Flexible and Adaptive Fairness-aware Learning in Non-stationary Data Streams

Wenbin Zhang[1], Mingli Zhang[2], Ji Zhang[3], Zhen Liu[4],
Zhiyuan Chen[1], Jianwu Wang[1], Edward Raff[5] and Enza Messina[6]
[1]University of Maryland, Baltimore County, MD, USA
{wenbinzhang, zhchen, jianwu}@umbc.edu
[2]McGill University, Montreal, Canada
mingli.zhang@mcgill.ca
[3]University of Southern Queensland, Toowoomba, Australia
ji.zhang@usq.edu.au
[4]Guangdong Pharmaceutical University, Guangzhou, China
liu.zhen@gdpu.edu.cn
[5]Booz Allen Hamilton, MD, USA
raff_edward@bah.com
[6]University of Milano-Bicocca, Milano, Italy
enza.messina@unimib.it

***Abstract*— Artificial intelligence (AI)-based decision-making systems are employed nowadays in an ever growing number of online as well as offline services–some of great importance. Depending on sophisticated learning algorithms and available data, these systems are increasingly becoming automated and data-driven. However, these systems can impact individuals and communities with ethical or legal consequences. Numerous approaches have therefore been proposed to develop decision-making systems that are discrimination-conscious by-design. However, these methods assume the underlying data distribution is stationary without drift, which is counterfactual in many real-world applications. In addition, their focus has been largely on minimizing discrimination while maximizing prediction performance without necessary flexibility in customizing the trade-off according to different applications. To this end, we propose a learning algorithm for fair classification that also adapts to evolving data streams and further allows for a flexible control on the degree of accuracy and fairness. The positive results on a set of discriminated and non-stationary data streams demonstrate the effectiveness and flexibility of this approach.**

***Keywords*—AI fairness, online classification, flexible fairness.**

I. Introduction

Artificial intelligence-based decision-making systems have become a necessity in both online as well as offline settings to render all sorts of decisions, such as screening of job application, loan credit approval, allocation of health resources and autonomous driving. This trend is likely to continue given that many AI systems are able to match or even surpass humans in accuracy and/or throughput [1]. However, without intervention these systems may make decisions dependent on the sensitive attributes (e.g., gender and ethnicity) and lead to discrimination against particular groups of people sharing one or more sensitive attributes. A growing body of such kind of discriminatory incidents caused by the AI-based decision-making systems have been observed and reported [2], [3], [4]. As a recent example, the AI algorithm behind Google's AdFisher tool has suggested signs of gender discrimination by displaying male job hunters significantly more higher-paying jobs than to female, even though the AI algorithm did not consider gender as an attribute [5]. Such incidents have pointed out the urgent need to consider the potential loss of fairness and accountability in AI-based decision-making systems, pulling in diverse scholars from civil organizations, policy-makers and legal experts.

A number of studies have therefore been proposed to address the growing concern on the fairness issue of AI models and to develop discrimination-conscious AI systems. These studies can be broadly categorized into pre-processing approaches [6], [7], in-processing approaches [8], [9] and post-processing approaches [10], [11] based on whether the elimination of discrimination are focusing on the data level, the algorithm level and the output of model, respectively. A common theme amongst all these prior works is the assumption of fairness as a static problem, which means the inappropriate discriminative correlations (e.g., gender with aptitude) is implicitly modeled as a constant and static property. This is unrealistic to the many domains that have underlying concept drift over time and leaves users with online-learning problems without tooling to mitigate these concerns [12], [13]. In addition, the focus of most current work is on minimizing discrimination performance while maximizing prediction accuracy, without a flexible control on the trade-off between fairness and accuracy [14], [15]. If a user wanted to alter the balance to meet a justified "business necessity" (a legally encoded concept to allow a balance between accuracy and performance) [16], they would generally be required to alter a number of hyper-parameters that have non-trivial interactions to obtain the desired balance [17]. In our approach we include a hyper-parameter designed to provide an easy and direct trade-off between these two factors so that users can more easily adjust to their own applications, with the goal of encouraging wider use through easy to obtain results.

Compared with the numerous approaches address static fairness, discrimination-aware learning in data stream is highly under-explored and brings unique challenges [18], [19]. In such applications, the model should be able to address the tightly coupled non-stationary issue simultaneously, and the evolution of target concepts is always accompanied and complicated by the biased decision regions. Transferring approaches among these two domains is therefore unfortunately not straightforward, sophisticated design is warranted. Motivated by these significant challenges, this paper introduces a fair and concept-adapting classifier for discriminated and time-evolving data streams. This work also takes a step forward by equipping it with flexible fairness control capability for application-driven fairness-aware learning. *To the best of our knowledge, this is the first work jointly considers non-stationary data distribution and flexible control on the trade-off between fairness and accuracy.* The contribution of this paper is three-fold:

- We define a new problem of fairness-aware learning in data streams with flexible fairness control. Then, we propose a discrimination-aware learner with add-on concept drift adaptation ability to handle discriminated and non-stationary data streams, and is capable of controlling the trade-off between fairness and accuracy in a flexible manner.
- We introduce the flexible fair information gain that jointly considers the information and fairness gain as well as customizes the trade-off between fairness and accuracy, thus providing more flexibility than the state of arts.
- The conducted experiments verify the capability of the proposed model in online settings and for application-driven fairness-aware learning.

The rest of the paper is organized as follows. Theoretical background knowledge and related studies regarding fairness-aware learning are first reviewed in Section II and III, respectively. We then discuss the vanilla Fairness-Aware Hoeffding Tree classifier which is the base model of our approach in Section IV. Next, Section V presents the proposed method, followed by the experimental results in detail in Section VI. Finally, we conclude the paper in Section VII.

II. Basic Notations and Problem Definition

Let D be the data stream consisting of a sequence of instances $x_1, x_2, \cdots, x_t$ that arrive continuously and possible infinitely over time. Each instance $x_t \in D$ is represented over the schema $(A_1, A_2, \cdots, A_n, Y)$ with their respective domains $dom(A_i)$ and $dom(Y)$, where A is a set of attributes and Y is the class label. Moreover, we further assume a special attribute $S \in A$, referred to as the *sensitive attribute*, e.g., gender and race, with a special value $s \in dom(S)$, referred to as the *sensitive value*, e.g., female and black, that define the unprivileged group.

An online classifier F is then a function from A to Y, denoted as F: $A \rightarrow Y$. The current model F_t is first trained given $x_1, x_2, \cdots, x_t$ and predicts the class of x_{t+1}, i.e., y_{t+1}. Once the prediction is made, the actual class label of x_{t+1} is revealed to F_t for model updating from F_t to F_{t+1} and the next instance x_{t+2} arrives afterwards. This setup is also known as first-test-then-train or prequential evaluation [20]. In addition, concept drift occurs when there are changes in the underlying data distribution such that $P_{t_1}(A, y) \neq P_{t_2}(A, y)$ for $t_1 \neq t_2$, and therefore F needs to be updated accordingly.

While maintaining an accurate and up-to-date classifier, online fairness-aware learning simultaneously requires F does not discriminate w.r.t the sensitive attribute, i.e., between the privileged and unprivileged groups. Up to now, more than twenty fairness measures have been proposed to assess such discriminative behavior [21]. One of the most widely used measures is the *statistical parity* [22] which measures the percentage difference between privileged and unprivileged group when being assigned a positive target class, for example allocating healthcare resources. The unprivileged can claim that they are discriminated when their positive classification percentage is lower than the privileged group's. Without loss of generality, we assume both the sensitive attribute and class label are binary-valued. Four respective groups representing privileged group receiving positive classification (PP), privileged group receiving negative classification (PN), unprivileged group receiving positive classification (UP), unprivileged group receiving negative classification (UN) can therefore created, and the statistical parity in online settings can be formulated as below:

$$Disc(D_t) = \frac{PP_t}{PP_t + PN_t} - \frac{UP_t}{UP_t + UN_t} \tag{1}$$

Compared with the statistical parity in offline settings, Equation (1) measures the cumulative discrimination up to time t and $Disc(D_t)$ might also evolve over time, which further complicates the learning. The aim of online fairness-aware learning is therefore to maintain an up-to-date F that makes accurate predictions for Y but also does not discriminate w.r.t S.

We make explicit that what constitutes "fair" or "discriminative" is dependent on many factors and context [23], as well as philosophical questions that have been researched long before the AI communities' interest [24]. Addressing such questions about when to use statistical parity over any other metric, and deeper questions, are critical but beyond the scope of this work. We select statistical parity because American user studies have found that it is a measure compatible with many user's intuition of what constitutes a "fair" decision [25]. As such we expect many applications could make use of our method, though an informed user should make a careful considerable about the relevant factors of their domain [26].

III. Related Work

This work focuses on fairness-aware learning in data streams, relevant studies therefore include fairness-aware learning in offline settings, data stream classification as well as the online fairness-aware learning that lies in the intersection of the previous two research directions.

A. Offline Fairness-aware Learning

Motivated by the increasing attentive concerns, booming approaches have been proposed to tackle AI fairness as a batch learning problem aiming at minimizing discrimination while preserving prediction performance [27], [2], [12]. Most of these studies assume offline settings by design and can be typically categorized into three main families: i) pre-processing approaches, ii) in-processing approaches and ii) post-processing approaches, based on whether they mitigate bias at the data level, the algorithm design or the output of model, respectively.

The first strategy, *pre-processing solutions*, consists of performing different data level operations such as transformation and augmentation to neutralize or eliminate the extent of inherited bias of the data. The rationality for such type of approaches is that classifiers trained on the fairly represented data could make fair predictions. These methods are model-agnostic and can be employed in conjunction with any applicable classifier after the pre-processing step. Representative works include massaging [6] and reweighting [7]. The former directly swaps the class labels of selected instances to change data distribution for the sake of balanced representation. The swapped instances are selected using a ranker based on the potential accuracy deterioration in order to minimize accuracy loss while reducing discrimination. While the latter, instead of intrusively relabeling the instances, assigns different weights to different communities to reduce discrimination. Instances belonging to the unprivileged group will receive higher weighs comparing to instances from the privileged group. However, methods in this category are typically not quite effective as standalone unless being used in conjunction with other methods with sophisticated design.

In contrast, the second category, *in-processing approaches*, consists of modifying existing algorithms, usually integrating fairness as a part of the objective function through constraints or regularization, to mitigate discrimination, and is therefore algorithm-specific. [11] is one of the seminal in-processing works, in which discrimination, reflected by the entropy w.r.t. sensitive attribute, is incorporated into the splitting criterion for fair tree induction. In [14], the measure of "decision boundary fairness" is leveraged to penalize discrimination in the formulation of a set of convex margin-based classifiers. More recently, Fang et al. [28] propose the notion of "fair-group construction" to emphasize sensitive attributes in the classification process for the sake of improving fairness in prediction outcomes. Our work belongs to this category by jointly considering data encoding and diminishing discrimination of the training data for an accuracy-driven as well as fairness-oriented model.

The last category, *postprocessing techniques*, consists of either adjusting the decision boundary of a model or directly changing the prediction labels. [22] processes with additional prediction thresholds to work against discrimination while the decision boundary of AdaBoost is shifted w.r.t. fairness in [29]. The latter approaches pay attention to the outcome of a classifier. In [11], for example, relabeling is performed on selected leaves of the decision tree to decrease discrimination while minimizing the effect on predictive accuracy. We emphasize that transferring such techniques to online settings is not straightforward as the boundary/prediction could evolve themselves due to the non-stationary distributions in online settings.

B. Data Stream Learning

In data stream mining, the data arrives sequentially and the underlying data distribution might also evolve over time, known as concept drift [20], [30], [31]. The learning algorithms therefore need to take the evolution of underlying data distribution into consideration, while remaining stable on historical but not outdated concepts. Such adaptation is normally enabled by: 1) incorporating new instances from the stream into the model [32], [33], and ii) forgetting the previous outdated knowledge from the model [34], [35].

The first type of adaptation calls for incremental algorithms. In [36], a probabilistic classifier that updates the probability distribution based on the new instances from the stream but does not forget the previous instances is proposed. The representative work is the Hoeffding Tree classifier [37], which scans each instance in the stream only once and stores sufficient information in its leaves in order to grow. The tree learned is also theoretical guaranteed to be asymptotically nearly identical to the tree induced by a conventional static learner. The second category calls for methods that are able to forget. There is a plethora of such approaches in the literature, these approaches can be further categorized into gradual forgetting [38], [39] and abrupt forgetting [40], [41] methods, depending on the rate at which concept drift presents in the stream.

C. Online Fairness-aware Learning

A number of studies have been proposed with regard to offline fairness-aware learning and data stream learning solely focusing on the elimination of discrimination and concept drift, respectively. However, fairness in online setting requires simultaneously taking the removal of prediction dependence on the sensitive attributes and the evolution of underlying data distribution into consideration. In [18], massaging and reweighting are extended for a chunk based fair stream classification approach in which these two pre-processing methods are applied before updating the online classifier focusing on concept drift adaptation. However, as discussed in Section III-A, pre-processing methods, even in the offline settings, are typically not quite effective as the standalone approach let alone in the more challenging online environments. The applicability of this approach is therefore unknown. More recently, Zhang et al. [12] improves the splitting strategy of [11] and operates their model in the online setting. This model was later extended for enhanced discrimination elimination and prompt response on concept drift [13]. However, research efforts in this direction have still been limited. Our work situates in this highly under-explored research direction by encapsulating the capability of drift detection and adaptation to tackle online fairness comprehensively, so as to provide fair online decision-making.

IV. Vanilla Fairness-Aware Hoeffding Tree (FAHT)

Our Flexible and Adaptive Fairness-Aware Hoeffding Tree (2FAHT) classifier extends the Fairness-Aware Hoeffding Tree (FAHT) classifier [12], which is built on top of the Hoeffding Tree (HT) classifier [37]. To mine high-speed data stream, FAHT or HT induces a decision tree from the given stream incrementally, briefly scanning each example in the stream only once and storing sufficient information in its leaves in order to grow when future data arrives. The critical decisions needed during the induction of the tree are when to split a node and with which example-discriminating test. To this end, the authors employ the Hoeffding bound [37] to guarantee that the tree learned converges (with high likelihood) to the conventional static tree built by a batch learner, given enough examples. In HT, these two decisions are based on the *information gain (IG)* [42], which is exclusively accuracy-oriented and does not consider fairness. The prior FAHT method solves the discrimination problem by introducing a new splitting criterion, called *fair information gain (FIG)*, that jointly considers the fairness gain and information gain of the introduction of an attribute split:

$$FIG(D,A) = \begin{cases} IG(D,A), & \text{if } FG(D,A)=0 \\ IG(D,A)\times FG(D,A), & \text{otherwise} \end{cases} \tag{2}$$

where FG refers to *fairness gain* that measures the difference in discrimination due to the split and is formulated as:

$$FG(D,A) = |Disc(D)| - \sum_{v\in dom(A)} \frac{|D_v|}{|D|} |Disc(D_v)| \tag{3}$$

where A is an attribute relative to the collection of instances D that stored in sufficient statistics, $D_v, v \in dom(A)$ are the partitions/subsets induced by A, and each corresponding discrimination value is gauged according to Equation (1). In *fairness gain*, each subset contributes with a weight factor relative to its cardinality, i.e., $\frac{|D_v|}{|D|}$, which is identical to information gain.

The authors [13] later reformulate FG by relaxing the fairness gain ratio so as to maximize the cumulative fairness:

$$FG(D,A) = |Disc(D)| - \sum_{v\in dom(A)} |Disc(D_v)| \tag{4}$$

The motivation is to encourages fair splits by giving priority to splitting candidates that result in a higher discrimination reduction regardless of their number of distinct values and less represented attribute values. This is also equivalent to assigning a bonus to attributes with multi-values and attribute values with small representation sizes but have a high discrimination reduction when being selected as the splitting attributes in Equation (3). We adopt this reformulated notion in our method to select splitting candidates resulting higher discrimination reductions, which simultaneously provide extra spaces controlling the level of fairness thus adding more flexibility on the trade-off between fairness and accuracy when needed.

The idea of FG is motivated by IG but from the discrimination perspective. Multiplication is favoured, when combining them as a conjunctive objective, over other operations, for example addition, as the values of these two metrics could be in different scales, and in order to promote fair splitting which results in a reduction in the discrimination after split, i.e., FG is a positive value.

The conducted experiments demonstrated the predictive and anti-discrimination capability of FAHT [12]. However, there are two limitations for FAHT. First, FAHT focuses on optimizing for fairness while maximizing accuracy and has no method to adjust the trade-off between accuracy and fairness. If FAHT's parameters do not confirm sufficient fairness, or if its accuracy is too low to be usable, there is no means to adjust FAHT's tree induction to fix these situations. Second, FAHT is not adaptive to concept drift. FAHT assumes the stream is independent and identically distributed to converge to the same tree in the static non-streaming case, leaving it ineffective for situations where the correlations of the sensitive attribute and target label change over time.

In this work we will address both of these limitations to FAHT to create the first method that can adjust the accuracy/discrimination trade-off and learn a useful model in the face of concept drift. This is achieved by extending the FAHT model in two ways: 1) by introducing a flexible fair splitting criterion that allows for flexible control on the fairness gain and information gain for splitting candidate determination when inducing the tree (c.f., Section V-A) ; and 2) by adding the ability to detect and act more promptly to the evolution of underlying distribution (c.f., Section V-B).

V. 2FAHT: Flexible and Adaptive Fairness-Aware Hoeffding Tree Classifier

This section first outlines the flexible fair information gain splitting criterion for application-wise fairness-aware learning, followed by the adaption of changes in the example-generating process for fairness-aware learning in non-stationary data stream settings. A number of refinements and modifications that instantiate the flexible and adaptive learning process are also specified.

A. The Flexible Fair Information Gain

FAHT integrates the fairness merit into the tree induction, results into an accuracy-driven and fairness-oriented induction of the tree. However, such induction process is internally working and does not allow for a fine-grained control on the trade-off between accuracy and fairness. To this end, we first introduce the *flexible fair information gain* as:

$$2FIG(D, A) = IG(D, A) \times e^{\gamma \times FG(D,A)} \qquad (5)$$

where γ is a tunable parameter, IG and FG stand for information gain and fairness gain respectively when considering attribute A as a potential split based on the sufficient statistics D, stored in leaves but also non-leaf node to enable concept drift adaptation (c.f., Section V-B). In this form the FG criterion becomes a gating mechanism for the information, and the hyper-parameter γ modules the flow of the gate. Discriminatory splits will receive a low FG and thus modulate the information gain, reducing the likelihood of being selected for splitting. Likewise non-discriminatory splits will receive large FG values which encourages the current IG score. Because tree induction is based on maximum score the gating like approach does not need to be normalized, so we use this simpler non-normalized equation.

We make note that the use of an exponential term to combine multiple factors in decision tree induction is a unique contribution. All prior work we are aware (e.g., [11], [12], [43]) considers only simple addition, subtraction, multiplication and division of competing factors. As we will show later, our exponential modulation allows for a smooth and intuitive practical means of performing this trade-off between accuracy and fairness.

The formulation design of 2FIG also comes with the following considerations. First, the exponential function in 2FIG is used for smoothing. For instance, suppose one attribute A_a has a FG of 0.1 and another attribute A_b has a FG of 0.01. Without the exponential function the weight of A_a will be 10 times of that of A_b assuming equal value of IG and γ = 1. This may overly enforce fairness thus result in underwhelming accuracy. With the exponential function the weight for A_a is 1.09 times of that of A_b. Second, when γ is set as 0, i.e., accuracy is the primary focus of the current application, 2FIG is identical to IG for the completely accuracy-driven model construction. For a positive γ value, the merit of a feature increases with the discrimination reduction of that splitting feature and decreases with the resultant uncertainty increase. That is to say, the increment of γ up-weights FG and correspondingly down-weights IG. So the model favors features that result in a higher discrimination reduction for the more fairness-oriented model construction. This proposed 2FIG is therefore used in replace of FIG for fine-grained fairness-aware learning.

B. 2FAHT: A Flexible and Adaptive Fairness-Aware Hoeffding Tree Classifier

To overcome the previous discussed second drawback of FAHT, we further endow 2FAHT with the ability of change detection and concept forgetting. To this end, 2FAHT keeps its model consistent with the example-generating process of the current stream, creates alternative decision subtrees when evolving data distribution is detected at a node, and replaces its corresponding branch when needed or prune the created alternative decision subtree. Such concept drift could be reflected by whether there is a change in the underlying data distribution resulting in performance deterioration in that node. When considering concept drift indications, compared with the previous online fairness studies [13] following literature from the data stream community which solely focus on predictive accuracy as the performance indication, 2FAHT takes *both* predictive accuracy and discrimination performance into consideration as they may have different non-stationary characteristics. For example, a subgroup could be receiving increased discrimination but has no discernible impact to overall accuracy.

In addition, instead of directly determining whether one node's performance deteriorates in terms of the numerical values of accuracy and fairness, 2FAHT monitors whether there is a new promising attribute at the non-leaf node to reflect new concepts and declares when branch replacement is necessary. We detect such signals by re-leveraging our previously introduced flexible fair information gain, and therefore jointly considers the implications of the evolving data distributions on accuracy as well as fairness. The sketch of 2FAHT is shown in Algorithm 1.

Algorithm 1 2FAHT induction algorithm

```
Input: a discriminated data stream $D$,
       confidence parameter $\delta$,
       the number of examples between checks for
       grownth $n_g$,
       the number of examples between checks for
       drift $n_d$.

2FAHT($D$, $\delta$, $\tau$)
Let $FAHT$ be a tree with a single leaf (the root)
Let $T_{alt}(L)$ be an initially empty set of alternative trees
  for non-leaf node
Let $n_L$ be be the number of examples seen at leaf or
  non-leaf node $L$
Init sufficient statistics at root
for each instance $x$ in $D$ do
    Sort example into leaf $l$ using $2FAHT$
    Update sufficient statistics in $l$ and nodes traversed in the sort
    Increment $n_L$
    2FAHTGrow($x$, FAHT, $\delta$, $\tau$, $n_g$)
    for traversed node that has an alternate tree $T_{alt}$ do
        2FAHTGrow($x$, FAHT, $\delta$, $\tau$, $n_g$)
    end for
    if $n_L$ mode $n_d$ = 0 then
        CheckPromisingSplit(FAHT, $\delta$, $\tau$, $n_d$)
    end if
end for
```

2FAHT first sorts each example from the stream into an appropriate leaf (line 6), depending on the splitting tests presented in 2FAHT to that point. Compared to FAHT, 2FAHT also maintains sufficient statistics of the nodes traversed in the sort in order to update alternative branches (line 7-8) and grow alternative these branches (line 11). The growing procedure of 2FAHT is similar to FAHT (line 9). However, 2FAHT continuously monitors the quality of old search decisions with respect to the latest instances from the data stream (line 14), in order to keep the model it is learning in sync with changes in the example-generating process. Such monitoring is done by periodically checking for promising splits detailed in Algorithm 3. When checking for promising splits, 2FAHT creates an alternative subtree for each node that change in the underlying distribution is detected (line 10-20). Under the condition that an alternative subtree already exists, 2FAHT checks whether the alternative branch performs better than the old branch (line 3). The old branch will be replaced by the alternative one if so (line 4), otherwise the alternative branch will be pruned (line 6). The whole learning process is therefore fairness-aware, flexible and concept-adapting.

Algorithm 2 2FAHT growth algorithm

```
**2FAHTGrow**($x$, FAHT, $\delta$, $\tau$, $n_g$)

 1: if examples seen at $l$ are not all of the same class and $n_l$ mode $n_g$ = 0 then
 2:     Calculate $2FIG_l(A_i)$ for each attribute according to Equation (5)
 3:     Let $A_a$ be the attribute with highest $2FIG_l$
 4:     Let $A_b$ be the attribute with second-highest $2FIG_l$
 5:     Compute Hoeffding bound $\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n_l}}$
 6:     if $A_a \neq A_\emptyset$ and $(2FIG_l(A_a) - 2FIG_l(A_b) > \epsilon$ or $\epsilon < \tau)$ then
 7:         for each branch of the split do
 8:             Start a new leaf and initialize sufficient statistics
 9:         end for
10:     end if
11: end if
```

VI. Experimental Evaluation

Having introduced our 2FAHT algorithm, we will now show it's effectiveness. First we will describe the datasets and sensitive attribute used in our experiments. Second, we will keep γ fixed at 1 and compare with prior approaches. This shows that our method maintains the desirable property of being effective without parameter tuning, making it easier for others to use. Third, we will show that by altering γ we can interpolate between a model that favors accuracy vs discrimination in a continuous manner, giving practitioners the tools to adjust the results to their needs.

A. Datasets and Experimental Setup

The growing concern on the discrimination bias of AI model has motivated a number of studies for the development of discrimination-conscious AI system. However, there is still a lack of datasets and benchmarks [27]. With respect to the highly under-explored fairness-aware learning in massive data streams, this challenge is further amplified as most datasets being used in current fairness studies contain less than 1,000 instances, which does not meet the demanding requirement with respect to the number of instances and drift therein so as to simulate data stream environments [12]. In addition, the problem of discrimination is a very complex problem for learning as it may be due to different factors such as attributes or subsets that act as proxies to the sensitive attribute. Such a behavior cannot be easily replicated in other datasets nor synthetically generated [21]. Therefore, we focus on evaluating our approach on the real fairness datasets used in the recent work of this research direction [12], [18], the $Adult$ and the $Census$ [44] datasets both aiming at determining whether a person makes over 50K dollars per annum.

Algorithm 3 Check promising split algorithm

```
**CheckPromisingSplit**(FAHT, $\delta$, $\tau$, $n_d$)

 1: for each node L in FAHT that is not a leaf do
 2:     for each tree in $T_{alt}$ in $T_{alt}(L)$ do
 3:         if $T_{alt}$ is more accurate or fair then
 4:             replace current node with its $T_{alt}$
 5:         else
 6:             prune its $T_{alt}$
 7:         end if
 8:         CheckPromisingSplit(FAHT, $\delta$, $\tau$, $n_d$)
 9:     end for
10:     Let $A_L$ be the split attribute at $L$
11:     Let $A_a$ be the attribute with the highest $2FIG_L$ other than $A_L$
12:     Let $A_b$ be the attribute with second-highest $2FIG_L$ other than $A_L$
13:     if $2FIG_L(A_a) - 2FIG_L(A_b) \geq 0$ and $A_a$ has not been used as the root node in $T_{alt}(L)$ then
14:         Compute Hoeffding bound $\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n_L}}$
15:         if $A_a \neq A_\emptyset$ and $(2FIG_l(A_a) - 2FIG_l(A_b) > \epsilon$ or $\epsilon < \tau)$ then
16:             for each branch of the split do
17:                 Start a new leaf and initialize sufficient statistics
18:             end for
19:         end if
20:     end if
21: end for
```

The first $Adult$ dataset contains 48,843 instances described by 14 employment and demographic attributes (attribute “fnlwg” is removed as suggested). For fair comparison, we follow the same options of [12], [18] in our experiments by setting “gender” as the sensitive attribute with female being the sensitive value and an annual income of more than 50K as the target class, i.e., the positive classification. The discrimination level of the whole dataset is 19.45% according to Equation (1). The second $Census$ dataset has an identical prediction task as the $Adult$ dataset but is significantly larger in size including 299,285 instances and 41 attributes. The settings of sensitive attribute, sensitive value and positive classification remain the same, and the intrinsic discrimination is 7.63%.

Most existing fairness works process these two datasets in a static manner and there is no temporal information. We turn them into data stream by randomizing the order and process them in sequence. The first-test-then-train or prequential evaluation set up [20] is employed for evaluation, in which each incoming instance is first being predicted upon arrival then is available for model training and cannot be reaccessed or being stored.

B. Minimizing Discrimination While Maximizing Accuracy

This section first experiments 2FAHT's devised discrimination eliminating and adaptation capabilities when addressing discrimination in non-stationary data stream settings. Note that for fair comparison, the tunable parameter γ is fixed at the value of 1 to disable tuning. We compare against three recently proposed fairness-aware online learner FEI [18] as well as FAHT [12] and its extension FEAT [13] along with two baselines therein, the Hoeffding Tree (HT) and KHF which incorporates the discrimination-aware splitting criterion of [11] into HT. We do not compare with other competing fairness methods since none of them is capable of addressing fairness in online settings. We further implemented an exclusively concept-adapting oriented online learner, denoted HAT [34], from the data stream mining literature as a baseline. Table I summarizes the obtained results from 2FAHT and all baselines trained the same way for all datasets.

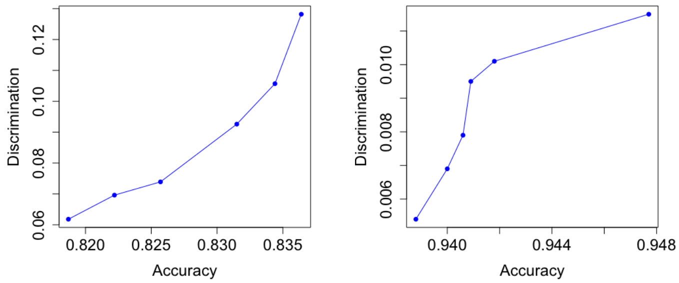
TABLE I
MINIMIZING DISCRIMINATION WHILE MAXIMIZING ACCURACY CAPABILITY BETWEEN 2FAHT AND BASELINE MODELS. THE BEST RESULTS ARE IN BOLD, AND PERCENTAGE IN PARENTHESIS IS THE RELATIVE DIFFERENCE OVER THE PERFORMANCE OF THE BEST BASELINE METHOD, WHICH DEMONSTRATED 2FAHT HAS THE LOWEST DISCRIMINATION WITH ONLY MINOR REDUCTION TO ACCURACY.

Metric Methods	**Adult dataset**		**Census dataset**	
	Disc%	Acc%	Disc%	Acc%
HT	22.59	83.91	6.84	95.06
FEI	22.16	75.51	6.34	81.26
FAHT	16.29	81.83	3.20	94.28
FEAT	15.26	84.01	1.25	95.03
HAT	22.30	**84.70**	6.54	**95.64**
KHT	22.61	83.92	6.59	94.82
2FAHT	**12.82** **(-15.99%)**	83.64 **(-1.25%)**	**1.15** **(-8.0%)**	94.77 **(-0.91%)**

As we can see from table I, 2FAHT achieves the lowest cumulative discrimination scores while maintaining fairly comparable predictive performance on all datasets. Specially, 2FAHT outperforms the best baseline by 15.99% and 8.0% discrimination reduction on $Adult$ and $Census$ dataset respectively, while the largest margin could be as high as 43.3% when comparing to the fairness performance of KHT. With respect to predictive performance, 2FAHT is the second best predictive classifier, and the most accurate baseline, i.e., HAT, narrowly outruns 2FAHT by respective 1.25% and 0.91%. This is expected as HAT is exclusively concept-adapting oriented by design and its discrimination levels are 42.51% and 82.42% higher than 2FAHT's. We can also observe that FEI is poorly performed although it is proposed for addressing online fairness. This verifies that a simple/direct combination of existing techniques from corresponding fairness-aware learning and data stream mining communities cannot handle the complex online fairness effectively. On the other hand, 2FAHT's theoretical design which jointly considers data encoding, discrimination reduction and concept-adapting indeed improves fairness with minimal accuracy loss in evolving online settings.

C. Minimizing Discrimination Under Accuracy Constrains

We then investigate 2FAHT's fine-grained control on the trade-off between fairness and accuracy for the sake of instantiating application-wise fairness-aware learning. To this end, we adjust the value of γ to minimize discrimination while controlling loss in accuracy so as to meet certain performance-related constraints such as the "business necessity" clause [16]. Figure 1 visualizes the results by showing the fairness and accuracy subject to accuracy constraints with different values of γ. As expected, as the value of γ increases, i.e., the current application focuses more on fairness, the fairness upvaluing is accompanied by the downvalued accuracy.



(a) Auldt dataset

(b) Census dataset

Fig. 1. The accuracy-vs-discrimination trade-off fined-grained by tunable parameter γ. The values of γ range from 100000 to 1 with the common ratio equals to 10.

VII. CONCLUSIONS

This paper focuses on the highly under-explored discrimination-aware learning in evolving data streams. To address this challenge, we propose 2FAHT with embedded flexible fair splitting criterion and endow it with the ability of change detection and concept forgetting to handle discriminated and non-stationary data streams. What's more, 2FAHT moves one step further to allow for application-wise fairness-aware learning. The positive results of conducted experiments show the flexibility and versatility of 2FAHT in online settings. One immediate future direction is to extend these results in conjunction with our previous works [45], [46] for the fair allocation of health care resources. We also plan to apply these results along with our initial work in [47] targeting the more challenging unsupervised fair clustering domain.

References

[1] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, “When will ai exceed human performance? evidence from ai experts,” *Journal of Artificial Intelligence Research*, vol. 62, pp. 729–754, 2018.

[2] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, “Data decisions and theoretical implications when adversarially learning fair representations,” *arXiv preprint arXiv:1707.00075*, 2017.

[3] M. Du, F. Yang, N. Zou, and X. Hu, “Fairness in deep learning: A computational perspective,” *IEEE Intelligent Systems*, 2020.

[4] I. Y. Chen, P. Szolovits, and M. Ghassemi, “Can ai help reduce disparities in general medical and mental health care?” *AMA journal of ethics*, vol. 21, no. 2, pp. 167–179, 2019.

[5] A. Datta, M. C. Tschantz, and A. Datta, “Automated experiments on ad privacy settings,” *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, 2015.

[6] F. Kamiran and T. Calders, “Classifying without discriminating,” in *2nd International Conference on Computer, Control and Communication*, 2009, pp. 1–6.

[7] T. Calders, F. Kamiran, and M. Pechenizkiy, “Building classifiers with independency constraints,” in *ICDMW*, 2009, pp. 13–18.

[8] S. Aghaei, M. J. Azizi, and P. Vayanos, “Learning optimal and fair decision trees for non-discriminative decision-making,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[9] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, “Fairness constraints: Mechanisms for fair classification,” *arXiv preprint arXiv:1507.05259*, 2015.

[10] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, “Discrimination-and privacy-aware patterns,” *Data Mining and Knowledge Discovery*, vol. 29, no. 6, pp. 1733–1782, 2015.

[11] F. Kamiran, T. Calders, and M. Pechenizkiy, “Discrimination aware decision tree learning,” in *ICDM*, 2010, pp. 869–874.

[12] W. Zhang and E. Ntoutsi, “Faht: an adaptive fairness-aware decision tree classifier,” in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 1480–1486.

[13] W. Zhang and A. Bifet, “Feat: A fairness-enhancing and concept-adapting decision tree classifier,” in *Proceedings of the 23rd International Conference on Discovery Science*. Springer, 2020.

[14] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, “Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment,” in *World Wide Web*, 2017, pp. 1171–1180.

[15] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, “Fairness constraints: A flexible approach for fair classification.” *J. Mach. Learn. Res.*, vol. 20, no. 75, pp. 1–42, 2019.

[16] S. Barocas and A. D. Selbst, “Big datas disparate impact,” *California Law Review*, vol. 104, no. 3, p. 671, 2016.

[17] J. Sylvester and E. Raff, “What about applied fairness?” in *Machine Learning: The Debates (ML-D) organized as part of the Federated AI Meeting*, 2018.

[18] V. Iosifidis, T. N. H. Tran, and E. Ntoutsi, “Fairness-enhancing interventions in stream classification,” in *International Conference on Database and Expert Systems Applications*. Springer, 2019, pp. 261–276.

[19] W. Zhang, X. Tang, and J. Wang, “On fairness-aware learning for non-discriminative decision-making,” in *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2019, pp. 1072–1079.

[20] J. Gama, *Knowledge discovery from data streams*. Chapman and Hall/CRC, 2010.

[21] S. Verma and J. Rubin, “Fairness definitions explained,” in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 2018, pp. 1–7.

[22] M. Hardt, E. Price, N. Srebro *et al.*, “Equality of opportunity in supervised learning,” in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[23] M. Skirpan and M. Gorelick, “The Authority of ”Fair” in Machine Learning,” in *FAT ML Workshop*, 2017. [Online]. Available: http://arxiv.org/abs/1706.09976

[24] R. Binns, “Fairness in machine learning: Lessons from political philosophy,” in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 149–159.

[25] M. Srivastava, H. Heidari, and A. Krause, “Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &amp; Data Mining*, 2019, pp. 2459–2468.

[26] J. Goldsmith and E. Burton, “Why Teaching Ethics to AI Practitioners Is Important,” in *The AAAI-17 workshop on AI, Ethics, and Society*, 2017, pp. 110–114.

[27] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi, “Putting fairness principles into practice: Challenges, metrics, and improvements,” *AAAI Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2019.

[28] B. Fang, M. Jiang, P.-y. Cheng, J. Shen, and Y. Fang, “Achieving outcome fairness in machine learning models for social decision problems,” in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 444–450.

[29] T. Calders and S. Verwer, “Three naive bayes approaches for discrimination-free classification,” *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[30] W. Zhang, “Phd forum: Recognizing human posture from time-changing wearable sensor data streams,” in *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2017, pp. 1–2.

[31] J. Wang, Z. Huang, W. Zhang, A. Patil, K. Patil, T. Zhu, E. J. Shiroma, M. A. Schepps, and T. B. Harris, “Wearable sensor based human posture recognition,” in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 3432–3438.

[32] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, “Ensemble learning for data stream analysis: A survey,” *Information Fusion*, vol. 37, pp. 132–156, 2017.

[33] W. Zhang and J. Wang, “A hybrid learning framework for imbalanced stream classification,” in *2017 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2017, pp. 480–487.

[34] A. Bifet and R. Gavaldà, “Adaptive learning from evolving data streams,” in *International Symposium on Intelligent Data Analysis*. Springer, 2009, pp. 249–260.

[35] J. Montiel, R. Mitchell, E. Frank, B. Pfahringer, T. Abdessalem, and A. Bifet, “Adaptive xgboost for evolving data streams,” *arXiv preprint arXiv:2005.07353*, 2020.

[36] A. Bifet and E. Frank, “Sentiment knowledge discovery in twitter streaming data,” in *International conference on discovery science*. Springer, 2010, pp. 1–15.

[37] P. Domingos and G. Hulten, “Mining high-speed data streams,” in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000, pp. 71–80.

[38] G. Forman, “Tackling concept drift by temporal inductive transfer,” in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 252–259.

[39] J. Read, N. Tziortziotis, and M. Vazirgiannis, “Error-space representations for multi-dimensional data streams with temporal dependence,” *Pattern Analysis and Applications*, vol. 22, no. 3, pp. 1211–1220, 2019.

[40] I. Khamassi, M. Sayed-Mouchaweh, M. Hammami, and K. Ghédira, “Discussion and review on evolving data streams and concept drift adapting,” *Evolving systems*, vol. 9, no. 1, pp. 1–23, 2018.

[41] P. Ksieniewicz, M. Woźniak, B. Cyganek, A. Kasprzak, and K. Walkowiak, “Data stream classification using active learned neural networks,” *Neurocomputing*, vol. 353, pp. 74–82, 2019.

[42] L. Zhang and W. Zhang, “A comparison of different pattern recognition methods with entropy based feature reduction in early breast cancer classification,” *European Scientific Journal*, vol. 3, pp. 303–312, 2014.

[43] E. Raff, J. Sylvester, and S. Mills, “Fair Forests: Regularized Tree Induction to Minimize Model Bias,” in *AAAI / ACM conference on Artificial Intelligence, Ethics, and Society*, 2018. [Online]. Available: http://arxiv.org/abs/1712.08197

[44] D. Dheeru and E. Karra Taniskidou, “UCI machine learning repository,” 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[45] W. Zhang and J. Wang, “Content-bootstrapped collaborative filtering for medical article recommendations,” in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1184–1188.

[46] W. Zhang, J. Tang, and N. Wang, “Using the machine learning approach to predict patient survival from high-dimensional survival data,” in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2016, pp. 1234–1238.

[47] W. Zhang, J. Wang, D. Jin, L. Oreopoulos, and Z. Zhang, “A deterministic self-organizing map approach and its application on satellite data based cloud type classification,” in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 2027–2034.