FABBOO - Online Fairness-aware Learning under Class Imbalance

Vasileios Iosifidis[1] and Eirini Ntoutsi[1]

L3S Research Center, Leibniz University of Hannover, Germany
{iosifidis, ntoutsi}@l3s.de

Abstract. Data-driven algorithms are employed in many applications, in which data become available in a sequential order, forcing the update of the model with new instances. In such dynamic environments, in which the underlying data distributions might evolve with time, fairness-aware learning cannot be considered as a one-off requirement, but rather it should comprise a continual requirement over the stream. Recent fairness-aware stream classifiers ignore the problem of class distribution skewness. As a result, such methods mitigate discrimination by "rejecting" minority instances at large due to their inability to effectively learn all classes. In this work, we propose FABBOO, an online fairness-aware approach that maintains a valid and fair classifier over a stream. FABBOO is an online boosting approach that changes the training distribution in an online fashion based on both stream imbalance and discriminatory behavior of the model evaluated over the historical stream. Our experiments show that such long-term consideration of class-imbalance and fairness are beneficial for maintaining models that exhibit good predictive- and fairness-related performance.

Keywords: data streams · fairness-aware classification · class-imbalance

1 Introduction

Data-driven decision support systems have become a necessity nowadays for many applications where huge amounts of historical data are available for analysis. Their performance in many tasks is comparable or has even surpassed human performance [15] and therefore, for many processes, human decisions are substituted by algorithmic ones. Such a replacement, however, has raised a lot of concerns [4] regarding the fairness, accountability and transparency of such methods in domains of high societal impact such as *risk assessment*, *recidivism*, *predictive policing*, etc. For example, Google's *AdFisher* online recommendation tool showed significantly more highly paid jobs to men than women [10]. Many similar incidents of algorithmic unfairness have been reported in recent years [1, 18, 26].

As a result of the ever-increasing interest in issues of fairness and responsibility of data-driven systems, a large body of work exists already in fairness-aware learning [17, 19–21, 23–25, 31]. Only a few recent works, however, investigate the problem of fair learning in non-stationary environments [22, 30]. Nonetheless, these methods ignore an important aspect of the learning problem, namely that the majority of (streaming) datasets suffer from class-imbalance. Class imbalance refers to the disproportion among

classes i.e., when one class, called *minority* class, has significantly fewer examples than another class, called *majority* class. If the imbalance problem is not tackled, the learner mainly learns the majority class and strongly misclassifies/rejects the minority. Such methods might appear to be fair for certain fairness definitions that rely on parity in the predictions between the protected and non-protected groups. In reality though the low discrimination scores are just an artifact of the low prediction rates for the minority class. This observation has been made in [21] but for the static case. We observe the same issue for the streaming case and propose an imbalance monitoring mechanism based on which we adapt the weighted training distribution.

Moreover, in a stream environment the decisions do not only have a short-term effect, but rather they might incur long-term effects. In case of discrimination, this means that discriminatory model decisions affect not only the immediate outcomes, but they might also affect future outcomes [9]. For example, [9] indicates small wage gaps between college-educated blacks and whites when they are first hired, but the pay gap increased over the years. To this end, we propose to define discrimination cumulatively over the stream rather than based only on the most recent outcomes. This is in contrast to recent stream fairness-aware approaches that focus only on short term outcomes, e.g., [22]. Our experiments verify that when treating for *short-term* discriminatory outcomes, the *cumulative* effects can be substantially higher over time and therefore, a cumulative approach is better.

Our contributions are summarized as follows: i) we propose FABBOO, a fairness and class imbalance-aware boosting method that is able to tackle class-imbalance as well as mitigate different parity-based discriminatory outcomes, ii) we introduce the notion of *cumulative fairness* in streams, which accounts for cumulative discriminatory outcomes, iii) our experiments, in a variety of real-world and synthetic datasets, show that our approach outperforms existing approaches that either do not consider class-imbalance or are based on short-term fairness evaluation.

2 Basic Concepts and Problem Definition

Let X be a sequence of instances $x_1, x_2, \cdots$, arriving over time at timepoints $t_1, t_2, \cdots$, where each instance $x \in \mathbb{R}^d$. Similarly, let y be a sequence of corresponding class labels, such that each instance in X has a corresponding class label in y. Without loss of generality, we assume a binary classification problem, i.e., $y = \{+1, -1\}$, and we denote by y^+ (y^-) the positive (negative, respectively) segments. We denote the classifier by $f : X \rightarrow y$. We follow the *online learning setting*, where new instances from the stream are processed one by one. For each new instance x arriving at t, its class label $f_{t-1}(x)$ is predicted by the current model f_{t-1}. The true class label of the instance is revealed to the learner before the arrival of the next instance, and it is used for model updating, thus resulting into the updated model f_t. This setup is known as first-test-then-train or prequential evaluation [14].

We assume that the underlying stream distribution is *non-stationary*, that is, the characteristics of the stream might change with time leading to *concept drifts*, i.e., changes in the joint distribution so that $P_{t_1}(X, y) \neq P_{t_2}(X, y)$ for two different timepoints t_1 and t_2. We are particularly interested in real concept drifts, that is when

$P_{t_1}(y|X) \neq P_{t_2}(y|X)$, as such changes make the current classifier obsolete and call for model update. Moreover, we consider the scenario where the stream population is *imbalanced*, that is, one of the classes dominates the stream impacting the learning ability of the classifiers that traditionally tend to ignore the minority to foster generalization and avoid overfitting [29]. We, however, do not require that the minority class is predefined and fixed over the course of the stream. Instead, we assume that this role might alternate between the two classes.

We also assume the existence of a sensitive feature *SA*, e.g., gender or race, which is binary with values $SA = \{z, \overline{z}\}$, e.g., gender=$\{$female, male$\}$; we refer to z, $\overline{z}$ as protected, non-protected group respectively [1]. Traditional *fairness-aware classification* aims to learn a mapping $f : X \rightarrow y$ that accurately maps instances x to their correct classes without discriminating between the protected and non-protected groups. The discrimination is assessed in terms of some fairness measure. Formalizing fairness is a hard topic per se, and there has already been a lot of work in this direction. For example, [27] overview more than twenty measures of fairness; however, there is no clear indication which measure is the most appropriate for classification tasks. In this work, we investigate *parity-based* notions of fairness such as the well-known *statistical parity* [23] and *equal opportunity* [16]; however, FABBOO can accommodate various parity-based fairness notions such as *disparate mistreatment* [31], *predictive quality* [27], and so on.

Statistical parity (S.P.) measures the difference in the probability of a random individual drawn from $\overline{z}$ to be predicted as positive and the probability of a random individual drawn from the complement z to be predicted as positive:

$$S.P. = P(f(x) = y^+|\overline{z}) - P(f(x) = y^+|z) \tag{1}$$

The S.P. values lie in the [-1, 1] range, with 0 meaning the decision does not depend on the sensitive value (aka fair), 1 meaning that the protected group is totally discriminated (aka discrimination), and -1 that the non-protected group is discriminated (aka reverse discrimination).

S.P. does not take into account the real class labels, and therefore may allow individuals to be assigned to the positive class, even though they do not satisfy the requirements, thus causing *reverse discrimination*. Equal opportunity (EQ.OP.) resolves this issue by measuring the difference in the True Positive Rates (TPR) between the two groups, i.e.,:

$$EQ.OP. = P(f(x) = y^+|\overline{z}, y^+) - P(f(x) = y^+|z, y^+) \tag{2}$$

Similar to S.P., EQ.OP's values lie in the [-1, 1] range.

Our work investigates the problem of fair classification in a stream environment. *Fairness-aware stream learning* refers to the problem of maintaining a valid and fair classifier over the stream. The term *valid* refers to the ability of the model to adapt to the underlying evolving population and deal with concept drifts. At the same time, the classifier should be fair according to the adopted S.P. or EQ.OP. fairness measures. Ensuring fairness is much harder in such an online environment comparing to the traditional batch setting. First, the model should be continuously updated to reflect the underling non-stationary population. The typically accuracy-driven update of the model

[1] SA definition could also be extended to cover feature combinations such as race and gender

cannot ensure fairness, so even if the initial model was fair, its discriminatory behavior might get affected by the model updates. Second, small amounts of unfairness at each time point might accumulate into significant discrimination as the learner typically acts as an amplifier of whatever biases exist in the data and furthermore, reinforces its errors. So, model update should consider fairness constraints and long term effects of discrimination beyond the point of its evaluation.

3 Related Work

Static Fairness-Aware Learning: Static fairness-aware approaches have received a lot of attention over the recent years. Literature in this area can be categorized in: i) pre-processing methods [5, 20, 23], where data are processed, transformed, or augmented to reduce discrimination or remove the correlation between various attributes and the sensitive attribute. ii) In-processing methods [21, 25, 31] focus on facilitating a fairness notion into a model's objective function. iii) Post-processing methods [12, 16, 19] alter a model's predictions or adjust a model's decision boundary to reduce unfairness.
Stream Fairness-Aware Learning: Stream fairness-aware approaches aim to remove unfair outcomes when data are presented sequentially. In [22], authors present a chunk based stream classification approach in which they apply pre-processing methods, such as *label swapping*, to remove discrimination, which is measured by statistical parity, from data before updating an online classifier; however, this approach accounts for short-term outcomes. In [30], they incorporate the notion of statistical parity into Hoeffding's Tree split criterion so that it accounts for cumulative discriminatory outcomes.
Stream Learning: In stream learning, data arrive sequentially and their distributions can change over time, the so-called concept drifts [14]. Concept drifts can be handled explicitly through *informed adaptation*, where the model adapts only if a change has been detected, or implicitly through *blind adaptation*, where the model is updated constantly to account for changes in the underlying data distributions. In addition, models developed for stream learning are categorized as *incremental* and *online* [28]. Incremental models are trained in batches [13], with the help of a chunk (window), while online models are updated continuously to accommodate newly incoming examples [7].

The goal of this paper is to highlight the importance of class-imbalance problem in fairness-aware stream learning; therefore, we select as competitors fairness-aware stream learners [22, 30] and omit class-imbalance stream learners.

4 Online Fairness- and Class Imbalance-aware Boosting

An overview of FABBOO, standing for online fairness and class imbalance-aware boosting, is shown in Figure 1. Our method consists of a *class-imbalance monitoring component* that keeps track of the class ratios over the stream and adjusts the weights of the new training instances accordingly to ensure that the learner properly learns both classes (Section 4.1), while adapting to concept drifts via *blind model adaptation* [14]. In addition, the *cumulative* discriminatory behavior of the learner is monitored, and when it exceeds a user-defined tolerance threshold ϵ, the decision boundary is adjusted to ensure that the learner does not incur discrimination (Section 4.2).

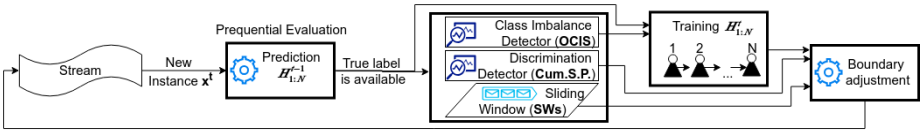

Fig. 1. An overview of FABBOO

4.1 Online Monitoring of Class Imbalance and Model Update

In evolving data streams, the role of minority and majority classes can exchange and what is now considered to be minority might turn later into a majority class or vice versa [28]. Knowing the class ratio over the stream is important for our method as it directly affects the instance weighting during training. Therefore, we keep track of the stream imbalance using the online class imbalance monitor (OCIS) of [28].

$$OCIS_t = W_t^+ - W_t^- \tag{3}$$

where W_t^y is the percentage of class y at timepoint t maintained in an online fashion. In particular, upon the arrival of a new instance x at timepoint t, the percentage of a class y is updated as follows:

$$W_t^y = \lambda \cdot W_{t-1}^y + (1-\lambda) \cdot \mathbb{I}[(y_t, y)] \tag{4}$$

where $\lambda \in [0,1]$ is a user-defined decay factor that controls the extent to which old class percentage information should be considered, and $\mathbb{I}[(y_t, y)]$ is an identity function which equals to 1 if the true class label of x_t is y, otherwise 0.

The imbalance index OCIS takes values in the $[-1,1]$ range, with 0 indicating a perfectly balanced stream and -1 or 1 indicating the total absence of one class.

Model adaptation: Our basic model is OSBoost [7] that generates smooth distributions over the training instances, and guarantees to achieve small error if the number of weak learners and training instances is large enough. We extend OSBoost to take into account class imbalance by changing the weighted instance distribution so that minority instances become more prominent during the training process.

The pseudocode of the algorithm is shown in Algorithm 1. OSBoost comes with a set of predefined parameters: $\gamma \in [0,1]$ that is an online analog of the "edge" of the weak learning oracle, and $N \in \mathbb{Z}^+$ that is the number of online weak learners. Upon the arrival of a new instance x at timepoint t, the class imbalance status is updated (line 2) according to Equation 3. Then, the weak learners are updated sequentially (lines 4-11) so that the predictions of model H_i^t (line 6) affect the training of its successor model H_{i+1}^t by changing the weight/contribution of instance x to the model accordingly. The weight of instance x is tuned per learner H_i^t based on the error of the predecessor model H_{i-1}^t on x, but also based on current class imbalance (lines 8-11).

To summarize, traditional OSBoost performs error-based instance weight tuning but does not adjust for class-imbalance. On the contrary, FABBOO adjusts the instance weights also based on the dynamic class ratio (c.f. Equation 3) so that minority instances receive extra "boosting" during training. Note that if the stream is balanced, i.e., $W_t^+ - W_t^- \approx 0$, the weights are only slightly affected.

Algorithm 1 FABBOO training procedure

procedure TRAIN($x_t, y_t, \gamma, H^{t-1}_{1:N}$) ▷ x_t: newly arrived instace, y_t: label of x_t, γ: learning rate, $H^{t-1}_{1:N}$: current ensemble
  $OCIS_t = W^+_t - W^-_t$ ▷ Update the class imbalance status
  $w_1 = 1, q_0 = 0$
  for $i = 1$ to N **do**
    Train H^t_i on x_t with weight w_i
    $q_i = q_{i-1} + y_t \cdot H^t_i(x_t) - \frac{\gamma}{2+\gamma}$
    $w_{i+1} = min\{(1-\gamma)^{q_i/2}, 1\}$
    if $x_t \in y_+ \; and \; OCIS_t < 0$ **then** ▷ y^+ is minority at timepoint t
      $w_{i+1} = \frac{w_{i+1}}{1+OCIS_t}$
    if $x_t \in y^- \; and \; OCIS_t > 0$ **then** ▷ y^- is minority at timepoint t
      $w_{i+1} = \frac{w_{i+1}}{1-OCIS_t}$
  return updated ensemble $H^t_{1:N}$

4.2 Online Monitoring of Cumulative Fairness and Boundary Adjustment

Methods which restore fairness only on *short-term* (recent) outcomes fail to mitigate discrimination over time as discrimination scores that might be considered negligible when evaluated individually (i.e., at a single time point) might accumulate into significant discrimination in the long run [9]. In this work, we aim to mitigate cumulative discrimination accumulated from the beginning of the stream in order to remove such long term discriminatory effects and adjust the decision boundary not only based on the recent behavior of the model, but rather on its historical performance.

Cumulative fairness monitoring accounts for discriminatory outcomes from the beginning of the stream until time point t. We introduce the cumulative fairness notion for non-stationary environments w.r.t. statistical parity and equal opportunity as follows:

Definition 1. *Cumulative Statistical Parity (Cum.S.P.)*

$$\frac{\sum_{i=1}^{t} 1 \cdot \mathbb{I}[f_i(x_i) = y^+ | x_i \in \bar{z}]}{\sum_{i=1}^{t} 1 \cdot \mathbb{I}[x_i \in \bar{z}] + l} - \frac{\sum_{i=1}^{t} 1 \cdot \mathbb{I}[f_i(x_i) = y^+ | x_i \in z]}{\sum_{i=1}^{t} 1 \cdot \mathbb{I}[x_i \in z] + l}$$

Definition 2. *Cumulative Equal Opportunity (Cum. EQ.OP.)*

$$\frac{\sum_{i=1}^{t} 1 \cdot \mathbb{I}[f_i(x_i) = y^+ | x_i \in \bar{z}, y^+_i]}{\sum_{i=1}^{t} 1 \cdot \mathbb{I}[x_i \in \bar{z}, y^+_i] + l} - \frac{\sum_{i=1}^{t} 1 \cdot \mathbb{I}[f_i(x_i) = y^+ | x_i \in z, y^+_i]}{\sum_{i=1}^{t} 1 \cdot \mathbb{I}[x_i \in z, y^+_i] + l}$$

where parameter l is employed for correction in the early stages of the stream. Cum.S.P. or Cum.EQ.OP. are maintained online using incremental counters updated with the arrival of new instances from the stream, and therefore, it is appropriate for stream applications where typically random access to historical stream instances is not possible.

The cumulative fairness notions are employed by FABBOO for discrimination monitoring. When their values exceed a user-defined discrimination tolerance threshold ϵ, the decision boundary should be adjusted i.e., $Cum.S.P. > \epsilon$ or $Cum.EQ.OP. > \epsilon$.

Decision boundary adjustment: Post-processing adjustment of the decision boundary for discrimination elimination has been investigated in the literature, e.g., [12, 16]. Closer to our approach is [12], where the authors adjust the decision boundary of an AdaBoost classifier based on the (sorted) confidence scores of misclassified instances of the protected group. However, in contrast to [12], we deal with stream classification, and therefore, we do not have access to historical stream instances in order to adjust the boundary accurately. Except for the access-to-the-data constraint, another reason for not considering the whole history for the adjustment of the boundary is the non-stationary nature of the stream. In such a case, adjusting the boundary based on the whole history of the stream will hinder the ability of the model to adapt to the underlying data and will eventually hurt predictive performance.

To overcome this issue, we use a sliding window model of a pre-defined size M for the adjustment. In particular, we maintain a sliding window of size M for each segment to allow for boundary adjustment for different parity-based notions based on each discriminated segment. In the case of statistical parity or equal opportunity, the only relevant sliding window is the one for the protected positive segment (denoted by SW_z^+). The number of examples (n_t) which are needed in order to mitigate discrimination at timepoint t is given by:

$$n_t = \left\lfloor \sum_{i=1}^{t} 1 \cdot \mathbb{I}[x_i \in z] \cdot \frac{\sum\limits_{i=1}^{t} 1 \cdot \mathbb{I}[f_i(x_i) = y^+ | x_i \in \bar{z}]}{\sum\limits_{i=1}^{t} 1 \cdot \mathbb{I}[x_i \in \bar{z}]} - \sum_{i=1}^{t} 1 \cdot \mathbb{I}[f_i(x_i) = y^+ | x_i \in z] \right\rfloor \quad (5)$$

Similar to statistical parity, to estimate the number of examples (n_t) for equal opportunity, we follow the same logic:

$$n_t = \left\lfloor \sum_{i=1}^{t} 1 \cdot \mathbb{I}[x_i \in z, y_i^+] \cdot \frac{\sum\limits_{i=1}^{t} 1 \cdot \mathbb{I}[f_i(x_i) = y^+ | x_i \in \bar{z}, y^+]}{\sum\limits_{i=1}^{t} 1 \cdot \mathbb{I}[x_i \in \bar{z}, y_i^+]} - \sum_{i=1}^{t} 1 \cdot \mathbb{I}[f_i(x_i) = y^+ | x_i \in z, y^+] \right\rfloor \quad (6)$$

Afterwards, the misclassified instances in SW_z^+ are sorted based on the confidence scores in a descending order. The decision boundary is adjusted according to the n^t-th instance of the sorted window (SW_z^+). In particular, if θ^{t-1} is the decision boundary value (original value θ^0 is 0.5) of the n_{t-1}-th, the fair-boundary is adjusted to θ^t. Note that in the early stage of the stream, where the sliding window does not contain a sufficient number of instances, the boundary is tweaked based on the misclassified instance with the highest confidence within the window.

4.3 FABBOO Classification

FABBOO is an online ensemble of sequential weak learners that tackles class imbalance and cumulative discriminatory outcomes in the stream. Moreover, FABBOO deals with

concept drifts, through *blind adaptation*, by employing a base learner that is able to react to concept drifts. In particular, we employ Adaptive Hoeffding Trees (**AHT**) [3] as weak learners; AHT is a decision-tree induction algorithm for streams that ensures DT model adaptation to the underlying data distribution by not only updating the tree with new instances from the stream, but also by replacing sub-trees when their performance decreases.

The classification of a new unseen instance at time point t, i.e., x_t, is based on weighted majority voting and depends on its membership to z. If the instance does not belong to z (i.e., it is a non-protected instance), then the standard boundary of the ensemble is used. Otherwise, the adjusted boundary is used. More formally:

$$f_t(x_t) = \begin{cases} y^+ & if\ x_t \in z\ and\ H^t_{1:N}(x_t) \geq \theta^t \\ H^t_{1:N}(x_t) & \text{otherwise.} \end{cases} \tag{7}$$

where N is the number of weak learners of the ensemble, and θ^t is the fair adjusted boundary at timepoint t. For Cum.S.P. and Cum.EQ.OP., only the boundary of the protected group is tweaked. Other parity-based notions (such as *Disparate Mistreatment* [31]) may also tweak the boundary of the non-protected group. Note that the adjustment of the boundary based on θ^t is applied at the ensemble level and not at each individual weak learner predictions.

5 Evaluation

In this section, we introduce the employed baselines as well as variants of FABBOO[2] that help us to demonstrate the behavior of FABBOO's individual components. The employed datasets as well as the performance measures are given below. For the experimental evaluation, in order to get the best γ, λ and M parameters, we performed a grid-search and selected $\gamma = 0.1, \lambda = 0.9$, $M = 2,000$ that showed an overall good performance across all datasets. We also set $N = 20$ for all the ensemble methods and a very small value $\epsilon = 0.0001$, which means no tolerance to discriminatory outcomes. Finally, for the prequential evaluation of the non-stream datasets, we report on the average of 10 random shuffles (same as in [22, 30]).

5.1 Competitors and Performance Measures

We evaluate FABBOO against two recent state-of-the-art fairness-aware stream classifiers [22, 30] and the fairness agnostic non-stationary OSBoost [7]. We also employ two variations of FABBOO to show the impact of its different components, namely class-imbalance and cumulative fairness. All methods employ AHTs as weak learners and therefore are able to handle concept drifts. The only exception is FAHT [30] which is an incremental Hoeffding Tree that not tackle concept drifts. An overview follows:

1. **Fairness Aware Hoeffing Tree (FAHT) [30]:** FAHT is an extension of the Hoeffding tree that accounts for statistical parity by alternating the node split procedure

[2] `https://iosifidisvasileios.github.io/FABBOO`

Table 1. An overview of the datasets.

	#Instances	#Attributes	Sen.Attr.	z	$\bar{z}$	Class ratio (+:-)	Stream	Positive class	Source
Adult Cen.	45,175	14	Gender	Female	Male	1:3.03	-	<50K	[2]
Bank	40,004	16	Marit. Status	Married	Single	1:7.57	-	subscription	[2]
Default	30,000	24	Gender	Female	Male	1:3.52	-	default payment	[2]
Kdd Cen.	299,285	41	Gender	Female	Male	1:15.11	-	<50K	[2]
Loan	21,443	38	Gender	Female	Male	1:1.26	✓	paid	[8]
NYPD	311,367	16	Gender	Female	Male	1:3.68	✓	felony	[6]
synthetic	150,236	6	synth.	synth.	synth.	1:3.13	✓	synth.	[22]

to facilitate information as well as fairness gain (statistical parity). FAHT grows according to the joint split of information and fairness gain, thus accounts for cumulative outcomes; however, it does not handle concept drifts nor class-imbalance.

2. **Massaging (MS) [22]:** a chunk based model-agnostic approach which minimizes S.P. on recent discriminatory outcomes. It detects and removes discrimination within the chunk by performing label swaps and retrains the model based on the "corrected" chunk. MS is dealing with concept drifts by blind adaptation (using an adaptive learner), but is considering short-term discrimination outcomes and does not account for class imbalance. We use the default chunk size of 1,000 instances.
3. **Online Smooth Boosting (OSBoost) [7]:** OSBoost does not consider fairness nor class imbalance.
4. **Online Fair Imbalanced Boosting (OFIB):** A variation of FABBOO that does not account for class imbalance i.e., it does not use OCIS during training. This variation helps to show the importance of tackling class imbalance.
5. **Chunk Fair Balanced Boosting (CFBB):** A variation of FABBOO that tackles short-term, instead of cumulative, discrimination. This variation helps to show the importance of long term fairness assessment. Instead of accounting for discrimination from the beginning of the stream, it monitors the 1,000 most recent instances.

To evaluate the performance of FABBOO and baselines, we employ a set of measures which are able to show the performance in the presence of class-imbalance. Same as in [11], we employ *gmean*, *recall*, and *balanced accuracy* (Bal.Acc.). For measuring discrimination, we report on *cumulative statistical parity* in Section 5.3 and *cumulative equal opportunity* in Section 5.4.

5.2 Datasets

To evaluate FABBOO, we employ a variety of real-world as well as synthetic datasets which are summarized in Table 1. The datasets vary in terms of class imbalance, dimensionality and volume. Same as in [22, 30], we use **Adult census** dataset (Adult) and **Kdd Census** dataset (Kdd Cen.) as well as **Bank** dataset, and **Default** dataset by randomly shuffling them, since they are not streaming datasets. We also employ **Loan**, **NYPD** and a **synthetic** dataset, all of which have temporal characteristics. For synthetic dataset, we follow the authors' initialization process [22], where each attribute corresponds to a different Gaussian distribution, and also inject class-imbalance and concept drifts to the stream. Concept drifts in this scenario are performed by shifting

Table 2. Overall predictive and fairness performance for Cum.S.P. (Winner in bold)

	Method	Bal. Acc. (%)	Gmean (%)	Recall (%)	Cum.S.P. (%)
Adult	FAHT	72.14±1.4	68.65±2.2	50.11±3.5	16.51±1.3
	MS	72.31±1.2	69.00±1.8	50.91±2.9	22.93±1.6
	OSBoost	73.90±0.5	71.11±0.8	53.73±1.3	18.05±0.6
	OFIB	74.21±0.3	72.92±0.4	60.01±1.0	0.26±0.1
	CFBB	74.12±0.6	73.74±0.6	66.95±1.1	-5.00±2.0
	FABBOO	**76.58±0.1**	**76.57±0.1**	**73.98±0.7**	**0.21±0.1**
Bank	FAHT	61.92±2.0	50.64±4.4	26.51±4.4	2.58±0.5
	MS	63.21±1.9	53.54±3.6	29.75±4.1	8.10±1.2
	OSBoost	64.41±0.6	55.54±1.1	31.81±1.3	3.37±0.2
	OFIB	67.90±0.7	62.04±1.2	40.21±1.6	**0.22±0.1**
	CFBB	78.37±0.5	78.08±0.6	71.24±1.5	-6.06±1.3
	FABBOO	**83.39±0.4**	**83.38±0.4**	**83.36±1.4**	**0.22±0.1**
Default	FAHT	62.72±0.6	53.48±1.2	29.95±1.4	1.80±0.4
	MS	63.76±0.5	55.53±1.4	32.4±2.0	12.16±1.5
	OSBoost	63.06±0.6	53.87±1.3	30.32±1.7	1.89±0.4
	OFIB	63.79±0.7	55.41±1.6	32.36±2.1	0.29±0.1
	CFBB	65.82±0.6	65.44±0.4	58.58±3.0	-7.74±2.2
	FABBOO	**67.49±0.6**	**66.89±0.5**	**58.66±2.8**	**0.17±0.1**
Kdd Cen.	FAHT	62.80±2.3	51.04±4.6	26.45±4.7	2.82±0.6
	MS	62.02±1.2	49.71±2.3	24.91±2.4	15.8±0.97
	OSBoost	65.55±0.8	56.28±1.3	31.97±1.5	3.62±0.3
	OFIB	67.55±0.9	60.48±1.5	37.59±1.9	0.13±0
	CFBB	78.40±0.5	77.58±0.6	66.60±1.1	1.34±0.5
	FABBOO	**81.48±0.3**	**81.41±0.4**	**77.98±0.6**	**0.04±0**
Loan	FAHT	62.61	60.14	70.21	6.41
	MS	61.44	59.64	69.31	60.13
	OSBoost	**63.84**	**60.31**	76.13	8.14
	OFIB	62.41	58.34	78.63	1.12
	CFBB	63.15	60.05	79.73	-2.72
	FABBOO	63.47	60.22	**79.91**	**0.51**
NYPD	FAHT	50.15	6.13	0.37	0.09
	MS	56.93	41.06	17.47	5.87
	OSBoost	52.24	24.33	6.01	0.75
	OFIB	52.32	24.96	6.36	0.05
	CFBB	62.48	59.48	43.63	-6.46
	FABBOO	**62.96**	**60.78**	**46.83**	**0.03**
synthetic	FAHT	57.12	42.56	18.90	8.31
	MS	62.43	53.81	30.90	15.26
	OSBoost	63.42	54.87	31.61	7.97
	OFIB	64.01	57.54	35.85	**-0.56**
	CFBB	65.93	64.75	53.75	-9.68
	FABBOO	**69.09**	**69.01**	**60.11**	0.66

the mean average of each Gaussian distribution (5 non-reoccuring concept drifts have been inserted at random points, see Figure 2 or 3).

5.3 Results on cumulative statistical parity

In this section, we compare our approach against the employed competitors for Cum.S.P., and report the overall results in Table 2. As we see, FABBOO is able to mitigate unfair outcomes and maintain the best performance in terms of balanced accuracy, gmean, and recall for all datasets. E.g., for *Adult Cen.*, the best balanced accuracy is achieved by FABBOO followed by OFIB (2.3%↓), the best gmean is achieved by FABBOO followed by CFBB (2.8%↓), and the best recall is achieved by FABBOO followed by

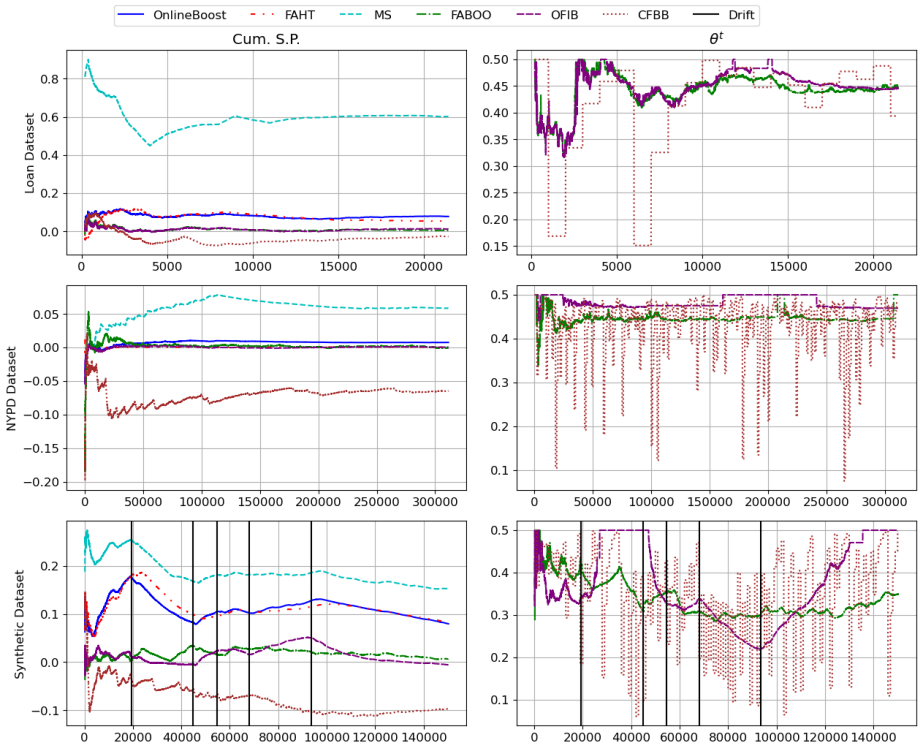

Fig. 2. Cum.S.P. and boundary adjusting for Loan (top), NYPD (middle) and Synthetic (bottom) datasets

CFBB (7%↓). OFIB is able to reduce discrimination, same as FABBOO, in expense of sacrificing 2.3%↓ balanced accuracy.

Overall, FABBOO achieves the best balanced accuracy, across all datasets, with an average score of 72.01%, followed by CFBB with an average score of 69.73%. In terms of discrimination, FABBOO is the clear winner, across all datasets, with an average score of 0.26%, followed by OFIB with an average score of 0.37%. Although the difference in terms of discrimination is small, OFIB has an average balanced accuracy score of 64.57%. CFBB achieves an average score of 5.57% in terms of Cum.S.P, while FAHT and MS achieve an average score of 5.49% and 20.02%, respectively.

To get a closer look at the over time performance of the different methods, we show in Figure 2 the Cum. S.P. (left) and the required decision boundary adjustment (right), i.e., the boundary threshold θ^t, for the datasets with temporal information. Looking at the Cum.S.P.(left), we see that for all datasets, CFBB is not able to mitigate discrimination; instead, it propagates *reverse discrimination* (negative Cum. S.P.) and discriminates the non-protected group. MS falls in the same pitfall; by "correcting" the data based solely on the chuck it is not able to tackle unfair cumulative outcomes. Both CFBB and MS results show that a short-term consideration of fairness is unable

to tackle discrimination propagation and reinforcement in the stream. The fairness-agnostic OSBoost is also not able to tackle discrimination. The only exception is the *NYPD* dataset. However, a closer look shows that the achieved low S.P. is only a result of vast rejecting the minority class (c.f., Table 2). On the other hand, FABBOO and OFIB (the FABOO variation that does not tackle class-imbalance) are able to tackle discrimination overtime, and outperform FAHT and MS.

Looking at the required adjustments of the decision boundary (right), we notice that OFIB tends to produce higher boundary values than FABBOO. This is caused due to OFIB's inability to learn the minority class effectively; therefore, it rejects more minority instances from both protected and non-protected groups. For Loan dataset FABOO and OFIB are performing similarly since the dataset is not severely imbalanced. Finally, we observe that CFBB has high fluctuation when adjusting the decision boundary due to its inability to adapt to underlying changes in data distributions w.r.t. fairness.

5.4 Results on cumulative equal opportunity

Table 3. Overall predictive and fairness performance for Cum.EQ.OP. (Winner in bold)

	Method	Bal.Acc. (%)	Gmean (%)	Recall (%)	Cum.EQ.OP. (%)
Adult	OSBoost	73.90±0.5	71.13±0.8	53.73±1.3	18.41±3.2
	OFIB	74.74±0.5	72.36±0.7	56.07±1.1	**3.14±1.4**
	CFBB	76.70±0.4	75.46±0.5	62.96±1.1	9.34±1.6
	FABBOO	**78.71±0.2**	**78.38±0.3**	**70.83±0.9**	3.27±1.6
Bank	OSBoost	64.41±0.6	55.54±1.1	31.81±1.3	5.51±1.1
	OFIB	65.42±0.6	57.46±1.2	34.17±1.4	1.5±0.8
	CFBB	76.74±0.8	75.27±1.1	61.88±2.0	-1.85±1.2
	FABBOO	**82.58±0.5**	**82.44±0.5**	**78.05±1.7**	**0.1±0.6**
Compass	OSBoost	**65.25±0.3**	**64.91±0.4**	58.74±1.4	29.81±1.7
	OFIB	64.58±0.2	64.53±0.2	62.51±1.5	4.84±2.5
	CFBB	64.76±0.4	64.69±0.4	62.07±1.3	14.73±3.3
	FABBOO	64.52±0.3	64.50±0.3	**64.40±1.5**	**4.76±2.9**
Default	OSBoost	63.06±0.6	53.87±1.3	30.32±1.7	79.01±0.9
	OFIB	63.14±0.6	54.06±1.5	30.57±1.8	**0.26±0.6**
	CFBB	66.61±0.3	65.75±0.4	56.31±2.6	-2.21±0.9
	FABBOO	**67.55±0.5**	**66.78±0.5**	**57.79±2.7**	0.93±0.7
Kdd Cen.	OSBoost	65.55±0.8	56.28±1.3	31.97±1.5	15.99±0.3
	OFIB	66.85±0.8	58.88±1.3	35.21±1.6	0.83±0.2
	CFBB	78.52±0.5	77.30±0.7	64.75±1.2	2.72±0.9
	FABBOO	**82.39±0.4**	**82.16±0.4**	**76.26±0.5**	**0.6±0.3**
Loan	OSBoost	**63.84**	**60.31**	76.13	1.25
	OFIB	61.51	58.59	78.31	0.12
	CFBB	62.61	59.84	79.03	12.89
	FABBOO	63.06	60.18	**80.73**	**0.07**
NYPD	OSBoost	52.24	24.33	6.01	1.25
	OFIB	52.31	24.75	6.22	0.12
	CFBB	62.17	58.84	42.08	12.89
	FABBOO	**62.65**	**60.38**	**45.92**	**0.07**
synthetic	OSBoost	63.42	54.87	31.61	5.18
	OFIB	63.75	56.67	34.55	-6.04
	CFBB	66.97	65.02	50.92	-18.10
	FABBOO	**69.13**	**68.17**	**57.68**	**-0.17**

For Cumul. EQ.OP., we report the results of OSBoost, OFIB, CFBB, and FABOO on Table 3. We exclude FAHT and MS since they are designed to mitigate unfair out-

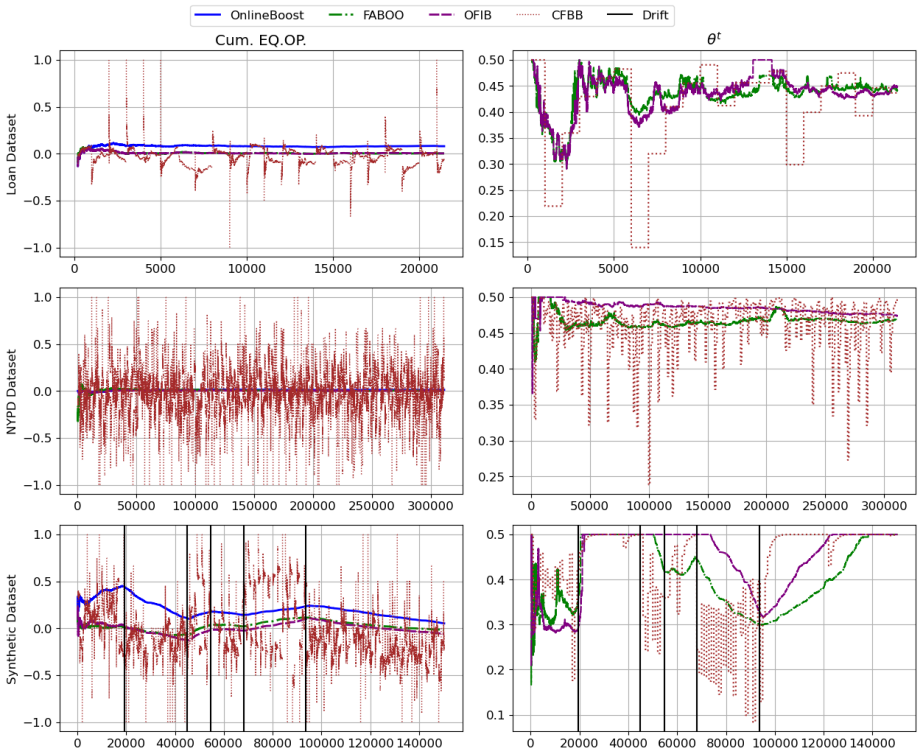

Fig. 3. Cum.EQ.OP. and boundary adjusting for Loan (top), NYPD (middle) and Synthetic (bottom) datasets

comes based on statistical parity. To the best of our knowledge, there are no fairness-aware stream learning methods that mitigate unfair outcomes based on equal opportunity.

The results indicate that FABBOO performs good in terms of balanced accuracy, gmean, and recall in all datasets except Compass and Loan, which are balanced datasets. E.g., for Adult Cen. dataset, the best balanced accuracy is achieved by FABBOO followed by CFBB (2%↓), the best Gmean is achieved by FABBOO followed by CFBB (2.9%↓), and the best recall is achieved by FABBOO followed by CFBB (7.9%↓). OFIB achieves slightly better Cumul. EQ.OP. than FABBOO (0.01%↓), however OFIB rejects more instances in the positive class. Similar behavior can be observed in all datasets, where FABBOO is able to tackle class imbalance and mitigate unfair outcomes better than the other methods. OSBoost fails to learn the positive (minority) class, thus under-performs in almost all datasets. In some cases, it produces low discriminatory outcomes; however, this is a result of misclassifying huge portions of the positive class.

We also demonstrate how Cumul. EQ.OP. and the decision boundary (FABBOO, OFIB and CFBB) vary over time for the stream datasets in Figure 3. In all datasets, we observe that CFBB's decision boundary is highly fluctuating in contrast to OFIB and

FABBOO. CFBB is also unstable in terms of Cumul. EQ.OP., since it is not mitigating cumulative unfair outcomes. OFIB tweaks the boundary less than FABBOO, while it fails to learn the minority class well enough, thus rejects more positive instances.

6 Conclusion

In this paper, we proposed FABBOO, an online fairness-aware learner for data streams with class imbalance and concept drifts. Our approach changes the training distribution online taking into account class-imbalance. Moreover, our method can facilitate different fairness notions by adjusting the decision boundary on demand. Our experiments show that our approach outperforms other methods in a variety of datasets w.r.t. both predictive- and fairness-performance. In addition, we show that recent fairness-aware methods reject the minority class at large to ensure fair results. On the contrary, our class-imbalance-oriented approach effectively learns both classes and fulfills different fairness criteria while achieving good predictive performance for both classes. Finally, we show that our cumulative definitions enable the model to mitigate long-term discriminatory effects, in contrast to a short-term definition like in CFBB and MS which are unable to deal with discrimination propagation and reinforcement in the stream. As part of our future work, we plan to embed the decision boundary adjustment directly into the training phase by altering the weighted training distribution, as proposed in [21]. Finally, we have assumed that the role of the minority class is not fixed over the stream; however, we have assumed that the protected group is fixed over the stream. We intend to waive this assumption and extend FABBOO to tackle *reverse discrimination* as well.

References

1. Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., Rieke, A.: Discrimination through optimization: How facebook's ad delivery can lead to skewed outcomes. arXiv preprint arXiv:1904.02095 (2019)
2. Bache, K., Lichman, M.: Uci machine learning repository (2013)
3. Bifet, A., Gavaldà, R.: Adaptive learning from evolving data streams. In: International Symposium on Intelligent Data Analysis. pp. 249–260. Springer (2009)
4. Calders, T., Žliobaitė, I.: Why unbiased computational processes can lead to discriminative decision procedures. In: Discrimination and privacy in the information society, pp. 43–57. Springer (2013)
5. Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Optimized preprocessing for discrimination prevention. In: Advances in Neural Information Processing Systems. pp. 3992–4001 (2017)
6. Chapman, D., Ryan, P., Farmer, J.P.: Introducing alpha.data.gov. Office of Science and Technology Policy (2013), www.whitehouse.gov/blog/2013/01/28/introducing-alphadatagov
7. Chen, S.T., Lin, H.T., Lu, C.J.: An online boosting algorithm with theoretical justifications. arXiv preprint arXiv:1206.6422 (2012)
8. Cortez, V.: Preventing discriminatory outcomes in credit models (2019), https://github.com/valeria-io/bias-in-credit-models
9. Council, N.R., et al.: Measuring racial discrimination. National Academies Press (2004)

10. Datta, A., Tschantz, M.C., Datta, A.: Automated experiments on ad privacy settings. Privacy Enhancing Technologies **2015**(1), 92–112 (2015)
11. Ditzler, G., Polikar, R.: Incremental learning of concept drift from streaming imbalanced data. IEEE transactions on knowledge and data engineering **25**(10), 2283–2301 (2012)
12. Fish, B., Kun, J., Lelkes, Á.D.: A confidence-based approach for balancing fairness and accuracy. In: Proceedings of the 2016 SIAM International Conference on Data Mining. pp. 144–152. SIAM (2016)
13. Forman, G.: Tackling concept drift by temporal inductive transfer. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 252–259. ACM (2006)
14. Gama, J.: Knowledge discovery from data streams. Chapman and Hall/CRC (2010)
15. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O.: When will ai exceed human performance? evidence from ai experts. Journal of Artificial Intelligence Research **62**, 729–754 (2018)
16. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems. pp. 3315–3323 (2016)
17. Hu, H., Iosifidis, V., Liao, W., Zhang, H., YingYang, M., Ntoutsi, E., Rosenhahn, B.: Fairnn-conjoint learning of fair representations for fair decisions. arXiv preprint arXiv:2004.02173 (2020)
18. Ingold, D., Soper, S.: Amazon doesn't consider the race of its customers. should it. Bloomberg, April (2016)
19. Iosifidis, V., Fetahu, B., Ntoutsi, E.: Fae: A fairness-aware ensemble framework. In: 2019 IEEE International Conference on Big Data (Big Data). pp. 1375–1380. IEEE (2019)
20. Iosifidis, V., Ntoutsi, E.: Dealing with bias via data augmentation in supervised learning scenarios. Jo Bates Paul D. Clough Robert Jäschke **24** (2018)
21. Iosifidis, V., Ntoutsi, E.: Adafair: Cumulative fairness adaptive boosting. CIKM (2019)
22. Iosifidis, V., Tran, T.N.H., Ntoutsi, E.: Fairness-enhancing interventions in stream classification. In: DEXA. pp. 261–276. Springer (2019)
23. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012)
24. Kamiran, F., Mansha, S., Karim, A., Zhang, X.: Exploiting reject option in classification for social discrimination control. Information Sciences **425**, 18–33 (2018)
25. Krasanakis, E., Xioufis, E.S., Papadopoulos, S., Kompatsiaris, Y.: Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In: WWW. pp. 853–862. ACM (2018)
26. Vafa, K., Haigh, C., Leung, A., Yonack, N.: Price discrimination in the princeton review's online sat tutoring service. JOTS Technology Science, Sep **1** (2015)
27. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 1–7. IEEE (2018)
28. Wang, S., Minku, L.L., Yao, X.: A learning framework for online class imbalance learning. In: 2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL). pp. 36–45. IEEE (2013)
29. Weiss, G.M.: Mining with rarity: a unifying framework. ACM Sigkdd Explorations Newsletter **6**(1), 7–19 (2004)
30. Wenbin, Z., Ntoutsi, E.: Faht: An adaptive fairness-aware decision tree classifier. arXiv preprint arXiv:1907.07237 (2019)
31. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1171–1180. WWW (2017)