Group Fairness in Case-Based Reasoning

Shania Mitra[1] Ditty Mathew[2] Deepak P[1,3(✉)] Sutanu Chakraborti[1]

[1] Indian Institute of Technology Madras, India
[2] University of Trier, Germany
[3] Queen's University Belfast, UK
shaniamitra9@gmail.com mathew@uni-trier.de
deepaksp@acm.org sutanuc@cse.iitm.ac.in

Abstract. There has been a significant recent interest in algorithmic fairness within data-driven systems. In this paper, we consider group fairness within Case-based Reasoning. Group fairness targets to ensure parity of outcomes across pre-specified sensitive groups, defined on the basis of extant entrenched discrimination. Addressing the context of binary decision choice scenarios over binary sensitive attributes, we develop three separate fairness interventions that operate at different stages of the CBR process. These techniques, called Label Flipping (LF), Case Weighting (CW) and Weighted Adaptation (WA), use distinct strategies to enhance group fairness in CBR decision making. Through an extensive empirical evaluation over several popular datasets and against natural baseline methods, we show that our methods are able to achieve significant enhancements in fairness at low detriment to accuracy, thus illustrating effectiveness of our methods at advancing fairness.

Keywords: Fairness · Group Fairness · Case-based Reasoning

1 Introduction

Algorithmic fairness [24] has attracted significant scholarly attention in recent times. While this has seen most interest in the case of machine learning (ML) [11], fairness has been explored within allied data-driven areas such as retrieval and recommenders [15] and natural language processing [10]. There has been emerging recent interest in fairness within Case-Based Reasoning (CBR) as well, with the first paper on algorithmic fairness in CBR appearing recently [8].

Fairness, as a concept with origins and a long legacy in the social sciences, is a deeply nuanced and contested construct. There are several different definitions of fairness [22], many of which are in conflict with one another. Two streams of fairness concepts, viz., individual and group fairness [14], have been subject to much study within data-driven learning. Individual fairness targets to ensure that all objects are treated uniformly, so that *similar* objects (similarity defined appropriate to the task) are accorded similar outcomes. In sharp contrast, group fairness is anchored on the notion of *sensitive attributes* (e.g., gender, race, nationality, religion), and seeks to ensure that outcomes are fairly

distributed across groups defined on the basis of such sensitive attributes. Sensitive attributes are not defined at a technical level, but chosen on the basis of extant evidence of historical and contemporary discrimination. Group fairness, often referred to variously as outcome fairness and distributive justice, is thus focused on ensuring that the workings of the algorithms are not configured in a way that some sensitive groupings are advantaged more than others. Adherence to group fairness may thus require that individuals who are similar on a task-level basis be treated differently (leading to possible violations of individual fairness), so the outcomes along the sensitive groupings are uniform. In contemporary society, the workings of meritocracy could be thought of as close to the spirit of individual fairness, whereas affirmative action and policies targeted to level off gender and race gaps (in pay, education, or other forms of achievement) are aligned with the notion of group fairness. It is also useful to note that there are latent similarities in the structures of these notions [6].

CBR systems, in sharp contrast to mainstream ML, are non-parametric models in that they do not involve the construction of a statistical model of a fixed and predetermined capacity. This has significant ramifications in analyzing, assimilating and mitigating algorithmic unfairness within them. As a case in point, a recent fairness intervention in a non-parametric local neighborhood-based outlier detection mechanism [12] was based more on statistical corrections using local neighborhood properties, as opposed to the usage of fairness optimization objectives in conventional ML tasks (e.g., [3]). The only work on fair CBR [8] is also divergent from the additional objective approach in that it adopts a metric learning approach, which involves modifying the similarity knowledge container to achieve the desired fairness goal. Their model, FairRet [8], is focused on mitigating underestimation bias towards minority protected groups. This, while using sensitive groups, may be seen as using a restricted form of group fairness.

In this paper, for the first time to our best knowledge, we initiate research into group fairness - defined as parity of outcomes across sensitive groups - in case-based reasoning. Our focus is on CBR systems that assign binary outcomes and on enhancing uniformity of outcomes across two groups defined over a binary sensitive attribute (e.g., male/female, white/non-white). We develop separate fairness-targeted interventions at three stages viz., *pre-processing*, *weighting* and *retrieval*. While our pre-processing intervention targets changing the data labelling in a targeted manner, our weighting approach assigns weights to individual cases to enhance fairness in outcomes. The third approach is an adaptation-stage approach where the aggregation mechanism is modified towards the fairness goal. Through extensive empirical validation over several real-world datasets, we illustrate the effectiveness of our methods in reducing the disparity of outcomes across the sensitive attributes.

2 Related Work

While there has been much work on fairness within ML [11], there has been, as we mentioned earlier, just one prior work on fairness in CBR [8]. Fairness

has also been explored within the context of information retrieval; this relates to CBR in that CBR also encompasses a retrieval step, though it goes much further than just presenting retrieval results. We briefly summarize related work within this section. We particularly focus on work relating to group fairness within this section, given the focus of this paper.

2.1 Fairness in Retrieval and Recommender Systems

Fairness considerations in retrieval and recommenders [15] have centred on ensuring diversity across sensitive attributes within the top-k retrieved results. This has often been termed as proportional fairness [30] given that the intent is to ensure that the top-k results reflect the proportions of sensitive attribute groups within the broader dataset. These are often realized using bespoke constraints such as diversity constraints [27]. Apart from demographic-sensitive groupings, fairness has also been explored over political bias [19] and popularity [31]. Another line of exploration has been to relax the query-level fairness proportionality constraint and ensure that there are no statistically significant deviations from proportionality across queries [29].

Apart from such generally applicable work on fairness in retrieval, fairness has been explored within specific contexts of retrieval, such as multi-stakeholder interactions within recommender systems viz., 2-sided fairness [23]. Work on group recommender systems [26][20] has focused on delivering better recommendations to groups of people by aggregating individual preferences of group members, modelling social factors such as personality awareness and trust between them. These approaches, however, do not concern themselves with the disparity between different groups and thus, have limited applicability in group-fair CBR. Novel fairness constructs such as attention fairness have been devised to account for the case that human users tend to focus on the top results within an ordered result display paradigm [5]. While the above interventions are interesting and pertinent to the retrieval stage, fairness in CBR systems would need to be conceptualized across the different stages. For example, ensuring proportional representations in the results may do little to further fairness, unless the downstream aggregation/adaptation step is able to make use of it. Thus, such retrieval-focused work has limited applicability within the context of fair CBR.

2.2 Fair Local Outliers

Local neighborhood-based approaches have been quite popular within the task of outlier detection [9, 18]. While these have very limited relevance to CBR beyond the usage of local neighborhood-based retrieval, their meta structure of retrieval followed by bespoke processing of retrieved sets to arrive at a decision resonates with the spirit of CBR. The downstream processing is very specific to outlier detection and is thus highly divergent from the intent and structure of CBR systems.

A recent work [12] considers a fairness-oriented adaptation of arguably the most popular local neighborhood-based outlier detection algorithm, LOF [9].

The proposed method, FairLOF [2], does statistical 'corrections' for fairness at three levels, viz., the diversity of local neighborhood as measured over sensitive attributes, the overall representational skew between groups defined over sensitive attributes, and corrections for the extent to which the similarity knowledge container embeds sensitive attribute knowledge within itself. The corrections are focused on fairness insofar as they relate to the downstream statistical processing of the LOF method, and are, thus, again of limited applicability to the considerations of CBR.

There has been some work in detecting outlying cases for case base maintenance viz., the Repeated Edited Nearest Neighbour approach (RENN) [1]. Here, a case is considered noisy and is removed if its class label differs from the majority of its k nearest neighbors. While this case base maintenance paradigm aligns with the spirit of the CW approach that we present, RENN does not relate to sensitive attributes, and is thus orthogonal to group fairness.

2.3 FairRet: Algorithmic Fairness in CBR

We now briefly summarize a recent work [8], the only extant work on algorithmic fairness in CBR to our best knowledge. This work focuses on addressing an intricate notion of algorithmic discrimination called *underestimation bias*. We illustrate this by using a simplistic example. Consider a binary choice scenario such as those encountered in job application shortlisting, where each application is to be either shortlisted or rejected. Let us suppose that the shortlisting success rates for males and females are 50% and 40%, respectively, reflecting gender discrimination as often observed in society. However, the algorithmic decision-making process may further accentuate this skew and offer an even lower success rate, say 30%, to female applicants. This is an instance of *underestimation bias*, the focus of *FairRet*. The underestimation, in this case, may be quantified as 0.75 (i.e., 30/40, expressed as a percentage) ([8], Sec 2.1), and is sought to be remedied by modifying the similarity knowledge container using metric learning approaches realized using multi-objective particle swarm optimization.

It may be noted that rectifying underestimation bias would bring the success rates for females to 40%, which is still significantly lower than the success rates for males, which stands at 50%. Thus, *FairRet* is aligned with fairness conceptualizations such as *separation* [4] that seek to equalize deviations or error rates (cf. success rates) for different sensitive groups, thus implicitly considering the decision-profile embodied in the labelled data as the reference standard. In other words, addressing underestimation bias would not achieve group fairness which is often conceptualized as *independence*, the notion that seeks equalizing success rates across sensitive groupings. Independence, often referred to as statistical parity [16], requires that there be a parity in the distribution of outcomes across sensitive attribute groups. Thus, group fairness (or statistical parity or independence), as instantiated within our example scenario, would target that the success rates for females be enhanced to 50%, or otherwise equalized across the gender groups; for example, by equating the success rates in decision making for both groups at 45%.

Our work, in contrast to *FairRet*, considers achieving *independence* or *group fairness*, which is about equalizing success rates across demographic groups. To summarize, in contrast to *FairRet* that uses the decision profile within the labelled data as the target for fairness, our focus is on using uniformized success rates across sensitive groups as the fairness target. Yet, given the high-level similarity in that biased behavior is sought to be mitigated and given that *FairRet* is the only extant work on fairness in CBR, we use *FairRet* as a baseline.

3 Problem Definition

We first outline the CBR decision making context we address, followed by the targeted fairness requirement and the fairness metric which we consider.

3.1 CBR Decision Making Scenario

Consider a dataset $\mathcal{X} = \{\ldots, X, \ldots\}$, where each element X corresponds to a case. X, as is typical of cases in CBR, comprises two parts, the data d and the label l. Additionally, each X is also associated with a value for a sensitive attribute denoted as s. Thus, $X = [d, l, s]$; d, l and s will have overloaded interpretations for ease of ensuing narration, but the intended interpretation will be clear from the context. As a first work towards group fairness, we restrict our attention to binary decision choices (so, $l \in \{0, 1\}$) and binary sensitive attributes (so, $s \in \{0, 1\}$). In a concrete scenario within a job shortlisting context, each X could comprise a historical job application (as d), the decision accorded to it (as l), with s denoting a sensitive demographic of the applicant that is known to be a facet of social discrimination (e.g., male/non-male, or white/non-white).

We now consider a CBR system $\mathcal{C}$ which makes use of the case base $\mathcal{X}$ to make decisions over an incoming stream of data points (job applications), which we will denote as $\mathcal{Y} = \{\ldots, Y, \ldots\}$. Each data point is, much like the case of $\mathcal{X}$, associated with a value for the sensitive attribute too. Thus, $Y = [d, -, s]$. The task of the CBR system is to fill up the missing label for elements in $\mathcal{Y}$ with its predictions. We will denote the labelling choice offered by the CBR system as $l = \mathcal{C}(Y)$. In the interest of general applicability across diverse scenarios, we would like the CBR system to not explicitly refer to the sensitive attribute of Y in making its decision; thus, $\mathcal{C}(Y)$ does not depend on s. Within the job shortlisting scenario, this amounts to creating a CBR system that will make choices over applications without explicitly referring to gender/race membership.

3.2 Group Fairness

Having laid out the CBR decision-making scenario and pertinent notations, we are now ready to present our group fairness consideration in technical terms. Consider that $\mathcal{C}$ has been applied over all elements of $\mathcal{Y}$. We would be able to measure the success rates for the separate demographic subsets of $\mathcal{Y}$ under the decisions offered by $\mathcal{C}$.

$$SR_{\mathcal{C}}(\mathcal{Y}, s=0) = \frac{\sum_{[d,-,s]\in\mathcal{Y}} \mathbb{I}(l=1 \wedge s=0)}{\sum_{[d,-,s]\in\mathcal{Y}} \mathbb{I}(s=0)} \tag{1}$$

where $\mathbb{I}(.)$ is the identity function which evaluates to 1 when the inner condition is satisfied and 0 otherwise. Simply stated, $SR_{\mathcal{C}}(\mathcal{Y}, s=0)$ is the success rates of the $s=0$ subset of $\mathcal{Y}$ under $\mathcal{C}$. Analogously, $SR_{\mathcal{C}}(\mathcal{Y}, s=1)$ is also defined. Our intent is to ensure that the success rates of the separate demographic groups are as similar as possible.

$$SR_{\mathcal{C}}(\mathcal{Y}, s=0) \quad \approx \quad SR_{\mathcal{C}}(\mathcal{Y}, s=1) \tag{2}$$

3.3 Disparity

Our focus is on ensuring that the design of $\mathcal{C}$ is such that the disparity between success rates of sensitive sub-groups is minimized as much as possible. Towards this, we use statistical disparity as the evaluation metric, defined as the following:

$$SDisp(\mathcal{C}, \mathcal{Y}) = |SR_{\mathcal{C}}(\mathcal{Y}, s=0) - SR_{\mathcal{C}}(\mathcal{Y}, s=1)| \tag{3}$$

This, or its variants, have been explored in various efforts in fair AI literature. For example, the above metric corresponds to the *violation of statistical parity* metric in [21] (Sec 2.3.1) and demographic (dis)parity in [25] (Sec 3.2). This will form our primary metric towards profiling the methods on group fairness in our empirical evaluation. As obvious lower values of $SDisp$ are more desirable.

4 GFCBR: Our Methods for Group Fair CBR

In this section, we now outline our suite of methods for group fairness within CBR, which we will denote as GFCBR. This comprises three approaches viz., *Label flipping* (LF) as a pre-processing method, *Case weighting* (CW) as a method to weigh cases based on their context, and *Weighted adaptation* (WA) as a method for aggregating/adapting retrieved results to form a labelling decision.

4.1 LF: Label Flipping

As outlined in Section 2.3, the target of group fairness may be misaligned with the labelling in the data within the case base $\mathcal{X}$. Even if the data labelling offers males and females success rates of 50% and 40% respectively, the intent of group fairness requires us to produce equalized success rates, one that cannot obviously be achieved by strict adherence to labelling patterns. In this backdrop, our label flipping technique targets to alter the ground truth labellings in the case base so that it becomes feasible for $\mathcal{C}$ to achieve group fairness.

CBR Model: To ensure generality of the label flipping-based pre-processing method, we assume a very simple design for a CBR decision maker which we

outline upfront. Given a similarity function to judge similarities between cases, the CBR decision for an object is given as follows:

$$\mathcal{C}_{\mathcal{X}}(d) = MajVote(top\text{-}k_{\mathcal{X}}(d)) \tag{4}$$

where $\mathcal{C}_{\mathcal{X}}$ denotes the CBR system working over the case base $\mathcal{X}$, $top\text{-}k_{\mathcal{X}}(d)$ denotes the top-k most similar data objects to d from within $\mathcal{X}$, and $MajVote(.)$ simply computes the majority vote (recollect we are dealing with binary labellings) from across the objects. To avoid ties, k may be set to an odd number.

Leave-one-out CBR: Towards ease of describing the label flipping approach, we introduce a leave-one-out instantiation of $\mathcal{C}$ which we denote as $\mathcal{C}^{L1O}$. The leave-one-out CBR system operating over $\mathcal{X}$ takes each element of $\mathcal{X}$, $X = [d, l, s]$ and determines a label for it using the *other* objects in $\mathcal{X}$ as the case base. This determination may be different from the object's label l since the neighbors of the object may mostly have the other label. Once decisions are made for each element of $\mathcal{X}$ using the leave-one-out approach, we can compute the $SDisp(\mathcal{C}^{L1O}, \mathcal{X})$ as the disparity between the success rates of the $s = 1$ and $s = 0$ subgroups within $\mathcal{X}$ when assessed using the decisions from the $\mathcal{C}^{L1O}$ model.

Label Flipping Approach: Having outlined the context and necessary background, we now describe our label flipping approach, which is outlined in Algorithm 1. The label flipping approach considers modifying the case base $\mathcal{X}$ by flipping some labels in its cases greedily, with an intent of choosing to flip the label of the object that reduces $SDisp(\mathcal{C}^{L1O}, \mathcal{X})$ most, at each step. Once the $SDisp(.,.)$ stabilizes - i.e., further label flips cannot decrease the disparity any further - the flipping process is stopped. We also introduce an additional parameter called the budget b, which allows for stopping earlier as necessary. The budget is specified as a percentage of the case base (e.g., 1%) which restricts the label flipping approach to making at most b% label flips in the dataset, even if convergence is not achieved by then. The high-level intuition behind this label flipping approach is that a case base over which $\mathcal{C}^{L1O}$ is able to achieve low disparity would facilitate group fair decisions for new cases too.

Once the label flipping process is complete, we end up with a modified case base $\mathcal{X}'$ which differs from $\mathcal{X}$ in that some object labels have been flipped. The modified CBR system is simply $\mathcal{C}_{\mathcal{X}'}$, which differs from $\mathcal{C}_{\mathcal{X}}$ in that it works over the modified dataset. This modified CBR system $\mathcal{C}_{\mathcal{X}'}$ is now ready to be applied over a new stream of cases - such as an unseen dataset $\mathcal{Y}$ - and that it works over a label-flipped dataset would aid it in achieving lower $SDisp$ over $\mathcal{Y}$. We note that this approach may be seen as creating an alternative case base which would be used for decision making. The original experiences in the case base may be maintained separately as a version of record.

4.2 CW: Case Weighting

Complementary to actually changing labellings in the data as the strategy in LF, CW adopts a different strategy towards the same goal of enhancing group fairness. The strategy within CW is to augment each case with a numeric weight

Algorithm 1: Label Flipping (LF)

Input: Case base $\mathcal{X}$, k, budget b
Output: A modified case base $\mathcal{X}'$
$\mathcal{X}' = \mathcal{X}$
while $SDisp(\mathcal{C}^{L1O}, \mathcal{X}')$ *has not converged and budget* b *has not been reached* **do**
  $X^* = \arg\min_{X \in \mathcal{X}'} SDisp(\mathcal{C}^{L1O}, \mathcal{X}' \cup \{labelflip(X)\} - \{X\})$
  $\mathcal{X}' = \mathcal{X}' \cup \{labelflip(X^*)\} - \{X^*\})$
end
return $\mathcal{X}'$

within $[0, 1]$ in such a way that cases that are aligned with group fairness get a higher weighting than others. We use the leave-one-out mechanism introduced in Section 4.1 in determining case weights.

Advantaged Group: In our scenario of binary decision choices, the existence of disparity entails that one of the sensitive groups has a higher success rate than the other. In typical scenarios involving systemic discrimination, the advantaged group is often consistent. This could be *males* in the case of gender as the sensitive attribute or *white* in the case of ethnicity. Without loss of generality, we will assume that $s = 1$ is the advantaged group.

Neighborhood Misaligned Cases: Consider the leave-one-out mechanism, $\mathcal{C}^{L1O}$, applied over a case $X = [d, l, s] \in \mathcal{X}$. The decision by $\mathcal{C}^{L1O}$, denoted $l^{L1O} = \mathcal{C}^{L1O}(d)$, could be different from the actual label associated with X, i.e., l. These cases, where $l \neq l^{L1O}$, indicate that they are, to some extent, misaligned with their neighborhood. The decision preference of their neighboring cases, as reflected through $\mathcal{C}^{L1O}$, is different from their own label. Towards designing CW, we posit that such cases with neighborhood misalignment could be differentially weighted towards enhancing group fairness.

Differentiated Weighting: We now outline the differentiated weighting heuristic, which is at the core of the CW technique. Cases that are aligned with their neighborhood, i.e., with $l = l^{L1O}$, are assigned a weight of unity. For neighborhood misaligned cases, we set weights based on how well their neighborhood is aligned with the goal of group fairness. We will illustrate this briefly.

On considering the scenario of cases from the disadvantaged group, denoted as $s = 0$, if the neighborhood decision (l^{L1O}) indicates a positive outcome, but the actual label is negative, it implies that the case is in a neighborhood that supports group fairness, since the prediction favours the selection of disadvantaged cases. Group fairness means that cases from the disadvantaged group should be assigned positive outcomes more often than what the labels suggest. We assign a weight of $\lambda \in [0, 1]$ to such cases, since, the prediction albeit incorrect is one that promotes the selection of minorities. Second, if the neighborhood decision is in favour of a negative outcome, but the actual case label is positive, we judge this to be an outlier, and assign a weight of 0.

In other words, our case weight is dependent on how well the neighborhood, whose decision is reflected in l^{L1O}, is aligned with the notion of group fairness. The above logic is flipped in the case of the advantaged group, $s = 1$, since we would like them to be assigned positive decisions at a lower rate than that supported by the labels. This weighting scheme is summarized in Eq 5. λ serves as a hyperparameter to this approach which would need to be pre-specified.

$$w(X = [d, l, s]) = \begin{cases} 1 & l^{L1O} = l \\ \lambda & s = 0 \wedge l^{L1O} = 1 \wedge l = 0 \\ 0 & s = 0 \wedge l^{L1O} = 0 \wedge l = 1 \\ 0 & s = 1 \wedge l^{L1O} = 1 \wedge l = 0 \\ \lambda & s = 1 \wedge l^{L1O} = 0 \wedge l = 1 \end{cases} \tag{5}$$

Algorithm 2 is then a simple application of this weighting scheme in round-robin fashion across the cases in the case base.

Algorithm 2: Case Weighting (CW)

Input: Case base $\mathcal{X}$, k
Output: Weights for each case in $\mathcal{X}$, denoted $w(X), \forall X \in \mathcal{X}$
for $X = [d, l, s] \in \mathcal{X}$ **do**
  $l^{L1O} = \mathcal{C}_{\mathcal{X}}^{L1O}(X)$
  Set $w(X)$ as in Eq. 5
end
return $\{w(X) | X \in \mathcal{X}\}$

CW-Weighted CBR: The weights assigned to cases would need to be exploited in decision making, should a CBR system working over a weighted case base is to provide enhanced group fairness. The natural way would be to aggregate the labels from the top-k neighbors for a new case using weighted aggregation, and choose the label associated with the highest aggregate weight. This is illustrated as below:

$$\mathcal{C}_{\mathcal{X}}^{CW}(d) = \underset{label \in \{0,1\}}{\arg\max} \sum_{X=[d,l,s] \in top-k_{\mathcal{X}}(d)} w(X) \times \mathbb{I}(label = l) \tag{6}$$

In our empirical evaluation, we will consider such a CBR system while profiling the effectiveness of CW.

4.3 WA: Weighted Adaptation

Our third technique, WA, adopts a group-fairness oriented strategy that operates much more downstream than either LF or CW. In particular, the case base $\mathcal{X}$ is kept as such, without being subject to label modifications or apriori neighborhood-based weightings. However, cases that are retrieved as similar

ones to a query case are accorded weights based solely on their (l, s) combination, while adapting their labels to make a decision on the query case.

Weight Formulation: The weight formulation in WA, unlike that in CW, is not dependent on the neighborhood of the case and is solely determined by the combination of (l, s) values associated with a case. Consider a particular (l, s) combination, such as $(l = 1, s = 0)$; this denotes a case where a positive outcome is assigned to a data object from the disadvantaged group. Similarly, $(l = 1, s = 1)$ denotes a case where a positive outcome is associated with a data object from the advantaged group. In the interest of ensuring group fairness, we would naturally like the former to have a higher influence in any decision-making process. Similarly, among $(l = 0, s = 0)$ and $(l = 0, s = 1)$, we may want the former to have a lower weighting in our interest to push up the success rate for the disadvantaged group. Additionally, we would like the differentiated weighting to reflect the quantum of the extant disparity in success rates across sensitive groups, as estimated using the leave-one-out mechanism. Our weighting scheme is outlined below:

$$w(l,s) = \frac{\frac{1}{|\mathcal{X}|} \times \sum_{X \in \mathcal{X}} \mathbb{I}(\mathcal{C}^{L1O}(X) = l)}{\frac{1}{\sum_{X \in \mathcal{X}} \mathbb{I}(X.s = s)} \times \sum_{X \in \mathcal{X}} \mathbb{I}(\mathcal{C}^{L1O}(X) = l \wedge X.s = s)} = \frac{p_{L1O}(l)}{p_{L1O}(l|s)} \quad (7)$$

The weighting for a case with $(l = 1, s = 0)$ would simply be the ratio of the rate of $l = 1$ decisions by the $\mathcal{C}^{L1O}$ system, to the rate of $l = 1$ decisions by the same system for $s = 0$ data objects. The shorthand representation at the right end of Eq. 7 illustrates the notion in more intuitive notation using probabilities and conditional probabilities. Suppose the overall success rate is 50%, with the disadvantaged group recording a success rate of 40% and the advantaged group recording 60%, the weight associated with positive-labelled data objects from the disadvantaged and advantaged group would respectively be 1.25 ($= \frac{0.5}{0.4}$) and 0.83 ($= \frac{0.5}{0.6}$). Thus, disadvantaged objects bearing a positive label get a higher say in the process, in alignment with the spirit of group fairness. If the overall success rate of 50% is borne out of a higher disparity - say, 70% and 30% for the advantaged and disadvantaged groups - the analogous weightings become more divergent, at 1.67 and 0.71. This illustrates how the quantum of extant disparity factors into the weightings.

Given that we address the binary decision choice scenario with binary sensitive attributes, there are only four distinct (l, s) combinations. Thus, each case would be associated with one of four weights, making WA an extremely simple weighting formulation.

WA-Weighted CBR: The WA weights, as introduced above, are incorporated into the decision-making process analogously to the case of CW weights.

$$\mathcal{C}^{WA}_{\mathcal{X}}(d) = \underset{label \in \{0,1\}}{\arg\max} \sum_{X=[d,l,s] \in top-k_{\mathcal{X}}(d)} w(l,s) \times \mathbb{I}(label = l) \quad (8)$$

As in the case of CW, we will use $\mathcal{C}^{WA}$ in our empirical evaluation to profile group fairness.

4.4 Discussion

All of the above three techniques involve the usage of the leave-one-out scheme, often multiple times. Given that all three techniques keep the similarity measure (i.e., the similarity knowledge container) consistent, one may pre-compute the nearest neighbor set for each object in the case base beforehand, and simply look-up nearest neighbors, rather than computing the top-k neighbors afresh. Such pre-computation would render the implementation of the techniques quite inexpensive in computational terms. Further, all our methods are best implemented as a pre-processing scheme, which may be performed once upfront at the CBR system deployment time, and do not have any further bearing on query response times. This allays computational cost considerations.

5 Experimental Evaluation

We now present an empirical analysis of our proposed approaches on several real-world datasets against the *FairRet* baseline.

5.1 Datasets, Setup and Evaluation Setting

We first start by describing our datasets and experimental/evaluation setup.

Datasets/Setup: We use four popular binary labelled datasets that have been popular in fairness-related studies viz., COMPAS violent recidivism [13], UCI Credit Card [28], Census [17], and Exemplar [7]. The COMPAS dataset is used to predict violent recidivism, while the UCI Credit Card (UCICC) dataset predicts the risk of credit card default. The Census dataset, often called the Adult Income dataset, aims to predict income, and the Exemplar dataset predicts income brackets based on age (already binarized in the dataset, perhaps as *young* and *elderly*). In all of these datasets, gender/sex is a protected attribute, except for Exemplar, where age is used as the protected attribute. Table 1 summarizes the key statistics of each dataset, including the number of samples and the proportion of the minority group. Towards using these datasets in our empirical validation, we split each dataset into three parts: 70% as the case base (i.e., $\mathcal{X}$), 10% for validation, and 20% for evaluation (i.e., $\mathcal{Y}$). Unless mentioned otherwise, we consistently set $k = 5$ and $b = 2\%$ (LF), with λ tuned based on the validation set. Apart from comparing against *FairRet*, our main baseline, we also indicate the results achieved by a simple majority-voting based CBR system over the dataset; this will be denoted as *Base*.

Dataset	Number of Records	Protected Attribute	Percentage Disadvantaged
COMPAS	4743	Sex	19.8%
UCICC	30000	Gender	39.6%
Census	48842	Sex	33.2%
Exemplar	37607	Age	32.8%

Table 1. Dataset Summary

Evaluation Metrics: Our primary evaluation metric, as introduced in Sec 3.3, is the disparity in the outcome (measured as success or failure rate within the context of binary decision scenarios), which we would like to be as low as possible. We express disparity as the difference in percentage success rate and thus may be interpreted as a percentage. Typical fairness-agnostic algorithms target to achieve as high an accuracy as possible. In seeking to heed an additional constraint, that of fairness, it is natural to expect that the attention to accuracy, and thus, the accuracy achieved, will be affected. Yet, we can potentially claim some success as long as the fairness gains are significantly higher than the accuracy detriment. Thus, disparity and accuracy are the focus of our empirical evaluation.

5.2 Disparity Results

Dataset	Label	Base	FairRet	LF	CW	WA
COMPAS	8.6%	6.3%	2.9%	**0.3%**	5.2%	0.6%
UCICC	3.4%	1.8%	1.7%	0.5%	1.1%	**0.2%**
Census	19.5%	17.8%	2.6%	7.9%	2.6%	**1.9%**
Exemplar	13.7%	8.1%	6.7%	6.3%	**0.5%**	2.2%

Table 2. Summary of Disparity Results

Our summary of disparity results appears in Table 2. The base system is seen to record a high disparity, even recording a high of ≈18% in the Census dataset. The disparity as assessed using the ground truth labels over the test set, is also shown for reference under the column *Label* in Table 2. *FairRet* is seen to bring down the disparity significantly from *Base* and *Label* in most cases. However, in each dataset, one of our methods ends up recording the lowest disparity, sometimes achieving levels as low as 0.2%. In particular, *WA* and *CW* beat the *FairRet* method in each dataset. This is quite expected, since the target of *FairRet* is just to rectify underestimation bias, whereas our methods target to go further, and towards full parity of success (failure) rates.

The performance of the *LF* method deserves closer attention. The *LF* method is extremely effective for COMPAS and UCICC, whereas the effectiveness in the other datasets is not yet to desirable levels; in fact, it fares worse than *FairRet* on the Census dataset. *LF*, it may be recollected, is a post-processing method which makes use of the simple CBR system; thus, it is quite limited in its ability to address the fairness consideration, unless the extant unfairness may be attributed to a small number of labels. Its greedy choice of the label to flip is another limiting factor, as we will see in a later section.

In summary, our methods are able to achieve very low levels of disparity, and their effectiveness in advancing the group fairness consideration is very apparent from the results in Table 2. It is notable that the results are at $< 1\%$ level in two datasets, indicating that the results are already very close to perfect group fairness.

5.3 Accuracy Results

Table 3 records the accuracy profile of our methods and the baselines. As expected, the *Base* CBR system achieves the best accuracy, with being truthful to labels being its sole focus. The key analysis point in this case is that of the comparison between *FairRet* and our methods. While our methods record, for most cases, a significant improvement over *FairRet* and *Base* in terms of fairness, we would expect that there would be an analogous drop in accuracy for our methods. The effectiveness of our methods depends on how low that drop is, when measured against *FairRet*. We can see that the drop in accuracy for our methods against *FairRet* is limited to, in most cases in the $\approx 1\%$ range, which contrasts favourably against the significant fairness gains analyzed earlier. This indicates that our methods are able to operate at an accuracy configuration similar to *FairRet*, while achieving moderate to significant fairness gains over it. In one case, that of Census, it is notable that our methods achieve an improvement in accuracy over *FairRet*, which is more than encouraging.

Dataset	Base	FairRet	LF	CW	WA
COMPAS	**91.59%**	86.60%	84.31%	86.05%	84.27%
UCICC	**81.65%**	79.39%	78.92%	78.84%	78.24%
Census	**81.82%**	77.00%	79.78%	79.31%	79.10%
Exemplar	**86.46%**	79.62%	84.12%	79.56%	78.40%

Table 3. Summary of Accuracy Results

5.4 Parameter Sensitivity Analyses

Having established the effectiveness of our methods to improve fairness achievement at low costs to accuracy across a number of datasets, we now turn our attention to analyzing the sensitivity of our methods to their hyper-parameters. One may recollect that there are two hyper-parameters among our methods, the budget b which determines the number of label flips in LF, and the parameter λ within CW which determines the weights of some cases. This is in addition to the CBR neighborhood size parameter, i.e., k.

Dataset	#Flips	% of Dataset
COMPAS	41	0.86%
UCICC	37	0.12%
Census	78	0.16%
Exemplar	54	0.14%

Table 4. Analysis of Flips by LF

Parameters within Our Methods

We now analyze the two parameters within our methods viz., flipping budget b (LF) and weighting parameter λ (CW).

Label Flipping Budget: While we set $b = 2\%$ to allow for up to 2% label flips, the label flipping procedure stopped far earlier due to convergence in disparity

(recall the stopping condition in Alg 1). As shown in Table 4, LF stopped after a few scores of flips even in our larger datasets. While it is promising to note that the fairness improvements achieved by LF are at the expense of just a few label flips, deepening the effectiveness of the LF technique may require devising a new non-greedy heuristic to continue label flips even when the disparity converges within the context of a single decision step.

Weighting Parameter in CW, λ: Another parameter within our methods is the weighting parameter in CW, denoted as λ. As indicated earlier, this was determined based on the validation set. Yet, across the widely varying datasets, a $\lambda \approx 0.1$ was found to be the most suitable value. This indicates that λ is not highly sensitive to changes in datasets, and a reasonably small value, one that has the ability to signal the direction of the intended change (wrt disparity), is what matters.

Neighborhood Size

The generic CBR parameter, that of the choice of neighborhood size, has been consistently set to $k = 5$ in our experiments. We did observe a small but consistent trend across variations in k, across datasets. Higher values of k led to improved fairness at the cost of reduced accuracy and vice versa. This is intuitively explained in that higher values of k lead to attention to wider neighborhoods, expanding the remit of the fairness interventions that are employed. These trends remained consistent across datasets; this could mean that k could work as a knob parameter to tune the attention to fairness.

6 Conclusions and Future Work

We considered the task of designing group fair CBR schemes, which seek to ensure that the disparity in outcomes across demographic groups is minimized. We developed three techniques, a pre-processing based label flipping scheme (LF), a contextual case weighting scheme (CW) and a weighted case-adaptation methodology (WA). We outlined how they would facilitate group fairness in CBR through various distinct ways. Further, in an empirical evaluation across multiple real-world datasets, we illustrated the empirical improvements in fairness that our methods achieve. Our experiments illustrate that improved fairness is achieved at low cost to accuracy, making them effective fairness-oriented CBR techniques.

Future Work

While our separate methods were designed for fairness interventions at different stages of the CBR process, it would be interesting to understand the complementarity between these methods and exploit them for application in scenarios where the user has control over all stages of the CBR process. We are considering extending this to cover multi-choice and structured decision scenarios and multi-valued sensitive attributes.

References

1. An experiment with the edited nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics **SMC-6**(6), 448–452 (1976). https://doi.org/10.1109/TSMC.1976.4309523
2. Abraham, S.S.: Fairlof: fairness in outlier detection. Data Science and Engineering **6**, 485–499 (2021)
3. Abraham, S.S., Sundaram, S.S., et al.: Fairness in clustering with multiple sensitive attributes. EDBT (2020)
4. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org (2019), http://www.fairmlbook.org
5. Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of attention: Amortizing individual fairness in rankings. In: The 41st international acm sigir conference on research & development in information retrieval. pp. 405–414 (2018)
6. Binns, R.: On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 514–524 (2020)
7. Blanzeisky, W., Cunningham, P., Kennedy, K.: Introducing a family of synthetic datasets for research on bias in machine learning (07 2021)
8. Blanzeisky, W., Smyth, B., Cunningham, P.: Algorithmic bias and fairness in case-based reasoning. In: Case-Based Reasoning Research and Development: 30th International Conference, ICCBR 2022, Nancy, France, September 12–15, 2022, Proceedings. pp. 48–62. Springer (2022)
9. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data. pp. 93–104 (2000)
10. Chang, K.W., Prabhakaran, V., Ordonez, V.: Bias and fairness in natural language processing. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts (2019)
11. Chouldechova, A., Roth, A.: A snapshot of the frontiers of fairness in machine learning. Communications of the ACM **63**(5), 82–89 (2020)
12. Deepak, P., Abraham, S.S.: Fair outlier detection. In: 21th International Conference on Web Information Systems Engineering: WISE 2020. pp. 447–462 (2020)
13. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Science Advances **4**(1), eaao5580 (2018). https://doi.org/10.1126/sciadv.aao5580, https://www.science.org/doi/abs/10.1126/sciadv.aao5580
14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
15. Ekstrand, M.D., Burke, R., Diaz, F.: Fairness and discrimination in retrieval and recommendation. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1403–1404 (2019)
16. Hertweck, C., Heitz, C., Loi, M.: On the moral justification of statistical parity. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 747–757 (2021)
17. Kohavi, R.: Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. KDD (09 1997)
18. Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Loop: local outlier probabilities. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 1649–1652 (2009)

19. Kulshrestha, J., Eslami, M., Messias, J., Zafar, M.B., Ghosh, S., Gummadi, K.P., Karahalios, K.: Search bias quantification: investigating political bias in social media and web search. Information Retrieval Journal **22**, 188–227 (2019)
20. Kunaver, M., Porl, T.: Diversity in recommender systems a survey. Know.-Based Syst. **123**(C), 154–162 (may 2017). https://doi.org/10.1016/j.knosys.2017.02.009, https://doi.org/10.1016/j.knosys.2017.02.009
21. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(3), e1452 (2022)
22. Narayanan, A.: Translation tutorial: 21 fairness definitions and their politics. In: Proc. conf. fairness accountability transp., New York, USA. vol. 1170, p. 3 (2018)
23. Patro, G.K., Biswas, A., Ganguly, N., Gummadi, K.P., Chakraborty, A.: Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In: Proceedings of the web conference 2020. pp. 1194–1204 (2020)
24. Pessach, D., Shmueli, E.: Algorithmic fairness. arXiv preprint arXiv:2001.09784 (2020)
25. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Computing Surveys (CSUR) **55**(3), 1–44 (2022)
26. Quijano-Sanchez, L., Recio-Garcia, J.A., Diaz-Agudo, B., Jimenez-Diaz, G.: Social factors in group recommender systems. ACM Trans. Intell. Syst. Technol. **4**(1) (feb 2013). https://doi.org/10.1145/2414425.2414433, https://doi.org/10.1145/2414425.2414433
27. Yang, K., Gkatzelis, V., Stoyanovich, J.: Balanced ranking with diversity constraints. arXiv preprint arXiv:1906.01747 (2019)
28. Yeh, I.C., hui Lien, C.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications **36**(2, Part 1), 2473–2480 (2009). https://doi.org/https://doi.org/10.1016/j.eswa.2007.12.020, https://www.sciencedirect.com/science/article/pii/S0957417407006719
29. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: Fa* ir: A fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 1569–1578 (2017)
30. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking: A survey. arXiv preprint arXiv:2103.14000 (2021)
31. Zhang, Y., Feng, F., He, X., Wei, T., Song, C., Ling, G., Zhang, Y.: Causal intervention for leveraging popularity bias in recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 11–20 (2021)