ISSN 2395-1621

A Novel Methodology for Direct and Indirect Discrimination Prevention in Data Mining

[#1]Karan Sisodiya, [#2]Mahesh Todkari, [#3]Prashant Kesarkar, [#4]Abhishek Sadhu
[#5]Prof.Y. B. Gurav.

[1]karan.sisodiya@live.com
[2]mtodkari@gmail.com

[#1234]Department of Computer
[#5]Professor, Department of Computer
P.V.P.I.T Pune, Maharashtra, India

ABSTRACT

Data Mining is a useful process for extracting a knowledge-full hidden data from a large collection of data. There are some negative social perception about data mining, among which are privacy and discrimination. Automatically data collection and data mining process like classification rule mining have cover the way to making automatic decision, like loan granting, insurance premium computation, etc. if the data set are based on discriminatory attributes like gender, race, religion, etc., discrimination decision may happen. Due to this reason, antidiscrimination techniques including discrimination and prevention have been introduced in data mining. Discrimination can be done in two ways first direct and second indirect. Direct discrimination occur when decision are made based on sensitive attributes. Indirect discrimination occur when decision are made based on non-sensitive attributes. In this paper, we doing discrimination prevention and process new techniques applicable for direct and indirect discrimination prevention individually or at the same time. We also propose new metrics to evaluate the utility of the proposed approaches and we compute these approaches. The experimental evaluation demonstrate that the proposed techniques are effective for removing direct and/or indirect discrimination from the original data set while preserving data quality.

Keywords— **Data mining, antidiscrimination, direct and indirect discrimination, rule protection, rule generalization, privacy**

ARTICLE INFO

Article History

Received :20th October 2015

Received in revised form :

21th October 2015

Accepted : 25th October , 2015

Published online :

28th October 2015

I. INTRODUCTION

Basically Discrimination is a treatment of an individual based on the membership in a certain group or category. It involves denying to member of one group opportunities that are available to other groups. There is law of antidiscrimination act, which are designed to prevent discrimination on the basis of attributes (e.g., race, gender, marital status, disability and age) in various settings (e.g., employment and training, access to public service, credit and insurance, etc.). There are some laws against discrimination, all of them are reactive, not reactive. Technology can add proactivity to legislation by contributing discrimination discovery and prevention techniques.

Now in the information society services allow for automatic and routine collection of data in large amount. Those data are often used to train association/classification rules in view of making automated decision, like loan granting/denial, personnel selection, insurance premium computation, etc. At first automatic decision give sense of fairness: classification rules don't guide them by personal preferences.

Discrimination is done in two ways direct and indirect. Direct discrimination consist of rules or procedure that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership. Indirect discrimination consists of rules or procedure that, while not explicitly mentioning discriminatory attributes, intentionally or un intentionally could generate discriminatory decisions. Indirect

discrimination happen because of the availability of some background knowledge(rules), for example, that a certain zip code correspondence to a deteriorating area. The background knowledge may be accessible from publicly available data (e.g., census data) or might be obtained from the original data set itself because of the existence of non-discriminatory attributes that are highly correlated with the sensitive ones in the original data set.

II. BACKGROUND

At here we briefly review the background knowledge require in the reminder of this paper. First we recall some basic definition related to data mining. After that we elaborate on measuring and discovering discrimination.

A. Basic definitions

- A data set is collection of data objects (records) and their attributes. LetDB be the original data set.
- An item is an attribute along with its value.
- An item set, i.e., X, is a collection of one or more items, e.g., {Foreign worker =Yes, City=London}.
- A classification rule is an expression $X \to C$, where C is class and X is an item set containing no class item, e.g., {Foreign worker =Yes, City=London} → Hire = no. X is called premise of the rule.
- The support of an item set, $supp(X)$, is the fraction of records that contain the item set X. we say that rule $X \to C$ is completely supported by a record if both $X$and Cappear in the record.
- The confidence of a classification rule, $conf(X \to C)$, measure how often the class item C appears in records that contain X. Hence, if $supp(X) > 0$ then

$$conf(X \to C) = \frac{supp(X, C)}{supp(X)}. \qquad (1)$$

  Support and confidence range over [0,1].
- A frequent classification rule is a classification rule with support and confidence greater than respective specified lower bounds. Support is a measure of statistical significance, whereas confidence is a measure of the strength of the rule. Let FR be the database of frequent classification rules extracted from DB.
- The negated item set, i.e., $\neg X$ is an item set with the same attributes as X, but the attributes in $\neg X$ take any value except those taken by attributes in X. in this paper, we use the $\neg$ notation for item sets with binary or non-binary categorical attributes. For a binary attributes, e.g., {Foreign worker = Yes/No}, if X is {Foreign worker = Yes}, then $\neg X$ is {Foreign worker = No}.

B. Discriminatory and Non-discriminatory Classification Rule

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

1. A classification rule $X \to C$ is potentially discriminatory (PD) when $X=A,B$ with $A \subseteq DI_s$ a nonempty discriminatory item set and B a non-discriminatory item set. For example, {Foreign worker = Yes, City = London} → Hire = No.
2. A classification rule $X \to C$ is potentially non-discriminatory (PND) when $X = D, B$ is a non-discriminatory item set. For example, {Zip = 1045, City = NYC} → Hire = No, or {Experience = Low, City = NYC} → Hire = No

The word "potentially" means that PD rule could probably lead to discriminatory decision. Therefore, some measures are needed to quantify the direct discrimination potential. Also, a PND rule could lead to discriminatory decision in combination with some background knowledge.

C. Direct Discrimination Measure

Pedreschi et al. [3], translated the qualitative statements in existing laws, regulation, and legal cases into quantitative formal counterparts over classification rules and they introduced a family of measures of the degree of discrimination of PD rule. One of these measures is the extended lift *(elift)*.

Definition 1: Let $A, B \to C$ be a classification rule such that $conf(B \to C) > 0$. The extended lift of the rule is

$$elift(A, B \to C) = \frac{conf(A, B \to C)}{conf(B \to C)}. \qquad (2)$$

The idea here is to evaluate the discrimination of a rule as the gain of confidence due to presence of the discriminatory items (i.e., A) in the premise of the rule. Whether the rule is to be considered discriminatory cabn be accessed by thresholding*elift*as follows.

Definition 2: Let $\alpha \in R$be a fixed $threshold$[1] and let A be a discriminatory item set. A PD classification rule $c = A, B \to C$ is α-protective w.r.t *elift(c)* < α. Otherwise, c is α-discriminatory.

The purpose of direct discrimination discovery is to identify α-discriminatory rules. In fact, α-discriminatory rules indicate biased rules that are directly inferred from discriminatory items (e.g., Foreign worker = Yes). We call these rules direct α-discriminatory rules. In addition to *elift*, two other measures *slift* and *olift* were proposed by Pedreschi et al. in [3].

D. Indirect Discrimination Measure

The purpose of indirect discrimination discovery is to identify redlining rules. In fact, redlining rules indicate biased rules that are indirectly inferred from non-discriminatory items (e.g., Zip = 10451) because of their correlation with discriminatory ones. To determine the redlining rules, Pedreschi et al. in [3] stated the theorem below gives us a lower bound for α-discrimination of PD classification rules, give information available in PND rules (γ, δ), and information available from background rules (β_1, β_2). They assume that background knowledge takes the form of classification rules relating a non-

discriminatory item set *D* to a discriminatory item set *A* within the context *B*.

Theorem 1:*letr: D, B→Cbe a PND classification rule, and let*

$$\gamma = conf(r: D, B \to C)\delta = conf(B \to C) > 0.$$

Let A be a discriminatory item set, and let, β_1, β_2 *such that*

$$conf(r_{b1}: A, B \to D) \geq \beta_1$$
$$conf(r_{b2}: D, B \to A) \geq \beta_2 > 0$$

Call

$$f(x) = \beta_1/\beta_2(\beta_2 + x - 1)$$
$$elb(x, y) = \begin{cases} f(x)/y, & if(x) < 0 \\ 0, & otherwise \end{cases}$$

It holds that, for α≥0, if elb(γ,δ)≥α, the PD classification rule r^1*: A, B→C is α-discriminatory.*

Based on the above theorem, the following formal definitions of redlining and non-redlining rules are presented:

Definition 3: A PND classification rule *r: D, B→C* is a redlining rule if it could yield any α-discriminatory rule $r': A, B \to C$ in combination with currently available background knowledge rules of form $r_{b2}: D, B \to A$, where *A* is a discriminatory item set. For example, {Zip = 10451, City = NYC} → Hire = No.

Definition 4: A PND classification rule *r: D, B→C* is a non-redlining rule if it cannot yield any α-discriminatory rule $r': A, B \to C$ in combination with currently available background knowledge rules of form $r_{b2}: D, B \to A$, where *A* is a discriminatory item set. For example, {Experience = Low, City = NYC} → Hire = No.

We call α-discriminatory rule that ensue from redlining rules indirect α-discriminatory rules.

III. DIRECT AND INDIRECT DISCRIMINATION PREVENTION

In this section, 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.

Our approach for 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 dived 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).
- **Data transformation:** Transform the original data *DB* in such a way to remove direct and/or indirect discriminatory biases, with minimum impact on the data and on legitimate decision rules, 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.

As mentioned before, background knowledge might be obtained from the original data set itself because of the existence of non-discriminatory attributes that are highly correlated with the sensitive ones in the original data set. Let *BK* be a database of background rules that is defined as $BK = \{r_{b2}: D, B \to A | A\ discriminatory\ item\ set\ and\ supp(D, B \to A) \geq ms\}$.

In fact, *BK* is the set of classification rules $D, B \to A$ with a given minimum support *ms* that shows the correlation between the discriminatory item set *A* and the non-discriminatory item set *D* with context *B*. although rules of the form $r_{b1}: A, B \to D (in\ theorem\ 1)$ are not include in *BK*, $conf(r_{b1}: A, B \to D)$ could be obtained as $supp(r_{b2}: D, B \to A)/supp(B \to A)$.

IV. ALGORITHMS

A. Direct discrimination prevention Algorithms

We start with direct rule protection. Algorithm 1 details method 1 for DRP. For each direct α-discriminatory rule *r'* in *MR* (step 3), after finding the subset DB_c should be changed until the direct rule protection requirement (Step 10) is met for each respective rule (Steps 10-14).

Algorithm 1: DIRECT RULE PROTECTION (METHOD1)

Inputs: $DB,\ FR,\ MR, \alpha, DI_s$
Output: DB'(transformed data set)

1) for each $r': A,\ B \to C \in MR$ do
2) $FR \leftarrow FR - \{r'\}$
3) $DB_c \leftarrow$All records completely supporting $\neg A,\ B \to \neg C$
4) for each $db_c \in DB_c$do
5) Compute $impact(db_c)$ $\left|\left\{r_a \in FR \middle| \begin{matrix} db_c\ supports \\ the\ premise\ of\ r_a \end{matrix}\right\}\right|$
6) end for
7) Sort $DB_c$by ascending impact
8) While $conf(r') \geq \alpha\ .\ conf(B \to C)$ do
9) Select first record in DB_c
10) Modify discrimination item set of db_c from $\neg A\ to\ A$ in *DB*
11) Re-compute $conf(r')$
12) end while
13) end for
14) Output $DB' \to DB$

Among the records of DB_c , one should change those withlowest impact on the other (α-protective or non-redlining)rules. Hence, for each record $db_c \in DB_c$ the number of ruleswhose premise is supported by db_c is taken as the impact of db_c (Step 5), that is $impact(db_c)$; the rationale is thatchanging db_c impacts on the confidence of those rules.Then, the records db_c with minimum $impact(db_c)$ areselected for change (Step 7), with the aim

of scoring wellin terms of the utility measures proposed in the nextsection. We call this procedure (Steps 4-7) impact minimizationand we reuse it in the pseudocodes of the rest ofalgorithms specified in this paper.

Algorithm 2 details Method 2 for DRP. The parts of Algorithm 2 to find subset DB_c and perform impact minimization are the same as in Algorithm 1. However, the transformation requirement that should be met for each α-discriminatory rule in MR (Step 8) and the kind of data transformation are different (Steps 8-12).

- **Algorithm 2:** DIRECT RULE PROTECTION (METHOD 2):

  Inputs: $DB,\ FR,\ MR, \alpha, DI_s$
  Output: DB'(transformed data set)

1) for each $r'\colon A,\ B \rightarrow C \in MR$ do
2) $FR \leftarrow FR - \{r'\}$
3) $DB_c \leftarrow$ All records completely supporting $\neg A,\ B \rightarrow \neg C$
4) for each $db_c \in DB_c$do
5) Compute $impact(db_c) = \left|\{r_a \in FR \middle| db_c\ supports\ the\ premise\ of\ r_a\}\right|$
6) end for
7) Sort $DB_c$by ascending impact
8) While $conf\ (B \rightarrow C) \leq \frac{con(r')}{\alpha}$ do
9) Select first record in DB_c
10) Modify the class item of db_c from $\neg C\ to\ C$ in DB
11) Re-compute $conf(B \rightarrow C)$
12) end while
13) end for
14) Output $DB = DB'$

Rule generalizationcannot be applied alone for solving direct discriminationprevention, but it can be used in combination withMethod 1 or Method 2 for DRP.
Algorithm 3 should be run to combine rulegeneralization and one of the two direct rule protectionmethods.

- **Algorithm 3:**DIRECT RULE PROTECTIONAND RULE GENERALIZATION

  Inputs: $DB,\ FR,\ TR,\ p \geq 0.8, \alpha, DI_s$
  Output: DB'(transformed data set)

1) For each $r'\colon A, B \rightarrow C \in TR$ do
2) $FR \leftarrow FR - \{r'\}$
3) if $TR_{r'} = RG$ then
4) // Rule Generation
5) $DB_c \leftarrow$ All records completely supporting $A, B, \neg D \leftarrow C$
6) Steps 6-9 Algorithm 1
7) while $conf(r') > \frac{conf(r_b:D,B\rightarrow C)}{p}$ do
8) select first record in DB_c
9) modify class item of db_c from $C\ to\ \neg C\ in\ DB$
10) Re-compute $conf(r')$
11) end while
12) end if
13) if $TR_{r'} = DRP$ then
14) // Direct rule Protection
15) steps 3-12 Algorithm 1 OR steps 2-12 Algorithm 2
16) end if
17) end for
18) Output: $DB' = DB$

Algorithm 3 takes as input *TR*, which is the output of the algorithm, containing all $r' \in MR$ andtheir respective $TR_{r'}$andr_b. For each α-discriminatory rule*r'* in *TR*, if $TR_{r'}$ shows that rule generalization should be performed (Step 3), after determining the records thatshould be changed for impact minimization (Steps 5-6), theserecords should be changed until the rule generalization requirement is met (Steps 7-11). Also, if $TR_{r'}$ shows that direct rule protection should be performed (Step 13), basedon either Method 1 or Method 2, the relevant sections ofAlgorithms 1 or 2 are called, respectively (Step 15).

B. Direct and indirect discrimination prevention Algorithms

Algorithm 4 details our proposed data transformationmethod for simultaneous direct and indirect discriminationprevention. The algorithm starts with redlining rules. Fromeach redlining rule $(r\colon X \rightarrow C)$, more than one indirect **α**-discriminatory rule $(r'\colon X \rightarrow C)$ might be generatedbecause of two reasons: 1) existence of different ways togroup the items in X into a context item set B and a non-discriminatory item set D correlated to some discriminatoryitem set A; and2) existence of more than one item inDI_s. Hence, as shown in Algorithm 4 (Step 3), given aredlining rule r, proper data transformation should beconducted for all indirect α-discriminatory rules $r'\colon (A \subseteq DI_s), (B \subseteq X) \rightarrow C$ ensuing from *r*.

- **Algorithm 4:**DIRECT AND INDIRECT SISCRIMINATION PREVENTION.

  Inputs: $DB, FR, RR, MR, \alpha, DI_s$
  Output: DB' (transformed data set)

1) For each $r\colon X \rightarrow C \in RR$, where $D, B \subseteq X$ do
2) $\gamma = conf(r)$
3) for each $r'\colon (A \subseteq DI_s),\ (B \subseteq \mathrm{X}) \rightarrow C\ \in RR$ do
4) $\beta_2 = conf(r_{b2}\colon X \rightarrow A)$
5) $\Delta_1 = supp(r_{b2}\colon X \rightarrow A)$
6) $\delta = conf(B \rightarrow c)$
7) $\Delta_2 = supp(B \rightarrow A)$
8) $\beta_1 = \frac{\Delta_1}{\Delta_2}$ // $conf\left(r_{b1}\colon A,\ B \rightarrow D\right)$
9) Find DB_c: all records in DB that completely support $\neg A, B, \neg D \rightarrow \neg C$
10) Steps 4-7 Algorithm 1
11) If $r' \in MR$ then
12) while $(\delta \leq \frac{\beta_1(\beta_2+\gamma-1)}{\alpha*\beta_2})$ and $(\delta \leq \frac{conf(r')}{\alpha})$ do
13) select first record in DB_c
14) modify class item of db_c from $\neg C\ to\ C\ in\ DB$
15) Re-compute $\delta = conf(B \rightarrow C)$
16) end while

17) else
18) while $\delta \leq \frac{\beta_1(\beta_2+\gamma-1)}{\alpha*\beta_2}$ do
19) repeat step 13-15
20) end while
21) end if
22) end for
23) end for
24) For each r': $(A, B \rightarrow C) \in MR \backslash RR$ do
25) $\delta = conf(B \rightarrow C)$
26) find DB_c: all records in DB that completely support $\neg A, B \rightarrow \neg C$
27) repeat step 10
28) while ($\delta \leq \frac{conf(r')}{\alpha}$) do
29) repeat steps 13-15
30) end while
31) end for
32) Output $DB' = DB$

If some rules can be extracted from DB as both direct andindirect α-discriminatory rules, it means that there isoverlap between *MR* and *RR*; in such case, datatransformation is performed until both the direct and theindirect rule protection requirements are satisfied (Steps 11-16).

V. CONCLUSION AND FUTURE SCOPE

Along with privacy, discrimination is a very important issue when considering the legal and ethical aspects of data mining. The purpose of this paper was to develop a new pre-processing 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.

REFRENCES

[1] Sara Hajian and Josep Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination in data mining" Proc. IEEE transaction on knowledgeand engineering, vol. 25, no. 7, july2013.

[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]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.

[4] S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: Discrimination Discovery in Databases," Proc. ACM Int'l Conf. Management of Data (SIGMOD '10), pp. 1127-1130, 2010.

[5] 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.

[6] 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.

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

[10] 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.

