Fairness-Aware Classification Algorithms

In recent years, researchers have began raising the issue of fairness and bias in classification. [2]. There have been concerns about classification algorithms or some models created with those algorithm with some datasets being "unfair".

Classification is a supervised learning technique that relies on using existing data to construct a model to use for predictions. The use of data points and construction of a model based on them necessitates distinction between different instances (i.e. discrimination). The idea of classification is inherently discriminatory, however, in the context of this essay, to be unfair, discriminatory or bias will refer to classification based on *protected attributes*. [1]. There are ethical and legal issues associated with this problem and some decisions that are made based on the output of classifiers may seriously affect people's lives (e.g. credit score). Therefore, the data mining community has began investigating ways in which we can remedy it which this essay will discuss.

A Naive Solution

A naive solution to the problem of sensitive features would be to remove them from the dataset. E.g. if we were in the business of estimating an individual's credit score and there was a dataset with columns: "FirstName", "LastName", "Age", "Race", "Salary", "HistoryOfIllness" ... based on which the credit score was to be computed, we would remove "Age" and "Race" to ensure that these characteristics don't influence the final decision. Researchers have pointed out a flaw in this reasoning, namely that these removed sensitive features *still* influence other features in an indirect fashion (e.g. age may be correlated with salary) . This effect has been termed *red-lining effect* and is an instance of indirect discrimination. [4, p. 36] [3, p. 378] [3, p. 380] [3, p. 381]

Measuring Fairness

A major problem within the area of fairness-aware classification is measuring the extent of bias and fairness. [9, p.281] Various suggestions have been put forward.

Indre in his survey discusses the various measures that have emerged to calculate the degree of bias. He categorises them into: statistical tests, absolute measures, conditional measures and structural measures. He argues that to determine whether the classifier is fair, we must have access to protected characteristics to be able to assess the degree of fairness of predictions with respect to them. [8, p. 8]

Kamishima et al. suggest that it may be formalized as lack of correlation between sensitive features and the outcome. [5, p. 259].

A measure of fitness invented by Calders and Verwer's was *Calders and Verwer's Discrimination Score* (CV score) [9, p. 281]. It has been used by a number of researchers to estimate discrimination. [4, p. 44] [3, p. 381]

[1] In the UK The Equality Act lists: age, gender reassignment, being married or in a civil partnership, being pregnant or on maternity leave, disability, race including colour, nationality, ethnic or national origin, religion or belief, sex, sexual orientation as *protected attributes*

CV score is defined formally as: [2]

$$1 - (P[\hat{Y} = 1|S = 1] - P[\hat{Y} - 1|S \neq 1])$$

An approach might be taken to combat bias in classification is taking a measure of bias such as CV score [3] and tweaking the model in such a way that the score begins to approach 0.

Proposed Improvements

The fundamental issue is that the traditional algorithms used in data mining are not aware of issues of social equality. They treat all features in a uniform manner. To remedy that a number of modification to existing algorithms have been proposed.

Calders and Verwer's two-naive-Bayes (CV2NB)

One of the algorithms that makes use of CV score is the CV2NB invented by Calders and Verwer [9]. In this approach when dealing with a binary sensitive attribute, two models are constructed which are trained on a portion of the dataset where the attribute have a value of 0 and 1. E.g. one model for males and one for females. When making predictions to algorithm makes a choice about which model to choose based on the value of the sensitive attribute. [9, p. 283]. A downside of the paper was that their comparison discussed 3 modifications to the Naive Bayes algorithm but did not compare them to the original algorithm which made it impossible to see whether their alterations fixed the issue of discrimination and how much accuracy was lost.

Prejudice Remover Regularizer

A different approach was taken by Kamishima et al. The researchers used PI as a measure of discrimination and proposed a regularizer which attempts to remove indirect prejudice by modifying the original equation of the classifier [4]. The modification requires "modest" resources and is usable with "any prediction algorithm with probabilistic discriminative model" [4, p. 35]. Kamishima et al. compared their results with CV2NB and managed to obtain promising results. Their regularizer manages to decrease discrimination so that it becomes competitive with CV2NB. By setting $\eta = 20$, [5] the researchers were able to decrease discrimination to 0.010 on the CV scale but at the expense of accuracy which dropped to 0.769. This is compared to 0.842 accuracy with $\eta = 5$.

Importantly, the use of those interventions, in both cases, resulted in decreased accuracy. On the other hand, they did decrease discrimination as measured by CV and prejudice index. [4, p. 44] [9] This is problematic since the ultimate aim of classification is to make *accurate* predictions.

Conclusion

Efforts have been made to address the issue of fairness in classification. These include: invention of measures to estimate discrimination and modification of existing algorithms to reduce bias. Two such modifications were discussed here. There seems to be a trade-off between accuracy and fairness which was evident in the lowering of accuracy as discrimination decreased. The fact that different researchers used different measures to calculate bias suggests there does not seem to be an agreement in the research community on: which measure to use and if any of the measures is a good reflection of the extent of "unfairness".

[2] where P is the probability, S is the sensitive attribute, Y is the actual value and $\hat{Y}$ is the predicted value

[3] There are also other measures such as *Disparate Impact (DI)* and Prejudice Index (PI).

[4] the authors use logistic regression to demonstrate it

[5] η regulated the amount of discrimination removed

References

[1] *A comparative study of fairness-enhancing interventions in machine learning*, A. Friedler, Sorelle & Scheidegger, Carlos & Venkatasubramanian, Suresh & Choudhary, Sonam & P. Hamilton, Evan & Roth, Derek, 2018.

[2] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini, *Discrimination-aware data mining*, In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Pages 560 - 568, 2008.

[3] T. Kamishima, S. Akaho, H. Asoh and J. Sakuma, *Considerations on Fairness-Aware Data Mining*, IEEE 12th International Conference on Data Mining Workshops, Brussels, Pages 378 - 385, 2012.

[4] Kamishima, Toshihiro and Akaho, Shotaro and Asoh, Hideki and Sakuma, Jun, *Fairness-Aware Classifier with Prejudice Remover Regularizer* Machine Learning and Knowledge Discovery in Databases, Pages 35 - 50, Springer Berlin Heidelberg, 2012.

[5] Kamishima, Toshihiro and Akaho, Shotaro and Asoh, Hideki and Sakuma, Jun, *Model-based and actual independence for fairness-aware classification*, Data Mining and Knowledge Discovery, Volume 32, Pages 258 - 286, 2018.

[6] T. Kamishima, S. Akaho, H. Asoh and J. Sakuma, *The Independence of Fairness-Aware Classifiers*, IEEE 13th International Conference on Data Mining Workshops, Dallas, TX, Pages 849 - 858, 2013.

[7] Aditya Krishna Menon and Robert C. Williamson, *The Cost of Fairness in Binary Classification*, Proceedings of Machine Learning Research, Conference on Fairness, Accountability, and Transparency, Pages 1 - 12, 2018.

[8] Zliobaite, Indre, *A survey on measuring indirect discrimination in machine learning*, 2015.

[9] Toon Calders and Sicco Verwer, *Three Naive Bayes Approaches for Discrimination-Free Classification*, Data Mining journal, 2010.