Discovering gender discrimination in project funding

Andrea Romei, Salvatore Ruggieri and Franco Turini
Dipartimento di Informatica, Università di Pisa
Largo B. Pontecorvo 3, 56127 Pisa, Italy
{romei,ruggieri,turini}@di.unipi.it

***Abstract*—The selection of projects for funding can hide discriminatory decisions. We present a case study investigating gender discrimination in a dataset of scientific research proposals submitted to an Italian national call. The method for the analysis relies on a data mining classification strategy that is inspired by a legal methodology for proving evidence of social discrimination against protected-by-law groups.**

I. Introduction

Social discrimination, and in particular gender discrimination, can be hidden in many situations. In the academic world, the processes that can be suspected of hiding discriminatory practices are hiring and promoting, paper rejection/acceptance in conferences and journals, and project funding. As for the last issue, a few studies from social sciences have provided no evidence of discrimination in grant awarding between male and female applicants (e.g., see [1]–[4]). However, such studies look for global differences in success, without considering the possibility that discrimination is not simply bound to the gender attribute in isolation. Instead, gender discrimination may pop up in specific contexts, or even in narrow niches. The discovery of such contexts is, on the contrary, one of the features of methodologies based on knowledge discovery.

This paper presents a case study on the discovery of gender discrimination out of a database of applications to a funding program for young researchers implemented by the Italian Ministry of Research and University. The method is rooted in the use of a data mining technique, a variant of k-NN classification designed in our previous work [5]. It allows us to simulate a well-known legal technique, *situation testing*, used to provide evidence of discrimination.

The organization of the paper is as follows. Section II presents some background on methods for discrimination discovery. Section III describes the case study and the data available for the analysis. Section IV reports on the design and on the findings of the data mining experiments. Comparison with related work from social sciences is reported in Section V. Finally, Section VI summarizes the contribution of the paper and opens directions for future work.

II. Background

In this section, we briefly review recent approaches that use data mining for discrimination discovery and discrimination prevention. Then, a deeper introduction of the data analysis technique of situation testing, and of its implementation as a data mining algorithm is presented. Related work from social and economic sciences will be discussed in Section V. A multi-disciplinary annotated bibliography on discrimination data analysis can be found in [6].

	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \quad p_2 = c/n_2$$

$$RD = p_1 - p_2 \quad RR = \frac{p_1}{p_2} \quad RC = \frac{1-p_1}{1-p_2} \quad OR = \frac{RR}{RC} = \frac{a/b}{c/d}$$

Fig. 1: 4-fold contingency table and discrimination measures

A. Data mining for discrimination discovery

Discrimination discovery from data consists in the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records. The aim is to unveil contexts of possible discrimination on the basis of *legally-grounded* measures of the degree of discrimination suffered by protected-by-law groups (from now on, simply *protected* groups) in such contexts. The legal principle of under-representation has inspired existing approaches for discrimination discovery based on pattern mining. A common tool for statistical analysis is provided by a 2×2, or 4-fold, contingency table, as shown in Figure 1. Different outcomes between groups are measured in terms of the proportion of people in each group (p_1 for the protected group, and p_2 for the unprotected one) with a specific outcome (benefit denial). Differences and rates of those proportions are commonly adopted as the formal counterpart of group under-representation. They are known in statistics as *risk difference* (RD), or as *absolute risk reduction*; *risk ratio* or *relative risk* (RR); *relative chance* (RC), also known as *selection rate*; *odds ratio* (OR).

Starting from a dataset of historical decision records, [7], [8] propose to extract classification rules such as

RACE=BLACK, PURPOSE=NEW_CAR → CREDIT=NO

called potentially discriminatory rules, to unveil contexts (here, people asking for a loan to buy a new car) where the protected group (here, black people) suffered from under-representation

of the favorable decision or from over-representation of the unfavorable decision. The approach has been implemented on top of an Oracle database by relying on tools for frequent itemset mining [9], and extended in [10] to take into account statistical significance of the discrimination measures.

The main limitation of such an approach is that measuring group under-representation by aggregated values over *undifferentiated* groups results in no control of the characteristics of the protected group, versus, or as opposed to others in this context. As an example, assume a high value for $RD = p_1 - p_2$, 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. This results in an overly large number of PD rules that need to be further screened to remove explainable discrimination. [5] overcomes this limitation by exploiting the legal methodology of situation testing, which will be presented in detail in Sect. II-B.

Finally, we only mention the related research line of *discrimination prevention* in data mining, where the objective is to extract models (typically, classifiers) that trade off accuracy for non-discrimination. Recent works on discrimination prevention include [11]–[13].

B. Situation testing and k-NN

Situation testing follows a quasi-experimental approach to investigate for the presence of discrimination by controlling the factors that may influence decision outcomes. In a legal setting, pairs of research assistants, called *testers*, 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. Observing significant difference in the selection outcome between testers is a *prima facie* evidence of discrimination, i.e., a proof that, unless rebutted, would be legally sufficient to prove the claim of discrimination. For reviews of the usage of situation testing, we refer to [14], covering employment discrimination in the U.S., and to [15], covering the E.U. member States context.

In [5], the idea of situation testing is exploited for discrimination discovery just inverting the point of view. Given past records of decisions taken in some context, for each member of the protected group with vector of attributes $\mathbf{r}$ suffering from a negative decision outcome (someone who may claim to be a victim of discrimination), we look for $2k$ testers with similar characteristics. Such characteristics are legally admissible in affecting the decision, apart the one of being or not in the protected group. Similarity is modelled via a distance function; here, we adhere to [5] by fixing the Manhattan distance of z-scores for continuous attributes in $\mathbf{r}$, and the percentage of mismatching attributes for discrete ones. An individual $\mathbf{r}$ is labeled as discriminated or not discriminated as follows: if we can observe significantly different decision outcomes between the k-nearest neighbors of $\mathbf{r}$ belonging to the protected group and the k-nearest neighbors belonging to the unprotected group, we can ascribe the negative decision to a bias against the protected group, hence labeling the individual $\mathbf{r}$ as discriminated. This approach resembles the k-nearest neighbor (k-NN) classification model, where the class of an individual is predicted as the most frequent class among its k-nearest neighbors. Difference in decision outcomes between the two groups of neighbors is measured by any of the functions from Figure 1, calculated over the proportions for the two sets of testers. Here, we fix risk difference $\mathit{diff}(\mathbf{r}) = p_1 - p_2$ with the intuitive reading that it represents the difference in the frequency p_1 of negative decisions in the neighbors of the protected group with respect to frequency p_2 in the neighbors of the unprotected group. A value $\mathit{diff}(\mathbf{r}) > 0$ 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[1]. $\mathit{diff}(\mathbf{r})$ is a measure of the strength of the discrimination bias. By fixing a threshold t, individuals $\mathbf{r}$ of the protected group can be labeled as discriminated or not on the basis of the condition $\mathit{diff}(\mathbf{r}) \geq t$. The problem of describing who was discriminated boils down then to a standard classification problem.

III. Problem and data understanding

This section reports the description of the case study, and some data exploration and data preprocessing steps of the discrimination analysis.

A. The FIRB "Future in research" call

In 2008, the Italian Ministry of University and Research published a call for scientific research projects under the Basic Research Investment Fund (FIRB) reserved to young scientists – the *FIRB "Future in research"* call. The scientific scope of the call is very broad, ranging from social sciences and humanities, to physical sciences and engineering, to life sciences. Research proposals are submitted by a consortium of one or more research units, with a *principal investigator* (from now on, PI) and zero or more *associate investigators* heading each unit and affiliated to an Italian university or to a public research organization. Research proposals are distinguished in two programs, depending on whether the PI holds a non-tenured position and she is at most 33 years old at the time of the call (program P_1), or she holds a tenured position and she is at most 39 years old (program P_2). Each program has its own total budget, but the submission forms and the evaluation processes are the same for both of them. Electronic proposal

[1] In this sense, "discrimination is the remaining racial [more generally, group] difference after statistically accounting for all other race-related [group-related] influences on the outcome" [16]. However, it is difficult to know that all important characteristics of individuals have been taken into account: a recurring problem known as *the omitted-variable bias*.

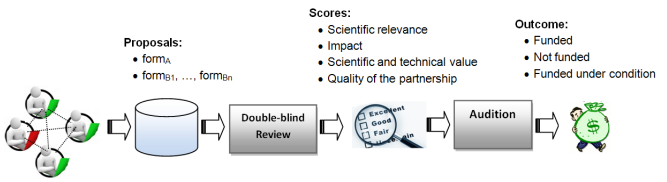


Fig. 2: A two-steps review process.

submissions consist of a description of work and a budget for each research unit, called the *B forms*, and of a description of work and a budget for the whole proposal, called the *A form*. The global budget is basically the sum of the budgets of the research units participating in the project. The A form also contains the curriculum vitae of the PI, a list of her main publications, and an abstract of the proposal.

Project proposals are subject to a two-steps evaluation process, as shown in Figure 2. The first step consists in a double-blind peer-review by national and international reviewers resulting in four scores about:

(S1) scientific relevance of the proposal (score 0 to 8);
(S2) impact of the proposal (score 0 to 7);
(S3) scientific and technical value of the proposal (score 0 to 15);
(S4) quality of the partnership (score 0 to 10).

Only project proposals that receive the best score in all of the four evaluation criteria (i.e., a total score of 40) are admitted to the second step, which consists in an audition of the PI in front of a panel of national experts. The panel ranks the proposals into three classes: "to be funded", "to be funded if additional budget were available"[2] and "not to be funded". For the first two classes, the panel also decides a budget cut with respect to the budget required in the project proposal.

B. Data preparation and exploration

Anonymized data on project proposals and evaluation results were made available to us as an Oracle relational database, for a total of 15 tables. The data understanding and preparation phases consisted in deriving a single table for the analysis, including both original and derived features. For each submitted proposal, we constructed four groups of features, summarized in Table I.

Features on the principal and associate investigators. These include gender, age, and title of the PI; number of publications and average number of authors in publications of the PI; region, city and type of her institution; and number of female (principal or associate) investigators in the project proposal.

Project costs. Several costs are considered: total cost of the project, requested grant (both absolute and relative), number and cost of young researchers, number and cost of good reputation researchers. At least one young researcher (i.e., a post-doc or a post-degree of at most 32) per project proposal is required by the call. Invitation of good reputation researchers from abroad to spend some period working on the project is, instead, an option.

[2]At that time, an increase of the budget of the call was under consideration by the Ministry. Unfortunately, the final decision was not to increase it. In the rest of the paper, we treat proposals ranked "to be funded if additional budget were available" as if they did not pass the second step of reviewing.

peer-review P_1

applic.	rejected	passed	
female	892	31	923
male	838	43	881
	1730	74	1804

$p_1 = 892/923 = 0.966$
$p_2 = 838/881 = 0.951$
$RD = 0.015$ $RR = 1.02$
$RC = 0.69$ $OR = 1.48$

audition P_1

	rejected	passed	
	14	17	31
	18	25	43
	32	42	74

$p_1 = 14/31 = 0.452$
$p_2 = 18/43 = 0.419$
$RD = 0.033$ $RR = 1.08$
$RC = 0.94$ $OR = 1.15$

peer-review P_2

applic.	rejected	passed	
female	761	31	792
male	1094	100	1194
	1855	131	1986

$p_1 = 761/792 = 0.961$
$p_2 = 1094/1194 = 0.916$
$RD = 0.045$ $RR = 1.05$
$RC = 0.46$ $OR = 2.24$

audition P_2

	rejected	passed	
	19	12	31
	49	51	100
	68	63	131

$p_1 = 19/31 = 0.613$
$p_2 = 49/100 = 0.49$
$RD = 0.123$ $RR = 1.25$
$RC = 0.76$ $OR = 1.65$

Fig. 3: 4-fold contingency tables and discrimination measures.

Research area. In addition to the research program a project proposal is submitted to, the proposal includes up to three research domains, using the European Research Council (ERC) classification. Such a classification consists of a three-level hierarchy. The top level includes Social sciences and Humanities (SH), Physical sciences and Engineering (PE), and Life Sciences (LS). The second and third levels include 25 and 3,792 sub-categories respectively.

Project evaluation. The following attributes are included: the scores (S1)-(S4) received at the peer-review, whether the project passed the first evaluation phase (i.e., the peer-review), whether the project passed the second evaluation phase (i.e., the audition), the actual amount granted after budget cut.

Let us summarize some aggregations and statistics on the input dataset. Actually, since research proposals of programs P_1 and P_2 are evaluated in isolation (due to the distinct budget of each program), we act as if there are two datasets, one per program. Program P_1 received 1804 submissions, 923 of which are from female PIs; program P_2 received 1986 ones, 792 of which from female PIs. Table II summarizes the proportion of genders in the two phases of the evaluation process: peer-review and audition. It is readily checked that, for both programs, the proportion of females passing the peer-review (resp., having the project funded at the second step) decreases compared to the proportion of female applicants (resp., passing the peer-review at the first steps).

We can quantify such a decrease by resorting to the discrimination measures from Figure 1. The 4-fold contingency tables of passing peer-review and of having the project funded for programs P_1 and P_2 are shown in Figure 3. Consider first the peer-review phase. Recall that the measures of risk difference (RD) and risk ratio (RR) compare the proportions of rejected proposals. Due to the small fraction of projects

Name	Description	Type	Range/Nominal values	Mean/Mode
Features on the principal and associate investigators				
gender	Gender of principal investigator (PI)	Nominal	{Male, Female}	Male
region	Region of the institution of the PI	Nominal	{North, Center, South}	Center
city	City of the institution of the PI	Nominal	{Aosta, Aquila, . . . , Trento}	Rome
inst_type	Type of the institution of the PI	Nominal	{Univ, Consortium, Other}	Univ
title	Title of the PI	Nominal	{Researcher, Prof., Other, PhD}	PhD
age	Age of the PI	Numeric	[26, 39]	32.8
pub_num	Number of publications of the PI	Numeric	[1, 156]	16.4
avg_aut	Average number of authors in publications of the PI	Numeric	[1, 87.1]	4.8
f_partner_num	Number of female principal or associate investigators	Numeric	[0, 3]	0.86
Project costs (absolute values are in €)				
tot_cost	Total cost of the project	Numeric	[300000, 2000000]	971792
fund_req	Requested grant	Numeric	[83720, 1260000]	506205
fund_req_perc	Percentage of requested grant over total cost	Numeric	[26, 63]	51.6
yr_num	Number of young researchers	Numeric	[1, 10]	2.1
yr_cost	Cost of young researchers	Numeric	[60000, 981261]	240557
yr_perc	Percentage of young researcher costs over total cost	Numeric	[3, 63]	25.5
grr_num	Number of International good repute researchers	Numeric	[0, 8]	1.5
grr_cost	Cost of good reputation researchers	Numeric	[0, 610000]	61863
grr_perc	Percentage of good reputation researchers cost	Numeric	[0, 35]	6.1
Research area				
program	Program the project was submitted to	Nominal	{P1, P2}	P2
d1_lv1, d2_lv1, d3_lv1	1^{st}, 2^{nd} and 3^{rd} domain at the 1^{st} level of the ERC hierarchy	Nominal	{LS, SH, PE}	PE
d1_lv2, d2_lv2, d3_lv2	1^{st}, 2^{nd} and 3^{rd} domain at the 2^{nd} level of the ERC hierarchy	Nominal	{LS_1, LS_2, . . . , PE_8}	PE_6
d1_lv3, d2_lv3, d3_lv3	1^{st}, 2^{nd} and 3^{rd} domain at the 3^{rd} level of the ERC hierarchy	Nominal	{LS_1_1, LS_1_2, . . . , PE_8_15}	PE_6_17
Project evaluation				
s1	Scores S1 received at the peer-review	Numeric	[1, 8]	6.6
s2	Scores S2 received at the peer-review	Numeric	[1, 7]	5.7
s3	Scores S3 received at the peer-review	Numeric	[1, 15]	11.8
s4	Scores S4 received at the peer-review	Numeric	[1, 10]	8.1
peer-review	Whether the project passed the peer-review (1st evaluation step)	Nominal	{passed, rejected}	rejected
funded	Whether the project was funded (2nd evaluation step)	Nominal	{passed, rejected, conditionally}	rejected
grant	The actual granted amount after budget cut	Numeric	[228000, 750100]	429990

TABLE I: Dataset attributes for discrimination discovery.

	Applicants		Peer-Review Passed		Project Funded	
Program	Male	Female	Male	Female	Male	Female
P_1	881 (48.8%)	923 (51.2%)	43 (58.1%)	31 (41.9%)	25 (59.5%)	17 (40.5%)
P_2	1194 (60.1%)	792 (39.9%)	100 (76.3%)	31 (23.7%)	51 (81%)	12 (19%)

TABLE II: Aggregated data on gender differences.

passing the phase, however, it turns out that RD and RR cannot highlight differences in the outcome of the phase. All in all, the vast majority of both males and females proposals are rejected. In fact, RR is only 1.02 for P_1 and 1.05 for P_2; RD is only 1.5% for P_1, and a modest 4.5% for P_2. On the other hand, since relative chance (RC) compares the success rates, it highlights major differences: the chance of passing the peer-review for a female is only 69% of the chance of a male for program P_1, and only 46% for program P_2. Finally, since the odds risk (OR) is the ratio of RR and RC, it highlights differences in both rejection and success rates. Consider now the second phase. Rejected and funded projects are now more evenly distributed. The discrimination measures highlight no significant difference for program P_1. On the contrary, there are some for program P_2: RD is 12.3% and RR is 1.25.

Finally, the study of the distributions of single features in isolation (i.e. the age of the PIs, the number of her publications and of some costs) along gender of the PI and program of the research proposal does not show significant differences between males and females. Broadly speaking, in both programs we found that: (i) there is no difference in age between genders; (ii) males have a slightly higher productivity than females in terms of number of publications; and (iii) proposals led by females require slightly lower total cost as well as costs for young and good reputation researchers. Even though such a preliminary analysis may provide some hints on differences between project proposals, it is still too gross grained to draw any conclusion on discrimination. Aggregations at the level of the whole dataset may hide differences in smaller niches of data. This is precisely the objective of our discrimination discovery analysis.

IV. Discrimination discovery experiments

In this section, we report on the application and on the findings of the discrimination discovery methodology described in Section II-B on data consisting of project proposals from Program 1 (resp., Program 2) with reference to the peer-review decision. We do not consider the second evaluation phase for three reasons. First, the number of proposals involved in the second phase is much lower, hence we run the risk of drawing no statistically significant conclusion. Second, the discrimination measures in Figure 3 highlight higher differences between genders in the peer-review results than in the audition results. Third, and more importantly, the set of features available in Table I appears adequate to model the first phase but not the second one. In fact, peer-reviewers had

access only to the proposal text, to the PI curriculum and list of publications, and to the budget data. This is approximatively the set of features listed in Table I. On the contrary, the panel of national experts "entered in personal contact" with the PI during the audition, so their decision was affected by additional factors not recorded in the data, e.g., physical characteristics of the PI, her proficiency in speaking, her motivation, the appropriateness of her answers to questions. In summary, the omitted-variable bias in analysing data with reference to the second phase would be considerably higher than for the first phase.

A. Studying the risk difference distributions

Let us recall the approach described in Section II-B in our context. Let $\mathbf{r}$ be a project proposal led by a female PI that did not pass the peer-review phase. The function $diff(\mathbf{r}) = p_1 - p_2$ measures the risk difference between the percentage p_1 of its k-nearest neighboring proposals headed by female PIs and the percentage p_2 of its k-nearest neighboring proposals headed by male ones. Distance is measured on the basis of proposal's characteristics that are legally admissible in affecting the peer-reviewers' decision. We consider all the features of Table I apart from the project evaluation features (the decision itself) and the gender of the PI. The higher $diff(\mathbf{r})$ is, the more the negative decision on proposal $\mathbf{r}$ is unexplainable by those characteristics. The residual explanation is then the gender of the PI, which implies gender discrimination, or other causes not recorded in the data (the *omitted variables*). A basic decision in computing $diff(\mathbf{r})$ is the choice of the k constant. How many neighbors should be compared? A large k means that every proposal is a neighbor, hence $diff(\mathbf{r})$ collapses to the risk difference RD of the overall dataset (see Figure 3). Conversely, for a small k, we run the risk that the observed difference $p_1 - p_2$ is affected by randomness, hence loosing statistical support in drawing any conclusion. Figure 4 (a,b) shows the distribution of $diff()$ for $k = 4, 8, 16, 32$ with reference to proposals from Program 1 and 2. In both cases, the distributions follow the trend expected as k increases. From now on, we fix $k = 8$, which means comparing each proposal with 0.9% of proposals in Program 1 ($= 16/1804$, where 16 is $2k$, and 1804 is the number of proposals), and with 0.8% of proposals in Program 2.

Before describing the mining phase, it is interesting to study the social phenomenon of *favoritism* using the same approach. Figure 4 (c) shows the distribution of a sub-group that is suspected of being favored, namely males professors in Program 2. The decision outcome for which risk difference is calculated is now passing the peer-review phase, rather than being rejected. The distribution shown in the figure highlights then a higher chance for male professors of passing the peer-review phase with respect to the rest of PIs (females or non-professors). As for any discrimination conclusions, this is only a *prima facie* evidence of favoritism. It could be well explained by the higher experience or skills of male professors in writing project proposals than the rest of PIs.

B. A classification model for describing discrimination

Consider now the datasets of proposals led by female PIs that did not pass the peer-review phase. They amounts at 892 instances for program P_1 and 761 instances for program P_2. By fixing a threshold value t to the maximum admissible risk difference, the approach of [5] allows us for labelling a proposal $\mathbf{r}$ in the set above as discriminated or not by testing the condition $diff(\mathbf{r}) \geq t$. The choice of the threshold t should be supported by laws or regulators[3]. The only legal reference that we were able to find is the *fourth-fifth rule* in the U.S. [17], which states that a job selection rate (the RC measure in Table 1) lower than 80% represents a *prima facie* evidence of adverse impact. From the distributions of risk difference (see Figure 4 (a,b)), we fix from now on the threshold $t = 0.10$. We add a binary attribute $disc$ to the two sets under analysis, precisely defined as $disc(\mathbf{r}) = true$ iff $diff(\mathbf{r}) \geq 0.10$. This allows us for reducing the problem of discovering contexts of discrimination to the standard problem of inducing a classification model – where the class attribute is the newly introduced attribute $disc$. The datasets under analysis resulted in a 26-74% and in a 38-62% distribution of $disc = true$ and $disc = false$ class values for program P_1 and program P_2 respectively.

Experimental settings. Recall that the intended use of the extracted classification model is descriptive, to show the conditions under which a proposal led by a female PI was rejected at the peer-review phase with a risk difference of 0.10 or above. In this sense, we restrict the search space to classification models that are easily interpretable, namely decision trees (C4.5 [18]) and classification rules (RIPPER [19] and PART [20]).

Performances of classifiers are evaluated by a 10-fold cross validation, with the actual classification model extracted from the whole dataset. This approach is motivated by the small number of proposals in the two datasets under analysis. In addition, it is also due to the fact that we are actually interested in extracting a classification model to describe the conditions where discrimination occurred in the *whole* dataset under analysis. Evaluation measures calculated by cross validation include accuracy, precision, recall and f-measure for the class $disc = true$.

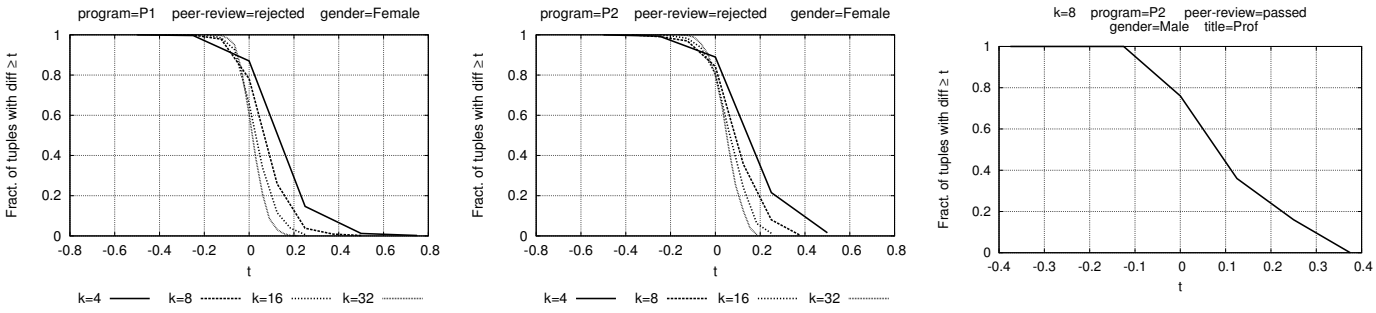
In order to overcome well-known problems in inducing classifiers from an unbalanced distribution of class values, we experimented a few standard approaches, including resampling of the training folds[4], cost-sensitive induction of classifiers[5], and meta-classification approaches (bagging and boosting).

Finally, we varied also the set of predictive attributes used to extract the classifiers. On the one side, this is a standard approach when searching for a (local) optimal classifier. On the other side, this is specific of the discrimination analysis

[3] A relevant question is how data analysis and social analysis techniques could help law makers and regulators in defining a reasonable value for t.

[4] Training set size was left unchanged, but the class values were resampled to a uniform distribution.

[5] With a cost of misclassifying actual $disc = true$ set to 2.5 times the cost of misclassifying $disc = false$.



(a) (b) (c)

Fig. 4: Cumulative distribution of $diff()$.

Id	Extraction parameters					Performance measures on P_1					Performance measures on P_2				
	Set of Attrs.	Alg.	Cost Sensitive	Meta Class.	Resamp.	Accuracy (%)	Prec. (%)	Recall (%)	F-measure	Size	Accuracy (%)	Prec. (%)	Recall (%)	F-measure	Size
1	D_{13}	Jrip	Yes	No	Yes	48.4	33.1	97.0	0.49	35	45.2	40.2	93.4	0.56	23
2	D_{19}	C4.5	No	No	Yes	77.6	54.3	83.1	0.66	302	70.7	58.7	74.5	0.66	184
3	D_{19}	C4.5	Yes	No	Yes	56.4	36.6	93.5	0.53	78	53.0	44.0	92.3	0.6	73
4	D_{19}	Jrip	Yes	No	Yes	53.5	35.4	97.0	0.52	21	45.4	40.4	95.5	0.57	11
5	D_{19}	PART	No	No	Yes	72.3	47.9	77.9	0.59	74	66.6	54.5	67.8	0.6	33
6	D_{19}	PART	Yes	No	Yes	61.9	39.7	90.9	0.55	74	57.3	46.4	88.8	0.61	87
7	D_{27}	C4.5	No	No	Yes	82.2	62.3	78.8	0.7	2051	74.5	63.4	75.9	0.69	4645
8	D_{27}	C4.5	Yes	No	Yes	50.7	33.5	92.2	0.49	9	37.6	37.6	100.0	0.55	325
9	D_{27}	Jrip	Yes	No	Yes	50.0	33.7	96.1	0.5	12	46.5	41.0	96.9	0.58	11
10	D_{27}	PART	Yes	No	Yes	61.2	38.1	80.1	0.52	128	53.1	43.7	86.7	0.58	113

TABLE III: Performances of the top 10 classifiers.

problem. In fact, in addition to the issue of the omitted-variable bias (i.e., not enough control variables), the literature on discrimination analysis also accounts for *included-variable bias* [21], namely for the presence of control variables already affected by gender discrimination. The effect of the included-variable bias is that part of observed risk difference is explainable in terms of predictive attributes that actually bear gender discrimination. For instance, the attribute $f\_partner\_num$ (the number of female investigators in the project proposal) could be one of such attributes. We denote by D_n subsets of n attributes from Table I defined as follows:

- D_{13} includes: age, $title$, pub_num and avg_aut as features of the PI; yr_num, yr_cost, grr_num, grr_cost, tot_cost and $fund\_req$ as features on project costs; $d1\_lv1$ and $d1\_lv2$ as feature on the research area; and, finally, the class attribute $disc$;
- D_{19} includes additional feature on the PI ($inst\_type$, $region$ and $f\_partner\_num$), and on project costs ($fund\_req\_perc$, yr_perc and grr_perc);
- D_{27} also includes all attributes at each level of the ERC hierarchy and the attribute $city$.

Extracting classification models. Table III shows the performances of the top 10 classifiers extracted in a large set of experiments conducted by varying all the parameters described in the previous subsection. Each row specifies the method and parameters used (attribute set, algorithm, cost-sensitive classification, meta-classification, resampling) and the performance measure values for the datasets of both program P_1 and P_2. A few comments on the lessons learned in tuning models and parameters for obtaining (local) optimal performances follow.

First, resampling the training set towards a uniform distribution of the class attribute reveals an effective techniques to improve performances, both in term of accuracy and f-measure, irrespectively of the model type and set of attributes. Using in addition cost-sensitive or meta-classifiers does not improve further. Compare for instance rows 2 vs 3, 5 vs 6, 7 vs 8 from Table III, where the only difference between the pairs is in the usage or not of misclassification costs.

Second, the effect of the set of predictive attributes is dependent on the classification model. Jrip and PART seem to benefit from larger sets as per accuracy and f-measure when moving from D_{13} to D_{19}, but then the additional attributes in D_{27} worsen the performances. Contrast for example rows 1 vs 4 vs 9, and 6 vs 10. This holds also for C4.5 models when using misclassification costs (see rows 3 vs 8). However, when using resampling only, there is an improvement from D_{19} to D_{27} (rows 2 and 7). C4.5 with resampling on D_{27} (row 7) is the best model w.r.t. both accuracy and f-measure.

Finally, we stress the importance of extracting models that trade performance with simplicity. We measure the structural complexity of a classification rule model (Jrip and PART) by the number of rules it contains, and the one of a decision tree model (C4.5) by the number of leaves in it – which boils down to the number of rules the model can be equivalently expressed with. Table III shows that the best model (row 7) is, unfortunately, the most complex one as well. The *global* description it provides is accurate but sparse in too many conditions. This motivates the search for a few *local* contexts

of discrimination.

Discovering contexts of discrimination from the extracted classification models. The actual discovery of discriminatory situations and practices may reveal an extremely difficult task. Due to time and cost constraints, an anti-discrimination analyst need to put under investigation a limited number of local contexts of possible discrimination. In our study, we proceeded with the selection of a few classification rules from the models of Table III. Such rules are either explicitly contained in classification models, as in the case of rule-based ones, or they are readily derived from, as in the case of paths in decision tree models. We consider rules of the following form:

```
(gender=female) and (cond_1) and ... and (cond_n)
    => disc=yes [prec] [rec] [diff]
```

where the item `gender=female` is implicitly part of any classification model, since models are extracted from project proposals headed by females. We evaluate interestingness of rules by standard measures of precision `[prec]` (proportion of discriminated proposals among those satisfying the antecedent), recall `[rec]` (proportion of the overall discriminated proposals covered by the antecedent), and f-measure. In addition, we consider the average risk difference of proposals satisfying the antecedent of the rule `[diff]`, as a measure of the degree of discrimination suffered by them. Finally, readability and interpretability of rules is also taken into account by preferring rules with fewer items in the antecedent. We adopt all such measures to let interesting rules emerge from the extracted models. In the following, we discuss two rules that ranked in the top positions.

The first classification rule, extracted with reference to proposals of the program P_1, highlights a context for a specific research area at top level of the ERC hierarchy, namely life sciences (LS):

```
R1: (d1_lv1 = LS) and (yr_cost >= 244,000) and
    (yr_num >= 2) and (avg_aut >= 8.4) and
    (pub_num <= 12) => disc=yes
    [prec=1.0] [rec=0.095] [diff=0.165]
```

The rule covers a considerable number of proposals among those labeled as discriminated, namely 9.5% of the total (27% of those in the LS research area), with a precision of 100%. The average risk difference is 16.5%. In addition to attributes on budget, rule R1 includes a test on the skills of PIs. In fact, the antecedent of the rule denotes research proposals (led by a female PI) requiring two or more young researchers, having a cost for them of 244,000€or more, and such that the PI has at most 12 publications with a mean number of authors of 8.4 or more. The lack of knowledge about the skills of an individual, as in presence of few publications and many co-authors, could be compensated by the peer-reviewers of LS through a prior knowledge of the average performances of the group or category the individual belongs to, in our case the gender. This is known as *statistical discrimination* or *rational racism* – as opposed to *taste-based* discrimination or *prejudice*.

Turning on proposals from program P_2, a large context of possible discrimination is highlighted by the rule:

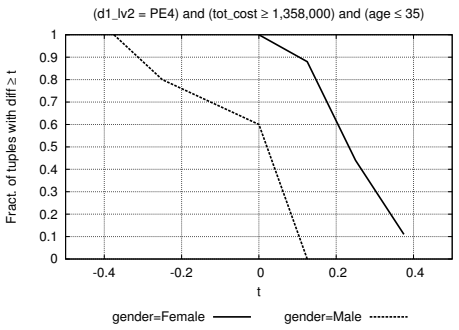


Fig. 5: Risk difference by gender for the antecedent of R2.

```
(d1_lv2 = PE4)
     => disc=yes [prec=0.55] [rec=0.147] [diff=0.086]
```

where PE4 is the "Physical and Analytical Chemical Sciences" panel, at the second level of the ERC hierarchy. Precision and average risk difference are not high, but recall is impressive: 1 out of 7 of the project proposals labeled as discriminated occurs in panel PE4, whilst only 1 out of 15 of all proposals from P_2 are in the panel. A specialization of the rule is:

```
R2: (d1_lv2 = PE4) and (tot_cost >= 1,358,000)
    and (age <= 35)
     => disc=yes [prec=1.0] [rec=0.031] [diff=0.194]
```

which concentrates on proposals with high budget led by young PIs. Recall is now only 3.1%, but precision is 100% and average risk difference is 19.4%. Intuitively, this can be read as if peer-reviewers of panel PE4 trusted young females requiring high budgets less than males leading similar projects.

Figure 5 reports the cumulative distributions of $diff()$ for proposals satisfying the antecedent of the rule R2 distinguishing female and male led projects. First, this is more informative than simply the average risk difference reported in the rules above. Second, it highlights the dual face of discrimination, namely favoritism: proposals led by males exhibit very low or even negative risk differences. Stated otherwise, they have been favored in comparison to similar projects led by females.

V. Related work from social sciences

Let us differentiate our approach from the extensive literature from social and economic sciences on gender discrimination in scientific peer-review (see [4] for a survey). On the side of the analytical methodology, linear regression (or some variants such as logistic or tobit regression) is the main tool adopted in those studies. The coefficient of the independent variable coding the gender is a measure of how gender affects the independent variable, which is typically the probability of a decision outcome. The issues of omitted-variable and included-variable bias are common problem here as well. Recent studies adopting regression include [22], which investigates gender and nationality discrimination in the selection of doctoral and post-doctoral research fellowships, and [23], investigating the interaction between nepotism, gender and productivity of applicants to grading procedures. Another

interesting methodology is meta-analysis [3], [24]: a regression model is not built from data on individual applications, but rather from aggregated data collected from other studies. The objective of meta-analysis is to detect which conditions (e.g., application type, research area of the applications, country of the applications) influence the log of the odds ratio measure (see Figure 1). The idea of searching "contexts" of possible discrimination makes then meta-analysis closer to our approach. However, even when gender is studied in association with other attributes, the approaches in the literature test only some prior hypothesis of discrimination. An example is *in-group* discrimination analysis, where the gender of applicants is related to the gender of peer-reviewers [22], but no other attribute is adopted to further stratify the data.

On the contrary, in our approach we mine data precisely to let previously unknown or unforeseen contexts of possible discrimination emerge. As an example, the rule R1 reported in Section IV-B unveils *prima facie* evidence of discrimination when certain project costs are above a threshold value and, at the same time, the age of the PI is below 35 years old. Both the cost and age attribute and the threshold values come out as a result of the analysis – there was no a prior hypotheses about them to be verified. The discovery of *niches of discrimination* is the major advancement over related work, and, to the best of our knowledge, this is the first paper that uses data mining techniques in discrimination studies on scientific review data.

VI. Conclusions and future work

We have presented a case study on the discovery of gender discrimination in a project funding program by discussing the available data, the issues tackled, the process followed and some *prima facie* evidences found. The methodology adopted relies on an implementation of situation testing using a variant of k-NN, and then on extracting and reasoning about a classification model. The proposed study can be refined in several directions. Our first objective is to formalize a knowledge discovery process in support of discrimination discovery. While the k-NN algorithm remains the core component, of particular interest is the definition of the *deductive* component, in which the extracted patterns are filtered, refined, validated and transformed into useful knowledge. To this aim, we are currently adapting the XQuake system [25]. A second objective, more related to the case study, is to enrich the available dataset with additional features measuring the scientific productivity of applicants and their professional network. As for now, in fact, it has been possible to take into account only the raw number of publications, since our input data was anonymized.

References

[1] M. Brouns, "The gendered nature of assessment procedures in scientific research funding: The Dutch case," *Higher Education in Europe*, vol. 25, no. 2, pp. 193–199, 2000.

[2] H. Marsh, U. Jayasinghe, and N. Bond, "Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability," *American Psychologist*, vol. 63, no. 3, pp. 160–168, 2008.

[3] H. W. Marsh, "Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches," *Review of Educational Research*, vol. 79, no. 3, pp. 1290–1326, 2009.

[4] S. J. Ceci and W. M. Williams, "Understanding current causes of women's underrepresentation in science," *Proceedings of the National Academy of Sciences*, vol. 108, no. 8, pp. 3157–3162, 2011.

[5] B. T. Luong, S. Ruggieri, and F. Turini, "k-NN as an implementation of situation testing for discrimination discovery and prevention," in *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2011)*. ACM, 2011, pp. 502–510.

[6] A. Romei and S. Ruggieri, "Discrimination data analysis: A multi-disciplinary bibliography," in *Discrimination and Privacy in the Information Society*, ser. Studies in Applied Philosophy, Epistemology and Rational Ethics, B. H. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky, Eds. Springer, 2012, pp. 109–135.

[7] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2008)*. ACM, 2008, pp. 560–568.

[8] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *ACM Trans. on Knowledge Discovery from Data*, vol. 4, no. 2, p. Article 9, 2010.

[9] ——, "DCUBE: Discrimination discovery in databases," in *Proc. of the ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2010)*. ACM, 2010, pp. 1127–1130.

[10] ——, "Integrating induction and deduction for finding evidence of discrimination," *Artificial Intelligence and Law*, vol. 18, no. 1, pp. 1–43, 2010.

[11] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining & Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[12] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, pp. 1–33, 2012.

[13] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE Transactions on Knowledge and Data Engineering*, p. to appear, 2012.

[14] M. Bendick, "Situation testing for employment discrimination in the United States of America," *Horizons Stratégiques*, vol. 3, no. 5, pp. 17–39, 2007.

[15] I. Rorive, "Proving Discrimination Cases - the Role of Situation Testing," 2009, Centre For Equal Rights & Migration Policy Group, `http://www.migpolgroup.com`.

[16] L. Quillian, "New approaches to understanding racial prejudice and discrimination," *Annual Review of Sociology*, vol. 32, no. 1, pp. 299–328, 2006.

[17] Equal Employment Oppotunity Commission, "Uniform guidelines on employee selection procedure," 1978, 43 FR 38295, `http://www.justice.gov`.

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

[19] W. W. Cohen, "Fast effective rule induction," in *Proc. of Int. Conf. on Machine Learning (ICML 1998)*. Morgan Kaufmann, 1995, pp. 115–123.

[20] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in *Proc. of Int. Conf. on Machine Learning (ICML 1998)*. Morgan Kaufmann, 1998, pp. 144–151.

[21] M. R. Killingsworth, "Analyzing employment discrimination: From the seminar room to the courtroom," *American Economic Review*, vol. 83, no. 2, pp. 67–72, 1993.

[22] L. Bornmann and H. Daniel, "Gatekeepers of science – effects of external reviewers' attributes on the assessments of fellowship applications," *Journal of Informetrics*, vol. 1, no. 1, pp. 83–91, 2007.

[23] U. Sandström and M. Hällsten, "Persistent nepotism in peer-review," *Scientometrics*, vol. 74, no. 2, pp. 175–189, 2008.

[24] L. Bornmann, R. Mutz, and H.-D. Daniel, "Gender differences in grant peer review: A meta-analysis," *Journal of Informetrics*, vol. 1, no. 3, pp. 226–238, 2007.

[25] A. Romei and F. Turini, "XML Data Mining," *Software: Practice and Experience*, vol. 40, no. 2, pp. 101–130, 2010.