A Causal Framework for Discovering and Removing Direct and Indirect Discrimination

Lu Zhang, Yongkai Wu, and Xintao Wu
University of Arkansas
{lz006,yw009,xintaowu}@uark.edu

Abstract

In this paper, we investigate the problem of discovering both direct and indirect discrimination from the historical data, and removing the discriminatory effects before the data is used for predictive analysis (e.g., building classifiers). The main drawback of existing methods is that they cannot distinguish the part of influence that is really caused by discrimination from all correlated influences. In our approach, we make use of the causal network to capture the causal structure of the data. Then we model direct and indirect discrimination as the *path-specific effects*, which accurately identify the two types of discrimination as the causal effects transmitted along different paths in the network. Based on that, we propose an effective algorithm for discovering direct and indirect discrimination, as well as an algorithm for precisely removing both types of discrimination while retaining good data utility. Experiments using the real dataset show the effectiveness of our approaches.

1 Introduction

Discrimination refers to unjustified distinctions in decisions against individuals based on their membership in a certain group. Laws and regulations have been established to prohibit discrimination on several grounds, such as gender, age, sexual orientation, race, religion, and disability, which are referred to as the *protected attributes*. Various predictive models have been built around the collection and use of historical data to make important decisions like employment, credit and insurance. If the historical data contains discrimination, the predictive models are likely to learn the discriminatory relationship present in the historical data and apply it when making new decisions. Therefore, it is imperative to ensure that the data goes into the predictive models and the decisions made with its assistance are not subject to discrimination.

In the legal field, discrimination falls into direct and indirect discrimination. Direct discrimination occurs when individuals receive less favorable treatment explicitly based on the protected attributes. An example would be rejecting a qualified female applicant in applying a university just because of her gender. Indirect discrimination refers to the situation where the treatment is based on apparently neutral non-protected attributes but still results in unjustified distinctions against individuals from the protected group. A well-known example of indirect discrimination is redlining, where the residential Zip Code of the individual is used for making decisions such as granting a loan. Although Zip Code is apparently a neutral attribute, it correlates with race due to the racial composition of residential areas. Thus, the use of Zip Code may indirectly lead to racial discrimination.

Discrimination discovery and removal from historical data has received an increasing attention over the past few years in data science [Hajian and Domingo-Ferrer, 2013; Kamiran and Calders, 2012; Ruggieri *et al.*, 2010; Romei and Ruggieri, 2014; Feldman *et al.*, 2015]. Many approaches have been proposed to deal with both direct and indirect discrimination but significant issues exist. For discrimination discovery, the difference in decisions across the protected and non-protected groups is a combined (not necessarily linear) effect of direct discrimination, indirect discrimination, and explainable effect that should not be considered as discrimination (e.g., the difference in average income of females and males caused by their different working hours per week). However, existing methods cannot explicitly and correctly identify the three different effects when measuring discrimination. For example, the classic metrics *risk difference*, *risk ratio*, *relative chance*, *odds ratio*, etc. [Romei and Ruggieri, 2014] treat all the difference in decisions as discrimination. [Žliobaitė *et al.*, 2011] realized the explainable effect but failed to distinguish the effects of direct and indirect discrimination. For discrimination removal, a general requirement is to preserve the data utility while achieving non-discrimination. As we shall show in the experiments, a crude method that totally removes all connections between the protected attribute and decision (e.g., in [Feldman *et al.*, 2015]) can eliminate discrimination but may suffer significant utility loss. To maximize the preserved data utility, it is necessary to first accurately measure the discriminatory effects and then precisely remove them.

The causal modeling based discrimination detection has been proposed most recently [Zhang *et al.*, 2016c; 2016b] for improving the correlation based approaches. However, these work also do not tackle indirect discrimination. In this paper, we develop a framework for discovering and removing both direct and indirect discrimination based on the causal network. A causal network is a directed acyclic graph