Booz | Allen | Hamilton

FAIR FORESTS:
REGULARIZED TREE INDUCTION TO MINIMIZE MODEL BIAS

Jared Sylvester, Edward Raff & Steven Mills

AIES: FEBRUARY 2018

PROBLEM

GOAL: ACCURATE DECISIONS THAT ARE BLIND TO PROTECTED ATTRIBUTES

- e.g., Predict credit-worthiness, recidivism, job performance, etc. but do not consider race, gender, nationality, etc. in our decision.
- "Fairness through unawareness" is insufficient
  - Even if the protected attribute is completely removed from the dataset, other features may be highly correlated with it and function as proxies.

SOLUTION: FAIR RANDOM FORESTS

- New fair learning algorithms
- New measures of discrimination applicable to:
  - Multinomial features
  - Continuous features
  - Regression problems
- New evaluation procedure for fair algorithms

FAIR TREE INDUCTION

- Standard decision trees pick features based on Information Gain (IG)
  - i.e., "How easy does this feature make it to predict the target?"
- We introduce a new criteria:

$$IG_{\text{fair}}(T, b) = IG^{y}(T, b) - IG^{a}(T, b)$$

IG with respect to the target
minus
IG with respect to the protected attribute

- Encourages the selection of features which make it easy to predict the target but hard to predict the protected attribute.
- Can define IG_{fair} to protect *numeric features* as well
- Maintains the advantages of Decision Tree / Random Forest approach
  - Easy to use, interpretable, powerful

FAIRNESS METRICS

- Discrimination (a.k.a. "group fairness")
  - Intuition: difference between average outputs for groups of individuals with each protected attribute value should be small
  - Previous definitions are limited to binary features
  - We extend Discrimination to *multinomial classification* and to *regression*

  These two extensions of Discrimination can be used to evaluate any fair learning algorithms in the future.

- We also measure:
  - Inconsistency (a.k.a. "individual fairness")
    - Intuition: individuals with similar feature vectors should get similar scores
  - Accuracy
  - Delta = Accuracy – Discrimination

RESULTS

- On three test datasets Fair Forests (RF^F)
  - *Reduce discrimination to 0.0*
  - *Reduce inconsistency to 0.0*
- On 2 of 3 datasets, RF^F also achieves the best accuracy
- Similar results for continuous protected features
  - No need to discretize
  - Zero discrimination & zero inconsistency w/o loss of accuracy
- Similar results for regression problems

	German Credit			
	Acc	Delta	Discrim	Incon
RF^F	**0.7000**	**0.7000**	**0.0**	**0.0**
NB^F	0.6888	0.6314	0.0574	0.3132
LR	0.6790	0.5517	0.1273	0.3050
LR^F	0.5953	0.5842	0.0111	0.1284
LFR	0.5909	0.5867	0.0042	0.0592

(See the paper for full results: arxiv.org/abs/1712.08197)

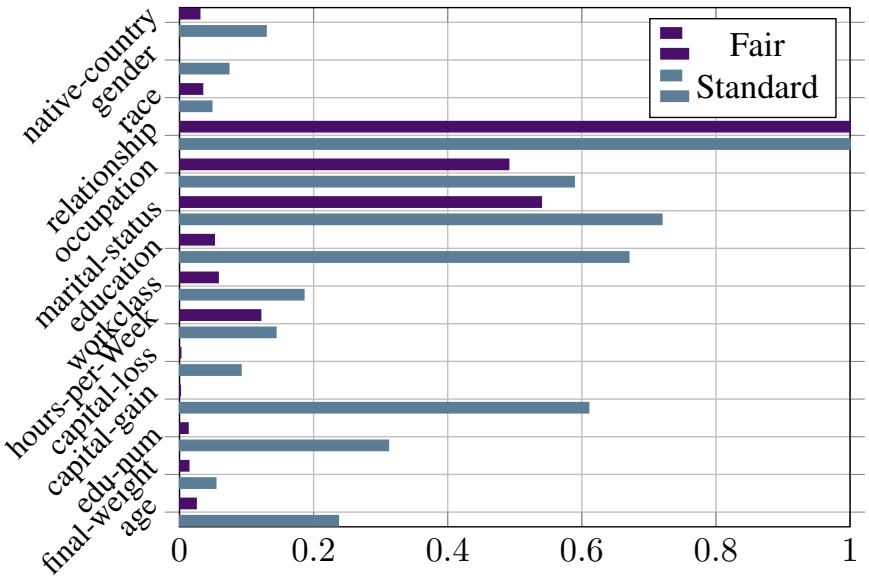
*Fair Forests are good at eliminating discrimination & inconsistency without sacrificing accuracy**

* We caution that improved accuracy should not be a general expectation. The protected attribute may be uniquely predictive of the target variable, in which case we would expect accuracy to decrease.

INTERPRETABILITY

- Examine which features were selected by standard vs fair RF
- *Gender importance goes to zero**
- Capital-gain & capital-loss also go to zero
  - Seems like it should be good for predicting creditworthiness
  - Too highly correlated w/ gender?
  - Problem w/ data collection?
  - Something else?

Relative feature importance, Adult Income data set*



Fair Forests can be used to tell us something interesting is going on even if we don't use them to make final decisions.

(See the paper for full results: arxiv.org/abs/1712.08197)

* Gender was defined as the protected attribute

FULL EVALUATION PROCESS

- Evaluate Fair Forests' ability to *protect any single attribute*
  - Ensure the approach is not overly sensitive to choice of attribute (why protect gender when predicting income but age when predicting credit?)
  - First time this evaluation process has been used
  - Strong demonstration of robustness
  - Future-proofing: the list of attributes worthy of protection has changed over time, and will continue to change, so test all of them for fairness.

This technique can be used to evaluate any fair learning algorithm in the future.

RESULTS

- We can *reduce discrimination to zero for any attribute* in German & Health datasets
- ...and the accuracy always comes out the same.

CONTRIBUTIONS

- New fair learning algorithms
- New measures of discrimination applicable to:
  - Multinomial features
  - Continuous features
  - Regression problems
- New evaluation procedure for fair algorithms

FAIR DECISION TREE & FAIR FOREST ALGORITHMS:

- Non-linear
- Interpretable
- Easy to use (no hyper-parameter tuning)
- Applicable to numeric features & classes
- Demonstrated high accuracy, group fairness & individual fairness on standard datasets

THANK YOU

Booz | Allen | Hamilton

For more information about Fair Forests, please contact me at sylvester_jared@bah.com

Or find our paper online at arxiv.org/abs/1712.08197

Jared Sylvester

@jsylvest

Edward Raff

@edwardraffml

Steven Mills

@stevndmills

BOOZALLEN.COM/MACHINEINTELLIGENCE

CONSULTING | ANALYTICS | DIGITAL SOLUTIONS | ENGINEERING | CYBER