Undergraduate Thesis Prospectus

Making Models Fully Forget: A Proven Method to Remove Data

(technical research project in Computer Science)

The Predictive Policing Controversy

(sociotechnical research project)

by

Chase Lemley

November 2, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.
Chase Lemley

Technical advisor: Yuan Tian, Department of Computer science

STS advisor: Peter Norton, Department of Engineering and Society

General Research Problem

How may the accuracy of predictive models be improved?

In an experiment, Cao and Yang (2015) entered bad data into a predictive spam filter; thereafter its true positive rate fell from 93 to 38 percent. Training data can embed bias, which the algorithm then encodes Courtland (2018). Because algorithms are no more free of bias than the societies that develop them, they do not absolve humans of responsibility for their consequences.

Making Models Fully Forget: A Proven Method to Remove Data

How can developers confirm that data that have been removed from a trained model no longer affects it?

This is an independent research project in the Computer Science department. It is led by Professor Yuan Tian, and graduate student Jianfeng Chi. There is a focus on finding algorithms to give a true benchmark on the impact of the data remaining, and effectiveness of the model after removal.

In machine learning data is used to train models that will be used to interpret new data. Once the model is trained it is impossible to tell the effects each training data point has on the model, and each retraining of the model is an updated version of the previous model unless it is starting from the beginning. This means that if bad data has gone into the model it can be very difficult to fully remove this data without destroying what has been built and starting from the beginning. This presents the need for algorithms that will fully erase the data and its effects in a more efficient manner.

Recently a new unlearning method (SISA) employed a micro retraining method as opposed to starting the whole model from scratch, Bourtoule et al (2020). The proposed method offers a new way to remove data from a model that would guarantee any removed data point has no effect on a model. This assurance would fix a weak point in most unlearning methods, where typically the data will have some residual effects even after its removal (Golatkar et al 2020; Cao, Yang 2015). The new method is faster than deleting the model and retraining from nothing, which is typically needed to get full assurance the data has been fully forgotten by the model.

The SISA method is theoretically capable of the performance enhancements stated in the previous paragraph, but it does not have clearly proven benchmarks to assure the data has been fully forgotten. Some potential benchmarks include using models not trained with the data point and seeing how the two models compare, statistical analysis and algorithms to show upper bounds for forgotten data and effect on accuracy Golatkar et al (2020). Once these benchmarks are in place we will use the python code for the SISA method, and produce results on local machines to show empirical evidence of the claims made in the paper.

The completion of this project will give a testing suite of algorithms that can be used to show numerically how well data has been forgotten, and the impact of the removed data on the accuracy of the model. The plan to develop this testing suite is to run the experiments outlined in unlearning research, use other unlearning research to find tests, and use these tests on the experiments to find the true forgotten rate. In the future this testing suite could become standardized and used to test all machine unlearning methods.

The Predictive Policing Controversy

How have the critics and the defenders of predictive policing advanced their respective agendas?

In predictive policing, police departments use data analytics to allocate law enforcement resources. The entered data originate from case files. Perry et al (2013) states predictive policing can improve law enforcement efficacy. More police departments are adopting the technology, knowing that it can help them with crime rates. As the rapid adoption continues there are growing concerns over the embedded biases within these algorithms. Social groups are competing to determine the future of predictive policing and the extent of its use in law enforcement.

Aguirre et al. (2019) examined advocacies' efforts to limit police departments' use of predictive policing, and to compel departments to disclose their practices. Duursma et al (2018) recommend that the U.N. use similar predictive models to accelerate its responses in unstable areas. The research shown in this paper points to the idea that by using predictive models the U.N. can act earlier on districts that are shown to be a higher risk, helping to keep areas safer than previously possible.

The algorithmic justice league, an organization showing the impacts of artificial intelligence, is finding innate biases in predictive algorithms. Buolamwini (2019) The leader of this group is making "scorecards" for facial recognition algorithms. The goal is to show how much bias is embedded within these algorithms. Buolamwini believes that by having standard tests officials can work to regulate predictive policing, and mitigate the biases.

PredPol, a predictive policing company developed with the LAPD, is fighting to say their technology is fair. Moravec (2019) quotes both the LAPD and the CEO of PredPol, Josh

Rubenstein. Rubenstein claims that this technology does not use demographic characteristics, therefore it does not violate any civil rights, or perpetuate racism. Predpol has published their patented algorithms to allow for observations from the public. PredPol's technology is being used by enough police departments that one out of 33 people are in its sphere of influence PredPol (2020).

In a RAND Corporation study, Shapiro (2017) found "no statistical evidence that crime was reduced more" in study districts with predictive policing "than in the control districts."

Police departments typically value predictive policing systems. John Williams, the crime analysis manager of the Minnesota police department, disputes the systems' critics Vlahos (2012). The Los Angeles Police Department believes that predictive policing will allow them to accurately police larger areas Puente (2019).

But the deputy director of the department's Office of Data Analysis, Research and Evaluation disagrees with Williams: "I'm very against using government money for black-box solutions where I can't tell my community what we're doing" Courtland (2018).

Some advocates are fighting to restrict predictive policing. In 2016 the Brennan Center for Justice, a bipartisan law group operating out of the New York University Law School, sued the New York Police Department to compel the department to disclose its predictive policing methods Brennan Center v. NYPD (2017).

Another advocate, the High-level Independent Panel on Peace Operations (HIPPO), an independent group finding emerging needs in peacekeeping for the U.N., favors predictive policing. One emerging need identified is the strengthening of analytical technology United Nations (2015). They have since urged the U.N. to start adopting these technologies.

References

Aguirre, K., Badran, E., & Muggah, R. (2019). Annex 1.: Selected applications of predictive policing and evaluation results (FUTURE CRIME:, pp. 12–13). *Igarape Institute*. JSTOR

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2020). Machine Unlearning. *ArXiv:1912.03817 [Cs]*. http://arxiv.org/abs/1912.03817

Brennan Center v. NYPD (2017). Brennan Center for Justice at New York University School of Law v. New York City Police Department and James P. O'Neill. Supreme Court of the State of New York County of New York.

Buolamwini, J. (2019). The Algorithmic Justice League. *Medium.* https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148

Cao, Y., & Yang, J. (2015). Towards Making Systems Forget with Machine Unlearning. *2015 IEEE Symposium on Security and Privacy*, 463–480. https://doi.org/10.1109/SP.2015.35

Courtland, R. (2018). Bias detectives: The researchers striving to make algorithms fair. *Nature*, *558*(7710), 357–360. https://doi.org/10.1038/d41586-018-05469-3

Duursma, A., & Karlsrud, J. (2018). Predictive peacekeeping: Opportunities and challenges. *Norwegian Institute of International Affairs* (NUPI). JSTOR

Golatkar, A., Achille, A., & Soatto, S. (2020). Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. *ArXiv:1911.04933 [Cs, Stat]*. http://arxiv.org/abs/1911.04933

Moravec, S. (2019) Do Algorithms Have a Place in Policing? *The Atlantic*. https://www.theatlantic.com/politics/archive/2019/09/do-algorithms-have-place-policing/596851/

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). Using Predictions to Support Investigations of Potential Offenders. In *Predictive Policing* (pp. 81–114). RAND Corporation.

PredPol (2020). *PredPol*. https://www.predpol.com/about/

Puente, M. (2019, July 3). LAPD pioneered predicting crime with data. Many police don't think it works. *Los Angeles Times*. https://www.latimes.com/local/lanow/la-me-lapd-precision-policing-data-20190703-story.html

Shapiro, A. (2017). Reform predictive policing. *Nature News*, *541*(7638), 458. https://doi.org/10.1038/541458a

United Nations, Uniting our Strengths for Peace - Politics, Partnership and People : Report of the High-Level Independent Panel on Peace Operations. (2015). https://www.refworld.org/docid/558bb0134.html

Vlahos, J. (2012). The Department Of Pre-crime. *Scientific American*, *306*(1), 62–67. JSTOR