Anti-discrimination learning: a causal modeling-based framework

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

***Abstract*—Anti-discrimination learning is an increasingly important task in data mining and machine learning fields. Discrimination discovery is the problem of unveiling discriminatory practices by analyzing a dataset of historical decision records, and discrimination prevention aims to remove discrimination by modifying the biased data and/or the predictive algorithms. Discrimination is causal, which means that to prove discrimination one needs to derive a causal relationship rather than an association relationship. Although it is well-known that association does not mean causation, the gap between association and causation is not paid enough attention by many researchers. The aim of this tutorial is to point out the limitations existing in current association-based approaches, introduce a causal modeling-based framework for anti-discrimination learning, and suggest potential future research directions.**

I. Importance and Target Audience

Discrimination discovery and prevention have been an active research area in data science due to increasing worries of discrimination as data analytic technologies could be used to digitally unfairly treat unwanted groups, such as customers, employees, tenants, or recipients of credit. In 2014, U.S. President Obama called for a 90-day review of data collecting and analyzing practices. An important conclusion from the resulting report [1] is that "*Big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups*". In May 2016, the Executive Office of the President made recommendations to "support research into mitigating algorithmic discrimination, building systems that support fairness and accountability, and developing strong data ethics frameworks" [2]. How to ensure non-discrimination in social computing and behavioral modeling & prediction is an important and challenging topic.

The tutorial targets the researchers interested in studying the issue of discovering and preventing discrimination caused by data mining and machine learning algorithms from the causal modeling perspective. The audience is assumed to be familiar with the fundamental concepts in data mining and machine learning, especially in predictive learning. No special requirement is needed on software and hardware.

II. Outline

The tutorial is organized as six parts, including an introduction part, a literature review part, three main technical parts, and a concluding part. The tutorial is based in part on our position paper [3].

1. Introduction, motivation and challenges
   - 1.1. Legal definitions and principles of discrimination
   - 1.2. Motivation of anti-discrimination learning
   - 1.3. Challenge in discrimination discovery and prevention
2. Association-based anti-discrimination literature review
   - 2.1. Approaches for discrimination discovery
   - 2.2. Approaches for discrimination prevention
   - 2.3. Gap between association and causation
3. Causal modeling-based anti-discrimination framework
   - 3.1. Causal modeling background
   - 3.2. Discrimination categorization and anti-discrimination framework overview
4. System-level discrimination discovery and removal
   - 4.1. Modeling of direct and indirect discrimination
   - 4.2. Quantitative discrimination criterion
   - 4.3. Discrimination removal algorithms
   - 4.4. Ensuring non-discrimination in prediction
5. Group and individual-level discrimination
   - 5.1. Approach for group-level direct discrimination
   - 5.2. Approach for individual-level direct discrimination
6. Challenges and directions for future research

III. Tutors' short bio

Dr. Lu Zhang is a postdoctoral researcher in the Computer Science and Computer Engineering Department, University of Arkansas. He received the BEng degree in computer science and engineering from the University of Science and Technology of China, in 2008, and the PhD degree in computer science from Nanyang Technological University, Singapore in 2013. His research interests include data mining algorithms, discrimination-aware data mining, and causal inference.

Yongkai Wu is a Ph.D. student in the Department of Computer Science and Computer Engineering at the University of Arkansas.

Dr. Xintao Wu is a Professor in the Department of Computer Science and Computer Engineering at the University of Arkansas. His major research interests include data privacy, bioinformatics and discrimination-aware data mining.

References

[1] Big data: Seizing opportunities, preserving values. Whilte House (2014)

[2] Munoz, C., Smith, M., Patil, D.: Big data: A report on algorithmic systems, opportunity, and civil rights. Executive Office of the President (2016)

[3] Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework, JDSA, to appear.