A KDD Process for Discrimination Discovery

Salvatore Ruggieri(✉) and Franco Turini

Dipartimento di Informatica, Università di Pisa,
Largo B. Pontecorvo 3, 56127 Pisa, Italy
{ruggieri,turini}@di.unipi.it

Abstract. The acceptance of analytical methods for discrimination discovery by practitioners and legal scholars can be only achieved if the data mining and machine learning communities will be able to provide case studies, methodological refinements, and the consolidation of a KDD process. We summarize here an approach along these directions.

1 The Way Ahead

Data mining and machine learning approaches to social discrimination discovery from historical decision records have recently gained momentum – see the surveys [1,6,8]. Most of the proposals are restricted to investigations of novel algorithms and models. In our opinion, the field still need major advancements towards: first, experimentation with real data; second, methodological refinements in compliance with legal rules and ethical principles; and third, the consolidation of a KDD process of discrimination discovery. Solving these issues is essential for the acceptance of discrimination discovery methods based on data mining and machine learning by practitioners and legal scholars. In the paper [7] we contributed in all those aspects by presenting: a case study on a real dataset about gender discrimination in scientific research proposals; an instantiation of the methodological approach of [4] based on the legal methodology of situation testing; a generalization of the case study to a KDD process in support of discrimination discovery. This is a summary of the last contribution.

2 Not only an Algorithm: An Analytical Process

Since personal data in decision records are highly dimensional, i.e., characterized by many multi-valued variables, a huge number of possible contexts may, or may not, be the theater for discrimination. In order to extract, select, and rank those that represent actual discriminatory behaviors, an anti-discrimination analyst should apply appropriate tools for pre-processing data, extracting prospective discrimination contexts, exploring in details the data related to the context, and validating them both statistically and from a legal perspective. Discrimination discovery consists then of an iterative and interactive process. Iterative because, at certain stages, the user should have the possibility of choosing different algorithms, parameters, and evaluation measures or to iteratively repeat some

B. Berendt et al. (Eds.): ECML PKDD 2016, Part III, LNAI 9853, pp. 249–253, 2016.
DOI: 10.1007/978-3-319-46131-1_28

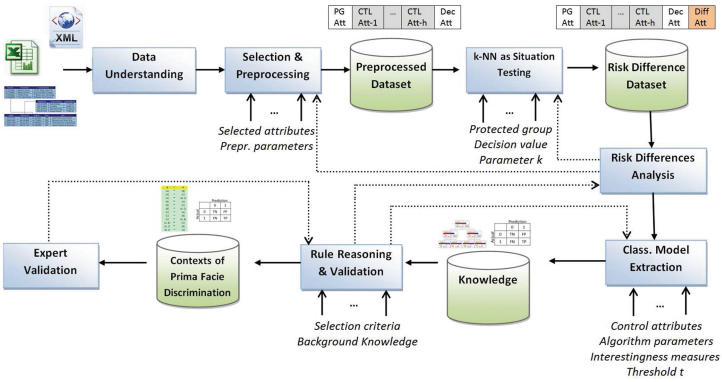

Fig. 1. The KDD process of situation testing for discrimination discovery.

steps to unveil meaningful discrimination patterns. Interactive because several stages need the support of a domain expert in making decisions or in analysing the results of a previous step. We propose in [7] to adopt the process reported in Fig. 1, which is specialized in the use of the situation testing for extracting contexts of possible discrimination. The process has been abstracted from the case study presented in the paper regarding gender discrimination in a dataset of scientific research proposals, and it consists of four major steps.

Data Understanding and Preparation. We assume a collection of data sources storing historical decisions records in any format, including relational, XML, text, spreadsheets or any combination of them. Standard data pre-processing techniques (selection, cleansing, transformation, outlier detection) can be adopted to reach a pre-processed dataset consisting of an *input relation* as the basis for the discrimination analysis. The grain of tuples in the relation is that of an individual (an applicant to a loan, to a position, to a benefit). Three groups of attributes are assumed to be part of the relation:

protected group attributes: one or more attributes that identify the membership of an individual to a protected group. Attributes such as sex, age, marital status, language, disability, and membership to political parties or unions are typically recorded in application forms, curricula, or registry databases. Attributes such as race, skin color, and religion may be not available, and must be collected, e.g., by surveying the involved people;

decision attribute: an attribute storing the decision for each individual. Decision values can be nominal, e.g., granting or denying a benefit, or continuous, e.g., the interest rate of a loan or the wage of a worker;

control attributes: one or more attributes on control factors that may be (legally) plausible reasons that may affect the actual decision. Examples include the financial capability to repay a loan, or the productivity of an applicant worker.

Risk Difference Analysis. Randomized experiments are the gold-standard for inferring causal influences in a process. However, randomized experiments are not possible or not cost-effective in discrimination analysis. An example of quasi-experimental approaches is situation testing [2], which uses pairs of testers who have been matched to be similar on all characteristics that may influence the outcome except race, gender, or other grounds of possible discrimination. In a legal setting, the tester pairs are then sent into one or more situations in which discrimination is suspected. In observational studies, [4] proposes to simulate the approach by contrasting the decisions of the tuple neighbors. For each tuple of the input relation denoting an individual of the protected group, the additional attribute *diff* is calculated as the risk difference between the decisions of its k nearest-neighbors of the protected group and the decisions for its k nearest-neighbors of the unprotected group (see Fig. 2). Risk difference is a measure of the degree of discrimination suffered by an individual. We call the output of the algorithm the *risk difference relation.* The value k is a parameter of the algorithm. A study of the distribution of *diff* for a few values of k is required. This means iterating the calculation of the *diff* attribute. Exploratory analysis of *diff* distributions may also be conducted to evaluate risk differences at the variation of: the protected group under consideration, e.g., discrimination against women or against youngsters; the compound effects of multiple discrimination grounds, e.g., discrimination against young women vs discrimination against women or youngsters in isolation; the presence of favoritism towards individuals of a dominant group, e.g., nepotism.

Discrimination Model Extraction. By fixing a threshold value t, an individual $\mathbf{r}$ of the protected group is then *labeled* as discriminated or not on the basis of the condition $diff(\mathbf{r}) \geq t$. We introduce a new boolean attribute *disc* and set it to true for a tuple $\mathbf{r}$ meeting the condition above, and to false otherwise. A global description of who has been discriminated can now be extracted by resorting to a standard classification problem on the dataset of individuals of the protected group, where the class attribute is the newly introduced *disc* attribute. Accuracy of the classifier is evaluated with objective interestingness measures, e.g., precision and recall over the $disc = true$ class value. The choice of the value t should then be supported by laws or regulators. For instance, the *four-fifths rule* by the US states that a job selection rate lower than 80 % represents a*prima facie* evidence of adverse impact. Since the intended use of the extracted classifier is descriptive, classification models that are easily interpretable by (legal) experts and whose size is small should be preferred. In other words, one should trade

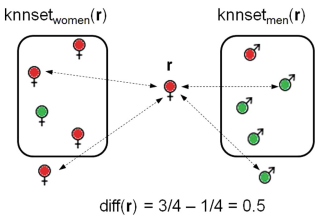

Fig. 2. Example of risk difference $diff(\mathbf{r})$ for $k = 4$. Women are the protected group, $knnset_{women}(\mathbf{r})$ (resp., $knnset_{men}(\mathbf{r})$) is the set of female (resp., male) k-nearest neighbors of $\mathbf{r}$. Red labels denote benefit denied, green labels denote benefit granted. (Color figure online)

accuracy for simplicity. Classification rules and decision trees are natural choices in this sense, since rules and tree paths can easily be interpreted and ranked. The extracted classification models provide a global description of the *disc* class values. They are stored in a knowledge base, for comparison purposes and for the filtering of specific contexts of discrimination – as described next.

Rule Reasoning and Validation. The actual discovery of discriminatory situations and practices may reveal itself as an extremely difficult task. Due to time and cost constraints, an anti-discrimination analyst needs to put under investigation a limited number of contexts of possible discrimination. In this sense, only a small portion of the classification models can be analysed in detail, say the top rules or the top paths of a decision tree [5]. We concentrate on rules of the form: `(cond_1) and ... and (cond_n) => disc=yes [prec] [rec] [diff]`, where `(cond_1) and ... and (cond_n)` is obtained from a classification model. Rules are ranked on the basis of one or more interestingness measures, including: precision `[prec]`, recall `[rec]`, average value of *diff* `[diff]`. Statistical validation is accounted for by relying on logistic regression, which is a well-known tool in the legal and economic research communities. Earlier studies on discrimination discovery, instead, relied upon simple association or correlation measures. Recently, the discrimination-aware data mining community has recognized the importance of causal analysis [3,9].

3 Conclusion

The lesson learned by developing the case study in [7] is above all that discrimination discovery needs a structured process around an algorithmic approach, and a solid compliance with legal rules and ethical principles. Not only this will provide guidance to data scientists and decision makers, but it is the only way we may hope to get acceptance of data mining and machine learning methods by the users of such methods: legal communities, civil rights and digital rights societies, regulation authorities, (inter)national agencies, and professional associations.

References

1. Barocas, S., Selbst, A.D.: Big data's disparate impact. California Law Rev. **104** (2016). SSRN: http://ssrn.com/abstract=2477899
2. Bendick, M.: Situation testing for employment discrimination in the United States of America. Horiz. Stratégiques **3**(5), 17–39 (2007)
3. Foster, S.R.: Causation in antidiscrimination law: beyond intent versus impact. Houston Law Rev. **41**(5), 1469–1548 (2004)
4. Luong, B.T., Ruggieri, S., Turini, F.: k-NN as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), pp. 502–510. ACM (2011)
5. Pedreschi, D., Ruggieri, S., Turini, F.: A study of top-k measures for discrimination discovery. In: Proceedings of ACM SAC 2012, pp. 126–131 (2012)

6. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. Knowl. Eng. Rev. **29**(5), 582–638 (2014)
7. Romei, A., Ruggieri, S., Turini, F.: Discrimination discovery in scientific project evaluation: a case study. Expert Syst. Appl. **40**(10), 6064–6079 (2013)
8. Žliobaitye, I.: A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148v1 (2015)
9. Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: a causal inference approach. In: Proceedings of IJCAI (2016, to appear)