Increasing Trust By Detecting Hidden Prediction Bias and Mistakes

Workshop on Trustworthy Algorithmic Decision-Making 2017

Zhe Zhang[1,*], *Daniel B. Neill*[1,2]
[1] *Carnegie Mellon University, Event and Pattern Detection Laboratory / Heinz College*
[2] *New York University, Center for Urban Science and Progress*
[*] *PhD student*
zhezhang@cmu.edu, neill@cs.cmu.edu

As noted in the prompt for this workshop, there is an increasing use of data-driven decision making and predictions in many parts of society, business, and governance. While these can provide benefits of improved and automated predictions, it is also important to know and ensure they are not making serious mistakes. In our motivation for trustworthy algorithms, a key issue is if an algorithm consistently makes serious, unanticipated prediction mistakes. We can increase trust in an algorithm if we can identify where an algorithm is making such mistakes, or have evidence that it is not making any.

However, it is not easy to always check for such mistakes. There are exponentially many ways to define a subpopulation, and — besides checking a few well-known subgroups beforehand — it is both computationally and statistically difficult to check all of them for potential bias. For example: in predicting criminal repeat offending, we may check and show accurate predictions on African-American defendants overall, but hidden in a subgroup of African-Americans, be seriously over-estimating the risk.

In our work, we seek to address this hurdle of trustworthiness. We are developing tools which both detect if an algorithm is making serious mistakes on subgroups, and what those subgroups are. Specifically, we consider 3 types of serious mistakes: Consistently (1) over-estimating the risk or predicted value for a subgroup, compared to the observed ground truth, (2) similarly under-estimation, and (3) making anomalously more prediction errors that expected. In our methods, we are also able to penalize the complexity of the detected subgroups, so that we can simplify the detected group for practitioners without notably reducing the seriousness of the detected mistakes.

The benefits of such a tool are three-fold. First, before utilizing algorithmic decision support, practitioners can audit their predictions and detect if such serious mistakes on subgroups occur. This is a warning system for practitioners, especially those in the policy or business domains, where hidden mistakes on subgroups that are realized after implementation can be damaging. Second, by identifying such subgroup anomalies in prediction, this identifies hypotheses for practitioners to better understand why such subgroups may be so poorly predicted. Third, we could utilize this method repeatedly to adjust our original algorithm to maintain similar prediction performance, but without having serious mistakes on subpopulations of the data.

We describe the basics of our methodology in a short paper, focused on binary classification algorithms, presented at the Fairness, Accountability, and Transparency in ML (FAT-ML) Workshop at KDD 2017, available online[1]. In summary, we utilize fast subset scan based anomaly-detection

[1] "Identifying Significant Predictive Bias in Classifiers", https://arxiv.org/abs/1611.08292

methods and develop extensions to focus on detecting subgroups with anomalously poor predictions. This overcomes two challenges: (1) computationally, we are able to efficiently consider the exponentially many number of possible subgroups, which naively would be infeasible, and (2) statistically, we use bootstrap techniques to estimate if the detected subgroups are indeed statistically significantly anomalous, or perhaps are just due to statistical noise.

Crime Recidivism Prediction. As a case study, we apply our bias scan method to the COMPAS crime recidivism risk prediction dataset provided by ProPublica ($n = 6172$). COMPAS predicts decile scores $\in \{1, 2, \ldots, 10\}$ for each individual, so we initially fit a logistic regression based on those decile scores. From this, we find notable biases by the COMPAS prediction that we have not seen noted elsewhere. Using bias scan, we find that the COMPAS decile scores clearly is biased on subgroups defined by counts of prior offenses. Defendants with >5 priors are significantly under-estimated (mean predicted rate of 0.60, observed rate of 0.72, $n = 1215$), while those with 0 priors are significantly over-estimated (mean predicted rate of 0.38, observed rate of 0.29, $n = 2085$).

If we refit the model to account for priors offenses, we again identify two significant subgroups of classifier bias. Young ($<$ 25 years) males are under-estimated (regardless of race) ($p < 0.005$); with an observed recidivism rate of 0.60 and a predicted rate of 0.50 ($n = 1101$). Additionally, females, whose initial arrested crimes were misdemeanors, and of half the COMPAS decile scores $\in \{2, 3, 6, 9, 10\}$ are over-estimated ($p = 0.035$); with an observed recidivism rate of 0.21 and a predicted rate of 0.38 ($n = 202$). In Figure 1, we compare the original COMPAS decile model with the logistic model that accounts for priors and these two detected subgroups.

Consumer Credit & Loan Delinquency. We also apply our bias scan method to a loan delinquency prediction dataset, "Give Me Some Credit", provided by Kaggle. In this dataset, using the cross-validation lasso classifier, we identify an interesting group whose delinquency risk is significantly over-estimated ($p < 0.01$). This group are consumers who are above the median in credit utilization and who have at least 1 occurrence of a late payment of 30-59, 60-89, and 90+ days late (i.e., on a 3 separate payments). This group is 1.7% of the total dataset, but an important group though. Of the 496 top 1% riskiest consumers in the dataset, almost all (470) belong to this subgroup. If we adjust the model to recognize this subgroup, then only 286 consumers from that subgroup would then be ranked in the top 1%. In this same data, we also detect a high classification error subgroup where the predicted and observed rate are the same (61%), but the model is too overly confident on the low and high-risk consumers.

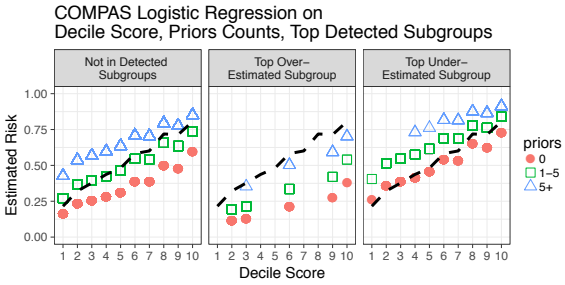

Figure 1: *Baseline COMPAS predictions are black line, the subgroup-adjusted predictions are colored points.*