UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

# Measuring discrimination in algorithmic decision making

**Zliobaite, Indre**

**2017**
Springer
http://hdl.handle.net/10138/307578
Zliobaite, I 2017, 'Measuring discrimination in algorithmic decision making', Data Mining and Knowledge Discovery, vol. 31, no. 4, pp. 1060-1089. https://doi.org/10.1007/s10618-017-0506-1

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

**Noname manuscript No.**
(will be inserted by the editor)

# Measuring discrimination in algorithmic decision making

**Indrė Žliobaitė**

Received: 2016 07 05; Revised 2016 10 19; Revised 2nd 2017 01 30

**Abstract** Society is increasingly relying on data-driven predictive models for automated decision making. This is not by design, but due to the nature and noisiness of observational data, such models may systematically disadvantage people belonging to certain categories or groups, instead of relying solely on individual merits. This may happen even if the computing process is fair and well-intentioned. Discrimination-aware data mining studies of how to make predictive models free from discrimination, when the historical data, on which they are built, may be biased, incomplete, or even contain past discriminatory decisions. Discrimination-aware data mining is an emerging research discipline, and there is no firm consensus yet of how to measure the performance of algorithms. The goal of this survey is to review various discrimination measures that have been used, analytically and computationally analyze their performance, and highlight implications of using one or another measure. We also describe measures from other disciplines, which have not been used for measuring discrimination, but potentially could be suitable for this purpose. This survey is primarily intended for researchers in data mining and machine learning as a step towards producing a unifying view of performance criteria when developing new algorithms for non-discriminatory predictive modeling. In addition, practitioners and policy makers could use this study when diagnosing potential discrimination by predictive models.

**Keywords** Discrimination-aware data mining · Fairness-aware machine learning · Accountability · Predictive modeling · Indirect discrimination

I. Žliobaitė
Dept. of Computer Science, University of Helsinki, Finland;
Dept. of Geosciences and Geography, University of Helsinki, Finland E-mail: indre.zliobaite@helsinki.fi

## Contents

1 Introduction . . . 2
2 Background . . . 4
2.1 Discrimination and law . . . 4
2.2 Discrimination-aware data mining . . . 6
2.3 Definition of fairness for data mining . . . 7
2.4 Principles for making predictive models non-discriminatory . . . 9
3 Discrimination measures . . . 10
3.1 Statistical tests . . . 11
3.1.1 Regression slope test . . . 12
3.1.2 Difference of means test . . . 12
3.1.3 Difference in proportions for two groups . . . 13
3.1.4 Difference in proportions for many groups . . . 13
3.1.5 Rank test . . . 13
3.2 Absolute measures . . . 13
3.2.1 Mean difference . . . 14
3.2.2 Normalized difference . . . 14
3.2.3 Area under curve (AUC) . . . 14
3.2.4 Impact ratio . . . 15
3.2.5 Elift ratio . . . 15
3.2.6 Odds ratio . . . 15
3.2.7 Mutual information . . . 16
3.2.8 Balanced residuals . . . 16
3.2.9 Relation between two variables . . . 17
3.2.10 Measuring for more than two groups . . . 18
3.3 Conditional measures . . . 18
3.3.1 Unexplained difference . . . 19
3.3.2 Propensity measure . . . 20
3.3.3 Belift ratio . . . 20
3.4 Situation measures . . . 21
3.4.1 Situation testing . . . 21
3.4.2 Consistency . . . 21
4 Experimental analysis of core measures . . . 22
4.1 Symmetry and boundary conditions . . . 23
4.2 Performance of Difference measures . . . 24
4.3 Performance of Ratios . . . 27
5 Recommendations for researchers and practitioners . . . 28

## 1 Introduction

Nowadays, increasingly many decisions for people and about people are made using predictive models built on historical data, including credit scoring, insurance, personalized pricing and recommendations, automated CV screening of job applicants, profiling of potential suspects by the police, and many more cases. The penetration of data mining and machine learning technologies, as well as decisions informed by big data has raised public awareness that data-driven decision making may lead to discrimination against groups of people [1,10,18,21,23,33,44]. Such discrimination may often be unintentional and unexpected, assuming that algorithms must be inherently objective. Yet decision making by predictive models may discriminate against people, even if the computing process is fair and well-intentioned [5,14,15]. This is because most data mining methods are based upon assumptions that historical datasets are

correct, and accurately represent population, which often appears to be far from reality.

Discrimination-aware data mining is an emerging discipline that studies how to prevent potential discrimination due to algorithms. It is assumed that non-discrimination regulations prescribe which personal characteristics are considered sensitive, or which groups of people are to be protected. The regulations are assumed to be defined externally, typically by national or international legislation. The research goal in discrimination-aware data mining is to translate those regulations mathematically into non-discrimination constraints, and develop predictive modeling algorithms that would be able to take into account those constraints, and at the same time be as accurate as possible. These constraints prescribe how much of differences between groups can be considered explainable. In a broader perspective, research needs to be able to computationally explain the roots of such discrimination events before increasing public concerns lead to unnecessarily restrictive regulations against data mining.

In the last few years researchers have been developing discrimination-aware data mining algorithms using a variety of performance measures. Yet there is a lack of consensus of how to define the fairness of predictive models, and how to measure their performance in terms of non-discrimination. Often research papers propose new ways to quantify discrimination, and new algorithms that would optimize that measure. The existing variety of evaluation approaches makes it difficult to compare results and assess progress in the discipline; furthermore, the variety of measures makes it difficult to recommend computational strategies to practitioners and policy makers.

The goal of this survey is to develop a unifying view towards discrimination measures in data mining and machine learning, and analyze the implications of optimizing one or another measure in predictive modeling. Therefore, it is essential to develop a coherent view early in the development of this research field, in order to present task settings in a systematic way for follow up research, to enable systematic comparison of approaches, and to facilitate a discussion hopefully aimed at reaching a consensus among researchers in terms of the fundamentals of the discipline. For this purpose we review and categorize measures that have been used in data mining and machine learning, and also discuss measures from other disciplines, such as feature selection, which in principle could be used for measuring discrimination. We complement the review by experimental analysis of core measures.

Several surveys on different aspects of discrimination-aware data mining already exist and are complementary to this survey. A previous review [48] presents a multi-disciplinary context for discrimination-aware data mining. The review [48] focuses on approaches to solutions across different disciplines (law, economics, statistics, computer science), rather than analysis and comparison of measures. A yet earlier study [47] discusses a number of measures in relation to association rule discovery task, which in principle can be applied to any classification algorithm. This study discussed four measures that we current categorize under Absolute measures. A recent review [5] discusses the

legal aspects of potential discrimination by machine learning, mainly focusing on American anti-discrimination laws in the context of employment, as well as discussing how big data and machine learning can lead to discrimination attributable to algorithmic effects regardless of jurisdiction. A classical handbook on measuring racial discrimination [8] focuses on surveying and collecting evidence for discrimination discovery. The book does not consider discrimination by algorithms, it only considers discrimination by human decision makers, and therefore presents inspiring ideas, but not solutions for measuring algorithmic discrimination, which is the focus of our survey. Interactions between human and algorithmic decision making is experimentally investigated in a recent study [6].

## 2 Background

The root of the word 'discrimination' is the Latin for *distinguishing*. While distinguishing is not undesirable as such, discrimination has a negative connotation when referring to adversary treatment of people based on belonging to some group rather than their individual merits. Initially associated with racism, nowadays discrimination may refer to a wide range of grounds, such as, race, ethnicity, gender, age, disability, sexual orientation, religion and more. Data mining is not aiming to decide what is the right or wrong reason for distinguishing, but considers sensitive characteristics to be externally decided by social philosophers, policy makers and society itself. The notion of sensitive characteristics can depend on the context and can change from case to case. The role of data mining is to understand generic principles and provide technical expertise on how to guarantee non-discrimination in algorithmic decision making.

### 2.1 Discrimination and law

Public attention to discrimination prevention is increasing, national and international anti-discrimination legislation are expanding the scope of protection against discrimination, and extending discrimination grounds. For instance, the EU is developing a unifying "Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation".

Adversary discrimination is undesired from the perspective of basic human rights, and in many areas of life non-discrimination is enforced by international and national legislation, to allow all individuals an equal prospect to access opportunities available in a society [24]. Enforcing non-discrimination is not only for the benefit of individuals. Considering individual merits rather than group characteristics is expected to benefit decision makers leading to more informed, and likely more accurate decisions.

From the regulatory perspective discrimination can be described by three main concepts: (1) *what actions*, (2) *in which situations*, and (3) *towards whom*

are actions considered to be discriminatory. Actions are forms of discrimination, situations are areas of discrimination, and grounds of discrimination describe the characteristics of the people who may be discriminated against.

The EU legal framework for anti-discrimination and equal treatment is constituted by several directives, including the Race Equality Directive (2000/43/EC), the Employment Equality Directive (2007/78/EC), the Gender Recast Directive (2006/54/EC) and the Gender Goods and Services Directive (2006/113/EC) [62]. The main grounds for discrimination defined in European Council directives [17] (2000/43/EC, 2000/78/EC) are: race and ethnic origin, disability, age, religion or belief, sexual orientation, gender and nationality. There is no general directive stating which attributes can and cannot be used for which types of decision-making [62]. Multiple discrimination occurs when a person is discriminated on a combination of several grounds. The main areas of discrimination are: access to employment, access to education, employment and working conditions, social protection and access to supply of goods and services.

Discriminatory actions may take different forms, the two main being known as *direct discrimination* and *indirect discrimination*. Direct discrimination occurs when a person is treated less favorably than another person would be treated in a comparable situation on protected grounds. For example, property owners not renting to a racial minority tenant. Indirect discrimination occurs where an apparently neutral provision, criterion or practice would put persons of a protected ground at a particular disadvantage compared with other persons. For example, the requirement to produce ID in the form of a driver's license for entering a club may discriminate against visually impaired people, who cannot have a driver's license. A related term *statistical discrimination* [2] is often used in economic modeling. It refers to inequality between demographic groups occurring even when economic agents are rational and non-prejudiced.

Data-driven decision making refers to using predictive models learned on historical data for decision support. Data-driven decision making is prone to indirect discrimination, since data mining and machine learning algorithms produce decision rules or decision models, which then may put persons of some groups at a disadvantage as compared to other groups. When decisions are made by human judgement, biased decisions may occur on a case-by-case basis. Rules produced by algorithms are applied to every case, and hence may discriminate more systematically and on a larger scale than human decision makers. Discrimination due to algorithms is sometimes referred to as *digital discrimination* [57].

The current non-discrimination legislation has been set up to guard against discrimination by human decision makers. The basic principles of the non-discrimination legislation generally apply to algorithmic decision making as well, the specifics of algorithmic decision making are yet to be taken into national and international legislation. Ideally, algorithmic discrimination measures should be universal in a sense that they would not be tied to any specific legislation.

The current EU directives do not specify particular discrimination measures or tests to be used to judge whether there has been a discrimination. Rather, statistical measures of discrimination are used on case-by-case bases to establish *prima facie* evidence, which then shifts the responsibility of proving discrimination from the person who is being discriminated against to the discriminating party.

The general population, and even some data scientists may think that since data mining is based on data, models produced by data mining algorithms must be objective by nature. In reality models are as objective as the data on which they are built, and as long as the assumptions behind the models are perfectly matched in the data. In practice, assumptions are rarely perfectly matched. Historical data may be biased, incomplete, or record past discriminatory decisions that can easily be transferred to predictive models, and reinforced in new decision making [14]. Lately, awareness of policy makers and public attention to potential discrimination has been increasing [10, 21, 23, 33, 44], but there are many research questions which must be answered in order to fully understand in which circumstances algorithms do or do not become discriminatory, and how to prevent them being so by computational means.

### 2.2 Discrimination-aware data mining

Discrimination-aware data mining is a discipline at an intersection of computer science, law and the social sciences. It has two main research directions: *discrimination discovery*, and *discrimination prevention*. Discrimination discovery aims at finding discriminatory patterns in data using data mining methods. A data mining approach for discrimination discovery typically extracts association and classification rules from data, and then evaluates those rules in terms of potential discrimination [28,39,40,47,49,53,54]. A more traditional statistical approach to discrimination discovery typically fits a regression model to the data including the protected characteristics (such as race or gender), and then analyzes the magnitude and statistical significance of the regression slopes at the protected attributes (e.g. [22]). If those slopes appear to be significant, then discrimination is flagged. The majority of discrimination discovery approaches are based on finding correlations, whereas there is a growing body of research aimed at demonstrating causation [9, 60], which is necessary for legal actions. Exploratory discrimination-aware data mining [6] is an emerging direction that aims to discover insights about new or changing forms of or grounds for discrimination. Discrimination-aware data mining relates to privacy-aware data mining (e.g. [29, 52]) with a common understanding that securing privacy and non-discrimination come with a cost of information loss, and the objective is to minimize information loss while ensuring a desired level of privacy and fairness.

Discrimination prevention algorithms have been developed to produce non-discriminatory predictive models with respect to externally given sensitive characteristics. The objective is to build a model or a set of decision rules

that would obey non-discrimination constraints. Typically, such constraints directly relate to some selected discrimination measure. Algorithmic solutions for discrimination prevention fall into three categories: data preprocessing, model post-processing, and model regularization. Data preprocessing modifies historical data such that it no longer contains unexplained differences across the protected and the unprotected groups, and then uses standard learning algorithms with this modified data. Data preprocessing may modify the target variable [34,36,40], or modify input data [25,59], or both [28,29]. Model post-processing produces a standard model and then modifies this model to obey non-discrimination constraints, for instance, by changing the labels of some leaves in a decision tree [12, 35], or removing selected rules from the set of discovered decision rules [30]. Model regularization forces non-discrimination constraints during the model learning process, for instance, by modifying the splitting criteria in decision tree learning [11, 35, 37]. Since the focus of this survey is on measuring discrimination, algorithmic solutions will be only briefly overviewed. An interested reader can find further details, for instance, in this edited book [19], this journal issue [43], or proceedings of specialized workshops [3, 4, 13].

Defining coherent discrimination measures is fundamental for both lines of research: discrimination discovery and discrimination prevention. Discrimination discovery requires some measure that can be used to judge whether there is any discrimination in data. Discrimination prevention requires some measure for use as an optimization criterion in order to sanitize predictive models. Direct discrimination by algorithms can be avoided by excluding the sensitive variable from decision making, but this unfortunately does not prevent the risk of indirect discrimination. In order to aid in establishing a basis for further research in the field, especially in algorithmic discrimination prevention, our main focus in this survey is to review indirect discrimination measures. While measuring direct discrimination is based on comparing individual to individual, measuring indirect discrimination is based on comparing group characteristics.

### 2.3 Definition of fairness for data mining

In the context of data mining and machine learning non-discrimination can be defined as follows: **(1) people that are similar in terms of non-protected characteristics should receive similar predictions, and (2) differences in predictions across groups of people can only be as large as justified by their non-protected characteristics.** To the best of our knowledge, in the data mining context these two conditions, expressed as Lipschitz condition and statistical parity, have been first formally discussed in [20].

The first condition is necessary but not sufficient for ensuring non-discrimination in decision making, because even though similar people are treated in a similar way, groups of similar people may be treated differently from other groups. The first condition relates to direct discrimination, which occurs when a person is

treated less favorably than another would be treated in a comparable situation, and can be illustrated by the *twin test*. Suppose gender is the protected attribute, and there are two identical twins who share all the characteristics, but gender. The test is passed if both individuals receive identical predictions by the model.

The second condition ensures that there is no indirect discrimination, which occurs when apparently neutral provision, criteria or practice would put persons of a protected ground at a particular disadvantage compared with other persons. The so called *redlining* practice [32] exemplifies indirect discrimination. The term relates to past practices by banks to deny loans for residents of selected neighborhoods. Race was not formally used as a decision criterion, but it appeared that the excluded neighborhoods had much higher populations of non-white people than average. Thus, even though people of different races ("twins") from the same neighborhood were treated equally, the lowering of positive decision rates in the non-white-dominated neighborhoods affected the non-white population in a worse way. Therefore, different decision rates across groups of similar people can only be as large as explained by non-protected characteristics. The second part of the definition controls for balance across the groups.

More formally, let $X$ be a set of variables describing non-protected characteristics of a person (a complete set of characteristics may not always be known or available, in such a case $X$ denotes a set of available characteristics), $S$ be a set of variables describing the protected characteristics, and $\hat{y}$ be the model output. A predictive model can be considered fair if: (1) the expected value for model output does not depend on the protected characteristics $E(\hat{y}|X,S) = E(\hat{y}|X)$ for all $X$ and $S$, that is, there is no direct discrimination; and (2) if non-protected characteristics and protected characteristics are not independent, that is if $E(X|S) \neq E(X)$, then the expected value for model output within each group should be justified by some fairness model, that is $E(\hat{y}|X) = F(\hat{y}|X)$, where $F$ is a fairness model. Defining and justifying $F$ is not trivial, that is where a lot of ongoing effort in discrimination-aware data mining currently is concentrated.

Discrimination by predictive models can occur only when the target variable is polar, that is, some predictions outcomes are considered superior to others. For example, getting a loan is better than not getting a loan, or the "golden client" package is better than the "silver", and "silver" is better than "bronze", or an assigned interest rate of 3% is better than 5%. If the target variable is not polar, there is no discrimination, because no treatment is superior or inferior to another treatment.

The protected characteristic (also referred to as the protected variable or sensitive attribute) may be binary, categorical or numeric, and it does not need to be polar. For example, gender can be encoded with a binary protected variable, ethnicity can be encoded with a categorical variable, and age can be encoded with a numerical variable. In principle, any combination of one or more personal characteristics may be required to be protected. Discrimination on more than one ground is known as *multiple discrimination*, and it may

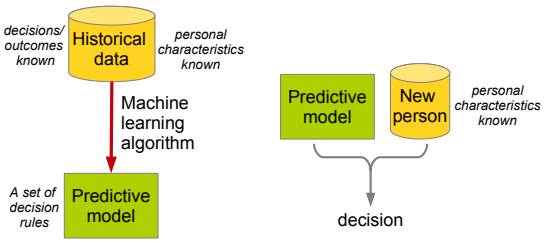



**Fig. 1** A typical machine learning setting.

be required to ensure the prevention of multiple discrimination in predictive models. Thus, ideally, algorithmic discrimination measures should be able to handle any type or a combination of protected variables. Finding out which characteristics are to be protected is outside the jurisdiction of data mining, the protected characteristics are to be given externally.

### 2.4 Principles for making predictive models non-discriminatory

Figure 1 depicts a typical machine learning process. A machine learning algorithm is a procedure used for producing a predictive model from historical data. A model is a collection of decision rules used for decision making for new incoming data. The model would take personal characteristics as inputs (for example, income, credit history, employment status), and produce a prediction (for example, credit risk level).

Learning algorithms as such cannot discriminate, because they are not used for decision making. The resulting predictive models (decision rules) would discriminate. Yet algorithms may be discrimination-aware by employing procedures to enforce non-discrimination constraints into the models. Hence, one of the main goals of discrimination-aware data mining is to develop discrimination-aware algorithms, that would guarantee that non-discriminatory models are produced.

There is a debate in the discrimination-aware data mining community about whether models should or should not use protected characteristics as inputs. For example, a credit risk assessment model may use gender as input, or may leave the gender variable out. Our position [62] is that protected characteristics, such as race, are necessary in the model building process in order to actively make sure that the resulting model is non-discriminatory. Of course, later when the model is used for decision making, it should not require protected characteristics as inputs. A data-driven decision model that does not use protected characteristics as inputs in principle cannot produce direct discrimination. By the first fairness condition, it would treat two persons that differ only in protected characteristics in the same way.

Ensuring that there is no indirect discrimination (the second fairness condition) is more tricky. In order to verify to what extent non-discrimination constraints are obeyed and enforce fair allocation of predictions across groups

**Table 1** Discrimination measure types

| Measures | Indicate what? | Type of discrimination |
|---|---|---|
| Statistical tests | presence/absence of discrimination | indirect |
| Absolute measures | magnitude of discrimination | indirect |
| Conditional measures | magnitude of discrimination | indirect |
| Situation measures | spread of discrimination | direct or indirect |

of people, learning algorithms must have access to the protected characteristics in the historical data. We argue that if protected information (e.g. gender or race) is not available during the model learning building process, the learning algorithm cannot be discrimination-aware, because it cannot actively control non-discrimination. The resulting models produced without access to sensitive information may be discriminatory, they may be not, but that is a chance rather than discrimination-awareness property of the algorithm.

Non-discrimination can potentially be measured in input data, on predictions made by models, or in models themselves. Measuring requires access to the protected characteristic. Yet this does not mean that algorithmic discrimination is always direct. The distinction between direct and indirect discrimination refers to using the protected characteristic in decision making, not to measuring discrimination. The following section presents a categorized survey of measures used in discrimination-aware data mining and the machine learning literature, and discusses other existing measures that could in principle be used for measuring the fairness of algorithms.

## 3 Discrimination measures

Discrimination measures can be categorized into (1) statistical tests, (2) absolute measures, (3) conditional measures, and (4) situation measures. We survey measures in this order due to historical reasons, which is more or less how they came into use. All four types are not alternative types to measure the same, but rather they measure different aspects of the problem, as summarized in Table 1.

Statistical tests indicate the presence or absence of discrimination at a dataset level, they do not measure the magnitude of discrimination, neither the spread of discrimination within a dataset. Absolute measures capture the magnitude of discrimination over a dataset (or a subset of interest) taking into account the protected characteristic, and the prediction decision; no other characteristics of individuals are considered. It is assumed that all individuals are alike, and there should be no differences in decision probability for people in the protected and in the general group, regardless of possible explanation. Absolute measures generally are not used alone in a dataset, but rather provide core principles for conditional measures, or statistical tests. Conditional measures capture the magnitude of discrimination, which cannot be explained by any non-protected characteristics of individuals. Statistical tests, absolute measures and conditional measures are designed for capturing indirect dis-

**Table 2** Summary of notation.

| Symbol | Explanation |
|---|---|
| $y$ | target variable, $y_i$ denotes the $i^{th}$ observation |
| $y^i$ | value of a binary target variable, $y \in \{y^+, y^-\}$ |
| $s$ | protected variable |
| $s^i$ | value of a categorical/binary protected variable, $s \in \{s^1, \ldots, s^m\}$ index 1 denotes the protected group, e.g. $s^1$ - ethnic minority, $s^0$ - majority |
| $X$ | set of input variables (predictors), $X = \{x^{(1)}, \ldots, x^{(l)}\}$ |
| $z$ | explanatory variable or stratum |
| $z^i$ | value of explanatory variable $z \in \{z^1, \ldots, z^k\}$ |
| $N$ | number of individuals in the dataset |
| $n_i$ | number of individuals in group $s^i$ |

crimination. Situation measures have been introduced mainly to accompany mining classification rules for the purpose of discovering direct discrimination. Situation measures do not measure the magnitude of discrimination, but the spread of discrimination, that is, the share of people in the dataset that are affected by direct discrimination.

The following notation summarized in Table 2, will be used throughout the survey. We will use the following short probability notation: $p(s = 1)$ will be encoded as $p(s^1)$, and $p(y = +)$ will be encoded as $p(y^+)$. Let $s^1$ denote the protected community, and $y^+$ denote the desired decision (e.g. positive decision to give a loan).

### 3.1 Statistical tests

Statistical tests are the earliest measures that are focused on indirect discrimination discovery in data. Statistical tests are formal procedures to accept or reject statistical hypotheses, which check how likely the result is to have occurred by chance. Typically, in discrimination analysis the null hypothesis is that there is no difference between the treatment of the general group and the protected group. The test checks how likely the observed difference between groups could have occurred by chance. If chance is unlikely then the null hypothesis is rejected and discrimination is declared.

Two limitations of statistical tests need to be kept in mind when using them for measuring discrimination.

1. Statistical significance does not mean practical significance; statistical tests do not show the magnitude of the difference between groups, which can be large or minor.
2. If the null hypothesis is rejected then discrimination is present, but if the null hypothesis cannot be rejected, that does not prove that there is no discrimination. It may be that the data sample is too small to declare discrimination.

Standard statistical tests are typically applied for measuring discrimination, such as Student's t-test, or the Chi-square test. The same tests are used

in clinical trials, marketing, and scientific research. Statistical tests are suitable for indirect discrimination discovery in data, but they do not necessarily directly translate into optimization constraints to be used in model learning to ensure discrimination prevention. Yet, statistical methods include methods for determining the effect size, which can in principle be translated to algorithmic discrimination measures and optimization constraints. The absolute measures, discussed in the next section (such as the mean difference), often derived from the statistical approaches for computing test statics.

*3.1.1 Regression slope test*

The test fits Ordinary Least Squares (OLS) regression to the data including the protected variable, and tests whether the regression coefficient of the protected variable is significantly different from zero. A basic version for discrimination discovery considers only the protected characteristic $s$ and the target variable $y$ [58]. Typically, in discrimination testing $s$ is binary, but in principle $s$ and $y$ can also be numeric. The regression may include only the protected variable $s$ as a predictor, but it may also include variables from $X$ that may explain some of the observed differences in decisions.

The test statistic is $t = b/\sigma$, where $b$ is the estimated regression coefficient of $s$, and $\sigma$ is the standard error, computed as

$$\sigma = \frac{\sqrt{\sum_{i=1}^{n}(y_i - f(y_i))^2}}{\sqrt{(n-2)}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}},$$

where $n$ is the number of observations, $f(.)$ is the regression model, $\bar{.}$ indicates the mean. The t-test with $n-2$ degrees of freedom is applied.

*3.1.2 Difference of means test*

The null hypothesis is that the means of the two groups are equal. The test statistic is

$$t = \frac{E(y|s^0) - E(y|s^1)}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}},$$

where $n_0$ is the number of individuals in the unprotected group, $n_1$ is the number of individuals in the protected group,

$$\sigma = \sqrt{\frac{(n_0-1)\delta_0^2 + (n_1-1)\delta_1^2}{n_0 + n_1 - 2}},$$

where $\delta_0^2$ and $\delta_1^2$ are the sample target variances in the respective groups. The t-test with $n_0 - n_1 - 2$ degrees of freedom is applied. The test assumes independent samples, normality and equal variances. Difference of means, although not formally used as a statistical test, has been used in the data mining literature, for instance in [11].

*3.1.3 Difference in proportions for two groups*

The null hypothesis is that the rates of positive outcomes within the two groups are equal. The test statistic is

$$z = \frac{p(y^+|s^0) - p(y^+|s^1)}{\sigma},$$

where

$$\sigma = \sqrt{\frac{p(y^+|s^0)p(y^-|s^0)}{n_0} + \frac{p(y^+|s^1)p(y^-|s^1)}{n_1}}.$$

The z-test is applied. Difference in proportions, although not formally used as a statistical test, has been used in a number of data mining studies [12, 34, 35, 46, 59].

*3.1.4 Difference in proportions for many groups*

The null hypothesis is that the probabilities or proportions are equal for all the groups. This can be used for testing many groups at once. For example, equality of decisions for different ethnic groups, or age groups. If the null hypothesis is rejected that means at least one of the groups has statistically significantly different proportion. The text statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np(y^+|s^i))^2}{p(y^+|s^i)},$$

where $k$ is the number of groups. The Chi-Square test is used with $k-1$ degrees of freedom.

*3.1.5 Rank test*

The Mann-Whitney U test [41] is applied for comparing two groups when the normality and equal variances assumptions are not satisfied. The null hypothesis is that the distributions of the two populations are identical. The procedure is to rank all the observations from the largest $y$ to the smallest. The test statistic is the sum of ranks of the protected group. For large samples the normal approximation can be used and then the z-test can be applied. A ranking approach for measuring discrimination, although without a formal statistical test, has been used in the data mining literature, for instance in [11].

## 3.2 Absolute measures

Absolute measures are designed to capture the magnitude of differences between (typically two) groups of people. The groups are determined by the protected characteristic (e.g. one group is males, another group is females). If more than one protected group is analyzed (e.g. different nationalities), typically each group is compared separately to the most favored group.

*3.2.1 Mean difference*

The mean difference measures the difference between the means of the targets of the protected group and the general group,

$$d = E(y^+|s^0) - E(y^+|s^1).$$

If there is not difference then it is considered that there is no discrimination. The measure relates to the difference of means, and difference in proportions test statistics, except that there is no correction for the standard deviation.

The mean difference for binary classification with a binary protected variable,

$$d = p(y^+|s^0) - p(y^+|s^1),$$

is also known as the discrimination score [12], or *sliftd* [46].

The mean difference has been the most popular measure in early work on discrimination-aware data mining and machine learning [11, 12, 34, 35, 46, 59].

*3.2.2 Normalized difference*

The normalized difference [61] is the mean difference for binary classification normalized by the rate of positive outcomes,

$$\delta = \frac{p(y^+|s^0) - p(y^+|s^1)}{d_{max}},$$

where

$$d_{max} = \min\left(\frac{p(y^+)}{p(s^0)}, \frac{p(y^-)}{p(s^1)}\right).$$

This measure takes into account maximum possible discrimination at a given positive outcome rate, such that with the maximum possible discrimination $\delta = 1$, and $\delta = 0$ indicates no discrimination.

*3.2.3 Area under curve (AUC)*

This measure is related to rank tests. It has been used in [11] for measuring discrimination between two groups when the target variable is numeric (regression task),

$$AUC = \frac{\sum_{(s^i,y^i)\in D^0} \sum_{(s^j,y^j)\in D^1} \mathbf{I}(y_i > y_j)}{n_0 n_1},$$

where $\mathbf{I}(true) = 1$ and 0 otherwise.

For large datasets computation of AUC is time and memory intensive, since a quadratic number of comparisons to the number of observations is required. The authors did not mention it, but there is an alternative way to compute based on ranking, which, depending on the speed ranking algorithm, may be

faster. Assign numeric ranks to all the observations, beginning with 1 for the smallest value. Let $R_0$ be the sum of the ranks for the favored group. Then

$$AUC = R_0 - \frac{n_0(n_0+1)}{2}.$$

We observe that if the target variable is binary, and in case of equality half of a point is added to the sum, then AUC linearly relates to mean difference as

$$\begin{aligned} AUC &= p(y^+|s^0)p(y^-|s^1) + 0.5p(y^+|s^0)p(y^+|s^1) + 0.5p(y^-|s^0)p(y^-|s^0) \\ &= 0.5d + 0.5, \end{aligned}$$

where $d$ denotes discrimination measured by the mean difference measure.

*3.2.4 Impact ratio*

The impact ratio, also known as slift [46], is the ratio of positive outcomes for the protected group over the general group,

$$r = p(y^+|s^1)/p(y^+|s^0).$$

The inverse $1/r$ has been referred to as the likelihood ratio [25]. This measure is used in the US courts for quantifying discrimination, the decisions are deemed to be discriminatory if the ratio of positive outcomes for the protected group is below 80% of that of the general group. Also this is the form stated in the Sex Discrimination Act of U.K. $r = 1$ indicates that there is no discrimination.

*3.2.5 Elift ratio*

The elift ratio [45] is similar to the impact ratio, but instead of dividing by the general group, the denominator is the overall rate of positive outcomes

$$r = p(y^+|s^0)/p(y^+).$$

In principle the same measure is expressed as

$$\frac{p(y,s)}{p(y)p(s)} \leq 1 + \eta,$$

where the requirement should be satisfied for all values of $y$ and $s$, is referred to as $\eta$-neutrality [26].

*3.2.6 Odds ratio*

The odds ratio of two proportions is often used in natural, social and biomedical sciences to measure the association between exposure and outcome. The measure has a convenient relation with the logistic regression. The exponential function of the logistic regression coefficient translates one unit increase in the odds ratio. Odds ratio has been used for measuring discrimination [46] as

$$r = \frac{p(y^+|s^0)p(y^-|s^1)}{p(y^+|s^1)p(y^-|s^0)}.$$

*3.2.7 Mutual information*

Mutual information (MI) is popular in information theory for measuring mutual dependence between variables. In the discrimination literature this measure has been referred to as the normalized prejudice index [26], and used for measuring the magnitude of discrimination. Mutual information is measured in bits, but it can be normalized such that the result falls into the range between 0 and 1. For categorical variables

$$MI = \frac{I(y,s)}{\sqrt{H(y), H(s)}},$$

where

$$I(s,y) = \sum_{(s,y)} p(s,y) \log \frac{p(s,y)}{p(s)p(y)},$$
$$H(y) = -\sum_{y} p(y) \log p(y).$$

For numerical variables the summation is replaced by an integral.

*3.2.8 Balanced residuals*

While the previous approaches measure discrimination in model outputs, no matter what the actual accuracy of the predictions is, balanced residuals measure builds on the accuracy of predictions. This measure characterizes the difference between the actual outcomes recorded in the dataset, and the model outputs. The requirement is that under-predictions and over-predictions should be balanced within the protected and unprotected groups. Calders et al [11] proposed balanced residuals as a criteria of non-discrimination, originally it was not intended as a measure. That is, the average residuals are required to be equal for the protected group and the unprotected group. In principle this approach could be used as a measure of discrimination

$$d = \frac{\sum_{i \in D^1} y_i - \hat{y}_i}{n_1} - \frac{\sum_{j \in D^0} y_j - \hat{y}_j}{n_0},$$

where $y$ is the true target value, $\hat{y}$ is the prediction. Positive values of $d$ would indicate discrimination towards the protected group.

One should, however, use and interpret this measure with caution. If the learning dataset is discriminatory, but the predictive model makes ideal predictions such that all the residuals are zero, this measure would show no discrimination, even though the predictions would be discriminatory, since the original data is discriminatory. Suppose, another predictive model makes a constant prediction for everybody, and the constant prediction is equal to the mean of the unprotected group. If the training dataset contains discrimination, then the residuals for the unprotected group would be smaller than for

the protected group, and the measure would indicate discrimination, however, a constant prediction for everybody means that everybody is treated equally, and there should be no discrimination detected.

Another measure related to the prediction errors, called the Balanced Error Rate (BER) was introduced in [25]. The approach is to measure the average error rate of predicting the sensitive variable $s$ from the other input variables $X$. In our interpretation this is not a measure of discrimination, but a measure of the potential for redlining, that is, how much information about the sensitive characteristic (e.g. race) is carried by the legitimate input variables (e.g. zip code, occupation or employment status).

A recent study by Hardt et al [31] introduces two accuracy related non-discrimination criteria: equalized odds and equal opportunity. These are not measures, but alternative fairness definitions, although they may be turned into measures by taking a difference or a ratio of the equation components. Equalized odds require the prediction conditioned on the true outcome to be the same for any group of people (with respect to the sensitive characteristic): $p(\hat{y}|s = 1, y) = p(\hat{y}|s = 0, y)$ for any $y$. Equal opportunity is a weaker version of equal odds. Equal opportunity requires the predictions only within the subset of positive true outcomes: $p(\hat{y}|s = 1, y = 1) = p(\hat{y}|s = 0, y = 1)$.

A forthcoming study by Kleinberg et al [38] specifies three fairness conditions for binary classification with the binary protected characteristic variable, which are closely related to equalized odds. In summary the conditions require the distribution of the prediction scores to be the same for all groups of people within the positive true label and within the negative true label data.

We argue that incorporating the true label into a fairness criteria implicitly assumes that the true labels are objective, that is, that historical data contains no discrimination. This assumption is realistic for datasets with objective labels, such as, for instance, credit scoring, where the label denotes whether the person has actually repaid the loan or not. But the assumption may be overoptimistic for datasets that record human decisions in the past. For example, if a dataset records who has been hired for a job based on candidate CVs, hiring decisions in the past may not necessarily have been objective. In such cases fairness criteria that depend on the true labels in the dataset should be considered with caution.

#### *3.2.9 Relation between two variables*

There are many established measures in the feature selection literature [27] for measuring the relation between two variables, which, in principle, can be used as absolute discrimination measures. The stronger the relation between the protected variable $s$ and the target variable $y$, the larger the absolute discrimination.

There are three main groups of measures for the relation between variables: correlation based, information theoretic, and one-class classifiers. Correlation based measures, such as the Person correlation coefficient, are typically used

for numeric variables. Information theoretic measures, such as mutual information mentioned earlier, are typically used for categorical variables. One-class classifiers present an interesting option. In discrimination the setting would be to predict the target $y$ solely on the protected variable $s$, and measure the prediction accuracy. We are not aware of such attempts in the discrimination-aware data mining literature, but it would be a valid option to explore.

#### *3.2.10 Measuring for more than two groups*

Most of the absolute discrimination measures are for two groups (protected group vs. unprotected group). Ideas, how to apply those for more than two groups, can be borrowed from multi-class classification [7], the multi-label classification [56], and one-class classification [55] literature. Basically, there are three options for obtaining sub-measures: measure pairwise for each pair of groups ($k(k-1)/2$ comparisons), measure one against the rest for each group ($k$ comparisons), measure each group against the unprotected group ($k-1$ comparisons). The remaining question is how to aggregate the sub-measures. Based on personal conversations with legal experts, we advocate for reporting the maximum from all the comparisons as the final discrimination score. Alternatively, all the scores could be summed weighing them by the group sizes to obtain an overall discrimination score.

Even though absolute measures do not take into account any explanations of possible differences of decisions across groups, they can be considered as core building blocks for developing conditional measures. Conditional measures do take into account explanations for differences, and measure only discrimination that cannot be explained by non-protected characteristics.

Table 3 summarizes the applicability of absolute measures in different machine learning settings. Straightforward extensions would be as follows. To apply the measures to categorical variables one would measure each group against the rest in a binary way and then average over the resulting measures. To extend balanced residuals to a binary target one would need to use raw probability scores of class label given the data, which can be produced by the most classifiers.

### 3.3 Conditional measures

Absolute measures take into account only the target variable $y$ and the protected variable $s$. Absolute measures consider all the differences in treatment between the protected group and the unprotected group to be discriminatory. The conditional measure, on the other hand, tries to capture how much of the difference between the groups is explainable by other characteristics of individuals, recorded in $X$, and only the remaining differences are deemed to be discriminatory. For example, part of the difference in acceptance rates for natives and immigrants may be explained by differences in education levels. Only the remaining unexplained difference should be considered as discrimination.

**Table 3** Summary of absolute measures. The checkmark (✓) indicates that it is directly applicable in a given machine learning setting. The tilde (∼) indicates that a straightforward extension exists.

| | Protected variable | | | Target variable | | |
|---|---|---|---|---|---|---|
| Measure | Binary | Categorical | Numeric | Binary | Ordinal | Numeric |
| Mean difference | ✓ | ∼ | | ✓ | | ✓ |
| Normalized difference | ✓ | ∼ | | ✓ | | |
| Area under curve | ✓ | ∼ | | ✓ | ✓ | ✓ |
| Impact ratio | ✓ | ∼ | | ✓ | | |
| Elift ratio | ✓ | ∼ | | ✓ | | |
| Odds ratio | ✓ | ∼ | | ✓ | | |
| Mutual information | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Balanced residuals | ✓ | ∼ | | ∼ | ✓ | ✓ |
| Correlation | ✓ | | ✓ | ✓ | | ✓ |

Let $z = f(X)$ be an explanatory variable. For example, if $z^i$ denotes a certain education level. Then all the individuals with the same level of education will form a strata $i$. Within each strata the acceptance rates are required to be equal.

### *3.3.1 Unexplained difference*

Unexplained difference [36] is measured, as the name suggests, as the overall mean difference minus the differences that can be explained by another legitimate variable. Recall that the mean difference is

$$d = p(y^+|s^0) - p(y^+|s^1).$$

Then the unexplained difference is

$$d_u = d - d_e,$$

where

$$d_e = \sum_{i=1}^{m} p^\star(y^+|z^i)\big(p(z^i|s^0) - p(z^i|s^1)\big),$$

where $p^\star(y^+|z^i)$ is the desired acceptance rate within strata $i$. The authors recommend using

$$p^\star(y^+|z^i) = \frac{p(y^+|s^0, z^i) + p(y^+|s^1, z^i)}{2}.$$

In the simplest case $z$ may be equal to one of the variables in $X$. The authors also use clustering on $X$ to take into account more than one explanatory variable at the same time. Then $z$ denotes a cluster, one strata is one cluster.

The related Cochran-Mantel-Haenszel test [16, 42] is a formal statistical counterpart for hypothesis testing.

*3.3.2 Propensity measure*

Propensity models [51] are typically used in clinical trials or marketing for estimating the probability that an individual would receive treatment. Given the estimated probabilities, individuals can be stratified according to similar probabilities of receiving treatment, and the effects of treatment can be measured within each strata separately. Propensity models have been used for measuring discrimination [11], in this case a function was learned to model the protected characteristic based on input variables $X$, that is $s^1 = f(X)$. A logistic regression was used for modeling $f(.)$. Then the estimated propensity scores $\hat{s}^1$ were split into five ranges, where each range formed one strata. Discrimination was measured within each strata, treating each strata as a separate dataset, and using the absolute discrimination measures discussed in the previous section. The authors did not aggregate the resulting discrimination into one measure, but in principle the results can be aggregated into one measure, for instance, using the unexplained difference formulas, reported above. In such a case each strata would correspond to one value of an explanatory variable $z$.

*3.3.3 Belift ratio*

The belift ratio [40] is similar to the elift ratio in absolute measures, but here the probabilities of positive outcome are also conditioned on input attributes,

$$belift = \frac{p(y^+|s^1, X^r, X^a)}{p(y^+|X^a)},$$

where $X = X^r \cup X^{\not r}$ is a set of input variables, $X^r$ denotes so called redlining attributes, the variables which are correlated with the protected variable $s$. The authors proposed estimating the probabilities via Bayesian networks. A possible difficulty for applying this measure in practice may be that not everybody, especially non-machine learning users, are familiar enough with Bayesian networks to the extent needed for estimating the probabilities. Moreover, construction of a Bayesian network may be different even for the same problem depending on the assumptions made about interactions between the variables. Thus, different users may get different discrimination scores for the same application case.

A simplified approximation of belift could be to treat all the attributes as redlining attributes, and instead of conditioning on all the input variables, condition on a summary of input variables $z$, where $z = f(X)$. Then the measure for strata $i$ would be

$$\frac{p(y^+|s^1, z^i)}{p(y^+)}.$$

The measure has a limitation that neither the original version, nor the simplified version allow differences to be explained by variables that are correlated with the protected variable. That is, if a university has two programs,

say medicine and computer science, and the protected group, e.g. females, are more likely to apply for a more competitive program, then the programs cannot have different acceptance rates. That is, if the acceptance rates are different, all the differences are considered to be discriminatory.

## 3.4 Situation measures

Situation measures are targeted at quantifying direct discrimination [50] The main idea behind situation measures is for each individual in the dataset to identify whether s/he is discriminated against and then analyze how many individuals in the dataset are affected.

### *3.4.1 Situation testing*

Situation testing [39] measures which fraction of individuals in the protected group are considered to be discriminated against as

$$f = \frac{\sum_{u_i \in D(s^1)} \mathbf{I}(\mathit{diff}(u_i) \geq t)}{|D(s^1)|},$$

where $D(s^1)$ is the subset of data containing all the individuals in the protected group, $u_i$ denotes an individual, $t$ is a user defined threshold of maximum tolerable difference, $\mathbf{I}$ is the indicator function that takes 1 if true, 0 otherwise. The situation testing for an individual $i$ is computed as

$$\mathit{diff}(u_i) = \frac{\sum_{u_j \in D(s^0, \kappa | u_i)} y_j}{\kappa} - \frac{\sum_{u_j \in D(s^1, \kappa | u_i)} y_j}{\kappa},$$

where $D(s^0, \kappa | u_i)$ is a subset of data containing the nearest neighbors of $u_i$ belonging to the protected group indicated by $s^0$, $\kappa$ is the user defined parameter indicating the number of neighbors, $y_j$ is the decision outcome for the individual $u_j$. Positive and negative discrimination is handled separately.

The idea is to compare each individual to the opposite group and see if the decision would be different. In that sense, the measure relates to propensity scoring (Section 3.3), used for identifying groups of people who are similar according to the non-protected characteristics, and requiring for decisions within those groups to be balanced. The main difference is that propensity measures would signal indirect discrimination within a group, and situation testing aims at signaling direct discrimination for each individual in question.

### *3.4.2 Consistency*

The consistency measure [59] compares the predictions for each individual with his/her nearest neighbors.

$$C = 1 - \frac{1}{\kappa N} \sum_{i=1}^{N} \sum_{y_j \in D(\kappa | u_i)} |y_i - y_j|,$$

where $D(\kappa|u_i)$ is the subset of data containing $\kappa$ nearest neighbors of $u_i$, $y_i$ is the decision outcome for the individual $u_i$.

The consistency measure is closely related to situation testing, but considers nearest neighbors from any group (not from the opposite group). Due to this choice, the consistency measure should be used with caution in situations where there is a high correlation between the protected variable and the legitimate input variables. For example, suppose we have only one predictor variable - location of an apartment, and the target variable is to grant a loan or not. Suppose all non-white people live in one neighborhood (as in the redlining example), and all the white people in another neighborhood. Unless the number of nearest neighbors to consider is very large, this measure will show no discrimination, since all the neighbors will get the same decision, even though all non-white residents will be rejected, and all white will be accepted. That would show a perfect consistency, in spite of the fact that discrimination is at its maximum. In their experimental evaluation the authors have used this measure in combination with the mean difference measure.

## 4 Experimental analysis of core measures

In this section we computationally analyze a set of absolute measures, and discuss their properties to provide a better understanding of implications of choosing one measure over another. Absolute measures are naive in the sense that they do not take possible explanations of different treatments into account, and due to that may show more discrimination that there actually is, these measures provide core mechanisms and a basis for measuring indirect discrimination. Conditional measures are typically built upon absolute measures, and statistical tests are often directly related to absolute measures.

We analyze the following measures, introduced in Section 3.2: mean difference, normalized difference, mutual information, impact ratio, elift and odds ratio. From these measures the mean difference and area under a curve can be directly used in regression tasks. Our main emphasis is on binary classification with a binary sensitive variable, since this scenario has been studied more extensively in the discrimination-aware data mining and machine learning literature, and there are more measures available for classification than for regression; the regression setting, except for a recent work [11], remains a subject for future research, and therefore is beyond the scope of this survey paper.

An important question to consider is to what extent we can control the ground truth of how much discrimination is in the data, even when we generate data synthetically. We argue that when considering absolute measures the ground truth of no discrimination is one and always the same – equal treatment for the groups no matter what possible justifications of the differences between the groups may be. If some differences are present, then different absolute measures may indicate different amounts of discrimination, that is a matter of convention. A simple analogy may be to describing the intensity of rain. It is

**Table 4** Limiting values of the selected measures.

| Measure | Maximum discrimination | No discrimination | Reverse discrimination |
|---|---|---|---|
| Differences | | | |
| Mean difference | 1 | 0 | $-1$ |
| Normalized difference | 1 | 0 | $-1$ |
| Mutual information | 1 | 0 | 1 |
| Ratios | | | |
| Impact ratio | 0 | 1 | $+\infty$ |
| Elift | 0 | 1 | $+\infty$ |
| Odds ratio | 0 | 1 | $+\infty$ |
| AUC | | | |
| Area under curve (AUC) | 1 | 0.5 | 0 |

clear if there is no rain the measures should agree that there is no rain, but if there is some rain, different measures may give different rain scores relative to different baselines. In our experimental analysis we consider the ground truth extent of discrimination to scale linearly between no discrimination and the maximum possible discrimination.

The experimental analysis is meant to support two main messages: when interpreting the absolute amount of discrimination (1) one needs to keep in mind the distinction between symmetric measures (differences) and asymmetric measures (ratios), and (2) one needs to keep in mind that some measures are sensitive to the rate of positive outputs and classifiers that output a different number of positive decisions may not be directly comparable to each other under certain measures. The following experimental analysis is aimed at providing analytical insights into why this happens.

### 4.1 Symmetry and boundary conditions

First we consider the boundary conditions of the selected measures, as summarized in Table 4. In the difference based measures *zero* indicates an absence of discrimination, in the ratio based measures *one* indicates an absence of discrimination, in AUC 0.5 indicates an absence of discrimination. The boundary conditions are reached when one group gets all the positive decisions (e.g. the unprotected group), and the other group (e.g. the protected group) gets all the negative decisions.

The selected measures fall into two categories: symmetric and asymmetric. In Table 4 Differences and AUC represent symmetric measures, and Ratios represent asymmetric measures. With the symmetric measures discrimination and reverse discrimination are measured in the same units. For example, in the case of the mean difference, where 0 denotes no discrimination, 0.2 and $-0.2$ would indicate the same amount of discrimination, but towards different groups of people. In contrast, the cases of the impact ratios 1.1 and 0.9 would indicate different amount of discrimination towards different groups,

even though both values appear to be at the same distance from no discrimination (1.0 denotes the absence of discrimination).

### 4.2 Performance of Difference measures

Next we experimentally analyze the performance of the selected measures. We leave out AUC from the experiments, since in the classification it is equivalent to the mean difference measure. The goal of the experiments is to demonstrate how the performance depends on variations in the overall rate of positive decisions, balance between classes and balance between the unprotected and protected groups of people in the data. The key point of this experiments is to demonstrate that we can only compare different classifiers with respect to discrimination if they are outputting the same rate of positive decisions, or otherwise we have to normalize the measure with respect to the rate of positive decisions, as proposed in [61].

For this analysis we need to generate data where we know and can control the level of discrimination. We argue that the following reasoning represents the ground truth for the purpose. A typical classification procedure performed by humans can be though of consisting of a ranking mechanism that ranks the candidates from presumably the best to presumably the worst, and a decision threshold deciding how many of the best candidates are accepted, or how good the candidates need to be in order to be accepted. In data mining and machine learning for decision support a machine is doing the ranking, whereas the threshold is supposed to be given externally by humans depending on available resources (e.g. how many places are available for university admission or how much money is available to be given as credits). Therefore, we argue, that a data mined or machine learned model is discriminatory if the rankings that it is providing are discriminatory.

As a toy example, suppose that a ship is sinking, passengers are first put in a queue for who should be saved and then starting from the first passenger in the queue as many passengers are saved as there are boats. The queue can be formed by a machine learned model. The number of boats is external and does not depend on the model. Thus, even if we see only which passengers were saved and which not, our discrimination measure should be able to reconstruct and capture the process of putting the passengers into the queue. We will experimentally demonstrate to what extent it is possible with the current measures.

Suppose that the goal is to measure discrimination against males in the queueing in the sinking ship. Clearly, if males and females are put into the queue at random, then there is no discrimination with respect to gender. On the other hand, maximum possible discrimination occurs when all the females are before all the males in the queue. For intermediate values of discrimination we adopt the concept from situation measures, that is, if for instance 50% of the individuals are discriminated against, and 50% are not discriminated against, then the discrimination measure should indicate 50%. In the sinking

ship example 50% can be achieved by splitting all the passengers randomly into two equal groups, the first group is ordered into a queue at random with respect to gender, and the second group is ordered in the fully discriminatory way – all the females first and then all the males. Then the final queue is formed by randomly merging those two queues while keeping the original order of people from the small groups. Thus, the final queuing reflect 50% random order and 50% discriminatory order. We generate our synthetic data following this scheme in order to know the ground truth, and then analyze how much of that information we can recover by only knowing classification outcomes – who was saved and who was not, but not knowing the actual queue.

The data generation takes four parameters: the proportion of individuals in the protected group $p(s^1)$, the proportion of positive outputs $p(y^+)$, the underlying discrimination $d \in [-100\%, 100\%]$, and the number of data points $n$. The data is generated as follows. First $n$ data points are generated assigning a score in $[0, 1]$ uniformly at random, and assigning group membership at random according to the probability $p(s^1)$. This data contains no discrimination, because the scores are assigned at random. For a given level of desired discrimination $d$ we select $dn$ observations at random, sort them according to their scores, and then permute group assignments within this subsample in such a way that the highest scores get assigned to the unprotected group, and the lowest scores get assigned to the protected group. Finally, to translate the scores to classification decisions, we round the scores to 0 or 1 in such a way that the proportion of ones is as desired by $p(y^+)$. For each parameter setting we generate $n = 10000$ data points, and average the results over 100 such runs[1].

Figure 2 shows the performance of mean difference, normalized difference and mutual information on the datasets generated for different ground truth levels of discrimination following the described scheme. Ideally, the measures should vary with variation in the balance of the groups ($p(s^10)$) and the proportion of positive outputs ($p(y^+)$), that is, run along the diagonal line in the plots.

From the plots we can see that the normalized difference captures this, as expected. It is not surprising, since the normalization factors have been specifically designed [61] to correct the biases of the classical mean difference. We can see that the classical mean difference captures the trends, but the indicated discrimination highly depends on the balance of the classes and balance of the groups, therefore, this measure should be interpreted with care when the data is highly imbalanced. The same holds for mutual information. For instance, at $p(s^1) = 90\%$ and $p(y^+) = 90\%$ the true discrimination in the data may be near 100%, i.e. nearly the worst possible, but both measures would indicate that discrimination is nearly zero.

In addition, we see that the mean difference and normalized difference are linear measures, while mutual information is non-linear, and would underesti-

[1] The code for our experiments is available at `https://github.com/zliobaite/paper-fairml-survey`.

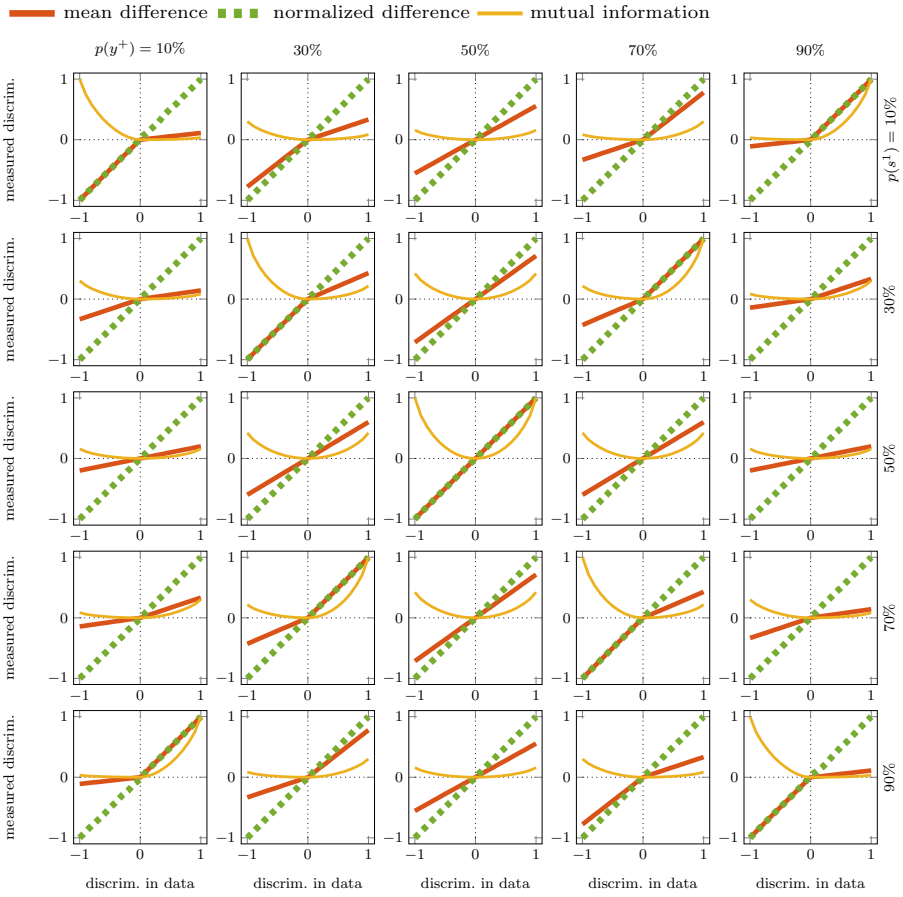



**Fig. 2** Analysis of the measures based on differences: discrimination in data vs. measured discrimination.

mate discrimination in the medium ranges. Moreover, mutual information does not indicate the sign of discrimination, that is, the outcome does not indicate whether discrimination is reversed or not. For these reasons, we do not recommend using mutual information for the purpose of quantifying discrimination. We advocate the normalized difference, which was designed to correct for biases due to imbalances in data. The normalized difference is somewhat more complex to compute than the mean difference, which may be a limitation for practical applications outside research. Therefore, if data is nearly balanced in terms of groups and positive-negative outcomes, then the classical mean difference will suffice.

### 4.3 Performance of Ratios

Figure 3 presents similar analysis of the measures based on ratios: impact ratio, elift and odds ratio. We do not expect these measures to follow the diagonal line, because the scaling of ratios is different to translate to the fraction of the population being discriminated. Nevertheless, such analysis across a range of conditions allows us to analyze the sensitivity of the measures to different settings. In other words, we expect a stable ratio to follow similar patterns across all the settings, even though the pattern is not linear. If the patterns produced by the same measure vary across the settings (the panels of the figure), that would indicate instability of the measure with respect to data imbalance.

We can see from the tilts in the lines that the odds ratio and the impact ratio are very sensitive to imbalances in groups and positive outputs, the patterns vary a lot across the panels. The elift is more stable, except for deviations at very high acceptance with very few protected people and the opposite extreme. It is notable that the measured discrimination by all ratios grows very fast at low rates of positive outcome (e.g. see the plot $p(y^+) = 10\%$ and $p(s^1) = 90\%$), while there is little discrimination in the data according to the ground truth model. We also can see how all the ratios are asymmetric in terms of reverse discrimination. One unit of measured discrimination is not the same as one unit of reverse discrimination. This makes the ratios somewhat more difficult to interpret than differences, analyzed earlier, especially at large scale explorations and comparisons of, for instance, different computational methods for prevention of discrimination. Due to these reasons, we do not recommend using ratio based discrimination measures, since they are more difficult to interpret correctly. Instead we recommend using and building upon the difference based measures, discussed in Figure 2.

The core measures that we have analyzed form a basis for assessing fairness of predictive models, but it is not enough to use them directly, since they do not take into account possible legitimate explanations of differences between the groups, and instead consider any differences between the groups of people undesirable. The basic principle is to try to stratify the population in such a way that in each stratum contains people that are similar in terms of their legitimate characteristics, for instance, have similar qualifications if the task is candidate selection for job interviews. Propensity score matching, reported in Section 3.3, is one possible way for stratification, but it is not the only one, and outcomes may vary depending on internal parameter choices. Thus, the principle for measuring is available, but there are still open challenges ahead to make the approach more robust for different users, and more uniform across different task settings, such that one could diagnose potential discrimination or declare fairness with more confidence.

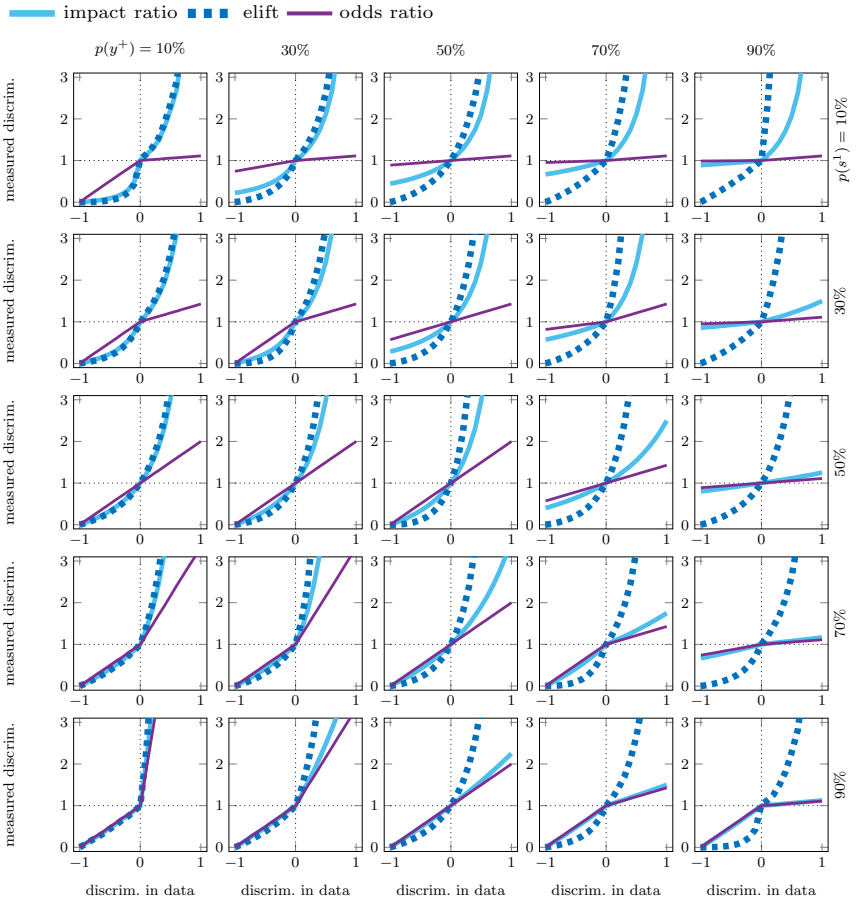



**Fig. 3** Analysis of the measures based on ratios: discrimination in data vs. measured discrimination.

## 5 Recommendations for researchers and practitioners

While the attention of researchers, the media and general public to potential discrimination is growing, it is important to measure the fairness of predictive models in a systematic and accountable way. We have surveyed measures used (and potentially usable) for measuring discrimination in data mining and machine learning, and experimentally analyzed the core discrimination measures in classification. Based on our analysis we generally recommend using the normalized difference, and in cases where the classes and groups of people in the data are well balanced, it may be sufficient to use the classical mean difference. We suggest using ratio measures with caution due to challenges associated with interpretation of their results in different situations.

We would like to emphasize that the absolute measures stand alone are not enough for measuring fairness. These measures can only be applied to uniform

populations where everybody within the population is equally qualified to get a positive decision. In reality this is rarely the case, for example, different salary levels may be explained by different education levels. Therefore, the main principle of applying the core measures should be by first segmenting the population into more or less uniform segments according to their qualifications, and then applying core measures within each segment. Discrimination-aware data mining is a young and rapidly developing discipline. The current state-of-the art-measures of algorithmic discrimination have their limitations. While the absolute measures are already well understood, the conditional measures are to a large extent open for research. A particularly challenging question is how decouple legitimate information and sensitive information carried by the same variable, such as zip code.

It is desired, but hardly possible to find any notion that covers all possible legal requirements. Moreover, the current legislation on non-discrimination has been designed to account for decision making by humans. The general principles apply to algorithmic decision making as well, but the nuances of algorithmic decision making are different. The legal base will need to be updated to account for algorithmic decision making. Input and expertise from computer science research is needed for incorporating algorithmic nuances into the legislation.

We hope that this survey can establish a basis for discussions and further research developments in this growing topic. Most of the research so far has concentrated on binary classification with binary protected characteristic. While this is a base scenario that is relatively easy to deal with in research, many technical challenges for future research lie in addressing more complex learning scenarios with different types and multiple protected characteristics, in multi-class, multi-target classification and regression settings, with different types of legitimate variables, noisy input data, potentially missing protected characteristics, and many more situations.

## References

1. Angwin, J., Larson, J.: Bias in criminal risk scores is mathematically inevitable, researchers say. ProPublica (2016). URL `https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say`
2. Arrow, K.J.: The theory of discrimination. In: O. Ashenfelter, A. Rees (eds.) Discrimination in Labor Markets, pp. 3–33. Princeton University Press (1973)
3. Barocas, S., Friedler, S., Hardt, M., Kroll, J., Venkatasubramanian, S., Wallach, H. (eds.): 2nd International Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) (2015). URL `http://www.fatml.org`
4. Barocas, S., Hardt, M. (eds.): International Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) (2014). URL `http://www.fatml.org/2014`
5. Barocas, S., Selbst, A.D.: Big data's disparate impact. California Law Review **104** (2016)
6. Berendt, B., Preibusch, S.: Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence. Artif. Intell. Law **22**(2), 175–209 (2014)
7. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc. (2006)

8. Blank, R.M., Dabady, M., Citro, C.F., on Methods for Assessing Discrimination, N.R.C.U.P.: Measuring racial discrimination. National Academies Press (2004)
9. Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. CoRR **abs/1510.00552** (2015)
10. Burn-Murdoch, J.: The problem with algorithms: magnifying misbehaviour. The Guardian (2013). URL `http://www.theguardian.com/news/datablog/2013/aug/14/problem-with-algorithms-magnifying-misbehaviour`
11. Calders, T., Karim, A., Kamiran, F., Ali, W., Zhang, X.: Controlling attribute effect in linear regression. In: Proc. of the 13th Int. Conf. on Data Mining, ICDM, pp. 71–80 (2013)
12. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. **21**(2), 277–292 (2010)
13. Calders, T., Zliobaite, I. (eds.): IEEE ICDM 2012 International Workshop on Discrimination and Privacy-Aware Data Mining (DPADM) (2012). URL `https://sites.google.com/site/dpadm2012/`
14. Calders, T., Zliobaite, I.: Why unbiased computational processes can lead to discriminative decision procedures. In: Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases, pp. 43–57. Springer (2013)
15. Citron, D.K., Pasqualle III, F.A.: The scored society: Due process for automated predictions. Washington Law Review **89** (2014)
16. Cochran, W.G.: Some methods for strengthening the common chi2 tests. Biometrics **10**(4), 417–451 (1954)
17. Commission, E.: How to present a discrimination claim: Handbook on seeking remedies under the EU Non-discrimination Directives. EU Publications Office (2011)
18. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S.: A computer program used for bail and sentencing decisions was labeled biased against blacks. its actually not that clear. The Washington Post (2016). URL `\url{https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.4ded14bf289e}`
19. Custers, B., Calders, T., Schermer, B., Zarsky, T. (eds.): Discrimination and Privacy in the Information Society. Data Mining and Profiling in Large Databases. Springer (2013)
20. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: Proc. of Innovations in Theoretical Computer Science, pp. 214–226 (2012)
21. Dwoskin, E.: How social bias creeps into web technology. The Wall Street Journal (2015). URL `http://www.wsj.com/articles/computers-are-showing-their-biases-and-tech-firms-are-concerned-1440102894`
22. Edelman, B.G., Luca, M.: Digital discrimination: The case of airbnb.com. Working Paper 14-054, Harvard Business School NOM Unit (2014)
23. Editorial: More accountability for big-data algorithms. Nature **537**(7621), 449 (2016)
24. European Union Agency for Fundamental Rights: Handbook on European non-discrimination law. EU Publications Office (2011)
25. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 259–268 (2015)
26. Fukuchi, K., Sakuma, J., Kamishima, T.: Prediction with model-based neutrality. In: Proc. of European conference on Machine Learning and Knowledge Discovery in Databases, pp. 499–514 (2013)
27. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3**, 1157–1182 (2003)
28. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng. **25**(7), 1445–1459 (2013)
29. Hajian, S., Domingo-Ferrer, J., Farras, O.: Generalization-based privacy preservation and discrimination prevention in data publishing and mining. Data Min. Knowl. Discov. **28**(5-6), 1158–1188 (2014)
30. Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., Giannotti, F.: Discrimination and privacy-aware patterns. Data mining and Knowledge Discovery **29**(6), 1733–1782 (2015)

31. Hardt, M., Price, E., , Srebro, N.: Equality of opportunity in supervised learning. In: Proc. of Advances in Neural Information Processing Systems 29, pp. 3315–3323 (2016)
32. Hillier, A.: Spatial analysis of historical redlining: A methodological explanation. Journal of Housing Research **14**(1), 137–168 (2003)
33. House, T.W.: Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. Executive Office of the President (2016)
34. Kamiran, F., Calders, T.: Classification without discrimination. In: Proc. nd IC4 conf. on Computer, Control and Communication, pp. 1–6 (2009)
35. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: Proc. of the 2010 IEEE International Conference on Data Mining, ICDM, pp. 869–874 (2010)
36. Kamiran, F., Zliobaite, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl. Inf. Syst. **35**(3), 613–644 (2013)
37. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Proc. of European Conference on Machine Learning and Knowledge Discovery in Databases, ECMLPKDD, pp. 35–50 (2012)
38. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: Proc. 8th Conf. on Innovations in Theoretical Computer Science (2017)
39. Luong, B.T., Ruggieri, S., Turini, F.: k-NN as an implementation of situation testing for discrimination discovery and prevention. In: Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD, pp. 502–510 (2011)
40. Mancuhan, K., Clifton, C.: Combating discrimination using bayesian networks. Artif. Intell. Law **22**(2), 211–238 (2014)
41. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics **18**(1), 50–60 (1947)
42. Mantel, N., Haenszel, W.: Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute **22**(4), 719–748 (1959)
43. Mascetti, S., Ricci, A., Ruggieri, S. (eds.): Special issue: Computational Methods for Enforcing Privacy and Fairness in the Knowledge Society, vol. 22. Artificial Intelligence and Law (2014)
44. Miller, C.C.: When algorithms discriminate. New York Times (2015). URL `http://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html`
45. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD, pp. 560–568 (2008)
46. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records. In: Proc. of the SIAM Int. Conf. on Data Mining, SDM, pp. 581–592 (2009)
47. Pedreschi, D., Ruggieri, S., Turini, F.: A study of top-k measures for discrimination discovery. In: Proc. of the 27th Annual ACM Symposium on Applied Computing, SAC, pp. 126–131 (2012)
48. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. Knowledge Eng. Review **29**(5), 582–638 (2014)
49. Romei, A., Ruggieri, S., Turini, F.: Discrimination discovery in scientific project evaluation: A case study. Expert Syst. Appl. **40**(15), 6064–6079 (2013)
50. Rorive, I.: Proving discrimination cases the role of situation testing (2009). URL `\url{http://migpolgroup.com/public/docs/153.ProvingDiscriminationCases_theroleofSituationTesting_EN_03.09.pdf}`
51. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **1**, 41–55 (1983)
52. Ruggieri, S.: Using t-closeness anonymity to control for non-discrimination. Trans. Data Privacy **7**(2), 99–129 (2014)
53. Ruggieri, S., Hajian, S., Kamiran, F., Zhang, X.: Anti-discrimination analysis using privacy attack strategies. In: Proc. of European Conference on Machine Learning and Knowledge Discovery in Databases, ECMLPKDD, pp. 694–710 (2014)
54. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM Trans. Knowl. Discov. Data **4**(2), 9:1–9:40 (2010)

55. Tax, D.: One-class classification. Ph.D. thesis, Delft University of Technology (2001)
56. Tsoumakas, G., Katakis, I.: Multi-label classification: an overview. International Journal of Data Warehousing & Mining **3**(3), 1–13 (2007)
57. Wihbey, J.: The possibilities of digital discrimination: Research on e-commerce, algorithms and big data. Journalist's resource (2015). URL \url{https://journalistsresource.org/studies/society/internet/possibilities-online-racial-discrimination-research-airbnb}
58. Yinger, J.: Measuring racial discrimination with fair housing audits: Caught in the act. The American Economic Review **76**(5), 881–893 (1986)
59. Zemel, R.S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: Proc. of the 30th Int. Conf. on Machine Learning, pp. 325–333 (2013)
60. Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: A causal inference approach. In: Proc. of the 25th Int. Joint Conference on Artificial Intelligence, IJCAI, pp. 2718–2724 (2016)
61. Zliobaite, I.: On the relation between accuracy and fairness in binary classification. In: The 2nd workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15 (2015)
62. Zliobaite, I., Custers, B.: Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artif. Intell. Law **24**(2), 183–201 (2016)