Università di Pisa
Dipartimento di Informatica

Technical Report: TR-11-04

k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention

Binh Thanh Luong Salvatore Ruggieri Franco Turini

February 20, 2011

Address: Largo B. Pontecorvo 3, 56127 Pisa, Italy. Tel: +39 050 2212700 Fax: +39 050 2212726

k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention

Binh Thanh Luong
Institute for Advanced Studies
Lucca, Italy
l.thanhbinh@imtlucca.it

Salvatore Ruggieri
University of Pisa
Pisa, Italy
ruggieri@di.unipi.it

Franco Turini
University of Pisa
Pisa, Italy
turini@di.unipi.it

ABSTRACT

With the support of the legally-grounded methodology of situation testing, we tackle the problems of discrimination discovery and prevention from a dataset of historical decisions by adopting a variant of k-NN classification. A tuple is labeled as discriminated if we can observe a significant difference of treatment among its neighbors belonging to a protected-by-law group and its neighbors not belonging to it. Discrimination discovery boils down to extracting a classification model from the labeled tuples. Discrimination prevention is tackled by changing the decision value for tuples labeled as discriminated before training a classifier. The approach of this paper overcomes legal weaknesses and technical limitations of existing proposals.

Keywords

Discrimination discovery and prevention, k-NN classification

1. INTRODUCTION

Decisions based on categorization or social sorting may be discriminatory, in the socially negative sense of the unfair or unequal treatment of people based on membership to a category, group or minority, without regard to individual characteristics. This problem is exacerbated by the fact that increasingly sophisticated data analysis and data mining techniques support knowledge discovery from human activity data, enabling the extraction of models, patterns, profiles, and rules of human behavior in support of (automated) decision making. Human right laws [2, 9, 23, 24] prohibit discrimination against protected groups on the grounds of race, color, religion, nationality, sex, marital status, age and pregnancy; and in a number of settings, including credit and insurance; sale, rental, and financing of housing; personnel selection and wages; access to public accommodations, education, nursing homes, adoptions, and health care. Since most of the decisions in the knowledge society era are taken on the basis of historical data, there is the need of developing models, methods and technologies for modelling the processes of discrimination analysis in order to discover and prevent discrimination phenomena. *Discrimination discovery* from data consists in the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records. The process of data analysis must then be supported by tools that exploit *legally-grounded* measures and reasonings. *Discrimination prevention* in data mining consists of extracting models (typically, classifiers) that do not lead to discriminatory decisions even if trained from a dataset containing them. In fact, mining from historical data may mean to discover traditional prejudices that are endemic in reality (*taste-based discrimination* [3]), or to discover patterns of lower performances, skills or capacities of protected-by-law groups (*statistical discrimination* [10], also called *rational racism*), and to assign to such practices the status of general rules, maybe unconsciously, as these rules can be deeply hidden within a data mining classifier.

Recent data mining proposals for discrimination discovery [20, 21] and prevention [5] have followed the legal principle of *under-representation*. Unfortunately, they suffer both from legal weaknesses, due to the use of aggregation measures over undifferentiated sets of people, and technical limitations, such as the restriction to nominal attributes and decisions and to local models of discrimination. In this paper, we overcome both the legal and the technical drawbacks. Our approach is inspired by the legal experimental procedure of *situation testing*, which looks for pairs of people with similar characteristics apart from membership to a protected-by-law group. We approximate situation testing by a variant of the k-nearest neighbor (k-NN) classification, which labels each tuple in a dataset as discriminated or not. Discrimination discovery then reduces to build a classifier providing a global description of the conditions where discrimination occurs. Discrimination prevention is tackled by a pre-processing step, changing the historical decision for tuples labeled as discriminated, before training a classifier.

This paper is organized as follows. In Sect. 2 we review the legal grounds of discrimination analysis, highlighting deficiencies of the existing data mining approaches. The legal methodology of situation testing is also presented. In Sect. 3, we recall the notion of distance functions and summarize the experimental setup. The approach based on k-NN is presented in Sect. 4, and applied in Sect. 5 to discrimination discovery, and in Sect. 6 to discrimination prevention. Finally, Sect. 7 summarizes the contributions of the paper.

2. DISCRIMINATION ANALYSIS

Discrimination analysis from data should build over the large body of existing legal and economic studies [7, 14, 18].

.

group	benefit: not granted	benefit: granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \qquad p_2 = c/n_2 \qquad p = m_1/n$$

$$\mathit{diff}(c) = p_1 - p_2 \qquad \mathit{slift}(c) = \frac{p_1}{p_2}$$

$$\mathit{elift}(c) = \frac{p_1}{p} \qquad \mathit{olift}(c) = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{ad}{cb}$$

Figure 1: Discrimination measures

In this section, we review the under-representation principle that has inspired previous data mining proposals, and the situation testing methodology, which provides the legal grounds for the approach proposed in this paper.

2.1 The Under-Representation Principle

Accordingly to laws, discrimination occurs when a group is treated "less favorably" [9, 23] than others, or such that "a higher proportion of people not in the group is able to comply" [2] to a qualifying criterium. A general principle is then to consider *group under-representation* in obtaining a benefit [7, 14] as a quantitative measure of discrimination against a protected-by-law group (briefly, a *protected* group). Consider a dataset of historical decisions about granting or not a benefit (e.g., a loan, a job, a wage increase). Let p_1 (resp., p_2) be the proportion of people in the protected group (resp., not in the protected group) that were not granted the benefit, and let p be the proportion of all people (both protected and not) that were not granted the benefit. Group under-representation can be measured as the difference $p_1 - p_2$, adopted in the U.K. legislation [23]; or as the ratio p_1/p_2, called the *selection lift* and adopted in the U.S. legislation [24] (d); or as one of the measures defined in Fig. 1 over a four-fold contingency table. Higher values of those measures denote higher under-representation of the protected group.

The under-representation principle has inspired the existing approaches for discrimination discovery and prevention. [20] proposes to extract classification rules of the form $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ to unveil subsets $\mathbf{B}$ of the dataset where the protected group $\mathbf{A}$ suffered from under-representation with respect to the decision $\mathbf{C}$. The approach is parametric to one of the measures in Fig. 1, and it is implemented on top of an Oracle database [21]. [5] investigates three approaches for preventing discriminatory predictions in a Naïve Bayes classifier. Discrimination is measured as the difference $p_1 - p_2$ calculated on the whole set of predictions over a test set.

Unfortunately, these approaches suffer from both legal weaknesses and technical limitations. On the legal side, measuring group under-representation by aggregated values over *undifferentiated* groups is opposable in a court of law. As an example, assume a high value for $p_1 - p_2$ in a dataset of historical decisions, with women as the protected group and job hiring as the benefit. Since p_1 and p_2 are aggregated values, they mix decisions for people that may be very different as per skills required for the job. A typical legal argument is that of *genuine occupational requirement*. For the example at hand, if the job requires a special driving licence, which most of the male applicants have and most of the female applicants do not have, then the non-hiring rates p_1 and p_2 cannot be adopted for comparing people with different, legally admissible, attributes.

On the technical side, the discrimination discovery approach [20] has two limitations. First, since it relies on frequent itemset mining, it deals with nominal attributes and nominal decisions only. Interval-scaled attributes (age, income) and decisions (loan rate, wage) must be discretized as a pre-processing step. Second, the result of the knowledge discovery process is a (possibly large) set of classification rules, which provide *local* niches of possible discrimination: a *global* description of who is discriminated and who is not is lacking. The discrimination prevention approach [5] shares the same limitation on nominal attributes. In addition, it considers discrimination at top level, i.e., $p_1 - p_2$ is controlled only for the whole set of decisions. However, discrimination may still occur in some subset, e.g., as when discrimination of a bank branch manager against a minority group remains hidden in the large set of decisions of the whole bank.

2.2 Situation Testing

In the legal field, *situation testing* is a systematic research procedure for creating controlled experiments analyzing decision maker's candid responses to applicant's personal characteristics. In situation testing, pairs of research assistants undergo the same kind of selection, for example they apply for the same job, they present themselves at the same night club, and so on. Within each pair, applicant characteristics likely to be related to the situation (characteristics related to a worker's productivity on the job in the first case, look, age and the like in the second case) are made equal by selecting, training, and credentialing testers to appear equally qualified for the activity. Simultaneously, membership to a protected group is experimentally manipulated by pairing testers who differ in membership – for example, a black and a white, a male and a female, and so on. Situation testing is being experimented worldwide as one of the tools that can assist victims to establish that discrimination may have occurred [4, 19].

In this paper, we intend to exploit the idea of situation testing just inverting the point of view: given past records of decisions taken in some context, for each member of the protected group with a negative decision outcome (someone who may claim to be a victim of discrimination) we look for testers with similar, legally admissible, characteristics, apart from being or not in the protected group. If we can observe significantly different decision outcomes between the testers of the protected group and the testers of the unprotected group, we can ascribe the negative decision to a bias against the protected group. Similarity is modelled via a distance function. Testers are searched among the k-nearest neighbors. And difference is measured by one of the functions from Fig. 1 calculated over the two sets of testers.

3. EXPERIMENTAL SETTING

3.1 Distance Functions

Let $\mathbf{r}$ and $\mathbf{s}$ be two tuples of n attributes. A *distance function* $d()$ measures the dissimilarity between $\mathbf{r}$ and $\mathbf{s}$. In general, $d(\mathbf{r}, \mathbf{s})$ is a non-negative real number, close to 0 when $\mathbf{r}$ and $\mathbf{s}$ are highly similar or "near" each others, and becoming larger the more they differ. We admit the unknown value in domains, syntactically represented by "?", to model missing tuple elements. In the following, we present the distance functions adopted in experiments for interval-scaled, nominal and ordinal domains. Notice that the theoretical foundations of our approach are parametric to the actual

dataset	size	No. of attributes		
		interval	nominal	ordinal
credit	1,000	6	12	3
credit-d	1,000	0	12	9
crimes	1,994	24	0	0
adult	48,842	5	8	1
census-income	299,285	7	32	1

Table 1: Datasets: size and attribute type

distance function adopted. Let i be an attribute index, and x, y two values in the domain of the ith attribute.

Interval-scaled values are first standardized using the z-score $z_i(x) = (x - m_i)/s_i$, where m_i is the mean value, and s_i is the mean absolute deviation. Distance between x, y is measured by the absolute difference of their z-scores: $d_i(x, y) = |z_i(x) - z_i(y)|$. We deal with unknown values by setting $|z_i(x) - z_i(y)| = 3$ if $x =?$ or $y =?$.

For nominal domains, distance is a binary function testing equality: $d_i(x, y) = 0$ if $x = y$, and $d_i(x, y) = 1$ otherwise. For unknown values, we set $d_i(x, y) = 1$ if $x =?$ or $y =?$.

Ordinal domains with ranked values $v_1, \ldots, v_M$ are first mapped into interval-scaled values $m_i(v_j) = (j - 1)/(M - 1)$. Distance is then computed by resorting to the absolute difference $d_i(x, y) = |m_i(x) - m_i(y)|$. We deal with unknown values by setting $d_i(x, y) = max\{m_i(x), 1 - m_i(x)\}$ if $x =?$ and $y \neq ?$, and vice-versa; and $d_i(x, y) = 1$ if $x = y =?$.

Finally, distance between tuples is defined as:

$$d(\mathbf{r}, \mathbf{s}) = \frac{\sum_{i=1}^{n} d_i(\mathbf{r}_i, \mathbf{s}_i)}{n}$$

For tuples of interval-scaled attributes only, it boils down to the Manhattan distance of z-scores (modulo division by a constant); and for tuples of nominal attributes only, it boils down to the percentage of mismatching attribute values.

3.2 Datasets

As a running example, we use the German credit dataset, which consists of 1000 records over bank account holders. It includes attributes on personal properties, past/current credits, employment status, and on personal status. The `class` attribute represents the good/bad creditor decision assigned to the bank account holder. Attributes classified by their types are the following.

Interval-scaled: duration, credit_amount, installment_commitment, age, existing_credits, num_dependents.

Nominal: credit_history, purpose, personal_status, other_parties, residence_since, property_magnitude, housing, job, other_payment_plans, own_telephone, foreign_worker, class.

Ordinal: checking_status, savings_status, employment.

We consider two versions of the dataset: *credit*, the original dataset, with interval-scaled attributes; and *credit-d*, where interval-scaled attributes are discretized into 5 bins of equal width, and the resulting attributes are assigned the ordinal type. In addition to the German credit, the larger datasets shown in Fig. 1 will be considered. Strictly speaking, they are not concerned with decisions, but rather with census information related to people income (*adult*, *census-income*) or to communities & crimes (*crimes*). Therefore, the objective of discrimination analysis on these datasets is to discover or prevent forms of statistical discrimination. All datasets are publicly available from [12]. Notice that the decision attribute is binary for all of them, apart for *crimes* where it is interval-scaled being the "total number of violent crimes per 100K population".

4. K-NN AS SITUATION TESTING

4.1 k-NN

k-nearest neighbor (k-NN) is a lazy instance-based classification method. The classification model simply consists of storing the training set. Given a tuple $\mathbf{r}$ to be classified, k-NN: (1) first searches the k tuples in the training set closest to $\mathbf{r}$ w.r.t. a distance measure $d()$, i.e., its k-nearest neighbors; (2) then assignes as class value to $\mathbf{r}$ the most frequent class value among its k-nearest neighbors.

Throughout the paper, we represent a dataset as a collection $\mathcal{R}$ of tuples with a superscript, e.g., as in $\mathbf{r}^i$ where $i \in [1, |\mathcal{R}|]$ is the tuple id. For a tuple $\mathbf{r}$, we assign to every $\mathbf{r}^i \in \mathcal{R}$ a rank (as a neighbor of $\mathbf{r}$) on the basis of its distance from $\mathbf{r}$ (or, for equal distances, on the tuple id). Formally, we define:

$$rank_{\mathcal{R}}(\mathbf{r}, \mathbf{r}^i) = |\{j \mid d(\mathbf{r}, \mathbf{r}^j) < d(\mathbf{r}, \mathbf{r}^i) \vee (d(\mathbf{r}, \mathbf{r}^j) = d(\mathbf{r}, \mathbf{r}^i) \wedge j \leq i\}|$$

The k-NN set for a given tuple is the set of top k tuples w.r.t. ranking. A refined version may include an additional constraint on the maximum allowable distance.

Definition 4.1. *For a dataset $\mathcal{R}$, we define:*

$$kset_{\mathcal{R}}(\mathbf{r}, k) = \{\mathbf{r}^i \in \mathcal{R} \mid rank_{\mathcal{R}}(\mathbf{r}, \mathbf{r}^i) \leq k\}$$
$$kset_{\mathcal{R}}(\mathbf{r}, k, d) = \{\mathbf{r}^i \in \mathcal{R} \mid rank_{\mathcal{R}}(\mathbf{r}, \mathbf{r}^i) \leq k \wedge d(\mathbf{r}, \mathbf{r}^i) \leq d\}$$

In k-NN classification, $\mathbf{r}$ is the tuple to be classified, and $\mathcal{R}$ is the training set. In our discrimination analysis context, we take a different approach – which is closer to the situation testing methodology. First, we need fixing some inputs.

4.2 Inputs of the Discrimination Analysis

In addition to a dataset $\mathcal{R}$, our analysis of discrimination, either for discovery or for prevention, will require the following inputs: the group under analysis, a distance function over a set of attributes, and the decision attribute.

Protected-by-law groups. Civil rights laws explicitly identify the groups to be protected against discrimination, e.g., women or black people. We assume that a protected-by-law group is specified as input of the discrimination analysis, and call it the protected group. Also, we assume that the dataset under analysis includes attributes to decide whether a tuple refers to a member of the group. Syntactically, we model this by a predicate *protected*. In the previous example, we require then that sex (resp., race) is an attribute of data, and set *protected*($\mathbf{r}$) iff $\mathbf{r}[\texttt{sex}] = \texttt{female}$ (resp., $\mathbf{r}[\texttt{race}] = \texttt{black}$).

We do not impose any syntactic restriction on the definition of *protected*: for instance, in our prototype implementation, it can be any boolean expression over comparison operators $<, \leq, =, \neq, \geq, >$ for interval-scaled and ordinal attributes; and over $=, \neq$ for nominal attributes.

Using *protected*, we can separate tuples of people in the protected group from those of people not in the protected group – which we call the *unprotected* group.

Definition 4.2. *We define $P(\mathcal{R}) = \{\mathbf{r}^i \in \mathcal{R} \mid protected(\mathbf{r}^i)\}$, and $U(\mathcal{R}) = \{\mathbf{r}^i \in \mathcal{R} \mid \neg protected(\mathbf{r}^i)\}$.*

Notice that $\mathcal{R} = P(\mathcal{R}) \cup U(\mathcal{R})$ does not necessarily hold, since tuples with unknown values to be tested by *protected* are not included in $P(\mathcal{R})$ nor in $U(\mathcal{R})$, e.g., as for $\mathbf{r}[\texttt{sex}] =?$ in the previous example.

Legally-grounded attributes for distance measurement. We assume a distance function $d()$ between tuples. As in k-NN classification, distance will be used to search for neighbors of a given tuple. As in situation testing, such neighbors are searched with reference to attributes that are legally-grounded for being adopted in taking the decision. Therefore, we assume that $d()$ is defined on a subset of the attributes of the dataset, e.g., those that are legally admissible in a discrimination litigation. Additional attributes may be present in the dataset for other purposes, e.g., as shown in Sect. 5, for extracting a description of cases where discrimination occurs. Let $G \subseteq \{1, \ldots, n\}$ be the set of attribute indexes (or, equivalently, attribute names) that are legally-grounded. We write $\pi_G(\mathbf{r})$ to denote the projection of tuple $\mathbf{r}$ over attribute indexes in G, e.g., $\pi_{\{1,3\}}(\langle 3, 5, 4\rangle) = \langle 3, 4\rangle$. We make the following syntactic assumption, which helps in taking notation simple.

REMARK 4.3. *Distance is computed with reference to attributes indexes in G. In symbols, when writing $d(\mathbf{r}, \mathbf{s})$ we actually restrict to consider $d(\pi_G(\mathbf{r}), \pi_G(\mathbf{s}))$.*

In particular, when writing $kset_{\mathcal{R}}(\mathbf{r}, k)$ we now intend the k-NN set w.r.t. distance over attribute indexes in G.

Decision attribute. Obviously, the dataset $\mathcal{R}$ includes an attribute recording the historical decision. We write $dec(\mathbf{r})$ to denote the value for $\mathbf{r}$ of the decision attribute. Such an attribute is sometimes called the *class* attribute, because classifiers built on the dataset try and learn it as the class. As such, the decision attribute does not include unknown values. We admit nominal, interval-scaled and ordinal decision attributes. For nominal (typically binary, i.e., two-valued) decision attributes, we denote by $\ominus$ the negative decision (deny of benefit), and by $\oplus$ the positive decision (grant of benefit). Examples of interval-scaled decision attributes include wage in a dataset of personnel, and interest rate in a dataset of loans. In the rest of the paper, we will first state definitions and results for nominal decision attributes, and then describe what differs for interval-scaled ones. For lack of space, ordinal decision attributes are not explicitly commented. Basically, they are treated as interval-scaled attributes once ordinal values are mapped into interval-scaled ones using the standard mapping recalled in Sect. 3.1.

4.3 k-NN as Situation Testing

Consider a tuple $\mathbf{r}$ in the protected group $P(\mathcal{R})$. In the words of situation testing, a person feels discriminated if she observes significantly different decision outcomes for people who are similar to her apart from belonging or not to the protected group. Let us formalize this as follows. Let $\mathcal{K}_1$ be the set of k-nearest neighbors of $\mathbf{r}$ in $P(\mathcal{R}) \setminus \{\mathbf{r}\}$, where $\mathbf{r}$ has been removed in order not to be included in $\mathcal{K}_1$. Also, let $\mathcal{K}_2$ be the set of k-nearest neighbors of $\mathbf{r}$ in $U(\mathcal{R})$. $\mathcal{K}_1$ and $\mathcal{K}_2$ represent persons with attributes close to the ones of $\mathbf{r}$ apart for being in the protected group or in the unprotected group respectively. Notice that the distance function is defined on attributes that are legally-grounded, i.e., legally admissible for being adopted in taking the decision of granting or not a benefit. For a nominal decision attribute, the probability of the decision outcome for $\mathbf{r}$ can be estimated from $\mathcal{K}_1$ (resp., $\mathcal{K}_2$) as the proportion p_1 (resp., p_2) of tuples whose decision value is the same of $\mathbf{r}$. The difference between the observed values p_1 and p_2 represents then the bias of the decision for $\mathbf{r}$ due to membership to the protected group. We measure such a difference as $p_1 - p_2$, although any discrimination measure from Fig. 1 can be adopted. Let us provide a formal definition.

DEFINITION 4.4. *For $\mathbf{r} \in P(\mathcal{R})$, we define $diff(\mathbf{r}) = p_1 - p_2$, where: $p_1 = |\{\mathbf{r}' \in kset_{P(\mathcal{R})\setminus\{\mathbf{r}\}}(\mathbf{r}, k) \mid dec(\mathbf{r}') = dec(\mathbf{r})\}|/k$ and $p_2 = |\{\mathbf{r}' \in kset_{U(\mathcal{R})}(\mathbf{r}, k) \mid dec(\mathbf{r}') = dec(\mathbf{r})\}|/k$.*

Assume that a negative decision is assigned to $\mathbf{r}$, namely $dec(\mathbf{r}) = \ominus$. A value $diff(\mathbf{r}) = t \geq 0$ means that the decision is more frequent in the neighbors of the protected group with respect to the neighbors of the unprotected group by a percentage difference of t. This implies that the negative decision for $\mathbf{r}$ is not explainable on the basis of the legally-grounded attributes used for distance measurement, but rather it is biased by group membership. Whether the bias was intentional or not is irrelevant: laws also sanction *unintentional discrimination*. t is a measure of the strength of the discrimination bias. A value t lower than 0 means that the negative decision for $\mathbf{r}$ is not explainable by a worse treatment of the neighbors in the protected group. Hence, no discrimination conclusion can be drawn.

Conversely, assume a positive decision $dec(\mathbf{r}) = \oplus$. By reasoning as above, $diff(\mathbf{r}) \geq 0$ means a bias towards the positive decision due to group membership. This could be the result of *affirmative actions*, also called *positive actions* or *reverse discrimination*, consisting in a range of policies or quotas to overcome and to compensate for past and present discrimination [22]. Dually, $diff(\mathbf{r}) < 0$ means that the positive decision is not explainable by a general better treatment of the neighbors in the protected group.

EXAMPLE 4.5. Consider the *credit* dataset. We fix the protected group to non-single women by setting $protected(\mathbf{r})$ to $\mathbf{r}[\texttt{personal_status}] = \texttt{female div/sep/mar}$. Also, we set the decision attribute to `class`, namely to the credit classification, with $\ominus = \texttt{bad}$ and $\oplus = \texttt{good}$. All the remaining attributes are included, for the moment, in G, i.e., they are used in distance measurement. Fig. 2 (a) shows the cumulative distribution of $diff()$ for people in the protected group that have received the bad credit classification, namely for tuples $\mathbf{r}$ such that $protected(\mathbf{r})$ and $\mathbf{r}[\texttt{class}] = \texttt{bad}$. The plot shows the distributions for different values of k.

It is immediate to see that the distributions are shifted towards the positive half of the spectrum. More than 60% of the people have $diff(\mathbf{r}) \geq 0.1$, which means that bad-debtors are at least 10% more frequent among the k-most similar persons in the protected group than among the k-most similar persons in the unprotected group. Hence, the decision of classifying $\mathbf{r}$ as bad-debtor appears to be biased by her membership to the protected group.

The discrimination scenario becomes clearer if we consider the same distributions for people in the protected group having received the good-debtor decision, namely for tuples $\mathbf{r}$ such that $protected(\mathbf{r})$ and $\mathbf{r}[\texttt{class}] = \texttt{good}$. They are shown in Fig. 2 (b). Now the distributions are shifted towards the negative half of the spectrum. For $k = 16$, it turns out that $diff(\mathbf{r}) \geq 0.1$ in less than 7% of cases, which means that only a small percentage of the good-debtor classification could be attributed to a bias in favor of non-single women.

Finally, observe that the distributions in Fig. 2 (a,b) tend to shrink as k increases. This is intuitive, since for $k \to \infty$ the k-NN sets of protected and unprotected groups tend to include all the elements of the group, and then $diff(\mathbf{r}) \to \hat{p}_1 - \hat{p}_2$ where $\hat{p}_1$ (resp., $\hat{p}_2$) is the proportion of decision $dec(\mathbf{r})$ in the whole protected (resp., unprotected) group.

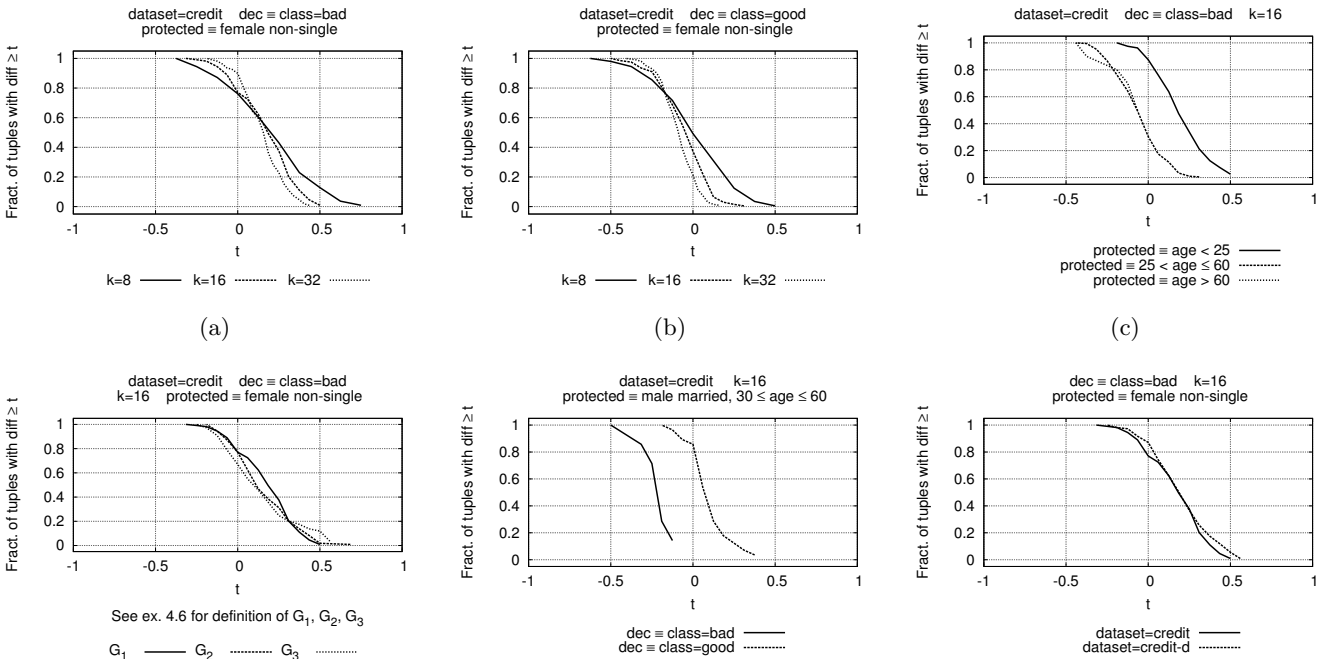

Figure 2: Cumulative distributions of *diff*() for datasets *credit* and *credit-d*

Example 4.6. It is worth studying how the distributions vary for different parameters in the previous example.

Fig. 2 (c) shows the cumulative distribution of *diff*() for protected groups defined on ranges of age. The plot clearly shows that youngers suffer from a higher bias towards the bad-debtor classification than middle-age or older people.

Fig. 2 (d) shows how the distribution varies with the set G of attributes used in distance measurement (see Remark 4.3). The plot refers to three sets. G_1 includes all attributes except `sex`, used to define the protected group, and `class`, used as the decision attribute. G_1 is the set used so far. G_2 includes attributes related to credit (history, purpose, amount, existing) and to properties (savings, property, housing, third parties) – but nothing about job. Finally, G_3 includes only attributes related to credit. From the plot, we can conclude that the distributions are not dramatically sensible to the set of attributes, although by restricting the set of attributes, high values of *diff*() tend to be even higher.

Fig. 2 (e) highlights the dual concept of *favoritism*, namely discrimination in favor of a group, where the "protected" group under analysis here consists of married men in the age range between 30 and 60. Compared to Fig. 2 (a,b), the distributions for the bad and the good decision are swapped, in the sense that there is no bias *against* the group in bad-debtor classification decisions, and there is bias *in favor* of the group members in good-credit classification decisions.

Finally, Fig. 2 (f) shows the distribution of *diff*() for the *credit-d* dataset, which is obtained by discretizing interval-scaled attributes in *credit*. Protected group, decision attribute and attributes in G are kept the same as in Fig. 2 (a). Technically, discretization affects the distance function, possibly resulting in different k-NN sets. The plot, however, closely resembles the distribution of the the original dataset, with a slight shift towards higher values.

Statistical significance of the *diff*() measure should also be taken into account, and it is actually customary in legal cases [13, 16]. In our context, we can interpret the observed proportions p_1 and p_2 from Def. 4.4 as the result of an experiment. What is the chance that the observed value $p_1 - p_2$ is affected by randomness in decisions of the dataset at hand? A confidence interval provides us with a range for the true value over the entire population (of decisions), at a certain significance level. Let us denote by π_1 and π_2 the true proportions over the protected and the unprotected neighbors. The Wald confidence interval for $\pi_1 - \pi_2$ at $100(1-\alpha)\%$ level of significance is $[(p_1 - p_2) - d, (p_1 - p_2) + d]$ where:

$$d_\alpha = Z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{k} + \frac{p_2(1-p_2)}{k}}.$$

We refer the reader to [1, 11] for details, and to [15] for a comparison of several methods for interval estimation. We mention here that when k is very low, the Wald interval becomes imprecise. Exact methods have been proposed in the statistics literature, where "exact" means that the actual discrete distribution of the statistical parameter is adopted in computing the confidence intervals, instead of approximating it with a normal distribution as done in Wald intervals. In our context, we extend the *diff*() function as follows:

$$\mathit{diff}(\mathbf{r}, \alpha) = \begin{cases} max\{0, p_1 - p_2 - d_\alpha\} & \text{if } p_1 - p_2 \geq 0 \\ min\{0, p_1 - p_2 + d_\alpha\} & \text{otherwise} \end{cases}$$

Intuitively, non-negative values of *diff*() are corrected to the lower bound of the confidence interval, and negative values are corrected to its upper bound. $\mathit{diff}(\mathbf{r}, \alpha) = 0$ when the null hypothesis $\pi_1 - \pi_2 = 0$ cannot be rejected.

Example 4.7. In the *adult* dataset the decision attribute `income` is not properly a "decision" but a range of income (lower or equal than \$50K , and higher than \$50K). Here,

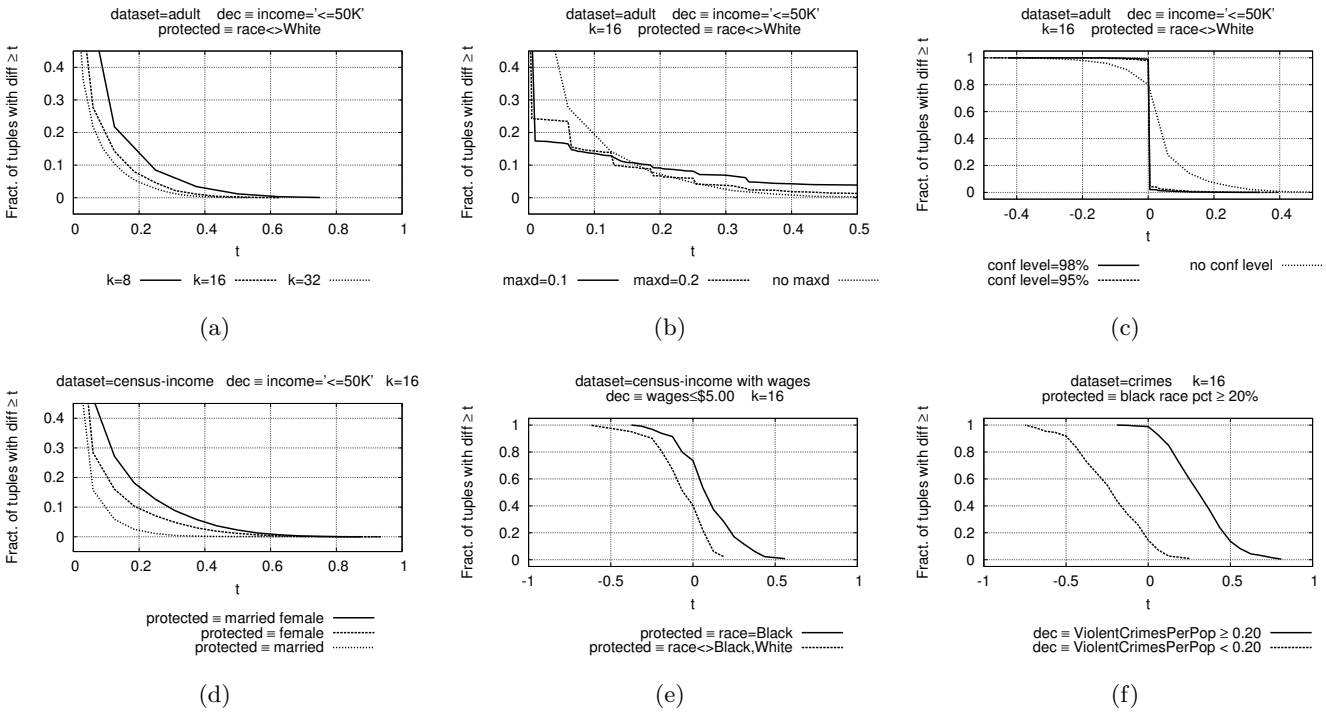

Figure 3: Cumulative distributions of *diff*() for datasets *adult*, *census-income* and *crimes*

the objective of the discrimination analysis is to discover whether people in a protected group suffer from low income not for their own specific characteristics but rather due to membership to the group – a form of the so-called *social* discrimination. Fig. 3 (a,b,c) show the distributions of *diff*() for non-white people with low income at the variation of three parameters of the analysis: (a) the number k of nearest neighbors; (b) the additional constraint *maxd* on maximum allowable distance of the nearest neighbors (see Def. 4.1); and (c) the level of significance in statistical validation. As one would expect, the distributions tend to shrink as k increases (already observed in Ex. 4.5) and as the confidence level increases. Fig. 3 (c) shows that very few cases could be brought in a court of law with the support of strong statistical arguments. Finally, putting a maximum distance threshold when looking for neighbors results in a longer right tail distribution, as shown in Fig. 3 (b). The maximum distance parameter affects those tuples $\mathbf{r}$ that are somehow isolated (ether because of their own attributes, or by effect of the *curse of dimensionality* on the distance function) by excluding "distant neighbors" from their k-NN set.

The dataset *census-income* also contains census data, with the same decision attribute as *adult*. However, it is larger both in the number of tuples and in the number of attributes. In Fig. 3 (d) we highlight the analysis of *multiple discrimination* [8], namely discrimination on the grounds of two or more factors. The plot shows the distributions of *diff*() for the low income class of three protected groups: women, married people, and married women. As it can be readily observed, society leads married women with low income to experience a higher difference from the unprotected group (not-married or men) than the difference experienced as women (w.r.t. men) or as married people (w.r.t. not married ones) alone.

The definition of *diff*() extends to interval-scaled decision attributes by setting:

$$\begin{aligned} p_1 &= |\{\mathbf{r}' \in kset_{P(\mathcal{R})\setminus\{\mathbf{r}\}}(\mathbf{r},k) \mid dec(\mathbf{r}') \geq dec(\mathbf{r})\}|/k \\ p_2 &= |\{\mathbf{r}' \in kset_{U(\mathcal{R})}(\mathbf{r},k) \mid dec(\mathbf{r}') \geq dec(\mathbf{r})\}|/k \end{aligned} \qquad (\star)$$

Assume that higher decision values mean more negative outcomes, e.g., as when the decision attribute is the interest rate to be paid for a loan or a mortgage. Intuitively, p_1 (resp., p_2) measures the proportion of the neighbors in the protected (resp., unprotected) group that have a higher decision value. A difference $diff(\mathbf{r}) = p_1 - p_2$ greater than 0 means a negative bias due to membership to the protected group. For the interest rate example above, $diff(\mathbf{r}) = 0.3$ means that there are 30% more neighbors in the unprotected group, compared to the protected group, with a granted interest rate better[1] than $dec(\mathbf{r})$. Finally, assume that *lower* decision values mean more negative outcomes, e.g., as when the decision attribute is the percentage of salary increase. In order to have that positive values of $diff(\mathbf{r}) = p_1 - p_2$ denote bias against the protected group, the comparison predicates in $(\star)$ must be changed from $\geq$ to $\leq$.

Example 4.8. The *census-income* dataset includes the attribute `wages` recording the wage per hour. We have selected the tuples with known values of `wages`. For the resulting dataset, we can set `wages` as the (interval-scaled) decision attribute and study the distribution of *diff*(). As for nominal decision attributes, we concentrate on negative decisions. For interval-scaled attributes, the analogous of $dec(\mathbf{r}) = \ominus$ is, for some threshold value t, $dec(\mathbf{r}) \geq t$ if higher decision values mean more negative outcomes, and $dec(\mathbf{r}) \leq t$ otherwise. Fig. 3 (e) shows the distribution for

[1]This statement can be better understood by observing that $diff(\mathbf{r}) = p_1 - p_2 = (1 - p_2) - (1 - p_1)$,

```
DiscoveryN(R, t) {
  L = ∅
  for r ∈ P(R) {
    if( dec(r) = ⊖ and
                diff(r) ≥ t )
      r[disc] = yes
    else
      r[disc] = no
    L = L ∪ {r}
  }
  build a classifier on L
}
```

```
DiscoveryI(R, l, t) {
  L = ∅
  for r ∈ P(R) {
    if( dec(r) ≥ l and
                diff(r) ≥ t )
      r[disc] = yes
    else
      r[disc] = no
    L = L ∪ {r}
  }
  build a classifier on L
}
```

Figure 4: Procedures for discrimination discovery

`wages` up to \$5.00 for the protected groups of black people, and of other minorities (non-black & non-white people). It is readily checked that blacks with low wages observe a bias towards their group that is higher than the bias observed by the other minorities.

Finally, the decision attribute in the *crimes* dataset is `ViolentCrimesPerPop`, the number of violent crimes in a community per 100K individuals. As already observed, it is not actually a "decision" taken by someone. Rather, here the interest is to understand whether some conclusions can be drawn from the dataset that relate high values of `ViolentCrimesPerPop` to the percentage of minorities, e.g., blacks, in the community. Such conclusions could drive forms of statistical discrimination against communities with large presence of black people. Fig. 3 (f) shows the distribution of *diff*() for the protected communities having 20% or more of blacks. Those of such communities that have a high number of crimes (`ViolentCrimesPerPop`≥ 0.20) observe a significant difference in crimes with communities that are similar to them apart from having less than 20% of blacks. In other words, a high number of crimes is "discriminatorily" present in communities with a high percentage of blacks! To further support this, Fig. 3 (f) shows the distribution for communities with a low number of crimes (`ViolentCrimesPerPop`< 0.20). Now, communities with 20% or more of blacks and a low number of crimes observe no bias for such low number of crimes due to the high presence of blacks.

5. DISCRIMINATION DISCOVERY

In this section, we devise an approach for discovering and characterizing discrimination. From the examples of the last section, it is clear that for a tuple $\mathbf{r}$ of the protected group *diff*$(\mathbf{r})$ measures the discrimination bias. By assuming a non-negative threshold value for *diff*$(\mathbf{r})$, we can label a tuple as discriminated or not.

Definition 5.1. *Let $t \in [0, 1]$ be a threshold value. We say that $\mathbf{r}$ is t-discriminated, and write $disc(t, \mathbf{r})$, if $dec(\mathbf{r}) = \ominus$ and $protected(\mathbf{r})$ and $diff(\mathbf{r}) \geq t$.*

Notice that we require that $dec(\mathbf{r}) = \ominus$, namely that the decision value for $\mathbf{r}$ is negative, in order to distinguish discrimination from favoritism, such as the one resulting from affirmative actions. In fact, if $dec(\mathbf{r}) = \oplus$ and $diff(\mathbf{r}) \geq t$, then $\mathbf{r}$ and its protected neighbors were granted a benefit in higher proportion than its unprotected neighbors. Also, we require that $protected(\mathbf{r})$ holds because we are interested in characterizing under which conditions a member of the protected group was discriminated or not. Obviously, also members of the unprotected group might be discriminated (again, for instance, because of affirmative actions), but labeling both protected and unprotected group members would mix different causes of discrimination.

How should t be chosen? The answer depends on the law. For instance, the U.K. legislation [23] for sex discrimination, sets $t = 0.05$, namely a 5% difference. The U.S. legislation [24] for employment discrimination, sets a threshold (known as the "four-fifths rule") of 1.25 for the measure of selection lift *slift*(). We take a general approach, and leave t as a parameter for the analyst. We are now in the position to provide a global description of the tuples that are discriminated by reducing the discrimination discovery problem to a classification problem.

Definition 5.2. *The t-labeled version of a dataset $\mathcal{R}$ is the dataset obtained: (1) by adding a binary attribute, called* `disc`*, assuming, for a tuple $\mathbf{r} \in \mathcal{R}$ the boolean value $disc(t, \mathbf{r})$; and (2) by restricting to tuples of the protected group only.*

Discrimination discovery reduces now to the extraction of an accurate classifier over a labeled version of $\mathcal{R}$ with `disc` as the class attribute. Accuracy will be evaluated with standard measures, e.g., precision and recall over the class value `disc=yes`. The intended use of the classifier is descriptive, namely to provide the analyst with a description of the conditions under which discrimination occurred. Classification models that can be interpreted clearly and easily should be adopted, such as decision trees and classification rules. The overall procedure for discrimination discovery over nominal decision attributes is shown in Fig. 4 as **DiscoveryN()**, where the t-labeled version of $\mathcal{R}$ is computed in $\mathcal{L}$.

Condition (2) in Def. 5.2 follows from the fact that for tuples not in the protected group $disc(t, \mathbf{r})$ is always false, so any classification model would derive assertions such as "if not $protected(\mathbf{r})$ then `disc=no`", assuming that the model is able to express the condition defining *protected*.

Example 5.3. Consider the *credit* dataset, with the settings of Fig. 2 (a) and $k = 32$. Using a C4.5 decision tree classifier [17], the procedure **DiscoveryN(***credit*, 0.10**)** yields the following model and evaluation measures[2].

```
num_dependents <= 1
|   credit_amount <= 2631: disc=yes (59.0/9.0)
|   credit_amount > 2631: disc=no (44.0/15.0)
num_dependents > 1: disc=no (6.0)

disc=yes:  Precision  0.847   Recall  0.769
```

A pair (x, y) at a leaf node means that x tuples reach the leaf, y of which are incorrectly classified. y is omitted if it is 0. Intuitively, the bad debtor class (the decision) is *discriminatorily* assigned to a female non-single (the protected group) when she has zero or one dependents and asks for a credit amount up to \$2,631. This is a concise, clear, and global characterization of the cases when discrimination occurred. It covers 77% of the protected group (recall), and it is accurate at 85 % (precision). On the same dataset, the RIPPER rule classifier [6] yields a slightly better recall.

```
(credit_amount >= 3190) => disc=no (39.0/12.0)
(installment_commitment <= 2) and (residence_since >= 3)
                                        => disc=no (10.0/2.0)
 => disc=yes (60.0/9.0)

disc=yes:  Precision  0.85   Recall  0.785
```

[2]Since the purpose of the classifier is descriptive, precision and recall are calculated over the training set. Notice that a predictive procedure simply consists of checking $disc(t, \mathbf{r})$.

```
PreventionN(𝒯, 𝒱, t) {
  𝒯′ = ∅
  for r ∈ 𝒯 {
    r′ = r
    if( dec(r) = ⊖ and
        protected(r) and
            diff(r) ≥ t )
      r′[dec] = ⊕
    𝒯′ = 𝒯′ ∪ {r′}
  }
  build classifiers on 𝒯 and 𝒯′
  compare them on 𝒱
}
```

```
PreventionI(𝒯, 𝒱, l, t) {
  𝒯′ = ∅
  for r ∈ 𝒯 {
    r′ = r
    if( dec(r) ≥ l and
        protected(r) and
            diff(r) ≥ t )
      r′[dec] = l − ϵ
    𝒯′ = 𝒯′ ∪ {r′}
  }
  build classifiers on 𝒯 and 𝒯′
  compare them on 𝒱
}
```

Figure 5: Procedures for discrimination prevention

The model can be read as follows. Discrimination occurs in all cases, except when the credit requested is at least \$3,190, or when there are up to 2 installment payments (short term loans) of a resident since at least 3 years. Of course, changing a parameter of the approach may yield different models. The following decision tree is obtained for $k = 8$.

```
num_dependents <= 1
|   installment_commitment <= 2
|   |   age <= 23: disc=no (9.0)
|   |   age > 23
|   |   |   employment = unemployed: disc=yes (1.0)
|   |   |   employment = <1: disc=no (8.0/1.0)
|   |   |   employment = 1<=X<4: disc=yes (8.0/1.0)
|   |   |   employment = 4<=X<7: disc=yes (1.0)
|   |   |   employment = >=7: disc=yes (4.0/2.0)
|   installment_commitment > 2: disc=yes (72.0/19.0)
num_dependents > 1: disc=no (6.0/1.0)

disc=yes:  Precision  0.744   Recall  0.970
```

Discrimination occurs when there are zero or one dependents and mid to long-term loans, or short term loans to applicants older than 23 that are either unemployed or employed since at least 1 year.

Finally, the overall procedure for discrimination discovery over interval-scaled decision attributes is shown in Fig. 4 as **DiscoveryI()**. A further parameter is now required, namely the threshold l such that $dec(\mathbf{r}) \geq l$ denotes the negative decision outcome (see Ex. 4.8). Notice that the class attribute `disc` in the labeled dataset $\mathcal{L}$ is still nominal, binary valued.

6. DISCRIMINATION PREVENTION

Starting from a dataset of historical decision records, classification models are typically extracted with the intent to learn and predict the decision $dec(\mathbf{r})$ (the class attribute) starting from the other attributes of a tuple $\mathbf{r}$. Preventing discrimination when training a classifier consists of balancing these two contrasting objectives: maximize accuracy of the extracted classification model; and minimize the number of predictions that are discriminatory. Within our framework, a prediction is discriminatory if the classified tuple is t-discriminatory, for some fixed threshold t. Why a classification model should be discriminatory? The main reason is due to *statistical discrimination*, namely the fact that the classification algorithm may recognize patterns in the data where, either directly or indirectly, the membership to a protected group is a proxy for lower performances (e.g., capacity to pay the loan installments). We now propose a simple pre-processing step, orthogonal to the classification algorithm, for tackling the discrimination prevention problem. The method consists of changing the decision value for tuples in the training set that are t-discriminated.

	No pre-processing		0.10-correction	
classifier	accuracy	0.10-discr.	accuracy	0.10-discr.
C4.5	85.60%	4.24%	84.94%	1.07%
Naïve Bayes	82.46%	4.06%	82.33%	2.23%
Logistic	85.28%	6.61%	84.70%	0.61%
RIPPER	84.42%	5.24%	83.98%	3.94%
PART	85.20%	12.62%	84.00%	2.3%

Table 2: Discrimination prevention on dataset *adult*

Definition 6.1. *The t-corrected version of a training set $\mathcal{T}$ is the dataset obtained by changing $dec(\mathbf{r})$ from $\ominus$ to $\oplus$ if $disc(t, \mathbf{r})$ holds.*

Given a protected group and a threshold t, in order to evaluate the effectiveness of the pre-processing method, we build two classifiers: one on the training set $\mathcal{T}$, and the other on its t-corrected version $\mathcal{T}'$. Both classifiers are evaluated on a test set $\mathcal{V}$ with respect to two measures: accuracy, and t-discrimination. Accuracy is measured as the percentage of correct predictions. t-discrimination is measured as follows. Consider the dataset $\mathcal{V}$ where the decision attribute is set to the prediction of the classifier. t-discrimination is the percentage of tuples $\mathbf{r}$ with $diff(\mathbf{r}) \geq t$ among the tuples in the protected group with negative decision. Graphically, we look at the value of the cumulative distribution of $diff()$ (such as those in Fig. 2, 3) for the x-axis equal to t. The overall procedure is shown in Fig. 5 as **PreventionN()**.

Example 6.2. Consider the dataset *adult* from Ex. 4.7, where the protected group consists of non-white people. By splitting the dataset into 2/3 training and 1/3 test sets, Table 2 shows accuracy and 0.10-discrimination for various types of classifiers, including decision trees (C4.5), Naïve Bayes, logistic regression, and rule induction (RIPPER and PART). If no pre-processing correction is performed, we can observe that the dataset of predictions exhibit a 0.10-discrimination measure ranging from 4.24% to 12.62%. Overall, C4.5 exhibits the best trade-off between accuracy and non-discrimination. If the training set is pre-processed by a 0.10-correction, all the classifiers have a modest decrease in accuracy and a significant gain in non-discrimination. The logistic regression model exhibits now the best trade-off.

Finally, the overall procedure for discrimination prevention over interval-scaled decision attributes is shown in Fig. 5 as **PreventionI()**. As for discrimination discovery, a threshold l, such that $dec(\mathbf{r}) \geq l$ denotes the negative decision outcome, is required as a further input. Notice that the change of the decision for t-discriminated tuples is now implemented by setting the decision value to a minimum $l - \epsilon$ that keeps the tuple below the limit of negative decisions.

7. CONCLUSIONS

We have modelled the discrimination analysis problems by a variant of k-NN classification that implements the legal methodology of situation testing. Major advancements over existing proposals consist in providing: a stronger legal ground, overcoming the weaknesses of aggregate measures over undifferentiated groups; a global description of who is discriminated and who is not in discrimination discovery; a discrimination prevention method that is independent from the classification model at hand; an approach admitting interval-scaled and ordinal attributes and decisions.

Download. The software implementing the discrimination discovery and prevention procedures is available at:

http://www.di.unipi.it/∼ruggieri.

8. REFERENCES

[1] A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2002.

[2] Australian Legislation. (a) Equal Opportunity Act – Victoria State, 2010, (b) Anti-Discrimination Act – Queensland State, 1991.

[3] G. S. Becker. *The Economics of Discrimination*. University of Chicago Press, 2nd edition, 1971.

[4] M. Bendick. Situation testing for employment discrimination in the United States of America. *Horizons Strategiques*, 5:17–39, 2007.

[5] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining & Knowledge Discovery*, 21(2):277–292, 2010.

[6] W. W. Cohen. Fast effective rule induction. In *Proc. of ICML 1995*, pages 115–123. Morgan Kaufmann, 1995.

[7] E. Ellis. *EU Anti-Discrimination Law*. Oxford University Press, 2005.

[8] ENAR. European Network Against Racism, Fact Sheet 33: Multiple Discrimination, 2007. http://www.enar-eu.org.

[9] European Union Legislation. (a) Race Equality Directive, 2000; (b) Employment Equality Directive, 2000; (c) Equal Treatment of Persons, 2009.

[10] H. Fang and A. Moro. Theories of statistical discrimination and affirmative action: A survey. In *Handbook of Social Economics, Vol 1B*. North-Holland, 2010.

[11] J. L. Fleiss, B. Levin, , and M. C. Paik. *Statistical Methods for Rates and Proportions*. Wiley, 2003.

[12] A. Frank and A. Asuncion. UCI machine learning repository, 2011. http://archive.ics.uci.edu/ml.

[13] J. L. Gastwirth. Statistical reasoning in the legal setting. *The American Statistician*, 46(1):55–69, 1992.

[14] N. Lerner. *Group Rights and Discrimination in International Law*. Martinus Nijhoff Publishers, 1991.

[15] R. G. Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17:873–89, 1998.

[16] M. J. Piette and P. F. White. Approaches for dealing with small sample sizes in employment discrimination litigation. *Journal of Forensic Economics*, 12:43–56, 1999.

[17] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[18] W. M. Rodgers, editor. *Handbook on the Economics of Discrimination*. Edward Elgar Publishing, 2006.

[19] I. Rorive. *Proving Discrimination Cases - the Role of Situation Testing*. Centre For Equal Rights & Migration Policy Group, 2009.

[20] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Trans. on Knowledge Discovery from Data*, 4(2):Article 9, 2010.

[21] S. Ruggieri, D. Pedreschi, and F. Turini. DCUBE: Discrimination discovery in databases. In *Proc. of SIGMOD 2010*, pages 1127–1130. ACM, 2010.

[22] T. Sowell, editor. *Affirmative Action Around the World: An Empirical Analysis*. Yale University Press, 2005.

[23] U.K. Legislation. (a) Sex Discrimination Act, 1975, (b) Race Relation Act, 1976.

[24] U.S. Federal Legislation. (a) Equal Credit Opportunity Act, 1974; (b) Fair Housing Act, 1968; (c) Employment Act, 1967; (d) Equal Pay Act, 1963; (e) Pregnancy Discrimination Act, 1978; (f) Civil Right Act, 1964, 1991.