Information Sciences 425 (2018) 18–33

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Exploiting reject option in classification for social discrimination control

Faisal Kamiran[a,*], Sameen Mansha[b,a], Asim Karim[c,a], Xiangliang Zhang[d]

[a] Information Technology University, Lahore, Pakistan
[b] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia
[c] Department of Computer Science, Syed Babar Ali School of Science and Engineering, Lahore University of Management Sciences, Lahore, Pakistan
[d] Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology, KSA, Saudi Arabia

ARTICLE INFO

Article history:
Received 30 March 2016
Revised 1 September 2017
Accepted 27 September 2017
Available online 28 September 2017

Keywords:
Discrimination-aware data mining
Fairness in machine learning
Classification
Decision theory

ABSTRACT

Social discrimination is said to occur when an unfavorable decision for an individual is influenced by her membership to certain protected groups such as females and minority ethnic groups. Such discriminatory decisions often exist in historical data. Despite recent works in discrimination-aware data mining, there remains the need for robust, yet easily usable, methods for discrimination control. In this paper, we utilize reject option in classification, a general decision theoretic framework for handling instances whose labels are uncertain, for modeling and controlling discriminatory decisions. Specifically, this framework permits a formal treatment of the intuition that instances close to the decision boundary are more likely to be discriminated in a dataset. Based on this framework, we present three different solutions for discrimination-aware classification. The first solution invokes probabilistic rejection in single or multiple probabilistic classifiers while the second solution relies upon ensemble rejection in classifier ensembles. The third solution integrates one of the first two solutions with situation testing which is a procedure commonly used in the court of law. All solutions are easy to use and provide strong justifications for the decisions. We evaluate our solutions extensively on four real-world datasets and compare their performances with previously proposed discrimination-aware classifiers. The results demonstrate the superiority of our solutions in terms of both performance and flexibility of applicability. In particular, our solutions are effective at removing illegal discrimination from the predictions.

1. Introduction

Social discrimination is said to occur when a decision in favor of or against a person is made based on the group, class, or category to which that person belongs to rather than on merit. Discriminatory practices suppress opportunities for members of deprived groups in employment, income, education, finance, and other benefits/services on the basis of their age, gender, skin color, religion, race, language, culture, marital status, economic condition, and other non-merit factors.

* Corresponding author.
E-mail addresses: faisal.kamiran@itu.edu.pk (F. Kamiran), s.mansha@uqconnect.edu.au (S. Mansha), akarim@lums.edu.pk (A. Karim), xiangliang.zhang@kaust.edu.sa (X. Zhang).

https://doi.org/10.1016/j.ins.2017.09.064
0020-0255/

Today, discrimination is considered unacceptable from social, ethical, and legal perspectives. Many anti-discrimination laws [3,11,27,28] have been enacted and many anti-discrimination organizations (e.g., ENAR [1]) are working for the eradication of discrimination. The consequences of discriminatory practices can range from legal prosecution to a variety of social problems like high unemployment rate, frustration, low productivity, and social unrest.

The discrimination-aware classification problem studies the construction and application of classifiers learned from discriminatory or biased data. The do-nothing approach of simply using a classifier learned from discriminatory data will propagate, if not exacerbate, discriminatory decisions, which is undesirable for decision makers at financial institutions, hiring agencies, and social service providers. Thus, this do-nothing approach can lead to litigations and penalties.

In recent years, several methods have been proposed for discrimination-aware classification. However, these methods have one or both of the following shortcomings. First, they require that either the discriminatory data is processed to remove discriminatory patterns before learning a classifier or a specific classifier's learning algorithm is modified to make it discrimination-aware. Second, they are usually 'brute force' techniques with limited control over overall and illegitimate (unexplainable) discrimination removal.

These shortcomings of existing methods have hindered their adoption by practitioners. A direct consequence of the first shortcoming is that whenever discrimination w.r.t. a different sensitive attribute needs to be addressed, the historical data or classifier needs to be processed again. Our experience with the *Dutch Research and Documentation Center* (WODC) associated with the Ministry of Security and Justice and *Statistics Netherlands*, the national census body, confirms the importance of tackling discrimination w.r.t. multiple factors including age, gender, and race [18]. Being restricted to a specific discrimination-aware classifier (e.g., naive Bayes [7], decision tree [17]) is also an issue because that classifier may not be the best performing classifier for a given dataset. The second shortcoming can lead to reverse discrimination whereby deprived group individuals are favored without a legitimate or plausible explanation. This issue has been studied by the authors of Zliobaite et al. [32]. They split overall discrimination into legal and illegal parts and claim that if the discrimination (e.g., high income of male employees as compared to female employees) can be explained by some reasonable factors (e.g., longer working hours of males), then it is acceptable and legitimate 'discrimination' rather than illegal discrimination. On the other hand, it would be illegal to discriminate on the basis of sensitive factors (e.g., gender, race) without any plausible explanation. The current state-of-the-art methods either deal with the overall discrimination or illegal discrimination and are not flexible enough to prevent both overall and illegal discrimination simultaneously.

In this paper, we develop and evaluate a methodology for making single and ensembles of classifiers discrimination-aware w.r.t. overall and illegal discrimination. This methodology is based on the decision theoretic notion of reject option where instances with highly uncertain labels are not given one in classification (i.e., they are given the reject label). Previously, it has been hypothesized that discriminatory decisions are often made close to the decision boundary because of decision maker's bias [16]. Our proposed methodology formalizes this into practically usable solutions for discrimination-aware classification. Furthermore, the rejected instances represent potentially discriminated or favored instances in the biased dataset. Thus, our methodology also serves as a model-based discrimination discoverer in biased datasets.

We present three rejection strategies and corresponding rules for discrimination control in predictions. The first solution called Probabilistic Rejection (PR), rejects instances with uncertain posterior probabilities, thus enabling it to be used with any probabilistic classifier or ensemble of classifiers. Our second rejection strategy, called Ensemble Rejection (ER), identifies instances that are not unanimously labeled by an ensemble of classifiers, thus emulating the natural decision making process by a group of experts. Our third rejection strategy, called Situational Rejection (SR), combines probabilistic rejection or ensemble rejection with situation testing to identify discriminated instances. Situation testing is a legally admissible procedure for verifying discrimination cases by comparing them with other similar cases. All strategies/solutions include relabeling rules with parametric control over the resulting discrimination. We perform extensive experiments to verify the superior performance of our methodology. In particular, we also demonstrate that our methodology prefers removing illegal discrimination over explainable discrimination while reducing overall discrimination. Thus, it addresses a common criticism that discrimination prevention methods disregard explainable discrimination while removing overall discrimination.

The rest of the paper is organized as follows. Section 2 discusses the related work in discrimination-aware classification. Section 3 defines the problem setting and measures for overall and illegal discrimination. We present our reject option based methodology and specific solutions in Section 4. Section 5 presents experimental evaluations and discussions of our solutions. We summarize and conclude our contribution in Section 6.

2. Related work

Data mining techniques can assist with the discovery of discriminatory patterns from data and with preventing discriminatory decisions based on biased data. The topic of social discrimination in data mining was introduced by Pedreschi et al. [24]. Since then many researchers have focused on discrimination detection and prevention in data mining. A multidisciplinary survey of discrimination analysis methods is given by Romei and Ruggieri [25] while an edited book provides a summary of the research works for discrimination discovery and prevention [8]. The book also deals with the legal and ethical issues of discrimination and profiling.

Proposed methods for discrimination prevention requiring learning model adaptation include those for decision trees [17], naive Bayes classifiers [7], logistic regression [20], and support vector machines (SVM) [31]. All these methods require that the learning model or algorithm is tweaked, and these methods are specific to their respective classifiers. For example,