Annotated Bibliography

Reference

Altenburger, Kristen M, and Daniel E Ho. "When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions." *Journal of Institutional and Theoretical Economics*, 2018, pp. 1–25., doi:10.1628/jite-2019-0001.

Category:

Summary

The authors address two goals in their paper. The first goal is to highlight the issues surrounding the use of consumer data (i.e., Yelp reviews, consumer complaints) in predictive algorithms for food safety regulation/enforcement. The authors demonstrate how consumer data is biased against Asian restaurants regardless of food safety scores by inspectors made prior to complaints (e.g., suspected food poisoning). As such, using consumer data in predictive analytics for food safety regulation/enforcement unfairly and negatively targets Asian restaurants. The second goal of the paper is to explore a solution by Pope and Sydnor (2011) to debias regulatory targeting of Asian restaurants while allowing the use of questionable predictors such as consumer reviews. The Pope and Sydnor (P&S) solution employs all predictors (i.e., socially acceptable predictors (SAPs) like food safety inspection scores, contested predictors (CPs) like Yelp reviews, and socially unacceptable predictors (SUPs) such as race) while marginalizing out SUPs. The authors address the limitations of the approach using Monte Carlo simulation and how a restricted random forest model demonstrates improved predictive accuracy

Altenburger and Ho's investigation uses data from New York City (77,661 routine health inspections of 22,096 Asian and non-Asian establishments and food poisoning complaints between 2010 to 2017) and King County in Seattle Washington (1,756 Seattle restaurants matched to 13,299 food safety inspections and 152,153 Yelp reviews from 2006-2013). The New York City data demonstrate a negative correlation between health inspection scores and complaints for Asian establishments. Asian restaurants with the same inspection score averaged 42% more consumer complaints. Similarly, the King County data illustrated a higher probability of suspicious terms (e.g., negative reviews, complaints) for Asian restaurants than non-Asian restaurants (9.8% vs. 7.7%) with the same inspection history. Asian establishments from both data sets are subjected to more negative/suspicious terms/complaints about suspected food poisoning while holding inspection scores (made prior to consumer complaints) constant. The results from this investigation highlights the discrimination against Asian establishments by consumers and using such data would unfairly target Asian establishments for further unnecessary inspections.

The authors explore a way to debias algorithms used regulatory enforcement through Pope and Sydnor (P&S) approach to debiasing by marginalizing socially unacceptable predictors (SUPs) The approach addresses the problem of contentious predictors (i.e., consumer complaints) potentially serving as a proxy for race.

Reference

Mirko Bagaric, Dan Hunter, and Nigel Stobbs, Erasing the Bias Against Using Artificial Intelligence to Predict Future Criminality: Algorithms are Color Blind and Never Tire, 88 U. Cin. L. Rev. 1037 (2020) Available at: https://scholarship.law.uc.edu/uclr/vol88/iss4/3

Category: Fairness

Summary

Authors of this article define recidivism as the act of taking measures that will reduce the length of prison terms for some offenders and consequently lower the number of inmates in federal prisons. This article focuses on three components in the criminal justice system where they believe the evaluation of future offending is relevant. These three components are **sentencing**, **bail**, and **parole**. The authors explain the purpose of these three systems and their roles in the criminal justice system.

Suggested Solutions

This article highlights the importance of monitoring the factors used in risk assessments. Each factor must be defined, expressed and identified. Each piece of evidence must be relevant to the sentencing process. Irrelevant factors should not be used in the calculation of sentencing decisions. All algorithms should be properly coded and transparent. People have shown to trust the use of an algorithm when they have seen how it works and how well it determines correct outcomes. One way algorithms could be transparent is by auditing and looking to the external inputs and outputs of the process of the decision. The article concludes by suggesting workings of risk assessments being

---

Reference

Picard, Sarah, et al. *Beyond the Algorithm: Pretrial Reform, Risk Assessment*. Formal Report . New York: Center for Court Innovation, 2019. Report. Link:

https://www.courtinnovation.org/sites/default/files/media/documents/2019-06/beyond_the_algorithm.pdf.

Category: Fairness

Summary

Beyond the Algorithm addresses risk assessments and their usefulness to society, while also addressing concerns over racial fairness in the criminal justice system. Their attempt isn't to argue for removing risk assessments, but to target the **"kinds or errors" made** by the algorithm. This article uses ProPublica's analysis of errors made by COMPAS as an example of how risk algorithms could contain racial biases towards minority groups. ProPublica conducted their own study for research purposes to highlight the consequences of implementing risk assessment tools in the criminal justice system. Their assessment was based on nine risk factors:

- **Criminal History**
  - 1. Prior convictions
  - 2. Prior jail or prison sentence
  - 3. Prior failure to appear in court
  - 4. Probation status
- **Current Case Characteristics**
  - 5. Charge type
  - 6. Charge severity
  - 7. Concurrent open cases
- **Demographic Characteristics**
  - 8. Age
  - 9. Gender

Each feature is assigned a weight. The weights represent the number of risk points associated with each risk factor. Scoring placed the defendant in a risk category: minimal, low, moderate, moderate-high, or high-risk. Area under the curve (AUC) statistics revealed the predictive accuracy for Raw Risk score to be 0.745 and Risk Categories scores 0.733. These scores show successfulness by being over 70% accurate. Additionally, their scores were slightly better than COMPAS scores. However, many of its findings were similar to COMPAS in that there were higher false positive rates for African American and Hispanic defendants opposed to White defendants.

Suggested Solutions

To combat the racial disparities in algorithmic systems, they've found the most success in their *Risk-Based Approach* (RBA). This approach reduced the overall detention rate by nine percentage points when compared to another model known as *Business as Usual* (BAU). In the BAU approach, false positive rates were relatively high for all groups. The RBA approach reduced the overall false positive rate to less than 10 percent by improving the accuracy of pretrial detention decisions. Another successful approach was

their *Hybrid Charge - and Risk-Based Approach*. This approach reduced the overall detention rates and lessened the racial disparities. Ultimately, they found that if pretrial detention was restricted only to defendants who are charged with violent crimes and who fall into higher-risk categories, such a policy may both reduce incarceration overall and alleviate racial disparities

Website: https://www.courtinnovation.org/about

Instructions on how to find article from website

- Areas of Focus
- Improving Decision-Making
- Risk Assessments
- Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness (Publication)

Reference

Skeem, J, and C Lowenkamp. "Using Algorithms to Address Trade-Offs Inherent in Predicting Recidivism." *Behav Sci Law*, vol. 38, 2020, pp. 259–278., doi:https://doi.org/10.1002/bsl.2465.

Category: Fairness

Summary

Risk assessment algorithms evaluate the likelihood of recidivism by weighing risk factors such as criminal history and drug use. There are concerns that any benefits of reducing incarceration rates are offset by racial justice costs as these algorithms have displayed racial disparities in incarceration. This study compares how alternative strategies for "debiasing" risk assessment algorithms affect fairness trade-offs inherent in predicting violent recidivism when the rate of recidivism is unequal across races. Fairness in this study is viewed as balance in error rates for outcomes and balance in accuracy rates for predictions. ProPublica's investigation demonstrates that COMPAS, a risk assessment instrument, misclassified Black defendants twice as much as Whites as medium or high risk to reoffend which subsequently could result in harsher treatment of Black defendants. The algorithm displays bias where fairness is focused on a balance in error rates for outcomes.

Reference

Zavrsnik, Ales. “Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings.” *European Journal of Criminology*, 2019, pp. 1–20., doi:10.1177/1477370819876762.

Summary

This article highlights the societal trend of ‘algorithmic governance’ reflected in the rising use of big data, algorithms and machine learning in the criminal justice sector. The author provides a brief review of algorithmic governance and its growing use in policing and the courts (i.e., predictive policing and algorithmic justice). Through the use of examples and past research, the purposes of this paper are 1) to examine the problematic shift and heavy reliance on big data, algorithms and machine learning in the criminal justice systems for understanding crime and how to address it and 2) demonstrate how algorithmic justice can conflict with civil liberties, legal doctrines, and criminal procedure laws.

In addressing the first goal of the paper, Zavrsnik discusses how the use of big data, algorithms, and machine learning are producing new knowledge, thresholds and justifications to address crime that may impinge upon civil liberties, contradict concepts in criminal procedure rules, and allow abuses of power. Citing examples of the use of algorithms and machine learning to identify “terrorist sleepers” in Germany, track “persons of interest” through sentiment analysis (predicting future personal states/behavior) using a Twitter analysis tool in Slovenia during the Occupy Movement, and track consumers in order to influence their spending habits, the author points out how
people’s civil rights and liberties are abused and blurring boundaries in the areas of security and criminal procedures.

In addressing the second purpose of the paper, the author addresses the risks and downfalls of algorithmic justice. First, the existence of false positives in predictive policing is discussed, Second, there are problems in creating reliable and valid databases and predictive algorithms if data used is unreliable and invalid and unintentional problems can arise even with the creator’s best intentions. Third, creating good algorithms requires examining the purpose for its creation and whether it betters society. Fourth, there is a need to address runaway algorithms that result in unintended consequences such as racial discrimination in prison sentencing. Fifth, blurring probability causality and certainty when using predictive algorithms in judicial decisions. Sixth, Zavrsnik discusses the issues with machine bias and the argument for more empathic human judgment in the justice system and how de-biasing may not be desirable. Seventh, there are limitations algorithms that conflict with changes in judicial cases. Eighth, he discusses the pitfalls of algorithms as judicial cases change whereby algorithms may hinder judicial evolution. Ninth, relying on algorithms (even ones that

produce errors) without human judgment is discussed. Lastly, human rights implications due to over policing and under policing as a result of algorithms and the rights of convicted persons to question the algorithms is explored.

AI Fairness 360 Demo

The AI Fairness 360 Demo provides users with a sample COMPAS dataset which allows us to explore areas where biases are located. Four mitigation strategies are provided. The successfulness of each strategy is measured based on five bias metrics.

1. **Sex and race are the focused attributes**
   - Protected Attribute: Sex

     Privileged Group: ***Female***, Unprivileged Group: ***Male***

   - Protected Attribute: Race

     Privileged Group: ***Caucasian***, Unprivileged Group: ***Not Caucasian***

2. **Bias mitigation algorithms**

"The choice of which to use depends on whether you want to fix the data (pre-process), the classifier (in-process), or the predictions (post-process)

Definitions

- **Reweighing** - Weights the examples in each (group, label) combination differently to ensure fairness before classification. (pre-process)

- **Optimized Pre-Processing** - Learns a probabilistic transformation that can modify the features and the labels in the training data. (pre-process)

- **Adversarial Debiasing** - Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit. (in-process)

- **Reject Option Based Classification -** Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty. (post-process).

3. There are 5 bias metrics: Statistical Parity Difference, Equal Opportunity Difference, Average Odds Difference, Disparate Impact, and Theil Index
4. Each Bias mitigation algorithm is compared to the original bias metrics found by COMPAS

Relates to project

I recommend the statistics team take a close look at steps (3) and (4) to further see which algorithm provides a more balanced approach toward fairness. At this point in the project, we are still unsure which direction we want to head in. Mitigation in one area could also lead to unfairness in another. Becoming attentive to tradeoffs early on could help us as we move forward.

Accessing Demo

- Click link: https://aif360.mybluemix.net/
- "Try a Web Demo"
- Choose Compas (ProPublica recidivism) dataset
- Click "Next"

Reference

Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, Ruchir Puri. *BIAS MITIGATION POST-PROCESSING FOR INDIVIDUAL AND GROUP FAIRNESS*. Brighton, UK: Institute of Electrical and Electronics Engineers, 2018. Report. Link: https://arxiv.org/pdf/1812.06135.pdf

Category: Post-Processing

Summary

This article proposes methods for increasing both individual and group fairness. Group fairness is defined by splitting a population into protected attributes (such as gender) and seeks for some statistical measure to be equal across groups. Individual fairness in seeks for similar individuals to be treated similarly. There are three post-processing algorithms used in this study: (1) Individual Group Debiasing (IGD), (2) Equalized Odds Post-processing (EOP), and (3) Reject Option Classification (ROC). Each of these algorithms are compared by three measures: (a) individual bias, (b) disparate impact, and (c) balanced classification accuracy. The AI Fairness 360 toolkit was used in this analysis for both the EOP and ROC algorithms. Both algorithms are used to mitigate bias in predictions. The EOP algorithm modifies the predicted labels using an optimization scheme to make predictions fairer while the ROC algorithm changes predictions from a classifier to make them fairer.

Relates to project

We know that post-processing is one approach we can look upon as we think about how we can mitigate bias. This article provides empirical results and explains in detail the trade-offs between each of these algorithms. Majority of the trade-offs are mentioned on page 4. The statistics team should take a look at these.

Reference

Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. "OptimizedData Pre-Processing for Discrimination Prevention." 2017. Report. Link: https://arxiv.org/pdf/1704.03354.pdf

Category: Pre-processing

Summary

Pre-processing is one of the three areas of focus when considering mitigating bias in algorithms. Many publications discuss several approaches (such as *disparate impact* in relation to group fairness, individual fairness, and equal error rates) towards making algorithmic systems fairer. This article proposes an optimization formula comprised of three goals: controlling discrimination (trade-offs), limiting distortion in individual data samples, and preserving utility. The distortion constraint included in this model distinguishes it from the previous approaches. Their goal is to determine a randomized mapping that transforms the given dataset into a new dataset, which may be used to train a model, and similarly transforms data to which the model is applied, i.e., test data. The optimizations in this study utilizes a probabilistic framework for discrimination-prevention pre-processing in supervised learning (pages 2-9). Additionally, ProPublica COMPAS recidivism data is used in this dataset and compared to their results (page 9-12).

Relates to project

This study has shown decreases in recidivism for certain groups. Looking forward, if we decide to take the pre-processing approach towards the project, we could possibly draw ideas from here. The mathematical steps are included in this article. The Domain and Statistics team can continue to read over these steps and make sense of it. It could help the Development team to view this article as well to see if the methods used here are applicable for us.