A METHODOLOGY FOR DIRECT AND INDIRECT DISCRIMINATION PREVENTION IN DATA MINING

V.Rajkumar
Master of Computer Applications, Madha Engineering College, Chennai.
Email: rajvenkatesan1993@gmail.com

Abstract - Along with privacy, discrimination is a very important issue when considering the legal and ethical aspects of data mining. It is more than observable that the majority people do not want to be discriminated because of their gender, nationality, religion, age and so on, particularly when those aspects are used for making decisions about them like giving them a occupation, loan, insurance, etc. determining such possible biases and eliminating them from the training data without harming their decision-making utility is therefore extremely popular. For this reason, antidiscrimination methods containing discrimination detection and prevention have been introduced in data mining. Discrimination prevention consists of suggest models that do not lead to discriminatory decisions even if the original training datasets are essentially biased. In this section, by focusing on the discrimination prevention, we present taxonomy for classifying and examining discrimination prevention schemes. Then, we begin a group of pre-processing discrimination prevention schemes and indicate the special features of each approach and how these approaches deal with direct or indirect discrimination. A production of metrics used to estimate the performance of those approaches is also specified. In conclusion, we finish our learn by specifying interesting future directions in this research body.

1. Introduction

In social sense, discrimination refers to an action based on prejudice resulting in unfair treatment of people, where the distinction between people is operated on the basis of their membership to a category or minority, without regard to individual merit or circumstances. Examples of social discrimination include racial/ethnic, religious, gender, nationality, disability, and age-related discrimination; a large body of international laws and regulations prohibit discrimination in socially-sensitive decision making tasks, including credit scoring/approval, house lending, and personnel selection. In order to prove (or disprove) a discrimination charge before a court, or to perform a social analysis of discrimination in a given context, it is clearly needed to rely on quantitative measures of the phenomenon under study: for this reason, discrimination has been the subject of a large body of research in legal, economic and social sciences, as well as the subject of empirical analysis in a large number of juridical cases.

For example, the European Union implements the principle of equal treatment between men and women in the access to and supply of goods and services in [3] or in matters of

employment and occupation in [4]. Although there are some laws against discrimination, each and every one are reactive, not positive. Technology can include proactivity to legislation by contributing discrimination discovery and prevention techniques.

Services in the information society allow for automatic and routine collection of large amounts of data. Those data are often used to train association/classification rules in view of production automated decisions, similar to loan granting/denial, insurance premium computation, personnel selection, etc. At initial prospect, automating decisions may provide a sense of fairness: classification rules do not guide themselves by personal preferences. Though, at an earlier seem, one realizes that classification rules are actually learned by the system (e.g., loan granting) from the training data. If the training data are essentially biased for or against a particular community (e.g., foreigners), the learned model may show a discriminatory biased activities. In additional terms, the scheme may assume that just being foreign is a legitimate reason for loan denial. Discovering such potential biases and eliminating them from the training data without harming their decision making utility is therefore highly desirable. One must prevent data mining from becoming itself a source of discrimination, owing to data mining responsibilities producing discriminatory models from biased data sets as part of the automated decision making. In [12], it is demonstrated that data mining can be both a source of discrimination and a means for discovering discrimination.

Discrimination can be either direct or indirect (also called systematic). Direct discrimination consists of rules or procedures that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership. Indirect discrimination consists of rules or methods that, while not explicitly point out discriminatory aspects, purposely or unintentionally could generate discriminatory decisions. Redlining by financial institutions (refusing to grant mortgages or insurances in urban areas they consider as deteriorating) is an archetypal example of indirect discrimination, although certainly not the only one. With a slight abuse of language for the sake of compactness, in this paper indirect discrimination will also be referred to as redlining and rules causing indirect discrimination will be called redlining policies. Indirect discrimination could occur since of the accessibility of some background knowledge (rules), for example, that a certain zip code corresponds to a deteriorating area or an area with mostly black population. The surroundings information might be available from publicly available data (e.g., census data) or might be obtained from the original data set itself because of the existence of nondiscriminatory attributes that are highly correlated with the sensitive ones in the original data set.

2. Related Work

The discovery of discriminatory decisions was first proposed by Pedreschi et al. [12], [15]. 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 US Equal Pay Act [18] states that: a

selection rate for any competition, gender, or national group which is less than four-fifths of the rate for the group with the highest rate will generally be considered as evidence of difficult conflict. This evolution has been absolute to include statistical importance of the extracted patterns of discrimination in [13] and to reason about affirmative action and favoritism [14]. Moreover it has been implemented as an Oracle-based tool in [16]. Current discrimination discovery methods consider each rule individually for measuring discrimination without considering other rules or the relation between them. However, in this paper we also take into account the relation between rules for discrimination discovery, based on the existence or nonexistence of discriminatory attributes. Discrimination prevention, the other major anti discrimination aim in data mining, consists of inducing patterns that do not lead to discriminatory decisions even if the original training data sets are biased. Three approaches are conceivable:

Preprocessing:

Preprocessing approaches of data sanitization and hierarchy-based generalization from the privacy-preserving literature. Along this line, adopts a controlled distortion of the training set. 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 preprocessing approach is useful for applications in which a data set should be published and/or in which data mining needs to be performed also by external parties (and not just by the data holder).

In-processing:

Change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules. For example, an alternative approach to cleaning the discrimination from the original data set is proposed in [2] whereby the nondiscriminatory constraint is embedded into a decision tree learner by changing its splitting criterion and pruning strategy through a novel leaf relabeling approach. But, it is observable 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 data set or changing the data mining algorithms. For example, in [13], a confidence-altering approach is proposed for classification rules inferred by the CPAR algorithm. The post-processing approach does not allow the data set to be published: only the modified data mining models can be published (knowledge publishing), hence data mining can be performed by the data holder only.

Although some methods have already been proposed for each of the above-mentioned approaches (preprocessing, in-processing, post-processing), discrimination preclusion continues

a mostly unexplored research avenue. In this paper, we concentrate on discrimination prevention based on preprocessing, because the preprocessing approach seems the most flexible one: it does not require changing the standard data mining algorithms, unlike the in-processing approach, and it allows data publishing (rather than just knowledge was publishing), unlike the post-processing approach.

3. Prevention of Direct And Indirect Discrimination

We present our approach, including the data transformation methods that can be used for direct and/or indirect discrimination prevention. For each method, its algorithm and its computational cost are specified. Direct and indirect discrimination prevention can be described in terms of two phases:

Discrimination measurement

Direct and indirect discrimination discovery includes identifying α-discriminatory rules and redlining rules. To this end, first, based on predetermined discriminatory items in DB, frequent classification rules in FR are divided in two groups: PD and PND rules. Second, direct discrimination is measured by identifying α-discriminatory rules among the PD rules using a direct discrimination measure (elift) and a discriminatory threshold (α). Third, indirect discrimination is measured by identifying redlining rules among the PND rules combined with background knowledge, using an indirect discriminatory measure (elb), and a discriminatory threshold (α). Let MR be the database of direct α-discriminatory rules obtained with the above process. In addition, let RR be the database of redlining rules and their respective indirect α-discriminatory rules obtained with the above process.

Data transformation

Transform the original data DB in such a way to remove direct and/or indirect discriminatory partialities, with lowest collision on the data and on genuine decision policy, so that no unfair decision rule can be mined from the transformed data. In the following sections, we present the data transformation methods that can be used for this purpose.

Direct Discrimination of Data Transformation

The proposed solution to prevent direct discrimination is based on the fact that the data set of decision rules would be free of direct discrimination if it only contained PD rules that are α-protective or are instances of at least one non-redlining PND rule. Therefore, a suitable data transformation with minimum information loss should be applied in such a way that each α-discriminatory rule either becomes α-protective or an instance of a non-redlining PND rule. We call the first procedure direct rule protection (DRP) and the second one rule generalization.

Direct Rule Protection

There are two methods that could be applied for direct rule protection. One method (Method 1) changes the discriminatory item set in some records (e.g., gender changed from male to female in the records with granted credits) and the other method (Method 2) changes the class item in some records (e.g., from grant credit to deny credit in the records with male gender). Similar data transformation methods could be applied to obtain direct rule protection with respect to other measures (i.e., slift and olift).

4. Rule Generalization

Rule generalization is an additional data revolution method for direct discrimination prevention.

Case 1: In this case, r0 is a p-instance of r for p $\geq 0:8$ and no transformation is required.

Case 2: In this case, the PND rule rb in Dpn should be selected which requires the minimum data transformation. A smaller difference between the values of the two sides for each r in Dpn indicates a smaller required data transformation. In this case, the α-discriminatory rule is transformed by rule generalization.

Case 3: No rule in Dpn satisfies. In this case, rule generalization is not possible and direct rule protection should be performed.

For the α-discriminatory rules to which rule generalization can be concerned, it is feasible that rule protection can be achieved with a smaller data transformation. For these rules the algorithm should select the approach with minimum transformation.

Indirect Discrimination of Data Transformation:

The proposed solution to prevent indirect discrimination is based on the fact that the data set of decision rules would be free of indirect discrimination if it contained no redlining policy. To accomplish this, a appropriate data transformation with minimum information loss should be applied in such a way that redlining rules are converted to non-redlining rules. We call this procedure indirect rule protection (IRP).

Indirect Rule Protection

There are two methods that could be applied for indirect rule protection. One method (Method 1) changes the discriminatory item set in some records (e.g., from non-foreign worker to foreign worker in the records of hired people in NYC city with Zip 6¼ 10451) and the other method (Method 2) changes the class item in some records (e.g., from "Hire yes" to "Hire no" in the records of non-foreign worker of people in NYC city with Zip 6¼ 10451).

Both Direct and Indirect Discrimination of Data Transformation

We deal here with the key problem of transforming data with minimum information loss to prevent at the same time both direct and indirect discrimination. We will give a preprocessing solution to simultaneous direct and indirect discrimination prevention. First, we explain when direct and indirect discrimination could simultaneously occur. This depends on whether the original data set (DB) contains discriminatory item sets or not. To provide both direct rule protection (DRP) and indirect rule protection (IRP) at the same time, an important point is the relation between the data transformation methods. Any data transformation to eliminate direct α-discriminatory rules should not produce new redlining rules or prevent the existing ones from being removed. Also any data transformation to eliminate redlining rules should not produce new direct α-discriminatory rules or prevent the existing ones from being removed.

5. Prevention Algorithms for Direct Discrimination

We start with direct rule protection. Algorithm 1 details Method 1 for DRP. For each direct α-discriminatory rule r´ in MR, after finding the subset DBc, records in DBc should be changed until the direct rule protection requirement is met for each respective rule.

Algorithm 1. **Direct Rule Protection and Rule Generalization**

```
1: Inputs: DB, FR, TR, p ≥ 0.8, α, DIₛ
2: Output: DB′ (transformed data set)
3: for each r′ : A, B → C ∈ TR do
4:     FR ← FR − {r′}
5:     if TRᵣ′ = RG then
6:         // Rule Generalization
7:         DBc ← All records completely supporting
           A, B, ¬D → C
8:         Steps 6-9 Algorithm 1
9:         while conf(r′) > conf(rb:D,B→C)/p do
10:            Select first record in DBc
11:            Modify class item of dbc from C to ¬C in DB
12:            Recompute conf(r′)
13:        end while
14:    end if
15:    if TRᵣ′ = DRP then
16:        // Direct Rule Protection
17:        Steps 5-14 Algorithm 1 or Steps 4-9 Algorithm 2
18:    end if
19: end for
20: Output: DB′ = DB
```

Algorithm 1 takes as input TR, it containing all r´ Є MR and their respective TRr´ and rb. For each α-discriminatory rule r´ in T R, if TRr´ shows that rule generalization should be performed, after determining the records that should be changed for impact minimization, these records should be changed until the rule generalization requirement is met.

Prevention Algorithms for Indirect Discrimination

A detailed algorithm implementing Method 2 for IRP is provided, from which an algorithm implementing Method 1 for IRP can be easily derived. For the sake of brevity and due to similarity with the previous algorithms.

Prevention Algorithms for Direct and Indirect Discrimination

Algorithm 2 details our proposed data transformation method for simultaneous direct and indirect discrimination prevention. The algorithm starts with redlining rules. From each redlining rule (r: X C), more than one indirect α- discriminatory rule (r´: A, B C) might be generated because of two reasons: 1) existence of different ways to group the items in X into a context item set B and a nondiscriminatory item set D correlated to some discriminatory item set A; and 2) existence of more than one item in DIs. Hence, as shown in Algorithm, given a redlining rule r, proper data transformation should be conducted for all indirect α-discriminatory rules.

Algorithm2. Direct and Indirect Discrimination Prevention:

1: Inputs: $\mathcal{DB}, \mathcal{FR}, \mathcal{RR}, \mathcal{MR}, \alpha, DI_s$
2: Output: $\mathcal{DB}'$ (transformed data set)
3: **for each** $r : X \rightarrow C \in \mathcal{RR}$, where $D, B \subseteq X$ **do**
4: $\quad \gamma = conf(r)$
5: $\quad$ **for each** $r' : (A \subseteq DI_s), (B \subseteq X) \rightarrow C \in \mathcal{RR}$ **do**
6: $\quad\quad \beta_2 = conf(r_{b2} : X \rightarrow A)$
7: $\quad\quad \Delta_1 = supp(r_{b2} : X \rightarrow A)$
8: $\quad\quad \delta = conf(B \rightarrow C)$
9: $\quad\quad \Delta_2 = supp(B \rightarrow A)$
10: $\quad\quad \beta_1 = \frac{\Delta_1}{\Delta_2} \quad // conf(r_{b1} : A, B \rightarrow D)$
11: $\quad\quad$ Find $\mathcal{DB}_c$: all records in $\mathcal{DB}$ that completely support $\neg A, B, \neg D \rightarrow \neg C$
12: $\quad\quad$ Steps 6-9 Algorithm 1
13: $\quad\quad$ **if** $r' \in \mathcal{MR}$ **then**
14: $\quad\quad\quad$ **while** $(\delta \leq \frac{\beta_1(\beta_2 + \gamma - 1)}{\beta_2 \cdot \alpha})$ and $(\delta \leq \frac{conf(r')}{\alpha})$ **do**
15: $\quad\quad\quad\quad$ Select first record db_c in $\mathcal{DB}_c$
16: $\quad\quad\quad\quad$ Modify the class item of db_c from $\neg C$ to C in $\mathcal{DB}$
17: $\quad\quad\quad\quad$ Recompute $\delta = conf(B \rightarrow C)$
18: $\quad\quad\quad$ **end while**
19: $\quad\quad$ **else**
20: $\quad\quad\quad$ **while** $\delta \leq \frac{\beta_1(\beta_2 + \gamma - 1)}{\beta_2 \cdot \alpha}$ **do**
21: $\quad\quad\quad\quad$ Steps 15-17 Algorithm 4
22: $\quad\quad\quad$ **end while**
23: $\quad\quad$ **end if**
24: $\quad$ **end for**
25: **end for**
26: **for each** $r' : (A, B \rightarrow C) \in \mathcal{MR} \setminus \mathcal{RR}$ **do**
27: $\quad \delta = conf(B \rightarrow C)$
28: $\quad$ Find $\mathcal{DB}_c$: all records in $\mathcal{DB}$ that completely support $\neg A, B \rightarrow \neg C$
29: $\quad$ Step 12
30: $\quad$ **while** $(\delta \leq \frac{conf(r')}{\alpha})$ **do**
31: $\quad\quad$ Steps 15-17 Algorithm 4
32: $\quad$ **end while**
33: **end for**
34: Output: $\mathcal{DB}' = \mathcal{DB}$

6. Conclusions

Along with privacy, discrimination is a very important issue when considering the legal and ethical aspects of data mining. It is other than observable that the majority people do not desire to be discriminated because of their gender, religious conviction, ethnic group, age, and so on, particularly when those aspects are used for making decisions about them like giving them a job, loan, insurance, etc. The purpose of this paper was to develop a new preprocessing

discrimination prevention methodology including different data transformation methods that can prevent direct discrimination, indirect discrimination or both of them at the same time. To attain this objective, the first step is to measure discrimination and identify categories and groups of individuals that have been directly and/or indirectly discriminated in the decision-making processes; the second step is to transform data in the proper way to remove all those discriminatory biases. Finally, discrimination- free data models can be produced from the transformed data set without seriously damaging data quality. The experimental results reported demonstrate that the proposed techniques are quite successful in both goals of removing discrimination and preserving data quality.

In future we extend the work by combining the other attributes. For example consider load granting application, the manager can reject the application based on their sensitive or non-sensitive attributes. So we introduce a new classification rule by means other than sensitive and non-sensitive attributes, we will add insurance policy details also. The proposed method mainly prevents the indirect discrimination process.

REFERENCES

[1]. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases, pp. 487-499, 1994.

[2]. T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.

[3]. European Commission, "EU Directive 2004/113/EC on Anti- Discrimination," http://eurlex.europa.eu/LexUriServ/LexUriServ.do?ui=OJ:L:2004:373:0037:0043:EN:PDF, 2004.

[4]. European Commission, "EU Directive 2006/54/EC on Anti-Discrimination," http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:204:0023:0036:en:PDF, 2006.

[5]. S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection," Proc. IEEE Symp. Computational Intelligence in Cyber Security (CICS '11), pp. 47-54, 2011.

[6]. S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Rule Protection for Indirect Discrimination Prevention in DataMining," Proc. Eighth Int'l Conf. Modeling Decisions for Artificial Intelligence (MDAI '11), pp. 211-222, 2011.

[7]. F. Kamiran and T. Calders, "Classification without Discrimination," Proc. IEEE Second Int'l Conf. Computer, Control and Comm. (IC4 '09), 2009.

[8]. F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf. Belgium and The Netherlands, 2010.

[9]. F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," Proc. IEEE Int'l Conf. Data Mining (ICDM '10), pp. 869-874, 2010.

[10]. R. Kohavi and B. Becker, "UCI Repository of Machine Learning Databases," http://archive.ics.uci.edu/ml/datasets/Adult, 1996.

[11]. D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases," http://archive.ics.uci.edu/ml, 1998.

[12]. D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 560-568, 2008.

[13]. D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," Proc. Ninth SIAM Data Mining Conf. (SDM '09), pp. 581-592, 2009.

[14]. D. Pedreschi, S. Ruggieri, and F. Turini, "Integrating Induction and Deduction for Finding Evidence of Discrimination," Proc. 12th ACM Int'l Conf. Artificial Intelligence and Law (ICAIL '09), pp. 157- 166, 2009

[15]. S. Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 2, article 9, 2010.