WP. 31
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (v): Privacy for new types of microdata: sequence data and mobility data

Anti-discrimination and privacy protection in released datasets

Prepared by Sara Hajian and Josep Domingo-Ferrer, Universitat Rovira i Virgili, Spain

Anti-discrimination and privacy protection in released datasets

Sara Hajian and Josep Domingo-Ferrer
Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili
UNESCO Chair in Data Privacy, {sara.hajian,josep.domingo}@urv.cat

Abstract:

Unfairly treating people on the basis of their belonging to a specific group (race, ideology, gender, etc.) is known as discrimination and is legally punished in many democratic countries. Automated data collection and a plethora of data mining techniques such as classification rule mining have been designed and are currently widely used for making automated decisions, like loan granting/denial, insurance premium computation, etc. If the training datasets are inherently biased in what regards discriminatory attributes like gender or other sensitive attributes, discriminatory decisions might ensue. Then, similar to privacy, discrimination also has a negative impact on the social perception about data mining. It is obvious that most people do not want to be discriminated because of their gender, religion, nationality, age and so on, especially when this information is going to be used for making decisions about them. In this paper, we tackle discrimination discovery and prevention in data mining.

1 Introduction

Unfairly treating people on the basis of their belonging to a specific group, namely race, ideology, gender, etc., is known as discrimination. In law, economics and social sciences, discrimination has been studied over the last decades and anti-discrimination laws have been adopted by many democratic governments. Some examples are the US Employment Non-Discrimination Act [1], the UK Sex Discrimination Act [1] and the UK Race Relations Act [2]. There are several decision-making tasks which lend themselves to discrimination, *e.g.* loan granting, education, health insurances and staff selection. In many scenarios, decision-making tasks are supported by information systems. Given a set of information items on a potential customer, an automated system decides whether the customer is to be recommended for a credit or a certain type of life insurance. Automating such decisions reduces the workload of the staff of banks and insurance companies, among other organizations. The use of information systems based on data mining technology for decision making has attracted the attention of many researchers in the field of computer science. In consequence, automated data collection and a plethora of data mining techniques such as association/classification rule mining have been designed and are currently widely used for making automated decisions.

At first sight, automating decisions may give a sense of fairness: classification rules do not guide themselves by personal preferences. However, at a closer look, one realizes that classification rules are actually learned by the system based on training

data. If the training data are inherently biased for or against a particular community (*e.g.* foreigners), the learned model may show a discriminatory prejudiced behavior. For example, in certain loan granting organization, foreign people might systematically have been denied access to loans throughout the years. If this biased historical dataset is used as training data to learn classification rules for an automated loan granting system, the learned rules will also show biased behavior toward foreign people. In other words, the system may infer that just being foreign is a legitimate reason for loan denial.

Despite the wide deployment of information systems based on data mining technology in decision making, the issue of anti-discrimination in data mining did not receive much attention until 2008 [3]. After that, some proposals were oriented to the discovery and measure of discrimination. Others dealt with the prevention of discrimination. The discovery of discriminatory decisions was first proposed by [3]. The approach is based on mining classification rules (the inductive part) and reasoning on them (the deductive part) on the basis of quantitative measures of discrimination that formalize legal definitions of discrimination. For instance, the U.S. Equal Pay Act [5] states that: "a selection rate for any race, sex, or ethnic group which is less than four-fifths of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact".

Discrimination can be either direct or indirect (also called systematic). Direct discriminatory rules indicate biased rules that are directly inferred from discriminatory items (*e.g.* Foreign worker = Yes). Indirect discriminatory rules (redlining rules) indicate biased rules that are indirectly inferred from non-discriminatory items because of their correlation with discriminatory ones (*e.g.* the zipcode is non-discriminatory, but if one knows that Zip = 10451 is mostly inhabited by foreigners, indirect discrimination based on the zipcode may occur).

2 Discrimination Discovery

A dataset is a collection of data objects (records) and their attributes. Let *DB* be the original dataset. An item is an attribute along with its value, *e.g.* Race=black. An itemset, *i.e.* *X*, is a collection of one or more items, *e.g.* {Foreign worker=Yes, City=NYC}. A classification rule is an expression $X \underset{\rightarrow}{} C$, where *C* is a class item (a yes/no decision), and *X* is an itemset containing no class item, *e.g.* {Foreign worker=Yes, City=NYC}→Hire=no. *X* is called the premise of the rule. A frequent classification rule is a classification rule with a support or confidence greater than a specified lower bound.

Let *DIs* be the set of predetermined discriminatory items in *DB* (*e.g.* DIs = {Foreign worker=Yes, Race=Black, Gender=Female}). Frequent classification rules fall into one of the following two classes:

1) A classification rule $X \rightarrow C$ is potentially discriminatory (PD) when $X = A, B$ with $A \subseteq DIs$ a non-empty discriminatory itemset and B a non-discriminatory itemset. For example {Foreign worker=Yes; City=NYC}→Hire=No.
2) A classification rule $X \rightarrow C$ is potentially non-discriminatory (PND) when $X = D, B$ is a non-discriminatory itemset. For example {Zip=10451, City=NYC} → Hire=No, or {Experience=Low; City=NYC} → Hire=No.

As mentioned before, Pedreschi et al.[3], [6] translated the qualitative statements in existing laws, regulations and legal cases into quantitative formal counterparts over classification rules and they introduced a family of measures of the degree of discrimination of a PD rule (i.e. *elift*) for direct discriminatory discovery and a PND (i.e. *elb*) for indirect discrimination discovery. Then, whether the PD rule is potentially directly discriminatory can be assessed by thresholding *elift*. Based on this measure, PD rules could be *discriminatory* or *protective.* In addition, whether the PND rule is potentially indirectly discriminatory can be assessed by thresholding *elb*. Based on this measure, PND rules could be *redlining* or *non-redlining (legitimate).*

3 Discrimination Prevention

Beyond discrimination discovery, preventing knowledge-based decision support systems from making discriminatory decisions (discrimination prevention) is a more challenging issue. It can be even more difficult when we want to prevent not only direct discrimination but also indirect discrimination or both at the same time.

The classification of discrimination prevention methods is related to the way of eliminating discrimination and also to the phase of the data mining process in which discrimination prevention is done. Based on this criterion the discrimination prevention methods fall into three groups [4]:

- *Pre-processing*. Transform the source data in such a way that the discriminatory biases contained in the original data are removed so that no unfair decision rule can be mined from the transformed data and apply any of the standard data mining algorithms. The pre-processing approaches of data transformation and hierarchy-based generalization can be adapted from the privacy preservation literature. Along this line, [7], [8] perform a controlled distortion of the training data from which a classifier is learned by making minimally intrusive modifications leading to an unbiased dataset.
- *In-processing*. Change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules [9], [10]. For example, an alternative approach to cleaning the discrimination from the original dataset is proposed in [9] whereby the nondiscriminatory constraint is embedded into a decision tree learner by changing its splitting criterion and pruning strategy through a novel leaf re-labeling approach. However, it is obvious that in-processing discrimination

prevention methods must rely on new special-purpose data mining algorithms; standard data mining algorithms cannot be used.

- *Post-processing*. Modify the resulting data mining models, instead of cleaning the original dataset or changing the data mining algorithms. For example, in [6], a confidence-altering approach is proposed for classification rules inferred by the CPAR algorithm.

Although some methods have already been proposed for each of the above mentioned approaches (pre-processing, in-processing, post-processing), discrimination prevention stays a largely unexplored research avenue.

References

[1] Parliament of the United Kingdom, Sex Discrimination Act, 1975. http://www.opsi.gov.uk/acts/acts1975/ PDF/ukpga 19750065 en.pdf.

[2] Parliament of the United Kingdom, Race Relations Act, 1976. http://www.statutelaw.gov.uk/content.aspx? activeTextDocId=2059995.

[3] D. Pedreschi, S. Ruggieri and F. Turini, "Discrimination-aware data mining", Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560-568. ACM, 2008.

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

[5] United States Congress, US Equal Pay Act, 1963. http://archive.eeoc.gov/epa/anniversary/epa-40.html.

[6] D. Pedreschi, S. Ruggieri and F. Turini, "Measuring discrimination in socially-sensitive decision records", Proc. of the 9th SIAM Data Mining Conference (SDM 2009), pp. 581-592. SIAM, 2009.

[7] F. Kamiran and T. Calders, "Classification without discrimination", Proc. of the 2nd IEEE International Conference on Computer, Control and Communication (IC4 2009). IEEE, 2009.

[8] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling", Proc. of the 19th Machine Learning conference of Belgium and The Netherlands, 2010.

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

[10] F. Kamiran, T. Calders and M. Pechenizkiy, "Discrimination aware decision tree learning", Proc. of the IEEE International Conference on Data Mining (ICDM 2010), pp. 869-874. ICDM, 2010.