Mitigating Racial Bias in Recidivism Prediction:

A Machine Learning Approach

Yijia Lyu

April 2020

Literature Review

Crime has been a significant disruptive factor in social security and our daily life. Keeping track of crime occurrences and criminals is therefore essential to build safer communities. Washington courts reported that 63.3% of the sentences in 2007 involved cases related to recidivism. This means that if our systems are able to capture and supervise criminals who have high-risks to re-commit crimes, the crime rate may drop significantly. According to the definition put by Urahn (2011), recidivism is the 'act of re-engaging in criminal offending'. Prior research has incorporated the key factors of recidivism as demographic factors including race, age, gender, etc. (Langan & Levin, 2002), criminal history like the term of imprisonment (Blumstein, Cohen, & Farrington, 1988; Piquero, Farrington &Blumstein, 2003), other individual-level factors such as antisocial attitudes, associates, personality (Serin, Lloyd, Helmus, Derkzen, & Luong, 2013), social bonds (Yang, Liu, & Coid, 2010) or socioeconomic status (Hanson & Harris, 2000). To reduce safety concerns and better allocate policing and other resources, evaluating who is likely to reoffend after release and classifying these high-risk offenders seem to be relatively valuable.

1.1 Statistical Risk Assessment on Recidivism Prediction

In fact, criminologists have long proposed and attempted to estimate and predict the recidivism risk of criminals, especially for cases that may cause extraordinary harm to the society. Since 1980s, estimations on sexual and violent offence recidivism have been a frequent topic in criminology. Many estimation studies start from the psychopathy point of view. For example, the U.S. Sentencing Commission (2005) employed a prediction tool called CHC for federal judges to ‘measure offender culpability, deter criminal conduct, and protect the public from further crime of the defendant’. Similarly, in 1995, the Canadian forensic researcher, Quinsey, combined the Psychopathy Checklist (PCL-R; Hare, 1991) with a number of relevant variables (Rice et al., 1990) to perform multivariate statistics and calculation of actuarial estimates of risk. Then in 2006, collaborating with his colleagues, Quinsey developed a prediction instrument, Sex Offender Risk Appraisal Guide (SORAG), which was later adopted by many scholars, to assess criminals’ risk score of violent and sexual recidivism based on fourteen items. Each item is scored individually and aggregated together following an assigned weight of each item. Multiple datasets gathered from German, USA, Canada, Belgium are used to test the validity of SORAG and many replication studies are built upon this guide (Rettenberger & Eher, 2007; Ducro & Pham, 2006). However, the magnitude of such sample size is often restricted to only hundreds, the risk factors are static, ignoring the dynamic predictors, and the scope for such prediction only limits to a few extreme crime categories. Most importantly, the coefficients of each item should be validated with external data, otherwise it may lead to inappropriate or even erroneous causal relationship explanation. With these concerns on accuracy and reliability, although these predictive studies were frequently discussed in academia, the variables to evaluate the recidivism were regarded only as a guideline while the risk scores themselves were rarely applied to jurisdiction at this stage.

1.2 Extending the Scope: Machine Learning Predictions on Recidivism

With the wide application of machine learning, the predictive tools in recidivism have also transited from clinical judgement to algorithm decision-making. A surge of novel data mining techniques including logistic regression, random forests, support vector machines, neural networks and the search algorithm are found to outperform the traditional methods (Attewell &Monaghan, 2015). These novel methods not only enlarge the scale of data being fed in, but also extend the scope to a wider range of crime types by decreasing unexplained variables in the dependent variables (ibid). However, a statistically expected outcome may not be a perfect match in an actual policy, predictive power is accompanied with errors and the cost of these errors needs to be evaluated (Berk, 2012). When assessing the performance of a predictive algorithm, apart from forecasting accuracy, the ratio of false positives (false alarms) to false negatives (missing cases) is another key metric. Bradley (1997) put that the cost of misclassification is more important than the rate of misclassification. Overestimating the false positives can lead to great amount of resource waste, leaving the low-risk defendants take unfair consequences including loss of freedom, decreased life quality or loss in future employment. But on the contrary, underestimating the false negatives may also put many lives in danger. The trade-off between the false positives and false negatives is at practitioners' discretion and vary between jurisdictions (Barnes & Hyatt, 2012). Researchers suggest practitioners to have the rate pre-determined on an agreeable level, such as 5:1 (Berk et al., 2005).

1.3 Real-world Application, Algorithmic Bias and Unfairness

Despite such discussion on prediction errors, many county-level jurisdictions in the United States have adopted machine learning or deep learning algorithms as a sentencing reference. By identifying which criminals are at high risks of re-committing crimes and predicting what types of crimes they may commit, the judges are referring to the result of this risk assessment to determine the final sentence imposed on the defendants. In 2012, the Wisconsin Department of Corrections launched COMPAS, an algorithmic software developed by Northpointe, and used it in each step in the prison system from sentencing to parole. This software was also employed in the jurisdiction systems in New York State, California, Florida and some others, but in neither of the states or country was the tool evaluated statistically-carefully. Brennan and his two colleagues (2009) published a validation study of COMPAS and found that the accuracy rate of the tool was 68%, however, COMPAS was 67% accurate in black men while it has a 69% accuracy in white men – although this algorithm did not include race as a variable. In a later analysis on COMPAS produced by Larson et, al (2016), the result showed that black defendants who did not recommit crimes over a two-year period were nearly twice as likely to be mistakenly labeled as higher risks compared to white counterparts (45% vs 23%). White defendants who were misclassified as low risk re-offenders almost twice as the black (48% vs 28%). In violent recidivism, compared to the black defendants, the white violent recidivists were 63% more likely to be misclassified as low risk.

Although compared to the early stage predictions, the accuracy and reliability of recidivism prediction seem to be dramatically improving with machine learning models, various validation studies on the widely-adopted prediction software COMPAS have proved us that the black communities suffer from significant algorithm unfairness. Even if we have removed the explicit race factor as an input variable, the systematic inequality still profoundly affects the individuals, groups and society. Accuracy is no longer the only concern in predictive models as existing bias towards certain groups might be further perpetuated through advanced machine learning algorithms. Algorithm fairness, defined as anti-classification (protected attributes like gender, race should not be used to make decisions), parity (the ratio of false positives and negatives should be equal across protected attribute groups) and calibration (conditional estimates are independent from protected attributes), is therefore crucial to measure model performance (Davies & Goel, 2018).

Therefore, examining and reducing the algorithmic bias is an urgent task if we decide to apply such machine prediction result in pretrial, parole, and sentencing decisions. Addressing these issues, this study will explore the reasons that lead to such algorithm unfairness in recidivism, build harm-reduction framework in machine learning models that mitigate such racial disparities to improve algorithm fairness while maintaining accuracy.

Reference List

Attewell, P., & Monaghan, D. (2015). Data mining for the social sciences: An introduction. Oakland: University of California Press.

Berk, R. A. (2012). Criminal justice forecasts of risk: A machine learning approach. New York, NY: Springer.

Blumstein, A., Cohen, J., & Farrington, D. P. (1988). Criminal career research: Its value for criminology. Criminology, 26(1), 1–35.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145–1159.

Brennan, T., Dieterich W., Ehret, B. (2009) Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. Available at: https://doi.org/10.1177/0093854808326545

Davies, S.C. & Goel, S. (2018) The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.

Ducro, C., & Pham, T. (2006). Evaluation of the SORAG and the Static-99 on Belgian sex offenders committed to a forensic facility. *Sexual Abuse: A Journal of Research and Treatment*, *18*(1), 15.

Hare, R.D. (1991). Manual for the revised Psychopathy Checklist. Toronto: Multi-Health Systems.

Langan, P., & Levin, D. (2002). Recidivism of prisoners released in 1994. Washington, DC.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm**.** Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Piquero, A., Farrington, D., & Blumstein, A. (2003). The criminal career paradigm. Crime and Justice, 30, 359–506.

Quinsey, V.L., Harris, G.T., Rice, M.E. & Cormier, C.A. (2006) 2nd Ed. Violent Offenders: Appraising and Managing Risk. Washington D.C: American Psychological Association.

Rettenberger, M., & Eher, R. (2007). Predicting reoffense in sexual offender subtypes: A prospective validation study of the German version of the Sexual Offender Risk Appraisal Guide (SORAG). *Sexual Offender Treatment*, *2*(2), 1-12

Rice, M.E., Harris, G.T. & Quinsey, V.L. (1990). A foloowup of rapists assessed in a maximum security psychiatric facility. Journal of Interpersonal Violence, 5, 435-448

Sentencing Guidelines Commission State of Washington (2008). Recidivism of Adult Felons

Serin, R. C., Lloyd, C. D., Helmus, L., Derkzen, D. M., & Luong, D. (2013). Does intraindividual change predict offender recidivism? Searching for the Holy Grail in assessing offender change. Aggression and Violent Behavior, 18(1), 32–53.

Urahn, S. (2011). State of recidivism: the revolving door of America’s prisons. The PEW Center on the States.

U.S. Sentencing Commission. (2005). A comparison of the federal sentencing guidelines criminal history category and the U.S. Parole Commission Salient Factor Score. Washington, DC.

Yang, M., Liu, Y., & Coid, J. (2010). Applying neural networks and other statistical models to the classification of serious offenders and the prediction of recidivism. UK Ministry of Justice Research Series 6/10.