Survey on Anti discrimination in Data Mining

Shilpa Gonbare[1], Satishkumar Varma[2],Manjusha Deshmukh[3]

[1]Department of Computer Engineering ,PIIT, New Panvel, India, shilpa_gonbare@yahoo.co.in
[2]Department of Information Technology ,PIIT, New Panvel, India, vsat2k@mes.ac.in
[3]Department of Computer Engineering ,PIIT, New Panvel ,India, mdeshmukh@mes.ac.in

---***---

Abstract- ***Data mining refers to the extracting knowledge from large amount of data. Data Mining automates the detection of relevant patterns in databases. Classification rule is one of the techniques used in data mining for making automated decision, like loan granting, staff selection. However, if the training data sets are unfair in what regards discriminatory attributes like age, caste, gender, nationality, etc., discriminatory decisions may ensue. To prevent such situation, antidiscrimination techniques comprising discrimination discovery and prevention have been introduced in data mining. The primary goal of this survey paper is to review and identify advantages and limitations of discrimination discovery and discrimination prevention techniques.***

Key Words:*Anti-discrimination, Data Mining, Discrimination, Discrimination Discovery, Discrimination Prevention.*

1.INTRODUCTION

Data mining refers to the extracting knowledge from large amount of data. The process of performing data analysis may reveal important data patterns,that could lead to adapt business strategies, knowledge bases, and scientific and medical research.

Data Mining automates the detection of relevant patterns in databases. For example, a pattern might indicate that married males are twice as likely to take loan than unmarried males. It uses well-established techniques to build models that predict customer behaviour.

Data mining models produces one or more output values for a given set of inputs. Analyzing data is often the process of building an appropriate model for the data. Models in Data Mining are either Predictive or Descriptive [8].

Discrimination is the prejudicial treatment of an different categories of people, especially on ground of age, nationality, etc. It involves denying opportunities to members of one group that are available to other groups. For instance, individuals may be discriminated because of their age, gender, etc. especially when these attributes are used for making decisions like offering them a job, loan, finance, etc. Person may not be selected for interview just because he is belonging to a particular religion.

Discrimination is commonly classified into two types, direct or indirect.Rules or procedures that clearly indicate deprived groups, based on discriminatory items related to particular group are known as direct discriminatory rules. Rules or procedures that, do not mention discriminatory items, but may generate discriminatory decisions are known as indirect discriminatory rules.

Indirect discrimination could happen because of the availability of some background knowledge (rules), for example, that a certain zip code corresponds to an area with mostly minority population. The background knowledge might be accessible from publicly available data or might be obtained from the original data itself because of the existence of non-discriminatory attributes that are highly correlated with the sensitive ones in the original data set [6].

Figure 1 illustrates the process of extracting discriminatory and non-discriminatory decision rules. Data analysis can lead to discriminatory rule extraction if the original dataset *DB* is biased, which could lead to automated unfair decisions. On the contrary, if the dataset *DB* goes through an anti-discrimination process ,the learned rules will be free of discrimination, as a result, fair and genuine automated decisions are enabled [7].

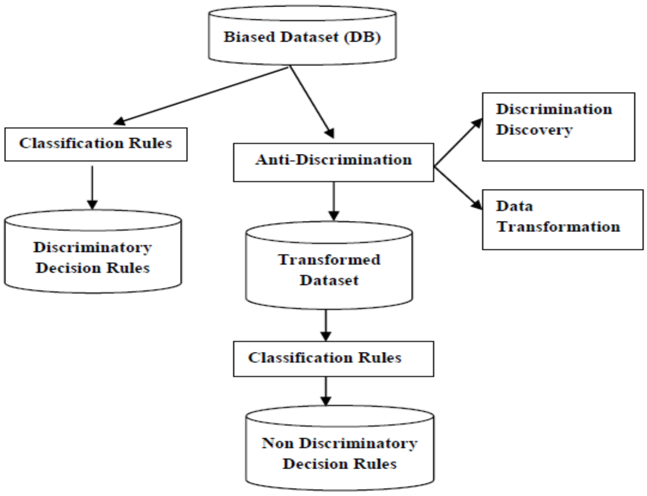


Figure 1: The Process of extracting discriminatory and non-discriminatory decision rules.

2. RELATED WORK

Even though there is a wide development in the information system based on data mining technology in decision making, the issue of antidiscrimination in data mining did not receive much attention. In this section we cite the relevant past literature that use the various anti-discrimination techniques. Some proposals are oriented to the discovery and measure of discrimination, others deals with the prevention of discrimination.

2.1 DISCRIMINATION DISCOVERY

Discrimination discovery from data consist of actual discovery of discriminatory situations and practices which are hidden in the large amount of historical decision records. The basic problem in the analysis of discrimination is to quantify the degree of discrimination suffered by a given group.

D. Pedreschi, S. Ruggieri, and F. Turini [1] addressed the discrimination problem in data mining. They gave idea about how discrimination can be discovered by measuring the discrimination through a measure known as generalization of lift. They have introduced concept of α-protection which is a threshold used to decide whether the classification rules containing one or more discriminatory items is discriminatory or non-discriminatory.

S. Ruggieri, D. Pedreschi, and F. Turini [2] have presented the discrimination discovery in databases in which unfair practices against minorities are hidden in a dataset of historical decisions. They formalized the processes of direct and indirect discrimination discovery where discrimination occurs in a classification rule based syntax. The actual discovery of discriminatory circumstances and practices is an extremely difficult task mainly because of two reasons. First, personal data in decision records are highly dimensional, i.e., characterized by many multi-valued variables: as a consequence, a huge number of possible contexts may, or may not, be the theatre for discrimination. The second source of complexity is indirect discrimination: often, the discriminatory feature e.g., the caste or nationality, is not directly recorded in the data. Figure 2.shows the discrimination discovery process [2].

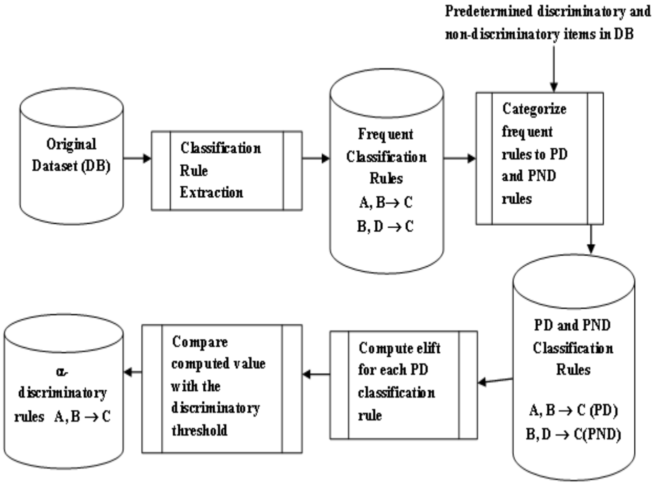

Predetermined discriminatory and non-discriminatory items in DB
Original Dataset (DB)
Classification Rule Extraction
Frequent Classification Rules A, B→ C B, D → C
Categorize frequent rules to PD and PND rules
PD and PND Classification Rules A, B → C (PD) B, D → C(PND)
Compute elift for each PD classification rule
Compare computed value with the discriminatory threshold
α-discriminatory rules A, B → C

Figure 2: Discrimination Discovery Process

2.1.1 Potentially Discriminatory and Nondiscriminatory Classification Rules

Assuming that discriminatory items in a dataset DB are predetermined e.g.(age=young, gender=female), discriminatory rules falls into one of the following two categories with respect to discriminatory and non-discriminatory items in dataset.

1. A classification rule X→C is potentially discriminatory (PD) when X=A,B with A, a nonempty discriminatory itemset and B is a non-discriminatory itemset.e.g.{age=young,Credit_history=Bad}→grant_credit=No
2. A classification rule X→C is potentially nondiscriminatory (PND) when X=D,B is a nondiscriminatory itemset.. E.g.{zip=400004,city=Mumbai}→grant_credit=No.

The word "potentially" means that a PD rule could possibly lead to discriminatory decisions. Also, a PND rule could lead to discriminatory decisions in combination with some background knowledge; *e.g.*, if the premise of the PND rule contains the zipcode as attribute and one knows that zipcode 400XXX is mostly inhabited by minority people. Hence, to quantify the direct discrimination potential and indirect discrimination potential some measures are required [6].

2.1.2 Measuring Direct Discrimination

One of measure of degree of discrimination of a PD rule is extended lift of a rule.

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

$$elift(A, B \rightarrow C) = \frac{conf(A,B \rightarrow C)}{conf(B \rightarrow C)}$$

elift evaluates the discrimination of a rule as the gain of confidence due to presence of discriminatory items (i.e. A) in the presence of a rule.

To check whether the rule is to be considered discriminatory some threshold is required. Let α€ R be a fixed threshold and let A be a discriminatory item set. A PD classification rule c = A,B→C is α-protective w.r.t. elift if elift(c) <α. Otherwise, c is α-discriminatory.

α-discriminatory rules are used to discover direct discrimination. α-discriminatory rules indicate biased rules that are directly inferred from discriminatory items [1].

2.1.3 Measuring Indirect Discrimination

The objective of indirect discrimination discovery is to identify PND rules that are to certain extent equivalent to α-discriminatory rules i.e. identifying redlining rules. Redlining rules indicate biased rules that are indirectly inferred from non-discriminatory items (e.g. zip=10451) because of their correlation with discriminatory ones.

Formal definitions of redlining and nonredlining rules are:
A PND classification rule r : 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 r_{b1} : A,B → D and r_{b2} : D,B → A, where A is a discriminatory item set. For example,{Zip = 400004; City =Mumbai}→ Hire= No.

A PND classification rule r :D,B → C is a nonredlining rule if it cannot yield any α-discriminatory rule r' :A,B→ C in combination with currently available background knowledge rules of the form r_{b1} : A,B → D andr_{b2} : D,B → A, where A is a discriminatory item set. For example{Experience = Low; City = Mumbai}→ Hire = No [1].

2.2. DISCRIMINATION PREVENTION

Discrimination prevention is a more challenging issue than discrimination discovery. The difficulty increases when we want to prevent not only direct discrimination but also indirect discrimination or both at the same time. As shown in Figure 3 discrimination prevention methods fall into three groups: pre-processing,in-processing and post-processing approaches.

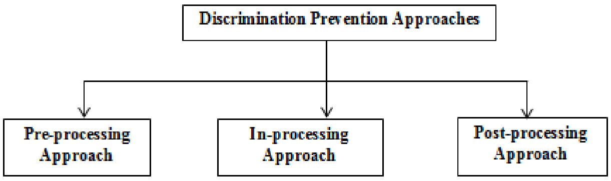


Figure 3: Discrimination Prevention Approaches

2.2.1 Pre-processing Approach.

Methods in this group tend to transform the source data in such a way that the discriminatory biases contained in the original data are removed, leading to unfair decision rule extraction from the transformed data. The preprocessing approach is useful for applications in which a data set should be published as well as in which data mining needs to be performed also by external parties and not just by the data holder.

Some of the work done in preprocessing approaches are:
F Kamiran, T Calders [3] had introduced a new classification scheme for learning unbiased models on biased training data, which is referred to as Classification with no discrimination (CND).

Their method massages the dataset by making the least intrusive changes which lead to an unbiased dataset. Massaging is done by changing class labels of selected objects in the training data so as to obtain a unbiased dataset. For massaging the data, first a ranking functions is learned on the biased data. This ranker is then used to rank data object according to their probability of being in the desired class. The class labels of the most likely victims (discriminated community with a negative label but a high positive class probability) and profiteers (favored community with a positive label but a low positive class probability) are changed. Then based on the sanitized data, a non-discriminatory model is learned which reduces the prejudicial behavior for future classification. Numerical attributes and group of attributes are not considered as sensitive attribute.

F. Kamiran et al.[4],in "Classification with no Discrimination by Preferential Sampling" introduced a Preferential Sampling (PS) scheme to make dataset bias free. To make the dataset discrimination free, they changed the distribution of data objects close to decision boundaries. Identification of the borderline object was done by learning a ranker on the training data. This ranker was used by PS to class the data objects of DP (*D*iscriminated community with *P*ositive class labels) and PP (*P*rivileged community with *P*ositive class labels) in ascending order, and the objects of DN (*D*iscriminated community with *N*egative class labels) and PN (*P*rivileged community with *N*egative class labels) in descending order; both w.r.t. the positive class probability. The data objects closer to the borderline have higher rank. Starting from the original dataset PS iteratively duplicates (for the groups DP and PN) and removes objects (for the groups DN and PP) in the following way: Decreasing the size of a group is always done by removing the data objects closest to the borderline. Increasing the sample size is done by duplication of the data object closest to the borderline. When an object has been duplicated, together with its duplicate, it is moved to the bottom of the ranking. The procedure is repeated until the desired number of objects is obtained.

PS works in the following steps:
(i) Data object is divided into the four groups, DP, DN, PP, and PN.
(ii) Any ranking algorithm is used on a complete training data to arrange the data object according to their probability of being in class +. This ranking will be used to identify the borderline data objects.
(iii) Expected size for each group is calculated to make the dataset bias free.
(iv) Finally sampling with replacement is applied to increase the size of DP and PN and decrease the size of DN and PP. Now this modified dataset is used for learning a discrimination free classifier.
However

- Discrimination in the original data was detected only for one discriminatory item and based on a single measure.
- Only direct discrimination was considered.
- They lack measure to evaluate how much discrimination has been removed and how much information loss has been incurred.

S. Hajian and J. Domingo-Ferrer [6] have proposed a new techniques applicable for direct or indirect discrimination prevention individually or both at the same time. They have discussed the cleaning of training data sets and outsourcing the datasets in such a way that direct and/or indirect discriminatory decision rules are converted to legitimate (nondiscriminatory) classification rules. They have also proposed new metrics to evaluate and compare of the proposed approaches.

This method can be described in terms of two phases:
Discrimination measurement-Direct and indirect discrimination discovery includes identifying α discriminatory rules and redlining rules.

(i) Based on predetermined discriminatory items in DB, frequent classification rules in FR are divided in two groups: PD and PND rules.
(ii) Direct discrimination is measured by identifying α-discriminatory rules among the PD rules using a direct discrimination measure (elift) and a discriminatory threshold (α).
(iii) 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).

Data transformation- Transform the original data DB in such a way to remove direct and/or indirect discriminatory biases, with minimum effect on the data and on legitimate decision rules, so that no unfair decision rule can be mined from the transformed data.
Transformation Method*:* There are two transformation method used in both direct and indirect discrimination removal.
(i) Direct Rule Protection– Which convert each α-discriminatory rule into a α-protective rule, based on the direct discriminatory measure. *elift (r′) <α*
(ii) Indirect Rule Protection– Which convert a redlining rule into an non-redlining rule, based on the indirect discriminatory measure [6],we should enforce the following inequality for each redlining ruler: D,B→C in RR:*elb (γ, δ) <α*

These two data transformation method for used simultaneous direct and indirect discrimination prevention.

Utility Measures:
These techniques should be evaluated based on two aspects.

- To measure the success of the method in removing all evidence of direct and/or indirect discrimination from the original data set.
- To measure the impact of the method in terms of information loss

To measure discrimination removal, four metrics were used:
(i) Direct discrimination prevention degree (DDPD):
This measure quantifies the percentage of α-discriminatory rules that are no longer α-discriminatory in the transformed data set.
(ii)Direct discrimination protection preservation(DDPP):
This measure quantifies the percentage of the α-protective rules in the original data set that remain α-protective in the transformed data set.
(iii) Indirect discrimination prevention degree (IDPD):
This measure quantifies the percentage of redlining rules that are no longer redlining in the transformed data set.
(iv)Indirect discrimination protection preservation(IDPP):
This measure quantifies the percentage of nonredlining rules in the original data set that remain nonredlining in the transformed data set.

2.2.2 In-processing Approach

Methods in this group change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules

F. Kamiran, T. Calders and M. Pechenizkiy [5] gave an approach in which the nondiscriminatory constraint is pushed deeply into a decision tree learner by changing its splitting criterion and pruning strategy by using a novel leaf relabeling approach. Two techniques, *Dependency-Aware Tree Construction &Leaf Relabeling* for incorporating discrimination awareness into the decision tree construction process was proposed

However, in-processing discrimination prevention methods must rely on new special purpose data mining algorithms; standard data mining algorithms cannot be used

2.2.3 Post-processing Approach

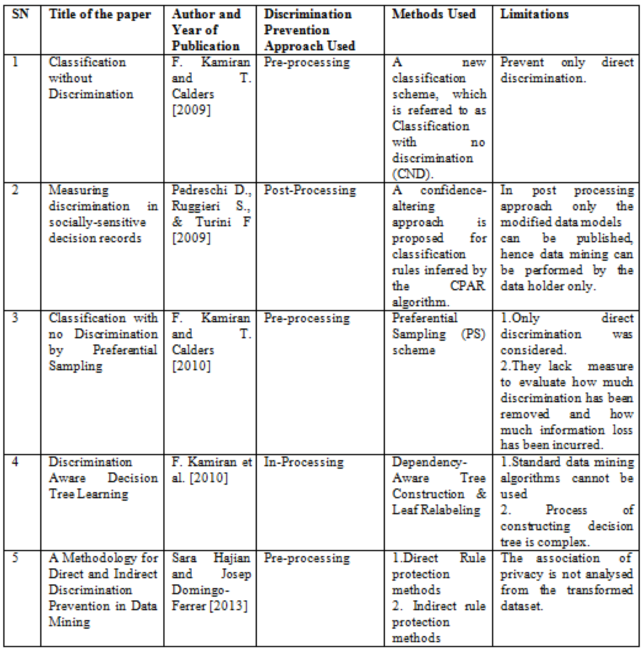
These methods modify the resulting data mining models, instead of cleaning the original data set or changing the data mining algorithms. For example, in , a confidence-altering approach is proposed for classification rules inferred by the rule-based classifier: CPAR(classification based on predictive association rules) algorithm. The post-processing approach does not allow the data set to be published: only the modified data mining models can be published, hence data mining can be performed by the data holder only.

Pedreschi et al. [9], proposed the extraction of classification rules of the form A, B →C, called potentially discriminatory (PD) rules, to unveil contexts B of the dataset where the protected group A suffered from under representation w.r.t the positive decision C or from over-representation w.r.t. the negative decision C. A is a non-empty itemset, whose elements belong to a fixed set of protected groups. C is a class item denoting the negative decision, e.g., credit denial, application rejection, job firing, and so on. Finally, B is an itemset denoting a context of possible discrimination. The degree of over-representation is measured by the ER measure (called extended lift). For example:RACE =BLACK, PURPOSE = NEWCAR → CREDIT = NO; is a PD rule about denying credit (the decision C) to blacks (the protected group A) among those applying for credit in order to buy a new car (the context B).PD rules are ranked according to their measure value.

3. COMPARATIVE STUDY

The comparison of existing work done in discrimination prevention is given in Table 1. It gives idea about which approach is used, which technique is used for discrimination preventions. It also gives the limitations of those work.

Table 1: Comparison of existing work



SN	Title of the paper	Author and Year of Publication	Discrimination Prevention Approach Used	Methods Used	Limitations
1	Classification without Discrimination	F. Kamiran and T. Calders [2009]	Pre-processing	A new classification scheme, which is referred to as Classification with no discrimination (CND).	Prevent only direct discrimination.
2	Measuring discrimination in socially-sensitive decision records	Pedreschi D., Ruggieri S., & Turini F [2009]	Post-Processing	A confidence-altering approach is proposed for classification rules inferred by the CPAR algorithm.	In post processing approach only the modified data models can be published, hence data mining can be performed by the data holder only.
3	Classification with no Discrimination by Preferential Sampling	F. Kamiran and T. Calders [2010]	Pre-processing	Preferential Sampling (PS) scheme	1.Only direct discrimination was considered. 2.They lack measure to evaluate how much discrimination has been removed and how much information loss has been incurred.
4	Discrimination Aware Decision Tree Learning	F. Kamiran et al. [2010]	In-Processing	Dependency-Aware Tree Construction & Leaf Relabeling	1.Standard data mining algorithms cannot be used 2. Process of constructing decision tree is complex.
5	A Methodology for Direct and Indirect Discrimination Prevention in Data Mining	Sara Hajian and Josep Domingo-Ferrer [2013]	Pre-processing	1.Direct Rule protection methods 2. Indirect rule protection methods	The association of privacy is not analysed from the transformed dataset.

4. CONCLUSION

Discrimination is crucial issues when considering the legality and morality of data mining. People never like to be distinguished because of their gender, caste, nationality, age, and so on, specifically while making decisions like offering them a job, house, insurance, etc.

The most concentration is on producing training data which are free or nearly free from discrimination. In order to control discrimination in a dataset, a first step consists of discovering whether there exists discrimination. If any discrimination is found, the dataset will be modified until discrimination is brought below a certain threshold or is entirely eliminated.

In this paper, we reviewed preprocessing, in-processing and post-processing approaches of discrimination prevention in data mining and found that discrimination prevention in data mining is extremely difficult and challenging.

REFERENCES

[1] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination Aware Data Mining", Proc. 4th ACMInternationalConference knowledge Discovery and Data Mining (KDD '08), pp. 560-568, 2008.
[2] S. Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery," ACM Transactions on Knowledge Discovery from Data, vol. 4, no. 2, article 9, 2010.
[3] F. Kamiran and T. Calders, "Classification without Discrimination," Proc. IEEE Second International ConferenceComputer, Control and Comm.(IC4 '09), 2009.
[4] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf. Belgium and The Netherlands,2010.
[5] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," Proc. IEEE International ConferenceData Mining(ICDM '10), pp. 869-874, 2010.
[6] Sara Hajian and Josep Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 7,pp. 1445-1459, July 2013.
[7]Sara Hajian and Josep Domingo-Ferrer, "Direct and Indirect Discrimination Prevention Methods", Discrimination and Privacy in the Information Society, Springer Berlin Heidelberg, pp. 241-254, 2013.
[8] M.H. Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education, first edition, 2008.
[9] Pedreschi D., Ruggieri S., and Turini F., "Measuring discrimination in socially-sensitive decision records", 9th SIAM Conference on Data Mining (SDM 2009), pp. 581-592. 2009.