Fair Outlier Detection

Deepak P[1,2] and Savitha Sam Abraham[2]

[1] Queen's University Belfast, UK
[2] Indian Institute of Technology Madras, India
deepaksp@acm.org   savithas@cse.iitm.ac.in

Abstract. An outlier detection method may be considered fair over specified sensitive attributes if the results of outlier detection are not skewed towards particular groups defined on such sensitive attributes. In this task, we consider, for the first time to our best knowledge, the task of fair outlier detection. In this work, we consider the task of fair outlier detection over multiple multi-valued sensitive attributes (e.g., gender, race, religion, nationality, marital status etc.). We propose a fair outlier detection method, *FairLOF*, that is inspired by the popular *LOF* formulation for neighborhood-based outlier detection. We outline ways in which unfairness could be induced within *LOF* and develop three heuristic principles to enhance fairness, which form the basis of the *FairLOF* method. Being a novel task, we develop an evaluation framework for fair outlier detection, and use that to benchmark *FairLOF* on quality and fairness of results. Through an extensive empirical evaluation over real-world datasets, we illustrate that *FairLOF* is able to achieve significant improvements in fairness at sometimes marginal degradations on result quality as measured against the fairness-agnostic *LOF* method.

1 Introduction

There has been much recent interest in incorporating fairness constructs into data analytics algorithms, within the broader theme of algorithmic fairness [12]. The importance of fairness in particular, and democratic values in general, cannot be overemphasized in this age when data science algorithms are being used in very diverse scenarios to aid decision making that could affect lives significantly. The vast majority of fair machine learning work has focused on supervised learning, especially on classification (e.g., [17,30]). There has also been some recent interest in ensuring fairness within unsupervised learning tasks such as clustering [1], retrieval [31] and recommendations [23]. In this paper, we explore the task of fairness in outlier detection, an analytics task of wide applicability in myriad scenarios. To our best knowledge, this is the first work on embedding fairness within outlier detection. The only work so far on fairness and outlier detection [13] focuses on analyzing outlier detection algorithms for fairness, a significantly different task, and proposes a method involving manual intervention.

Outlier Detection and Fairness: The task of outlier detection targets to identify *deviant* observations from a dataset, and is usually modelled as an unsupervised task; [8] provides a broad review of outlier detection methods. The

classical characterization, due to Hawkins [16], defines an outlier as 'an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different process'. Applications of outlier detection range across varied application domains such as network intrusions [18], financial fraud [24] and medical abnormalities [21]. Identification of non-mainstream behavior, the high-level task that outlier detection accomplishes, has a number of applications in new age data scenarios. Immigration checks at airports might want to carry out detailed checks on *'suspicious'* people, while AI is likely used in proactive policing to identify *'suspicious'* people for stop-and-frisk checks. In this age of pervasive digitization, *'abnormality'* in health, income or mobility patterns may invite proactive checks from healthcare, taxation and policing agencies. Identification of such *abnormal* and *suspicious* patterns are inevitably within the remit of outlier detection. The nature of the task of outlier detection task makes it very critical when viewed from the perspective of fairness. Even if information about ethnicity, gender, religion and nationality be hidden (they are often not hidden, and neither is it required to be hidden under most legal regulations) from the database prior to outlier identification, information about these attributes are likely spread across other attributes. For example, geo-location, income and choice of professions may be correlated with ethnic, gender, religious and other identities. The identification of non-mainstream either falls out from, or entails, an analogous and implicit modelling of mainstream characteristics in the dataset. The mainstream behavior, by its very design, risks being correlated with majoritarian identities, leading to the possibility of minority groups being picked out as outliers significantly more often. Interestingly, there have been allegations of racial discrimination in many such settings[3].

Outlier Detection and the Web: Web has emerged, over the past decades, as a rich source of unlabelled digital data. Thus, the web likely presents the largest set of scenarios involving outlier detection. Each user on the web leaves different cross-sections of digital footprints in different services she uses, together encompassing virtually every realm of activity; this goes well beyond the public sector applications referenced above. In a number of scenarios, identified as an outlier could lead to undesirable outcomes for individuals. For example, mobility outliers may receive a higher car insurance quote, and social media outliers may be subjected to higher scrutiny (e.g., Facebook moderation). It is important to ensure that such undesirable outcomes be distributed fairly across groups defined on protected attributes (e.g., gender, race, nationality, religion etc.) for ethical reasons and to avoid bad press[4].

Our Contributions: We now outline our contributions in this paper. First, we characterize the task of fair outlier detection under the normative principle of disparate impact avoidance [4] that has recently been used in other unsupervised learning tasks [11,1]. Second, we develop a fair outlier detection method, *FairLOF*, based on the framework of LOF [7], arguably the most popular outlier

---

[3] https://www.nyclu.org/en/stop-and-frisk-data
[4] https://www.cnet.com/features/is-facebook-censoring-conservatives-or-is-moderating-just-too-hard/

detection method. Our method is capable of handling multiple multi-valued protected attributes, making it feasible to use in sophisticated real-world scenarios where fairness is required over a number of facets. Third, we outline an evaluation framework for fair outlier detection methods, outlining quality and fairness metrics, and trade-offs among them. Lastly, through an extensive empirical evaluation over real-world datasets, we establish the effectiveness of *FairLOF* in achieving high levels of fairness at small degradations to outlier detection quality.

2   Related Work

Given the absence of prior work on fair outlier detection methods, we cover related work across outlier detection and fairness in unsupervised learning.

Outlier Detection Methods: Since obtaining labelled data containing outliers is often hard, outlier detection is typically modelled as an unsupervised learning task where an unlabelled dataset is analyzed to identify outliers within it. That said, supervised and semi-supervised approaches do exist [8]. We address the unsupervised setting. The large majority of work in unsupervised outlier detection may be classified into one of two families. The first family, that of *global methods*, build a dataset-level model, and regard objects that do not conform well to the model, as outliers. The model could be a clustering [29], Dirichlet mixture [15] or others [14]. Recent research has also explored the usage of auto-encoders as a global model, the reconstruction error of individual data objects serving as an indication of their outlierness; *RandNet* [10] generalizes this notion to determine outliers using an ensemble of auto-encoders. The second family, arguably the more popular one, is that of *local methods*, where each data object's outlierness is determined using just its neighborhood within a relevant similarity space, which may form just a small subset of the whole dataset. The basic idea is that the outliers will have a local neighborhood that differs sharply in terms of characteristics from the extended neighborhood just beyond. LOF [7] operationalizes this notion by quantifying the contrast between an object's local density (called local reachability density, as we will see) and that of other objects in its neighborhood. Since LOF, there has been much research into local outliers, leading to work such as SLOM [9], LoOP [20] and LDOF [32]. Schubert et al [27] provide an excellent review of local outlier detection, including a generalized three phase meta-algorithm that most local outlier detection methods can be seen to fit in. Despite advancements over the last two decades, LOF remains the dominant method for outlier detection, continuously inspiring systems work on making it efficient for usage in real-world settings (e.g., [3]). Accordingly, the framework of LOF inspires the construction of *FairLOF*.

Fairness in Unsupervised Learning: There has been much recent work on developing fair algorithms for unsupervised learning tasks such as clustering, representation learning and retrieval. Two streams of fairness are broadly used; *group fairness* that targets to ensure that the outputs are fairly distributed across groups defined on sensitive attributes, and *individual fairness* which strives to limit possibilities of similar objects receiving dissimilar outcomes. Individual

fairness is agnostic to the notion of sensitive attributes. Our focus, in this paper, is on group fairness in outlier detection. For group fairness in clustering, techniques differ in where they embed the fairness constructs; it could be at the pre-processing step [11], within the optimization framework [1] or as a post-processing step to re-configure the outputs [6]. FairPCA [22], a fair representation learner, targets to ensure that objects are indistinguishable wrt their sensitive attribute values in the learnt space. Fair retrieval methods often implement group fairness as parity across sensitive groups in the top-k outputs [2]. The techniques above also differ in another critical dimension; the number of sensitive attributes they can accommodate. Some can only accommodate one binary sensitive attribute, whereas others target to cater to fairness over multiple multi-valued sensitive attributes; a categorization of clustering methods along these lines appears in [1].

In contrast to such work above, there has been no work so far in fair outlier detection. The only related effort in this space so far, to our best knowledge, is that of developing a human-in-the-loop decision procedure to determine whether the outputs of an outlier detection is fair [13]. This focuses on deriving *explanations based on sensitive attributes* to distinguish the outputs of an outlier detection method from the *'normal'* group. If no satisfactory explanation can be achieved, the black-box outlier detection method can be considered fair. The human is expected to have domain knowledge of the task and data scenario to determine parameters to identify what is unfair, and interpret explanations to judge whether it is indeed a case of unfairness.

3   Problem Definition

Task Setting: Consider a dataset $\mathcal{X} = \{\ldots, X, \ldots\}$ and an object pairwise distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is deemed relevant to the outlier detection scenario. Further, each data object is associated with a set of sensitive attributes $\mathcal{S} = \{\ldots, S, \ldots\}$ (e.g., gender, race, nationality, religion etc.) which are categorical and potentially multi-valued, $V(S)$ being the set of values that a sensitive attribute, S, can take. $X.S \in V(S)$ indicates the value assumed by object X for the sensitive attribute S. Thus, each multi-valued attribute S defines a partitioning of the dataset into $|V(S)|$ parts, each of which comprise objects that take the same distinct value for S.

Outlier Detection: The task of (vanilla) outlier detection is that of identifying a small subset of objects from $\mathcal{X}$, denoted as $\mathcal{O}$, that are deemed to be outliers. Within the *local outliers* definition we adhere to, it is expected that objects in $\mathcal{O}$ differ significantly in local neighborhood density when compared to other objects in their extended neighborhood. In typical scenarios, it is also expected that $|\mathcal{O}| = t$, where t is a pre-specified parameter. The choice of t may be both influenced by the dataset size (e.g., t as a fixed fraction of $|\mathcal{X}|$) and guided by practical considerations (e.g., manual labour budgeted to examine outliers).

Fair Outlier Detection: The task of fair outlier detection, in addition to identifying outliers, considers ensuring that the distribution of sensitive attribute

groups among $\mathcal{O}$ reflect that in $\mathcal{X}$ as much as possible. This notion, referred to interchangeably as representational parity or disparate impact avoidance, has been the cornerstone of all major fair clustering algorithms (e.g., [11,1,6]), and is thus a natural first choice as a normative principle for fair outlier detection. As a concrete example, if gender is a sensitive attribute in $\mathcal{S}$, we would expect the gender ratio within $\mathcal{O}$ to be very close to, if not exactly equal to, the gender ratio in $\mathcal{X}$. Note that fairness is complementary and often contradictory to ensuring that the top neighborhood-outliers find their place in $\mathcal{O}$; the latter being the only consideration in (vanilla) outlier detection. Thus, fair outlier detection methods such as *FairLOF* we develop, much like fair clustering methods, would be evaluated on two sets of metrics: (a) *'Quality'* metrics that measure how well objects with distinct local neighborhoods are placed in $\mathcal{O}$, and (b) *Fairness* metrics that measure how well they ensure that the dataset-distribution of sensitive attribute values are preserved within $\mathcal{O}$. We will outline a detailed evaluation framework in a subsequent section. Good fair outlier detection methods would be expected to achieve good fairness while suffering only small degradations in quality when compared against their vanilla outlier detection counterparts.

Motivation for Representational Parity: It may be argued that the distribution of sensitive attribute groups could be legitimately different from that in the dataset. For example, one might argue that outlying social media profiles that correlate with crime may be legitimately skewed towards certain ethnicities since propensity for crimes could be higher for certain ethnicities than others. The notion of representational parity disregards such assumptions of skewed apriori distributions, and seeks to ensure that the inconvenience of being classed as an outlier be shared proportionally across sensitive attribute groups. This argument is compelling within scenarios of using outlier detection in databases encompassing information about humans. In particular, this has its roots in the distributive justice theory of *luck egalitarianism* [19] that distributive shares be not influenced by arbitrary factors, especially those of *'brute luck'* that manifest as membership in sensitive attribute groups (since individuals do not choose their gender, ethnicity etc.). The normative principle has been placed within the umbrella of the *'justice as fairness'* work due to John Rawls [26] that underlies most of modern political philosophy. Further, since outlier detection systems are often used to inform human decisions, it is important to ensure that outlier detection algorithms do not propagate and/or reinforce stereotypes present in society by way of placing higher burden on certain sensitive groups than others.

4 Background: Local Outlier Factor (LOF)

Our method builds upon the pioneering LOF framework [7] for (vanilla) outlier detection. LOF comprises three phases, each computing a value for each object in $\mathcal{X}$, progressively leading to LOF: (i) k-distance, (ii) local reachability density (LRD), and (iii) local outlier factor (LOF).

k-distance: Let $N_k(X)$ be the set of k nearest neighbors[5] to X (within $\mathcal{X}$), when assessed using the distance function $d(.,.)$. The k-distance for each $X \in \mathcal{X}$ is then the distance to the kth nearest object.

$$k\text{-}distance(X) = \underset{X' \in N_k(X)}{\arg\max}\ d(X, X') \tag{1}$$

Local Reachability Density: The local reachability density of X is defined as the inverse of the average distance of X to it's k nearest neighbors:

$$lrd(X) = 1 / \left(\frac{\sum_{X' \in N_k(X)} rd(X, X')}{|N_k(X)|} \right) \tag{2}$$

where $rd(X, X')$ is an assymetric distance measure that works out to the true distance, except when the true distance is smaller than $k\text{-}distance(X')$:

$$rd(X, X') = max\{k\text{-}distance(X'), d(X, X')\} \tag{3}$$

This lower bounding by $k\text{-}distance(X')$ - note that $k\text{-}distance(X')$ depends on $N_k(X')$ and not $N_k(X)$ - makes the $lrd(.)$ measure more stable. $lrd(X)$ quantifies the density of the local neighborhood around X.

Local Outlier Factor: The local outlier factor is the ratio of the average lrds of X's neighbors to X's own lrd.

$$lof(X) = \left(\frac{\sum_{X' \in N_k(X)} lrd(X')}{|N_k(X)|} \right) / lrd(X) \tag{4}$$

An $lof(X) = 1$ indicates that the local density around X is comparable to that of it's neighbors, whereas a $lrd(X) >> 1$ indicates that it's neighbors are in much denser regions than itself. Once $lof(.)$ is computed for each $X \in \mathcal{X}$, the top-t data objects with highest $lof(.)$ scores would be returned as outliers.

5   FairLOF: Our Method

5.1   Motivation

In many cases, the similarity space implicitly defined by the distance function $d(.,.)$ bears influences from the sensitive attributes and groups defined over them. The influence, whether casual, inadvertent or conscious, could cause the sensitive attribute profiles of outliers to be significantly different from the dataset profiles. These could occur in two contrasting ways as outlined below.

Under-reporting of Large Sensitive Groups: Consider the case where $d(.,.)$ is aligned with groups defined by S. Thus, across the dataset, pairs of objects that share the same value for S are likely to be judged to be more proximal than those that bear different values for S. Consider an example scenario of a dataset comprising 75% males and 25% females. Such skew could occur in real-world

---

[5] $|N_k(X)|$ could be greater than k in case there is a tie for the kth place.

cases such as datasets sourced from populations in a STEM-oriented college or in certain professions (e.g., police force[6]). Let us consider the base/null assumption that *real outliers* are also distributed as 75% males and 25% females. Now, consider a male outlier (M) and a female outlier (F), both of which are equally eligible outliers according to human judgement. Consider the case of M; M is likely to have a *quite cohesive* and *predominantly male k*NN neighborhood due to both: (i) males being more likely in the dataset due to the apriori distribution, and (ii) $d(.,.)$ likely to judge males as more similar to each other (our starting assumption). Note that the first factor above does not apply to F; thus, F's neighborhood would be both less cohesive and less gender homogeneous. This would translate to $lrd(M) > lrd(F)$, and thus $lof(M) < lof(F)$ (Ref. Eq. 4) despite them being both equally eligible outliers. In short, *when $d(.,.)$ is aligned with groups defined over S, the smaller groups would tend to be over-represented among the outliers and vice versa.*

Over-reporting of Large Sensitive Groups: Consider a domain-tuned distance function designed for a health records agency who would like to ensure that records be not judged similar just due to similarity on gender; such fine-tuning, as is often done with the intent of ensuring fairness, might often be designed with just the *'main groups'* in mind. In the case of gender, this would ensure a good spread of male and female records within the space; however, this could result in minority groups (e.g., LGBTQ) being relegated to a corner of the similarity space. This would result in a tight clustering of records belonging to the minority group, resulting in the LOF framework being unable to pick them out as outliers. Thus, *a majority conscious design of $d(.,.)$ would result in over-representation of minority groups among outliers and vice versa.*

5.2   FairLOF: The Method

The construction of *FairLOF* attempts to *correct* for such kNN neighbohood distance disparities across object groups defined over sensitive attributes. *FairLOF* distance correction is based on three principles; (i) **neighborhood diversity** (object-level correction), (ii) **apriori distribution** (value-level), and (iii) **attribute asymmetry** (attribute-level). We outline these for the first scenario in Sec 5.1, where $d(.,.)$ is aligned with the sensitive attribute, S, resulting in *minority over-representation among outliers*; these will be later extended to the analogous scenario, as well as for multiple attributes in $\mathcal{S}$.

Neighborhood Diversity: Consider the case of objects that are embedded in neighborhoods comprising objects that take different values of S than itself; we call this as a S-diverse neighborhood. These would be disadvantaged with a higher k-*distance*, given our assumption that $d(.,.)$ is aligned with S. Thus, the k-*distance* of objects with highly diverse neighborhoods would need to be corrected *downward*. This is an object-specific correction, with the extent of the correction determined based on S-diversity in the object's neighborhood.

---

[6] https://www.statista.com/statistics/382525/share-of-police-officers-in-england-and-wales-gender-rank/

Apriori Distribution: Consider objects that belong to an S group that are very much in minority; e.g., LGBTQ groups for $S = gender$. Since these objects would have an extremely diverse neighborhood due to their low apriori distribution in the dataset (there aren't enough objects with the same S value in the dataset), the neighborhood diversity principle would correct them deeply downward. To alleviate this, the neighborhood diversity correction would need to be discounted based on the sparsity of the object's value of S in the dataset.

Attribute Asymmetry: The extent of $k\text{-}distance$ correction required also intuitively depends on the extent to which $d(.,.)$ is aligned with the given S. This could be directly estimated based on the extent of minority over-representation among outliers when vanilla LOF is applied. Accordingly, the attribute asymmetry principle requires that the correction based on the above be amplified or attenuated based on the extent of correction warranted for S.

The above principles lead us to the following form for $k\text{-}distance$:

$$\left(\underset{X' \in N_k(X)}{\arg\max}\ d(X,X')\right)\left(1 - \lambda \times W_S^{\mathcal{X}} \times D_{X.S}^{\mathcal{X}} \times Div(N_k(X), X.S)\right) \quad (5)$$

where $Div(N_k(X), X.S)$, $D_{X.S}^{\mathcal{X}}$ and $W_S^{\mathcal{X}}$ relate to the three principles outlined above (respectively), $\lambda \in [0,1]$ being a weighting factor. These terms are constructed as below:

$$Div(N_k(X), X.S) = \frac{|\{X'|X' \in N_k(X) \wedge X'.S \neq X.S|}{|N_k(X)|} \quad (6)$$

$$D_{X.S}^{\mathcal{X}} = \frac{|\{X'|X' \in \mathcal{X} \wedge X'.S = X.S\}|}{|\mathcal{X}|} \quad (7)$$

$$W_S^{\mathcal{X}} = c + |D_{v^*}^{\mathcal{X}} - D_{v^*}^{\mathcal{R}_{LOF}}| \quad \text{where} \quad v^* = \underset{v \in V(S)}{\arg\max}\ D_v^{\mathcal{X}} \quad (8)$$

Eq. 6 measures diversity as the fraction of objects among $N_k(X)$ that differ from X on it's S attribute value. Eq. 7 measures the apriori distribution as the fraction of objects in $\mathcal{X}$ that share the same S attribute value as that of X. For Eq. 8, $D_v^{\mathcal{R}_{LOF}}$ refers to the fraction of $S = v$ objects found among the top-t results of vanilla LOF over $\mathcal{X}$. Thus, $W_S^{\mathcal{X}}$ is a constant factor (i.e., c) added to the asymmetry extent measured as the extent to which the largest S-defined group in the dataset is underrepresented in the vanilla LOF results. While we have used a single S attribute so far, observe that this is easily extensible to multiple attributes in $\mathcal{S}$, yielding the following refined form for $k\text{-}distance$:

$$\left(\underset{X' \in N_k(X)}{\arg\max}\ d(X,X')\right)\left(1 - \lambda \sum_{S \in \mathcal{S}}\left(W_S^{\mathcal{X}} \times D_{X.S}^{\mathcal{X}} \times Div(N_k(X), X.S)\right)\right) \quad (9)$$

While we have been assuming the case of $d(.,.)$ aligned with S and minority over-representation among outliers, the opposite may be true for certain attributes in $\mathcal{S}$; recollect the second case discussed in Section 5.1. In such cases, the $k\text{-}distance$ correction would need to be performed upward, as against downward. We incorporate that to yield the final $k\text{-}distance$ formulation for $FairLOF$.

$$k\text{-}distance_{FairLOF}(X) =$$
$$\left(\underset{X' \in N_k(X)}{\arg\max} \ d(X, X') \right) \left(1 - \lambda \sum_{S \in \mathcal{S}} \left(\mathbb{D}(\mathcal{X}, S) \times W_S^{\mathcal{X}} \times D_{X.S}^{\mathcal{X}} \times Div(N_k(X), X.S) \right) \right)$$

$$(10)$$

where $\mathbb{D}(\mathcal{X}, S) \in \{-1, +1\}$ denotes the direction of correction as below:

$$\mathbb{D}(\mathcal{X}, S) = \begin{cases} +1 & \text{if } D_{v^*}^{\mathcal{X}} > D_{v^*}^{\mathcal{R}_{LOF}} \quad \text{where} \quad v^* = \arg\max_{v \in V(S)} \ D_v^{\mathcal{X}} \\ -1 & otherwise \end{cases} \quad (11)$$

This modified formulation of $k\text{-}distance$ warrants a similar correction in the construction of $rd(.,.)$ to put the pairwise distance on level ground with the modified $k\text{-}distance$ for the max comparison.

$$rd_{FairLOF}(X, X') = max\Big\{ k\text{-}distance_{FairLOF}(X'),$$
$$d(X, X') \times \Big(1 - \lambda \sum_{S \in \mathcal{S}} \big(\mathbb{D}(\mathcal{X}, S) \times W_S^{\mathcal{X}} \times D_{X.S}^{\mathcal{X}} \times \mathbb{I}(X.S \neq X'.S) \big) \Big) \Big\} \quad (12)$$

The second term in $rd(.,.)$ is corrected in the same manner as for $k\text{-}distance$, except that the diversity term is replaced by a simple check for inequality, given that there is only one object that X is compared with.

These distance corrections complete the description of $FairLOF$, which is the LOF framework from Section 4 with $k\text{-}distance(.,.)$ and $rd(.,.)$ replaced by their corrected versions from Eq. 10 and Eq. 12 respectively. While we omit the whole sequence of $FairLOF$ steps to avoid repetition with Sec. 4, we will use $flof(.)$ to denote the final outlier score from $FairLOF$, analogous to $lof(.)$ for LOF. The $FairLOF$ hyperparameter, λ, determines the *strength* of the fairness correction applied, and could be a very useful tool to navigate the space of options $FairLOF$ provides, as we will outline in the next section.

Note on Complexity: Eq. 7 and Eq. 8 can be pre-computed at a an exceedingly small cost of $\mathcal{O}(|\mathcal{S}| \times m)$ where m is $max_{S \in \mathcal{S}}|V(S)|$. Eq. 6 needs to be computed at a per-object level, thus multiplying the LOF complexity by $|\mathcal{S}| \times m$. With typical values of $|\mathcal{S}| \times m$ being in the 1000s at max (e.g., 3-5 sensitive attributes with 10-20 distinct values each), and outlier detection being typically considered as an offline task not requiring real-time responses, the overheads of the $k\text{-}distance$ adjustment may be considered as insignificant.

6   Evaluation Framework for Fair Outlier Detection

Enforcing parity along S-groups among outliers, as discussed, often contradicts with identifying high-LOF outliers. This trade-off entails two sets of evaluation measures, inspired by similar settings in fair clustering [1].

Quality Evaluation: While the most desirable quality test for any outlier detection framework would be accuracy measured against human generated outlier/non-outlier labels, public datasets with such labels are not available, and far from feasible to generate. Thus, we will use the assumption that LOF, being a hugely popular outlier detection technique, may be considered as gold standard. Based on that, we devise the following measures:

$$Jacc(\mathcal{R}_{LOF}, \mathcal{R}_{FairLOF}) = \frac{|\mathcal{R}_{LOF} \cap \mathcal{R}_{FairLOF}|}{|\mathcal{R}_{LOF} \cup \mathcal{R}_{FairLOF}|} \tag{13}$$

$$Pres(\mathcal{R}_{LOF}, \mathcal{R}_{FairLOF}) = \frac{\sum_{X \in \mathcal{R}_{FairLOF}} lof(X)}{\sum_{X \in \mathcal{R}_{LOF}} lof(X)} \tag{14}$$

where $\mathcal{R}_{LOF}$ and $\mathcal{R}_{FairLOF}$ are top-t outliers (for any chosen k) from LOF and $FairLOF$ respectively. $Jacc(.,.)$ computes the jaccard similarity between the result sets. Even in cases where $\mathcal{R}_{FairLOF}$ diverges from $\mathcal{R}_{LOF}$, we would like to ensure that it does not choose objects with very low $lof(.)$ values within $\mathcal{R}_{FairLOF}$; $Pres(.,.)$ computes the extent to which high $lof(.)$ scores are preserved within $\mathcal{R}_{FairLOF}$, expressed as a fraction of the total $lof(.)$ across $\mathcal{R}_{LOF}$. For both these, higher values indicate better quality of $FairLOF$ results.

Fairness Evaluation: For any particular sensitive attribute $S \in \mathcal{S}$, we would like the distribution of objects across its values among outliers (i.e., $\mathcal{R}_{FairLOF}$) be similar to that in the dataset, $\mathcal{X}$. In other words, we would like the $\mathcal{D}_S^{\mathcal{R}_{FairLOF}} = [\ldots, D_v^{\mathcal{R}_{FairLOF}}, \ldots]$ vector ($V \in V(S)$, and Ref. Eq. 7 for computation) to be as similar as possible to the distribution vector over the dataset for S, i.e., $\mathcal{D}_S^{\mathcal{X}} = [\ldots, D_v^{\mathcal{X}}, \ldots]$, as possible. We would like this to hold across all attributes in $\mathcal{S}$. Note that this fairness notion is very similar to that in fair clustering, the only difference being that we evaluate the outlier set as against each cluster. Thus, we adapt the fairness metrics from [1,28] as below:

$$ED(\mathcal{R}_{FairLOF}) = \sum_{S \in \mathcal{S}} Euclidean\_Distance(\mathcal{D}_S^{\mathcal{R}_{FairLOF}}, \mathcal{D}_S^{\mathcal{X}}) \tag{15}$$

$$Wass(\mathcal{R}_{FairLOF}) = \sum_{S \in \mathcal{S}} Wasserstein\_Distance(\mathcal{D}_S^{\mathcal{R}_{FairLOF}}, \mathcal{D}_S^{\mathcal{X}}) \tag{16}$$

where $ED(.)$ and $Wass(.)$ denote aggregated Euclidean and Wasserstein distances across the respective distribution vectors. Since these measure deviations from dataset-level profiles, lower values are desirable in the interest of fairness.

Quality-Fairness Trade-off: It may be noted that all the above metrics can be computed without any external labellings. Thus, this provides an opportunity for the user to choose different trade-offs between *quality* and *fairness* by varying the *FairLOF correction strength* hyper-parameter λ. We suggest that a practical way of using *FairLOF* would be for a user to try with progressively higher values of λ from $\{0.1, 0.2, \ldots\}$ (note that *FairLOF* becomes equivalent to *LOF* at $\lambda = 0$) using a desired value of *Jaccard* similarity as a pilot measure. For example, we may want to retain a Jaccard similarity (i.e., $Jacc(.)$ value) of approximately 0.9 or 0.8 with the original *LOF* results. Thus, the user may stop when that is achieved. The quantum of fairness improvements achieved by *FairLOF* over *LOF* at such chosen points will then be indicative of *FairLOF*'s effectiveness.

Single Sensitive Attribute and a Quota-based System: As noted upfront, *FairLOF* is targeted towards cases where there are multiple sensitive attributes to ensure fairness over; this is usually the case since there are often many sensitive attributes in real world scenarios (e.g., gender, ethnicity, caste, religion, native language, marital status, and even age in certain scenarios). For the case of a single sensitive attribute with a handful of possible values, there is a simple and effective strategy for fair outlier detection. Consider $S = gender$; we take the global list of objects sorted in descending order of $lof(.)$ scores and splice them into *male* list, *female* list etc. Based on the desired distribution of gender groups among outliers (as estimated from the dataset), *'quotas'* may be set for each gender value, and the appropriate number of top objects from each gender-specific list is then put together to form the outlier set of t objects. This is similar to the strategy used for job selection in India's affirmative action policy (*aka* reservation[7]). The extension of this quota-based strategy to multiple sensitive attributes by modelling them as one giant attribute taking values from the cross-product, is impractical due to multiple reasons. First, the cross-product may easily exceed t, leading to practical and legal issues across scenarios; for example, with just $\mathcal{S} = \{nationality, ethnicity\}$, we could have the cross product approaching 2000+ given there are 200+ nationalities and at least 10+ ethnicities, and practical values of t could be in the 100s due to manual perusal considerations. Thus, the quota system, by design, would exclude the large majority of rare combinations of sensitive attribute values from being represented among outliers however high their LOF scores may be; such a policy is unlikely to survive any legal or ethical scrutiny to allow practical uptake. Second, the quota based system offers no way to control the trade-off between fairness and quality, making it impractical to carefully choose trade-offs as outlined earlier. Third, even a simpler version of the extension of the quota-based system to multiple attributes has been recently shown to be NP-hard [5].

---

[7] https://en.wikipedia.org/wiki/Reservation_in_India

[8] http://archive.ics.uci.edu/ml/datasets/Adult

[9] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[10] https://archive.ics.uci.edu/ml/datasets/wiki4he

[11] https://archive.ics.uci.edu/ml/datasets/Student+Performance

Dataset	Domain	$|\mathcal{X}|$	Sensitive Attributes Used
Adult[8]	US 1994 Census	48842	*marital status, race, sex, nationality*
CC[9]	Credit Card Default	30000	*sex, education, marital status*
W4HE[10]	Wikipedia HE Use	913	*gender, disciplinary domain, uni name*
St-Mat[11]	Student Maths Records	649	*gender, age*

Table 1. Dataset Information (Note: The table footnotes may be in adjacent pages)

7   Experimental Evaluation

Datasets and Experimental Setup: There are only a few public datasets with information of people, the scenario that is most pertinent for fairness analysis; this is likely due to person-data being regarded highly personal and anonymization could still lead to leakage of identifiable information[12]. The datasets we use along with details are included in Table 6. The datasets encompass a wide variety of scenarios, and vary much in sizes as well as the sensitive attributes used. We set t (to get top-t results) to 5% of the dataset size capped at 500, $k = 5$ and c (Ref. Eq. 8) to be $1/s$ where s is the number of sensitive attributes.

Dataset	Guide Point	Quality				Fairness			
		Jacc	*Det%*	*Pres*	*Det%*	*ED*	*Impr%*	*Wass*	*Impr%*
Adult	LOF	1.0	N/A	1.0	N/A	0.2877	N/A	0.5328	N/A
	0.9	0.8939	10.61%	0.9977	**00.23%**	0.1906	**33.75%**	0.3468	**34.91%**
	0.8	0.7986	20.14%	0.9906	**00.94%**	0.1714	**40.42%**	0.2372	**55.48%**
CC	LOF	1.0	N/A	1.0	N/A	0.2670	N/A	0.2152	N/A
	0.9	0.9011	09.89%	0.9976	**00.24%**	0.2235	16.29%	0.2112	01.86%
	0.8	0.7921	20.79%	0.9879	01.21%	0.1568	**41.27%**	0.2012	06.51%
W4HE	LOF	1.0	N/A	1.0	N/A	0.2121	N/A	0.3305	N/A
	0.9	0.8776	12.24%	0.9987	**00.13%**	0.0966	**54.46%**	0.2989	09.56%
	0.8	0.8039	19.61%	0.9951	**00.49%**	0.1820	14.19%	0.2498	*24.42%*
St-Mat	LOF	1.0	N/A	1.0	N/A	0.4174	N/A	0.9196	N/A
	0.9	0.9047	09.53%	0.9970	**00.30%**	0.3467	16.94%	0.8962	02.54%
	0.8	0.8182	18.18%	0.9896	01.04%	0.3467	16.94%	0.8962	02.54%

Table 2. FairLOF Effectiveness Study. The *FairLOF* results at *guide points* set to $Jacc = 0.9$ and $Jacc = 0.8$ are shown along with LOF results for each of the datasets. Since we use coarse steps for λ, the precise guide point value for *Jacc* may not be achieved; so we choose the closest *Jacc* that is achievable to the guide point. The deteriorations in Quality metrics and improvements in Fairness metrics are indicated in percentages. Fairness improvements of 20%+ are italicized, and those that are 30%+ are shown in bold, whereas Quality deteriorations < 1% are indicated in bold.

---

[12] https://en.wikipedia.org/wiki/AOL_search_data_leak

7.1   FairLOF Effectiveness Study

The effectiveness of *FairLOF* may be assessed by considering the quantum of fairness achieved at low degradations to quality. *It may be noted that higher values are better on the quality measures (Jacc and Pres) and lower values are better on the fairness measures (ED and Wass).* We follow the quality-fairness trade-off strategy as outlined in Section 6 with a λ search step-size of 0.1 and choose 0.9 and 0.8 as *guide points* for *Jacc*. The detailed results are presented in Table 7, with details of the result formatting outlined in the caption therein. Broadly, we observe the following:

- **Fairness Improvements:** *FairLOF* is seen to achieve significant improvements in fairness metrics at moderate degradations to quality. The ED measure is being improved by 30% on an average at the chosen guide points, whereas $Wass$ is improved by 12% and 22% on an average at the guide points of 0.9 and 0.8 respectively. These are evidently hugely significant gains translating to corresponding improvements in fairness of outlier detection.
- **Trends on Pres:** Even at *Jacc* close to 0.9 and 0.8, the values of *Pres* achieved by *FairLOF* are seen to be only marginally lower than 1.0, recording degradations of less than 1.0% in the majority of the cases. This indicates that while the *LOF* results are being altered, *FairLOF* is being able to replace them with other objects with substantively similar $lof(.)$ values. This, we believe, is a highly consequential result, indicating that *FairLOF* remains very close in spirit to *LOF* on result quality while achieving the substantive fairness gains.

In addition to the above, we note the following trends on *FairLOF* performance. First, $Wass$ is significantly harder to optimize for, as compared to ED; this is because $Wass$ prefers the gains to be equally distributed across sensitive attributes. Second, for small datasets where there is relatively less room for re-engineering outlier results for fairness, *FairLOF* gains are seen to saturate quickly. This is most evident for *St-Mat* in Table 7.

7.2   FairLOF Parameter Sensitivity Study

One of the key aspects is to see whether *FairLOF* effectiveness is smooth against changes in λ, the only parameter of significant consequence in *FairLOF*. In particular, we desire to see consistent decreases on each of *Jacc*, *Pres*, *ED* and *Wass* with increasing λ. On each of the datasets, such gradual and smooth trends were observed, with the gains tapering off sooner in the case of the smaller datasets, W4HE and St-Mat. The trends on Adult and CC were very similar; for Adult, we observed that the Pearson product-moment correlation co-efficient [25] against $\lambda \in [0, 1]$ to be -0.900 for *Jacc*, -0.973 for *Pres*, -0.997 for *ED* and -0.959 for *Wass* indicating a graceful movement along the various metrics with changing λ. We observed similar consistent trends for increasing c (Eq. 8) as well. *FairLOF* was also observed to be quite stable with small changes of k and t around the chosen values for the experiments in the previous section.

8   Conclusions and Future Work

In this paper, for the first time (to our best knowledge), we considered the task of fair outlier detection. Fairness is of immense importance in this day and age when data analytics in general, and outlier detection in particular, is being used to make and influence decisions that will affect human lives to a significant extent. We consider the paradigm of local neighborhood based outlier detection, arguably the most popular paradigm in outlier detection literature. We outlined the task of fair outlier detection over a plurality of sensitive attributes, basing our argument on the normative notion of luck egalitarianism, that the costs of outlier detection be borne proportionally across groups defined on protected/sensitive attributes such as gender, race, religion and nationality. We observed that using a task-defined distance function for outlier detection could induce unfairness when the distance function is not fully orthogonal to all the sensitive attributes in the dataset. We develop an outlier detection method, called *FairLOF*, inspired by the construction of *LOF* and makes use of three principles to nudge the outlier detection towards directions of increased fairness. We outline an evaluation framework for fair outlier detection, and use that in evaluating *FairLOF* extensively over real-world datasets. Through our empirical results, we observe that *FairLOF* is able to deliver substantively improved fairness in outlier detection results, at very marginal costs to result quality as assessed against *LOF*. This illustrates the effectiveness of *FairLOF* in achieving fairness in outlier detection. **Future Work:** In this work, we have limited our attention to local neighborhood based outlier detection. Extending notions of fairness to global outlier detection would be an interesting future work. Further, we are considering extending *FairLOF* to the related task of identifying groups of anomalous points.

References

1. Abraham, S.S., P, D., Sundaram, S.S.: Fairness in clustering with multiple sensitive attributes. In: EDBT. pp. 287–298 (2020)
2. Asudeh, A., Jagadish, H., Stoyanovich, J., Das, G.: Designing fair ranking schemes. In: SIGMOD (2019)
3. Babaei, K., Chen, Z., Maul, T.: Detecting point outliers using prune-based outlier factor (plof). arXiv preprint arXiv:1911.01654 (2019)
4. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)
5. Bei, X., Liu, S., Poon, C.K., Wang, H.: Candidate selections with proportional fairness constraints. In: AAMAS (2020)
6. Bera, S.K., Chakrabarty, D., Flores, N., Negahbani, M.: Fair algorithms for clustering. In: NeurIPS. pp. 4955–4966 (2019)
7. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: SIGMOD. pp. 93–104 (2000)
8. Chandola, V., Banerjee, A., Kumar, V.: Outlier detection: A survey. ACM Computing Surveys **14**, 15 (2007)
9. Chawla, S., Sun, P.: Slom: a new measure for local spatial outliers. Knowledge and Information Systems **9**(4), 412–429 (2006)

10. Chen, J., Sathe, S., Aggarwal, C., Turaga, D.: Outlier detection with autoencoder ensembles. In: SDM (2017)
11. Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvitskii, S.: Fair clustering through fairlets. In: NIPS (2017)
12. Chouldechova, A., Roth, A.: A snapshot of the frontiers of fairness in machine learning. Commun. ACM **63**(5), 8289 (Apr 2020)
13. Davidson, I., Ravi, S.: A framework for determining the fairness of outlier detection. In: ECAI (2020)
14. Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J.: A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition **74**, 406–421 (2018)
15. Fan, W., Bouguila, N., Ziou, D.: Unsupervised anomaly intrusion detection via localized bayesian feature selection. In: ICDM (2011)
16. Hawkins, D.M.: Identification of outliers, vol. 11. Springer (1980)
17. Huang, L., Vishnoi, N.K.: Stable and fair classification. arXiv:1902.07823 (2019)
18. Jabez, J., Muthukumar, B.: Intrusion detection system (ids): anomaly detection using outlier detection approach. Procedia Computer Science **48**, 338–346 (2015)
19. Knight, C.: Luck egalitarianism: Equality, responsibility, and justice. EUP (2009)
20. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Loop: local outlier probabilities. In: CIKM (2009)
21. Kumar, V., Kumar, D., Singh, R.: Outlier mining in medical databases: an application of data mining in health care management to detect abnormal values presented in medical databases. IJCSNS International Journal of Computer Science and Network Security pp. 272–277 (2008)
22. Olfat, M., Aswani, A.: Convex formulations for fair principal component analysis. In: AAAI. vol. 33, pp. 663–670 (2019)
23. Patro, G.K., et al.: Incremental fairness in two-sided market platforms: On updating recommendations fairly. In: AAAI (2020)
24. Pawar, A.D., Kalavadekar, P.N., Tambe, S.N.: A survey on outlier detection techniques for credit card fraud detection. IOSR J. Comp. Engg. **16**(2), 44–48 (2014)
25. Pearson, K.: Vii. note on regression and inheritance in the case of two parents. proceedings of the royal society of London **58**(347-352), 240–242 (1895)
26. Rawls, J.: A theory of justice. Harvard university press (1971)
27. Schubert, E., Zimek, A., Kriegel, H.P.: Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data Mining and Knowledge Discovery **28**(1), 190–237 (2014)
28. Wang, B., Davidson, I.: Towards fair deep clustering with multi-state protected variables. arXiv preprint arXiv:1901.10053 (2019)
29. Yu, D., Sheikholeslami, G., Zhang, A.: Findout: Finding outliers in very large datasets. Knowledge and Information Systems **4**(4), 387–412 (2002)
30. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. arXiv preprint arXiv:1507.05259 (2015)
31. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: Fa* ir: A fair top-k ranking algorithm. In: CIKM. pp. 1569–1578 (2017)
32. Zhang, K., Hutter, M., Jin, H.: A new local distance-based outlier detection approach for scattered real-world data. In: PAKDD (2009)