A Survey on Direct and Indirect Discrimination Prevention in Data Mining

Ancy Daniel
Dept. of Computer Science
College of Engineering poonjar

Sreekumar K
Dept. of Computer Science
College of Engineering poonjar

Minu KK, Ph.D.
Dept. of Mathematics
College of Engineering poonjar

ABSTRACT
Data mining is an important technology for extracting useful information from large collections of data in the database. Data mining techniques like classification rule mining and automated data collections have given the way to making automated decisions for loan granting or denial, personal selection etc. If the training data sets are biased with discriminatory or sensitive attributes like gender, age, religion, color etc., discriminatory decisions may ensue. That cause potential privacy invasion and potential discrimination. Later one consists of unfairly treating people on the basis of their belonging to a specific group. The anti-discrimination techniques named discrimination discovery and prevention have been introduced in data mining to solve these problems. Discrimination is divided into two, Direct and Indirect and it tackles discrimination prevention in data mining and propose new techniques applicable for direct, indirect and both at the same time. It also describes how to clean training data sets and outsourced data sets in such a way that direct and/or indirect discriminatory decision rules are converted to legitimate (nondiscriminatory) classification rules and a number of papers mention measures of utility too. This survey paper is aimed at understand the existing discrimination prevention techniques and the utility measures discussed so far.

General Terms
Data mining, Discrimination Prevention, Direct and Indirect Discrimination

1. INTRODUCTION
Data mining involves the extraction of implicit previously unknown and potentially useful knowledge from large databases. The important issue in data mining is discrimination. Discrimination can be viewed as the act of unfairly treating people on the basis that they belong to a specific group. For instance, individuals may be discriminated because of their ideology, gender, age etc. In Economics and Social Sciences, discrimination has been studied for over half a century. There are several decision-making tasks which lend themselves to discrimination, *e.g.* loan granting insurance premium computing and staff selection. So the decision making utility must be discrimination free.

Discrimination is of two types, direct or indirect (systematic). Direct discrimination includes a set of rules(laws) or procedures(events) that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership. Indirect discrimination includes rules or procedures that are not explicitly mentioning discriminatory attributes, but could generate discriminatory decisions. The literature has given evidence of unfair treatment in racial profiling and redlining, mortgage discrimination, personnel selection discrimination and wages discrimination.

Redlining by financial institutions is an example of indirect discrimination. This indirect discrimination will also be referred to as redlining and rules causing indirect discrimination will be called redlining rules. Indirect discrimination occurred because of the availability of some background knowledge (rules). The background knowledge might 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 nondiscriminatory attributes that are highly correlated with the sensitive ones in the original data set.

2. REVIEW OF EXISTING APPROACHES

2.1 Decision Theory for Discrimination-aware Classification
This method[2] propose two flexible and easy-to-use solutions for discrimination-aware classification based on an intuitive hypothesis: discriminatory decisions are often made close to the decision boundary because of decision maker's bias. This hypothesis is implemented via decision theoretic concepts of prediction confidence and ensemble disagreement. The first solution ROC (Reject Option based Classification), exploits the low confidence region of a single or an ensemble of probabilistic classifiers for discrimination reduction. That means, ROC invokes the reject option and labels instances belonging to deprived and favored groups in a manner that reduces discrimination. And second solution, called Discrimination-Aware Ensemble (DAE), exploits the disagreement region of a classifier ensemble to re label deprived and favored group instances for reduced discrimination.

Real world Datasets: Adult, Communities and Crimes.

Advantages: Both ROC and DAE ensure discrimination-aware classifications at run-time without data modification or algorithm tweaking. Moreover, both solutions provide the decision maker with easy control over the resulting discrimination.

Drawbacks: A uniform strategy is applied to all rejected instances.

2.2 Classification with No Discrimination by Preferential Sampling
This method [3] deals with Classification with No Discrimination (CND) problem by preferential sampling. This preferential sampling (PS) approach changes the distribution of different data objects bias free. The data objects nearer to the decision boundaries are more biased, make changes the distribution of these borderline objects to make the dataset discrimination free. PS starts by learning a ranker on the training data. PS uses this ranker to class the data objects of DP and PP in ascending order, and the objects of DN and PN

in descending order; both with respect to the positive class probability. Such understanding of data objects makes sure that the higher the rank an element occupies, the closer it is to the borderline.PS starts from the original training dataset and iteratively duplicates (for the groups DP and PN) and removes objects. The size of a group can be decreased by removing the data objects closest to the borderline. And increase the size by duplication of the data object closest to the borderline.

Figure 1 gives an illustration of Preferential Sampling (PS), showing 40 data point. Data points of the desired class and the negative class are represented by + and – symbols respectively [3].

PS works in the following steps:

(i) Divide the data objects into the four groups, DP, DN, PP, and PN.

(ii) Any ranking algorithm may be used for calculating the class probability of each data tuple. This ranking will be used to identify the borderline data objects.

(iii) Calculate the expected size for each group to make the dataset bias free.

(iv) Finally apply sampling with replacement to increase the size of DP and PN. And decrease the size of DN and PP.

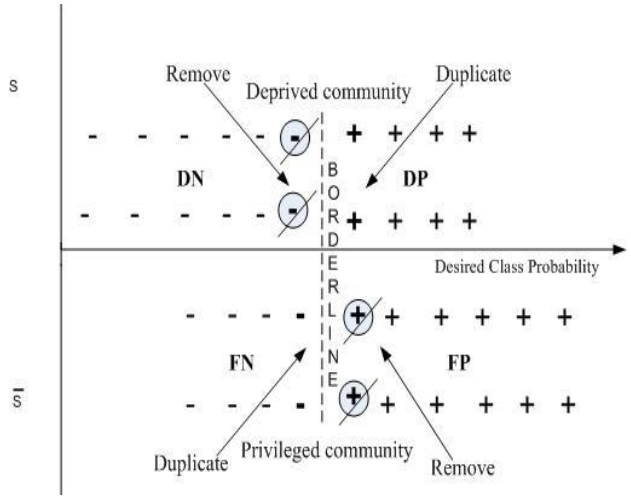

Fig: 1. Representation preferential sampling

Dataset: The Census Income dataset in the UCI ML-repository (Asuncion &Newman, 2007).

Advantages: Preferential Sampling provide discrimination free data and it does not require the change of any class label.

Drawback: Low data utility rate and minimum discrimination removal. This PS is not applicable for Indirect discrimination.

2.3 Discrimination-aware Approach

This data mining method focused on discrimination aware mining. Here [1] α- protection is introduced as a measure of the discrimination power of a potentially discriminatory PD classification rule. It defines a measure that relative gain in confidence of the rule due to the presence of the discriminatory item sets. The measure called, α- parameter is the key for tuning the desired level of protection against discrimination. It uses *elift* for direct discrimination measure and *glift* for indirect discrimination measure.

α- protection

[α-protection] Let c = A,B → C is PD classification rule, where A is a PD and B is a PND itemset, and let:

γ = con *f*(A,B → C) δ = con *f*(B → C) > 0.

For a given threshold α ≥ 0, we say that c is α-protective if *elift*(γ,δ) = γ /δ.

c is called α-discriminatory if *elift*(γ,δ) ≥ α.

Strong α- Protection

[Strong α-protection] Let c= A, B → C be a PD classification rule, where A is a PD and B is a PND itemset, and let:

γ = con *f*(A,B → C) δ = con *f*(B → C) > 0.

For a given threshold α ≥ 1, we say that c is strongly α – protective if *glift*(γ,δ) < α, where:

$$glift(\gamma,\delta) = \begin{cases} \gamma/\delta & \text{if } \gamma \geq \delta \\ (1-\gamma)/(1-\delta) & otherwise \end{cases}$$

If *glift*(γ,δ) ≥ α. We say that c is strongly α-discriminatory.

The *glift*() function ranges over [1, ∞]. If classification rules with a minimum support

ms > 0 are considered, it ranges over [1, 1/*ms*]. Moreover, for 1 > δ > 0:

glift(γ,δ) = max{*elift*(γ,δ), *elift*(1 - γ, 1 - δ)}.

Dataset: German credit dataset

Advantages: Direct and indirect discrimination are measured using strong α-protection and α-protection.

Drawbacks: This method does not measure direct and indirect discrimination at the same time.

2.4 Three naive Bayes approaches for discrimination-free Classification

This method use naïve bayes classifier for discrimination aware classification problem. Navies bayes model[4], Latent variable model, and Modified naives bayes are used. This is done by Modifying probability of positive decisions, i.e the probability distribution *P(C|S)* of the sensitive attribute values *S* given the class values *C*. The joint distribution over the class *C*, sensitive *S*, and all other *A*1 . . . *An* attributes becomes

*P(C, S, A*1, . . . , *An) = P(S)P(C|S)P(A*1|*C) . . . P(An|C)*

And train one model for every sensitive attribute value and balance them. Then add a latent variable L that represents an unbiased, discrimination-free label and optimize the model parameters for likelihood using expectation maximization.

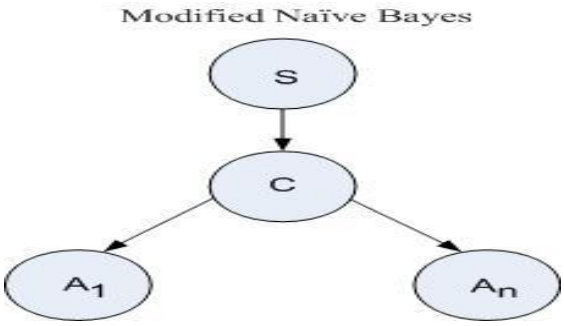

Fig2. Graphical representation of modified naïve bayes approach

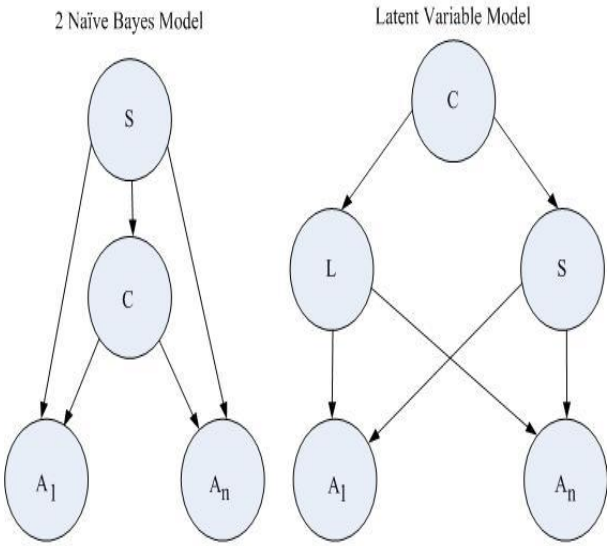

Fig 3: Pictorial representation of 2 Naïve Bayes model and Latent Variable Model

Dataset: Census income

Advantage: provide high accuracy with low discrimination.

Drawbacks: The problem of this model is that it was based on assumptions that might not always hold in practice.

2.5 Discrimination Prevention Approach for Intrusion and Crime Detection

This approach define a new discrimination prevention method called α- discrimination rule. This α-discrimination rule[5] is divided into two groups, first α–discriminatory rules such that there is at least one PND rule leading to same result and α-discriminatory rules such that there is no such PND rule. Second group, also a suitable data transformation with minimum information loss should be applied in such a way that α- discriminatory rules are converted to α-protective rules based on the definition of the discriminatory measure (*i.e. elif t*).

Data Transformation Method

The α-discriminatory rules with the first transformation requirement inequality in the equation $conf(A,B \rightarrow C) \leq conf(D,B \rightarrow C)/p$). The α-discriminatory rules with the second transformation requirement (inequality $conf(A,B \rightarrow D) \geq p$), the value of the right-hand side of the inequality is fixed. Then value of the left-hand side could be increased independently. And the α-discriminatory rules with the third transformation requirement (inequality $conf(A,B \rightarrow C) < \alpha \cdot conf(B \rightarrow C)$), it required both inequality sides are dependent.

Discrimination removal measured by using these utilities; Discrimination Prevention Degree (DPD), Discrimination Protection Preservation (DPP), Misses Cost (MC), and Ghost Cost (GC).

Datasets: Intrusion Detection Systems.

Advantages: The measures that evaluate the degree of discrimination and information loss.

Drawbacks: The method does not consider background knowledge.

Method	Advantages	Drawback	Dataset
Decision theory for discrimination aware classification	Ensure discrimination aware classifications at runtime.	A uniform strategy applied to all rejected instances.	Adult, communities and criminal
Classification with no discrimination by preferential sampling	Provides discrimination free data and does not require any change in the class label.	Low data utility rate and minimum discrimination removal	Senses income
Discrimination aware approach	Direct and indirect discriminations are measured	It does not measure direct and indirect discrimination at the same time.	German credit card
Three naïve bayes approaches for discrimination free classification.	Provide high accuracy with low discrimination	It was based on assumptions that might not allows hold in practice	Censes income
Discrimination prevention approach for intrusion and crime detection.	Measures evaluate the degree of discrimination and information loss.	It does not consider background knowledge.	Intrusion detection system.

3. CONCLUSION

In this paper a wide survey of the different approaches for discrimination prevention, and analyses of major algorithms available for discrimination prevention method is carried out and pointed out the drawbacks of direct and indirect discrimination prevention methods, utility measures. We need to further improve those approaches or develop some efficient novel methods. There is lot of scopes for improving the current system by handling direct and indirect discrimination at same time. And this method can be implement in all real time database.

4. REFERENCE

[1] Dino Pedreschi, Salvatore Ruggieri and Franco Turini "Discrimination-aware Data Mining" Dipartimento di Informatica, Università di Pisa, 2008 ACM, Las Vegas, Nevada, USA

[2] Faisal Kamiran, Asim Karim, and Xiangliang Zhang "Decision Theory for Discrimination-aware Classification" King Abdullah University of Science and Technology (KAUST), The Kingdom of Saudi Arabia, IEEE 12th International Conference on Data Mining 2012.

[3] Faisal Kamiran and Toon Calders "Classification with No Discrimination by Preferential Sampling" Eindhoven University of Technology, Netherlands,

19th Machine Learning conference of Belgium and The Netherlands. 2010

[4] Toon Calders and Sicco Verwer "Three naive Bayes approaches for discrimination-free classification", This article is published with open access at Springerlink.com, 2010

[5] Sara Hajian, Josep Domingo-Ferrer and Antoni Mart´ınez-Ballest´e "Discrimination Prevention in Data Mining for Intrusion and Crime Detection" Universitat Rovira i VirgiliDept. of Computer Engineering and Maths, UNESCO Chair in Data Privacy, Tarragona, Catalonia, 2011 IEEE

[6] S. Hajian and J. Domingo-Ferrer. "A methodology for direct andindirect discrimination prevention in data mining" 2012 IEEE

[7] European Commission, "EU Directive 2004/113/EC on Anti- Discrimination," 2004.

[8] European Commission, "EU Directive 2006/54/EC on Anti- Discrimination," 2006.

[9] P.N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining" Addison-Wesley, 2006.

[10] Sara Hajian, Josep Domingo-Ferrer, and Antoni Mart´ınez-Ballest "Rule Protection for Indirect Discrimination Prevention in Data Mining" Springer 2011