Discrimination Aware Classification[1]

Faisal Kamiran Toon Calders

Eindhoven University of Technology
The Netherlands

In classification, models for predicting the class of future examples are learned on the basis of training data. The quality of these models depends critically on the quality of this training data. Often, however, training data is biased in an unacceptable way towards certain groups or classes of objects. Consider, e.g., the following situation: throughout the years, in a certain organization systematically *Black* people have been denied from jobs. As such, the historical employment information of this company concerning recruitment decisions is biased towards giving jobs to white people while denying jobs from black people. In order to reduce this type of racial discrimination, states enacted new laws requiring equal job opportunities, and all organization are enforced to employ minimum quota for *Black* employees. Suppose now that this same company wants to partially automate its recruitment strategy by learning a classifier that predicts the most likely candidates for a job. As the historical recruitment data of the company is biased, a model trained on this data may show unlawfully prejudiced behavior in future predictions. This partial attitude of the learned model leads to discriminatory outcomes for future unlabeled data objects. Clearly, even though the data contains a lot useful information, not taking into account this discrimination will lead to an unacceptable classifier. In this paper we tackle exactly this problem: *How can we train an unbiased classifier when the training data is biased?* Some other real-world situations where this problem is relevant include:

- Even though there is clear historical evidence showing higher accident rates for male drivers, insurance companies are not allowed to discriminate based on gender in many countries. In this case the historical data is biased towards assigning a higher risk class to male drivers.
- Often, salaries of women are lower than those of men. Nevertheless, when training a classifier in order to decide in which salary scale to employ a new-hire, it is undesirable to have this inequality in the learned model.

In above mentioned cases, the training data is biased. Classification models trained on such data will not fulfill the future requirements. Future data objects must follow a different class label distribution than that of the training data. So, sometimes impartial classification results are required for future data objects in spite of having discriminatory training data.

Most of the classification models, however, deal with all the attributes equally when classifying data objects and take no care about the sensitivity of attributes. Simply removing these discriminatory attributes (e.g., *Ethnicity*) from the training data is not enough to solve this problem because often other attributes will still allow for the identification of the discriminated community. For example, the ethnicity of a person might be strongly linked with the postal code of his residential area, leading to a classifier with indirect racial discriminatory behavior based on postal code. This effect and its exploitation is often referred to as *redlining*, stemming from the practice of denying or increasing services such as, e.g., mortgages or health care to residents in certain often racially determined areas. The term redlining was coined in the late 1960s by community activists in Chicago[2]. Different experiments conducted by the authors of [4] and ourselves support this claim: even after removing the discriminatory attributes from the dataset discrimination persists.

In this paper, we introduce a classification model which is learnt on biased training data but works impartially for future data and refer to this model as *Classification with No Discrimination (CND)*. *CND* assumes that historical data knowing to contain discrimination is available. Our approach consists of first

[1] Extended abstract of the paper: F. Kamiran and T. Calders. Classifying without discriminating. In *IEEE International Conference on Computer, Control & Communication (IEEE-IC4)*. IEEE press, 2009.

[2] Source: `http://en.wikipedia.org/wiki/Redlining`, September 30th, 2008

"massaging" the data to remove the discrimination with the least possible changes. For massaging the data, *CND* learns a (biased) ranker for predicting the class attribute without taking into account the discriminatory attribute. This ranker will then be used to rank the data objects according to their probability of being in the desired class, e.g., job = yes. Any ranking algorithm may be used, but for the experiments in [2], we used a Naive Bayesian classifier for calculating the class probability of each data tuple. To this end, the class labels of the most likely *victims* (training instances of the discriminated community with a negative label but a high positive class probability) and *profiters* (training instances of the favored community with a positive label but a low positive class probability) will be changed. In our job application example, the list of *Victims* will consist of all *Black* applicants with good probability of getting the job but negative label. Similarly the list of *profiters* will contain those *white* people with positive label but low probability. The modified data is then used for learning a classifier with no discrimination for future decisions. The fact that the final model is then learned on the cleaned, non-discriminatory data reduces the prejudicial behavior for future classification. Obviously, changing the training data might result in lower accuracy scores. Nevertheless, as we try to keep the changes as minimal and least intrusive as possible, the trade-off between accuracy and non-discrimination will be minimal.

The CND method was implemented and tested on a the German Credit Dataset available in the UCI ML-repository [3] for our experiments. The dataset has 1000 instances which classify the bank account holders into credit class *Good* or *Bad*. In our experiments, we compare the following two approaches:

1. Our proposed approach; i.e., we will use *CND* for massaging the training data to make it discrimination free. The ranking function will be based on a Naive Bayesian model learned on the raw data. Then we learn a Naive Bayesian classifier *CND* on the discrimination-free data.

2. For reasons of comparison, we also learn a Naive Bayesian classifier directly on the original data without massaging. We refer to this second approach as "*Classification without the Massaging*". We further explore the problem of discrimination in [1].

We find that *CND* classifies the future data with minimum discrimination. Though the discriminatory behavior of the classification models is affected by the change of discrimination level in the data, *CND* always shows more impartiality as compared to the *Classification without Massaging*.

So, we conclude that the notion of discrimination is non trivial and poses ethical and legal issues as well as obstacles in practical applications. *CND* provides us with a simple yet powerful starting point for the solution of the discrimination problem. *CND* classifies the future data (both discriminatory and non discriminatory) with minimum discrimination and high accuracy. It also addresses the problem of redlining.

References

[1] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE ICDM Workshop on Domain Driven Data Mining*. IEEE press. Accepted for publication, 2009.

[2] F. Kamiran and T. Calders. Classifying without discriminating. In *IEEE International Conference on Computer, Control & Communication (IEEE-IC4)*. IEEE press, 2009.

[3] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. (uci) repository of machine learning databases. 1998.

[4] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.