ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 3, September 2014

A Survey on Anti-Discrimination Techniques

Naveena M.S, Robert S
M. Tech Student, Asst. Professor, Dept. of CSE
Marian Engineering College, Marian Engineering College
Trivandrum

Abstract- Discrimination is the unfair treatment of people on the basis of number of attributes like race, religion, gender, age, etc. Discrimination can be either direct or indirect. Direct discrimination occurs when the decisions are made based on the sensitive attributes. Indirect discrimination occurs when the decisions are made based on the nonsensitive attribute which are strongly correlated with sensitive attribute. Automated data collection and data mining techniques such as classification rule mining have paved the way to making automated decisions. If the training data sets are biased in discriminatory attributes, discriminatory decisions may arise. For this reason, antidiscrimination techniques including discrimination discovery and prevention have been introduced in data mining. The focus of this paper is to provide a brief survey on the techniques used for the discrimination prevention and their comparisons.

***Index Terms*-Data Mining; Discrimination Discovery; Discrimination Prevention; Preprocessing Techniques.**

I. INTRODUCTION

Data mining is the technology of extracting useful knowledge hidden in large collections of data. But, extracting of the knowledge without the violation of privacy and non-discrimination is the most challenging task. This is mainly because of the data mining techniques such as classification rules are learned by the system from the training data and the training data sets itself are biased in discriminatory attributes like age, gender, caste, etc. Discrimination is the prejudicial treatment of an individual or a group on the basis of their membership on a particular category. Discrimination can be classified into direct or indirect (systematic, [2]). Direct discriminatory rules indicate biased rules that are directly inferred from discriminatory items (e.g. Foreigner = Yes). Indirect discriminatory rules (redlining rules) indicate biased rules that are indirectly inferred from nondiscriminatory items (e.g. Zip = 10451) because of the correlation with discriminatory ones. Indirect discrimination could happen because of the availability of background knowledge. It might be accessible from publicly available data (e.g. census data) or might be obtained from the original data set itself.

II. DISCRIMINATION DISCOVERY

The accessing to hidden historical data concerning decisions without violation such as privacy and non-discrimination is the introductory point for discovering discrimination. It is a difficult task. The reasons are as:

- ***Personal data in decision records are highly dimensional:*** Due to this, a huge number of possible contexts may be the theater for discrimination.
- ***Complexity in indirect discrimination:*** the feature that may be the object of discrimination is not directly recorded in the data.

The problem of discovering discrimination in the records of decisions taken by human decision makers, and in the recommendations provided in the classification models, or their combinations are described in [3] and it extends the methodology used in [2]. The notions of discriminatory rules are introduced as a criterion to identify and analyze the potential risk of discrimination. By mining all the classification rules from a data set of decision records, [3]offers a sound and practical method to discover niches of direct and indirect discrimination as well as the criterion to measure discrimination.

III. DISCRIMINATION PREVENTION

The prevention of the knowledge based decision support systems from making discriminatory decisions is the most challenging issue beyond discrimination discovery [2]. It can be even more difficult when we want to prevent not only direct discrimination but also indirect discrimination or both at the same time simultaneously. The classification of the discrimination prevention methods is related to the way of eliminating discrimination and also to the phase of the data mining process in which the discrimination prevention is done [3]. Based on this criterion, the discrimination prevention methods are classified into three groups:

Preprocessing:

Removing of discrimination from original source data in such a way that no unbiased rule can be mined from the transformed data and applying any standard algorithm. This preprocessing approach is useful in such cases where data set should be published and performed by external parties.

Inprocessing:

Changing of the knowledge discovery algorithm in such a way that resulting model do not contain biased decision rules. In processing discrimination prevention depends on new special purpose algorithm. In this standard data mining algorithm cannot be used.

Post processing:

Instead of removing biases from original data set or modify the standard data mining algorithm, resulting data mining models are modified. This approach does not allow the data set to be published; only modified mining models can be published. So this can be performed only by data holder. The preprocessing approach is used for discrimination prevention since it is flexible [3]. It does

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 3, September 2014

not require changing of the standard data mining algorithms, unlike in inprocessing approach, and it allows data publishing (rather than just knowledge publishing), unlike the post processing approach. The different approaches used to solve the discrimination aware classification problem is stated below:

A. Massaging of data

The idea of classification with no discrimination and a solution based on massaging the data by changing class labels of selected objects in the training data with least possible changes is described in [5]. For massaging the data, first a ranker for predicting the class attribute without taking in account the discrimination is learnt. This ranker is then used to rank the data objects according to their probability of being in the desired class. The class labels of the most likely victims (training instances of the discriminated community with a negative label but a high positive class probability) and profiteers (training instances of the favored community with a positive label but a low positive class probability) are changed. The modified data is then used for learning a classifier with no discrimination for future decisions. The drawback of massaging the data is that it is very intrusive.

B. Preferential sampling

Preferential Sampling (PS) changes the distribution of different data objects for a given data to make it discrimination free. The idea is that the data objects close to the decision boundaries are more prone to the victim of discrimination. Then the distribution of this borderline objects is changed to make the dataset discrimination free. To know the least certain elements, [6] use a ranking function, learned on original data, to identify the data objects close to the borderline. PS uses this ranker to class the data objects of DP (Discriminated community with Positive class labels) and PP (Privileged community with Positive class labels) in ascending order, and the objects of DN (Discriminated community with Negative class labels) and PN (Privileged community with Negative class labels) in descending order; both w.r.t 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 (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.

Data points of the desired class and the negative class are represented by + and – symbols respectively. Then, based on the sanitized data, a nondiscriminatory model can be learned [6]. Since the model is learned on non-discriminatory data, it reduces the prejudicial behavior for future classification. This approach gives similar performance to "massaging" [5] but without changing the data set and always outperforms the "reweighing" scheme of previous method. Classification with No Discrimination by Preferential Sampling [6] is an excellent solution to the discrimination problem. It gives promising results with both stable and unstable classifiers give more accurate results but do not reduce the discrimination. The drawbacks are low data utility rate and minimum discrimination removal. This PS is also not applicable for indirect discrimination.

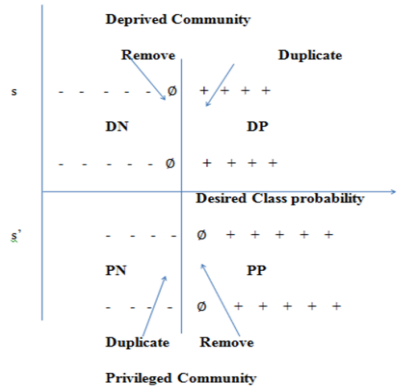

Fig. 1. Pictorial representation of Preferential sampling [6]

C. Decision Tree Learning Approach

The removing of the sensitive attribute from the training data does not work, as other attributes may be correlated with the suppressed attribute [4]. It was observed that classifiers tend to pick up these relations and discriminate them indirectly. Therefore, the preprocessing methods that will cleanse away the discrimination were proposed. In [9] another solution based on the integration of discrimination awareness into the model induction process of a decision tree were explored. Particularly, the following two techniques for incorporating discrimination awareness into the decision tree construction process:

Dependency-Aware Tree Construction. When evaluating the splitting criterion for a tree node, not only its contribution to the accuracy, but also the level of discrimination caused by this split is evaluated.

Leaf Relabeling. In a decision tree, the label of a leaf is determined by the majority class of the tuples that belong to this node in the training set. In leaf relabeling we change the label of selected leaves in such a way that discrimination is lowered with a minimal loss in accuracy. This method gives high accuracy and low discrimination scores when applied to nondiscriminatory test data. The enrichment in discrimination reduction with the relabeling method is very satisfying [9]. The relabeling reduces the discrimination to almost zero under certain conditions. The relabeling methods outperform the baseline in almost all cases. As such it is reasonable to

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 3, September 2014

say that the straightforward solution is not satisfactory and the use of the dedicated discrimination-aware techniques are justified. These methods will significantly improve the current state-of-art techniques [6] w.r.t. accuracy discrimination trade off. Discrimination removal is very low using relabeling method.

D. Data Transformation Methods

The discrimination prevention can be done under two phases:

Discrimination measurement: *B*ased on the identified α-discriminatory rules and the redlining rules, the frequent classification rules are grouped into Potentially Discriminatory (PD) and Potentially Non Discriminatory (PND). Then calculate the respective discrimination measure for direct and indirect discrimination.

Rule Protection for Indirect Discrimination Prevention: A method for indirect discrimination prevention based on data transformation that considers several discriminatory attributes and their combinations are introduced in [8]. The major drawback of this method is that only preliminary experiments are conducted.

Unified Preprocessing Discrimination Prevention

The new data transformation methods are based on both direct and indirect discrimination prevention simultaneously or both at the same time. The metrics which specify which records should be changed, how many records should be changed and how those records should be changed during data transformation are described in [1]. Extensive experimental results and utility measures are carried on Adult Data Sets and German Credit Data Sets. Both the values of direct discrimination removal and indirect discrimination removal measures shows high success in discrimination prevention. The data quality measures sound less information loss by implementing this method. The drawbacks addressed by this method are that nonbinary attributes are not considered and privacy preservation is not mentioned.

Data Transformation: Transform the original data to remove discriminatory biases either direct/indirect, with minimum impact on the data [1], [7], [8]. The data transformation methods are based on direct and indirect discrimination prevention methodology. Transformation methods include rule protection (RP) and rule generalization (RG).

Data mining for intrusion and crime detection: Automated data collection has encouraged the use of data mining for intrusion and crime detection. A new discrimination prevention method based on data transformation and the measures to evaluate the success in discrimination prevention and its impact on data quality are introduced in [7]. Antidiscrimination in the context of cyber security is considered. The drawback found is that only direct discrimination was addressed

Table 1: Comparison of Discrimination Prevention Techniques

Discrimination Prevention Techniques	Advantages	Disadvantages
Massaging the data set	High utility rate	Intrusive in nature
Preferential Sampling	High accuracy level	Concentrates in border regions more
Decision Tree Learning	Lower discrimination score	Construction of decision tree is complex
Transformation methods 1) Discrimination prevention for intrusion detection	Detects crime intrusion	Only direct discrimination prevention
2) Rule protection for indirect discrimination	Discovers discrimination in databases	No experimental proof
3) Unified approach	Prevents direct and indirect discrimination	Privacy preservation not mentioned

V. CONCLUSION AND FUTURE WORK

The result of [1] shows that discrimination prevention whether direct or indirect can be tackled easily. The main issue is that the association or interaction between discrimination prevention and privacy preservation, which is not included in [1]. By establishing privacy protection, the extent to which how much discrimination can be removed can be done as the future work.

REFERENCES

[1] Sara Hajian, 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, July 2013.

[2] 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.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 3, September 2014

[3] Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery," ACM Trans. Knowledge Discovery from Data, vol. 4, no. 2, article 9, 2010.

[4] 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.

[5] Kamiran and Calders, "Classification without Discrimination,"Proc. IEEE Second Int'l Conf. Computer, Control and Comm. (IC4 '09), 2009.

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

[7] S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection," Proc. IEEE Symp. Computational Intelligence in Cyber Security (CICS '11), pp. 47-54

[8] 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, and 2011.

[9] 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.