Classification Rule Mining Supported by Ontology for Discrimination Discovery

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

***Abstract*—Discrimination discovery from data consists of designing data mining methods for the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records. Approaches based on classification rule mining consider items at a flat concept level, with no exploitation of background knowledge on the hierarchical and inter-relational structure of domains. On the other hand, ontologies are a widespread and ever increasing means for expressing such a knowledge. In this paper, we propose a framework for discrimination discovery from ontologies, where contexts of *prima-facie* evidence of discrimination are summarized in the form of generalized classification rules at different levels of abstraction. Throughout the paper, we adopt a motivating and intriguing case study based on discriminatory tariffs applied by the U.S. Harmonized Tariff Schedules on imported goods.**

I. Introduction

In social and legal sense, discrimination occurs in situations when members of a minority are treated unequally or less favorably than the ones of the majority group without regard to individual merits. Unfair behaviors have been observed in racial profiling and redlining, mortgage lending, consumer market, credit and housing, personnel selection and wages. Even though this problem has been surveyed for a long time by economists, sociologists and legal scholars, it has been only recently studied from the viewpoint of data mining – see the surveys [1]–[3].

A few studies focus on extracting knowledge discovery models to unveil and represent discriminatory treatments, e.g., in the form of classification rules ranked by legally-grounded interestingness measures. Their results however, are limited to considering all the attributes of a dataset at flat level, without taking into account the (implicit or elicited) hierarchical and inter-relational structure of data domains, which may lead to low expressivity. We claim that there is the need to discover discriminatory treatments at multiple levels of refinement, and possibly with support of semantics. The intriguing case study of the *U.S. Harmonized Tariff Schedules* (HTS) presented in the paper will support our claim.

On the other hand, ontologies are an intuitive, flexible, and effective means for the categorization of objects in a given domain. The distinguished advantage of ontology engineering is the offer of reasoning services, which can answer rich semantic questions about concepts and individuals on ontology, that cannot be solved by normal SQL queries. Therefore, ontology engineering appears to be a promising and convenient support for extending classification rule mining for discrimination discovery, especially in the semantics extent.

In this paper, we introduce a framework for the discovery of discrimination hidden in a dataset, in which an ontology is used to represent the domain of data under analysis. Basically, the ontology provides the knowledge, hierarchically organized, out of which patterns of discrimination are extracted according to the legal methodology of situation testing. Such a methodology has already been exploited by a few approaches for discrimination discovery [4]–[6]. The idea is to find out pairs of individuals with essentially the same characteristics except for some sensitive attribute, such as gender, which are treated differently. Closeness of characteristics is evaluated through a similarity measure that takes into account the distance between the individuals in a pair with respect to the hierarchy in the ontology. We also exploit the capability of ontology management systems of answering queries, specifically SWRL queries, to compute interestingness measures, namely support and confidence, of the extracted rules. While this is not an immediate technical advancement over re-using existing classification rule extraction algorithms, it demonstrates that the whole process of analysis can be coded within existing ontology management systems. Our overall system is implemented as a plug-in of the Protégé knowledge base framework.

This paper is organized as follows. In Section II, we review the related work on ontology and on discrimination discovery. Section III introduces concepts and notation. Section IV presents the case study and the theoretical basis of our approach. Section V illustrates the implementation aspects, while experiments are reported in Section VI. Finally, Section VII summarizes the contributions of the paper.

II. Related Work

Since ontology engineering is a method of conceptual modeling, in which semantic relationships among categories of beings are also formally defined, it has been used in a variety of knowledge mining systems in supporting more elaborated results, especially in the semantic extent. [7] initiated this direction by a suggestion of an object-oriented implementation for an explicit representation of a "dynamic" hierarchy to support the evolution of concept hierarchies during its cycle. A central trend in ontology mining is the exploitation of semantics obtained from the user's information in the semantic web for achieving more expressive mining models [8], [9].

On the other hand, extensions of classic data mining models to reason over ontologies have been proposed (*semantic data mining* [10]), and the overall data mining process can itself be supported by ontologies (*semantic meta-mining* [11]).

This work does not aim at building novel theories for the foundations of ontology engineering/mining. Instead, it uses ontology as a representation for the domain of analysis of discrimination discovery, in order to provide semantic support during the extraction of patterns of discrimination. First, hierarchies among concepts are exploited to extract patterns of discriminatory behaviors at different levels of abstraction. Second, the distance between two concepts in the hierarchy is adopted as a measure of similarity in comparing (different) treatments between similar individuals.

The research on discrimination discovery from data consists of designing data mining methods for the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records [1]. The aim is to unveil contexts of possible discrimination suffered by protected-by-law groups in such contexts. The legal principle of under-representation has inspired existing approaches for discrimination discovery based on pattern mining. [12] proposes to extract classification rules such as RACE=BLACK, PURPOSE=NEW_CAR $\rightarrow$ CREDIT=NO, called *potentially discriminatory* (PD) 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 with respect to the positive decision (here, credit granting). The approach has been implemented on top of an Oracle database by relying on tools for frequent itemset mining [13]. The impact of the choice of the discrimination measure at hand is discussed in [14]. The main limitation of the approach is that there is no control of the characteristics of people from the protected group vs the rest of people in the context, e.g., in the above example, the capacity to repay the loan. This results in an overly large number of PD rules that need to be further screened. [4], [5] exploit the idea of situation testing: for each individual of the protected group with a negative decision outcome, one looks for testers in the dataset with similar, legally admissible, characteristics, apart from being or not in the protected group. If one can observe significantly different decision outcomes between the testers of the protected group and the testers of the unprotected group, one can ascribe the negative decision of the individual to a bias against the protected group, hence labeling the individual as discriminated. This paper follows a similar approach and extends it significantly by exploiting ontologies and their data structuring capabilities in organizing data and their background knowledge. Our contribution is orthogonal to other extensions of situation testing, such as the causality extensions of [6].

Finally, we mention two related topic, which are out of the scope of this paper. One is discrimination prevention (or fairness) in data mining, where the issue is to extract data mining models (typically, classifiers) that trade off accuracy for non-discrimination. The other is indirect discrimination discovery, which tackles the problem under the further assumption that the dataset under analysis does not record information about membership of individuals to protected-by-law groups. We refer to the surveys [1]–[3] for details on such two problems.

III. Preliminaries

A. (Generalized) Association Rules

Association rules were originally introduced in the context of relational databases. Given a relation $\mathcal{R}$, an item is a term $a = v$, where a is an attribute of $\mathcal{R}$ and v belongs to the domain of values of a. An itemset is a set of items. As usual in the literature, we write I, J for the itemset $I \cup J$. Tuples in $\mathcal{R}$ can be readily represented as itemsets. The support of an itemset I is the fraction of tuples in $\mathcal{R}$ covering I: $supp(I) = |\{\ t \in \mathcal{R} \mid I \subseteq t\ \}|/|\mathcal{R}|$, where $|\ |$ is the cardinality operator.

An association rule is an expression $I \rightarrow J$, where I and J are itemsets, with $I \cap J = \emptyset$. I is called the *antecedent* and J is called the *consequent* of the rule. We say that $I \rightarrow J$ is a *classification rule* if J is a singleton $a = v$, where a is the class attribute in the relation $\mathcal{R}$. The support of $I \rightarrow J$ is defined as: $supp(I \rightarrow J) = supp(I, J)$; and its confidence is: $conf(I \rightarrow J) = supp(I, J)/supp(I)$. Support and confidence range over $[0, 1]$.

Generalized association rules extend association rules by exploiting an *is-a* hierarchy over items. Antecedent and consequent can now include items at different levels of the hierarchy, and rules can be compared on the basis of the levels of items appearing in them. We say that an itemset $\hat{I}$ is an *ancestor* of an itemset I if $\hat{I} \neq I$ and $\hat{I}$ can be obtained by repeatedly replacing one or more items in I with a common ancestor in the *is-a* hierarchy. The rules $\hat{I} \rightarrow J$, $\hat{I} \rightarrow \hat{J}$, and $I \rightarrow \hat{J}$ are called *ancestors* of the rule $I \rightarrow J$.

Since the seminal papers introducing association rules [15] and generalized association rules [16], many well explored algorithms have been designed in order to extract (generalized) association rules with a user-specified minimum support (*minsupp*), and minimum confidence (*minconf*). A survey of frequent pattern mining is [17]; a summary of interestingness measures for association rules is reported in [18]; a repository of implementations is [19].

B. Ontology

The definition of an ontology is still debated, but the mostly quoted one is that an *ontology* is a formal, explicit specification of a domain conceptualization [20]. For our purposes, an ontology is a description logic (DL) knowledge base $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$. The TBox (Terminological Box) $\mathcal{T}$ defines hierarchies over a set of *concepts* $C_1, C_2, \ldots, C_p$, denoting classes of individuals, and *roles* or *object properties* $R_1, R_2, \ldots, R_m$, denoting binary relationships between individuals. Assertions take the form of:

- *inclusions* between concepts and roles: $C_i \sqsubseteq C_j$ denoting that C_j is more general (the super-class) than C_i (the sub-class), i.e., every individual satisfying C_i also satisfy C_j; and similarly for roles, $R_i \sqsubseteq R_j$;
- *equalities*: $C_i \equiv C_j$, stating that C_i and C_j comprise the same individuals, and similarly for roles $R_i \equiv R_j$.

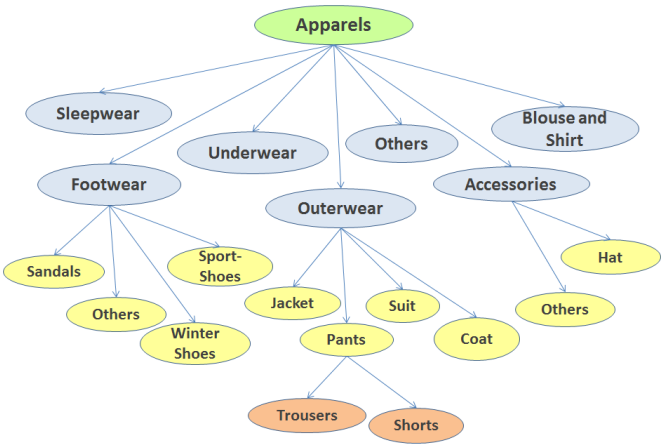

Fig. 1. The HTS hierarchy.

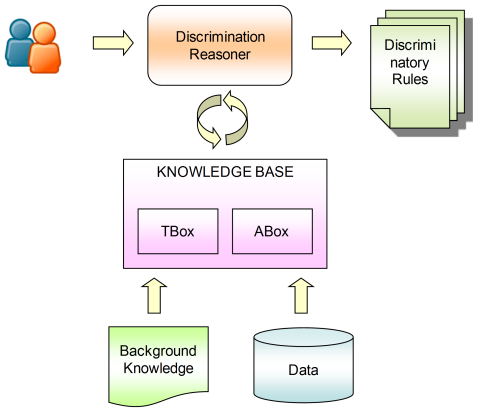

Fig. 2. A framework of discrimination discovery.

The ABox (Assertional Box) $\mathcal{A}$ contains factual assertions about concrete individuals:

- *concept assertions*: $C(a)$ states that the individual a belongs to concept C;
- *role assertions*: $R(a_1, a_2)$ states that a_2 is the filler of the role R for a_1.

We reason on an ontology by querying the knowledge base for *instance checking*: $\mathcal{O} \models C(a)$ (is the individual a an instance of class C?); *relation checking*: $\mathcal{O} \models R(a, b)$ (does the role R include the pair (a, b)?); and *subsumption checking*: $\mathcal{O} \models C_i \sqsubseteq C_j$ (is a concept C_i included in another C_j?). We say that a is a *direct instance* of a concept C_j if $\mathcal{O} \models C_j(a)$ and $\mathcal{O} \not\models C_i(a)$ for all C_i such that $\mathcal{O} \models C_i \sqsubseteq C_j$, namely for all sub-classes of C_j.

IV. Discrimination Discovery from Ontologies

A. The HTS Case Study

Throughout the paper, we will be illustrating concepts, problems and proposed solutions by means of a running example on the publicly available U.S. Harmonized Tariff Schedules (HTS) dataset [21]. The HTS is a tariff classification system for merchandize imported in the U.S., including nomenclatures (names), descriptions for goods, and formulae for calculating tariff rates. It consists of nearly 900 different categories of apparel. The structure of the domain of items is formally conceptualized by the hierarchy (TBox) in Fig. 1. The newspaper article [22] has firstly pointed out that tariffs on men's and women's garments are different for no apparent reason. Globally, the U.S. government imposes a 14 percent tariff on women's garments, but only 9 percent on men's ones. [23] calculated that U.S. importers, and ultimately U.S. female consumers, overpaid more than 1.3 billion dollars in the biennium 2005-2006 due to discriminatory tariffs. The legal context of gender discrimination in HTS tariffs is discussed in [24], which covers the *Totes-Isotoner Corp. v. U.S.* case, and in [25], which covers the *Rack Room Shoes Inc. and Forever 21 Inc. vs U.S.* case. This last case was concluded in 2014, when the U.S. Supreme Court and the U.S. Court of Appeals for the Federal Circuit affirmed the decision of the Court of International Trade to dismiss challenges to alleged gender discrimination in HTS tariffs. The plaintiffs' had argued that US Federal Government tariffs on apparel and footwear were discriminatory since those tariff rates were based on gender, rather than non-gender factors like the composition of materials, the weight of materials, the size of an article, or the function of an article. As observed in [26], "the courts may have concluded that Congress had no *discriminatory intent* when ruling the HTS, but there is little doubt that gender-based tariffs have *discriminatory impact*".

B. A Framework for Discrimination Discovery

We propose a framework for discrimination discovery whose main components are reported in Fig. 2. The framework is centered around a knowledge base in the form of an ontology. Data is populated from relational databases or external resources. Background knowledge is modelled by a domain expert, while patterns of possible discrimination are specified by an anti-discrimination analyst by defining a few concepts of interest, as described in Section IV-C. The system for discrimination discovery, called discrimination reasoner and described in Section IV-D, extracts from the ontology individuals that satisfy patterns of possible discrimination and that, according to the legal methodology of situation testing, can be considered as discriminated. Starting from such a set, generalized classification rules summarizing contexts of discrimination are calculated, and ranked on the basis of the number of discriminated individuals (rule support), and on the precision of the summarization (rule confidence).

C. Ontology Structure for Discrimination Discovery

Typically, an ontology $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$ is designed by domain experts (as per the TBox), and populated (as per the ABox) from external data sources by means of schema-mapping or ETL processes. As an alternative, the data under analysis is already stored in an ontology. For the HTS running example, relational data has been first converted into XML format,

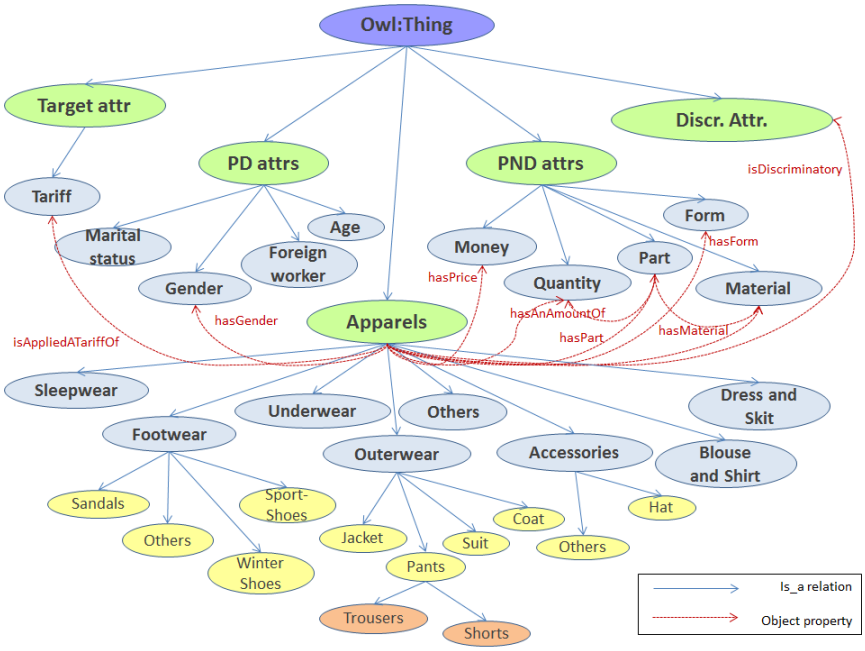

Fig. 3. HTS ontology

which, in turn, has been loaded as individuals of an ontology of taxed garments shown in Fig. 3. Categories of clothing items are gathered in the *Apparels* concept, while other relational attributes are modelled as object properties, e.g., *hasGender*, *hasMaterial*, *hasPrice*. We assume that the concepts described next are at top level in an ontology designed for supporting discrimination discovery. They are set up by the anti-discrimination analyst as a preliminary step of the analysis.

Relevant concepts. While ontologies often contain a large set of concepts, which intertwine to describe a knowledge domain, we assume a main concept that is the subject of the discrimination discovery problem. In our running example, this is the *Apparels* concept, because we are interested in unveiling disparate taxation practices in the HTS system. We call such a main concept and its sub-classes the *relevant concepts*.

PD and PND attributes. Object properties of relevant concepts link to *Potentially Discriminatory* (PD) or *Potentially Non-Discriminatory* (PND) groups of concepts, on the basis of whether they refer or not to sensitive or non-sensitive personal attributes respectively. The terminology PD-PND is borrowed from [12]. According to anti-discrimination laws, PD attributes include gender, age, marital status, nationality, ethnicity, and so on. In our running example, the only sensitive attribute is the gender the garment is produced for (specifically, we will consider the female gender, which is protected by the anti-discrimination laws), and the object property *hasGender* connects *Apparels* to the PD concept *Gender*. As another example, the property *hasPrice* connects *Apparels* to the PND, or non-sensitive, concept *Money*.

Target attribute. A specific concept, called the *target attribute*, is assumed to model the decision that may have discriminatory impacts. In our example, this is the taxation applied to a garment, expressed as a percentage value. In Fig. 3, the *Tariff* concept models the target attribute, and the *isAppliedATariffOf* object property specifies the amount of tax applied to a garment. We use the meta-variable τ to represent the decision value for an individual. The domain of τ can be discrete, e.g., a yes/no decision to a loan application, or continuous, as in our running example. In the HTS example, a taxation tariff τ_1 is worse than τ_2 if $\tau_1 > \tau_2$.

Finally, we assume a *discriminatory attribute* concept, taking only yes/no values, which is linked to relevant concepts via the $isDiscriminatory$ object property. Intuitively, such a property is intended to label as discriminated or not each individual belonging to relevant concepts on the basis of its PD, PND, and target attributes. In our example, it labels as discriminated or not each apparel on the basis of the gender it is produced for, of its characteristics (material, form, quantity, etc.), and on the basis of the tariff applied to it. The formal definition of the $isDiscriminatory$ property is presented in the next subsection.

D. Modelling Discrimination

In social and legal sense, discrimination occurs in contexts when a person is treated unequally or less favorably than another in the same conditions. We model contexts by expressions Δ of the form:

$$\Delta = C \wedge R_1 \wedge ... \wedge R_t$$

where C is a relevant concept, at any level of the hierarchy, and $R_1, \ldots, R_t$ are object properties that link C to PND properties. As an example, the context:

$$Outerwear \wedge hasMaterial \wedge hasForm \quad (1)$$

mentions a kind of garments, $Outerwear$, that are made by some (not further specified) material and that have some (not further specified) form.

The legal methodology of *situation testing* [27], [28], also known as field experiments or auditing, follows a quasi-experimental approach to investigate for the presence of discrimination by controlling the factors that may influence decision outcomes. It consists of using pairs of *testers* (also called *auditors*), who have been matched to be similar on all characteristics that may influence the outcome except race, gender, or other grounds of possible discrimination. The tester pairs are then sent into one or more situations in which discrimination is suspected, e.g., to rent an apartment or to apply for a job, and the decision outcome is recorded. A different outcome between the paired testers is then considered a *prima-facie* evidence of discrimination.

We rephrase the situation testing approach as follows. The concept C in Δ is used to select individuals as candidate testers, while the object properties $R_1, \ldots, R_t$ in Δ are used to compare individuals to select pairs of similar (w.r.t. those properties) characteristics. More formally, we define the *realization set* of Δ over the ontology the set of pairs $(x, \{y_1, \ldots, y_t\})$, called *realizations*, such that:

- $\mathcal{O} \models C(x)$
- for $i = 1 \ldots t$, $\mathcal{O} \models R_i(x, y_i)$

Therefore, the individual x is a candidate tester, while $y_1, \ldots, y_t$ are its properties in the context Δ. We look for another realization that can be paired with it by introducing a notion of similarity over ontologies. The path-distance of two concepts C_1 and C_2 in a hierarchy is defined as the number of edges in the shortest path between C_1 and C_2 (and in particular, it is 0 when $\mathcal{O} \models C_1 \equiv C_2$). The path-distance of two individuals x and y, denoted as $p(x, y)$, is the path-distance of the concepts C_x and C_y, whose x and y are direct instances of. The basic similarity measure $sim()$ between individuals x, y is defined as a monotonically decreasing function of their path distance. In experiments, we set:

$$sim(x, y) = 2^{-p(x,y)}$$

It shows that if two individuals belong to the same concept, their similarity is 1, otherwise it exponentially decreases with their path-distance. We extend similarity to a pair of realizations $\zeta_1 = (x_1, \{y_i^1\}_{i=1}^t)$, $\zeta_2 = (x_2, \{y_i^2\}_{i=1}^t)$ as follows:

$$sim(\zeta_1, \zeta_2) = \frac{sim(x, y) + \sum_{i=1}^{t} ssim(y_i^1, y_i^2)}{t + 1}$$

where $ssim()$ is a similarity function between scalar values. For nominal domains, it boils down to an equality indicator: $ssim(a, b) = 1$ if $a = b$, and $ssim(a, b) = 0$ otherwise. For continuous values, normalized in the interval $[0, 1]$, we assume $ssim(a, b) = 1 - |a - b|$. It can be easily seen that sim ranges over $[0, 1]$.

The similarity measure sim is used to search for pairs of testers. Given a candidate ζ_1, we look for ζ_2 such that[1] $sim(\zeta_1, \zeta_2) \approx 1$, namely a realization ζ_2 with a similar individual and similar object properties as ζ_1. With reference to the description (1), we look for individuals of the $Outerwear$ concept that are similar with respect to the material they are made of (e.g., synthetic fiber) and form (e.g., knitted).

The legal notion of "different treatment", which is the basis for claiming discrimination, is interpreted as follows. Consider a realization $\zeta_1 = (x_1, \{y_i^1\}_{i=1}^t)$. Let s_{ζ_1} be the PD attribute value(s), and τ_1 the target attribute value for the individual x_1. In our example, s_{ζ_1} is the gender the garment x_1 is produced for, and τ_1 is the tariff applied. We label ζ_1 as discriminated if there exists a realization ζ_2 close to ζ_1 (namely, $sim(\zeta_1, \zeta_2) \approx 1$), with $\zeta_2 = (x_2, \{y_i^2\}_{i=1}^t)$, for which the PD attribute value s_{ζ_2} of x_2 is different from s_1 (in our example, it refers to another gender), and the target attribute value τ_{ζ_2} for x_2 is better than τ_{ζ_1} (in our example, taxation τ_{ζ_2} is lower than τ_{ζ_1}, i.e., $\tau_{\zeta_2} < \tau_{\zeta_1}$). We formalize such a situation testing reasoning by introducing the following *discriminatory indicator*:

$$\theta_\Delta(\zeta_1) = \begin{cases} \text{yes} & \text{if there exists } \zeta_2 \text{ such that} \\ & \quad sim(\zeta_1, \zeta_2) \approx 1 \text{ and} \\ & \quad s_{\zeta_2} \neq s_{\zeta_1} \text{ and } \tau_{\zeta_2} < \tau_{\zeta_1} \\ \text{no} & \text{otherwise} \end{cases}$$

We are now in the position to label an individual x_1 as discriminated or not by setting the object property $isDiscriminatory$ of x_1 to the value of $\theta_\Delta(\zeta_1)$, where ζ_1 is the realization whose first element is x_1 and the second element is the set of its object properties values w.r.t. Δ. With reference to (1), an outerwear garment produced for women is labeled as discriminated if the amount of taxation applied is higher than the one applied to the same, or similar as per material and form, garment produced for men. Summarizing, for a fixed context Δ, the $isDiscriminatory$ property is populated for individuals in its realization set according to the discriminatory indicator θ_Δ. However, we are still faced with the problem of extracting a high-level, intelligible, characterization of the individuals labeled as discriminated. We resort to (generalized) classification rule mining by extracting rules of the form:

$$C(?x) \wedge R_{i_1}(?x, v_1) \wedge \ldots \wedge R_{i_n}(?x, v_t) \rightarrow isDiscriminatory(?x, yes)$$

that we call *discriminatory classification rules*. They are extracted from the subset of realizations having PD property values (in our case, from garments produced for women) enriched with their discriminatory indicator value. Here, $v_1, \ldots, v_n$ are PND attribute values characterizing specific properties of an individual $?x$ belonging to the

[1] In our experiments on the HTS dataset, we have observed no significant difference between $sim(\zeta_1, \zeta_2) = 1$ and $sim(\zeta_1, \zeta_2) \geq 0.95$, and thus we simply require $sim(\zeta_1, \zeta_2) = 1$, namely we look for garments of the same type and with exactly the same characteristics. However, the sensitivity of the results to the approximation "≈ 1" may change from a domain to another.

for all relevant concepts C **do**
  let $R_1, \ldots, R_t$ be all object properties of C
  let $\Delta = C \wedge R_1 \wedge \ldots, R_t$
  let Γ = realizations(Δ)
  retract all $isDiscriminatory$
  for all PD $\zeta = (x, \{y_i\}_{i=1}^t) \in \Gamma$ **do**
    assert $isDiscriminatory(x, \theta_\Delta(\zeta))$
  end for
  extract rules with $supp \geq minsupp$
    and $conf \geq minconf$ of the form
      $C(?x) \wedge R_{i_1}(?x, v_1) \wedge ... \wedge R_{i_n}(?x, v_t)$
        $\rightarrow isDiscriminatory(?x, yes)$
end for

Fig. 4. Discrimination rule generation algorithm.

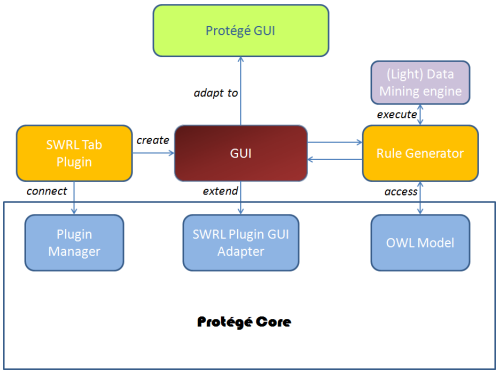

Fig. 5. Connecting to Protégé.

concept C. Object properties $R_{i_1}, \ldots, R_{i_n}$ are a subset[2] of $R_1, \ldots, R_t$, or, in symbols, $\{i_1, \ldots, i_n\} \subseteq \{1, \ldots, t\}$. Classical interestingness measures, such as support and confidence, are applicable. The support of a rule is the percentage of individuals that satisfy both sides of the rule, while the confidence is the percentage of individuals satisfying the left hand side that are labeled as discriminated. Notice that we adopt a human-readable SWRL (Semantic Web Rule Language) syntax [29]. In our running example, the rule:

$$Outerwear(?x) \wedge hasMaterial(?x, \text{“}synthetic\ fiber\text{”}) \wedge hasForm(?x, \text{“}knitted\text{”}) \rightarrow isDiscriminatory(?x, yes)$$
$$(supp = 5\%, conf = 100\%)$$

states that 5% of outerwear apparels produced for women are made of synthetic fiber, they are knitted and with a tariff that is higher than those produced for men; and that 100% of the outerwear apparels produced for women that are made of synthetic fiber and knitted incur in such a disparate taxation.

V. Algorithm and Implementation

A. Rule Extraction Algorithm

The pseudo-code of the overall rule generation algorithm is shown in Fig. 4. For each concept C that is specified by the user as relevant for the analysis, the following steps are executed. First, all object properties of the concept are retrieved from the ontology to build a pattern Δ. The set of realizations Γ of such a pattern is then calculated, consisting of all individuals belonging to concept C and of its PND attributes. For each realization with PD property values, the discriminatory indicator is calculated, and the object property $isDiscriminatory$ of the individual is set according. After the inner loop, each individual in the realization set is then labelled as discriminated or not, on the basis of the legal methodology of situation testing. A summarization of the conditions for which individuals are discriminated can be extracted via discriminatory classification rules, as computed in the algorithm of Fig. 4. It is worth noting that such rules could be computed by standard (generalized) classification/association rule mining algorithms by exporting the dataset consisting of tuples $(x, y_1, \ldots, y_t, \theta_\Delta(\zeta))$ for each realization $\zeta = (x, \{y_i\}_{i=1}^t)$. In our actual implementation, described next, we rather exploit the capability of ontology management systems of answering queries, specifically SWRL queries. We have adopted the SWRL syntax for classification rules exactly with the purpose of using such a query language for computing the basic support counting primitives of (Apriori-based) association rule mining. While this is not an immediate technical advancement over reusing existing classification rule extraction implementations, it demonstrates that the whole process of analysis can be coded within an ontology management system, thus supporting modern approaches to semantic meta-mining [11].

B. Implementation

The proposed approach for discrimination discovery has been implemented as an extension of Protégé (`http://protege.stanford.edu`), an open-source ontology editor and knowledge base framework. We have developed a plugin called *RuleGenerator* by exploiting Protege OWL APIs and the SWRLTab, a development environment for working with SWRL rules in Protege-OWL. The overall system is shown in Fig. 5. The RuleGenerator plugin accesses the Protégé OWLModel module for gathering information on the ontology's structure and individuals. The execution of the discriminatory rule extraction algorithm is demanded to a (light) data mining engine. In fact, while we currently exploit classification rules to represent patterns of discovered discrimination, we point out that other data mining models, in particular classification models, can be used to the purpose. In this sense, in the place of our particular implementation of classification rule mining, other data mining system and libraries could be adopted. Finally, we have implemented an extension of the Protégé GUI, via the SWRL GUI Adapter, to provide the user with a specialized sub-tab to interact with the discrimination discovery module.

[2]Since differences between individuals are controlled by means of the discriminatory indicator, a discriminatory rule involving only a subset of $R_1, \ldots, R_t$ can be safely read as the fact that the individuals in such a subset are discriminated (in proportion to the confidence of the rule). This conclusion cannot be made by the early approaches using classification rule mining [12].

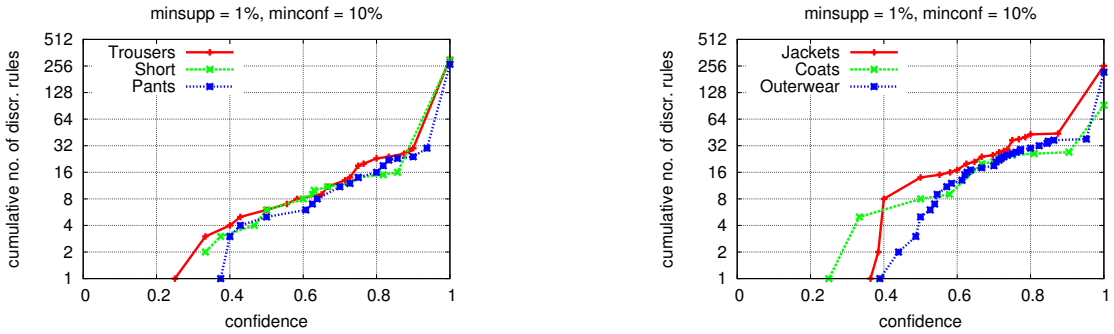

Fig. 6. Cumulative distribution of discriminatory classification rules by confidence.

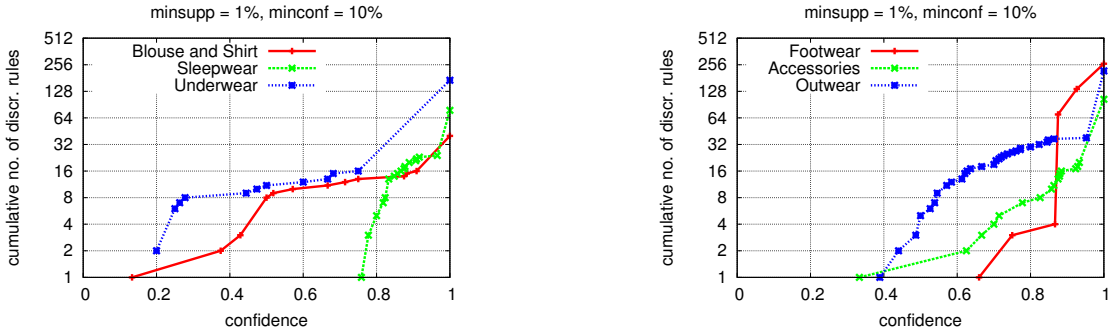

Fig. 7. Cumulative distribution of discriminatory classification rules by confidence.

VI. Analysing the HTS Dataset

In this section, we apply the rule extraction algorithm to the HTS ontology reported in Fig. 3, adopting a minimum support threshold of 1% and a minimum confidence threshold of 10%. Let us recall that there is only one PD attribute value, namely the female gender a garment is produced for, that the target attribute is the taxation tariff applied, and that support and confidence of a discriminatory classification rule can be respectively interpreted as the proportion of discriminatorily taxed apparels produced for women that are recalled by the rule; and the precision of the discrimination conclusion of the rule given that the antecedent holds. The immediate advantage of relying on an ontology is that discriminatory rules at different levels of abstraction can be considered. As an example, the extracted rule:

$$Shorts(?x) \wedge hasMaterial(?x, \text{“}fine\ animal\ hair\text{”}) \rightarrow isDiscriminatory(?x, yes)$$

with a confidence $conf = 66.67\%$ can be directly compared with its ancestor rule at the grand-parent level (the concept *Shorts* is a sub-class of *Outerwear*):

$$Outerwear(?x) \wedge hasMaterial(?x, \text{“}fine\ animal\ hair\text{”}) \rightarrow isDiscriminatory(?x, yes)$$

which has a lower confidence of $conf = 57.78\%$. Intuitively, this can be read as the fact that tariffs for shorts are more discriminatory than the ones at the level of outerwar. As observed in footnote 2, the two rules above concern pairs of similar garments, i.e., shorts/outerwar with the same (PND) object properties, but with different gender property. Hence, they already control for characteristics other than the one explicitly mentioned in the rule, i.e., $hasMaterial$, which now assumes the expected role of *summarizing* under which conditions gender discriminatory tariffs apply.

Fig. 6 and Fig. 7 show the cumulative number of discriminatory classification rules extracted for a minimum confidence of 10%. Fig. 6 (left) restricts to trousers and shorts and to their parent concept, namely pants (see the hierarchy in Fig. 3). Most of the rules have a confidence of 100%, namely they summarize contexts of garments with different male/female tariffs. Also, the three distributions are rather similar, which means that the conditions for discriminatory tariffs are not very specific of trousers or shorts or pants, but rather of their properties, such as the materials they are made of. Fig. 6 (right) shows a plot with characteristics similar to the previous one, but now comparing higher level concepts, namely jackets, coats, and their parent concept, i.e., outerwear. Fig. 7 instead reports the distribution of six concepts at the highest level. Some differences are now visible. Sleepwear and footwear appear to have the steepest distributions, which means that the contexts of discrimination summarized by their classification rules are very precise. In addition, footwear has the highest number of rules, hence exhibiting the largest number of

discriminatory contexts[3]. On the contrary, the blouse and shirt concept shows the lowest number of discriminatory contexts. Summarizing, the plots in Fig. 6 and Fig. 7 provide interesting hints to an anti-discrimination analyst on how to prioritize subsequent analyses.

Finally, we mention that, due to the small size of the HTS dataset, running time and memory occupation of the rule extraction algorithm are negligible on current PCs. In general, however, the bulk of the resources needed by the algorithm in Fig. 4 is due to the classification rule extraction phase. While our SWRL-based implementation is non-optimal, any optimized implementation can be adopted as the "Data mining engine" in Fig. 5.

VII. Conclusion

The use of data mining is revealing its strengths in the analysis of discrimination data, improving over traditional statistical techniques, such as regression and significance tests, towards the discovery of contexts of *prima-facie* evidence of discrimination. Such context have to be further investigated by the anti-discrimination analyst, possibly discussing each single context with a legal expert. It is then of primary importance that the number of contexts presented by a data mining system is: (1) limited in number; and, (2) as much as expressive as possible. Compared to the early approaches based on classification rule mining [12], recent advancements [4] have covered the former requirement, by relying on the legal methodology of situation testing. In this paper, in addition to adopting a form of situation testing as well, we aimed at satisfying also the latter requirement.

We have proposed here a general framework to discover discrimination in the form of generalized classification rules with an ontology support. Discrimination is formalized through the measure of similarity of individuals sharing a common background of non-sensitive properties. By exploiting the hierarchy and object properties of the ontology, different levels of abstraction can be obtained in the analysis. Positive experimental results on the HTS case study show the potential and flexibility of the framework. As a future work, a further exploitation of ontologies concerns a fuller use of their reasoning capabilities. The properties of individuals in the HTS case study are defined directly by means of ground values, but in general we can expect that they may be indirectly defined via rules, in particular in case of dynamically inferred properties.

References

[1] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 5, pp. 582–638, 2014.

[2] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, 2016, available at SSRN: http://ssrn.com/abstract=2477899.

[3] I. Žliobaitye, "A survey on measuring indirect discrimination in machine learning," *arXiv preprint arXiv:1511.00148v1*, 2015.

[4] 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 Int. Conf. on Knowledge Discovery and Data Mining (KDD 2011)*. ACM, 2011, pp. 502–510.

[5] A. Romei, S. Ruggieri, and F. Turini, "Discrimination discovery in scientific project evaluation: A case study," *Expert Systems with Applications*, vol. 40, no. 10, pp. 6064–6079, 2013.

[6] L. Zhang, Y. Wu, and X. Wu, "Situation testing-based discrimination discovery: A causal inference approach," in *Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI 2016)*, 2016, pp. 2718–2724.

[7] F. Scott and L. Ling, "An object-oriented approach to multi-level association rule mining," in *Proc. of CIKM 1996*. ACM, 1996, pp. 65–72.

[8] B. Glimm and H. Stuckenschmidt, "15 years of semantic web: An incomplete survey," *KI*, vol. 30, no. 2, pp. 117–130, 2016.

[9] P. Ristoski and H. Paulheim, "Semantic web in data mining and knowledge discovery: A comprehensive survey," *J. Web Sem.*, vol. 36, pp. 1–22, 2016.

[10] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," in *Proc. of the IEEE Int. Conf. on Semantic Computing (ICSC 2015)*, 2015, pp. 244–251.

[11] C. M. Keet, A. Lawrynowicz, C. d'Amato, A. Kalousis, P. Nguyen, R. Palma, R. Stevens, and M. Hilario, "The data mining OPtimization ontology," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 32, pp. 43 – 53, 2015.

[12] 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.

[13] ——, "DCUBE: Discrimination discovery in databases," in *Proc. of the ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2010)*, A. K. Elmagarmid and D. Agrawal, Eds. ACM, 2010, pp. 1127–1130.

[14] D. Pedreschi, S. Ruggieri, and F. Turini, "A study of top-k measures for discrimination discovery," in *Proc. of ACM Int. Symposium on Applied Computing (SAC 2012)*, ACM, 2012, pp. 126–131.

[15] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases." in *Proc. of Int. Conf. on Very Large Data Bases (VLDB 1994)*. Morgan Kaufmann, 1994, pp. 487–499.

[16] R. Srikant and R. Agrawal, "Mining generalized association rules," in *Proc. of Int. Conf. on Very Large Data Bases (VLDB 1995)*. Morgan Kaufmann, 1995, pp. 407–419.

[17] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 55–86, 2007.

[18] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys*, vol. 38, no. 3, 2006.

[19] B. Goethals, "Frequent itemset mining implementations repository," 2010, http://fimi.cs.helsinki.fi.

[20] T. R. Gruber, "A translation approach to portable ontologies," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[21] U.S. Federal Legislation, "U.S. Harmonized Tariff Schedules," 2008, 43 FR 38295, http://hts.usitc.gov.

[22] M. Barbaro, "In apparel, all tariffs aren't created equal," *New York Times*, April 28, 2007.

[23] M. Gersper and T. Gould, "Gender and age discrimination costs U.S. importers billions," Oct. 23, 2007, http://apparel.edgl.com/news.

[24] A. P. Bertero, "Gender discrimination in the United States tariff schedule: Does unequal tariff treatment of mens and womens products constitute an equal protection violation?" 2011, working Paper, http://ssrn.com.

[25] J. Lewis, "Gender-classified imports: Equal protection violations in the Harmonized Tariff Schedule of the United States," *Cardozo Journal of Law & Gender*, vol. 18, pp. 171–180, 2011.

[26] L. L. Taylor and J. Dar, "Fairer trade, removing gender bias in US import taxes," 2015, available from http://hdl.handle.net/1969.1/153774.

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

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

[29] W3C Member Submission, "SWRL: A semantic web rule language combining OWL and RuleML," 21 May 2004, http://www.w3.org/Submission/SWRL.

[3] Since the grain of the dataset is a tariff applied to a garment, this does not imply that footwear exhibits the largest discrimination in terms of male/female tariff *differences*, or in terms of *market-value* of the discriminated garments.