FAHT: An Adaptive Fairness-aware Decision Tree Classifier (IJCAI 19)

Wenbin Zhang[1] Eirini Ntoutsi[2]

[1]University of Maryland, Baltimore County, MD, USA

[2]Leibniz University Hannover & L3S Research Center, Hannover, Germany

Bias in Machine Learning

AI could be the key to ending discrimination in hiring, but experts warn it can be just as biased as humans

Business Insider

Aaron Holmes, Business Insider
8.10.2019, 18:14

FACEBOOK LINKEDIN TWITTER EMAIL PRINT

REUTERS/Rick Wilking

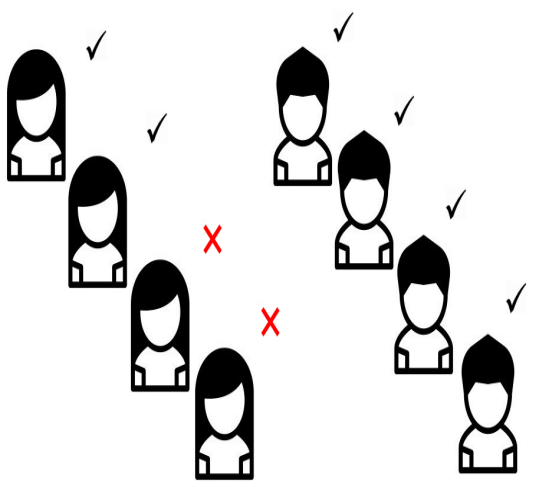
- Employers are increasingly turning to artificial intelligence-driven tools to carry out recruitment and hiring. Companies like Amazon, FedEx, Target, and Capital One have tested or used AI hiring software.

- Current studies tackle fairness as a static/batch problem, we focus on the online setting for data streams.

Notions of Fairness

- More than twenty fairness-related measures has been proposed [Verma and Rubin, 2018].
- We adopt the widely used **statistical parity**:

$$Disc(D) = \frac{FG}{FG + FR} - \frac{DG}{DG + DR}$$



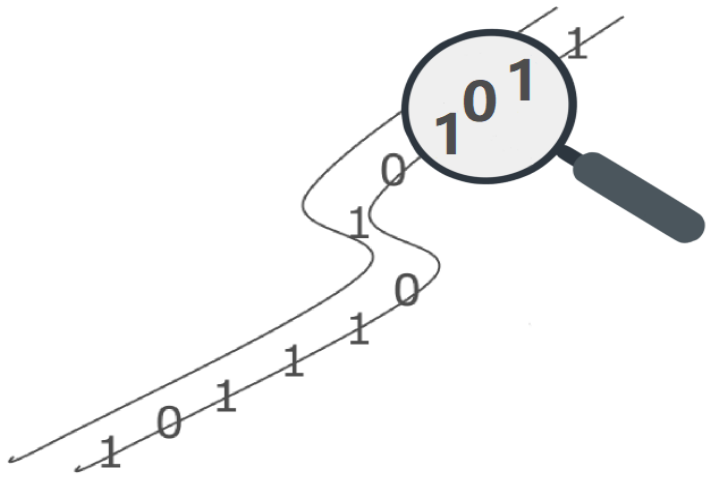
- **DR** (deprived-rejected): females rejected a benefit.
- **DG** (deprived-granted): females granted a benefit.
- **FR** (favored-rejected): males rejected a benefit.
- **FG** (favored-granted): males granted a benefit.

Fairness-aware Learning

- Add regularization terms to the Mixed-Integer Programming model to penalize discrimination [Aghaei et al., 2019].
- Closer to ours: introduce a splitting criterion w.r.t. sensitive attribute and class label [Kamiran et al., 2010].
- Two distinctions:
  - Fairness is directly defined in terms of the discrimination difference of the induction of a split, i.e., the fairness gain due to the split.
  - Our model operates in an online setting rather than upon a static/batch dataset.

Stream Classification

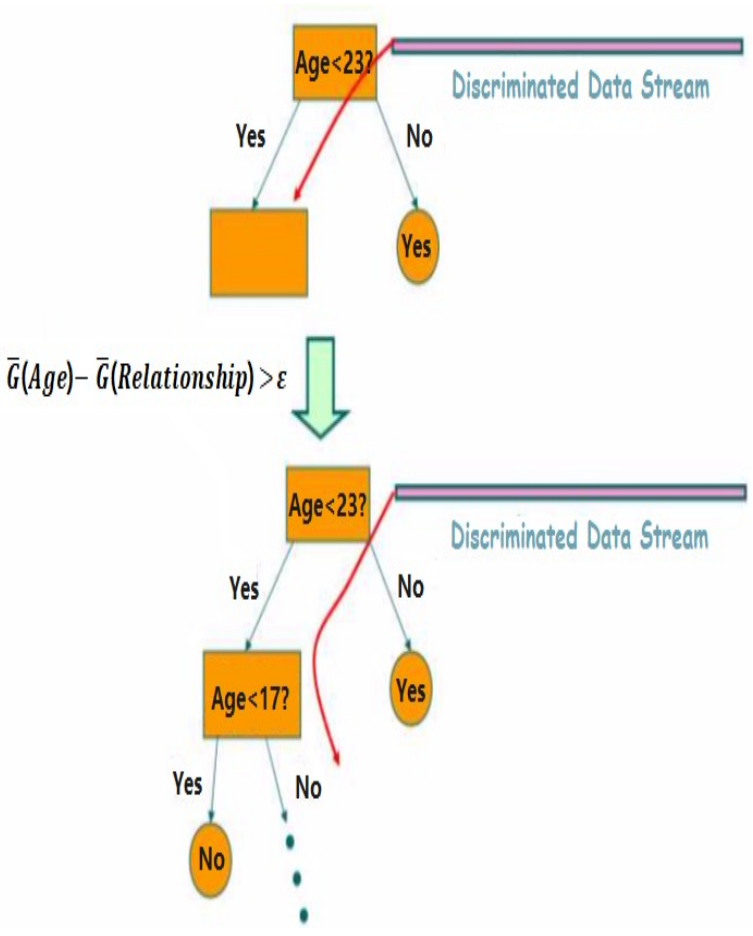
- Continuous flow of data.
- Main challenge: changes in the joint data distribution over time.



- The learning methods therefore should adapt to changes by learning incrementally from new instances [Krawczyk et al., 2017], and by carefully considering historical information into the model [Melidis et al., 2018].
  - Our approach integrates fairness-aware solution and the online approach to maintain a fair and up-to-date classifier for infinite data streams.

Vanilla Hoeffding Tree

- Our Fairness-Aware Hoeffding Tree (FAHT) builds upon the Hoeffding Tree (HT).
- HT scans each instance in the stream only once and stores sufficient information in the leaves for tree growing.
- The crucial decisions are when and how to split a node by Hoeffding bound:
  $$\triangle \overline{G} = \overline{G}(X_a) - \overline{G}(X_b) > \epsilon$$
- Such decisions are based on information gain to optimize predictive performance and do not consider fairness.



Fairness-Aware Hoeffding Tree Classifier

FAHT extends the HT model in two ways:

- Introduce a new splitting criterion that *jointly* considers the gain of an attribute split w.r.t. classification and also w.r.t. discrimination.
- Maintain sufficient statistics at each node to enable the computation of the new splitting criterion values.

The Fair Information Gain Splitting Criterion

HT model:

- Information gain (IG): exclusively accuracy-oriented and fairness is inconceivable.

The Fair Information Gain Splitting Criterion

- ~~Information gain (IG): exclusively accuracy-oriented and fairness is inconceivable.~~
- To overcome this, we propose to alter the splitting criterion to also consider the **fairness gain (FG)**:

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

  - $D_v, v \in dom(A)$ are the partitions induced by A (attribute).
  - $Disc(D_v)$ is computed based on *statistical parity*.
  - The idea of *FG* aligns with *IG* but focuses on discrimination and is also the higher the merrier.
  - Directly defined in terms of the reduction in discrimination rather than mediating between the entropy w.r.t. sensitive attribute.

The Fair Information Gain Splitting Criterion

Fair Information Gain (FIG):

- Combine *FG* and *IG* to a joint objective:

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

- Evaluate the suitability of a splitting attribute in terms of both accuracy and fairness.
- For split that does not change the distribution of discrimination, *FIG* is reduced to *IG*.
- Multiplication is favoured: two metrics are not necessary in the same scale and encourage fair splits.

The FAHT System

- **Pre-pruning.** For the null attribute of *FIG*, the current class distribution is used to represent the *IG* and the *FG* is evaluated as the current level of discrimination.
- **Sufficient statistics.** Keep track of the counts/maintain Gaussian distribution for discrete attributes and numeric attributes, respectively, to evaluate the *FG*.
- **Memory.** The required memory becomes $O((d+2)vc)$ from $O(dvc)$, which incurs negligible extra costs especially when $d \gg 2$ (d attributes with a maximum number of v values and c possible classes).

Evaluation Metrics and Goals

- The predictive- vs fairness-performance.
- Prequential evaluation: first test, on both aggregated measures and over the stream performance, then train.
- Understand the effects of proposed splitting criterion in the structure of the resulting decision tree models.

Datasets

- Still short of datasets for fairness-aware research [Hajian et al., 2016], this challenge is further magnified by the demanding requirement for big non-stationary datasets.
- The ones that best meet streaming requirements are the *Adult* and *Census* datasets both aiming to predict whether individual's annual income will exceed a certain amount.
- Render them as discriminated data streams by randomizing the order of the instances and processing them in sequence.

Accuracy vs. Fairness

- To the best of our knowledge, this is the first work to address discrimination in data stream classification, so we compare FAHT to HT and Kamiran's.

Metric / Methods	Adult dataset		Census dataset	
	Accuracy	Discrimination	Accuracy	Discrimination
HT	83.91%	22.59%	95.06%	6.84%
Kamiran's	83.92% (**+0.01%**)	22.61% (**+0.09%**)	94.82% (**-0.25%**)	6.59% (**-3.65%**)
FAHT	81.83% (**-2.48%**)	16.29% (**-27.89%**)	94.28% (**-8.20%**)	3.20% (**-53.22%**)

- HT induces discriminated trees, and Kamiran's method has little numerical differences comparing to HT.
- FAHT is capable of diminishing the discrimination to a lower level while maintaining a fairly comparable accuracy.

Accuracy vs. Fairness

HT \ FAHT	Adult dataset[1]		Census dataset[2]	
	Granted	Rejected	Granted	Rejected
Granted	527	310	824	963
Rejected	523	14,832	564	153,424

[1] Chi-squared = 53.954, df = 1, p-value = 2.052e-13

[2] Chi-squared = 103.74, df = 1, p-value $<$ 2.2e-16

Table: McNemar's test on deprived community between HT and FAHT applied to each dataset, testing whether FIG worked to benefit the positive classification of the deprived group.

- The anti-discrimination capability of FAHT is also statistically significant.

Applicability and Discrimination Propagation

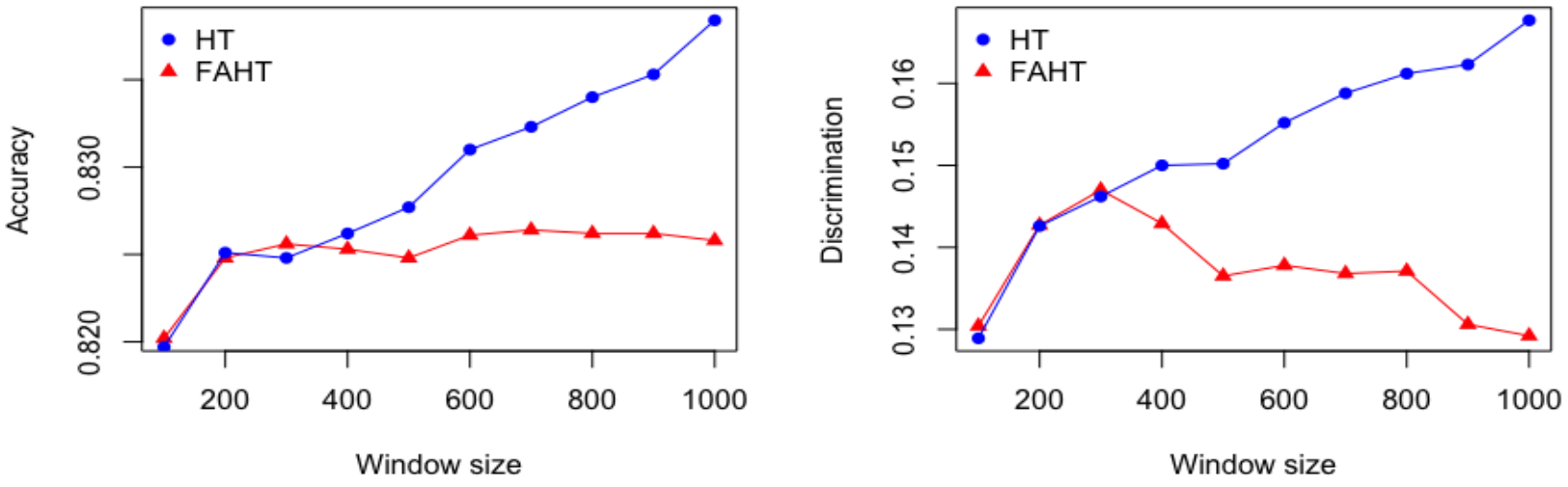


Figure: The Adult data stream is processed in sliding windows; Each window trains a base learner as the ensemble component, the oldest one will be replaced when the classifier window is full; The ensemble members stored in the classifier window will also get updated with the instances in the current sliding window.

- FAHT controls the discrimination propagation while maintaining a high prediction capability.

Structural Effects on the Tree Construction

Selected attributes on the tree construction:

- **HT:** capital-gain(root), capital-loss, relationship, native-country, education, age.
- **FAHT:** age(root), capital-gain, marital-status, relationship.
- Note that the selected splitting attribute is a candidate splitting attribute for the succeeding splitting selection as well.

Structural Effects on the Tree Construction

Entity	Sensitive attribute	Predicted boundary	Actual boundary
Sensitive attribute	1:1	-0.20 : -0.16	-0.21 : -0.21
Predicted boundary	-0.20 : -0.16	1:1	0.52 : 0.44
Actual boundary	-0.21 : -0.21	0.52 : 0.44	1:1

Table: Pearson Correlation coefficients between sensitive attribute, predicted decision boundary and actual decision boundary on Adult dataset. The values before colon are from the HT and after are from FAHT.

- FAHT selects attributes that balance encoding and diminish discrimination of the training data.

Structural Effects on the Tree Construction

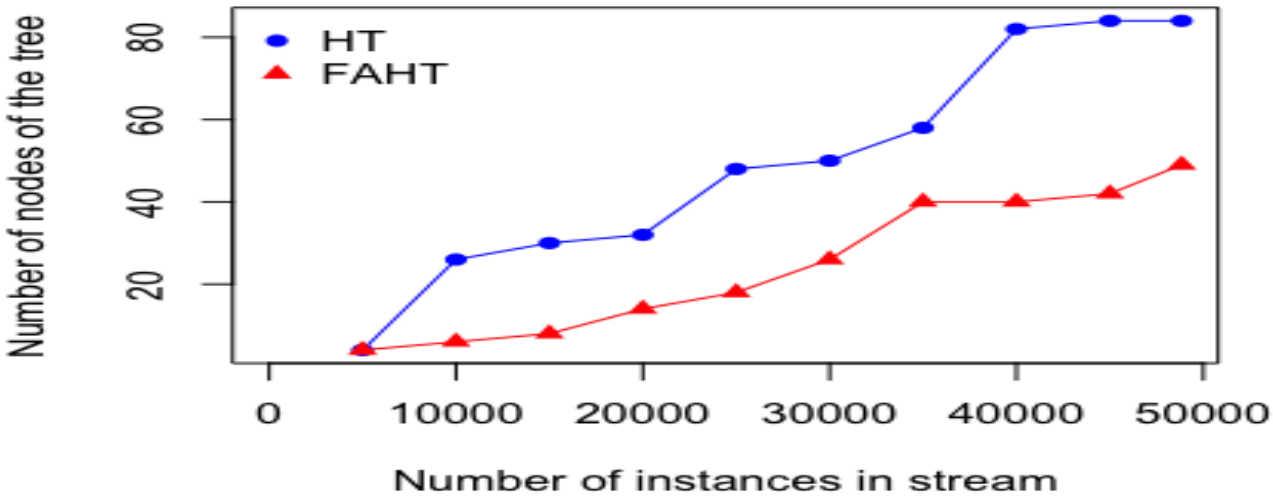


Figure: Adult dataset: Model complexity (number of nodes) over the stream.

- FAHT results in a shorter tree comparing to HT, as its splitting criterion FIG is more restrictive comparing to IG.

- The code and datasets are available at: `https://github.com/vanbanTruong/FAHT`.

- Th
  ht

Thanks! Questions?