Artif Intell Law (2014) 22:211–238
DOI 10.1007/s10506-014-9156-4

Combating discrimination using Bayesian networks

Koray Mancuhan · Chris Clifton

Published online: 17 February 2014

Abstract Discrimination in decision making is prohibited on many attributes (religion, gender, etc…), but often present in historical decisions. Use of such discriminatory historical decision making as training data can perpetuate discrimination, even if the protected attributes are not directly present in the data. This work focuses on discovering discrimination in instances and preventing discrimination in classification. First, we propose a discrimination discovery method based on modeling the probability distribution of a class using Bayesian networks. This measures the effect of a protected attribute (e.g., gender) in a subset of the dataset using the estimated probability distribution (via a Bayesian network). Second, we propose a classification method that corrects for the discovered discrimination without using protected attributes in the decision process. We evaluate the discrimination discovery and discrimination prevention approaches on two different datasets. The empirical results show that a substantial amount of discrimination identified in instances is prevented in future decisions.

Keywords Discrimination discovery · Discrimination prevention · Bayesian network · Data mining

1 Introduction

Discrimination is treating people unequally according to their membership in a specific group, class, or category. Membership criteria can encompass race, gender, native-country, religion, age, etc. A remarkable amount of legal regulations ban

K. Mancuhan (✉) · C. Clifton
Purdue University, 305 N. University St., West Lafayette, IN 47907, USA
e-mail: kmancuha@purdue.edu

C. Clifton
e-mail: clifton@cs.purdue.edu

discriminatory decisions on an individual (instance) basis. The European Union (EU) principle of equal treatment of individuals regardless of a variety of attributes is well established by an important body of Community law, in particular in Article 14 of the European Convention on Human Rights (Rome, 4.XI.1950):

> The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.

Council Directive 76/207/EEC of 9 February 1976 sets the implementation of the principle of equal treatment for men and women as regards access to employment, vocational training and promotion, and working conditions. Council Directive 2000/43/EC of 29 June 2000 implements the principle of equal treatment between persons irrespective of racial or ethnic origin. A general framework for equal treatment in employment and occupation is established in Council Directive 2000/78/EC of 27 November 2000. In the United States (US), Title VIII of the Civil Rights Act of 1968 (Fair Housing Act) prohibits discrimination in the sale, rental and financing of dwellings based on race, color, religion, sex or national origin. Title VII of the Civil Rights Act of 1964 prohibits employment discrimination based on race, color, religion, sex, or national origin. The Equal Credit Opportunity Act of 1974 states that it shall be unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction on the basis of race, color, religion, national origin, sex or marital status, or age provided the applicant has the capacity to contract.

Although the legal regulations against the discrimination of individuals are clear, recent cases in the EU and in the US show that there are still effective direct and indirect discrimination in professional life and in provided services. EU's Court of Justice ruled in March 2011 (Test-Achats Case 236/09) that different insurance premiums for women and men constitute sex discrimination and that they are not compatible with the EU's Charter of Fundamental Rights. In January 2013, the European Court of Human Rights (ECHR) ruled that the domestic authorities failed sufficiently to protect Nadia Eweida's right to manifest her religion, in breach of the positive obligation under Article 9 (Eweida v. The United Kingdom Case 48420/10). Despite not being a very strong case of anti-discrimination, this recent case is a real life example of indirect discrimination about religious freedom. In Jackson v. Birmingham Board of Education, a teacher and girls' basketball coach in an Alabama public high school, complained about sex discrimination in the school's program and was later removed from his coaching position. The court held that such retaliation is a form of intentional sex discrimination forbidden by the statute in Title IX (U.S. Supreme Court 544-167, 2005). The U.S. Supreme court is currently revisiting the issue of use of race in college admissions in *Fisher v. University of Texas at Austin* (U.S. Supreme Court No. 11-345). These are just a sample of cases occurring since 2005; many similar cases occur in many jurisdictions.

Therefore, mechanisms to discover discrimination towards specific individuals and to enforce equality for all individuals are needed. Data mining has a good potential for such mechanisms. We will next give some examples from E.U. and

U.S. law about legal challenges in using protected and non-protected attributes, give a set of definitions; and define formally the discrimination discovery problem.

1.1 Legal background: protected and non-protected attribute usage

1.1.1 Discriminatory decision cases (individuals)

Protected attribute usage It is becoming increasingly clear that any use of a protected attribute in making a decision about an individual is prohibited. Article 8 of Directive 95/46/EC was clear about the wide spread usage of protected attributes such as race, gender, medical information, … across all member states (Article 8, 95/46/EC):

> Whereas, in order to remove the obstacles to flows of personal data, the level of protection of the rights and freedoms of individuals with regard to the processing of such data must be equivalent in all Member States; whereas this objective is vital to the internal market but cannot be achieved by the Member States alone, especially in view of the scale of the divergences which currently exist between the relevant laws in the Member States and the need to coordinate the laws of the Member States so as to ensure that the cross-border flow of personal data is regulated in a consistent manner that is in keeping with the objective of the internal market as provided for in Article 7a of the Treaty; whereas Community action to approximate those laws is therefore needed

The EU Gender Directive of 13 December 2004 implemented the principle of equal treatment between men and women in the access to and supply of goods and services (Article 2, 2004/113/EC). It made distinction about the protected attribute usage (gender) between direct and indirect discrimination towards individuals:

- *direct discrimination:* where one person is treated less favorably, on grounds of sex, than another is, has been or would be treated in a comparable situation;
- *indirect discrimination:* where an apparently neutral provision, criterion or practice would put persons of one sex at a particular disadvantage compared with persons of the other sex, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary;

The March 2011 E.U. Court of Justice ruling on insurance premiums (Test-Achats Case 236/09) was quite explicit: Gender cannot be used to set insurance premiums, despite actuarial evidence that there are sound business reasons to do so. The court ruled that an exemption written into the law allowing use of gender in setting insurance premiums was only transitional, and any such use must end by 21 December 2012. The U.S. Equal Credit Opportunity Act is equally explicit:

> Specific rules concerning use of information (1) Except as provided in the Act and this regulation, a creditor shall not take a prohibited basis into account in any system of evaluating the creditworthiness of applicants. (12 CFR 202.6(b))

The exceptions are that special purpose programs for the benefit of an economically disadvantaged class of persons may require participants to possess one or more common protected characteristics, but this is a requirement rather than a factor in the decision. Specifically, such a program is allowed "… only if it was established and is administered so as not to discriminate against an applicant on any prohibited basis; however, all program participants may be required to share one or more common characteristics (for example, race, national origin, or sex) …" (12 CFR 202.8(b)(2)).

Special purpose programs (or other similar programs in different legislative contexts) are limited cases where use of a protected attribute may be acceptable, but such use must be carefully tailored to combat discrimination. For instance, the U.S. Supreme Court upheld the narrowly-tailored use of race as a factor in deciding admissions to improve diversity at the University of Michigan Law School in *Gruttinger v. Bollinger* (U.S Supreme Court 02-241). In contrast, the U.S. Supreme Court the same year banned a broader use that made race a decisive factor in their undergraduate admissions process (*Gratz v. Bollinger*, U.S. Supreme Court 02-516). While this suggests that techniques that adjust decisions or scores based on protected attributes may have some applicability, the U.S. Supreme Court's decision to revisit race in college admissions in *Fisher v. University of Texas at Austin* (U.S. Supreme Court No. 11-345) raises questions about the long-term applicability of any technique that uses an individual's protected attribute in making a decision about that individual. The legal cases, which are mentioned above, show that the usage of protected attribute for fairness might tend to have a broader usage in decision making about individuals. Broad usage might divert from providing equal treatment to disadvantaged groups, and eventually create new types of disadvantaged groups (*Gratz v. Bollinger*, U.S. Supreme Court 02-516).

1.1.2 Discriminatory decision making models

Protected attribute usage Staff interpretations of the U.S. Equal Credit Opportunity Act sets specifically the usage of protected attribute age in models learned from data:

> An empirically derived, demonstrably and statistically sound, credit scoring system may include age as a predictive factor (provided that the age of an elderly applicant is not assigned a negative factor or value). Besides age, no other prohibited basis may be used as a variable. (Supplement I to Part 202.2(p)4)

While the above statement makes it clear that protected attributes should not be used in decision making models, the above statements do not appear to prevent using a model that does not include protected attributes, even if that model is learned from historic data and as a result perpetuates past discrimination.

Non-protected attribute usage It is clear that non-protected attributes can have a high correlation with protected attributes, and use of such correlated attributes can

perpetuate discrimination. Perhaps the most famous example is mortgage redlining in the U.S., a process that began with the National Housing Act of 1934. This act resulted in the development of maps dividing neighborhoods into four categories, with "Type D" neighborhoods outlined in red and considered least desirable for lending. Because of restrictive covenants in newer neighborhoods, minorities were often confined to redlined neighborhoods and thus denied loans, even if race was not an explicit factor in the decision. Redlining was such a widespread and egregious practice that the term "redlining" is explicitly mentioned in U.S. law (12 CFR 27.4(a)(3)). While redlining on the basis of race, color, religion, or national origin was prohibited by Title VIII ("Fair Housing Act") of the Civil Rights Act of 1968 (P.L. 90-284 section 804), this did not directly address the question of use of non-protected attributes for a legitimate business purpose where such attributes correlated with protected attributes.

The Equal Credit Opportunity Act *is* explicit on this point, and the history and extent of this is captured in the following from the official staff interpretations of the act:

> Effects test. The effects test is a judicial doctrine that was developed in a series of employment cases decided by the U.S. Supreme Court under Title VII of the Civil Rights Act of 1964 (42 U.S.C. 2000e et seq.), and the burdens of proof for such employment cases were codified by Congress in the Civil Rights Act of 1991 (42 U.S.C. 2000e-2). Congressional intent that this doctrine apply to the credit area is documented in the Senate Report that accompanied H.R. 6516, No. 94-589, pp. 4–5; and in the House Report that accompanied H.R. 6516, No. 94-210, p. 5. The Act and regulation may prohibit a creditor practice that is discriminatory in effect because it has a disproportionately negative impact on a prohibited basis, even though the creditor has no intent to discriminate and the practice appears neutral on its face, unless the creditor practice meets a legitimate business need that cannot reasonably be achieved as well by means that are less disparate in their impact. (Supplement I to Part 202.6(a) 2.)

This makes it quite clear that even though a learned model does not use protected attributes, if that model (e.g., because of historical discrimination reflected in the training data) discriminates against a protected group, use of that approach is prohibited if an alternative approach obtains comparable outcomes (e.g., classifier accuracy) with less discrimination.

1.2 Definitions

We now give a set of definitions used to describe our discrimination discovery and prevention methods.

Definition 1 *Protected Attribute* is a dataset attribute that has been used to discriminate against individuals in instances. Example protected attributes include gender, race, religion, etc…

Next we define direct discrimination and indirect discrimination, based on EU legislation emphasized in EU council directive at December 13 2004 (2004/113/EC) for provided goods and services (Article 2).

Definition 2 *Direct Discrimination* is an event where one person is treated less favorably, on grounds of the protected attribute, than another is, has been or would be treated in a comparable situation

Definition 3 *Indirect Discrimination* is an event where an apparently neutral provision, criterion or practice would put persons of one protected attribute value at a particular disadvantage compared with persons of the other protected attribute value, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary.

We continue to the definitions by setting standard names for instances suffering from direct or indirect discrimination in a given set of instances. We also define here the discriminatory dataset and classifiers.

Definition 4 *Discriminated Instance* is an instance suffering from either direct or indirect discrimination among a set of instances. If a discriminated instance is corrected such that it does not suffer from either direct or indirect discrimination, then it is called a *corrected discriminated instance*.

Definition 5 *Non-Discriminated Instance* is an instance not suffering from both direct and indirect discrimination among a set of instances.

Definition 6 *Discriminatory dataset* is a dataset that has 1 or more discriminated instances. A dataset is *non-discriminatory* if it does not contain any discriminated instances (all instances are non-discriminated). Given a discriminatory dataset, if its instances are corrected discriminated instances, the dataset is called a *corrected discriminatory dataset*.

Definition 7 *Discriminatory classifier* is a classifier that is learned from the discriminatory dataset. A classifier is *non-discriminatory* if this classifier is either learned from a non-discriminatory dataset or learned from a corrected discriminatory dataset.

We conclude this section by defining redlining effect, redlining attributes and elift measure that have been used in the data mining literature to measure discrimination.

Definition 8 *Redlining Effect* (Calders and Verwer 2010) is the effect that a classifier has when a classifier learns discriminatory rules using features that correlate with the protected attribute. *Redlining attribute* is an attribute which is correlated to the protected attribute. Redlining effect is the reflection of indirect discrimination in a learning problem due to redlining attribute, because redlining attribute causes discriminatory classifiers despite ignoring protected attributes.

Definition 9 *Elift* (Pedreschi et al. 2008) is a discrimination measure for association rules. It literally calculates how many times the item set $A \subseteq \{a_1, a_2, \ldots, a_n\}$ increases an instances membership to a class $C \in \{c_1, c_2\}$ with respect to context item set $B \subseteq \{b_1, b_2, \ldots, b_m\}$ in a classification rule of the form $A, B \rightarrow C$.

$\{a_1, a_2, \ldots, a_n\} \cup \{b_1, b_2, \ldots, b_m\}$ is the set of attributes defined in the dataset. Formally, it is the ratio of the confidence (conf) of a classification rule $A, B \rightarrow C$ that has item set A in the antecedent part over the confidence of the context rule $B \rightarrow C$ that does not have item set A in the antecedent. *Elift* is defined as:

$$elift = \frac{conf(A, B \rightarrow C)}{conf(B)} \tag{1}$$

The context rule in the denominator of the formula is obtained from the original classification rule in the numerator by discarding the item set A in the antecedent part. *Elift* is used to identify if a classification rule discriminates with respect to protected attributes by setting item set A to contain the protected attributes. A classification rule is considered significantly discriminative when its *elift* is higher than a user defined threshold. Other types of measures and the statistical significance of these measures can be found in Pedreschi et al.'s paper (Pedreschi et al. 2008, 2009).

We can now formally define the problem addressed in this paper.

1.3 Problem formulation

Assume that the sets of attributes $A = \{a_1, a_2, \ldots, a_l\}, B = \{b_1, b_2, \ldots, b_m\}$, and $R = \{r_1, r_2, \ldots, r_n\}$ exist. The sets represent the protected attributes, the non-protected attributes, and the redlining attributes respectively. A, B and R have domains $\{dom(a_1), dom(a_2), \ldots, dom(a_l)\}, \{dom(b_1), dom(b_2), \ldots, dom(b_m)\}$ and $\{dom(r_1), dom(r_2), \ldots, dom(r_n)\}$ respectively. (We do *not* assume that which attributes fall in B and which fall in R is known, only that such (possibly empty) sets exists.) The binary class label (decision) is defined as $C = \{ -, + \}$. Binary class labels are considered throughout this paper.

An instance x is formed from the set of attributes and the binary class label:

$$x = (x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m, z_1, z_2, \ldots, z_n, c) \tag{2}$$

where $x_i \in dom(a_i), y_i \in dom(b_i), z_i \in dom(r_i)$ and $c \in C$ hold. The dataset D is defined as:

$$D = x^1, x^2, \ldots, x^k \tag{3}$$

where x^i stands for the ith instance in the dataset.

We define *belift* (Bayesian *elift*) as:

$$belift = \frac{P(C \mid a_1, a_2, \ldots, a_l, b_1, b_2, \ldots, b_m, r_1, r_2, \ldots, r_n)}{P(C \mid b_1, b_2, \ldots, b_m)} \tag{4}$$

such that

$$P(C \mid a_1, a_2, \ldots, a_l, b_1, b_2, \ldots, b_m, r_1, r_2, \ldots, r_n) > t > P(C \mid b_1, b_2, \ldots, b_m) \tag{5}$$

holds. We call Eq. 4 Bayesian *elift*, because its conditional probabilities are estimated via a Bayesian network. We will elaborate in the solution section (cf. Sect. 3).

t is the decision boundary between $c = -$ and $c = +$ in Eq. 5. For example, suppose that t is set to 0.5. An instance x^i will be assigned to the class label $-$ if

$P(c = -|x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m, z_1, z_2, \ldots, z_n) > 0.5$ for the $-$ class. Otherwise, if $P(c = -|x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m, z_1, z_2, \ldots, z_n)$ satisfies <0.5 for the class label $-$, then the instance will be assigned to the $+$ class label. ($P(c = +|x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m, z_1, z_2, \ldots, z_n) > 0.5$ holds for $c = +$ as we assume a binary classification problem).

The condition in Eq. 5 lets Eq. 4 capture class flips ($c = -$ to $c = +$ or $c = +$ to $c = -$) due to the usage of sets A and R in class probabilities. Informally, *belift* calculates how many times the usage of protected attributes (A) and the redlining attributes (R) increase the class probability for a given instance with respect to joint class probability distribution.

The reader may be concerned about the validity of Eq. 5. Given a binary protected attribute, the inequality $>$ is valid for one specific protected attribute value whereas the inequality $<$ is valid for the other protected attribute value (fixed class value). These two cases are complementary for discrimination. One means degradation while the other means favoring. We give an example to clarify this issue. Suppose that we have census data with binary class label *low income* and *high income*. We would like to define the discrimination problem using Eqs. 4 and 5. The binary protected attribute is gender (male, female). Assume that there is no redlining attribute (simple case). Females have low income because of just being female ($<$ in Eq. 5) when they deserve a high income regarding to merit and qualifications. Their incomes are *degraded*. Meanwhile, males have higher income because of "being male" ($>$ in Eq. 5). Incomes are *favored* towards males since they don't have low income (have high income) despite the lack of merit and qualification in protected attributes.

Equations 4 and 5 treat redlining attributes and protected attributes differently because of legal reasons and algorithm performance. Legislations ban decision makers using protected attributes in their decision models. However, there is no ban on the legitimate use of the redlining attributes. They thus can be used for learning a model as long as their correlation with protected attribute is removed (i.e., it does not result in indirect discrimination). Redlining attributes can be good predictors for classification algorithms whereas the protected attributes perpetuate classification algorithms with discriminatory bias (even if reverse bias).

In this problem setting, we assume that discrimination occurs randomly across dataset D. Each instance, that has a specific protected attribute value, is equally likely to be a victim of discrimination. A real life discrimination example would be bank credit decisions. Individuals who have same credit risk score, same occupation, same age etc.. can be discriminated randomly with respect to being foreigner or not. Randomness comes from the bank officer who confirms lending credit. While this makes the (likely incorrect) assumption that each bank officer is equally likely to discriminate provided the same sample of credit applicants, this is largely unavoidable unless the individual making the decision in the training data were included as an attribute.

The first advantage of *belift* is that the estimated joint class probability distribution is more accurate and realistic than the estimated joint class probability of classification rules. Bayesian networks capture the intra-correlation between the attributes. The second advantage of *belift* is that its condition (Eq. 5) force identification of the decision change in the decision making process due to the usage

of $A = \{a_1, a_2, \ldots, a_l\}$ and $R = \{r_1, r_2, \ldots, r_n\}$. It uses the decision boundary t to identify *decision flip*. The original *elift* measure does not consider a decision boundary in the underlying data. A third advantage of *belift* is that it targets direct and indirect discrimination since it considers the protected attribute and the redlining effect due to $A = \{a_1, a_2, \ldots, a_l\}$ and $R = \{r_1, r_2, \ldots, r_n\}$ respectively. *Elift* does not target indirect discrimination (Pedreschi et al. 2008, 2009).

Since we defined the problem setting and *belift* measure, we now define the discrimination discovery problem itself. Given a dataset D having a set of instances x^i, the problem is to determine a subset of D that are discriminated instances. The determination should be achieved using the *belift* measure. This subset should also be determined with respect to a given class label value (either + or −). There are two critical points of this problem. First, the estimation of the numerator and denominator in the *belift* formula. Second, the identification of redlining attributes R in the data so that we can apply *belift* for a given instance x^i. The protected attributes are assumed to be known for analysis (gender, religion, etc…), so identifying them is not a problem.

Section 2 gives a brief summary of existing work in the literature and lays out our contribution. We continue our explanation by introducing the discrimination discovery and prevention propositions in Sects. 3 and 4 (problem solution). We run a case study in the experiments to investigate gender discrimination on two realistic datasets from the UCI repository (Newman et al.). The results show that our discrimination discovery approach identifies discriminated instances and our discrimination prevention technique improves the enforcement of gender equality.

2 Related work and contributions

2.1 Related work

Data mining's potential for discrimination discovery and prevention, first discussed Pedreschi et al. (2008), has attracted the attention of the data mining community. Since Pedreschi et al., several data mining techniques have been proposed to address the discrimination issue in decision making. Luong et al. (2011) and Pedreschi et al. (2009) proposed k-NN and classification rule approaches for discrimination discovery. Ruggieri et al. implemented a discrimination discovery tool called DCUBE that measures discrimination of socially protected instances. DCUBE uses a classification rule approach to discover discrimination on instances (Ruggieri et al. 2011). Luong et al. (2011) and Pedreschi et al. (2009) don't use the protected attribute for decision making purposes, whereas they use the protected attribute to measure its effect on historical decisions. As the protected attribute is not used to produce new modified decisions (outcomes), they do not violate the E.U. and the U.S. legislation mentioned earlier. However, Luong et al. and Pedreschi et al. don't aim to identify indirect discrimination. Their proposition's scope is bounded with the discovery of direct discrimination.

Kamiran et al. and Calders et al. introduced discrimination prevention techniques. These techniques include correction of samples (Kamiran and Calders 2009)

and adjusting the scoring functions of classifiers including decision trees (Kamiran et al. 2010) and Naïve Bayes (Calders and Verwer 2010). The problem with these approaches is that they are not effective unless the protected attribute (e.g., gender, religion, etc…) is included in the decision process. Calders et al. emphasize this fact by testing their Non-Discriminatory Naïve Bayes approach with additional experiments (Calders and Verwer 2010). The algorithm uses the protected attribute to enforce fairness in these experiments. Thus, Calders et al. propose implicitly a Naïve Bayes modeling method that can ensure fairness with accurate predictions when the protected attribute is used (Calders and Verwer 2010). As we have seen, direct use of protected attributes in making individual decisions is often (although not always) prohibited.

Zliobaite et al. extend Kamiran et al.'s relabeling technique (Kamiran and Calders 2009) by the concept of explanatory attribute (Zliobaite et al. 2011). They assume that the unbalanced distribution of protected attributes among + labeled instances can be acceptable according to an explanatory attribute. However, their solution based on K-means clustering is a saddle point when multiple explanatory attributes are used to justify this unbalanced distribution. They propose using K-means clustering to define a new explanatory attribute that is concatenated from the set of explanatory attributes among the instances in the clusters. Using K-means clustering to handle this case may not be a good solution because each clustering run would give different sets of explanatory attributes and discrimination analysis from the same data. There is also the cost of the clustering algorithm in addition to twice training a classification model (training of ranker for correction and training of a discrimination-aware classifier on the corrected training set).

Recently, Kamiran and Calders survey and extend their data preprocessing techniques in Zliobaite et al. (2011), Kamiran et al. (2010) for discrimination-aware classification (Kamiran and Calders 2011). They discuss various suppression techniques, massaging techniques (relabeling discriminated instances in the data) and reweighing and resampling techniques. A Weka-based non-discriminatory classification tool is also introduced. Romei and Ruggieri (2013) provide a recent multidisciplinary survey about various discrimination analysis techniques, overviewing data analysis techniques developed in last 50 years. Kamiran et al. (2012) introduce the decision theory into discrimination-aware classification to avoid data tweaking and classifier modification. Hajian and Ferrer investigate how the current k-anonymization techniques in privacy community affects the datasets in terms of discrimination. Their case study shows that the anonymization techniques does not handle the existing historical discrimination in the data (Hajian and Domingo-Ferrer 2012). Hajian et al. (2012) also investigate the mitigation of privacy awareness and discrimination injection regarding to pattern discovery. Mancuhan and Clifton (2012) propose a discrimination prevention technique based on decision policies. Essentially, they use Luong et al.'s work (Luong et al. 2011) to model a decision policy and to use a relabeling technique for correcting discrimination in instances. They propose learning a classifier from the corrected instances. Dwork et al. (2012) suggest the usage of statistical parity to assure discrimination free (or fair) decisions. A very recent work of Zemel et al., which is an extension of Dwork et al., focuses on discrimination prevention using statistical parity (Zemel et al. 2013).

Essentially, their objective is to map data into a new space where data is partitioned into different prototypes. Mapping process is independent from the protected attribute while keeping the attribute information as much as possible.

2.2 Contributions

One problem with prior approaches is that they do not distinguish between discriminatory practices and legitimate business reasons why the class labels may be imbalanced. The closest work which is achieving this is Zemel et al.'s work. However, they assume that the discriminated instances are already identified in the dataset. The challenge of identifying these instances is not addressed. Even more critical, many other works fail to achieve fairness unless the protected attribute is used as part of the decision process. In Sect. 1.1, we explained the legal problems of using the protected attribute to provide fairness for individuals. Furthermore, most of the previous approaches fail to model the overall decision process for a given set of instances in both discovery and prevention. The closest work doing this task is Pedreschi et al.'s discrimination discovery approach (Pedreschi et al. 2009) and Calders et al.'s latent variable model (Calders and Verwer 2010). Despite of the fact that Calders et al. have an objective to determine the latent variable (fair class label) and to model the decision process, their proposed solution raises issues and is hard to generalize. They model latent variable either using EM clustering or using prior knowledge. EM clustering is not a stable method. Discrimination is tied to legal cases and requires justifications from stable methods. In addition, they maintain the naive assumption for Bayesian network structures simulating decision process. Mancuhan and Clifton used Pedreschi et al.'s decision discovery method to model a decision process and prevent discrimination in the predictions (Mancuhan and Clifton 2012) . However, their discovery process relies on the usage of classification rules. The classification rules fall short of modeling a decision process since they only consider the dependence between the antecedent attributes and the class label. These issues may also raise legal concerns.

We propose a new discrimination discovery process based on modeling decision process that targets both direct and indirect discrimination. To our knowledge, this is the first paper studying the discovery of both direct and indirect discrimination discovery. The discrimination discovery mechanism of this paper is based on the Bayesian networks. Bayesian networks consider the dependence between all the attributes and they estimate the joint probability distribution without any strong assumption. Bayesian networks thus can be generalized easily. This makes them an appropriate model for approximating the decision process. We also extend this discovery method using a relabeling technique similar to Kamiran et al.'s massaging proposition (Kamiran and Calders 2009), and Mancuhan and Clifton's correction approach (Mancuhan and Clifton 2012) in order to make non-discriminatory predictions.

As we shall see, the method proposed in this paper identifies specific individuals where discrimination occurs. Instead of using a latent variable like Calders et al., this paper uses two Bayesian networks to identify and prevent discrimination. One of the Bayesian networks is the benchmark for discrimination free classification

while the other Bayesian network has the discriminatory bias. Briefly, we learn a Bayesian Network (with discriminatory bias) that captures the decision processes, and identify nodes in that network that lead to high decision disparity relative to the protected attribute. These attributes are discounted, forcing the decision process to rely on other attributes (benchmark Bayesian network). The result is a classifier that reduces discrimination relative to the training data, but does not use an individual's protected attribute(s) in making a decision about that individual.

It is worth noting that Calders et al. also propose combining two Bayesian networks (Calders and Verwer 2010). However, they propose learning models on two different groups where each group belongs to a set of people having one specific protected attribute value (binary protected attribute). Their model makes predictions using a combination of these two group networks. The difference in our work is that we focus on capturing the protected attribute's effect by using two models. We don't make a weighted combination of them for final decision making. Both models are learned on the entire dataset so that we can measure both the protected and redlining attributes' effect on the decision making process.

We want to stress that we use the protected attribute only to discover direct and indirect discrimination, as with Pedreschi et al. (2009) and Luong et al. (2011). We do not use the protected attributes in the prediction process, and we never employ it for making predictions as is done by Calders et al. (Calders and Verwer 2010). Thus, the discovery process using the protected attribute and prediction process ignoring the protected attribute are part of the *measures enforcing equality*. This practice is legally grounded with respect to 76/207/EEC, 2004/113/EC, 2000/78/EC and 2000/78/EC E.U. council directives that ban the usage of the protected attribute for decisions made and which oblige all E.U. member states to take *measures enforcing equality* between different individuals with respect to gender, race, religion and other protected attributes.

3 Discrimination discovery

We now proceed to our discrimination discovery process. Our objective is to determine the total discrimination in the discovery process. The differentiation of explainable and non-explainable discrimination (Zliobaite et al. 2011) in discovery is a future extension of our approach. We also assume that the provided protected attributes are binary. Handling numerical and/or non-binary protected attributes is another potential future extension. Multiple protected attributes *are* handled in the discrimination discovery process.

Elift is a discrimination measure based on the $conf(A, B \rightarrow C)$ of $A, B \rightarrow C$ classification rule. *Conf* estimates essentially the $P(C|A, B)$ class probability. Although *elift* provides a good method for discrimination discovery, it suffers from one essential assumption of classification rules. It assumes that attributes in A, B item sets are independent from each other. However, this is a very strong assumption that does not generally hold in real life.

In contrast, Bayesian networks estimate the probability $P(A, B, C)$ by capturing the conditional dependencies between the attributes within the item sets A and

```
DiscDiscovery(p,c,thres,db,D)
Input: p protected attributes
         c class value
         thres threshold value for belift
         db decision boundary value for class probabilities
         D dataset of historical decision records
Output: (DI, NDI, net)
         DI discriminated instances
         NDI non-discriminated instances
         net Bayesian network learnt from dataset D
begin:
net<-buildBayesNetwork(D)
relative_net<-removeProtectedRedliningAttributes(net, p)
DI<-createEmptyDataset()
NDI<-create EmptyDataset()
for each instance in D:
         begin:
         classProb<-estimateProb(instance,c,net)
         relClassProb<-estimateProb(instance,c,relative_net)
         belift<-calculateBelift(classProb,relClassProb,db)
         if(belift>=thres):
                  begin:
                  DI<-addInstance(DI,instance)
                  end
         else:
                  begin:
                  NDI<-addInstance(NDI,instance)
                  end
         end
return (DI,NDI,net)
end
```

Fig. 1 Discrimination discovery strategy pseudo code

B. The Bayesian networks can be used to estimate $P(A, B, C)$ probability and the $P(C|A, B)$ class probability can be derived from the Bayes theorem. As a result, Bayesian networks are a better class of models than classification rules to define a decision process. *Elift* can be extended intuitively by calculating the numerator and the denominator probabilities with Bayesian networks.

Figure 1 lays out our Discrimination Discovery Strategy in pseudo code, solving the problem defined in Sect. 1.3. The process starts by building a Bayesian network from a given set of instances D. Then, a copy of the Bayesian network is created. The protected attributes, the parents of the protected attributes and the children of the protected attributes are deleted from the copied network using *removeProtectedRedliningAttributes* function (Fig. 3). (If the protected attribute has the class attribute as parent in the network, it is ignored in the deletion.) The parents and the children of the protected attributes define the redlining attributes in the *belift* formula. The reason why the neighboring nodes of the protected attributes are redlining attributes is that the highest correlation exists between the protected attributes and their neighboring attributes in the entire data. Thus, they could lead to indirect discrimination. We call this copy the relative Bayesian network, since it is a relative comparison baseline excluding the protected attributes and the redlining attributes in the probability distribution of the dataset. The original and relative Bayesian networks will be used later to estimate the numerator and denominator of the belift measure for each instance in the dataset.

```
calculateBelift(classProb,relClassProb,db)
Input: classProb
         relClassProb
         db decision boundary value for class probabilities
Output: belift
begin:
         belift<-(-1)
         if(classProb > db and relClassProb < db):
                  begin:
                  belift<-classProb/relClassProb
                  end
return belift
end:
```

Fig. 2 *calculateBelift* function pseudo code

The next step in the discrimination discovery process is the calculation of *belift* for each instance in the dataset D (Fig. 1). The calculation of *belift* is based on the class probabilities calculated from the Bayesian network and the relative Bayesian network using Bayes theorem. These probabilities are calculated for each instance (Fig. 1) and they are placed in the original formula of *belift* using the *calculateBelift* function (Fig. 2). The function *calculateBelift* returns -1 if there is no decision flip for a provided decision boundary. Otherwise, it verifies if there is discrimination in the instance's decision (class label) according to its *belift* value. If there is, then the instance is considered as discriminated. If there is not, the instance is considered non-discriminated. The *belift* calculation uses the provided class label value (either $+$ or $-$), the provided protected attribute and the provided decision boundary. They depend on the analysis type that data scientists would like to achieve and they are the provided parameters to the algorithm (Fig. 1). When all the instances in the dataset are evaluated, the discrimination discovery algorithm returns the Bayesian network, the discriminated instances, and the non-discriminated instances (Fig. 1).

The initial network structure assumptions are important since discrimination discovery algorithm relies on Bayesian networks. In this paper, we assume an initial Bayesian network structure with naive assumption (Naïve Bayes Network). Additional edges are added using the hill climbing search algorithm K2. We use a local search algorithm to deal with computational issues. For more detail about the K2 algorithm, see Cooper and Herskovits (1991).

The verification of discrimination for an instance's class label is also critical. The verification is done by setting a specific threshold value for the *belift* measure and by comparing the instance's *belift* value to this specific threshold value (Fig. 1). The choice of this specific threshold value depends on the legal environment of the analysis. The critical point is that higher values of threshold force a more strict criterion for the identification of discriminated instances. As the threshold value gets higher, the instances that are far from the decision boundary and which have decision flip are considered discriminated. This threshold value is provided as a parameter to the algorithm (Fig. 1).

In this paper, we propose 1 as the threshold value due to the E.U. and U.S. legislation. We call that an instance satisfies the *perfect equality point* if its *belift* value is equal to 1. The *perfect equality point* and the threshold value 1 are justified

```
removeProtectedRedliningAttributes(net, p)
Input: net Bayesian network
       p protected attributes
Output: relative_net
       relative_net Bayesian network obtained from net by
       deleting p nodes, parent nodes of p nodes and
       children nodes of p nodes
begin:
relative_net<-net
for each attr in p:
       begin:
       parent<-findParents(net,attr)
       children<-findChildren(net,attr)
       relative_net<-deleteNode(attr,relative_net)
       for each attr in parent:
              begin:
              relative_net<-deleteNode(attr,relative_net)
              end

       for each attr in children:
              begin:
              relative_net<-deleteNode(attr,relative_net)
              end
       end
return relative_net
end
```

Fig. 3 *removeProtectedRedliningAttributes* function pseudo code

by the definition of the direct and indirect discrimination in E.U. council directive 2004/113/EC (Article 2). E.U. and U.S. legislations require equal treatment (e.g., no decision flip) for an instance x^i in dataset D with respect to the protected attributes and the redlining attributes. *Belift* calculates essentially how many times the usage of protected and redlining attributes increase the chance of having a specific decision. Therefore, $belift = 1$ means that the protected and the redlining attributes do not increase the chances of having a specific decision for an instance (no violation of equal treatment).

We give an example application of the discrimination discovery algorithm in Fig. 1 to clarify how it works. This example simulates the case of car insurance premiums. Suppose that we have the hypothetical dataset for car insurance premiums in Table 1.

The predicted class label is Monthly_Premium. It is a binary class attribute with $\{<100, \geq 100\}$ values. The parameters provided to the discrimination discovery algorithm are given in Table 2. The implementation is done using the Weka data mining tool (Witten Ian and Frank 2011). SimpleEstimator is used to learn the contingency tables and the K2 algorithm is used to learn the network structure (Witten Ian and Frank 2011).

The discrimination discovery algorithm in Fig. 1 estimates the Bayesian network in Fig. 4 (The ID field is ignored during learning since it does not have predictive power). Figure 4 shows that there exists a strong correlation in this dataset between the Monthly_Premium attribute and the protected attribute gender since the nodes of both attributes have an edge between them. Furthermore, there also is a strong correlation between gender and occupation since there is an edge between these two

Table 1 Insurance premium dataset

Instance_ID	Accident_History	Car_Type	Occupation	Gender	Monthly_Premium
1	No accident	Saloon	Vet	Female	<100
2	No accident	Van	Admin	Female	≥100
3	No accident	Saloon	Vet	Female	≥100
4	No accident	Sports car	Admin	Female	≥100
5	Accident ≥1	Saloon	Vet	Female	≥100
6	No accident	Sports car	Engineer	Male	<100
7	No accident	Sports car	Doctor	Male	<100
8	No accident	Van	Doctor	Male	<100
9	Accident ≥1	Saloon	Doctor	Male	≥100
10	Accident ≥1	Sports car	Engineer	Male	≥100

Table 2 Discrimination discovery algorithm (cf. Fig. 1) parameters

Parameter name	Parameter value
p (protected attributes)	Gender
c (class value)	≥100
Threshold (belift threshold)	1.0 (perfect equality)
db (decision boundary)	0.5
D (dataset)	Insurance premium dataset (cf. Table 1)

attributes in Fig. 4. Thus, occupation is a redlining attribute that might cause indirect discrimination. The discrimination discovery algorithm (Fig. 1) creates the relative Bayesian network (Fig. 5) by removing the protected attribute gender and the redlining attribute occupation. Then, the algorithm starts passing through the insurance premium dataset identifying the instances violating the perfect equality point ($threshold = 1$).

$P(Monthly\ Premium \geq 100)$ is calculated for each instance of insurance premium dataset (Table 1) according to joint probability distributions in Fig. 4 and in Fig. 5. *Belift* value is derived using both former probabilities since the discrimination discovery is executed for the provided ≥100 class label (Table 2). Discrimination discovery algorithm identifies instances with IDs 2 and 4 as the discriminated instances violating the perfect equality point.

4 Non-discriminatory classification (discrimination prevention)

Given a discriminatory dataset D, the non-discriminatory classification problem is to learn a non-discriminatory classifier C using discriminatory dataset D. We propose in this section a non-discriminatory classification algorithm that aims to solve this problem (Fig. 6). The general framework of most prior work can be summarized as follows:

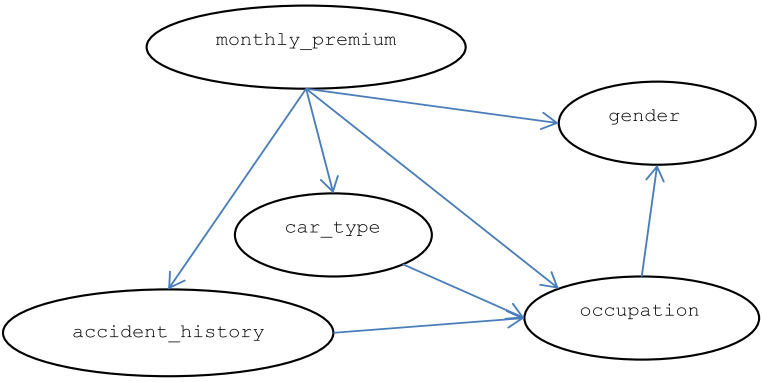

Fig. 4 Bayesian network of insurance premium dataset

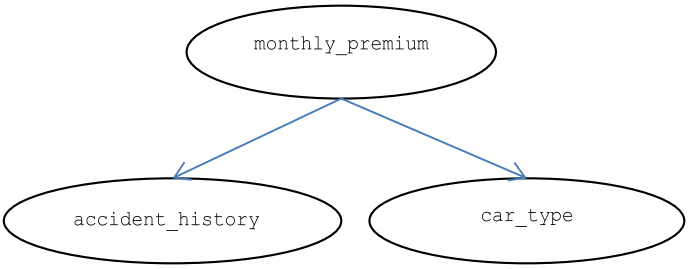

Fig. 5 Relative Bayesian network of instance premium dataset

Step 1: identify the discriminated instances in the discriminatory dataset D,
Step 2: correct the decision of the discriminated instances so that they turn into non-discriminated instances and eventually obtain a corrected discriminatory dataset $\overline{D}$ (preferential sampling, class relabeling, etc…),
Step 3: learn a non-discriminatory classifier C using the corrected discriminatory dataset $\overline{D}$.

The key difference of our work is in Steps 1 and 3. We combine both these steps by using the same Bayesian network structure (except for protected attribute) in the discovery and prediction phases. We do follow the class relabeling technique in step 2 as suggested by prior work (Calders and Verwer 2010; Zliobaite et al. 2011; Kamiran and Calders 2011; Mancuhan and Clifton 2012). Thus, the probability table updating cost is also trivial since it will only update the tables of the nodes that are adjacent to the class node in the Bayesian network.

The non-discriminatory classification algorithm assumes that the discrimination discovery algorithm in Fig. 1 is executed first. Thus, it receives the Bayesian network, discriminated instances and non-discriminated instances from the discovery algorithm. In addition, the protected attributes are also provided (as in discovery), because it is the chief source of discrimination. The discrimination

```
NonDiscClassification(DI, NDI, net, p)
Input:
    DI discriminated instances
    NDI non-discriminated instances
    net Bayesian network learnt from dataset D
    p protected attributes
Output:
    nd_net Non-discriminatory Bayesian network
begin:
nd_net<-net
nd_net<-removeProtectedAttributes(nd_net, p)
correctedDI<-flipClassLabels(DI)
D'<-concatenateDatasets(correctedDI, NDI)
nd_net<-updateProbabilityTables(nd_net, D')
return nd_net
end
```

Fig. 6 Non-discriminatory Bayesian network classifier algorithm

prevention method has the same assumptions about the protected attribute that the discrimination discovery method has (cf. Sect. 3).

We now explain the non-discriminatory classification algorithm in Fig. 6. The algorithm has five steps to make non-discriminatory predictions using discriminatory training data. First, it deletes the node of the protected attribute (with its edges) from the Bayesian network. The redlining attributes are not deleted from the Bayesian network, as they can be strong predictors for the decision process (giving a legitimate business reason for their use). Even though the objective is to learn a non-discriminatory classifier, achieving a high accuracy is also an important learning target. The redlining effect and its predictive power is a trade-off in the non-discriminatory classification problem that is already discussed in the discrimination-aware data mining literature by Calders and Verwer (2010).

Second, the classifier learning changes the label of the discriminatory instances identified by the discovery algorithm in Fig. 1, flipping their class label value to that assigned by the relative baseline Bayesian network (Sect. 3). This mitigates the effect of direct and indirect discrimination, thus mitigating the ability of the redlining attributes to perpetuate discrimination.

Third, the algorithm takes the union of the non-discriminated and corrected discriminated instances such that the corrected discriminatory dataset is obtained from the original discriminatory dataset.

Fourth, the probability tables of the Bayesian network are updated so that the class relabeling shows its effect in the model. Note that only the probability tables of the nodes that are adjacent to class node are updated; other partial distributions are intact from the modification.

Finally, the algorithm finishes the execution by returning the updated Bayesian network.

We extend the discovery example in Sect. 3 for the non-discriminatory Bayesian network classifier algorithm (Fig. 6). As mentioned, the second and forth instances of the insurance premium dataset (Table 1) are discriminated. So, the discrimination

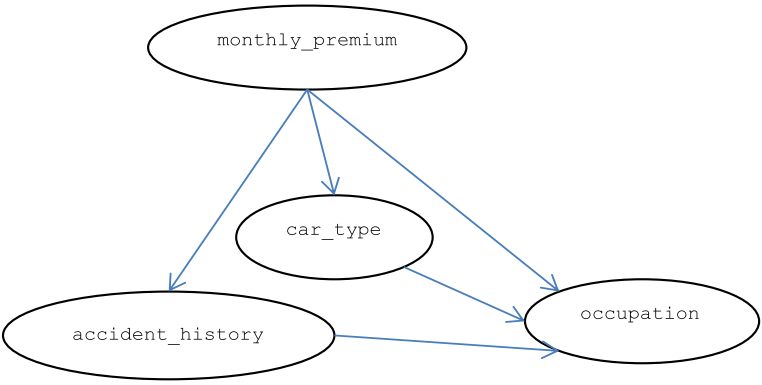

Fig. 7 Non-discriminatory Bayesian network classification model of insurance premium dataset

discovery algorithm (Fig. 1) returns the discriminated instances (instances with IDs 2 and 4), non-discriminated instances (instances with IDs 1, 3, and 5–10), and Bayesian network (Fig. 4). The non-discriminatory Bayesian network classifier algorithm (cf. Fig. 6) takes the three previous parameters and the protected attribute (gender) as the last parameter. It first deletes the node for the protected attribute gender and the structure of the non-discriminatory Bayesian network classifier is obtained (Fig. 7). Second, it flips the class label $\geq$100 of the discriminated instances to <100. Third, the algorithm takes the union of the corrected discriminated instances and the non-discriminated instances, to get the corrected discriminatory insurance premium dataset (Table 3). Finally, the probability tables of the network are updated according to the corrected discriminatory insurance premium dataset, and the updated non-discriminatory classification model is returned (Fig. 7).

5 Experiments

We start our discussion of experiments by presenting the datasets used and the experimental setup. We then present experiments first for discrimination discovery, then for discrimination prevention.

5.1 Experiment data and experimentation setup

We use the German Credit Dataset and the US Census Income Dataset from the UCI repository (Newman et al.) in our experiments. The US Census Income dataset is the same dataset that was used in Calders et al.'s paper (Calders and Verwer 2010). Both contain similar types of data, although the class attribute is significantly different (as we shall explain.)

The German Credit Dataset includes nominal (or discretized) attributes on personal properties: checking account status, duration, savings status, property

Table 3 Corrected discriminatory insurance premium dataset (corrections in italic form)

Instance_ID	Accident_History	Car_Type	Occupation	Gender	Monthly_Premium
1	No accident	Saloon	Vet	Female	<100
2	No accident	Van	Admin	Female	*<100*
3	No accident	Saloon	Vet	Female	≥100
4	No accident	Sports car	Admin	Female	*<100*
5	Accident ≥1	Saloon	Vet	Female	≥100
6	No accident	Sports car	Engineer	Male	<100
7	No accident	Sports car	Doctor	Male	<100
8	No accident	Van	Doctor	Male	<100
9	Accident ≥1	Saloon	Doctor	Male	≥100
10	Accident ≥1	Sports car	Engineer	Male	≥100

magnitude, type of housing; on previous/present credits and requested credit: credit history, credit request purpose, credit request amount, installment commitment, existing credit, other parties, other payment plan; on employment status: job type, employment since, number of dependents, own telephone; and on personal attributes: personal status and gender, age, residence since, and foreign worker. The class is if customer has been a good or bad credit risk, based on repayment history.

Similarly, the US Census Income Dataset includes nominal (or discretized) attributes on personal properties: education, number of years in education, capital gain, capital loss; on employment information: work class, occupation, hours per week; and on personal attributes: age, race, sex, relationship, marital status, native country. The class is high or low income (greater or less than \$50K). We ignore the final weight attribute as weighting the instances should really be considered a part of the learning process rather than as input.

We focus on gender discrimination in our experiments: The protected attributes are personal_status in the German Credit Dataset and sex in the US Census Data. The reason why we choose the personal_status as the protected attribute is that it has the gender information male and female in addition to marital status. We consider discrimination according to bad (credit default occurred) labeled instances in German credit dataset, and discrimination according to ≤50K (low income) labeled instances in US Census Data. The discrimination discovery experiments measure how being female affects an instance's credit risk score and income status respectively. The discrimination prevention experiments measure the extent to which historical discrimination towards females is reduced in classifier predictions where there is historical discrimination.

Because the credit risk measures actual repayment history (objective criterion) rather than a (potentially discriminatory) decision. We use the German credit dataset as a control dataset for our discrimination discovery approach. Significant discrimination is not expected in this dataset because it only has instances of people who were given credit (if any individuals were discriminated against, they would not have been given credit and thus would not appear in the data).

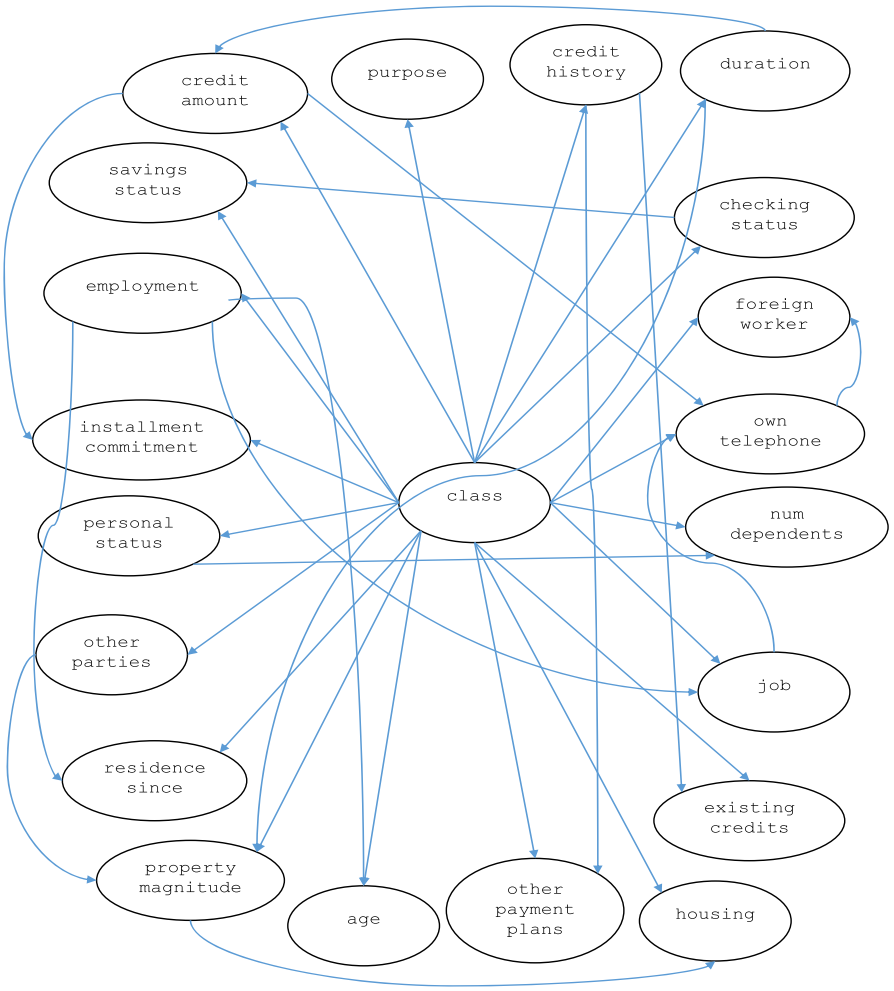

Fig. 8 German credit dataset Bayesian net

Furthermore, the class label is relatively objective. Our proposition thus would be incorrect if there were significant discrimination in this dataset.

Income is a different story. In the U.S., it is well known that females tend to have lower income than males, even when other factors are equal. As a sample of census data, it should be reflective of the overall population (as opposed to the German credit data, which only includes people who were extended credit.) As a result, we expect to find evidence of gender discrimination in this data.

In both the discovery and prevention experiments, the *belift* value has a threshold set to 1 which is the perfect equality point (Sect. 3). The decision boundary is 0.5 for the discovery experiments in both datasets. The datasets are discretized to ease probability calculation. Missing values are filled in with the mode of the attribute. Final results of discrimination prevention experiments are compared with the Latent

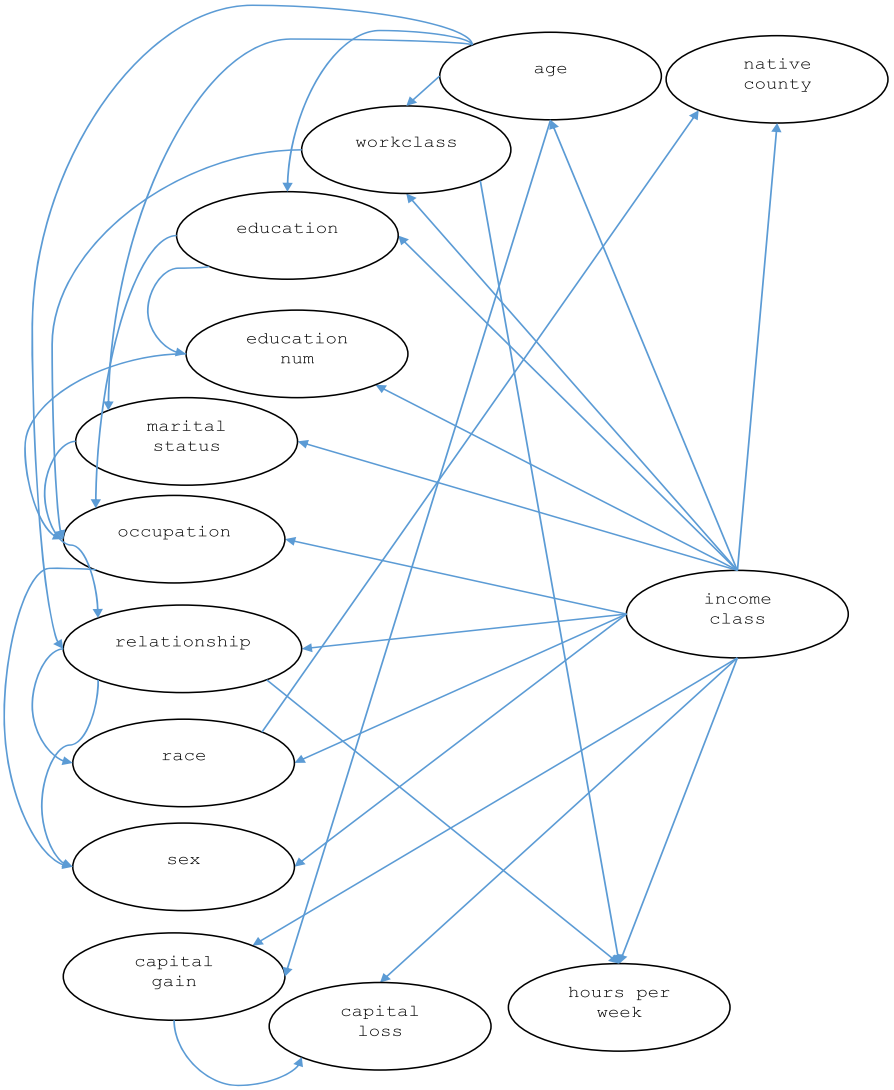

Fig. 9 US Census dataset Bayesian net

Variable Model results that are presented in Calders et al.'s work (Calders and Verwer 2010).

5.2 Discrimination discovery experiments

Our primary objective is to measure the number of discriminated instances in two datasets with respect to the discrimination discovery algorithm in Fig. 1. Our secondary objective is to determine if there is direct and indirect discrimination in

the analyzed dataset. We apply the discovery algorithm on two datasets to measure how decisions are affected for female individuals.

The Bayesian networks learned are given in Figs. 8 and 9. Our discrimination discovery algorithm identifies 10 discriminated instances among 1,000 instances in the German Credit dataset and 688 discriminated instances among 16,281 in US Census Income dataset. In terms of percentage, 1 % of the German Credit dataset is formed of discriminated instances whereas 4 % of the US Census Income dataset is formed of the discriminated instances. There is an even bigger difference in the significance of the discrimination, discussed below.

The Bayesian network for the German Credit dataset shows that there is indirect discrimination in the dataset with respect to attribute personal_status. There is an edge between the protected attribute node personal_status and non-class attribute num_dependents. Attribute num_dependents is the redlining attribute in this case. However, these observations might be related to a trivial fluctuation in the dataset. The *belift* distribution (Table 4) shows that the observed discrimination is insignificant. The *belift* values of the 10 discriminated instances range between 1.02 and 2.76. The quartile and median values show that the distribution is dense in the interval of (1.5,1.7) (with median 1.65). The discriminated instances in the German Credit dataset are very close to the decision boundary. This makes the measured discrimination insignificant and validates our discrimination discovery approach. It doesn't find significant discrimination in a dataset which has objective decisions.

The Bayesian network of US Census dataset show that there is indirect discrimination in the dataset with respect to attribute sex. This network structure exhibits that there are redlining attributes relationship, occupation, marital status; because the attributes relationship and occupation have an edge with protected attribute sex. The US Census dataset's *belift* distribution (Table 5) shows signs of significant discrimination. The distribution exhibits *belift* values of the 688 discriminated instances are between 1.03 and 20.2. The quartile and median values show that the *belift* distribution is dense in the interval of (1.03,2.6) (with median 2.02). There are a remarkable number of discriminated instances that are far from the decision boundary.

5.3 Discrimination prevention experiments

We test our non-discriminatory classification algorithm (Fig. 6) on US Census dataset, because there is significant discrimination in this dataset. (As expected, the discrimination in German Credit dataset was not significant.) Furthermore, the US

Table 4 Statistics for belift distribution in German credit dataset

Median	1.651010183
Max	2.76263988
Min	1.017787653
1st quartile	1.500872754
3rd quartile	1.700617184

Table 5 Statistics for belift distribution in US Census income dataset

Median	2.021246096
Max	20.18799216
Min	1.028860648
1st quartile	1.028860648
3rd quartile	2.601049449

Census dataset was used by Calders et al. (Calders and Verwer 2010). This lets us make accuracy and discrimination prevention comparison with their Latent Variable Models. The main success criterion is to reduce the number of discriminated instances in the predictions with minimal impact on classification accuracy.

We calculate the reduction in the number of discriminated instances using Eq. 6.

$$\frac{Number_{dc\_prediction}(DI) - Number_{ndc\_prediction}(DI)}{Number_{dc\_prediction}(DI)} \times 100 \tag{6}$$

Equation 6 calculates the percentage reduction of the discriminated instances (DIs) in the non-discriminatory classifier's (ndc) predictions relative to the discriminatory classifier's (dc) predictions. A positive percentage indicates that the non-discriminatory classifier's predictions have fewer DIs than the discriminatory classifier's predictions.

Tenfold cross validation is applied on the US Census dataset (as with Calders and Verwer 2010). The reduction percentage in the number of discriminated instances and the accuracy are calculated. The results are provided in Figs. 10 and 11.

The reduction in the number of discriminated instances is approximately 75 %, and above 70 % for nine of tenfolds (Fig. 10). The discriminatory classifier's average accuracy is around 0.836 while the non-discriminatory classifier's average accuracy is around 0.830 (Fig. 11); the worst case is only a 2 % reduction in accuracy. We would expect to see a slight decrease in accuracy, as the test data is based on data containing evidence of discrimination (which we are trying to correct for.) In other words, accuracy on the test dataset reflects our ability to predict the past, rather than give good decisions for a (non-discriminatory) future. With the U.S. census income dataset, accuracy is measured relative to historical gender discrimination. The accuracy decrease may well be that the non-discriminatory classifier is failing to predict discrimination, rather than failure to accurately predict income potential.

We calculate the amount of discrimination on predictions using the Eq. 7.

$$\frac{number\ of\ DIs\ in\ predictions}{test\ set\ size} \tag{7}$$

This ratio scales the number of discriminated instances (DI) in predictions over test set size. This lets us make a quantitative comparison with Calders et al.'s latent variable approach (Calders and Verwer 2010). Calders et al. defines the discrimination in terms of class disparity over test set size whereas this paper defines the discrimination in terms of instances having decision probability changes (DIs) over test set size. This proposition and latent variable approach target to measure 0 discrimination according to respective discrimination definitions while having

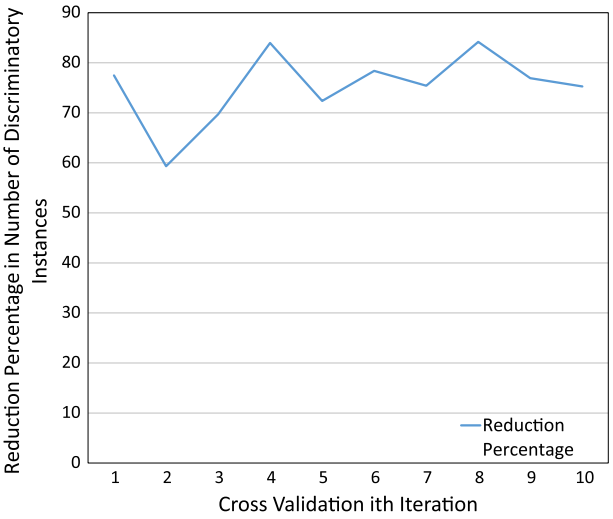

Fig. 10 Reduction percentage in number of discriminated instances

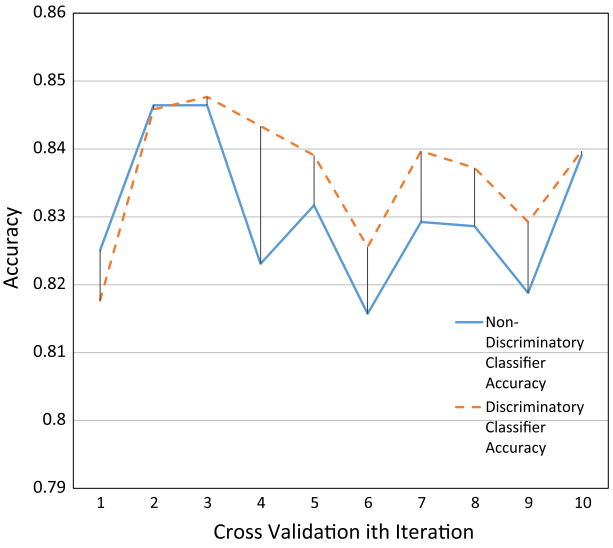

Fig. 11 Accuracies

minimum accuracy decrease. We make a comparison between this paper's proposition and latent variable approach so as to see how well the propositions remove the respective discriminations. Tables 6 and 7 give average values of discrimination and accuracy across 10 test folds.

The latent variable approaches (LVA), which use the protected attribute for preventing discrimination, outperform our proposition (Table 6). However, our

Table 6 Bayesian network approach (BNA) versus latent variable approach (LVA) (Calders and Verwer 2010)

Approach (LVA protected attribute known)	Discrimination	Accuracy
LVA NB	−0.003	0.813
LVA 2 models	−0.003	0.812
LVA EM	0.000	0.773
LVA EM prior	0.013	0.790
LVA EM stopped	−0.006	0.797
LVA EM prior stopped	−0.001	0.801
BNA	0.013	0.830

Table 7 Bayesian network approach (BNA) versus latent variable approach (LVA) (Calders and Verwer 2010)

Approach (LVA protected attribute unknown)	Discrimination	Accuracy
LVA NB	0.286	0.818
LVA 2 models	0.047	0.807
LVA EM	0.081	0.739
LVA EM prior	0.077	0.765
LVA EM stopped	0.061	0.792
LVA EM prior stopped	0.063	0.793
BNA	0.013	0.830

proposition have the lowest amount of discrimination and highest accuracy when the protected attribute is not used to make a decision on an individual (Table 7). In other words, the Bayesian network approach outperforms LVA if the protected attribute is cannot be used in the prediction process. Under this assumption, the best latent variable approach (LVA 2 models in Table 7) has 3–4 times more discrimination than Bayesian network approach. In addition, the Bayesian network approach has a very small decrease in accuracies (Fig. 11). Its accuracy decrease is only 0.006 on average (relative to discriminatory Bayesian network model). The Bayesian network approach is also more accurate than all LVA models. These results show two drawbacks of LVA. Firstly, LVA remove almost all the discrimination if and only if the protected attribute is used. The LVA depend mostly on the usage of protected attribute, which is a sticky legal point. We motivated earlier the legal restrictions about this matter. Secondly, LVA gives up significantly more accuracy to achieve discrimination removal (although, as we have pointed out, some of this poor accuracy may be a failure to perpetuate discrimination rather than an actual poor decision.)

6 Conclusion

We have proposed a discrimination discovery and a non-discriminatory classification technique using Bayesian networks. The key point of the discrimination

discovery technique is that it models the overall decision process in a set of instances and captures both direct and indirect discrimination. The non-discriminatory classification method mitigates both direct and indirect discrimination since it is the extension of the discovery method detecting both types of discrimination. Experiments show that the non-discriminatory classification can substantially reduce the number of discriminated instances in the predictions, while still maintaining similar accuracy.

A significant goal in developing this approach was to ensure compliance with legal mandates. The non-discriminatory classification technique never uses the protected attribute of an individual to make a decision affecting that individual, a legal requirement made quite clear in E.U. council directives 76/207/EEC of 9 February 1976, 2000/43/EC of 29 June 2000, 2000/78/EC of 27 November 2000, 2004/113/EC of 13 December 2004 and E.U. Supreme Court's Test Achats Case 236/09, and the U.S. Equal Credit Opportunity Act. The discrimination discovery technique uses the protected attribute to identify victims suffering from unequal treatment in a given set of instances. Victim identification does not produce a decision about that individual, although detecting violations could be used for compensation or reparation (2004/113/EC). The classification technique also targets indirect discrimination, reducing disparate impact without impacting accuracy. This provides a method to satisfy the judicial doctrine of the effects test, resulting in an approach that is not only effective, but could be considered mandated under laws such as the U.S. Equal Credit Opportunity Act.

Acknowledgments We wish to thank Alysa C. Rollock, J.D., Vice President for Ethics and Compliance at Purdue University, for discussion and pointers to relevant U.S. law. We also wish to thank journal reviewers for their helpful comments.

References

Calders T, Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. Data Min J (special issue with selected papers from ECML/PKDD)

Cooper GF, Herskovits E (1991) A Bayesian method for the induction of probabilistic networks from data. Mach Learn BMIR-1991-0293

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: ITCS, pp 214–226

Hajian S, Domingo-Ferrer J (2012) A study on the impact of data anonymization on anti-discrimination. In: IEEE ICDM international workshop on discrimination and privacy-aware data mining

Hajian S, Monreale A, Pedreschi D, Domingo-Ferrer J, Giannotti F (2012) Injecting discrimination and privacy awareness into pattern discovery. In: IEEE ICDM international workshop on discrimination and privacy-aware data mining

Kamiran F, Calders T (2009) Classifying without discriminating. IEEE Press, New York

Kamiran F, Calders T (2011) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst

Kamiran F, Calders T, Pechenizkiy M (2010) Discrimination aware decision tree learning. In: Proceedings IEEE ICDM international conference on data mining

Kamiran F, Karim A, Zhang X (2012) Decision theory for discrimination-aware classification. In: IEEE international conference on data mining

Luong BT, Ruggieri S, Turini F (2011) k-NN as an implementation of situation testing for discrimination discovery and prevention. In: 17th ACM international conference on knowledge discovery and data mining (KDD 2011). ACM, pp 502–510

Mancuhan K, Clifton C (2012) Discriminatory decision policy aware classification. In: IEEE ICDM international workshop on discrimination and privacy-aware data mining

Newman DJ, Hettich S, Blake CL, Merz CJ UCI repository of machine learning databases, http://archive.ics.uci.edu/ml/

Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: KDD conference

Pedreschi D, Ruggieri S, Turini F (2009) Measuring discrimination in socially-sensitive decision records. In: 9th SIAM conference on data mining (SDM 2009). SIAM, pp 581–592

Romei A, Ruggieri S (2013) A multidisciplinary survey on discrimination analysis. Knowl Eng Rev 1–57

Ruggieri S, Pedreschi D, Turini F (2011) DCUBE: discrimination discovery in databases. In: ACM international conference on knowledge discovery and data mining (KDD 2011). ACM, pp 502–510

Tan P.-N, Steinbach M, Kumar V (2006) Introduction to data mining. Addison-Wesley, Reading, MA, pp 227-246

Witten IH, Frank E (2011) Data mining: practical machine learning tools and techniques. 3rd edn. Morgan Kaufmann, Los Altos, CA

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: ICML

Zliobaite I, Kamiran F, Calders T (2011) Handling conditional discrimination. In: Proceedings IEEE ICDM international conference on data mining