Identifying Significant Predictive Bias in Classifiers

Zhe Zhang, Daniel B. Neill • [Fairness & Transparency in Machine Learning - KDD 2017]
Carnegie Mellon University (CMU) - Event and Pattern Detection Lab

Overview

- A novel anomaly detection method to detect if a classifier has statistically significant bias for some subgroup in the data — and identify characteristics of this subgroup.
  - *extensions:* penalize complexity, subgroups with anomalous high error rates
- Unlike other approaches, this method efficiently considers all exponentially possible subgroups.
  This allows consideration beyond interaction effects or subgroups of a priori interest; it enables grouping of weak, but related signals.
- By considering all subgroups, the method can outperform lasso and other methods in detection and prediction performance.

Why is Predictive Bias Important?

- Increasingly, data-driven tools like probabilistic classifiers are being used for decision support and risk assessment in many areas. It's important to check these for possible bias or discrimination.
  - e.g. ProPublica analysis of COMPAS crime risk predictions
- **Source of bias:** limited classifier flexibility or model misspecification. This can lead to some subgroup(s), S, being poorly estimated, with predictive bias:
  $$\mathbb{P}(Y = 1 | 1_{\{S\}}) < \hat{p}_S$$
  for over-estimation (and similar for under-estimation bias).
- Assessing bias in all possible subgroups is difficult: *computationally* and *statistically*.
  - Four features, with 5 categorical values, has $\prod_{m=1}^{4}(2^5 - 1) \approx 10^6$ subgroups
  - It is trivial to identify *some* measure of predictive bias — is it significant?

Subset Scan Methodology

We extend methods from anomaly detection, particularly fast, expectation-based subset scans (*Neill, 2012*).

$$S^* = MDSS(\mathcal{D}, \hat{p}, score_{bias})$$

S^* = approximately most biased subgroup of $\mathcal{D}$
MDSS = Multi-Dimensional Subset Scan (*Kumar, Neill 2012*)

We contribute a novel extension of *MDSS* algorithm:

1. a **new scoring function of bias** ($score_{bias}$) that statistically measures predictive bias and satisfies subset scan properties needed to find S^* in linear time.
2. an estimate of **statistical significance** of a detected subgroup (*parametric bootstrapping*).
3. **penalizing the complexity** of the detected subgroup, in linear time, enabling "elbow curve"-style penalty selection
4. extending to detect **anomalous high classification error** subgroups

$score_{bias}$ is based on:

$$H_0 : odds(y_i) = \frac{\hat{p}_i}{1 - \hat{p}_i} \; \forall i \in \mathcal{D}$$

$$H_1 : odds(y_i) = q\frac{\hat{p}_i}{1 - \hat{p}_i}, \text{ where } q > 1 \; \forall i \in s \text{ and } q = 1 \; \forall i \notin s$$

which results in this log-likelihood ratio:

$$score_{bias}(S) = \max_q \log \prod_{i \in S} \frac{Bernoulli(\frac{q\hat{p}_i}{1-\hat{p}_i+q\hat{p}_i})}{Bernoulli(\hat{p}_i)}$$

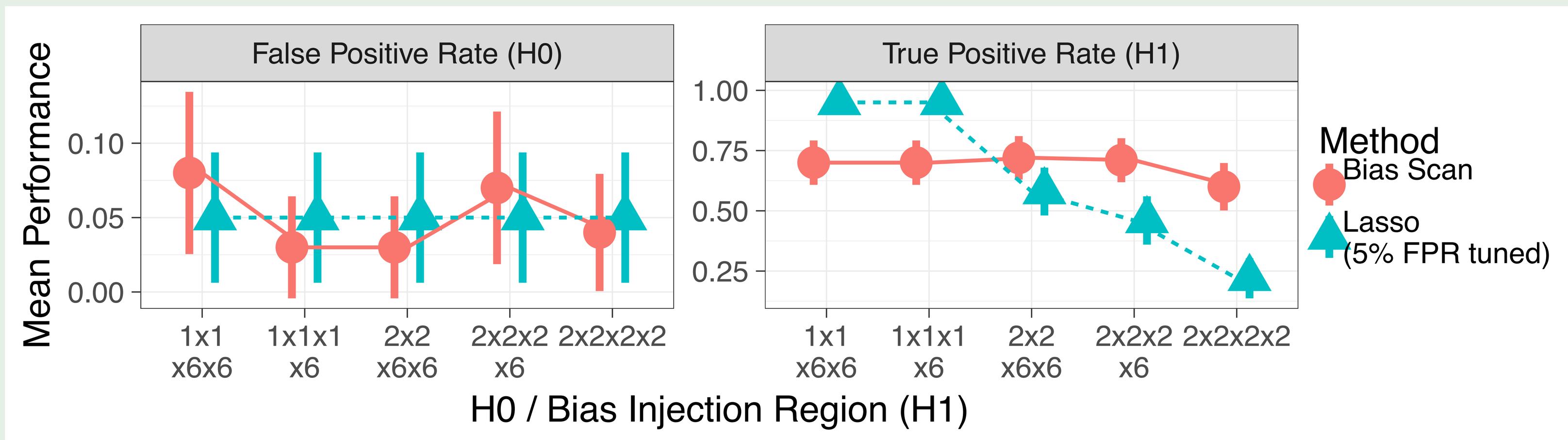
Future extensions: stepwise classifier, classifier disagreement scan, continuous outcomes Y

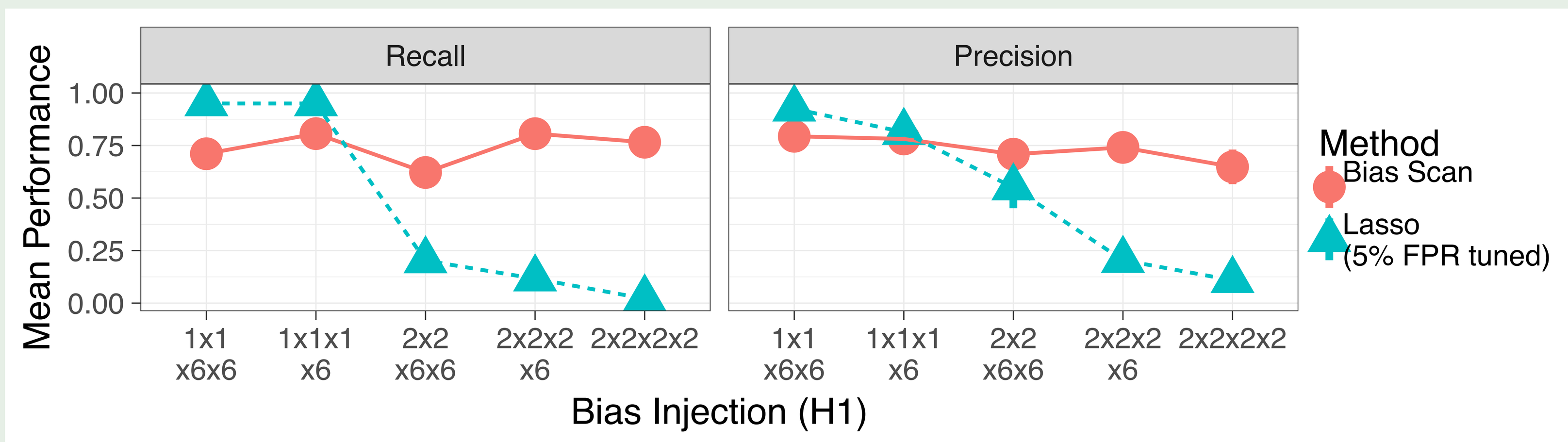
Existing Methods Have Difficulty Considering Subgroups

- high-dimensional interaction effects
  - penalized regression (e.g. lasso) on residuals — limited by inability to group related interactions, unless using prior knowledge (group lasso, *Yuan, Lin 2006*)
  - stepwise methods — too coarse of a consideration set
- F-test style methods using black-box methods — do not pinpoint where bias is present (*Shah, Buhlman 2017*)
- tree-style methods / clustering methods — difficult to obtain statistical significance, top-down greedy process separates subgroups

Synthetic Comparison with Lasso

Bernoulli data using additive log-odds model, with 4 categorial features (6 values). Bias is added to {one, several} interactions of {2, 3, 4} dimensions (*x-axis*). The number of biased observations is fixed at $n = 100$ — spreading these out across interactions makes **weak, but related signals**.



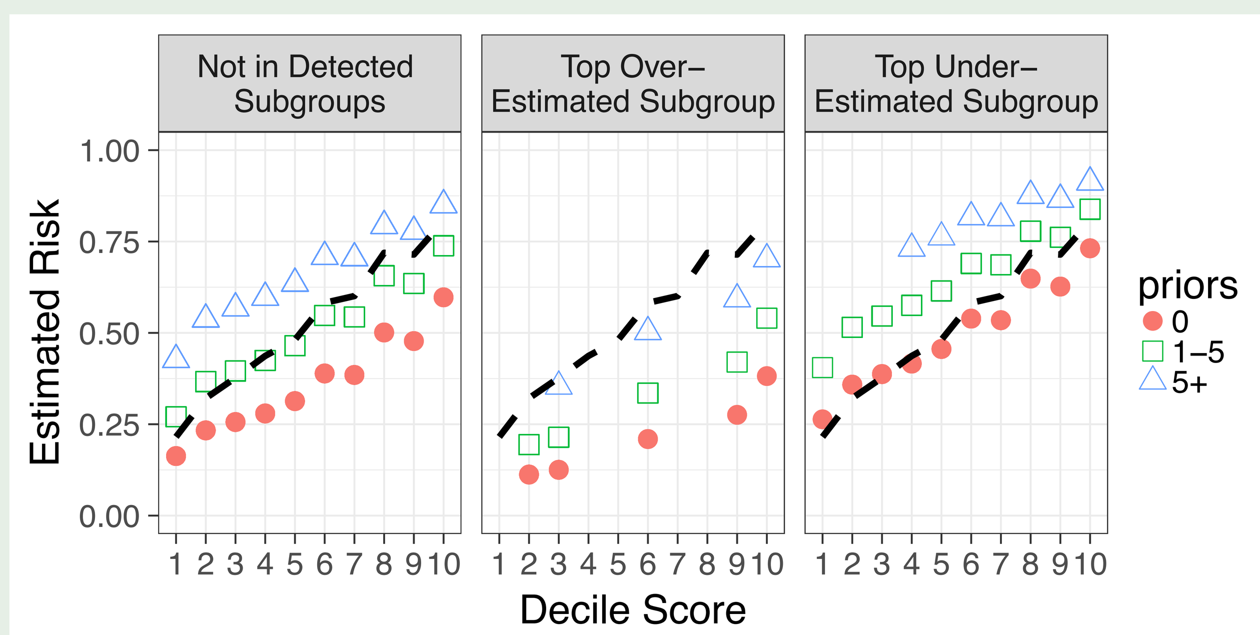


Recidivism Prediction, Credit Prediction

COMPAS re-offending risk prediction dataset. COMPAS's risk predictions are represented as decile groups (1-10).

We find new, notable biases in COMPAS predictions (using penalization):

1. COMPAS does not adequately account for prior offenses
2. *Under-estimated:* males, age ≤ 25 ($p < 0.005$)
   (mean $\hat{p}$ of 0.50; observed rate of 0.60; $n = 1101$)
3. *Over-estimated:* females, charged with misdemeanors, and in decile scores $\in \{2, 3, 6, 9, 10\}$ ($p = 0.035$).
   (mean $\hat{p}$ of 0.38; observed rate of 0.21; $n = 202$)



(dotted black line = base COMPAS prediction)

Credit delinquency data:

- 470 of the 496 (top 1%) riskiest consumers are in a significantly over-estimated subgroup. After detection and model correction, only 286 of those consumers fall in the top 1%.
- Abnormally high error subgroup: the logistic regression is over-confident for both low-risk and high-risk consumers.

Predictive bias in stop-and-frisk prediction data (Goel et. al. 2016) too.

This work was partially supported by NSF grant IIS-0953330