Indian Journal of Communication Engineering and Systems(IJCES)
Vol.2.No.1 2014 pp 37-42.
available at: www.goniv.com
Paper Received :08-03-2014
Paper Accepted:22-03-2014
Paper Reviewed by: 1. R. Venkatakrishnan 2. R. Marimuthu
Editor : Prof. P.Muthukumar

PREVENTING INDIRECT DISCRIMINATION IN DATA MINING OWED HEADED FOR BIASED TRAINING DATASETS

Baskaran P
Dept of Computer Science.,
Manonmanium Sundaranar University.,
Tirunelveli, Tamilnadu, India

Dr. Arulanandam K
Research Supervisor, HOD and Professor,
Dept. of Computer Applications,
G.T.M Govt. Arts and Science College, Gudiyatam, Tamilnadu, India2

ABSTRACT

Services in the information society to authorize frequently and on a regular basis collecting huge amounts of data. Those data are regularly used to train classification rules in view of making automated decisions, like loan granting/denial, insurance premium computation, etc. If the training datasets are biased in what regards responsive attributes like gender, race, religion, etc., discriminatory decisions may follow. Direct discrimination occurs when decisions are made based on biased responsive attributes. Indirect discrimination occurs when decisions are made based on non- responsive attributes which are strongly associated with biased sensitive attributes. This paper discusses how to clean training datasets and outsourced datasets in such a way that justifiable the classification rules can still be extracted but indirectly discriminating rules cannot.

Keywords: Anti-discrimination, Indirect discrimination, Discrimination prevention, Data mining, Privacy.

1. INTRODUCTION

Automated data collection in the information society facilitates automating decision making as well. Superficially, 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 trained on the collected data. If those training data are biased, the learned model will be biased. For example, if the data are used to train classification rules for loan granting and most of the Brazilians in the training dataset were denied their loans, the leaned rules will also show biased behavior toward Brazilian and it is a discriminatory reason for loan denial. 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.

2. DISCRIMINATION-AWARE DATA MINING

The literature in law and social sciences distinguishes direct and indirect discrimination (the latter is also called systematic). Direct discrimination consistsof rules or procedures that explicitly impose “disproportionate burdens” on minority or disadvantaged groups (i.e. discriminatory rules) based on sensitive attributes related to group membership (i.e. discriminatory attributes). Indirect discrimination consists of rules or procdures that, while not explicitly mentioning discriminatory attributes, impose the same disproportionate burdens, intentionally or unintentionally. This effect and its exploitation is often referred to as redlining and indirectly discriminating rules can be called redlining rules [1]. The term “redlining” was invented in the late 1960s by community activists in Chicago [2]. The authors of [1] also support this claim: even after removing the discriminatory attributes from the dataset, discrimination persists because there may be other attributes that are highly correlated with the sensitive (discriminatory) ones or there may be background knowledge from publicly available data (e.g. census data) allowing inference of the discriminatory knowledge (rules).

The existing literature on anti-discrimination in computer science mainly elaborates on data mining models and related techniques. Some proposals are oriented to the discovery and measure of discrimination [1,3,4,7]. Others deal with the prevention of discrimination. Although some methods have been proposed, discrimination prevention stays a largely unexplored research avenue. Clearly, a straightforward way to handle discrimination prevention would consist of removing discriminatory attributes from the dataset. However in terms of indirect discrimination, as stated in [1,2] there may be other attributes that are highly correlated with the sensitive ones or there may be background knowledge from publicly available data that allow for the inference of discrimination rules. Hence, one might decide to remove also those highly correlated attributes as well. Although this would solve the discrimination problem, in this process much useful information would be lost. Hence, one challenge regarding discrimination prevention is considering indirect discrimination other than direct discrimination and another challenge is to find an optimal trade-off between anti-discrimination and usefulness of the training data.

2.1 Involvement and Paper Association

The main contributions of this paper are as follows: (1) a new preprocessing method for indirect discrimination prevention based on data transformation that can consider several discriminatory attributes and their combinations; (2) some measures for evaluating the proposed method in terms of its success in discrimination prevention and its impact on data quality. Although some methods have recently been proposed for discrimination prevention [2,5,6,10]. However, such works only consider direct discrimination. Their approaches cannot guarantee that the transformed dataset is really discrimination-free, because it is known that discriminatory behaviors can be hidden behind non-discriminatory items. To the best of our knowledge this is the first work that proposes a discrimination prevention method for indirect discrimination. In this paper, Section 2 elaborates on the discovery of indirect discrimination. Section 3 presents our proposed method. Evaluation measures and experimental evaluation are presented in Section 4. Conclusions are drawn in Section 5. Rule Protection for Indirect Discrimination Prevention in Data Mining.

3. DISCOVERING DISCRIMINATION

In this section, we present some background concepts that are used throughout the paper. Moreover, we formalize the finding of indirect discrimination.

3.1 Background

A dataset is a collection of 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 is a collection of one or more items. A classification rule is an expression $X \rightarrow C$, where X is an itemset, containing no class items, and C is a class item, e.g. Class=bad. The support of an itemset, supp(X), is the fraction of records that contain the itemset X. We say that a rule $X \rightarrow C$ completely supported by a record if both X and C appear in the record. The confidence of a classification rule, conf($X\rightarrow C$), measures how often the class item C appears in records that contain X. A frequent classification rule is a classification rule with a support or confidence greater than a specified lower bound. Let FR be the database of frequent classification rules extracted from DB. With the assumption that discriminatory items in DB are predetermined (e.g. Race=black), rules fall into one of the following two classes with respect to discriminatory and non-discriminatory items in DB: (i) a classification rule is potentially discriminatory (PD) when X = A,B with A a non-empty discriminatory itemset and B a non-discriminatory itemset (e.g. {Race=black, City=NYC}→Class=bad); (ii) a classification rule is potentially non-discriminatory (PND)when X = D,B is a non-discriminatory itemset (e.g. {Zip=10451, City=NYC} → Class=bad). Let assume that the notation X(D,B) means X = D,B. Let PR a database of frequent classification rules with PDand PND classification rules. The word “potentially” means that a PD rule could probably lead to discriminatory decisions, so some measures are needed to quantify the discrimination potential (direct discrimination). Also, a PND rule could lead to discriminatory decisions if combined with some background knowledge (indirect discrimination); e.g., if the premise of the PND rule contains the Zip=10451 item set, rely on additional background knowledge one knows that zip 10451 is mostly inhabited by black people.

This will be introduced a family of measures of the degree of discrimination of a PD rule. One of these measures is extended lift measure (elif t):

$$\mathbf{elif\ t(A,B \rightarrow C) = conf(A,B \rightarrow C)/conf(B \rightarrow C)}$$

Whether the rule is to be considered discriminatory can be assessed by using a threshold:
Let α € R be a fixed threshold and let A be a discriminatory itemset. A PD classification rule c : A,B→ C is α -protective w.r.t. elif t if elif t(c)< α. Otherwise, c is α-discriminatory.

3.2 Indirect Discrimination Formalization

In terms of indirect discrimination, the purpose of discrimination discovery is identifying PND rules that are to a certain extent equivalent to !-discriminatory rules or, in other words, identifying redlining rules. To determine the redlining rules, This will be stated in the theorem below which gives a lower bound for !-discrimination of PD classification rules given information available in PND rules **(γ,δ)** and information available from background rules (**$\beta1$, $\beta2$**). 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 ([1]). Let r : X(D,B) → C be a PND classification rule, and let $\gamma = conf(D,B \rightarrow C)$ $\delta = conf(B \rightarrow C) > 0$.

Let A be a discriminatory itemset, and let $\beta1$, $\beta2$ such that

$$conf(rb1 : A,B \rightarrow D) \geq \beta1$$
$$conf(rb2 : D,B \rightarrow A) \geq \beta2 > 0.$$

Call

$$f(x) = \beta1/\beta2(\beta2 + x - 1)$$
$$elb(x, y) = \begin{cases} f(x)/y \text{ if } f(x) > 0 \\ 0 \text{ otherwise} \end{cases}$$

It holds that, for $\alpha \geq 0$, if $elb(\gamma, \delta) \geq \alpha$, the PD classification rule r'' : A,B → C is !-discriminatory. Based on the above theorem, we propose the following formal definitions of redlining and non redlining rules.

Definition 1. A PND classification rule r : X(D,B) →C is a redlining rule
if it could yield an α -discriminatory rule r'' : A,B → C in combination with currently available background knowledge rules of the form rb1 : A,B→ D and
rb2 : D,B →A, where A is a discriminatory itemset.

Definition 2. A PND classification rule r : X(D,B) → C is a non-redlining
rule if it cannot yield any α -discriminatory rule r'' : A,B → C in combination
with currently available background knowledge rules of the form rb1 : A,B → D
and rb2 : D,B → A, where A is a discriminatory itemset.

Note that the correlation between the discriminatory itemset A and the nondiscriminatory Item set D with context B indicated by the background rules rb1 and rb2 holds with confidences at least $\beta1$ and $\beta2$, respectively; however, it is not a completely certain correlation. Let RR be the database of redlining rules extracted from database DB.

4. A Proposal for Indirect Discrimination Prevention

In this section we present a new indirect discrimination prevention method. The method transforms the source data by removing indirect discriminatory biases so that no unfair decision rule can be indirectly mined from the transformed data. The proposed solution is based on the fact that the dataset of decision\ rules would be free of indirect discrimination if it contained no redlining rule. For discrimination prevention using preprocessing, we should transform data by removing all evidence of discrimination in the form of α -discriminatory rules and redlining rules.
We concentrated on direct discrimination and considered α -discriminatory rules. In this paper, we focus on indirect discrimination and consider redlining rules. For these rules, a suitable data transformation with minimum information loss should be applied in such a way that those redlining rules are converted to non-redlining rules. As mentioned above, based on the definition of the indirect discriminatory measure (i.e. elb), to convert redlining rules into non-redlining rules, we should enforce the following inequality for each redlining rule r : D,B→ C in RR:

$$\mathbf{(\gamma, \delta)<\alpha} \qquad (1)$$

By using the definitions in the statement of Theorem 1, Inequality (1) can be rewritten as (conf(rb1)/ conf(rb2)) * (conf(rb2) + conf(r : D,B→ C) − 1)

____ $< \alpha$

conf(B→ C) (2)

To enforce the above inequality, there can be two situations:

Case 1: Assume that discriminatory items (i.e. A) are removed from the original database (DB), and the rb1 and rb2 rules are obtained from publicly available data so that their confidences are constant. Let us rewrite

Inequality (2) in the following way

conf(r : D,B → C) < α · conf(B → C) · conf(rb2) − (conf(rb2) + 1) (3)

conf(rb1)

It is clear that Inequality (2) can be satisfied by decreasing the confidence of redlining rule (r : D,B→ C) to values less than the right-hand side of Inequality (3).

Case 2: Assume that discriminatory items (i.e. A) are not removed from the original database (DB), and the rules rb1 and rb2 might be obtained from DB so that their confidences might change by data transformation. This could be more useful to detect the non-discriminatory items that are highly correlated with the discriminatory ones and thereby discover the possibly discriminatory rules that could inferred from them. Let us rewrite Inequality (2) as Inequality (4), where the confidences of rb1 and rb2 rules are not constant.

conf(rb1)/conf(rb2) (conf(rb2) + conf(r : D,B → C) − 1)

conf(B → C) > (4)

α

Clearly, in this case Inequality (2) can be satisfied by increasing the confidence of the base rule (B ! C) of the redlining rule (r : D,B ! C) to values greater than the right-hand side of Inequality (4) without affecting either the confidence of the redlining rule or the confidence of the rb1 and rb2 rules.

The detailed process of our preprocessing discrimination prevention method for indirect discrimination is described by means of the following phases:

– Phase 1. Use Pedreschi's measure on each PND rule to discover the patterns of indirect discrimination emerged from the available data and also the background knowledge. It consists of the following steps: (i) extract frequent classification rules from DB using Apriori [9]; (ii) divide the rules into PD and PND, with respect to the predetermined discriminatory items in the dataset; (iii) for each PND rule, compute elb to determine the collection of redlining rules. Let RR be a database of redlining rules and their respective **α** -discriminatory rules ensuing from those rules through combination with background knowledge rules.

– Phase 2. Transform the original data to convert each redlining rule to a non-redlining rule without seriously affecting the data or other rules. Algorithms 1 and 2 show the steps of this phase.

– Phase 3. Evaluate the transformed dataset with the discrimination prevention and information loss measures gievn below, to check whether they are free of discrimination and useful enough.

The second phase will be explained in detail in the following subsection.

5. DATA TRANSFORMATION METHOD

The data transformation method should increase or decrease some rule confidences as proposed in the previous section with minimum impact on data quality. In terms of the measures defined in gievn below, we should maximize the discrimination prevention measures and minimize the information loss measures. It is worth mentioning that data transformation methods were previously used for knowledge hiding [8] in privacy-preserving data mining (PPDM). Here we propose a data transformation method for hiding discriminatory and redlining rules.

Algorithms 1 and 2 detail our proposed data transformation method for each of the aforementioned cases. Without loss of generality, we assume that the class attribute C is binary (any non-binary class attribute can be expressed as the Cartesian product of binary class attributes).

1. No discriminatory attributes in the dataset. For each redlining rule in this case, Inequality (3) should be enforced. Note that conf(rb2 : D,B → A) Rule Protection for Indirect Discrimination Prevention in Data Mining 217 and conf(rb1 : A,B → D) are constant. The values of both sides of Inequality (3)

are not independent; hence, a transformation is required that decreases the left-hand side of the inequality without any impact on the right-hand side. A possible solution for decreasing

supp(D,B,C)
conf(r : D,B → C)= supp(D,B) (5)
In inequality (3) to the target value is to perturb item D from ¬D to D in the subset DBc of all records of the original dataset which completely support the rule ¬D,B !¬C and have minimum impact on other rules to increase the denominator of Expression (5) while keeping the numerator and conf(B → C) fixed.

2. Discriminatory attributes in the dataset. For each redlining rule in this case, Inequality (4) should be enforced. Note that in this case conf(rb2 : D,B → A) and conf(rb1 : A,B → D) might not be constant. So it is clear that the values of both inequality sides are dependent; hence, a transformation is required that increases the left-hand side of the inequality without any impact on the right-hand side. A possible solution for increasing

conf(B → C) = supp(B,C) / supp(B) (6)

in Inequality (4) to the target value is to perturb item C from ¬C to C in the subset DBc of all records of the original dataset which completely support the rule ¬A,B, ¬D → ¬C and have minimum impact on other rules; this increases the numerator of Expression (6) while keeping the denominator and conf(rb1 : A,B → D), conf(rb2 : D,B → A), and conf(r : D,B → C) fixed.

In Algorithms 1 and 2, records in *DBc* should be changed until the transformation requirement is met for each redlining rule. Among the records of *DBc*, one should change those with lowest impact on the other (non-redlining) rules. Hence, for each record dbc " DBc, the number of rules whose premise is supported by dbc is taken as the impact of dbc, that is impact(dbc); the rationale is that changing dbc impacts on the confidence of those rules. Then the records dbc with minimum impact (dbc) are selected for change, with the aim of scoring well in terms of the four utility measures proposed in the next section. Background Information. In order to implement the proposed data transformation method for indirect discrimination prevention, we simulate the availability of a large set of background rules under the assumption that the dataset contains the discriminatory items. Let BKs be a database of background rules be defined as

BK = {rb2 : X(D,B) → A|A discriminatory itemset and supp(X → A) ≥ ms}

In fact, *BK* is the set of classification rules X → A with a given minimum support ms and A a discriminatory itemset. Although rules of the form rb1 :
Algorithm 1.

```
Inputs: DB, FR, RR,α , DIs
Output: DB’: the transformed dataset
for each r : X(D,B) →  C € RR do
  γ = conf(r)
for each r’ : (A ≤ DIs), (B ≤ X) →  C do
β2 = conf(rb2 : X →  A)
Δ1 = supp(rb2 : X →  A)
δ = conf(B →  C)
Δ 2 = Supp(B → A)
β 1= Δ 1 / Δ 2  //conf(rb1 : A,B → D)
Find DBc: all records in DB that completely support ¬D,B →  ¬C
for each dbc € DBc do
Compute impact(dbc) = |{ra € FR|dbc supports the premise of ra}|
end for
Sort DBc by ascending impact
while (γ≥ α.δ.β2/β1) – (β2+1)  do
Select first record dbc in DBc
Modify D item of dbc from ¬D to D in DB
 Recompute γ = conf(r : X → C)
end while
end for
end for

Output: DB’ = DB
```

A,B → D are not included in BK, conf(rb1 : A,B → D) could be obtained as :
supp(rb2 : D,B → A)/supp(B ! A).
From each redlining rule (r : X(D,B) → C) in combination with background knowledge, more than one α-discriminatory rule r’ : A,B → C might be generated because of two reasons:
1) existence of different sub-itemsets D,B≤ X such that X can be written as D,B and 2) existence of more than one item in the set of predetermined discriminatory items (DIs). Hence, given a redlining rule (r), proper data transformation should be conducted for all α-discriminatory rules
r' : (A ≤ DIs), (B≤ X) → C ensuing from r.

6. CONCLUSIONS

To the best of our knowledge, we have presented the first method for preventing indirect discrimination in data mining due to biased training datasets. Our contribution in this paper concentrates on producing training data which are free or nearly free from indirect discrimination while preserving their usefulness to data mining algorithms. In order to prevent indirect discrimination in a dataset, a first step consists in discovering whether there exists indirect discrimination. If any discrimination is found, the dataset is modified until discrimination is brought below a certain threshold or is entirely eliminated. In the future, we want to present a unified discrimination prevention approach based on the discrimination hiding idea that encompasses both direct and indirect discrimination.

REFERENCES

1. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560–568. ACM, New York (2008)

2. Kamiran, F., Calders, T.: Classification without discrimination. In: Proc. of the 2nd IEEE International Conference on Computer, Control and Communication (IC4 2009). IEEE, Los Alamitos (2009)

3. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data 4(2) Article 9

4. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in sociallysensitive decision records. In: Proc. of the 9th SIAM Data Mining Conference (SDM 2009

5. Kamiran, F., Calders, T.: Classification with No Discrimination by Preferential Sampling. In: Proc. of the 19th Machine Learning Conference of Belgium and, The Netherlands

6. Calders, T., Verwer, S.: Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21(2), 277–292.

7. Pedreschi, D., Ruggieri, S., Turini, F.: Integrating induction and deduction for finding evidence of discrimination. In: Proc. of the 12th ACM International Conference on Artificial Intelligence and Law (ICAIL 2009), pp. 157–166. ACM, New York (2009).

8. Verykios, V., Gkoulalas-Divanis, A.: A survey of association rule hiding methods for privacy. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy- Preserving Data Mining: Models and Algorithms. Springer, Heidelberg (2008).

9. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. of the 20th International Conference on Very Large DataBases.

10. Hajian, S., Domingo-Ferrer, J., Mart´ınez-Ballest´e, A.: Discrimination prevention in data mining for intrustion and crime detection. In: Proc. of the IEEE Symposium on Computational Intelligence in Cyber Security (CICS 2011), pp. 47–54. IEEE,Los Alamitos (2011).

11. Hajian, S., Domingo-Ferrer, J., Mart´ınez-Ballest´e, A.: Rule generalization and protection for discrimination prevention in data mining (submitted).

12. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998), http://archive.ics.uci.edu/ml