Data Science and Engineering (2021) 6:485–499
https://doi.org/10.1007/s41019-021-00169-x

FairLOF: Fairness in Outlier Detection

Deepak P[1,2] · **Savitha Sam Abraham**[3]

Received: 15 May 2021 / Revised: 30 June 2021 / Accepted: 8 August 2021 / Published online: 29 August 2021

Abstract
An outlier detection method may be considered fair over specified sensitive attributes if the results of outlier detection are not skewed toward particular groups defined on such sensitive attributes. In this paper, we consider the task of fair outlier detection. Our focus is on the task of fair outlier detection over multiple multi-valued sensitive attributes (e.g., gender, race, religion, nationality and marital status, among others), one that has broad applications across modern data scenarios. 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. We also show that a generalization of our method, named *FairLOF-Flex*, is able to open possibilities of further deepening fairness in outlier detection beyond what is offered by *FairLOF*.

Keywords Outlier detection · Fairness · Unsupervised learning

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, 34]). There has also been some recent interest in ensuring fairness within unsupervised learning tasks such as clustering [1], retrieval [35] and recommendations [25]. In this paper, we explore the task of fairness in outlier detection, an analytics task of wide applicability in myriad scenarios.

✉ Deepak P
deepaksp@acm.org

Savitha Sam Abraham
savitha.sam-abraham@oru.se

[1] Queen's University Belfast, Belfast, UK

[2] Indian Institute of Technology Madras, Chennai, India

[3] School of Science and Technology, Örebro University, 701 82 Örebro, Sweden

1.1 Outlier Detection and Fairness

The task of outlier detection targets to identify *deviant* observations from a dataset, and is usually modeled as an unsupervised task; [8] provides a review of outlier detection methods. The classical outlier characterization, due to Hawkins [16], considers outliers as 'observations that deviate so much from other observations as to arouse suspicion that they were generated by a different process'. Applications of outlier detection range across varied application domains such as network intrusions [19], financial fraud [26] and medical abnormalities [22]. Identification of non-mainstream behavior, the high-level task that outlier detection accomplishes, has a number of applications in new age data scenarios. Immigration officials 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 is inevitably within the remit of outlier detection.

The nature of the task of outlier detection, and the contexts it is used, makes it very critical when viewed from the perspective of fairness. When an abnormal behavior is flagged by an (automated) outlier detection system, it would naturally lead to concrete actions. In policing, it could lead to approaching the person involved to perform checks. In the financial sector, it could lead to de-activating a bank account or credit card temporarily. In certain other sectors, being classed as an outlier may be invisible, but could lead to significant inconvenience; for example, an insurance company may interpret the abnormality as a higher level of risk, and offer a higher premium. Across all the above cases, being classed as an outlier leads to moderate to significant levels of inconvenience. In the case of policing, the individual may be subjected to questioning or targeted surveillance, with the individual pushed to being defensive; this may be interpreted as skirting on the boundaries of violating the *presumption of innocence*, a human right enshrined in the universal declaration of human rights.[1] Further, it could also lead to frustration and a feeling of humiliation,[2] and could impact trust in policing among those who are subject to pro-active interrogation. Against this backdrop, consider a *sensitive attribute* such as ethnicity. An attribute which is often assigned to a person on the basis of chance (e.g., one does not choose one's ethnicity) and has a historical context of discrimination may be regarded as sensitive. While contexts of historical discrimination may differ in shape and size across varying geographies, the notion of sensitive attributes is often enshrined within legal and constitutional frameworks, as well as in affirmative action policies.[3] As an example, Title VII of the Civil Rights Act of 1964 in the United States declares that race, color, religion, sex, or national origin are *'protected'* attributes, whereas India's affirmative action system recognizes *caste* as a facet of social discrimination (and thus, a candidate for *sensitivity*). If the distribution of those who are classed as outliers is skewed toward particular ethnicities, say minorities as often happens, it directly entails that they are subject to much more inconvenience than others. This could lead to multiple issues:

- It would directly exacerbate targeting of minorities, since higher levels of pro-active surveillance would lead to higher rates of crime detection for minorities.
- Minorities being perceived as more likely to be targeted for pro-active surveillance could lead to a higher level of distrust in policing among them, potentially leading to social unrest.
- It would reinforce the stereotype that minorities are more likely to engage in crime.

These, and analogous scenarios in various other sectors, provide a compelling case to ensure that the inconvenience load stemming from outlier detection and other downstream processing be proportionally distributed across ethnicities. The same kind of reasoning holds for other sensitive attributes such as gender, religion and nationality.

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 inherently spread across other attributes. For example, geo-location, income and choice of professions may be correlated with ethnic, gender, religious and other identities. In fact, it has been argued that even a single attribute such as the postcode could reveal a lot of information about individuals [32], many of which are likely to be correlated with sensitive attributes. The identification of non-mainstream character either falls out from, or entails, an analogous and implicit modeling 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 patterns of racial prejudice in such settings.[4]

1.2 Outlier Detection and the Web

While we have discussed public sector scenarios to motivate outlier detection considerations so far, there are an abundance of other scenarios within the context of the web. Web has emerged, over the past decades, as a rich source of unlabeled 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

[1] https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[2] https://reason.com/2013/03/27/when-proactive-policing-becomes-harassme/.

[3] https://en.wikipedia.org/wiki/Affirmative_action.

[4] https://www.nyclu.org/en/stop-and-frisk-data.

(e.g., gender, race, nationality, religion and others) in such private sector settings for ethical reasons and to avoid bad press.[5]

1.3 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 [1, 11]. Second, we develop a fair outlier detection method, *FairLOF*, based on the framework of *LOF* [7], arguably the most popular outlier 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. We also outline a generalization of our method, called *FairLOF-Flex*, which allows usage of domain knowledge to customize *FairLOF*. 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. We also illustrate that our generalization, *FairLOF-Flex*, is able to open possibilities of further improving fairness in outlier detection outcomes.

2 Related Work

Given that there has only been very limited work in fair outlier detection, we start with covering related work across outlier detection and fairness in unsupervised learning, before moving on to discussing fair outlier detection.

2.1 Outlier Detection Methods

Since obtaining labeled data containing outliers is often hard, outlier detection is typically modeled as an unsupervised learning task where an unlabeled dataset is analyzed to identify outliers within it. That said, supervised and semi-supervised approaches do exist [8]. We address the unsupervised setting in our work. 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 [33], 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 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 the *LOF* proposal, there has been much research into local outliers, leading to work such as *SLOM* [9], *LoOP* [21] and *LDOF* [36]. Schubert et al [29] 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 much research 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 our *FairLOF* method.

2.2 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 restrict possibilities of similar objects receiving dissimilar outcomes. Individual fairness is typically 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* [23], a fair representation learner, targets to ensure that objects are indistinguishable with respect to 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].

[5] https://www.cnet.com/features/is-facebook-censoring-conservatives-or-is-moderating-just-too-hard/.

2.3 Fairness in Outlier Detection

In contrast to such several efforts into deepening fairness in unsupervised learning tasks, there has been very limited exploration into fair outlier detection. The main 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. This human-in-the-loop and explanation-oriented framework is only tangentially related to our remit of fairness in automated outlier detection. Apart from the above work, there has been a recent arXiv pre-print on fair outlier detection [30]. Their method, *FairOD*, is designed for the case of binary sensitive attributes and targets to achieve equal representation for the two groups among outlier results. This binary parity model does not generalize to multi-valued or numeric sensitive attributes. Further, the authors target not to use sensitive attributes at decision time; however, the decision is made by a neural auto-encoder model (a *global model*, among outlier detection families described in Sect. 2.1) that makes use of sensitive attributes in training. In contrast to such characteristics of [30], we address local neighborhood based outlier detection over data comprising sensitive attributes that could be of various types; including binary, multi-valued (e.g., ethnicity) or numeric (e.g., age).

3 Problem Definition

In this section, we outline the fair outlier detection task against the backdrop of (fairness-agnostic) outlier detection.

3.1 Task Setting

Consider a dataset $\mathcal{X} = \{\ldots, X, \ldots\}$ and an object pairwise distance function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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 and others) 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.

3.1.1 (Fairness-Agnostic) Outlier Detection

The task of (vanilla or fairness-agnostic) 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 neighborhoods. 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/or guided by practical considerations (e.g., manual labor budgeted to examine outliers).

3.1.2 Fair Outlier Detection

The task of fair outlier detection, in addition to identifying outliers, considers ensuring that the distribution of sensitive attribute groups among $\mathcal{O}$ reflects 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., [1, 6, 11]), 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:

- *'Quality'* metrics that measure how well objects with distinct local neighborhoods are placed in $\mathcal{O}$, and
- *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.

3.2 Motivation for Representational Parity

We outlined fair outlier detection using representational parity as our fairness objective. 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 toward 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, as argued in Sect. 1.1. 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* [20] 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 and ethnicity). The normative principle has been placed within the umbrella of the *'justice as fairness'* work due to John Rawls [28] 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[6] 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.

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

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 asymmetric 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 also 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

We first outline the motivation for our method, followed by the details of *FairLOF*.

5.1 Motivation

In many cases, the similarity space implicitly defined by the distance function $d(.,.)$ bears influences from the sensitive attributes and grouping of the dataset defined over such attributes. 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.

5.1.1 Under-Reporting of Large/Majority 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 a dataset comprising 75% males and 25% females. Such skew could occur in real-world cases such as datasets sourced from populations in a STEM college or certain professions (e.g., police[7]). 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. First, consider M; M is likely to have a *quite cohesive* and *predominantly male* kNN neighborhood due to both: (i) males being more likely in the

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

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

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 works in favor of a male-dominated neighborhood for F too; thus, F's neighborhood would be less gender homogeneous to itself, and thus less cohesive when measured using our S aligned $d(.)$. This would yield $lrd(M) > lrd(F)$, 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.*

5.1.2 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.*

5.2 FairLOF: The Method

The construction of *FairLOF* attempts to *correct* for such kNN neighborhood distance disparities across object groups defined over sensitive attributes. *FairLOF* distance correction is based on three heuristic 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 Sect. 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 $\mathcal{S}$.

5.2.1 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.

5.2.2 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.

5.2.3 Attribute Asymmetry

The extent of *k-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-distance*:

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

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:

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

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

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

Equation 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 representation 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 $\mathcal{X}$. $W_S^{\mathcal{X}}$ is computed as 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-distance*:

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

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 Sect. 5.1. In such cases, the *k-distance* would need to be corrected upward, as against downward. We incorporate that to yield the final *k-distance* formulation for *FairLOF*.

$$k\text{-}distance_{FairLOF}(X) = \left(max\{d(X,X')|X' \in N_k(X)\}\right) \times \left(1 - \lambda \sum_{S \in \mathcal{S}} \left(\mathbb{D}(\mathcal{X},S) \times W_S^{\mathcal{X}} \times D_{X.S}^{\mathcal{X}} \times Div(N_k(X), X.S)\right)\right) \tag{10}$$

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

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

This modification in *k-distance* warrants an analogous correction of $rd(.,.)$ to ensure level ground among the two terms determining $rd(.,.)$.

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

The second term in $rd(.,.)$ is corrected in the same manner as for *k-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 Sect. 4 with *k-distance*$(.,.)$ and $rd(.,.)$ replaced by their corrected versions from Eqs. 10 to 12, respectively. The overall process is outlined in Algorithm 1; 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.

Algorithm 1: FairLOF Method

input : Dataset $\mathcal{X}$, sensitive attributes $\mathcal{S}$, distance function $d(.,.)$
parameters: k, t (output size), λ, c

for $S \in \mathcal{S}$ **do**
  compute $W_S^{\mathcal{X}}$ using Eq. 8
  compute $\mathbb{D}(\mathcal{X},S)$ using Eq. 11
  for $s \in V(S)$ **do**
    compute $D_s^{\mathcal{X}}$ using Eq. 7
for $X \in \mathcal{X}$ **do**
  compute $N_k(X)$ using $d(.,.)$
  for $S \in \mathcal{S}$ **do**
    compute $Div(N_k(X), X.S)$ using Eq. 6
  compute $k\text{-}distance_{FairLOF}(X)$ using Eq. 10
  for $X' \in N_k(X)$ **do**
    compute $rd_{FairLOF}(X,X')$ using Eq. 12
  compute $fairlrd(X)$ using Eq. 2 where $rd(X,X')$ is replaced by $rd_{FairLOF}(X,X')$ above ;
  compute $flof(X)$ using Eq. 4 where $lrd(.)$ is replaced by $fairlrd(.)$
return *top-t* objects from $\mathcal{X}$ with the highest $flof(.)$ values

5.2.4 FairLOF Complexity

We briefly analyze the computational complexity of *FairLOF*. Equations 7 and 8 can be pre-computed at a an exceedingly small cost of $\mathcal{O}(|\mathcal{S}| \times m)$ where m is $max_{S \in \mathcal{S}}|V(S)|$. Equation 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-distance* adjustment may be considered as very light.

5.2.5 FairLOF and Sensitive Attribute Types

Our design of *FairLOF* is targeted toward multi-valued or categorical sensitive attributes. This is motivated by observing that the vast majority of sensitive attributes, such as race, color, religion, sex, national origin and caste, are multi-valued in nature. There are, however, some niche scenarios where numeric sensitive attributes, such as age, may need to be treated as sensitive. While treating them as categorical through discretization into categories is a way of incorporating such attributes into *FairLOF*, it does not capture the ordinal nature of those attributes in full essence. We believe that adapting Eqs. 6, 7 and 8 is feasible, to incorporate numeric sensitive attributes more organically; however, such adaptations remain outside the scope of this work.

5.3 FairLOF-Flex: Extending FairLOF

FairLOF uses three heuristic principles to *correct* distances in order to nudge the scoring to produce fair outlier detection results. To recap briefly, these are:

- *Neighborhood Diversity* The neighborhood diversity principle yields the factor $Div(N_k(X), X.S)$ in Eq. 10 and the factor $\mathbb{I}(X.S \neq X'.S)$ in Eq. 12.
- *Apriori Distribution* The apriori distribution principle maps to the factor $D^{\mathcal{X}}_{X.S}$ which figures in both Eqs. 10 and 12.
- *Attribute Asymmetry* The attribute asymmetry principle yields the factor $W^{\mathcal{X}}_{S}$ which is also used in both Eqs. 10 and 12.

FairLOF combines the above factors by *multiplying them together* with a direction factor $\mathbb{D}(\mathcal{X}, S)$ which, being in $\{-1, +1\}$ merely changes the direction of the correction without altering the extent of the correction. The product of those terms is computed on a per-attribute basis and summed up over all sensitive attributes to yield a discounting factor for distance computations, which is further weighted by λ. The product form to aggregate the contributions of the separate heuristic principles is informed by the expectation that they would be in similar ranges numerically. All of $Div(N_k(X), X.S)$, $\mathbb{I}(X.S \neq X'.S)$ and $D^{\mathcal{X}}_{X.S}$ are in [0, 1], whereas $W^{\mathcal{X}}_{S}$ is also expected to be in that range, though not strictly bounded above by 1.0. The product form places all the correction terms on an equal footing with respect to the extent to which they can inform the correction.

However, when *FairLOF* is applied within a particular application domain, we would naturally expect that these principles would apply to varying levels depending upon the specifics of the domain. As an example, consider using *FairLOF* in a spatial outlier detection scenario over a database of individuals in a town which happens to have deeply racially segregated areas. Given the domain knowledge that most individuals would be located in racially similar neighborhoods, we may want to limit the strength of the terms coming from the *neighborhood diversity* principle when race is used as a sensitive attribute, and leave the fairness heavylifting to be largely handled by remaining two factors. Different domains might require different levels of attenuation and amplification of the effects of the three different principles. Thus, a simple product form that offers equal weighting to the three terms would be considered as inflexible for specific domains. We do not attempt to outline a full suite of scenarios where the relative strengths of the factors from the separate heuristic principles need to be regulated in specific ways since that would inevitably depend on the domain knowledge from the application scenario. Instead, we attempt to outline that we recognize the need for controlling relative strengths of the correction terms, and describe a generalization of *FairLOF* to provide some flexibility in combining these factors.

We now outline a way to generalize *FairLOF* to incorporate such flexibility. We call the generalized method as *FairLOF-Flex*. The generalization is incorporated by modifying Eq. 10 as follows:

$$k\text{-}distance_{FairLOF\text{-}Flex}(X) = \left(max\{d(X, X') | X' \in N_k(X)\} \right) \times \left(1 - \lambda \sum_{S \in \mathcal{S}} \left(\mathbb{D}(\mathcal{X}, S) \times \left(W^{\mathcal{X}}_{S} \right)^{\alpha} \times \left(D^{\mathcal{X}}_{X.S} \right)^{\beta} \times Div(N_k(X), X.S) \right) \right) \tag{13}$$

This modification incorporates two hyperparameters, α and β, as exponents to $W^{\mathcal{X}}_{S}$ and $D^{\mathcal{X}}_{X.S}$, respectively. By varying α and β, we can vary the strengths of the corresponding terms, and thus, indirectly, the relative strength of the $Div(N_k(X), X.S)$ term as well. Analogous to the above, Eq. 12 is also modified as the following:

$$rd_{FairLOF\text{-}Flex}(X, X') = max\Bigg\{ k\text{-}distance_{FairLOF}(X'), d(X, X') \times \Bigg(1 - \lambda \sum_{S \in \mathcal{S}} \Big(\mathbb{D}(\mathcal{X}, S) \times \big(W_S^{\mathcal{X}}\big)^{\alpha} \times \big(D_{X.S}^{\mathcal{X}}\big)^{\beta} \times \mathbb{I}(X.S \neq X'.S) \Big) \Bigg) \Bigg\} \tag{14}$$

Given that $D_{X.S}^{\mathcal{X}}$ is in the [0, 1] range, $\beta > 1$ would in effect reduce the extent of the correction brought about by it (for example, 0.5^2 is less than 0.5). The same holds for $W_S^{\mathcal{X}}$ vis-a-vis α. *FairLOF-Flex* is thus simply the generalization of *FairLOF* to incorporate the relative strength hyperparameters, α and β. As evident, setting $\alpha = \beta = 1.0$ would make it equivalent to *FairLOF*. We do not set an analogous exponent for the diversity term, since the relative strength of the diversity term with respect to the other two can indirectly be controlled by setting α and β appropriately. For example, setting $\alpha = \beta = 2.0$ would implicitly be equivalent to giving a higher weighting to the diversity term, since the other two terms are discounted by squaring the corresponding terms. Therefore, setting $\alpha = \beta = 1.0$ is not equivalent to $\alpha = \beta = 2.0$.

As indicated earlier, we do not attempt to outline prescriptive ways on how the flexibility provided by way of the additional hyperparameters α and β could be used, since that is likely to intimately depend on how the specifics of the domain could be mapped to the terms coming from the three heuristic principles. Given that our work has been to come up with a general technique, we have not gathered expertise of the specific dataset domains to meaningfully reason and arrive at suitable parameter settings for those. We will illustrate, in our empirical analysis section, that effective usage of such parameters could plausibly lead to gains in fairness.

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].

6.1 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 measure how well *FairLOF* results align with the fairness-agnostic *LOF*, to assess quality of *FairLOF* results.

$$Jacc(\mathcal{R}_{LOF}, \mathcal{R}_{FairLOF}) = \frac{|\mathcal{R}_{LOF} \cap \mathcal{R}_{FairLOF}|}{|\mathcal{R}_{LOF} \cup \mathcal{R}_{FairLOF}|} \tag{15}$$

$$Pres(\mathcal{R}_{LOF}, \mathcal{R}_{FairLOF}) = \frac{\sum_{X \in \mathcal{R}_{FairLOF}} lof(X)}{\sum_{X \in \mathcal{R}_{LOF}} lof(X)} \tag{16}$$

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, and is thus a quantification of the extent to which LOF outliers find their place within the FairLOF results. If we consider the (fairness-agnostic) LOF results as 'true' quantifications of outlierness, the idea is that we would not want to diverge much from it while striving to ensuring fairness. Thus, high values of *Jacc*(., .) are desirable. Next, 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 $\mathcal{R}_{LOF}$. High values of *Pres*(., .) indicate that the FairLOF results are aligned well with the outlierness as estimated by the fairness-agnostic LOF method. As is obvious from the construction, higher values for *Pres*(., .) indicate better quality of *FairLOF* results.

The semantics of the *LOF* score suggests that objects with *lof*(.) scores less than 1.0 are *inliers* since they are in a higher density region than their local neighborhood. Even if $\mathcal{R}_{FairLOF}$ contains a few inlier objects, *FairLOF* could still score well in the case of *Jacc*(., .) as long as $\mathcal{R}_{FairLOF}$ contains several other objects from $\mathcal{R}_{LOF}$. Similarly, a $\mathcal{R}_{FairLOF}$ containing some inliers could still score well on *Pres*(., .) as long as there are other very high *lof*(.) scores among the objects it contains. In order to tease out the membership of inliers within FairLOF results more explicitly, we now outline another metric, *outlier-fraction* (*OF* for short), which measures the fraction of objects in $\mathcal{R}_{FairLOF}$ that are not inliers.

$$OF(\mathcal{R}_{LOF}, \mathcal{R}_{FairLOF}) = \frac{\sum_{X \in \mathcal{R}_{FairLOF}} \mathbb{I}(lof(X) \geq 1.0)}{|\mathcal{R}_{FairLOF}|} \tag{17}$$

where $\mathbb{I}$ is an indicator function that returns 1 or 0. This measures the fraction of objects that have a neighborhood that is equally or more sparse than its neighboring objects. It may be noted that for all the quality metrics outlined above, viz., *Jacc*(., .), *Pres*(., .) and *OF*(., .), 1.0 functions as an upper bound, and higher values indicate higher quality of *FairLOF* results.

6.2 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 once as against each cluster separately. Thus, we adapt the fairness metrics from [1, 31] as below:

$$ED(\mathcal{R}_{FairLOF}) = \sum_{S \in \mathcal{S}} Euclidean\_Distance(\mathcal{D}_S^{\mathcal{R}_{FairLOF}}, \mathcal{D}_S^{\mathcal{X}}) \tag{18}$$

$$Wass(\mathcal{R}_{FairLOF}) = \sum_{S \in \mathcal{S}} Wasserstein\_Distance(\mathcal{D}_S^{\mathcal{R}_{FairLOF}}, \mathcal{D}_S^{\mathcal{X}}) \tag{19}$$

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.

6.3 Quality-Fairness Trade-off

Note 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 $\lambda = 0.0$ yields *FairLOF* = *LOF*, with higher values reducing $Jacc(.)$, $Pres(.)$ and $OF(.)$ progressively) using a desired value of *Jaccard* similarity as a pilot point. 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, as well as the $Pres(.,.)$ and $OF(.,.)$ values at those configurations, will then be indicative of *FairLOF*'s effectiveness.

6.4 Single Sensitive Attribute and a Quota-Based System

As noted upfront, *FairLOF* is targeted toward 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 and other values for the gender attribute. 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[8]). The extension of this quota-based strategy to multiple sensitive attributes by modeling 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 Experimental Evaluation

We now describe our empirical evaluation assessing the effectiveness of *FairLOF* over several real-world datasets. We start by describing the datasets and the experimental setup, and then move on to presenting the experimental results and analyses.

7.1 Datasets

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

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

Table 1 Dataset information

Dataset	Domain	$\lvert\mathcal{X}\rvert$	Sensitive attributes used
Adult[a]	US 1994 census	48,842	Marital status, race, sex, nationality
CC[b]	Credit card default	30,000	Sex, education, marital status
W4HE[c]	Wikipedia HE use	913	Gender, disciplinary domain, uni name
St-Mat[d]	Student maths records	649	Gender, age
Tweets [18]	Twitter posts	47,560	Gender, ethnicity

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

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

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

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

Table 2 FairLOF Quality Results: 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

Dataset	Method	Guide Point	Quality measures					
			Jacc	*Det%*	*Pres*	*Det%*	*OF*	*Det%*
Adult	LOF		1.0		1.0		1.0	
	FairLOF	0.9	0.8939	10.61	0.9977	**00.23**	1.0	**0.0**
	FairLOF	0.8	0.7986	20.14	0.9906	**00.94**	1.0	**0.0**
CC	LOF		1.0		1.0		1.0	
	FairLOF	0.9	0.9011	09.89	0.9976	**00.24**	1.0	**0.0**
	FairLOF	0.8	0.7921	20.79	0.9879	01.21	1.0	**0.0**
W4HE	LOF		1.0		1.0		1.0	
	FairLOF	0.9	0.8776	12.24	0.9987	**00.13**	1.0	**0.0**
	FairLOF	0.8	0.8039	19.61	0.9951	**00.49**	1.0	**0.0**
St-Mat	LOF		1.0		1.0		1.0	
	FairLOF	0.9	0.9047	09.53	0.9970	**00.30**	1.0	**0.0**
	FairLOF	0.8	0.8182	18.18	0.9896	01.04	1.0	**0.0**
Tweets	LOF		1.0		1.0		1.0	
	FairLOF	0.9	0.8975	09.25	0.9986	**00.14**	1.0	**0.0**
	FairLOF	0.8	0.8051	19.49	0.9931	**00.69**	1.0	**0.0**

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 are indicated in percentages; Quality deteriorations < 1% are indicated in bold. The guide point and *Det%* only apply to FairLOF; thus, those cells against LOF are grayed out

identifiable information.[9] The datasets we use along with details are included in Table 1. While the *Adult* dataset is a dataset from a national Census exercise, *CC* is from the financial sector and comprises credit card default information. *W4HE* and *St-Mat* are both from the education sector, and relate to Wikipedia usage and student performance information, respectively. *Tweets*, on the other hand, forms the only text dataset in our analysis, and is a dataset of twitter posts collected for the purpose of hate speech identification; we have removed all records which have missing values for any sensitive attribute. Thus, it may be seen that the datasets encompass a wide variety of application domains and data types, and vary much in sizes as well as the sensitive attributes used. The choice of sensitive attributes has guided by what could be considered as those where the membership in those is not chosen by individuals. In the case of *W4HE*, we additionally assess that disciplinary domain and uni name are attributes worthy of ensuring fairness over.

7.2 Experimental Setup

For both *LOF* and *FairLOF*, we set t (to get top-t results) to 5% of the dataset size capped at 500 and set $k = 5$ consistently. For *FairLOF*, the parameter c (Ref. Eq. 8) is set to be $1/s$ where s is the number of sensitive attributes.

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

Table 3 FairLOF Fairness Results: Like for the case of Quality results, 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

Dataset	Method	Guide point	Fairness Measures			
			ED	*Impr%*	*Wass*	*Impr%*
Adult	LOF		0.2877		0.5328	
	FairLOF	0.9	0.1906	**33.75**	0.3468	**34.91**
	FairLOF	0.8	0.1714	**40.42**	0.2372	**55.48**
CC	LOF		0.2670		0.2152	
	FairLOF	0.9	0.2235	16.29	0.2112	01.86
	FairLOF	0.8	0.1568	**41.27**	0.2012	06.51
W4HE	LOF		0.2121		0.3305	
	FairLOF	0.9	0.0966	**54.46**	0.2989	09.56
	FairLOF	0.8	0.1820	14.19	0.2498	*24.42*
St-Mat	LOF		0.4174		0.9196	
	FairLOF	0.9	0.3467	16.94	0.8962	02.54
	FairLOF	0.8	0.3467	16.94	0.8962	02.54
Tweets	LOF		0.1248		0.0883	
	FairLOF	0.9	0.0598	**52.08**	0.0422	**49.34**
	FairLOF	0.8	0.0694	**44.39**	0.0491	**44.39**

The improvements in Fairness metrics are indicated in percentages. Fairness improvements of 20%+ are italicized, and those that are 30%+ are shown in bold. The guide point and *Det%* only apply to FairLOF; thus, those cells against LOF are grayed out

7.3 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, Pres and OF) and lower values are better on the fairness measures (ED and Wass)*. We follow the quality-fairness trade-off strategy as outlined in Sect. 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 for the quality measures are shown in Table 2, whereas those on the fairness measures are shown in Table 3. As observed in Sect. 6.3, the quality and fairness results are to be analyzed in tandem; high fairness improvements at small deterioration in quality metrics may be considered as a desirable point in the trade-off. Broadly, we observe the following:

- *Fairness Improvements FairLOF* is seen to achieve significant improvements in fairness metrics at reasonable 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 indicating that *FairLOF* achieves compelling fairness improvements.
- *Trends on Pres and OF* 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. The *OF*(., .) results are even more compelling; *FairLOF* is seen to score 1.0 at each of the settings. This indicates that there is not even one case where *FairLOF* includes an *LOF* inlier within its results. This further asserts the effectiveness of *FairLOF*.

In addition to the above observations, we note the following trends on *FairLOF* performance that throws light on the nature of *FairLOF*.

- *Wass versus ED FairLOF* is seen to achieve higher gains on *ED* as compared to *Wass*. Analyzing the nature of the relative character of these measures, we observe that *Wass* prefers the distances to be fairly distributed across attributes. It may be noted that the form of *FairLOF* considers corrections at the level of each sensitive attribute and aggregates them in a sum form (e.g., Eq. 10). Thus, it is designed to minimize aggregate fairness adherence, making it natural to expect higher gains on the *ED* measure than *Wass*.
- *Dataset Sizes and Gains* It is promising to note that *FairLOF* is able to achieve higher fairness gains in the large datasets such as *Adult*, *CC* and *Tweets* vis-a-vis the other smaller datasets. Larger datasets offer a larger search

space for solutions, and it may be interpreted that *FairLOF* is able to make use of the larger room accorded to it by the larger datasets to its advantage. Analogously, *FairLOF* gains are seen to saturate quickly for smaller datasets, a trend that is most evident for *St-Mat* in Table 3.

- *Trends on Guide Point* Our empirical study used two guide points, *0.9* and *0.8*. The latter allows for around twice as much deviation from *LOF* results as compared to the latter. *FairLOF* is seen to be able to exploit that additional wiggle room while moving from *0.9* to *0.8* in order to achieve improvements on the fairness metrics. Notable exceptions to this trend are seen in W4HE (for ED) and Tweets datasets; on further analysis, we found that such outlying trends could largely be attributed to noise effects than any consistent regularity.

Overall, the empirical analysis confirms the effectiveness of *FairLOF* is trading off small amounts of result quality in return for moderate to large improvements on fairness metrics.

7.4 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*, *OF*, *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 [27] 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. *OF* remained consistently at 1.0 even for high values of λ. We observed similar consistent trends for increasing c (Eq. 8) as well. *FairLOF* was also observed to be quite stable with changes of k and t.

7.5 FairLOF-Flex: Empirical Analysis

FairLOF-Flex, as discussed in Sect. 5.3, is a generalization of *FairLOF* through the introduction of two hyperparameters, α and β, which allow to control the relative strengths of the three kinds of corrections employed by *FairLOF*. However, setting α and β meaningfully requires eliciting deep domain insights and mapping those to the heuristic principles. We do not have such knowledge about the contexts from which our testbed datasets were sourced, and would thus not be able to verify whether setting these hyperparameters meaningfully would lead to good results. Thus, an alternative way to assess the utility of *FairLOF-Flex* flexibility is to ask the following question:

Does any setting of α and β, apart from the FairLOF defaults of $\alpha = \beta = 1.0$, yield better results than FairLOF on fairness and quality?

If we find that there are positions in the (α, β) hyperparameter space which are better than *FairLOF* on both the facets of evaluation (or alternatively, comparable on one facet and much better on the other), it indicates that *FairLOF-Flex* is likely to offer useful returns for investment in understanding the domain and using insights to adjust hyperparameters meaningfully.

Accordingly, we experimented with several (α, β) settings over two datasets, viz., the large *Adult* and *CC* datasets and the small *W4HE* dataset. We do not perform a comprehensive hyperparameter search over all datasets to further emphasize two points about the exploratory and cautious nature of this analysis. First, this comprehensive search in the parameter space is not a recommended strategy since there is a chance that such 'juicy' regions in the hyperparameter space may be incidental and not necessarily meaningful. For example, a brute force search could often uncover meaningless patterns, and capitalizing on such discoveries is often associated with p-hacking or data dredging.[10] That the comprehensive search uncovers juicy regions only serves to suggest that it is *plausible* that the flexibility offered by *FairLOF-Flex could* be used effectively. Second, it is not necessary or given that there would be other hyperparameter settings that outperform $\alpha = \beta = 1.0$. That a comprehensive parameter search for a dataset does not yield any better results than the *FairLOF* setting should not be interpreted as taking anything away from *FairLOF-Flex*. The malleability in the latter does not guarantee that it would outperform the former. The malleability is simply there for practitioners who have a deep knowledge of the data domain to be able to use it if they may like to.

Coming to the analysis, we found that variations in the hyperparameters often led to better results for *FairLOF-Flex* than *FairLOF*. It was also seen that the beneficial direction of variation from the *FairLOF* setting of $\alpha = \beta = 1.0$ differed across datasets; this further confirms our initial presumption that there is no domain-agnostic rule for such hyperparameter tuning, and that nuances in the domain should inform the hyperparameter settings. Some of the results are summarized in Table 4. The table illustrates settings for α and β where the *FairLOF-Flex* results are clearly superior to those achieved by *FairLOF*. In each of the rows in Table 4, *FairLOF-Flex* is seen to be better than *FairLOF* on fairness measures, while remaining comparable to *FairLOF* on

[10] https://en.wikipedia.org/wiki/Data_dredging.

Table 4 *FairLOF-Flex* Analysis: the table shows some parameter settings for α and β where *FairLOF-Flex* results are clearly superior to *FairLOF* results

Data set	GP*	Method	α	β	Quality			Fairness	
					Jacc	*Pres*	*OF*	*ED*	*Wass*
W4HE	0.8	FairLOF	1.0	1.0	0.8039	0.9951	1.0	0.1820	0.2498
		FairLOF-Flex	0.5	1.0	0.8039	**0.9967**	1.0	**0.1512**	**0.2281**
		FairLOF-Flex	0.5	0.5	0.7985	0.9915	1.0	**0.1679**	**0.2481**
Adult	0.9	FairLOF	1.0	1.0	0.8939	0.9977	1.0	0.1906	0.3468
		FairLOF-Flex	0.5	1.0	0.8832	0.9969	1.0	**0.1692**	**0.3054**
		FairLOF-Flex	1.0	1.5	0.8975	0.9975	1.0	**0.1861**	**0.3374**
CC	0.9	FairLOF	1.0	1.0	0.9011	0.9976	1.0	0.2235	0.2112
		FairLOF-Flex	1.5	1.5	0.9194	**0.9984**	1.0	0.2246	**0.2052**
		FairLOF-Flex	0.5	2.0	0.9084	**0.9982**	1.0	**0.2077**	**0.1932**

In all the cases above, there are visible improvements on fairness measures achieved by *FairLOF-Flex*, while the quality measures remain competitive with those of *FairLOF*. Measures where *FairLOF-Flex* outperforms *FairLOF* are indicated in bold

GP Guide point

quality measures. There are some regularities within datasets that are noteworthy. For example, *W4HE* prefers lower α, *Adult* prefers $\alpha < \beta$ and *CC* is seen to like higher values of β. During our empirical evaluation, we found a number of settings where *FairLOF-Flex* was seen to deteriorate, and thus, Table 4 is not meant to be a representative sample of our results. However, as mentioned above, the fact that *FairLOF-Flex* can achieve better results than *FairLOF* in certain configurations points to the possibility that deeper understanding of the domain and using the domain knowledge to tune the hyperparameters meaningfully holds much promise in deepening outlier detection fairness.

8 Conclusions and Future Work

In this paper, 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, especially within web data scenarios that operate at scale. 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 developed an outlier detection method, called *FairLOF*, inspired by the construction of *LOF* and makes use of three principles to nudge the outlier detection toward directions of increased fairness. We then outlined an evaluation framework for fair outlier detection, and used that in evaluating *FairLOF* extensively over real-world datasets. Through our empirical results, we observed that *FairLOF* is able to deliver substantively improved fairness in outlier detection results, at reasonable detriment to result quality as assessed against *LOF*. This illustrates the effectiveness of *FairLOF* in achieving fairness in outlier detection. We also designed a generalization of *FairLOF*, called *FairLOF-Flex*, which was seen to be able to deliver substantively improved fairness in outlier detection results in certain configurations, which indicates its promise in improving the fairness of outlier detection beyond *FairLOF*.

8.1 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, and other considerations of relevance to fair unsupervised learning [24].

Author Contributions DP was responsible for initiating the work and played a key role in conceptualization and refinement. DP also held the responsibility for implementing and managing the code base. SSA contributed the core ideas that led to the conception of the method being presented in the paper apart from actively engaging with DP in refining the method based on empirical insights. SSA also played a major role in identifying the datasets for empirical benchmarking. The manuscript was collaboratively authored across both the authors.

Funding This research was not supported by any funded Project.

Data availability All datasets used in the empirical evaluation are publicly available. The code will be made public upon acceptance of the manuscript.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

1. Abraham SS, Sundaram SS (2020) Fairness in clustering with multiple sensitive attributes. In: EDBT, pp 287–298
2. Asudeh A, Jagadish H, Stoyanovich J, Das G (2019) Designing fair ranking schemes. In: SIGMOD
3. Babaei K, Chen Z, Maul T (2019) Detecting point outliers using prune-based outlier factor (plof). arXiv preprint arXiv:1911.01654
4. Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104:671
5. Bei X, Liu S, Poon CK, Wang H (2020) Candidate selections with proportional fairness constraints. In: AAMAS
6. Bera SK, Chakrabarty D, Flores N, Negahbani M (2019) Fair algorithms for clustering. In: NeurIPS, pp 4955–4966
7. Breunig MM, Kriegel HP, Ng RT, Sander J (2000) Lof: identifying density-based local outliers. In: SIGMOD, pp 93–104
8. Chandola V, Banerjee A, Kumar V (2007) Outlier detection: a survey. ACM Comput Surv 14:15
9. Chawla S, Sun P (2006) SLOM: a new measure for local spatial outliers. Knowl Inf Syst 9(4):412–429
10. Chen J, Sathe S, Aggarwal C, Turaga D (2017) Outlier detection with autoencoder ensembles. In: SDM
11. Chierichetti F, Kumar R, Lattanzi S, Vassilvitskii S (2017) Fair clustering through fairlets. In: NIPS
12. Chouldechova A, Roth A (2020) A snapshot of the frontiers of fairness in machine learning. Commun ACM 63(5):82–89
13. Davidson I, Ravi S (2020) A framework for determining the fairness of outlier detection. In: ECAI
14. Domingues R, Filippone M, Michiardi P, Zouaoui J (2018) A comparative evaluation of outlier detection algorithms: experiments and analyses. Pattern Recognit 74:406–421
15. Fan W, Bouguila N, Ziou D (2011) Unsupervised anomaly intrusion detection via localized Bayesian feature selection. In: ICDM
16. Hawkins DM (1980) Identification of outliers, vol 11. Springer
17. Huang L, Vishnoi NK (2019) Stable and fair classification. arXiv:1902.07823
18. Huang X, Xing L, Dernoncourt F, Paul MJ (2020) Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. arXiv preprint arXiv:2002.10361
19. Jabez J, Muthukumar B (2015) Intrusion detection system (IDS): anomaly detection using outlier detection approach. Procedia Comput Sci 48:338–346
20. Knight C (2009) Luck egalitarianism: equality, responsibility, and justice. EUP
21. Kriegel HP, Kröger P, Schubert E, Zimek A (2009) Loop: local outlier probabilities. In: CIKM
22. Kumar V, Kumar D, Singh R (2008) Outlier mining in medical databases: an application of data mining in health care management to detect abnormal values presented in medical databases. IJCSNS Int J Comput Sci Netw Secur 8:272–277
23. Olfat M, Aswani A (2019) Convex formulations for fair principal component analysis. AAAI 33:663–670
24. Deepak P (2020) Whither fair clustering? In: AI for social good workshop
25. Patro GK et al (2020) Incremental fairness in two-sided market platforms: on updating recommendations fairly. In: AAAI
26. Pawar AD, Kalavadekar PN, Tambe SN (2014) A survey on outlier detection techniques for credit card fraud detection. IOSR J Comput Eng 16(2):44–48
27. Pearson K (1895) VII. Note on regression and inheritance in the case of two parents. Proc R Soc Lond 58(347–352):240–242
28. Rawls J (1971) A theory of justice. Harvard University Press
29. Schubert E, Zimek A, Kriegel HP (2014) Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data Min Knowl Discov 28(1):190–237
30. Shekhar S, Shah N, Akoglu L (2020) Fairod: fairness-aware outlier detection. arXiv preprint arXiv:2012.03063
31. Wang B, Davidson I (2019) toward fair deep clustering with multi-state protected variables. arXiv preprint arXiv:1901.10053
32. Webber R, Burrows R (2018) The predictive postcode: the geodemographic classification of British society. Sage
33. Yu D, Sheikholeslami G, Zhang A (2002) Findout: finding outliers in very large datasets. Knowl Inf Syst 4(4):387–412
34. Zafar MB, Valera I, Rodriguez MG, Gummadi KP (2015) Fairness constraints: mechanisms for fair classification. arXiv preprint arXiv:1507.05259
35. Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R (2017) Fa* ir: a fair top-k ranking algorithm. In: CIKM, pp 1569–1578
36. Zhang K, Hutter M, Jin H (2009) A new local distance-based outlier detection approach for scattered real-world data. In: PAKDD