Fairness-Aware Classification Algorithms

In recent years, researchers have begun raising the issue of fairness and bias in classification. The idea of fairness-aware data mining emerged in 2008 thanks to the contribution of Pedreshi et al. [5, p. 849] [1]. 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 biased 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). This essay will discuss ways in which we can tackle this issue by focusing on modifications to the Naive Bayes (NB) algorithm.

A Naive Solution

A naive solution to the problem of sensitive features would be to remove them from the dataset, in order to diminish the effect of sensitive attributes on the predictions. 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 has been termed the *red-lining effect*, and is an instance of indirect discrimination. [3, p. 36] [2, p. 378] [2, p. 380] [2, p. 381]. Removing sensitive features does diminish discrimination, as measured by a discrimination metric, but it does not remove it completely. This was seen in Kamishima et al.'s study which among other algorithms compared NB with NB without sensitive features. [5, p. 855]

Measuring Fairness

A major problem within the area of fairness-aware classification is measuring the extent of bias and fairness. [6, p.281] Various suggestions have been put forward.

Kamishima et al. suggest that it may be formalized as lack of correlation between sensitive features and the outcome. [4, p. 259].

A measure of fitness invented by Calders and Verwer's was *Calders and Verwer's Discrimination Score* (CV score) [6, p. 281]. It has been used by a number of researchers to estimate discrimination. [3, p. 44] [2, p. 381]. CV score is defined formally as: [2]

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

"[...] As this score increases, the members of unprotected group get favorable treatment more frequently while those in the protected group get favorable treatment less frequently. [...]" – Kamishima et al. [5, p .852]

Kamishima et al. used Normalized Prejudice Index (NPI) when comparing variants of NB algorithms [3]. It expresses discrimination as a real number in $[0, 1]$. The measure can be expressed formally as: [4]

[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*

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

[3] NPI is a modification of Prejudice Index (PI) [3]

[4] where H is an entropy function

$$NPI = \frac{I(\hat{Y}; S)}{\sqrt{(H(\hat{Y})H(S))}}$$

An approach might be taken to combat bias in classification is taking a measure of bias such as CV or NPI [5] 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 [6]. 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 the algorithm makes a choice about which model to choose based on the value of the sensitive attribute. [6, p. 283]. A downside of the paper was that their comparison discussed 3 modifications to the NB 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. That said, we know from other studies that it does perform very well. [5, p. 855] [3, p. 44]

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 [6]. The modification requires "modest" resources and is usable with "any prediction algorithm with probabilistic discriminative model" [3, 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 $\eta = 20$, [7] the researchers were able to decrease discrimination to 0.01 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. [3, p. 44] [6] 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 measures to use and if any of the measures are a good reflection of the extent of "unfairness".

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

[6] the authors use logistic regression to demonstrate it

[7] η regulated the amount of discrimination removed

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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