FEAT: A Fairness-enhancing and Concept-adapting Decision Tree Classifier

Wenbin Zhang[1] and Albert Bifet[2,3]

[1] University of Maryland, Baltimore County, MD 21250, USA
wenbinzhang@umbc.edu
[2] University of Waikato, Hamilton 3216, New Zealand
[3] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
albert.bifet@waikato.ac.nz

Abstract. Fairness-aware learning is increasingly important in socially-sensitive applications for the sake of achieving optimal and non-discriminative decision-making. Most of the proposed fairness-aware learning algorithms process the data in offline settings and assume that the data is generated by a single concept without drift. Unfortunately, in many real-world applications, data is generated in a streaming fashion and can only be scanned once. In addition, the underlying generation process might also change over time. In this paper, we propose and illustrate an efficient algorithm for mining fair decision trees from discriminatory and continuously evolving data streams. This algorithm, called FEAT (Fairness-Enhancing and concept-Adapting Tree), is based on using the change detector to learn adaptively from non-stationary data streams, that also accounts for fairness. We study FEAT's properties and demonstrate its utility through experiments on a set of discriminated and time-changing data streams.

Keywords: AI ethics · online fairness · online classification.

1 Introduction

Artificial Intelligence (AI)-based decision making systems are routinely being used in both online as well as offline settings to assist or even completely automate the decision-making. Yet, these automated data-driven tools may, even in the absence of intent, lead to a loss of fairness and accountability in the employed models. A plethora of such kind of AI-based discriminatory incidents have been observed and reported [1, 2, 7, 12]. As a recent example, the AI algorithm behind Amazon Prime has suggested signs of racial discrimination when deciding which areas of a city are eligible for advanced services [13]. Areas densely populated by black people are excluded from services and amenities even though race is blind to the AI algorithm. Such incidents have sparked heated debate on the bias and discrimination in AI decision systems, pulling in scholars from a diverse of areas such as philosophy, law and public policy.

The growing concern over discriminative behavior of AI models has motivated a number of approaches, ranging from defining discrimination to discrimination

discovery and prevention for the development of AI tools that are discrimination-conscious by-design. Up to now, more than twenty notions have been proposed to measure the discriminative behavior of AI models [19]. One of the most widely used measures is the *statistical parity* [19] which examines whether the probability of being assigned a positive target class, for example allocating healthcare resources, is the same for both privileged and unprivileged groups. Formally put:

$$Discrimination(D) = \frac{PP}{PP + PN} - \frac{UP}{UP + UN} \tag{1}$$

where D is the labeled dataset, PP and PN refer to privileged community receiving positive and negative classification, respectively. So are UP and UN for unprivileged community. Here, the attribute that distinguishes privileged groups from unprivileged ones is referred as the *sensitive attribute* with the *sensitive value* defining the unprivileged community. Take "race" as the *sensitive attribute* for example, then the *sensitive value* is "black" and the positive class value as allocating healthcare resources. The four communities PP, PN, UP and UN therefore represents "non-black" being allocated healthcare resources, "non-black" being denied healthcare resources, "black" receives healthcare resources and "black" does not receive healthcare resources, respectively.

The aim of fairness-aware learning is then to train a decision model which provides accurate predictions, yet does not unduly bias against unprivileged groups. That is to say, from *statistical parity* point of view, equally granting a benefit to both privileged and unprivileged groups. While a large number of methods have been proposed to achieve this goal, most of them tackle fairness as a static problem. In many applications, however, data is generated sequentially and its characteristics might also evolve over time. Therefore, fairness-aware learning for such sort of applications should also be able to adapt to non-stationary distribution simultaneously.

Compared with the booming approaches in static settings, fairness-aware learning in data stream is highly under-explored because of its significant challenges [23]. To address this issue, this paper introduces a fairness-enhancing classifier that also equips with drift adaptation capability. The contribution of this paper is three-fold:

- We define the problem of fairness-aware learning in non-stationary data distribution. Then, we propose FEAT, a discrimination-conscious learner with add-on concept drift adaptation ability to handle discriminated and non-stationary data streams.
- We introduce fair-enhancing information gain that also accounts for the local discrimination to maximize the cumulative fairness, thus providing enhanced fairness-awareness learning.
- The conducted experiments verify the capability of the proposed model in online settings. *To the best of our knowledge, this is the first work that jointly addresses fairness and concept drift.*

The rest of the paper is organized as follows. Related studies are first reviewed in Section 2. We describe the proposed FEAT in Section 3 and discuss the experimental results in detail in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

The tremendous societal importance of AI fairness has arose growing concern with ever increasing amount of discrimination-conscious models being proposed [1, 2, 24]. These approaches typically can be 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 rationale 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 [15] and reweighting [5]. 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 protected group will receive higher weighs comparing to instances from the unprotected group. In [14], these two methods have also been extended for online classification. However, methods in this category are typically not quite effective as standalone approaches 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. [16] 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 [20], the measure of "decision boundary fairness" is leveraged to penalize discrimination in the formulation of a set of convex margin-based classifiers. More recently, [23] improves the splitting strategy of [16] and operates their model in the online setting. However, research efforts in this direction have still been limited. Our work situates in this highly under-explored research direction to provide fair online decision-making.

The last category, *postprocessing techniques*, consists of either adjusting the decision boundary of a model or directly changing the prediction labels. [12] processes with additional prediction thresholds to work against discrimination while the decision boundary of AdaBoost is shifted w.r.t. fairness in [6]. The

latter approaches pay attention to the outcome of a classifier. In [16], 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.

Fairness in data streams further requires the addressing of non-stationary distributions, known as concept drift [4, 10, 22, 25]. The learning algorithms therefore should be able to remain stable on previously learned and not outdated concepts while adapting to such drifts. The adaptation is typically enabled by learning incrementally from new instances [11, 17] and by forgetting outdated information from the model [4, 18]. A significant amount of work has been done with respect to this specific issue. However, the combined approach of addressing both fairness and concept drift has enjoyed relatively little research. Our work situates in this research direction to enable fairness-aware learning in non-stationary data streams.

3 FEAT: Fairness-Enhancing and concept-Adapting Tree

This section first outlines the vanilla Hoeffding Tree (HT), then the reformulated fair information gain splitting criterion for fairness enhancement is introduced, followed by the adaption of changes in the example-generating process. A number of refinements and modifications that instantiate the fairness enhancement and concept-adapting learning are specified thereafter.

3.1 The Hoeffding Tree (HT) Classifier

Our Fairness-Enhancing and concept-Adapting hoeffding Tree (FEAT) is built on top of the Hoeffding Tree (HT) classifier [9]. To mine high-speed data stream, 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. The crux 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 Hoeddding bound [9] to guarantee that the tree learned probably converges to the conventional static tree built by a batch learner, given enough examples. In HT, these two decisions are based on the $information\ gain$, which is exclusively accuracy-oriented and does not consider fairness. In addition, the construction of tree assumes the distribution generating examples does not change over time.

In this work, to enable fairness-aware learning and concurrently adapt to non-stationary data distributions, we extend the HT model in two ways: i) by introducing an enhanced fair splitting criterion that enables the fairness-aware learning (c.f., Section 3.2) and ii) by adding the ability to detect and adapt to the evolution of underlying distribution (c.f., Section 3.3).

3.2 The Fair-Enhancing Information Gain

The *information gain* (IG) [21] splitting criterion measures the uncertainty reduction due to a split during the tree construction. It is proposed purely from the data encoding perspective without considering fairness of the tree construction. To address the fairness-free issue of IG, previous studies reformulate the IG by incorporating the discrimination gain into the splitting criterion of the decision tree construction [16, 23]. Inspired by these ideas, we propose the *fair-enhancing information gain (FEIG)* as follows,

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

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 FEG refers to *fair-enhancing gain (FEG)* that measures the difference in discrimination due to the split and is formulated as:

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

where each corresponding discrimination value *Disc* is gauged according to Equation (1).

In *fair-enhancing gain*, different from the previous proposed fair splitting criteria [16, 23], the gain in fairness is directly gauged according to the discrimination difference due to the split rather than entropy in regards to the sensitive attribute. In addition, in fairness-aware learning, it is expected that all groups being treated equally regardless of their population sizes. That is to say, discrimination is discrimination regardless the number of population being discriminated. To align with this idea, our splitting evaluation metric also cares for local discrimination to maximize the cumulative fairness by assigning equal weights to different discrimination representations. Specifically, each partition induced by the attribute A contributes equally to the cumulative fairness of A regardless the number and size of branches. In the general case, the higher reduction in discrimination the merrier, the *fair-enhancing gain* therefore would like a larger merit to be assigned when evaluating the fairness suitability of a candidate splitting attribute and ignores the number of its distinct values and of each specific value.

The FEG is then tied with IG through multiplication as the FEIG. 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., FEG is a positive value. In the end, this conjunctive metric would be used as the alternative fair-enhancing splitting criterion during the construction of the tree to enable discrimination-aware learning while maintaining predictive performance over the course of the stream.

3.3 The FEAT Algorithm

HT learns incrementally from the high-speed data streams by incorporating the incoming data in the stream into the model while simultaneously maintaining the performance of the classifier on the previous information. The tree is adapted, in practice grow, based on the newly available data in the stream and does not forget the obsolete concept that not following the current example-generating process. Therefore, HT assumes the distribution generating examples does not change over time and cannot adapt to the evolving example-generating process.

To overcome this drawback, we further extend HT and propose FEAT which maintains HT's capabilities of processing high speed data stream and data-driven encoding, also with enhanced fairness-aware learning by employing the previous introduced *fair-enhancing information gain* as well as the ability of change detection and concept forgetting.

To detect and react promptly to the evolution of the stream, FEAT keeps its model consistent with the example-generating process of the current stream, creates and replaces alternative decision subtrees when evolving data distribution is detected at a node. FEAT extends HT which is incremental, so the tree is adapted based on new instances. General speaking, the performance of such model, under stationary distribution without drift, improves over the course of the stream as it generalizes better after incorporating more examples into the model. Therefore, performance deterioration is a good indicator of drift. FEAT employs the sliding window size free ADWIN [3] to monitor the error rate of the non-leaf node and declare when branch replacement is necessary. ADWIN recomputes online whether two "large enough" subwindows of the most recent data exhibit "distinct enough" averages, and the older portion of the data is dropped when such distinction is detected. ADWIN therefore eases the burden of selecting a fixed window size that the distribution likely remains to be stationary within this window and adapts to the rate of change observed in the data itself. The use of ADWIN and the sketch of FEAT is shown in Algorithm 1.

FEAT grows similarly to HT (line 1-2 and 6-16). The difference is that HT depends on IG while FEAT employs FEIG to enable accuracy-oriented and fairness-enhanced construction of the tree. What's more, in order to keep the model it is learning in sync with changes in the example-generating process, FEAT continuously monitors the quality of old search decisions with respect to the latest instances from the data stream (line 17). FEAT creates an alternative subtree for each node that change in the underlying distribution is detected by ADWIN (line 19). Under the condition that an alternative subtree already exists, FEAT checks whether the alternative branch performs better than the old branch (line 21). The old branch will be replaced by the alternative one if so (line 22), otherwise the alternative branch will be pruned (line 24). Compared to HT, FEAT also maintains sufficient statistics of the nodes traversed in the sort in order to update alternative branches (line 3-5). The learning process is therefore fairness-enhancing and concept-adapting.

Algorithm 1 The FEAT induction algorithm

Input: a discriminated data stream D,
confidence parameter δ,
tie breaking parameter τ.

FEAT(D, δ, τ)

Let $FEAT$ be a tree with a single leaf (the root)
Init sufficient statistics at root
for each instance x in D **do**
  FEATGrow(x, FEAT, δ, τ)
end for

FEATGrow(x, $FEAT$, δ, τ)

Sort example into leaf l using $FEAT$
Update sufficient statistics in l and nodes traversed in the sort
for traversed node that has an alternate tree T_{alt} **do**
  FEATGrow(x, T_{alt}, δ, τ)
end for
if examples seen at l are not all of the same class **then**
  Calculate $FEIG_l(A_i)$ for each attribute according to Equation (2)
  Let A_a be the attribute with highest $FEIG_l$
  Let A_b be the attribute with second-highest $FEIG_l$
  Compute Hoeffding bound $\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n_l}}$
  if $A_a \neq A_\emptyset$ and ($FEIG_l(A_a) - FEIG_l(A_b) > \epsilon$ or $\epsilon < \tau$) **then**
    for each branch of the split **do**
      Start a new leaf and initialize sufficient statistics
    end for
  end if
end if
for non-leaf node that its $ADWIN$ detects change **do**
  if T_{alt}== null **then**
    Create an alternative subtree T_{alt}
  else
    if T_{alt} is more accurate **then**
      replace current node with its T_{alt}
    else
      prune its T_{alt}
    end if
  end if
end for

3.4 The FEAT System

Our FEAT induction algorithm is built on top of the HT classifier. FEAT therefore still holds HT's theoretical guarantees and theorems can be proven accordingly. Moreover, FEAT aims at enhancing fairness-aware learning while optimizing predictive performance by alleviating the discrimination bias towards the unprivileged group through the proposed fair splitting criterion, the *fair-enhancing information gain* (Section 3.2), and by equipping itself with the ability of change detection and concept forgetting (Section 3.3). The modifications and refinements being included to Algorithm (1) to instantiate the fairness-enhancing and concept-adapting learning over streams are discussed hereafter.

Pre-pruning. HT detects the case of not splitting a node benefits more than splitting by considering the merit of no split, represented by the null attribute $X_\emptyset$ at each node to enable pre-pruning. A node is thus only allowed to split when the candidate attribute is sufficiently better, according to the same Hoeffding bound test that determines differences among other attributes, than $X_\emptyset$. In the implementation of FEAT, the merit to be maximized is the previous introduced FEIG. Thus, the FEIG of the best split found should be sufficiently better than $X_\emptyset$'s. In terms of the FEIG of the null attribute, the current level of class distribution and discrimination are used to represent IG and FEG, respectively.

Sufficient statistics. HT briefly inspects each instance in the stream only once and store sufficient information in the leaves to enable the calculation of the splitting merit afforded by each possible split. In FEAT, the statistics required for the calculation of FEIG should also be maintained. For the *discrete attributes*, each node in the tree maintains a separate table per attribute, containing the counts of the class labels that apply for each attribute value for the calculation of IG, and the counts of unprivileged group and privileged group as well as receiving positive classification in unprivileged group and privileged group that apply for each attribute value for the calculation of FEG. The learning process updates appropriate entries based on the attribute value, sensitive attribute value and class of the examples over the stream accordingly. As for the *numeric attribute*, FEAT maintains a separate Gaussian distribution per class label that apply for each attribute. So are the four previous mentioned FEG calculation related statistics. The appropriate distribution statistics is updated according to the sensitive attribute value and class of the examples over the stream. The most appropriate binary split point for each distribution is evaluated based on the allowing test and the merit of each allowed threshold candidate is also calculated according to the proposed FEIG. With the selected split points, the weight of values to their either side are approximated for each class and four FEG calculation related statistics, and the FEIG merit of each

numeric attribute candidate is thus computed from these weights.

Memory management. Efficient storage of the sufficient statistics is crucial in stream environment. In case of the non-leaf node, FEAT prunes the alternative branch if its performance is inferior to the old one. FEAT also reduces the size of the sufficient statistics in each leaf by removing poor attributes when their FEIG is less than the current best attribute by more than the Hoeffding bound. The rationale is that, according to the bound, such attributes are unlikely to be selected in that leaf. In addition, assuming there are d attributes with a maximum number of v values per attribute and c possible classes in total, the required memory of FEAT is $O((d+2)vc)$ compared to the $O(dvc)$ of HT. FEAT therefore incurs negligible extra costs especially when $d \gg 2$.

4 Experimental Evaluation

In this section, we conduct experiments to evaluate the accountability and fairness of the proposed discrimination-aware data stream learner. To this end, we first investigate the enhanced discrimination reduction capability of the proposed fair splitting criterion. We also show a comprehensive quantitative evaluation to verify the concept adaptation capability of our approach.

4.1 Dataset

Contrary to a growing body of discrimination-conscious approaches motivated by the increasing attentive AI fairness, related datasets and benchmarks are still in a shortage [1]. With respect to the highly under-explored online fairness, this challenge is further magnified by the drift and the demanding requirement of the number of instances contained therein. We evaluate our approach on the datasets used in the recent work of this research direction [23], the *Adult* and the *Census* datasets [8] both targeting on identical learning task of determining whether a person earns more than 50K dollars per annum.

There are 48,843 instances in the *Adult* dataset and each instance is described by 14 employment and demographic attributes (attribute "fnlwg" is removed as suggested). We follow the same options in [23] by setting "gender" as the sensitive attribute with sensitive value equals to "female" being the protected group. The positive class is people making an annual income of more than 50K dollars. The *Census* dataset is significantly bigger in size including 299,285 instances and 41 attributes. It has an identical prediction task as the *Adult* dataset. So are the setting of sensitive attribute, sensitive value and positive classification. The intrinsic discrimination levels, according to Equation (1), of the these two datasets are 19.45% and 7.63%, respectively.

Existing works mostly address these two datasets from the static learning perspective [19, 24, 27]. In our experiments we randomize the order of the instances then process them in sequence to simulate discriminated data streams,

following [23]. The prequential evaluation [10] is employed in which each incoming instance is first being predicted upon arrival then is available for model training.

4.2 Justification of FEIG

The proposed FEIG is designed to enhance the learning idea of all groups being treated equally regardless of their population sizes for fair-enhancing learning. To validate this enhanced fairness-aware learning, we incorporate FEIG into the model proposed in [23] denoted as FAHT+ and FEAT- representing FEAT driven by the splitting criterion proposed in [23] and compare them respectively. We further incorporates the discrimination-aware splitting criterion of [16] into our model in replacing of FEIG, referred as Kamiran's. We do not incorporate FEIG into their model as it is designed for offline setting. Our motivation for using the identical classifiers is that, since our main interest at this stage is to compare the fair-enhancing learning of FEIG with other discrimination-aware splitting criteria, we would like to minimize the influences on the results from the bias of classifiers due to their versatile difference. The obtained results are shown in Table 1.

Table 1. Accuracy-vs-discrimination between FEIG and other discrimination-aware splitting criteria. Percentage in parenthesis is the relative difference over the performance of its corresponding comparing method.

Metric / Methods	**Adult dataset**		**Census dataset**	
	Discrimination	Accuracy	Discrimination	Accuracy
FAHT	16.29%	81.83%	3.20%	94.28%
FAHT+	15.62% (**-4.11%**)	81.01% (**-1.0%**)	2.61% (**-18.44%**)	92.82% (**-1.55%**)
FEAT-	19.14%	83.76%	2.20%	94.14%
FEAT	15.26% (**-20.27%**)	84.01% (**+0.3%**)	1.25% (**-43.18%**)	95.03% (**+0.95%**)
Kamiran's	22.61%	83.92%	6.59%	94.82%
FEAT	15.26% (**-32.51%**)	84.01% (**+0.11%**)	1.25% (**-81.03%**)	95.03% (**+0.22%**)

As shown in Table 1, it is clear that FEIG consistently enhances the fairness-aware learning by diminishing the discrimination to a lower level while maintaining a high prediction capability. The best discrimination reduction obtained by FEIG is 81.03% on *Census* dataset comparing with the discrimination-aware splitting criterion proposed by Kamiran et al [16]. FEIG's learning idea of all groups being treated equally therefore indeed pushes the discrimination to a lower level, which is consistent with its theoretical design. This enhanced anti-discrimination ability is also statistically verified, comparing to the more effective fair splitting criterion among the baseline criteria, as shown in Table 2.

Table 2. The McNemar's test on the datasets for two different splitting criteria: FEIG and FAHT, testing whether FEIG worked to enhance the positive classification of the unprivileged group.

FAHT+ / FAHT	Adult dataset[1]		Census dataset[2]	
	Granted	Rejected	Granted	Rejected
Granted	716	110	1,120	263
Rejected	173	15,193	468	153,924

[1] Chi-squared = 13.583, df = 1, p-value = 0.0002282
[2] Chi-squared = 56.93, df = 1, p-value 4.516e-14

FEAT / FEAT-	Adult dataset[3]		Census dataset[4]	
	Granted	Rejected	Granted	Rejected
Granted	1,127	80	1,331	359
Rejected	153	14,832	658	153,427

[3] Chi-squared = 22.249, df = 1, p-value = 2.395e-06
[4] Chi-squared = 87.32, df = 1, p-value < 2.2e-16

With respect to the attributes being selected for the construction of trees, both FAHT+ and FEAT select "marital status" as their root on *Adult* dataset, while FAHT and FEAT- are rooted on "age". Neither of these two attributes is discrimination-inclined compared to the root attribute "capital gain" of Kamiran's, which encodes the intrinsic discrimination bias of the historic data as members from the unprivileged group, i.e., the sensitive value is female, are less like to receive higher capital-gain than the privileged group's. On the other hand, age, generally speaking, is positively correlated with income per annum and holds that regardless of the sensitive attribute value. However, it is also possible that age could have local discrimination. That is to say, within a small age range, male could more likely to have a higher income than female as they tend to mature at different ages therefore differ in career age which could reflect income. FEIG's learning idea of all groups being treated equally regardless of their population sizes aims to detect and reflect such type of discrimination encoding. Such fair-enhancing attribute selection can also be concluded from the Pearson correlation coefficients between sensitive attribute and decision boundaries as shown in Table 3. As one can see, FEIG based models' predicted boundaries are less correlated with the sensitive attribute than FAHT's due to its fairness-enhancing ability. In addition, different from the tree induction in static setting, the selected attributes are still splitting candidates for the succeeding splitting selection, such fairness-enhancing decisions therefore have impacts on following decisions as well (feedback loops) and could further enhance fairness-aware learning.

4.3 Drift Adaptation Capability

FEAT is designed for enhanced fairness-aware learning with add-on concept drift adaptation ability to handle non-stationary discriminated data streams. For comparison, we implemented two recently proposed fairness-aware online

Table 3. Pearson analysis on sensitive attribute, predicted decision boundary and actual decision boundary. Comparison values within each cell are formatted by (FAHT: FAHT+‖ FEAT-: FEAT) with results on *Adult* and *Census* dataset in the above and below table, respectively.

Entity	Sensitive attribute	Predicted boundary	Actual boundary
Sensitive attribute	1:1 ‖ 1:1	-0.16:-0.14 ‖ -0.19:-0.14	-0.21:-0.21 ‖ -0.21:-0.21
Predicted boundary	-0.16:-0.14 ‖ -0.19:-0.14	1:1 ‖ 1:1	0.44:0.41 ‖ 0.49:0.50
Actual boundary	-0.21:-0.21 ‖ -0.21:-0.21	0.44:0.41 ‖ 0.49:0.50	1:1 ‖ 1:1

Entity	Sensitive attribute	Predicted boundary	Actual boundary
Sensitive attribute	1:1 ‖ 1:1	-0.09:-0.07 ‖ -0.07:-0.05	-0.16:-0.16 ‖ -0.16:-0.16
Predicted boundary	-0.09:-0.07 ‖ -0.07:-0.05	1:1 ‖ 1:1	0.56:0.53 ‖ 0.57:0.57
Actual boundary	-0.16:-0.16 ‖ -0.16:-0.16	0.56:0.53 ‖ 0.57:0.57	1:1 ‖ 1:1

learners FAHT [23] and FEI [14]. In addition, we compared against two baselines, the Hoeffding Tree (HT) and Kamiran's which incorporates the discrimination-aware splitting criterion of [16] into FEAT in replacing of FEIG. We also trained a concept-adapting learner, denoted HAT [4], as a baseline. All methods are trained in the same way for all datasets and the results are summarized in Table 4.

Table 4. Accuracy-vs-discrimination between FEAT and baseline models. The best performance of the compared baselines is marked in boldface. Percentage in parenthesis is the relative difference over the performance of the best baseline method.

Metric / Methods	**Adult dataset**		**Census dataset**	
	Discrimination	Accuracy	Discrimination	Accuracy
HT	22.59%	83.91%	6.84%	95.06%
Kamiran's	22.61%	83.92%	6.59%	94.82%
FAHT	**16.29%**	81.83%	**3.20%**	94.28%
FEI	22.16%	75.51%	6.34%	81.26%
HAT	22.3%	**84.7%**	6.54%	**95.64%**
FEAT	**15.26%** (**-6.32%**)	**84.01%** (**-0.7%**)	**1.25%** (**-60.94%**)	**95.03%** (**-0.64%**)

As one can see, FEAT consistently pushes the discrimination to lower values while maintaining fairly comparable predictive performance in all datasets. Compared with the best accuracy results, FEAT has a small drop of 0.7% and 0.64% on *Adult* and *Census* dataset, respectively. This is expected as HAT is exclusively accuracy-driven while FEAT optimizes for data encoding as well as enhanced discrimination reduction. In comparison with the most fair baselines, FEAT achieves 6.32% and 60.94% discrimination reduction on *Adult* and *Census* dataset, respectively. We also observe that FEI performances poorly although it is proposed for online setting. This verifies that online fairness cannot be trivially solved by a simple combination of existing techniques from corresponding

communities. We further posit that such theoretical design is fundamental to progress in fairness in evolving data streams and not ad hoc.

5 Conclusions

This paper focuses on the highly under-explored discrimination-conscious learning in evolving data streams. To address this challenge, we propose FEAT with embedded fair-enhancing splitting criterion and further equip it with the ability of change detection and concept forgetting to handle discriminated and non-stationary data streams. The positive results of conducted experiments show the versatility of FEAT in online settings. One immediate future direction is to have an ensemble as random forests based on FEAT. A different avenue is to extend these results in conjunction with our previous work [26] to situations where the class label is not available for fair clustering. Here there are multiple unique challenges including appropriately defining and assessing fairness in the unsupervised scenarios.

References

1. 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.
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. A. Bifet and R. Gavalda. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 443–448. SIAM, 2007.
4. A. Bifet and R. Gavaldà. Adaptive learning from evolving data streams. In *International Symposium on Intelligent Data Analysis*, pages 249–260. Springer, 2009.
5. T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.
6. T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.
7. I. Y. Chen, P. Szolovits, and M. Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.
8. D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.
9. 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*, pages 71–80. ACM, 2000.
10. J. Gama. *Knowledge discovery from data streams*. CRC Press, 2010.
11. H. M. Gomes, J. Read, and A. Bifet. Streaming random patches for evolving data stream classification. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 240–249. IEEE, 2019.

12. M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
13. D. Ingold and S. Soper. Amazon doesnt consider the race of its customers. should it. *Bloomberg News*, 2016.
14. V. Iosifidis, T. N. H. Tran, and E. Ntoutsi. Fairness-enhancing interventions in stream classification. In *International Conference on Database and Expert Systems Applications*, pages 261–276. Springer, 2019.
15. F. Kamiran and T. Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, pages 1–6, 2009.
16. F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE, 2010.
17. B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156, 2017.
18. J. Read, N. Tziortziotis, and M. Vazirgiannis. Error-space representations for multi-dimensional data streams with temporal dependence. *Pattern Analysis and Applications*, 22(3):1211–1220, 2019.
19. S. Verma and J. Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.
20. 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*, pages 1171–1180, 2017.
21. L. Zhang and W. Zhang. A comparison of different pattern recognition methods with entropy based feature reduction in early breast cancer classification. *COBISS. MK-ID 95468554*, page 304, 2014.
22. W. Zhang. Phd forum: Recognizing human posture from time-changing wearable sensor data streams. In *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 1–2. IEEE, 2017.
23. W. Zhang and E. Ntoutsi. Faht: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1480–1486. AAAI Press, 2019.
24. 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)*, pages 1072–1079. IEEE, 2019.
25. W. Zhang and J. Wang. A hybrid learning framework for imbalanced stream classification. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 480–487. IEEE, 2017.
26. 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)*, pages 2027–2034. IEEE, 2018.
27. I. Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.