Fairness Testing

Fairness Testing: Testing Software for Discrimination
ESEC/FSE 2017

http://tinyurl.com/FairnessPaper

Modern software influences critical decisions

Software can make bad decisions. Software can discriminate!

Modern software influences critical decisions

Software can make bad decisions. Software can discriminate!

Modern software influences critical decisions

Software can make bad decisions. Software can discriminate!

Modern software influences critical decisions

Software can make bad decisions. Software can discriminate!

Modern software influences critical decisions

Software can make bad decisions. Software can discriminate!

Modern software influences critical decisions

Software can make bad decisions. Software can discriminate!

Modern software influences critical decisions

Software can make bad decisions. Software can discriminate!
how people want to use vision software

today's goals

Define software discrimination.

Operationalize measuring discrimination through causal software testing.

Design software to be fair

Typically machine learning systems:
- Balance training sets
- Introduce training noise
- Constrain regression's loss function
- Split criteria on sensitive inputs

LOAN program

Fairness: prior definitions

1. Hide the data

2. Compare subpopulation proportions

Fairness: prior definitions

Ineffective because of data correlation. [Latanya Sweeney. Discrimination in online ad delivery. CACM 2013]

This talk is not about policy.
where group discrimination fails

Europe

Asia

recommend loans to all green and to no purple applicants

recommend loans to all purple and to no green applicants

Group discrimination can be 0.

Fairness: prior definitions

3. Correlation or mutual information

corr(race, \text{APPROVED}) = 0.8

MI(race, \text{APPROVED}) = 0.6

Correlation does not measure causation

What is fairness?

Sensitive inputs should not affect software behavior.

We want to measure causality!


causal testing

Sensitive inputs should not affect software behavior.

\text{LOAN}(?) = ?

Hypothesis testing:

NO

Yes

No need for an oracle!

Themis

automated test-suite generator

How much does my software discriminate with respect to …?

Does my software discriminate more than 10% of the time, and against what?

Themis generates a test suite or can use a manually written one

http://fairness.cs.umass.edu

discrimination measures

causal discrimination

\text{LOAN}(?) \overset{?}{=} \text{LOAN}(?)

\text{group discrimination}

\text{apparent discrimination (causal or group)}
Apparent discrimination can be group or causal, measured on a given test suite or operational profile.

Software may discriminate, but not for a given set of customers.

Fair software may appear to discriminate (e.g., Amazon same-day delivery).

How does Themis work?

- Adaptive, confidence-driven sampling
- Input schema
- Confidence error bound
- Sound pruning
- Themis

Evaluation

Eight open-source decision systems trained on two public data sets

<table>
<thead>
<tr>
<th>Decision system</th>
<th>Data set</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination-aware logistic regression</td>
<td>Census income dataset: financial data 45K people income &gt; $50K?</td>
<td>- More than 11% of the individuals had the output flipped just by altering the individual's gender.</td>
</tr>
<tr>
<td>Discrimination-aware decision tree</td>
<td>Statlog German credit dataset: credit data 1K people “good” or “bad” credit?</td>
<td>- Causal discrimination score was up to 21x higher!</td>
</tr>
<tr>
<td>Discrimination-aware naive Bayes</td>
<td></td>
<td>- Trying to avoid group discrimination may introduce other discrimination.</td>
</tr>
<tr>
<td>Discrimination-aware decision tree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrimination-aware naive Bayes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classifier</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

findings

Group discrimination is not enough.

Causal discrimination can capture significant differences from group discrimination.

Trying to avoid group discrimination may introduce other discrimination.
findings

- The more a system discriminates, the more efficient Themis is.
- On average, pruning reduced test suites by $148 \times$ for causal and $2,849 \times$ for group discrimination. Best improvement was $13,000 \times$.

Pruning is highly effective.

related work

Ways of measuring discrimination
- CV score [19]
- Correlation, mutual information [79]
- Output probability distributions [51]

Discrimination-aware algorithms
[18, 40, 88, 91]

Measuring discrimination with manually-written tests [79]

Causal model inference
[Maier et al., UAI '13]

Fairness verification
[Abarghoui et al., OOPSLA '17]

Contributions

- Causality-based definition and method for measuring software fairness
- Themis, an automated test-suite generator for fairness testing
- Provably-sound pruning test-suite reductions
- Evaluation on real-world software, demonstrating Themis’ effectiveness

http://fairness.cs.umass.edu