which side of the line they fall. Consider another example of a data pre-processing method for fair clustering; the fairlet clustering method [Chierichetti *et al.*, 2017], in a gender-balanced dataset, would create fairlets as pairs, each pair comprising one from each gender (assuming binary genders for narrative simplicity only). Data objects that do not have an object of the other gender in its vicinity would stand to lose out due to being paired with a far-off object with which it bears shallow resemblance. As from the above two cases and their comparative evaluation, the cost of the fairness adjustments are unlikely to be *random* and would be borne asymmetrically across dataset objects. Higher volatility, and thus higher benefits or detriments, would likely be placed on objects that deviate much from the implicit data pattern assumptions made within the clustering formulations. While such qualitative differences of fairness shortfalls would be hard to be done away with, a characterization of the fairness shortfall, through quantitative metrics or exemplars, would be necessary to inspire confidence that fair clustering formulations do not exacerbate secondary biases while alleviating major ones.

5 Application Space Information

The clusters in our job shortlisting scenario, we assumed, would be manually assigned one of three decisions, eventually leading to one of two decisions, shortlist or reject. Once this process is complete, we would obviously only care about whether there is representational parity on sensitive attributes over the *shortlisted* set (being just two sets, this would implicitly be equivalent to ensuring the same for the *rejected* set as well). In other words, the upstream clustering algorithm that tried to enforce representational parity in *each* of the several hundred clusters it generated, was, simply put, addressing a needlessly constrained problem. While clustering algorithms cannot foresee downstream cluster-level decisions, fair clustering formulations could be re-designed to provide interactive fairness guidance. For example, as soon as a cluster is chosen for the *shortlist* decision, the clustering could be re-run on the residual dataset with a different fairness target, that seeking fairness among the clusters *conditional on the choice(s) already made* (this is similar in spirit to the *alternative clustering* task [Bae and Bailey, 2006] at the high-level). Another handling of this would be for a one-shot clustering to produce, along with clusters, dependencies among clusters indicating that certain cluster pairs be assigned the same decision. This would be expected in cases of clusters that deviate from fairness in *different directions*, so this dependency constraint across them would help offset them. Such dependencies could also be envisioned as being one of *must-link* and *cannot-link* inspired by literature on semi-supervised clustering [Basu *et al.*, 2002]. In cases with multiple (discrete/continuous) decision choices in the positive-negative spectrum, fairness considerations may be higher in certain parts of the decision space than others. For example, we may want to ensure that the set of failed candidates in a course not be very homogeneous on gender or race, whereas these may be more relaxed at the higher grades, viz., 90% of first graders being of a particular ethnicity may be more tolerable than when 90% of fails come from the same ethnic background. In short, information on clustering usage would go a long way in providing computational leeway in the pursuit of the chosen fairness targets for the clustering method.

6 Concluding Notes

The above discussion was intended towards unravelling the diverse and inter-disciplinary possibilities in extending the scholarly frontier in fair clustering. We hope that researchers with interests in fair clustering would take note of such myriad research frontiers and diversify fair clustering research, an important task for data-driven decision making for the future.

References

[Abraham *et al.*, 2020] Savitha Sam Abraham, Deepak P, and Sowmya S. Sundaram. Fairness in clustering with multiple sensitive attributes. In *EDBT*, pages 287–298, 2020.

[Bae and Bailey, 2006] Eric Bae and James Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 53–62. IEEE, 2006.

[Barocas *et al.*, 2017] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

[Basu *et al.*, 2002] Sugato Basu, Arindam Banerjee, and Raymond Mooney. Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. Citeseer, 2002.

[Bera *et al.*, 2019] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems 32*, pages 4954–4965. 2019.

[Binns, 2018] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159, 2018.

[Chen *et al.*, 2019] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, volume 97, pages 1032–1041, 2019.

[Chierichetti *et al.*, 2017] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037, 2017.

[Deepak and Deshpande, 2015] P Deepak and Prasad M Deshpande. *Operators for similarity search: Semantics, techniques and usage scenarios*. Springer, 2015.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Dworkin, 2002] Ronald Dworkin. *Sovereign virtue: The theory and practice of equality*. Harvard university press, 2002.

[Lang, 2009] Gerald Lang. Luck egalitarianism, permissible inequalities, and moral hazard. *Journal of moral philosophy*, 6(3):317–338, 2009.

[MacQueen, 1967] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[Mahajan *et al.*, 2012] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. *Theoretical Computer Science*, 442:13–21, 2012.

[Nozick, 1974] Robert Nozick. *Anarchy, state, and utopia*, volume 5038. New York: Basic Books, 1974.

[O'neil, 2016] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[P and Abraham, 2020] Deepak P and Savitha Sam Abraham. Representativity fairness in clustering. In *ACM Web Science*, 2020.

[Rawls, 1971] John Rawls. *A theory of justice*. Harvard university press, 1971.

[Ziko *et al.*, 2019] Imtiaz Masud Ziko, Eric Granger, Jing Yuan, and Ismail Ben Ayed. Clustering with fairness constraints: A flexible and scalable approach. *arXiv preprint arXiv:1906.08207*, 2019.