Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes

Shen Yan
shenyan@usc.edu
University of Southern California

Hsien-te Kao
hsientek@usc.edu
University of Southern California

Emilio Ferrara
emiliofe@usc.edu
University of Southern California

ABSTRACT

Machine learning models are at the foundation of modern society. Accounts of unfair models penalizing subgroups of a population have been reported in domains including law enforcement, job screening, etc. Unfairness can spur from biases in the training data, as well as from class imbalance, i.e., when a sensitive group's data is not sufficiently represented. Under such settings, balancing techniques are commonly used to achieve better prediction performance, but their effects on model fairness are largely unknown. In this paper, we first illustrate the extent to which common balancing techniques exacerbate unfairness in real-world data. Then, we propose a new method, called *fair class balancing*, that allows to enhance model fairness without using any information about sensitive attributes. We show that our method can achieve accurate prediction performance while concurrently improving fairness.

CCS CONCEPTS

• **Social and professional topics** → **Socio-technical systems**;
• **Computing methodologies** → **Machine learning**.

KEYWORDS

fairness; bias; class balancing

ACM Reference Format:
Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411980

1 INTRODUCTION

Machine learning models pervade modern decision-making systems and are used in a plethora of application domains including law enforcement, medical and job screening, etc. The increasing impact of such systems on people's lives has raised concerns about the fairness and biases of algorithmic-driven decisions. Such models, at times, have been criticized for their opacity and lack of interpretability [8, 23]. Factors such as data imbalance, that can lead to the systematic under-representation of a certain group, may unintentional induce discrimination on model outcomes.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411980

Class imbalance is common in most real-world data collections. For example, there are fewer instances of data about subjects with a certain disease than that of healthy subjects. This is due the natural imbalance dictated by the prevalence of that disease in the target population. However, data imbalance may also spur from biased data collection or polluted training data: for example, it has been reported that decision-making systems used to predict recidivism, employed for parole or sentencing decisions, are trained on data where some sensitive demographic attribute (e.g., race) is over- or under-represented with respect to the prevalence of a model outcome (i.e., a criminal defendant's likelihood of committing a crime) [8]. Various class balancing techniques (e.g., resampling, synthetic samples generation) [5, 26] have been proposed to improve model accuracy, while their effects on model fairness are largely unknown.

The increasing concerns about the fairness of machine learning decision-making systems motivated various solutions, ranging from data pre-processing [4, 12] to post-processing [21]. However, most of the studies require access to sensitive attributes (e.g., gender or race of the subjects). In many real-world applications, sensitive attributes are not observable due to privacy concerns or legal restrictions. In the United States, there are laws and regulations prohibit to use and request sensitive and protected attributes in many decision making systems. For example, credit institutions cannot ask or access information about race to applicants who apply for credit [Equal Credit Opportunity Act: 15, 12 CFR §1002.5(b)]. Similarly, insurance companies can no longer request race information from the individuals they insure [11]. Thus, approaches that don't rely upon, and don't access sensitive attributes are better suited to real-world applications.

Contributions of this work

In this work, we conduct experiments on real-world datasets to investigate the effects of existing class balancing algorithms on fairness. Our analyses show that common class balancing techniques can exacerbate unfairness, also in part due to inherent properties of the data. Inspired by these observations, we propose a new class balancing method named *fair class balancing*.

Our *fair class balancing* method is a revised class balancing technique that is inspired by the K-Means SMOTE [13] algorithm. Our proposed method can enhance model fairness as well as prediction accuracy. It can be viewed as a pre-processing strategy to enhance model fairness without observing sensitive attributes. In summary:

- We propose two new bias measures to quantify two types of biases in data collection, which are also two sources of discrimination in model outcomes.
- We investigate the effects of common class balancing techniques on fairness and show how different data properties affect the interplay between balancing and fairness.

- We propose the *fair class balancing* method, which provides a way to improve the fairness of model outcomes when sensitive attributes are unobserved. Experimental results show that the proposed method yields state-of-the-art performance on both accuracy and fairness metrics.

2 RELATED WORK

The problem of fairness of machine learning algorithms has been drawing increasing research interests in recent years. Much work has been done to improve the fairness through different aspects of the modeling process, including feature selection [14], data preprocessing [4, 12], model adjustment [19, 28], and post-processing [21]. Different approaches also have been proposed for different machine learning applications including representation learning [27], clustering [22], natural language processing [3], etc.

All of the strategies above require the access to sensitive attributes in order to mitigate the bias. However, collecting that type of information might be difficult, or even forbidden by the law, in real-world applications. Recently, a few studies have explored different strategies to address the issue. One typical solution is using *non-sensitive information* as proxy for sensitive attributes. Previous work [16] has shown that non-sensitive information can be highly correlated with sensitive attributes. Kilbertus *et al.* proposed a method for selecting proxy groups by inferring causal relationships in the underlying data [20]. That framework is based on the assumption of causality between proxy features and sensitive attribute(s). Instead, *proxy fairness* [15] leverages the correlations between proxy features and true sensitive attributes. Proxy features are used as the alternative to sensitive attribute(s) when applying a standard fairness-improving strategy. The use of weighted estimators [7] is also discussed as the assessment for proxy models. Although the existence of proxy features gives the hope to improve fairness with unobserved sensitive attributes, identifying perfect proxy groups is still a challenging task.

3 PRELIMINARIES

In this paper, we study the problem of model fairness in supervised learning, where the goal is to predict a true outcome Y from a feature vector X based on labeled training data. The fairness of prediction $\hat{Y}$ is evaluated with respect to *sensitive groups* of individuals defined by *sensitive attributes* A, such as gender or race. Both outcome Y and sensitive attributes A are assumed to be binary, i.e., $Y \in \{0, 1\}$ and $A \in \{0, 1\}$, where $A = 1$ represents the privileged group (e.g., male), while $A = 0$ represents the unprivileged group (e.g., female).

As for the criteria to assess fairness, in addition to the traditional group-based fairness metrics, in this study we propose two new measures to assess the amount of existing bias.

3.1 Group Fairness Definitions and Metrics

We consider three group fairness definitions in literature: *Equal Opportunity* [17], *Equalized Odds* [17], and *Statistical Parity* [10].

Definition 3.1 (Equal Opportunity). Equal opportunity measures a binary predictor $\hat{Y}$ with respect to A and Y

$$\Pr\{\hat{Y} = 1|A = 1, Y = 1\} = \Pr\{\hat{Y} = 1|A = 0, Y = 1\}.$$

Definition 3.2 (Equalized Odds). The equalized odds of a binary predictor $\hat{Y}$ with respect to sensitive attribute A and outcome Y, is if $\hat{Y}$ and A are independent conditional on Y

$$\Pr\{\hat{Y} = 1|A = 1, Y = y\} = \Pr\{\hat{Y} = 1|A = 0, Y = y\},$$

where $y \in \{0, 1\}$. For the outcome $y = 1$, the constraint measures the differences of true positive rates across different sensitive groups. For $y = 0$, the constraint measures false positive rates.

Definition 3.3 (Statistical Parity). Statistical parity rewards the classifier for classifying each group as positive at the same rate. The statistical parity of a binary predictor $\hat{Y}$ is

$$\Pr\{\hat{Y} = 1|A = 1\} = \Pr\{\hat{Y} = 1|A = 0\}.$$

The three definitions cover different concepts of group fairness in real-world applications. *Equal opportunity* and *Equalized odds* are linked to the performance of the model, which ensures the model has similar accuracy across different sensitive groups. *Statistical parity* aims to equalize the outcomes for different groups: for example, a bank might be legally required to give loans at equal rates to both genders. Based on the above three definitions, we define the following metrics used in this work:

Metric 1 (Equal Opportunity Differences (E. Opp.)). Equal opportunity measures a binary predictor $\hat{Y}$ with respect to A and Y:

$$\Pr\{\hat{Y} = 1|A = 0, Y = 1\} - \Pr\{\hat{Y} = 1|A = 1, Y = 1\}.$$

Metric 2 (Average Equalized Odds Difference (E. Odds)). Average equalized odds differences computes the average difference of false positive rate (false positives / negatives) and true positive rate (true positives / positives) between unprivileged and privileged groups.

The average equalized odds difference of a binary predictor $\hat{Y}$ with respect to sensitive attribute A and outcome Y, is if $\hat{Y}$ and A are independent conditional on Y:

$$\frac{1}{2}[\Pr\{\hat{Y} = 1|A = 0, Y = 0\} - \Pr\{\hat{Y} = 1|A = 1, Y = 0\}] +$$
$$\frac{1}{2}[\Pr\{\hat{Y} = 1|A = 0, Y = 1\} - \Pr\{\hat{Y} = 1|A = 1, Y = 1\}].$$

Metric 3 (Statistical Parity Differences (SP)). Statistical parity rewards the classifier for classifying each group as positive at the same rate. The statistical parity of a binary predictor $\hat{Y}$ is

$$\Pr\{\hat{Y} = 1|A = 0\} - \Pr\{\hat{Y} = 1|A = 1\}.$$

For all the three metrics, values that are closer to zero have less differences between sensitive groups, which indicate fairer predictions. Negative values indicate the models bias against the unprivileged group. Positive values indicate that the models favor the unprivileged group.

3.2 Hardness and Distribution Biases

The fairness metrics that have been so far proposed in the literature are designed to measure the bias in model outcomes. Previous work [6] has shown that the properties of collected data have great impact on model fairness. Thus, it is important to quantify and measure biases in data collection. To capture and quantify the amount of different types of bias in data collection, we propose two new measures, namely *hardness bias* and *distribution bias*.

3.2.1 Hardness Bias. Different hardness measures have been proposed to indicate the hardness level of an instance to be correctly classified. In this work, we use *k-Disagreeing Neighbors* (kDN) [25] as the hardness metric. kDN measures the local overlap of an instance in the original task space in relation to its nearest neighbors. The kDN of an instance is the percentage of the k nearest neighbors (using Euclidean distance) for an instance that do not share its target class value.

$$\text{kDN}(x, y) = \frac{|(x', y') : x' \in \text{kNN}(x), y' \neq y|}{k},$$

where kNN(x) is the set of k nearest neighbors of the instance x, and y is the target class value for x.

Larger kDN (i.e., closer to 1) indicates that the instance is harder to be predicted correctly. If the instances within one group has higher kDN comparing to other groups, the prediction accuracy of that group will also be lower. Thus, the hardness bias is defined as the distribution difference of kDN between groups (e.g., instances in different classes, sensitive groups). In this work, we measure kDN when k=5, i.e., *5-Diagreeing Neighbors*.

Definition 3.4 (Hardness Bias). The hardness bias $\Gamma_A(y)$ of a dataset D is defined as the *Kullback–Leibler divergence* (KL) of the distribution of kDN of instances with different sensitive attributes $A \in \{0, 1\}$:

$$\Gamma_A(y) = \text{KL}(f(\{\text{kDN}(x, y)|A = 1\}) - f(\{\text{kDN}(x, y)|A = 0\})),$$

where $f(\{\text{kDN}(x, y)|A = a\})$ is the density of the kDN distribution of all instances with $A = a$ and $Y = y$.

The kDN distribution of each sensitive group affects the classification accuracy. Thus the hardness bias is linked to the accuracy differences between groups, which has the similar concept of *Equal opportunity* and *Equalized odds* measures.

3.2.2 Distribution Bias. In addition to the hardness bias between classes, the distribution of sensitive groups within each class also affects the model performance.

Definition 3.5 (Distribution Bias). The distribution bias $\Delta_A(y)$ of a dataset D is defined as the difference of probabilities of $Y = y$, conditioned upon values of the sensitive attribute $A \in \{0, 1\}$:

$$\Delta_A(y) = \Pr(Y = y|A = 1) - \Pr(Y = y|A = 0),$$

where $y \in \{0, 1\}$.

3.3 Datasets

In this work, as case studies we adopt two datasets that have been widely investigated in previous fairness research.

3.3.1 Adult. The *Adult* dataset [9] contains 32,561 records of yearly income (represented as a binary label: over or under $50,000) and twelve categorical or continuous features including education, age, and job types. The gender (binary represented, male or female) of each subject is considered as sensitive attribute.

3.3.2 COMPAS. The ProPublica *COMPAS* dataset [23] relates to recidivism, to assess if a criminal defendant will commit an offense within a certain future time. The dataset is gathered by ProPublica, with information on 6,167 criminal defendants who were subject to screening by *COMPAS*, a commercial recidivism risk assessment tool, in Broward County, Florida, in 2013–2014. Features in this dataset include number of prior criminal offenses, age of the defendant, etc. The race (binary, white/non-white) of the defendant is the sensitive attribute of interest.

3.3.3 Violent Crime. The violent recidivism version of the ProPublica data [23] describes the same scenario as the recidivism data described above, but where the predicted outcome is a rearrest for a violent crime within two years. 4,010 individuals are included. The race (binary, white/non-white) of the defendant is the sensitive attribute of interest.

4 EFFECTS OF CLASS BALANCING ON FAIRNESS

Class balancing techniques such as over-sampling and under-sampling are commonly adopted to modulate the class distribution of a dataset that exhibits class imbalance. Despite, typically, class balancing yields better prediction performance, the effect of such type of strategies on model fairness is largely unknown.

Next, we focus on five popular class balancing techniques, i.e., random over-sampling (ROS), random under-sampling (RUS), synthetic minority over-sampling technique (SMOTE) [5], cluster centroids (CC) [26], and K-Means SMOTE [13], and investigate the effects of their adoption on model fairness. All the algorithms described above are implemented in the reference Python library `imbalanced-learn`.[1]

For all our experiments, we use the `scikit-learn`'s *Logistic Regression Classifier* as classification model. Each dataset is randomly split into 80% development set and 20% test set with 50 randomized restart. Models are trained on development set with 10-fold cross validation parameter tuning. All sensitive attributes are removed before learning to avoid information leak.

4.1 Effects on Biases

Class balancing techniques are designed to address the imbalance of the target variables. However, the biases in the data may increase after balancing, due to the process being oblivious of the inherent properties of the datasets.

Figure 1 gives an illustration of the biases change before and after class balancing with Random Over Sampling (ROS). As Figure 1 shows, the three datasets have different properties. In the original *Adult* dataset, positive (i.e., high income) samples only compose 25% of the data. More than 20,000 positive samples are added after class balancing, where only 15% are female. Therefore, more male samples are added after class balancing, which further increases the distribution bias. Similarly, *Violent Crime* also has a significant lower ratio of positive samples, and only 4% positive samples are belong to the non-white populations. The distribution bias is increased 6 times after class balancing with ROS (*cf.*, Table 1). Differently from the two datasets, the *COMPAS* dataset is more balanced, with 54% of the instances being positive samples and 46% being negative; the ratios of different race groups within each class are also similar (*cf.*, Figure 1 (c)).

[1] https://imbalanced-learn.readthedocs.io

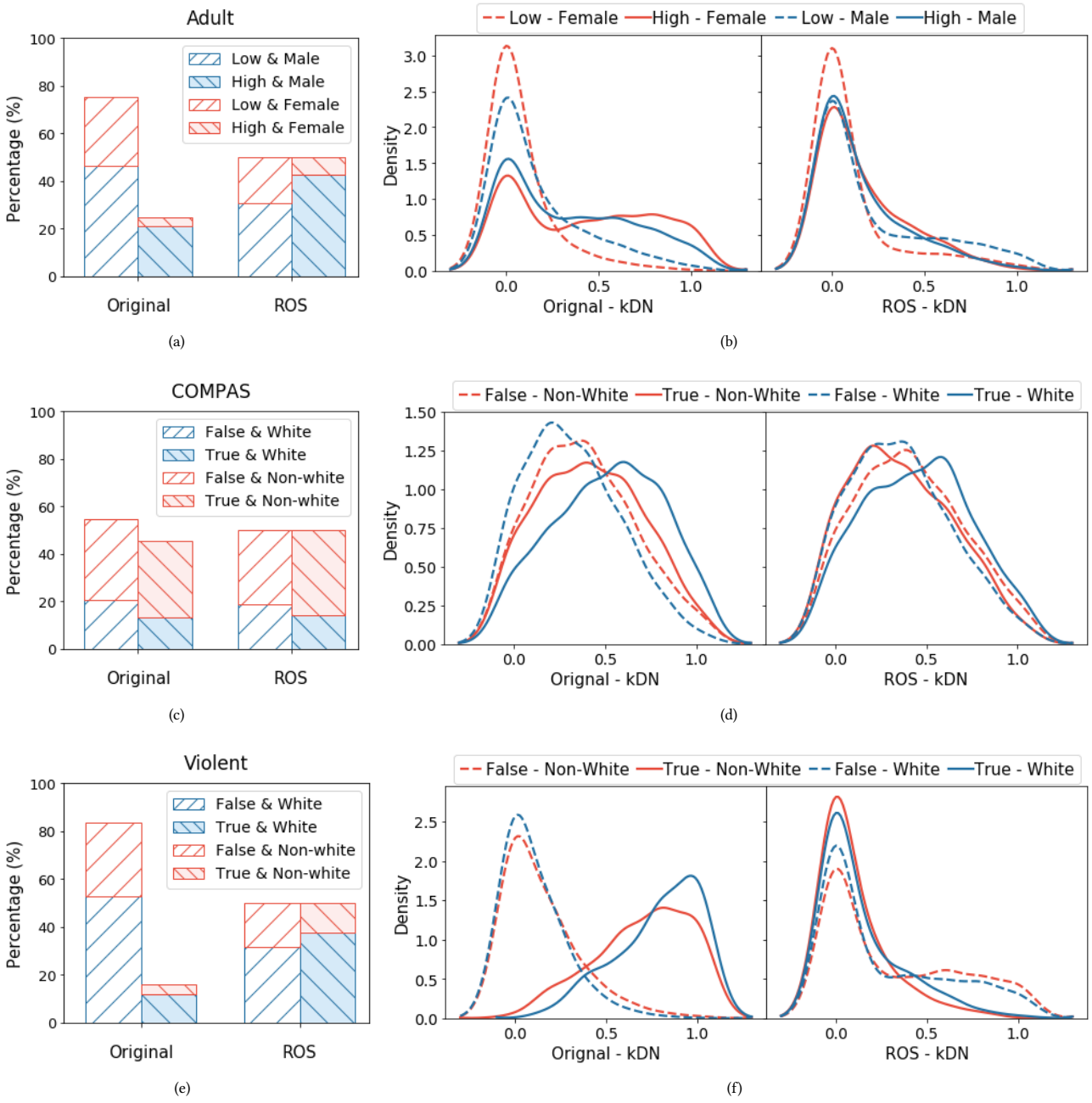

Figure 1: Examples of the bias changes before and after class balancing. (a), (c) and (e) illustrate the change of *distribution bias*. (b), (d) and (f) compare the *hardness bias* before and after class balancing for *Adult*, *COMPAS* and *Violent Crime* data.

Regarding the hardness bias, Figure 1 (b), (d) and (f) illustrate the distribution of kDN (i.e., 5DN). All three datasets have different patterns of the kDN distribution. Most of the samples of the *Adult* dataset have a small kDN, while the kDN distribution of the *COMPAS* dataset centralize around 0.5. For the *Violent Crime* dataset, positive samples have higher kDN than negative samples. After class balancing, the kDN distribution of all datasets have smaller differences between different classes. However, hardness bias still exist after class balancing. For example, $\Gamma_{race}(1)$ of the *COMPAS* dataset only decreases from 0.056 to 0.051 (*cf.*, Table 1).

In summary, all class balancing techniques can decrease the hardness bias but increase the distribution bias comparing to the original datasets.

Algorithm 1: Fair Class Balancing

Data: Original dataset $D = \{d_1, d_2, ..., d_n\}$,
Number of nearest neighbors k
Result: Balanced dataset $\hat{D}$
Clusters $C = \{c_1, c_2, ..., c_m\}$,
Silhouette scores $S = \{s_1, s_2, ..., s_n\}$ ⟵ clustering method M
Threshold $\theta = Q_1(S)$.
for $c_i \in C$ **do**
  $\tilde{c_i} = \{d_j \in c_i | s_j > \theta\}$
  Sampling Count ⟵majorityCount ($\tilde{c_i}$) - minorityCount ($\tilde{c_i}$);
  y_{min} = Minority class in $\tilde{c_i}$.
  if *Sampling Count > 0* **then**
    $nn = \{\}$
    for $d_j \in \{x \in \tilde{c_i} | y = y_{min}\}$ **do**
      $knn_j \longleftarrow k$-nearest neighbors of d_j;
      ▷ k-nearest neighbors are not limited to the minority class
      $nn \longleftarrow nn \cup knn_j$.
    generatedSamples ⟵ Generator (Sampling Count, nn);
    $\hat{c_i} \longleftarrow \tilde{c_i} \cup$ generatedSamples;
$\hat{D} = \{\hat{c_1}, \hat{c_2}, ..., \hat{c_m}\}$.

4.2 Effects on Group Fairness

Table 1 shows how the class balancing techniques impact the model performance on both utility and group fairness metrics. For all datasets, class balancing techniques increase the classification accuracy of the minority class (i.e., Acc. in Table 1). Generally, oversampling strategies perform better than under-sampling. In terms of fairness metrics, class balancing methods increase the discrimination between different sensitive groups for all datasets. Class balancing has less of an impact on fairness for the *COMPAS* dataset: almost all metrics have the same range as prior to the balancing.

Comparing with different class balancing methods, KMeans-SMOTE has the less negative impact on the fairness metrics.

5 FAIR CLASS BALANCING

To address the challenges posed by class balancing, and its effect on model fairness illustrated above, we here propose a new strategy, named the *fair class balancing* method. It is worth noting that our method not only enhances the fairness of the balancing approach while preserving prediction accuracy, but also paves the way for a pre-processing strategy to improve fairness when *sensitive attributes are unobserved.*

5.1 Proposed Method

We propose a *cluster-based* balancing method, named *fair class balancing*, that is guided by the *group structure* of the data, that is the natural occurrence of homogeneous subgroups with shared feature similarities, which can be identified via clustering in the feature space. Algorithm 1 provides a formal description of the proposed method.

Similarly to K-Means SMOTE, *fair class balancing* includes three steps: clustering, filtering, and oversampling. We re-design the filtering and oversampling steps to incorporate fairness constraints. An intuition of the proposed method is as follows:

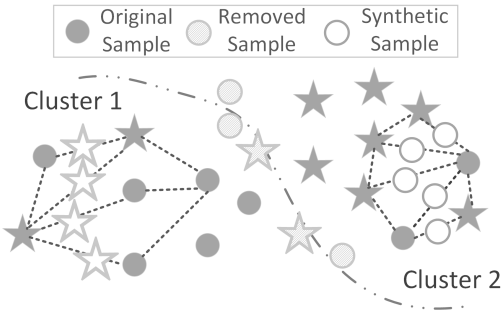

Figure 2: Illustration of how *fair class balancing* enhances fairness. Circle and star nodes represent samples with positive and negative labels, respectively. The hollow nodes are samples generated based on the nearest neighbors of the minority samples in each cluster.

1. Split the data into clusters according to a clustering algorithm M of choice, yielding samples in each cluster c_i that share similarity in the feature space. Then calculate the silhouette score s_j of each sample. We run the clustering algorithm by varying the number m of clusters, and for each parameter instantiation we use the average silhouette score to determine the goodness of clustering. Hence, we chose the best number of cluster m accordingly.
2. Remove the samples close to the cluster boundaries based on silhouette score. The samples with the lowest 25% (lower quartile) silhouette scores are filtered out from the data set.
3. For each cluster c_i, the minority class is the class that have less than half samples within the cluster. New samples of the minority class are generated based on the k-nearest neighbors of the minority samples, following the SMOTE [5] generation algorithm.

The proposed method above has a similar pipeline to *K-Means SMOTE* [13]. K-Means SMOTE generates the new samples based on the nearest neighbors within the minority class, while the nearest neighbors in our oversampling process are not limited to the minority samples. Figure 2 illustrate how *fair class balancing* works. In addition to clustering, we also filter out the samples that are close to the cluster boundaries. The samples close to the cluster boundary are difficult to separate in the feature space, which means that they have higher probabilities to be falsely assigned to one of the clusters. Thus, generating new instances based on those samples may further increase the distribution bias.

The strategy of *fair class balancing* is similar to *Preferential Sampling* [18], aiming to avoid the discrimination from borderline samples, i.e., samples close to the decision boundary are more likely to be discriminated against (or favored). In section §5.2, we will show how this approach enhances model fairness via balancing.

5.2 Biases after Fair Class Balancing

Table 2 compares the biases of original dataset, the dataset after *fair class balancing*, and the dataset after KMeans SMOTE balancing (which has the best fairness performance according to Table 1). Although traditional class balancing methods can also reduce the

Dataset	Methods	Metrics					Biases	
		F1	Acc.	E. Opp.	E. Odds	SP	$\Delta(y=1)$	$\Gamma(y=1)$
Adult	Original	**0.85±0.006**	0.57±0.04	**-0.12±0.04**	**-0.09±0.02**	**-0.17±0.01**	**0.20±0.01**	0.26±0.01
	RUS	0.79±0.01	0.87±0.01	-0.13±0.05	-0.17±0.06	-0.32±0.06	0.31±0.03	**0.02±0.03**
	CC	0.75±0.01	**0.89±0.01**	-0.14±0.03	-0.21±0.03	-0.37±0.03	0.29±0.02	0.05±0.02
	ROS	0.81±0.02	0.82±0.07	-0.14±0.03	-0.20±0.03	-0.36±0.04	0.29±0.02	0.08±0.01
	SMOTE	0.81±0.02	0.82±0.07	-0.15±0.05	-0.19±0.06	-0.33±0.03	0.32±0.03	0.03±0.02
	KMeans-SMOTE	0.82±0.006	0.56±0.02	-0.14±0.03	-0.12±0.02	-0.19±0.01	0.28±0.03	0.11±0.03
COMPAS	Original	**0.67 ±0.01**	0.52 ±0.04	**-0.18±0.03**	**-0.13±0.02**	**-0.15±0.01**	**0.09±0.01**	0.05±0.01
	RUS	0.66±0.01	0.64±0.02	-0.18±0.04	-0.14±0.02	-0.16±0.02	0.10±0.02	0.05±0.01
	CC	0.64±0.01	**0.66±0.02**	-0.21±0.03	-0.17±0.02	-0.19±0.02	0.11±0.02	0.06±0.01
	ROS	0.52±0.04	0.65±0.02	-0.18±0.03	-0.14±0.02	-0.17±0.02	0.11±0.03	0.05±0.02
	SMOTE	0.66±0.01	0.65±0.02	-0.18±0.03	-0.14±0.02	-0.17±0.02	0.10±0.03	0.05±0.02
	KMeans-SMOTE	0.67±0.01	0.60±0.02	-0.20±0.04	-0.15±0.02	-0.17±0.02	0.09±0.02	**0.01±0.02**
Violent Crime	Original	**0.84±0.01**	0.16±0.02	**-0.10±0.05**	**-0.05±0.02**	**-0.03±0.01**	0.06±0.02	0.13±0.03
	RUS	0.71±0.01	0.66±0.04	-0.21±0.08	-0.16±0.04	-0.15±0.02	0.10±0.02	0.09±0.03
	CC	0.41±0.02	**0.81±0.04**	-0.11±0.07	-0.11±0.03	-0.12±0.03	0.05±0.02	0.05±0.02
	ROS	0.72±0.01	0.64±0.02	-0.20±0.08	-0.15±0.04	-0.14±0.03	0.13±0.03	0.02±0.03
	SMOTE	0.74±0.01	0.47±0.04	-0.11±0.09	-0.08±0.04	-0.08±0.02	0.24±0.03	0.19±0.02
	KMeans-SMOTE	0.83±0.01	0.37±0.04	-0.16±0.07	-0.10±0.04	-0.09±0.02	**0.02±0.03**	**0.02±0.02**

Table 1: Effects of class balancing techniques on group fairness. For F1 and the accuracy of minority class (Acc.), the higher the better. For *Equal Opportunity* (E. Opp.), *Equal. Odds* (E. Odds), *Statistical Parity* (SP), and biases metrics, values close to zero indicate fairer outcomes.

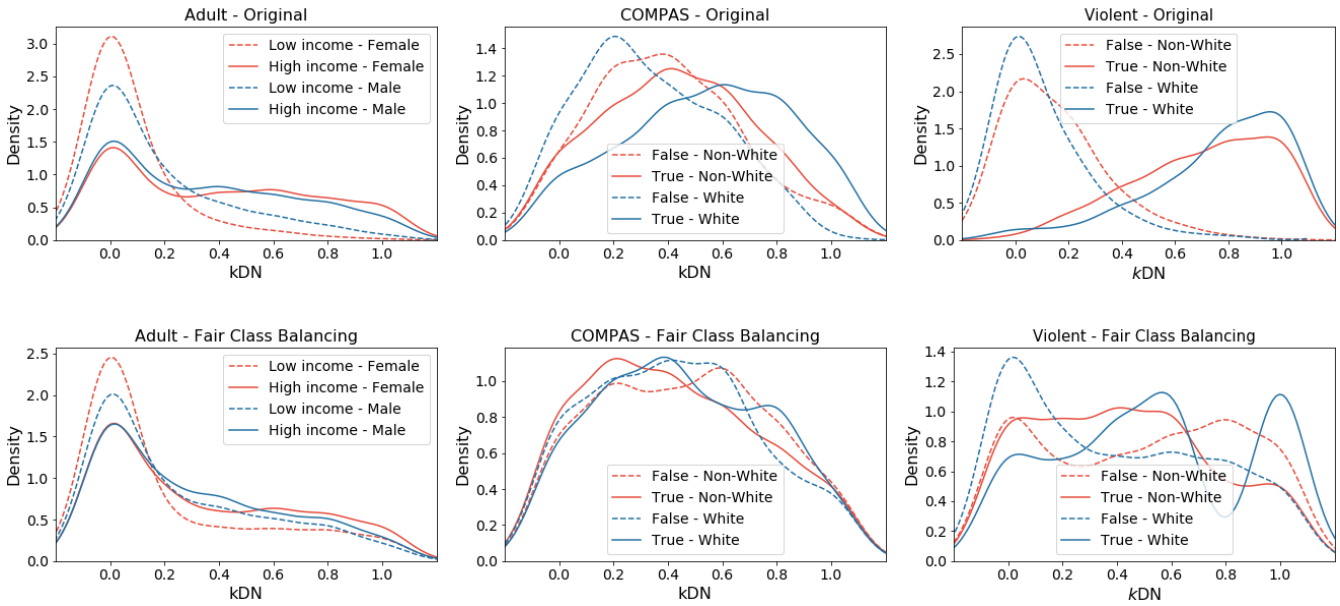

Figure 3: Examples of the hardness bias changes in test set before and after class balancing. The kDN for each test sample here is the percentage of the k nearest training samples that do not share the same target variable as the test sample.

hardness bias, our proposed method yield more balanced hardness distribution across sensitive groups, and prevent the increase of distribution bias. Additionally, in real applications, traditional class balancing techniques aim to balance the population differences between the target classes in the training set while not considering the potential biases in the test set.

Assume that both training set and test set are randomly sampled from the same data source and follow the same distribution. As we previously mentioned, samples close to the decision boundary are more likely to be discriminated against (or favored). Those borderline samples exist in both training set and test set. In traditional

(1) *Adult* Dataset

Income	$\Delta_{gender}(y)$			$\Gamma_{gender}(y)$		
	Orig.	KMeans SMOTE	Ours	Orig.	KMeans SMOTE	Ours
High (Y=1)	0.20	0.28	**0.09**	0.035	**0.022**	0.030
Low (Y=0)	-0.20	-0.28	**-0.09**	0.089	0.045	**0.043**

(2) *COMPAS* Dataset

Recidivism	$\Delta_{race}(y)$			$\Gamma_{race}(y)$		
	Orig.	KMeans SMOTE	Ours	Orig.	KMeans SMOTE	Ours
True (Y=1)	-0.09	-0.09	**-0.08**	0.056	0.018	**0.006**
False (Y=0)	0.09	0.09	**0.08**	0.034	0.009	**0.001**

(3) *Violent Crime* Dataset

Recidivism	$\Delta_{race}(y)$			$\Gamma_{race}(y)$		
	Orig.	KMeans SMOTE	Ours	Orig.	KMeans SMOTE	Ours
True (Y=1)	-0.06	**-0.02**	-0.05	0.131	0.088	**0.018**
False (Y=0)	0.06	**0.02**	0.05	0.035	0.016	**0.016**

Table 2: Bias measurement. Difference in bias of original data, after KMeans SMOTE class balancing, and after our proposed balancing method, as captured by the two metrics we propose, i.e., *distribution bias* Δ, and *hardness bias* Γ.

class balancing process, some of the borderline samples in the training set are oversampled based on the target variables' distribution, which makes the borderline samples in the test set are more likely to be classified as one of the class. Thus, the biases of the model outcomes are amplified.

To address the issue, *fair class balancing* generate new samples based on all neighboring samples, aiming to make sure that the borderline samples have similar number of neighbors in both classes, which in turn mitigates the biases in the predictions.

Figure 3 gives examples of the kDN distribution of the test sets. Each dataset is randomly split into 80% training set and 20% test set. The kDNs for test set here are calculated by locating the k nearest training samples for each test sample. Higher test kDN means that the samples are more likely to be falsely assigned to another class.

As the figure shown, samples from different classes have different kDN distribution. Among the three original datasets, the *Adult* dataset has lower test kDN with the mean of 0.18. The test kDN of the *COMPAS* dataset are more close to normal distribution with the mean of 0.41, indicating that most samples have similar probabilities to be assigned to both classes. Different from the first two datasets, the *Violent Crime* dataset shows a different pattern. Negative samples have significant higher kDN than positive samples in the *Violent Crime* dataset, which means negative samples are harder to predict accurately. In general, positive samples (i.e., the minority class) have higher test kDN than negative samples in all datasets, especially for the *Violent Crime* dataset. If one sensitive group has more negative samples, it would result the biases in model outcomes.

The kDN distribution differences are reduced after *fair class balancing*. For *Violent Crime* dataset, the mean values of test kDN between positive and negative samples are 0.72 and 0.13, respectively. After balancing, the mean values are 0.455 and 0.452. Regarding the sensitive groups, after balancing, the mean values of the test kDN change from 0.25 to 0.48 for non-white population and from 0.17 to 0.40 for white population, which also mitigates the biases in the original data.

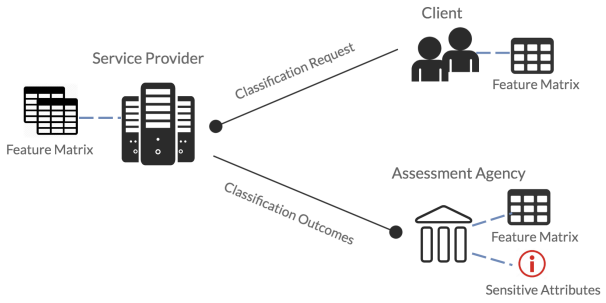

Figure 4: Prediction and assessment framework. In our setting, sensitive attributes are considered as unobservable to both the service provider and the client sides. Hence, fairness is judged by a third-party assessment agency.

6 EXPERIMENTS

6.1 Fairness Assessment Framework

This work focuses on scenarios where the sensitive attributes are not observable in both model training and prediction processes. To assess fairness in this scenario, we propose the *third-party fairness assessment framework* shown in Figure 4. The proposed framework has a similar structure to *black-box auditing* [1] which is designed to audit the biases of machine learning models without accessing the models' inside structure.

In our framework, sensitive attributes are unobservable for a *service provider* (i.e., the model training process) as well as on the *client side* (i.e., the testing process). Model fairness is evaluated by an *assessment agency* (e.g., government), which obtains both the feature matrix and sensitive attributes of the test set. This framework is practical in real-world applications and also provides a comparable setting with respect to the extant fairness research.

6.2 Experiment Settings

To illustrate the real fairness improvements of our proposed method and have a fair comparison with other strategies, we follow the *third-party fairness assessment framework* introduced in Section §6.1. All experiments are based on 80-20 dataset split with 50 randomized restart. All sensitive attributes are removed from the training and test sets.

We focus the following utility and fairness metrics in this work:

- **Utility:** F1-score (F1) and the classification accuracy of the minority class (Acc.).
- **Fairness:** *Equal Opportunity Differences* (E. Opp.), *Average Equalized Odds Differences* (E. Odds), and *Statistical Parity Differences* (SP).

We test our *fair class balancing* method with three widely-used clustering algorithms: *KMeans clustering*, *Agglomerative clustering* (Agg.), and *Spectral clustering* (Spec.). All metrics show similar performance for different classification algorithms (e.g., *Logistic*

Dataset	Method	Utility		Fairness		
		F1	Acc.	E. Opp.	E. Odds	SP
Adult	Baseline (Original data)	**0.84±0.003**	0.59±0.01	-0.08±0.03	-0.07±0.01	-0.17±0.006
	fair class balancing + KMeans (5-*nn*)	0.75±0.02	0.68±0.02	0.009±0.03	-0.005±0.01	-0.10±0.01
	fair class balancing + Agg. Clust. (5-*nn*)	0.75 ±0.01	0.71±0.04	**0.20±0.04**	**0.12±0.03**	**-0.03±0.03**
	fair class balancing + Spec. Clust. (5-*nn*)	0.72 ±0.02	**0.73±0.02**	0.02 ±0.04	0.02±0.02	-0.06±0.02
COMPAS	Baseline (Original data)	**0.67 ±0.01**	0.56±0.01	-0.20±0.04	-0.14±0.02	-0.16 ± 0.02
	fair class balancing + KMeans (5-*nn*)	0.62 ±0.02	**0.63 ±0.08**	-0.14 ±0.04	-0.11 ±0.03	-0.13 ±0.03
	fair class balancing + Agg. Clust. (5-*nn*)	0.64 ±0.01	0.56 ±0.03	**-0.13 ±0.04**	**-0.11 ±0.03**	**-0.13 ±0.02**
	fair class balancing + Spec. Clust. (5-*nn*)	0.64 ±0.01	0.61 ±0.09	-0.14 ±0.04	-0.11 ±0.03	-0.13 ±0.03
Violent Crime	Baseline (Original data)	**0.84 ±0.01**	0.16 ±0.02	-0.10±0.05	-0.06±0.02	-0.04±0.01
	fair class balancing + KMeans (5-*nn*)	0.65 ±0.02	0.44 ±0.06	-0.04 ±0.05	-0.02 ±0.05	-0.02 ±0.03
	fair class balancing + Agg. Clust. (5-*nn*)	0.61 ±0.01	**0.52 ±0.04**	**-0.01 ±0.08**	**-0.01 ±0.05**	**-0.02 ±0.03**
	fair class balancing + Spec. Clust. (5-*nn*)	0.63 ±0.02	0.52 ±0.05	-0.03 ±0.08	-0.04 ±0.05	-0.06±0.04

Table 3: Classification performance comparison among original dataset (Baseline) and *fair class balancing* with 5-NN parameters and different clustering algorithms: *KMeans clustering* (KMeans), *Agglomerative clustering* (Agg.), and *Spectral clustering* (Spec.). Results with the smallest bias against the unprivileged group (i.e., most positive results) are bolded.

Dataset	Method	Metrics				
		F1	Accuracy	E. Opp.	E. Odds	SP
Adult	Baseline	**0.84±0.003**	0.59±0.01	-0.08±0.03	-0.07±0.01	-0.17±0.006
	CEO [24]	0.81±0.0004	0.46±0.003	-0.09±0.004	-0.07±0.01	-0.13±0.006
	ROS + CEO	0.83±0.003	0.67±0.03	-0.13±0.007	-0.15±0.009	-0.27±0.01
	fair class balancing	0.75±0.02	**0.68±0.02**	0.009±0.03	-0.005±0.01	-0.10±0.01
	fair class balancing + CEO	0.79±0.002	0.54±0.008	**0.10±0.02**	**0.04±0.01**	**-0.08±0.01**
COMPAS	Baseline	**0.67 ±0.01**	0.56±0.01	-0.20±0.04	-0.14±0.02	-0.16 ± 0.02
	CEO [24]	0.65±0.01	0.48± 0.01	-0.17±0.03	-0.12±0.02	-0.14±0.02
	ROS + CEO	0.65±0.01	0.55±0.01	-0.17±0.03	-0.13±0.02	-0.16±0.02
	fair class balancing	0.63±0.01	**0.60±0.03**	-0.13 ±0.04	-0.11 ±0.03	-0.13 ±0.02
	fair class balancing + CEO	0.63±0.02	0.50±0.05	**-0.11±0.04**	**-0.09±0.02**	**-0.11±0.02**
Violent Crime	Baseline	**0.84 ±0.01**	0.16 ±0.02	-0.10±0.05	-0.06±0.02	-0.04±0.01
	CEO [24]	0.43±0.0004	0.06± 0.01	0.003±0.004	-0.01±0.003	-0.02±0.03
	ROS + CEO	0.52±0.001	0.22±0.005	-0.01±0.001	-0.02±0.001	-0.03±0.001
	fair class balancing	0.61 ±0.01	**0.52 ±0.04**	-0.01 ±0.08	-0.01 ±0.05	-0.02 ±0.03
	fair class balancing + CEO	0.50±0.01	0.17±0.03	**0.002±0.008**	**-0.006±0.005**	**-0.01±0.006**

Table 4: Combining class balancing and fairness-aware learning. Performance comparison among *fair class balancing*, proxy-based post-processing (CEO), and combining CEO with traditional class balancing and *fair class balancing*. Results with the smallest bias against the unprivileged group are bolded.

Regression, *Random Forest*, etc.). We only report the results of *Logistic Regression* due to the page limit.

6.3 Performance

Comparing to the performance of traditional class balancing techniques (shown in Table 1), our fair class balancing method can improve all the fairness metrics. In general, all clustering methods improve the performance. Especially for the *Adult* dataset, fair class balancing yields promising scores, e.g., E. Opp. is 0.009 (±0.01) and E. Odds equal to -0.005 (±0.01) with *KMeans clustering*. With *Agglomerative clustering*, the fairness metrics are further increased to 0.20 (±0.04) for E.Opp., which indicates that the biases against the unprivileged group are removed, and the model starts to favor those samples. Additionally, comparing to the baseline method, our fair class balancing method also yield higher accuracy for the minority class, which is a crucial metric in real-life applications.

In terms of the fairness metrics, different clustering methods show promising performance and *Agglomerative clustering* provides relatively better performance for all datasets.

Effects of fairness budget k-NN. In our method, we introduce the *fairness budget* parameter k-NN, which indicates the number of nearest neighbors for generating new samples. As k-NN increases, more noisy samples will be generated. This will further decrease the biases across sensitive groups. However, more noisy samples will decrease the model accuracy.

Figure 5 shows how different metrics change as a function of k-NN. The performances reported in Figure 5 are based on *fair class balancing* with K-Means clustering algorithm. All three datasets show increasing trend of all three fairness metrics when k-NN increase, indicating more biases against the unprivileged group are removed. As the trade-off, the model accuracy has a decreasing trend when k-NN increase.

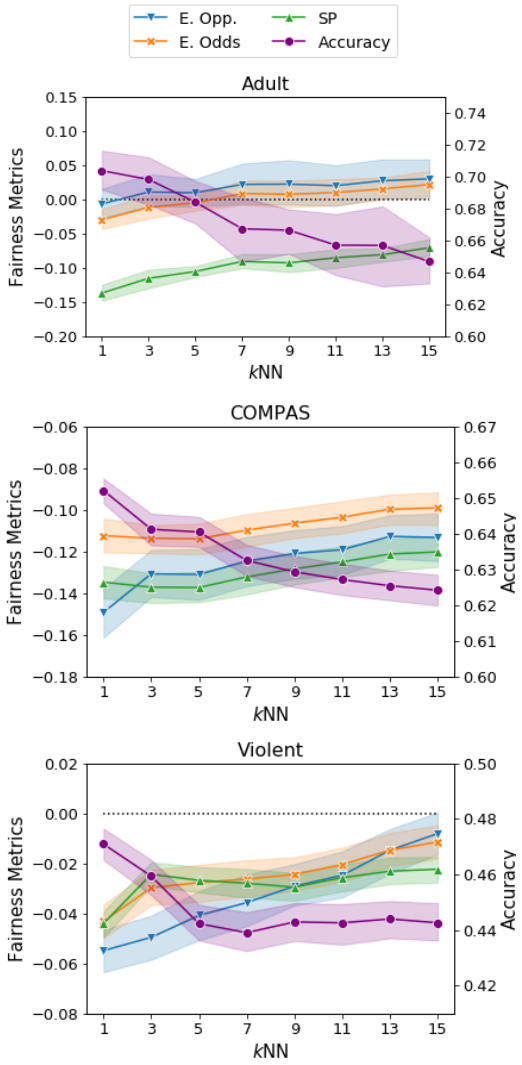

Figure 5: Effects of the fairness budget parameter kNN. The plots illustrate model performance and fairness, according to four metrics (*Equal Opportunities, Equalized Odds, Statistical Parity,* and *Accuracy*), for the *Adult* , *COMPAS*, and *Violent Crime* data.

6.4 Fair class balancing & Fairness-aware learning

Our proposed method modifies the training data, thus is amenable as a pre-processing step toward *fairness-aware model learning*, a set of strategies that aim at enhancing fairness in a post hoc fashion.

Next, we combine our proposed fair class balancing technique with a post-processing fairness-improving strategy to analyze how our method enables achieving even higher model fairness compared to the results presented earlier.

For our purposes, we use the *Calibrated Equalized Odds* (CEO) [24] as the post-processing method of choice. The algorithm is implemented by `AI Fairness 360 Open Source Toolkit` [2].

Since this work focus on the scenario where sensitive attributes are not observable, we follow the idea of using proxy features as the alternative to the true sensitive attributes [15]. We use "age" information to create proxy groups. "*Age* > 40" and "*age* <= 40" represent privileged and unprivileged groups, respectively.

Table 4 compares model performance when using different fairness-improving strategies. In general, our proposed *fair class balancing* always yields the most accurate outcomes. As a post-processing method, CEO optimizes on the E. Odds metric and scarifies the accuracy. Our experimental results show that combining *fair class balancing* and CEO can boost the accuracy of the results as well as achieve better fairness performance. Moreover, for the *Adult* and *COMPAS* dataset, just applying the fair class balancing technique can yield better performance in terms of E. Opp. and E. Odds compared to CEO. We also include the comparison of the combination of class balancing techniques and CEO, specifically the Random Over-Sampling (ROS) and our *fair class balancing*. Experimental results show that the combination with our method can further improve the fairness performance.

Overall, *fair class balancing* + CEO has the minimum biases against the unprivileged groups. Since CEO has negative impact on the model accuracy, considering the importance of model accuracy in real-life applications, our proposed *fair class balancing* can achieve the best balance between utility and fairness among all reported methods.

6.5 Discussion

According to the experimental results presented above, our proposed fair class balancing method can provide fairness improvements with unobserved sensitive attributes. In this section, we further discuss how different data properties impact the performance of our strategy.

As the results reported in Table 3, our proposed method may decrease the F1-score of the predictions, but improve the accuracy of the minority class and fairness metrics. Across all three datasets, the *COMPAS* dataset has the minimum improvements. The differences of the performance can be explained by the data biases exist in the original data. According to Figure 1, the *COMPAS* dataset has the most balanced sensitive groups, and the kDN distribution among all groups are similar. Our method is designed to mitigate the hardness bias and distribution bias, and thus the improvements are less significant than other two datasets.

7 CONCLUSIONS

Guaranteeing model fairness in data imbalanced settings is an open challenge in real-world machine learning applications. In this work, we investigated how class balancing techniques impact the fairness of model outcomes. Inspired by our findings showing how class balancing exacerbates unfairness, we proposed the *fair class balancing* method to enhance fairness, which also has the desirable property of assuming that the sensitive attributes are unobserved.

The proposed method aims to mitigate the biases come from the borderline samples, which are one of the main sources of the model unfairness according to the literature. *Fair class balancing* has a similar framework as cluster-based oversampling class balancing strategies except the synthetic samples are generated based on

the samples from both minority and majority classes. The cluster-based strategy identifies the real class imbalance within samples with similar feature space, and also effectively avoids generating too noisy samples. Furthermore, by generating synthetic samples based on both classes, our method ensures that unseen borderline samples in the test sets have similar probabilities to be assigned to both classes, which reduce the biases in model outcomes.

Experimental results on real-world datasets show that our *fair class balancing* method can improve all three fairness metrics of interest as well as the accuracy of the minority class. The improvements are not limited to the clustering algorithm of choice. *Fair class balancing* with different clustering algorithms all yield more fair predictions. As the fairness budget parameter kNN increase, the biases against the unprivileged groups are decreased with the accuracy decrease as the trade-off. With KMeans clustering and fairness budget parameter $k = 5$, the E. Odds metrics on the *Adult*, *COMPAS*, and *Violent Crime* datasets have 92%, 21%, and 66% improvements, respectively. Our method also can be used as the pre-processing step for other fairness-aware mechanisms, further improving both fairness and accuracy.

With the increasing concerns on data privacy, more and more real-life applications no longer collect sensitive attributes due to legal restrictions, which also limits the application of traditional fairness-aware algorithms. As a class balancing technique without using any information about sensitive attributes, *fair class balancing* can still be widely applied to data-driven decision making systems.

In this work, we discuss the performance of *fair class balancing* with three clustering algorithms and basic classification models (e.g., *Logistic Regression*, *Random Forest*, etc.). For future work, we will explore the performance of our method in more scenarios. Experiments on different datasets will be conducted to assess the performance of the methodology under explicit non-linearities in the data/label space. We will also study how the choice of algorithms affect performance in different scenarios, for example, the performance with more sophisticated models (e.g., deep learning models), the effects of different clustering algorithms with different data properties, and the optimal parameter choice in different applications. We also aim to take *subgroup fairness* into consideration, an *intersectional approach* where subgroups are jointly defined over multiple sensitive attributes (e.g., "black" and "female"). In real-life applications, data balancing also exhibits in continuous target variables. We will extend this work to regression tasks.

ACKNOWLEDGMENTS

This work is partly supported by DARPA grant no. D16AP00115.

REFERENCES

[1] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54, 1 (2018), 95–122.

[2] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[3] Shikha Bordia and Samuel Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 7–15.

[4] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[6] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *Advances in Neural Information Processing Systems*. 3539–3550.

[7] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 339–348.

[8] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[9] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

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

[11] Marc N Elliott, Allen Fremont, Peter A Morrison, Philip Pantoja, and Nicole Lurie. 2008. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research* 43, 5p1 (2008), 1722–1736.

[12] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[13] D Georgios, B Fernando, and L Felix. 2018. Oversampling for imbalanced learning based on K-means and SMOTE. *Inf. Sci.* 465 (2018), 1–20.

[14] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[15] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *arXiv preprint arXiv:1806.11212* (2018).

[16] Sara Hajian and Josep Domingo-Ferrer. 2012. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* 25, 7 (2012), 1445–1459.

[17] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[18] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[19] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[20] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[21] Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 247–254.

[22] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. 2019. Fair k-center clustering for data summarization. *arXiv preprint arXiv:1901.08628* (2019).

[23] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).

[24] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[25] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. 2014. An instance level analysis of data complexity. *Machine learning* 95, 2 (2014), 225–256.

[26] Show-Jane Yen and Yue-Shi Lee. 2009. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications* 36, 3 (2009), 5718–5727.

[27] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[28] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 335–340.