Direct Discrimination Aware Data Mining

Deepali Jagtap
Department of Computer Engineering, PG Student, KKWIEER, Nashik, University of Pune, India

Shirish S. Sane, Ph.D
Department of Computer Engineering, Head of Department, KKWIEER, Nashik, University of Pune, India

ABSTRACT
With the advent of data mining, in many applications the automated decision making systems are used to make fair decision, but there can be discrimination hidden in the decision made by system. Discrimination refers to treating person or entity unfairly based on their membership to a certain group. Discrimination can be observed not only in social sense but also in data mining. People do not want discrimination on the basis of gender, age, nationality, race etc. and many more; therefore it is important to prevent such discrimination. Discrimination prevention mainly consists of two steps: first is discrimination discovery and second is data transformation. The data transformation follows similar approach to that of data sanitization that is used in privacy preservation. Various discrimination measures can be used to analyze its effect on quality of the original dataset.

General Terms
Data mining, Privacy preservation

Keywords
Data sanitization, Data transformation, Discrimination, Discrimination discovery, Discrimination measures

1. INTRODUCTION
Discrimination is treating an individual unfairly based on their actual/perceived membership to a certain group or category. It restricts members of one group from opportunities or privileges that are available to another group, for example, in loan approval/ denial system discrimination on the basis of gender may takes place, not only this but discrimination on the basis of nationality, age, race can also be observed in many applications where automated decision making systems are used to derive rules. In data mining training data is used to learn classification model and based on that, decisions are made by decision makers. If training data that is used in classification modeling is biased towards or against certain group or group attribute then it may leads to a discriminatory decision. Many supervised classification models uses training data for learning process therefore it is important to remove or prevent such discrimination from training data to make it discrimination free.

Direct discrimination can be observed if premise part of the classification rule contains discriminatory attribute whereas in indirect discrimination no such discriminatory attributes are present in premise of classification rules but it contains some such attributes which are found to be indirectly correlated with discrimination attributes when mapped with some other publicly available data. The use of automated decision making systems may give sense of fair decision but in reality it is not always true. It may leads to discriminatory results due to biased training dataset. There exist many anti-discriminatory law but those are reactive not proactive. The use of DRP algorithm [9] as well as other techniques helps to add pro-activeness to it. For example, there exists law in Indian article 15 of law that prohibits discrimination. This anti-discriminatory system can be used in various applications where there is possibility of discriminatory mining model, for example, credit/insurance approval/ denial, job hiring, crime detection and many more.

Rest of the paper is organized as follows. Section one provides introduction, survey of literature along with pros and cons of some of the existing methods are discussed in section two. Section three highlights basic terminology associated with this topic and section four describes algorithm as well as block diagram for discrimination prevention. Section five contains results and discussion about data set and finally last section presents conclusion along with the future scope of system.

2. LITERATURE SURVEY
The various studies have been made by various authors to prevent or remove discrimination. Some of them have focused on discrimination measures while some authors have tried to prevent or remove discrimination from the original dataset. The work in this area can be tracked back to year 2008, done by Pedeschi for the very first time. Discrimination prevention can be done in three ways based on when and in which phase data or algorithm is to be changed. Three ways for Discrimination prevention are: Preprocessing method, Inprocessing method and Postprocessing method. Discrimination can be of 3 types: Direct, Indirect or combination of both, based on presence of discriminatory attributes and other attributes that are strongly related with discriminatory one. Thus depending on the existence of any of above three discrimination, the respective Discrimination Prevention Method is applied. Dino Pedreschi, Salvatore Ruggieri, Franco Turini [3] have developed a model that can be used for the analysis and reasoning of discrimination in Decision Support System which helps DSS owners and control authorities in the process of discrimination analysis [3]. S. Ruggieri, Pedreschi and F. Turini [13] have implemented the oracle based DCUBE tool to explore discrimination hidden in data.

2.1 Discrimination Prevention by Preprocessing Method
In preprocessing method, the original dataset is modified so that it will not result in discriminatory classification rule. In this method any data mining algorithm can be applied to get mining model. Sara Hajian and Josep Domingo-Ferrer [9] proposed another preprocessing method to remove direct and indirect discrimination from original dataset. It employees 'elift' as discrimination measure to prevent discrimination in crime and intrusion detection system [10]. Kamiran and Calder [4] proposed a method based on "*data massaging"* where class label of some of the records in the dataset is changed but as this method is intrusive, concept of

"*Preferential sampling*" was introduced where distribution of objects in a given dataset is changed to make it non-discriminatory [4]. It is based on the idea that, "Data objects that are close to the decision boundary are more vulnerable to be victim of discrimination." This method uses Ranking function and there is no need to change the class labels. This method first divides data into 4 groups that are DP, DN, PP, PN, where first letters D and P indicate Deprived and Privileged class respectively and second letters P, N indicates positive and negative class label. The ranker function then sorts data in ascending order with respect to positive class label. Later it changes sample size in respective group to make that data biased free.

Preprocessing method is useful in applications where data mining is to be performed by third party and data needs to publish for public usage [9].

2.2 Discrimination Prevention by Inprocessing Method

Faisal Kamiran, Toon Calders and Mykola Pechenizkiy [5] introduced inprocessing method based on decision Tree where data mining algorithm is modified instead of modifying original dataset. This approach consists of two techniques for the decision tree construction process, first is Dependency-Aware Tree Construction and another is Leaf Relabeling. The first technique focuses on splitting criterion for tree construction to build a discrimination-aware decision tree. In order to do so, it first calculates the information gain with respect to class & sensitive attribute represented by IGC and IGS respectively. There are three alternative criteria for determining the best split that uses different mathematical operation: (i) IGC-IGS (ii) IGC/IGS (iii) IGC+IGS. The second approach consists of processing of decision tree with discrimination-aware pruning and it relabel the tree leaves [5]. Unfortunately, inprocessing methods requires special purpose data mining algorithms.

2.3 Discrimination Prevention by Postprocessing Method

Unlike preprocessing and inprocessing method in postprocessing method resultant mining model is modified instead of modifying original data or mining algorithm but disadvantages of this method are, it doesn't allow original data to be published for public usage, also the task of data mining should be performed by data holder only. Toon Calders and Sicco Verwer [14] proposed approach where the Naive Bayes classifier is modified to perform classification that is independent with respect to a given sensitive attribute. There are three approaches in order to make the Naive Bayes classifier discrimination-free: (i) was modifying the probability of the decision being positive where the probability distribution of the sensitive attribute is modified. This method has disadvantage of either always increasing or always decreasing the number of positive labels assigned by the classifier, depending on how frequently the sensitive attribute is present in dataset, (ii) training one model for every sensitive attribute value and balancing them. This is done by splitting the dataset into two separate sets and the model is learned using only the tuples from the dataset that have a favoured sensitive value, (iii) adding a latent variable to the Bayesian model. This method models the actual class labels using a latent variable [14]. Sara Hajian, Anna Monreale, Dino Pedreschi ,Josep Domingo Ferrer [11] proposed postprocessing method that derive frequent classification rule and modifies the final mining model using α-Protective k-Anonymous pattern sanitization to remove discrimination from Mine Model.

3. BACKGROUND CONCEPT

In this section, basic terms in data mining are described in short as below:

The collection of records i.e. data object is known as dataset. An item is any attribute associated with its value, for example, age = old. An Item set is collection of such one or more attributes, for example age = old, gender = female.

A classification rule is represented as A→ C, where A is any item other than class item and C is a class item. A is called as premise of the classification rule and C as conclusion of the rule. *Support* for a rule of the form (A → C) indicates number of records in original dataset that contains both A and C. *Confidence* for a rule of the form (A → C) shows how often attribute C appears in transaction where A appears. It can be computed as fraction of Support (A → C) and Support (A). A *frequent classification rule* is a classification rule extracted having minimum support and minimum confidence greater than some specified value [7].

Pedreschi[2] introduced '*elift*' called extended lift as one of the discrimination measure. It provides gain in confidence due to presence of discriminatory item [1]. For a given classification rule, Extended lift can be calculated as below.

elift(A, B → C) = Confidence(A, B → C)/Confidence(B → C)

3.1 Direct discrimination

Direct discrimination consists of rules or procedures that explicitly mention disadvantaged or minority groups based on sensitive discriminatory attributes. For example, the rule r: (Foreign_worker = Yes, City = Nasik → Hire = No) shows direct discrimination as it contains discriminatory attribute Foreign_worker = yes.

3.2 Indirect discrimination

Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or un-intentionally could generate discriminatory decisions [9], for example, the rule r: (Pin_code = 422006, City = Nashik → Hire = No) shows indirect discrimination, as attribute Pin_code corresponds to area with mostly people belonging to particular religion.

3.3 PD Rule

A classification rule is said to be Potentially Discriminatory rule if it contains discriminatory item in premise of a given rule.

3.4 PND Rule

A classification rule is said to be Potentially Non-discriminatory rule if it doesn't contains any discriminatory item in premise of a given rule [1, 9].

4. DISSERTATION WORK

The proposed work uses preprocessing approach of discrimination prevention where different discrimination measures are used for discrimination discovery such as elift, slift, glift etc. The preprocessing approach of discrimination prevention mainly consists of two steps. First step emphasizes the discrimination discovery and second step performs data modification to make original dataset biased free [10].

4.1 Process Block Diagram

The block diagram for discrimination prevention is shown in figure 1. The system takes original dataset containing discriminatory items as an input.

4.1.1 Data Preprocessing and Discretization

The original dataset contains numerical values for some of the attributes, those attributes should be preprocessed and converted into categorical form i.e. discretization is performed.

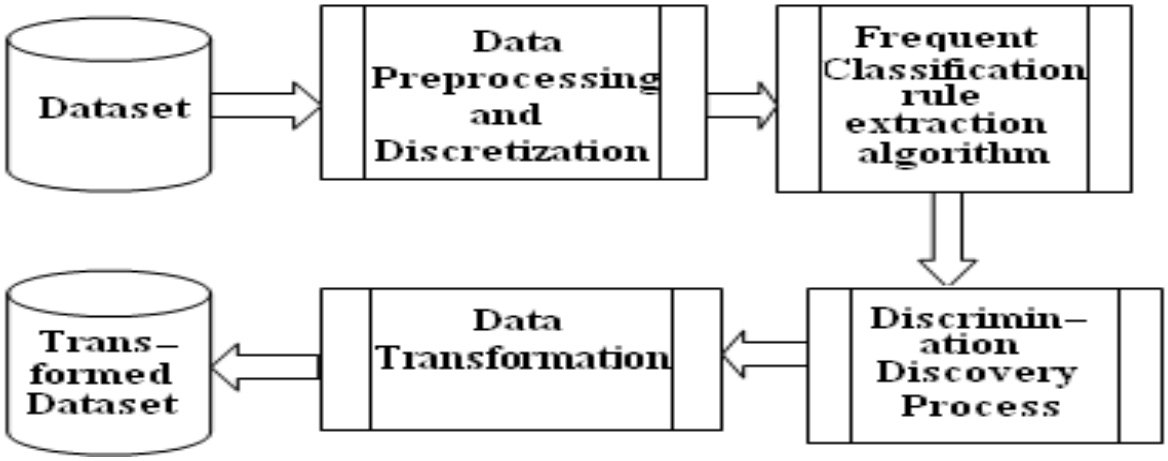

Fig 1: Block diagram for Discrimination Prevention

4.1.2 Frequent Classification Rule extraction algorithm

The Apriori algorithm is used to generate frequent item sets. In Apriori algorithm candidate set generation and pruning steps are performed and the resultant frequent item sets are used to generate frequent classification rules.

4.1.3 Discrimination Discovery Process

The frequent classification rules are then categorize into Potentially Discriminatory and Potentially Nondiscriminatory groups in discrimination discovery process that is shown in figure 2. For discrimination discovery each classification rule is examined and is placed into either PD or PND group based on presence of discriminatory items in premise of the rule. In the next step for every PD rule elift, glift and slift is calculated. If that calculated value is greater than or equal to threshold value (α) then that rule is considered as α-discriminatory.

4.1.4 Data Transformation

The α-discriminatory rules need to be treated further to prevent discrimination. For that purpose data transformation is carried out in the next step where class label of some of the records is perturbed to prevent discrimination. As a result of above process finally the transformed dataset is obtained as an output.

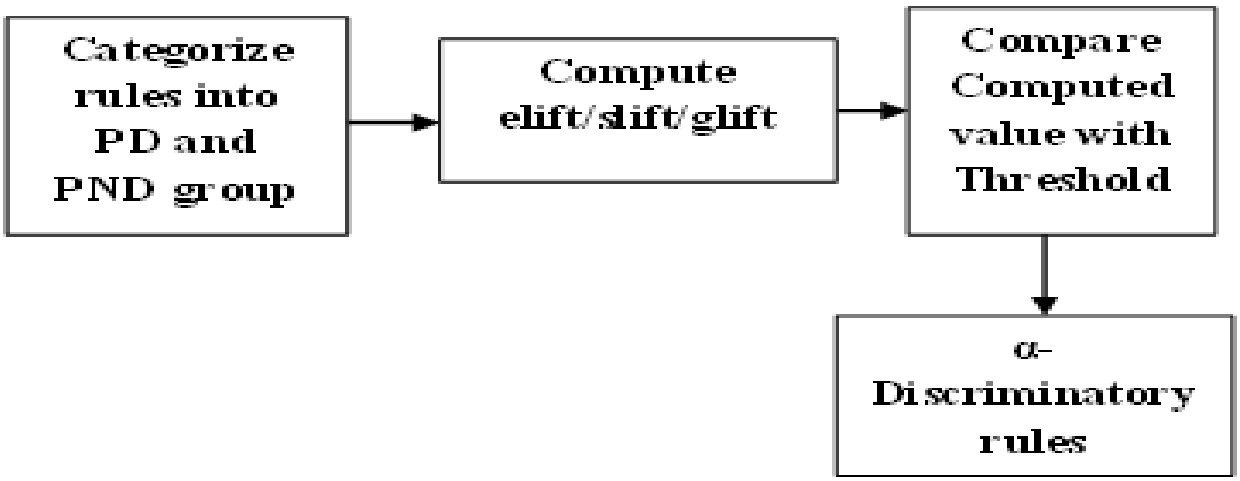

Fig 2: Discrimination Discovery Process

The data transformation is second step in discrimination prevention where the data is actually modified to make it biased free. In this step modifications are done using the definition of elift/ glift/ slift i.e. equality constraint on rule are enforced to satisfy the definition of corresponding discrimination prevention measure.

Direct Rule Protection algorithm is used here that converts α-discriminatory rules into α-protective rule using the definition of elift. It can be done in following way:
let r': α-discriminatory rule, condition enforced on r' is:

$= \text{elift}(r') < \alpha$

$= \frac{\text{Confidence}(A,B \rightarrow C)}{\text{Confidence}(B \rightarrow C)} < \alpha$

$= \text{confidence}(r': A, B \rightarrow C)/\text{confidence}(B \rightarrow C) < \alpha$
$= \text{confidence}(r': A, B \rightarrow C/\alpha < \text{confidence}(B \rightarrow C)$

Here one needs to increase confidence (B→C) in order to satisfy the condition in above equation, so change the class item from ¬C to C for all records in original DB that supports the rule of the form(¬A,B → ¬C). In this way this method changes the class label of class item in some records[9]. Similar method for slift as well as glift can be carried out.

4.2 Performance measures

To measure the success of the method in removing all evidence of Direct Discrimination and to measure quality of the modified data, following measures are used:

4.2.1 Direct discrimination prevention degree (DDPD)

The DDPD counts the percentage of α-discriminatory rules that are no longer α-discriminatory in the transformed data set.

4.2.2 Direct discrimination protection preservation (DDPP)

This measure counts the percentage of the α-protective rules in the original data set that remain α-protective in the transformed data set.

4.2.3 Misses cost (MC)

This measure helps to find the percentage of rules that are extractable from the original data set but cannot be extracted from the transformed data set. This is considered as side effect of the transformation process.

4.2.4 Ghost cost (GC)

This ghost cost quantifies the percentage of the rules that are extractable from the transformed data set but were not extractable from the original data set.

This MC and GC are the measures that are used in the context of privacy preservation. As similar approach of data sanitization is used in some methods for discrimination prevention, the same measures that are MC and GC can be applied to find out the information loss [15].

5. RESULTS AND DISCUSSION

5.1 German Credit Data set

This data set consists of 1000 records as well as 20 attributes. Out of those 20 attributes 7 are numerical and remaining 13 are categorical attributes. The class attributes indicates good or bad class for given bank account holder. Here the attribute foreign worker = Yes, Personal status = Female but not single and age = old are considered as discriminatory items.

The Table I show the partial results computed on German credit dataset containing total number of classification rules generated and number of Potentially Discriminatory rules and Potentially Non Discriminatory rules.

Table 1. German Credit dataset: Columns show the partial results for number of PD and PND classification rules

Total No. of Classification rules	No. of PD classification rules	No. of PND classification rules
8124	4293	3831

Table 2 shows the effect of different data mining algorithm, that are available in Weka, on German credit dataset where the number of correctly and incorrectly classified instances changes with presence and absence of discriminatory attributes in the given dataset.

Table 2. German Credit dataset: Result of effect of various Data Mining algorithm on classification instances with and without discriminatory attributes

Algo -rithm	With Discriminatory Attribute		Without Discriminatory Attribute	
	Correct classified instances	Incorrect classified instances	Correct classified instances	Incorrect classified instances
Simple Naïve Bayes classifier	**770**	**230**	**764**	**236**
Naïve Bayes classifier	**772**	**228**	**765**	**235**
J48 classifier	**855**	**145**	**854**	**146**

6. CONCLUSION

Discrimination can be observed not only in social sense but also in data mining. It is very important to remove such discrimination from original data. Only removing discriminatory attributes does not solve the problem. In order to prevent such discrimination, Discrimination Prevention by preprocessing technique is advantageous over the other two methods. The approach mentioned in this paper works in two steps: first is the discrimination discovery and the second is data transformation in that the original data is transformed to prevent direct discrimination. This second step follows similar approach of Data Sanitization that is used in privacy preservation context. Many such algorithms uses 'elift' as a measure of discrimination, but instead of that one may use slift, glift as a measure of discrimination. The performance measure metrics i.e. DDPD, DDPP, MC, GC analyses data to check whether discrimination has been remove completely from original data. As result of using different discrimination measures such as slift, glift the number of rules that are considered to be discriminatory is expected to be changed that may have varying impact on original data.

As future work, one may explore how rule hiding in privacy preservation or other privacy preserving algorithms helps to prevent discrimination.

7. ACKNOWLEDGMENTS

With deep sense of gratitude I thank to my guide **Prof. Dr. S. S. Sane**, Head of Department of Computer Engineering for guiding me and his constant support and valuable suggestions. I am ending this acknowledgement with deep indebtedness to my friends especially my classmates and my family members who have helped me in the successful completion of this paper.

8. REFERENCES

[1] 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. (Cited by 56)

[2] 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.

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

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

[5] 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.

[6] M.Kantarcioglu, J. Jin and C. Clifton. When do data mining results violate privacy? In KDD 2004, pp. 599-604. ACM, 2004.

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

[8] 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.

[9] S. Hajian and J. Domingo, "A Methodology for Direct and Indirect Discrimination prevention in data mining." IEEE transaction on knowledge and data engineeing, VOL. 25, NO. 7, pp. 1445-1459, JULY 2013. (Cited by 12)

[10] S. Hajian, J. Domingo-Ferrer, and A. Martnez 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.

[11] S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer and F. Giannotti. "Injecting discrimination and privacy awareness into pattern discovery," In 2012 IEEE 12th Inter-national Conference on Data Mining Workshops, pp. 360-369. IEEE Computer Society, 2012.

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

[13] 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.

[14] T. Calders and S. Verwer. "Three naive Bayes approaches for discrimination-free classification," Data Mining and Knowledge Discovery, 21(2):277-292, 2010. (Cited by 44)

[15] V. Verykios and A. Gkoulalas Divanis, "A Survey of Association Rule Hiding Methods for Privacy," Privacy-Preserving Data Mining: Models and Algorithms,C.C. Aggarwal and P.S. Yu, eds.,Springer, 2008.