QUALIFIER QUESTION

SCOTT LILLEBOE
CS 6460, GA TECH, FALL 2018

QUESTION

Bias in algorithms is a very hot topic these days, but bias in any form is often difficult to self-observe. What are the quantitative methods by which this bias can be detected and quantified so as to make the assessment more objective?

RESPONSE

INTRODUCTION

Bias in algorithms can be found in various forms and contain subtle differences in definition. Machine Learning (ML) and artificial intelligence (AI) algorithms require data to be modeled from in the form of datasets. Bias in the datasets can come from the type (images, language, etc.), how the dataset was sampled, features of the dataset used in the algorithm modeling, and correlations of the features. Another form of bias is in how the models fit the underlying datasets they are modeled from. The last form of bias to be explored is in the cognitive interpretation of the results of the algorithms.

DATASET BIAS

RECOGNITION

Here we are defining recognition datasets as image data used for recognition algorithms. Image data can suffer from various forms of bias such as: selection, capture, category/label, and negative set biases. A method to detect and measure recognition data bias is to perform cross-dataset generalization. This technique trains the recognition algorithm on one dataset but tests on another one. This isn't the same as slicing up a homogenous dataset into training, verification, testing sets. This measuring technique helps to detect how well the dataset allows for generalization of the target image data (Torralba & Efros, 2011). An example of biased image datasets can be seen in some facial recognition software that can't classify certain races correctly due to the algorithm being trained on a dataset that disproportionally represents a certain race (Buolamwini, 2016).

SAMPLE

ML and AI datasets can be created with sampling selection bias due to the dataset not correctly representing the underlying populations (Cortes, Mohri, Riley, & Rostamizadeh, 2008). This form of bias can be a result of non-randomized sample selection from an underlying dataset caused by a defect in the sample selection process. Sample selection bias is often detected using bivariate and multiple regression methods (Cuddeback, Wilson, Orme, & Combs-Orme, 2004).

FEATURES

Bias can be introduced from features of the dataset that should be omitted prior to the algorithms modeling the data. An example of bias introduced from features can be discrimination of certain minority or disadvantaged groups. Direct discrimination can be features explicitly mentioning those groups (i.e. race, gender, sexual orientation). Indirection discrimination can be features that may not explicitly mentioned those groups but produce results that do produce discriminatory decisions. Direct and indirect bias detection and measurement can be done through identifying a set of α-discriminatory (direct) and redlining (indirect) rules. Proposed processes also exist to measure the success of removal of direct and/or indirect discriminatory bias for a dataset. Four metrics used to make these measurements are direct discrimination prevention degree (DDPD), direct discrimination protection preservation (DDPP), indirect discrimination prevention degree (IDPD), and indirect discrimination protection preservation (IDPP) (Hajian & Domingo-Ferrer, 2013).

WORD EMBEDDING

There can be bias introduced into an algorithm through word embedding. Word embeddings encode words into a low dimensional continuous space to preserve semantic and syntactic information (Li, 2015). For example, gender bias is often undesirable and can be introduced through word embeddings. Identifying the gender subspace is a technique that can help identify direct and indirect gender bias. Measuring the direct bias in gender takes the identification of words considered gender-neutral and comparing those against the formulated gender subspace; however, this does not capture indirect gender bias. Indirect gender bias detection requires more subtle relationships between words to be identified and more vector comparisons to the gender subspace. These methods may be able to be expanded to other types of bias such as racial, ethnic and cultural stereotypes (Bolukbasi, 2016).

ALGORITHM DESIGN BIAS

Bias defined in algorithmic design is the error from flawed assumptions of the algorithm itself. ML and AI algorithms are generally trained, verified and tested with datasets. The algorithms are modeled from a training set that is a subset generated from the dataset. These models are then verified and tested from other subsets of the total dataset. The sizes of these various datasets are not always predetermined and depend on the model, dataset size, and other factors (James, Witten, Hastie, & Tibshirani, 2013). When bias in this context is considered high the algorithm can miss applicable associations and is considered to underfit the data. Bias and its counterpart variance can be interpreted graphically or automatically through software packages to find a tradeoff between the two metrics that allows for a model complexity of acceptable accuracy (Fortmann-Roe, 2012). Measuring and detecting potential bias in the sampling of the training, verification, and testing subsets can be done by calculating various metrics of the resulting model. Those metrics can include accuracy, precision, recall, F1 score, ROC curves (Powers, 2011).

RESULT INTERPRETATION BIAS

Biased interpreted results from ML and AI algorithms can be caused by the datasets and algorithm design; however, even with correct data and proper algorithms their interpretation can be cognitively biased. This form of bias of the results can be classified into different forms that distort the interpretation of the algorithm's results. There is no generally accepted measure of interpretability of ML model results. There is research into the concept of plausibility to measure interpretability (Kliegr, 2018).

There can be difficulties in measuring bias by those who didn't train and create the algorithms. Discovering bias in algorithms for these cases can be difficult due to many of the algorithms being proprietary and black boxes

(Savage, 2016). There are some proposed methods on measuring bias in algorithms where the algorithm is a black box and the only way to measure it is by feeding the algorithm data and interpreting the results. Pitoura distinguishes bias as two types: user or content. User bias is defined as bias against users receiving the information while content bias refers to bias in the information sent to the users. His teams bias measuring system claims bias if the measurements exceed a small threshold error value (Pitoura, 2018).

REFERENCES

Bolukbasi, T. C. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 4349-4357.

Buolamwini, J. (2016, November). *How I'm fighting bias in algorithms [Video file]*. Retrieved from TED: https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample Selection Bias Correction Theory. *Algorithmic Learning Theory, 5254*.

Cuddeback, G. S., Wilson, E. E., Orme, J. G., & Combs-Orme, T. (2004). Detecting and Statistically Correcting Sample Selection Bias. *Journal of Social Service Research, 30*(3).

Fortmann-Roe, S. (2012, June). *Understanding the bias-variance tradeoff*. Retrieved from http://scott.fortmann-roe.com/docs/BiasVariance.html

Hajian, S., & Domingo-Ferrer, J. (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering, 25*(7), 1445-1459.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

Kliegr, T. B. (2018). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *arXiv preprint arXiv:1804.02969*.

Li, Y. X. (2015). Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective. *IJCAI*, 3650-3656.

Pitoura, E. T. (2018). On Measuring Bias in Online Information. *ACM SIGMOD Record*, 16-21.

Powers, D. M. (2011). Evaluation: From Precision, Recall and F-Factor to ROC ... *Journal of Machine Learning Technologies, 2*(1).

Torralba, A., & Efros, A. (2011). Unbiased look at dataset bias. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1521-1528.