Unveiling Systematic Biases in Decisional Processes

An Application to Discrimination Discovery

Laura Genga
Eindhoven University of Technology
l.genga@tue.nl

Luca Allodi
Eindhoven University of Technology
l.allodi@tue.nl

Nicola Zannone
Eindhoven University of Technology
n.zannone@tue.nl

ABSTRACT

Decisional processes are at the basis of several security and privacy applications. However, they are often not transparent and can be affected by human or algorithmic biases that may lead to systematically misleading or unfair outcomes. To unveil these biases, one has to identify which information was used to make the decision and to quantify to what extent such information has influenced the process outcome. Two classes of techniques are widely used to determine possible correlation between variables within decisional processes from observational data: *(i)* econometric techniques, in particular *regression analysis*, and *(ii)* knowledge discovery techniques, in particular *association rules mining*. However, these techniques, taken individually, have intrinsic drawbacks that limit their applicability. In this work, we propose an approach for unveiling biases in decisional processes, which leverages association rule mining for systematic hypothesis generation and regression analysis for model selection and recommendation extraction. We demonstrate the proposed approach in the context of discrimination detection, showing that not only it provides 'statistically significant' evidence of discrimination but it also allows for a more efficient operationalization of the recommendations extracted, upon which the decision maker can operate.

CCS CONCEPTS

• **Security and privacy** → *Social aspects of security and privacy*; • **Computing methodologies** → *Reasoning about belief and knowledge*.

KEYWORDS

Decisional Process, Econometrics, Association Rule Mining

ACM Reference Format:
Laura Genga, Luca Allodi, and Nicola Zannone. 2019. Unveiling Systematic Biases in Decisional Processes: An Application to Discrimination Discovery. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS '19), July 9–12, 2019, Auckland, New Zealand*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3321705.3329856

1 INTRODUCTION

Several practical applications of Security and Privacy to real world problems require the collection of very large volumes of data, these being system and network logs, alarm events in a Security Operation Centers (SOCs), or records of sensitive customer or patient data. Operationally, security & privacy data is only as useful as our ability to discover (undesired) patterns in the data, for example to identify violations of a security policy, internal requirements for a business process, or laws. This is crucial to allow operational personnel, policy makers, or process coordinators to evaluate the overall performance of the process, and its compliance with security, privacy, and law requirements.

Security and privacy applications are especially weak in this respect due to the highly noisy data and large amount of false positives that the relevant '*data generation processes*' (e.g., network inspection by an Intrusion Detection System [4] or human decision making [13]) inject in the data. These inherent biases in the data generation, regardless of whether they are systematically introduced by data-crunching algorithms or humans, can be severely misleading for data analysts and decision makers, particularly in security domains such as vulnerability analysis [5, 9] or discrimination detection [12].

Uncovering biases from observational data is a broad and still an open problem that requires a thorough exploratory analysis and understanding of the data, as well as rigorous estimations of effect sizes. Whereas the literature generally considers association rule mining approaches for the former [1], the outcome typically consists of several thousand rules that cannot be easily operationalized. Similarly, statistical and econometric models are often used to evaluate effect sizes and rigorously evaluate evidence in the data [7], but are of no use without clearly defined hypotheses and a clear understanding of the data generation process.

In this work, we propose a novel methodology that leverages principles from both association rule mining and regression analysis to uncover systematic biases in decisional processes. We apply our methodology to the case of discrimination detection in large datasets, to uncover the systematic use of sensitive data in decision making. We use association rule mining to extract candidate hypotheses of biases from an exploration of data. These hypotheses are used to build regression models, which provide us with 'statistically significant' evidence for the presence/absence of biases in the decisional process, and effectively act as a cream-skimming mechanism to filter out hypotheses that are equivalent or that do not add significant information to uncover the data generation mechanism. This evidence can then be used by the analyst to take action and tackle the decision bias at the source. We evaluated the proposed methodology through experiments with synthetic data. In particular, we study the ability of the methodology to determine whether decisional processes are affected by biases that lead to the unfair treatment because of personal characteristics or membership to certain (protected) societal groups.

AsiaCCS '19, July 9–12, 2019, Auckland, New Zealand

ACM ISBN 978-1-4503-6752-3/19/07…$15.00
https://doi.org/10.1145/3321705.3329856

2 BACKGROUND

Decisional processes are at the core of most businesses; they rely on cognitive resources to help decision-makers in making appropriate and profitable decisions and accurate predictions. In this work, we model a decisional process as a set of records comprising a number of attributes (hearafter called *variables*) and the *outcome* of the process, i.e., the decision made on the basis of the variables.

Table 1 exemplifies the decisional process of a financial institute, aimed to determine whether a given individual should or should not be classified as a high risk individual, i.e. he is likely not able to repay a loan. Each record corresponds to a single observation of the process where `SubjID` simply provides the ID of the individual requesting the loan and `Employed`, `Income`, `Gender`, `Race` are variables characterizing the individual. `HighRisk` represents the variable describing the outcome of the process.

Decisional processes are often not transparent, and can be affected by (implicit or explicit) human or algorithmic biases, which may lead to systematically unfair outcomes. In the example above, the data could reveal to an observer whether being currently employed is a relevant criterion for the bank's decision making process. Whereas one could reasonably expect to find this relation in any decision process of this type, other complex dynamics can have a 'hidden' impact on the decision. For example, `Gender` or belief-related biases can interact with other variables, like `Race` or social status, to affect a decision.

In order to detect biases in decisional processes, one has to quantify to what extent the use of variables (e.g., `Gender`, `Race`) has influenced the process outcome. A main challenge lies in the fact that the decisional process is often unknown. Therefore, detecting biases in decisional processes requires reconstructing the decisional process from observational data and determining which variables were used for decision making.

Two classes of techniques are widely used for the exploration of observational data: *(i)* approaches that use statistical tools, in particular *regression analysis*, to determine which variables are more likely able to explain the process outcome and *(ii)* approaches based on knowledge discovery techniques, in particular *association rule mining*, to measure possible differences in the proportions of positive/negative decisions on different groups of observations. The following sections introduce the basic concepts underlying these two lines of research and discuss their shortcomings.

2.1 Association Rule Mining

Agrawal et al. [1] have formulated the problem of rule mining as follows. Let D be a set of records called the dataset, $\mathcal{V}$ a set of variables (e.g., `Employed`, `Income`, `Gender`, `Race` in Table 1) and Λ a set of items of the form $Var_i = x_j$ where $Var_i \in \mathcal{V}$ and x_j is a value Var_i's domain (e.g., `Employed` = Y). An *association rule* r is an implication of the form $r : X \rightarrow Y$, where X and Y are two itemsets (i.e., $X, Y \subseteq \Lambda$), respectively called *antecedent* and *consequent* of the rule. Intuitively, an association rule indicates that if X occurs in a record, then Y will also likely occur in that record. In this work, we consider *class association rules* [8], i.e., association rules whose consequent consists of a single class item representing the possible process outcome (e.g., `HighRisk` in Table 1). Hereafter, we use the term 'rule' to refer to a class association rule.

	Outcome	Variables			
SubjID	High Risk	Employed	Income	Gender	Race
1	1	N	2000	M	Black
2	0	Y	10000	F	White
⋮	⋮	⋮	⋮	⋮	⋮
100	1	N	5000	M	Asian

Table 1: Example decisional process of a finance institute

Since the number of rules that can be mined from a dataset is usually huge, a common practice in association rule mining is to filter out all rules whose *relevance* is below a predefined threshold. Two well known metrics for assessing the relevance of rules are *support*, which represents the percentage of records in the dataset covered by a rule, and *confidence*, which represents the percentage of records fulfilling the rules among those fulfilling its antecedent.

Given a dataset recording a decisional process, a set of potentially discriminated groups and a criterion of unlawful discrimination, association rule mining provides a means to find potentially discriminated groups [12]. In this regards, besides the standard association rules mining relevance metrics, several metrics tailored to measure the impact of sensitive itemsets on the class have been proposed. In a seminal work on discrimination discovery [12], Ruggeri et al. introduce the notion of *extended lift* to measure how the rule confidence varies with/without the discriminatory itemset, thus providing an evaluation of the relevance of this itemset. This approach, however, suffers two main drawbacks: *(i)* the lack of a *statistical validation* of the discovered evidence and *(ii)* possible *redundancy* among the mined rules.

To deal with the first issue, Pedreschi et al. [11] propose to exploit statistical tools to assess the significance of quantitative measures of discrimination. In particular, they compute *confidence intervals* for each measure, which describe the probability with which we can expect to find a given measure within a given range in repeated experiments. The level of discrimination is then assessed by considering the lower bound of the interval, rather than the measure directly.

The presence of redundant rules in the mined ruleset, i.e. rules that describe the same (or very similar) set of records, poses significant challenges on the reliability of the discovered evidence of discrimination. In fact, if a given set of records can be described both by a rule including sensitive itemsets and by a rule not including sensitive itemsets, then it is not possible to determine which itemsets were actually used in the decisional process. The problem of rule redundancy is well studied in the rule mining community [3, 6]. However, most of the existing approaches only aim to find the minimal ruleset that describes the given dataset without loosing relevant information. As such, they do not distinguish between sensitive and non sensitive itemsets and their outcome might still contain redundant rules with respect to the decisional process. An exception is the approach in [10], where Pedreschi and colleagues introduce the notion of p-instance. The underlying idea is to discard those rules for which there exists at least one rule has similar confidence, and involve a non-sensitive itemset that presents a significant correlation with a sensitive itemset.

2.2 Econometrics

The field of Econometrics deals with the problem of robustly estimating the effect of some variable of interest on an observable outcome. Researchers are generally interested in the coefficients that quantify the effect of a change in a so-called 'explanatory variable' on the 'dependent variable', and on the power of the model to *explain* (or, equivalently, *predict*) the phenomenon of interest. Econometric techniques can be applied, in principle, to any dataset relating 'explanatory variables' to an outcome; the relation between the two emerges from the definition of hypotheses that can be tested in the data. In the example process of Table 1, explanatory variables are `Employed`, `Income`, `Gender`, `Race`.

Econometric estimations. Econometric models are also called *regression equations*, and have the general form:

$$Y = c + \beta_1 Var_1 + \beta_2 Var_2 + \cdots + \beta_n Var_n + \epsilon \quad (1)$$

where Y is the outcome variable, c is the intercept, and $\beta_1, \ldots, \beta_n$ are the regression coefficients of the n explanatory variables, and ϵ is the error term. The estimation of the coefficients is the key aspect, and which regression function one uses depends on the nature of the data (e.g. a logit for binary outcomes, or a poissonian for count variables). Which model to choose boils down to the analyst's expertise and knowledge of the data.

Model parametrization and selection. The formulation (also called *parametrization*) of an econometric model is generally based on theoretical predictions, expectations, or hypotheses that the researcher wants to test in the data. In some cases, however, the parametrization of the model cannot be based on any specific theoretical prediction, and is derived in an *exploratory* setting to identify potentially interesting relations in the data [7]. This model selection can be automated by employing techniques that, based on the analysis of variance (ANOVA) and the principle of minimality (simpler models are preferred), can select the model that has the highest *power* in 'explaining' the data.

Regression output interpretation. The output of a regression is the estimation of which values of $c, \beta_1, \ldots, \beta_n$ provide the best prediction of Y. For example, the fictitious model on the binomial outcome variable `HighRisk`: $\texttt{HighRisk} = c + \beta_1 \texttt{Employed} \cdot \beta_2 \texttt{Gender}$ will generate an output of the following type:

Regressor	Coeff.	p-value
c	0.52	< 0.05
$\texttt{Employed} = N$	0.2	< 0.01
$\texttt{Gender} = F$	-1.2	0.10
$\texttt{Employed} = N \wedge \texttt{Gender} = F$	1.1	< 0.01

The output indicates that non employed (and male) subjects have a 22% ($\exp(0.2) = 1.22$)[1] higher probability of being assigned to the category `HighRisk` than to the category `LowRisk`. Being female (*and* employed) decreases chances by 70% ($\exp(-1.2) = 0.3$). The coefficient for $\texttt{Employed} = N \wedge \texttt{Gender} = F$ tells us that, however, being female *and* unemployed increases the baseline risk 3 times ($\exp(1.1) = 3.0$). The statistical significance of each coefficient serves as an indication to the analyst that the estimation is unlikely to have been generated by a random process; the smaller the probability of observing that estimation from 'random data' (the infamous *p-value*), the highest the confidence one can have in the value of the specific estimation. In the example above, the *p-values* suggest that all coefficients are statistically significant with a borderline effect for the variable `Gender`.

[1] As the outcome variable of this example is binary, a logistic regression should be used. For a *logit* regression, the outcome is $\log(p/(1-p))$, and so regression coefficients should be exponentiated to reveal the odds ratio.

	Assoc. Rule Mining	Econometrics
Statistical validation	◐	●
Data-agnostic	●	○
Hypothesis ranking	◐	●
No parameter tuning	○	●

Table 2: Comparison of Association Rule Mining and Econometric approaches where ● means "support", ◐ "partially support", ○ "no support".

2.3 Discussion

Uncovering biases from observational data requires a thorough exploratory analysis of the data, as well as rigorous estimations of effect sizes. Whereas the literature generally considers association rule mining approaches for the former, statistical and econometric models are often used to evaluate effect sizes and rigorously evaluate evidence in the data. However, both methods have intrinsic shortcomings in our application as summarized in Table 2.

3 METHODOLOGY

The goal of this work is to devise an approach to detect systematic biases in decisional processes, which lead for instance to discrimination. To this end, we propose to combine principles of rules mining with regression analysis. As discussed in Section 2, rule mining enables the discovery of potential relations between variables and the outcome of the decisional process. However, this method is prone in generating too many rules for its output to be actionable (e.g., by a policy maker investigating discrimination), and does not provide any obvious ranking to decide which rules to look at first. To address this issue, we look at econometrics for the evaluation of statistical evidence, and employ a set of methods to generate econometric models from the generated rules, select the most powerful models in explaining the data.

Our methodology is summarized in Figure 1. First, association rule mining is employed to generate the set of relevant rules (1). For our purpose, we treat each mined rule as a candidate hypothesis for bias. Then, we generate econometric models by considering the variables included in each selected rule and regress over them to generate model estimates of the considered outcome variable (2). A model comparison is performed to eliminate 'redundant' models that do not add relatively more information to the prediction than a simpler model does (3). This leaves us only with models that, among all evaluated candidates, provide the more convincing evidence for some effect, if any. Finally, we extract the coefficients of these selected models to identify (sub)populations of interest where there is statistically significant evidence of bias in the decision making (4). Next, we provide a detailed breakdown of each step of the methodology.

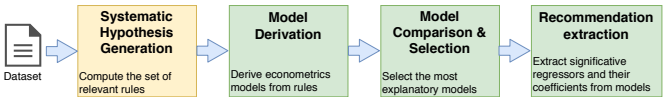

Figure 1: Depiction of the proposed methodology

3.1 Systematic Hypothesis Generation

Given a dataset D, we apply class association rule mining to derive the set of relevant rules R_{rel}, that is the set of rules that fit user-defined thresholds for the selected metric(s) of relevance. To measure the relevance of rules, we use *support* and *confidence*. Specifically, we say that a rule is relevant if its *support* and *confidence* levels are above some given thresholds ρ_{supp} and ρ_{conf} respectively. Each rule in R_{rel} is considered a candidate hypothesis of biases in the decisional process.

Example 3.1. Below, we show some example (relevant) rules that can be extracted by the application of association rule mining to the decisional process of Table 1:

$$\begin{aligned} r_1: &\ \texttt{Income} = 5000, \texttt{Race} = \textit{White} \rightarrow \texttt{HighRisk} = 0 \\ r_2: &\ \texttt{Gender} = M \rightarrow \texttt{HighRisk} = 1 \\ r_3: &\ \texttt{Gender} = M, \texttt{Employed} = N \rightarrow \texttt{HighRisk} = 1 \\ r_4: &\ \texttt{Income} = 2000, \texttt{Race} = \textit{White} \rightarrow \texttt{HighRisk} = 0 \end{aligned}$$

An analyst should investigate all the rules in R_{rel} to check whether they correspond to actual biases in the decisional process. However, as discussed earlier, association rule mining provides very little support for this. For instance, R_{rel} might contain rules that are not 'independent' from each other. In particular, many rules can be "subrules" of others, i.e., they extend other rules with additional itemsets as in the case of r_2 and r_3 above. Thus, the analyst might not know whether the (possible) bias concerns the population characterized by a given rule (e.g., employed males) or whether it affects a larger population as characterized by a subrule (e.g., all males).

To find more reliable evidence of biases, the analyst can only tune the thresholds for support (ρ_{supp}) and confidence (ρ_{conf}). Although these parameters has a significant impact on the size of the mined ruleset, they provide limited means to understand the actual impact of the variables in the antecedent of a rule on the process outcome. Also, too high values can lead to discard biases affecting small populations (see Section 2.3). While we will evaluate the choice of ρ_{supp} and ρ_{conf} experimentally in Section 4, in this work, we pursue a different direction and exploit econometric techniques to determine the statistical validity of the evidence found and use the obtained coefficients to extract recommendations that help the analyst in uncovering biases in the decisional process.

3.2 Model Derivation

As discussed above, rule mining techniques often return a large number of rules, which, with no clear prioritization or comparison criterion, are essentially useless from an operative standpoint. In this work, we exploit econometric techniques to evaluate statistical evidence of bias starting from the candidate hypotheses extracted using rule mining. In this step, we show how econometric models can be obtained from the set of relevant rules R_{rel}.

To derive econometric models from R_{rel}, we look at the variables occurring in the rules in R_{rel}. Let $V = \bigcup_{r_i \in R_{rel}} \{V_i \mid V_i = \text{var}(r_i)\}$ where $\text{var}(r_i)$ denotes the set of variables occurring in the antecedent of rule r_i. Given a rule $r_i : Var_{i,1} = x_{i,1}, \ldots, Var_{i,N} = x_{i,N} \rightarrow \texttt{Class} = Y$, we consider the set of explanatory variables $V_i = \{Var_{i,1}, \ldots, Var_{i,N}\} \in V$ and build a corresponding regression model M_i of the form $M_i : \texttt{Class} = c_i + \beta_{i,1} Var_{i,1} \cdot \beta_{i,2} Var_{i,2} \cdot \ldots \cdot \beta_{i,N} Var_{i,N}$.

Example 3.2. From the ruleset in Example 3.1, we can extract three models:

$$\begin{aligned} M_1: &\ \texttt{HighRisk} = c_1 + \beta_{1,1}\, \texttt{Income} \cdot \beta_{1,2}\, \texttt{Race} \\ M_2: &\ \texttt{HighRisk} = c_2 + \beta_{2,1}\, \texttt{Gender} \\ M_3: &\ \texttt{HighRisk} = c_3 + \beta_{3,1}\, \texttt{Gender} \cdot \beta_{3,2}\, \texttt{Employed} \end{aligned}$$

Note that some rules collapse in a single model as they contain exactly the same set of variables. In our example, this is the case for rules r_1 and r_4, which are *both* represented by model M_1. The reason for this lies in the fact that the regression evaluates the effect of a *change* in the explanatory variables on the dependent variable.

To efficiently compare the 'credibility' of the obtained models (next step), we organize the derived models in a hierarchical structure. To this end, we introduce a partial order relation over econometric models, resembling the subrule relation, and, given two econometric models M_i and M_j defined over the sets of explanatory variables V_i and V_j respectively, we say that M_j is *nested* in M_i if $V_i \subset V_j$. Based on this relation, we construct a forest of models whereby the model at the root of each tree is the simplest model (i.e., with the lowest number of variables on the right hand side of the equation), and each child is a nested model of its parent(s).

3.3 Model Comparison and Selection

Once the hierarchical structure is in place, we apply a model selection procedure by comparing each parent with all its child models. Our pruning strategy consists in checking whether the addition of variables to a child model adds information that leads to a better description of the data, or whether the simpler model is preferable (in that it describes the data indistinguishably well w.r.t. the more complex model). This can be operationalized through an ANOVA test comparing the model pairs.

The ANOVA test is a widely used statistical test that allows one to compare the fits of two regression models by comparing the (sum of squares of the) residuals (i.e., the errors) of the respective model predictions before and after the inclusion of the additional variables in the nested model [2]. If the output of the ANOVA test indicates that the more complex model provides a significantly better explanation of the variance in the prediction, we mark the nested model as preferable compared to the more general, simpler one. Otherwise, we mark the parent model as preferable. Once all comparisons have been performed, we discard a model if it is less preferable than all its nested models.

Example 3.3. The three models in Example 3.2 can be represented as two trees in which one tree only consists of model M_1 and the other consists of models M_2 and M_3 where M_3 is a nested model of M_2. Therefore, one should compare M_2 with M_3 to determine which of the two provides a better description of the data. Supposing the ANOVA test indicates that M_3 is preferable, the model selection procedure returns models M_1 and M_3.

3.4 Recommendation Extraction

From step 3 we obtain a set of selected regression models M_{sel} that provide the best 'explanation' of the dataset. Each model comprises a set of coefficients $C_i = \{\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,N}\}$ together with an output of a statistical test determining whether each element of C_i is significantly different from zero (i.e., whether the associated variable in V_i is likely to have a significant effect on the outcome variable). The minimum level of statistical significance generally considered is 5% ($p \leq 0.05$), with borderline significance indicated up to a 10% level ($0.05 < p \leq 0.10$).

By inspecting each model $M_i \in M_{sel}$, in this phase we extract 'significant' regressors and associated coefficients $\langle \beta_{i,j}, Var_{i,j} = x_{i,j} \rangle$ with $p_{i,j} \leq 0.05$. Each extracted pair conveys information on the *direction* and *size* of the bias towards the group $Var_{i,j} = x_{i,j}$, represented respectively by the sign and magnitude of the coefficient $\beta_{i,j}$. The interpretation of this coefficient, in the case of discrete (as opposed to continuous) variables, is to be interpreted w.r.t. a baseline value as the relative change in the outcome variable for subjects that belong to the relevant category. We refer to Section 2.2 for the representation of model M_2 and its interpretation.

4 EXPERIMENTS

We evaluate the ability of our approach to detect situations in which discrimination occurred against a set of experiments with synthetic data (allowing us to determine ground truth).

Generating synthetic data. Table 3 shows the variables we used along with their domain. We consider Age as the 'discriminatory variable', whereas the others are considered 'context variables'. We generated the synthetic data in two steps. First, we created discriminated groups by randomly selecting one or more context variables in addition to the discriminatory variable Age, along with a value from their domain. We generated 10 datasets by varying the number of discriminated groups (i.e., 1, 2 and 3) and the complexity of the dataset. The latter is defined by the presence/absence of *noise* (i.e., non-discriminated subjects whose context variables partially match values for the discriminated group), and the presence/absence of *overlapping* (i.e., discriminated groups can share at least a value of a context variable). Combining these two dimensions, we obtain four types of datasets: i) no noise, no overlapping; ii) no noise, with overlapping; iii) with noise, no overlapping; iv) with noise and overlapping.

For every dataset, we generated a total of 10000 samples; among them, every discriminated group covered the 25% of the dataset.

Evaluation metrics. To evaluate our methodology, we computed the *density* of acceptable recommendations returned by our methodology as the ratio between the number of significant regressors that describe (at least some of) true correlations among variables and the total number of regressors marked as significant in the models returned by the approach. To this end, we first derive for each model the set of significant regressors along with their coefficients. Then, we compare each regressor with the set of 'true' discriminating variables describing a discriminated group. This comparison can return four different outcomes: a) *Exact*, if the regressor involves all and only the variables characterizing one of the discriminated groups; b) *Too specific*, if variables in a regressor is a strict superset of the true discriminating variables; c) *Partial*, if variables in the regressor overlap with the set of discriminating variables (but it is not a superset); d) *Off target*, if the regressor does not involve any discriminating variable. The density of acceptable recommendations is computed as the number of *Exact*, *Too specific*, *Partial* regressors (rules) over all significant regressors (rules).

Instruction	Speak Language	Previous Role	Country	Age	Class
Bachelor	Y	Employee	USA	25-50	Y
Doctorate	N	Manager	India	>50	N
Master		Self-Employee	Europe		
HighSchool		Unemployed	SA		
			China		

Table 3: Variables used for the generation of the synthetic datasets along with their domain.

Results. Table 4 reports descriptive statistics for the results for the experiment runs. The first set of rows reports aggregate descriptive statistics of the results obtained using association rule mining; the second set reports results for model selection. We notice that the number of rules identified by association rule mining is significantly larger than both the number of selected models and that of extracted regressors. The average experimental run produces 24.8 rules, with the top 25% of runs producing in excess of 27 up to 131 rules. The relatively high standard deviation (w.r.t. the mean) indicates that rules in output can be expected to vary consistently in number across experimental setups. By contrast, the number of models selected per experimental setup is on average three times smaller, with only 3.6 regressors selected per model. The upper 25% of the distribution of selected models and regressors is in the ranges 7-16 and 5-28 respectively, both much smaller than the one returned by association rule mining. The low standard deviation indicates relatively stable output across experiments. Overall, this indicates that the model selection procedure appears to be removing a significant number of rules, but says little about the *correctness* of the process.

A first indication of this comes from the evaluation of *Exact, Too specific, Partial,* and *Off target* rules/regressors. As 'matches' are calculated on a per-rule basis, and multiple regressors can be generated from a single rule, a direct quantitative comparison between *Rules* and *Regressors* should not be attempted here. On the other hand, comparing the distributions can shed some light on the overall outcomes of the experiments. *Exact* and *Too specific* matches for rules are very few, as up to the third quartile (75% of the outcomes distribution) there is no match. On the other hand, in no case association rule mining returns an *off-target* result. The vast majority of results are *Partial*. Looking at the *Regressors* results, we observe a very similar distribution, whereby each rule leads to regressors that are most likely going to point to an over-estimation of the targeted group in the dataset. However, the econometric approach appears to be adding a small number of *off-target* matches by pointing to groups that are not part of the 'discriminated' group in our data. This is likely a product of random variation in the sample. Overall, we observe that the regressor analysis appears to lead to similar performances in terms of detected groups, while producing three to five fewer 'recommendations' as an output of the procedure.

	Metric	min	1st Q	avg	med	3rd Q	max	sd	n_obs
Rules	N_rules	1	6.0	24.8	14.0	27.0	131	29.6	563
	Exact	0	0.0	0.1	0.0	0.0	1	0.2	13966
	Too specific	0	0.0	0.0	0.0	0.0	1	0.2	13966
	Partial	0	1.0	0.9	1.0	1.0	1	0.3	13966
	Off target	0	0.0	0.0	0.0	0.0	0	0.0	13966
Regressors	N_rules	1	6.0	24.9	14.0	27.2	131	29.6	560
	N_models	1	5.0	7.4	7.0	9.0	16	3.9	792
	N_reg/model	1	1.0	3.6	2.0	5.0	28	4.7	792
	Exact	0	0.0	0.4	0.0	0.0	3	0.8	792
	Too specific	0	0.0	0.1	0.0	0.0	5	0.3	792
	Partial	1	1.0	2.6	1.0	3.0	25	3.9	792
	Off target	0	0.0	0.5	0.0	1.0	5	0.9	792

Table 4: Descriptive statistics of the experiment runs

Figure 2 displays the density of 'useful' indications provided by the two procedures. Results are arranged in a matrix, where each box represents a set of experiments with varying confidence and support levels. Within each cell, each square corresponds to one combination of support and confidence thresholds for that specific experiment setting and number of discriminated groups (rows). Lighter the color, higher the density. Cells that are not colored represent support-confidence combinations for which no rules were inferred.

Across all experimental setups we observe that the combination of the econometric approach and rule mining (top of the figure) produces a much higher density of rules covering the truth than rule mining alone. In both cases we observe performance to decrease with the addition of noise and overlap to the datasets, although the econometric approach appears to be more resilient. Similar observations can be made as discriminated groups increase. Rule mining appears to perform better with higher levels of confidence only in the presence of high support levels. This is in line with previous observations that rule mining is sensitive to parametrization, and choosing the correct parameter configuration largely depends on unknown structures in the data. By contrast, the econometric approach appears to perform well across the board, with decreasing performance in general only for higher levels of support, whereas it appears to be largely insensitive to varying confidence levels. This effectively removes the need for fine-tuning the support and confidence thresholds for rule selection.

5 CONCLUSION

In this work, we have proposed a methodology that leverages both association rule mining and regression analysis to uncover systematic biases in decisional processes. Our methodology uses association rule mining to systematically generate hypotheses of bias sources from an exploration of the data. These hypotheses are then used to build regression models that provide statistically significant evidence about the impact of variables on the process outcome.

The experiments show that our methodology overcomes the limitations of standard association rule mining. However, while being able to detect the population been discriminated, it tends to provide only an indication of the targeted set of observations, as opposed to giving a precise picture of targeted sub-groups. This is to be expected from any statistical analysis, as noisy data and sample

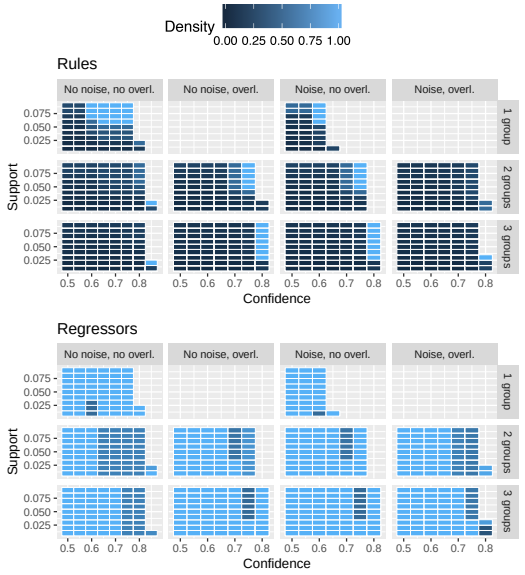

Figure 2: Density of 'useful' indication for varying levels of support and confidence

sizes affect clearly have an impact on the prediction. Nonetheless, the ability of filtering out a large number of overly-specific rules and focusing only on a few that are highly likely to cover the population of interest, enables policy makers and analysts to focus on groups of observations where future investigations and data collection are likely to uncover the specific effect.

Acknowledgements. This work is funded by the ECSEL project SECREDAS (783119) and the ITEA3 project APPSTACLE (15017).

REFERENCES

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining Association Rules Between Sets of Items in Large Databases. *SIGMOD Rec.* 22, 2 (1993), 207–216.
[2] Alan Agresti. 2003. *Categorical data analysis*. Vol. 482. John Wiley & Sons.
[3] Mafruz Zaman Ashrafi, David Taniar, and Kate Smith. 2007. Redundant Association Rules Reduction Techniques. *Int. J. Bus. Intell. Data Min.* 2, 1 (2007), 29–63.
[4] Stefan Axelsson. 2000. The base-rate fallacy and the difficulty of intrusion detection. *TISSEC* 3, 3 (2000), 186–205.
[5] Steve Christey and Brian Martin. 2013. *Buying into the bias: why vulnerability statistics suck.* Technical Report.
[6] Laurentiu Cristofor and Dan Simovici. 2002. Generating an informative cover for association rules. In *Proc. of Int. Conf. on Data Mining*. IEEE, 597–600.
[7] Andy Field. 2013. *Discovering statistics using IBM SPSS statistics*. Sage.
[8] Bing Liu Wynne Hsu Yiming Ma and Bing Liu. 1998. Integrating classification and association rule mining. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 80–86.
[9] Thanh Nguyen, Bram Adams, and Ahmed Hassan. 2010. A Case Study of Bias in Bug-Fix Datasets. In *Proc./ of Working Conference on Reverse Engineering*. IEEE.
[10] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Integrating induction and deduction for finding evidence of discrimination. In *Proceedings of International Conference on Artificial Intelligence and Law*. ACM, 157–166.
[11] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of International Conference on Data Mining*. SIAM, 581–592.
[12] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. Data mining for discrimination discovery. *TKDD* 4, 2 (2010), 9:1–9:40.
[13] Ankit Shah, Rajesh Ganesan, Sushil Jajodia, and Hasan Cam. 2018. Understanding Trade-offs Between Throughput, Quality, and Cost of Alert Analysis in a CSOC. *IEEE Transactions on Information Forensics and Security* (2018).