RESEARCH ARTICLE OPEN ACCESS

Discrimination Discovery and Prevention in Data Mining: A Survey

Jagriti Singh*, Prof. Dr. S. S. Sane**
*Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, University of Pune, Maharashtra - 422003, India
**Head of Department, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, University of Pune, Maharashtra - 422003, India

ABSTRACT
Data Mining is the computation process of discovering knowledge or patterns in large data sets. But extract knowledge without violation such as privacy and non-discrimination is most difficult and challenging. This is mainly because of data mining techniques such as classification rules are actually learned by the system from the training data and training data sets itself are biased in what regards discriminatory (sensitive) attributes like gender, race, religion, etc. As a result actual discovery of discrimination situations, practices may be extremely difficult task. The focus of this paper is to provide a brief survey of the researcher's works on discrimination discovery and prevention in the field of data mining.

Keywords **–** Discrimination Discovery, Discrimination Measure, Data Mining, Discrimination Prevention, Preprocessing Technique

I. INTRODUCTION

Data Mining is the computation process of discovering knowledge or patterns in large data sets. But extract knowledge without violation such as privacy and non-discrimination is most difficult and challenging. This is mainly because of data mining techniques such as classification rules are actually learned by the system from the training data and training data sets itself are biased in what regards discriminatory (sensitive) attributes like gender, race, religion, etc. As a result actual discovery of discrimination situations, practices may be extremely difficult task. Privacy refers to the individual right while discrimination refers to unfair or unequal treatment of people. From a legal perspective, discrimination arises only on application of different rules or of the same rule or practice to different situations or practices to comparable situations.

There are two types of discrimination, one is direct and another is indirect discrimination. Direct discrimination is pretty straightforward in most cases. It happens due to dealt with unfairly on the basis of one of the grounds (compared with someone who doesn't have that ground) and in one of the areas covered by the Act. Sometimes direct discrimination is also called as Systematic Discrimination.

Indirect discrimination is often less obvious. Sometimes, a policy, rule or practice seems fair because it applies to everyone equally, but a closer look shows that some people are being treated unfairly. This is because some people or groups of people are unable or less able to comply with the rule or are disadvantaged because of it. If this policy or practice is 'not reasonable', it may be indirect discrimination or sometime called as disparate impact. Government plays a vital role in the prevention and reduction of discriminations, by enforcing different type of anti-discrimination laws.
In this paper, we review the existing work on discrimination discovery and prevention techniques in data mining.

II. BACKGROUND

The computerization and automation have substantially enhanced our capabilities for both generating and collecting data from diverse sources. A large amount of data has been generated from almost every aspect of our lives. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. This has led to the generation of a promising and flourishing frontier in computer science called data mining.

But to extract knowledge without violation such as privacy and non-discrimination is most difficult and challenging. The reasons are as:

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

The word discrimination originates from the Latin discriminare, which means to "distinguish between". There is need of disruptive technologies for the construction of human knowledge discovery systems that, by design, over native technological safeguards against discrimination. To ensure this, these computational models should be free from discrimination and for the same researchers has suggested different technologies for prevention of discrimination in data mining.

There are three different approaches for discrimination prevention in data mining:

- *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.
- *In-processing:* Change of 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.
- *Postprocessing:* 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.

Although some of the methods have already been proposed for each of the above mentioned approach, but still this is a challenge to remove the discrimination from the original data set.

III. LITERATURE REVIEW

Discrimination prevention has been recognized as an issue in a tutorial by (Clifton, 2003) [1] where the danger of building classifiers capable of racial discrimination in home loans has been put forward. Data mining and machine learning models extracted from historical data may discover traditional prejudices for example, mortgage redlining can be easily recognized as a common pattern in loan data but so solution was provided in this tutorial.

The data mining techniques such as classification and association rules, when used for decision tasks such as benefit or credit approval, found that results are in discriminatory in nature. This deficiency of classification and association rules poses ethical and legal issues, as well as obstacles to practical application. D. Pedreschi, S. Ruggieri, and F. Turini [2] have presented the first kind of papers, which address the discrimination problem in data mining models in 2008. They have investigated how discrimination may be hidden in data mining models and also measured the discrimination through a generalization of lift. They have introduced α protection as a measure of the discrimination power of a classification rule containing one or more discriminatory items.

Pedreschi et al. (2008) [3]; propose 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 underrepresentation 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 groupA) among those applying for credit in order to buy a new car (the context B). PD rules are ranked according to their measure value. F Kamiran, T Calders [4] had tackled the problem of impartial classification by introducing a new classification scheme for learning unbiased models on biased training data in 2009. Their method is based on massaging the dataset by making the least intrusive modifications which lead to an unbiased dataset. Numerical attributes and group of attributes are not considered as sensitive attribute.

S. Ruggieri, D. Pedreschi, and F. Turini (2010) [5], have presented the discrimination discovery in databases in which unfair practices against minorities are hidden in a dataset of historical decisions. The DCUBE system, based on classification rule extraction and analysis implements the approach which is centering the analysis phase on an Oracle database. The proposed demonstration guides the audience through the legal issues about discrimination hidden in data, and through several legally-grounded analyses to unveil discriminatory situations. The SIGMOD attendees will freely pose complex discrimination analysis queries over the database of extracted classification rules, once they are presented with the database relational schema, a few ad-hoc functions and procedures, and several snippets of SQL queries for discrimination discovery. In another paper, they have also have presented a systematic framework for measuring discrimination, based on the analysis of the historical decision records stored out of a socially-sensitive decision task, e.g. insurance. They investigate whether evidence of direct and indirect discrimination can be found in a given set of decisions, by measuring the degree of discrimination of a rule that formalizes an

expert's hypothesis. They have also implemented LP2DD [6] approach by integrating induction and deduction for finding evidence of discrimination of the overall reference model. They had discussed integrating induction, through data mining classification rule extraction, and deduction, through a computational logic implementation of the analytical tools in 2009.

I. Zliobaitye, F. Kamiran, and T. Calders (2011) [7] have used historical data for supervised learning may contain discrimination. They have studied how to train classifiers on such data, so that they are discrimination free with respect to a given sensitive attribute; e.g., gender. Existing techniques did not take into account of the discrimination explainable by other attributes, such as, e.g., education level only dealt in removing all discriminations. They have analyzed and introduced the conditional non-discrimination in classifier design. They observed that in such cases, the existing discrimination aware techniques will introduce a reverse discrimination, which is undesirable as well. Therefore, they have developed local techniques for handling conditional discrimination when one of the attributes is considered to be explanatory. Experimental evaluation demonstrates that the new local techniques remove exactly the bad discrimination, allowing differences in decisions as long as they are explainable.

B Luong, S Ruggieri, F Turini [8] had modeled the discrimination discovery and prevention problems by a variant of k-NN classification that implements the legal methodology of situation testing in 2011. Major advancements over existing proposals consist in providing: a stronger legal ground, overcoming the weaknesses of aggregate measures over undifferentiated groups; a global description of who is discriminated and who is not in discrimination discovery; a discrimination prevention method that is independent from the classification model at hand; the cleaned dataset obtain by method is probably more desirable as it contain less "illegal inconsistencies." But for discrimination –aware classification, it is unclear if the obtained dataset is suitable for learning a discrimination-free classifier.

F. Kamiran and T. Calders (2012) [9] presented algorithmic solutions that preprocess the data to remove discrimination before a classifier is learned. They have proposed three preprocessing techniques i.e. *Massaging*, *Reweighing* and *Sampling* which applies on training dataset. These preprocessing techniques have been implemented in a modified version of Weka and presented the results of experiments on real-life data. These preprocessing methods for prevention of discrimination are as below:

- *Suppression:* Finding the attribute which correlate most with the sensitive attribute S. Remove S and most correlated attribute, to reduce the discrimination between the class levels and attribute.
- *Massaging the dataset:* Discrimination can be removed from the dataset by changing the labels of some objects in dataset. The best candidates for relabeling can be select with help of ranker.
- *Reweighing:* Instead of change in some of the labels of some objects, assigning the weights in training data set's tuples. By carefully assigning the weights, the training data set can be made discrimination free without changing the labels in the dataset.
- *Sampling:* This method can be used where weights cannot be used directly. Sample sizes for the 4 combinations of sensitive attribute S- and Class-values will make the dataset discrimination free. Applying stratified sampling on the four groups will make two of the groups as under sampled and two will be over sampled. Then with help of two techniques, *Uniform Sampling* and *Preferential Sampling* for selecting the objects to duplicate, and to remove.

S. Hajian and J. Domingo-Ferrer (2012) [10] 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 data sets 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. They have demonstrated that the proposed techniques are effective at removing direct and/or indirect discrimination biases in the original data set while preserving data quality with help of experiments.

A. Romei, S. Ruggieri [11] have published an annotated bibliography of the main references and recent approaches on discrimination data analysis in 2013.

IV. DISCUSSION AND FUTURE WORK

Since most of effective decision models of data mining are constructed on the basis of historical decision records e.g., in credit scoring procedures and in credit card fraud detection systems, there is no guarantee that the extracted knowledge does not incur discrimination. This may be because the data from which the knowledge is extracted contain patterns with discriminatory bias. Hence, data mining from historical data may lead to the discovery of traditional prejudices. Thus prevention of discrimination knowledge based decision support systems; discovery is a more challenging issue. Some of the proposed techniques have been revived and based on the above review this is common that to

some extent accuracy must be traded-off for lowering the discrimination. This trade-off was studied and confirmed theoretically.

Some of the future works in the area of discrimination prevention in data mining are to extend the discrimination prevention techniques to:

- A multiple class problem by simply assuming one class as the desired class value and the rest of the class values as the not-desired category and vice versa
- To measure and evaluate how much discrimination has been removed by the above techniques

V. CONCLUSION

In this paper, we reviewed the existing work on discrimination discovery and prevention techniques in data mining and found that discrimination prevention in data mining is extremely difficult and challenging. Since most of us do not want to be discriminated based on our gender, religion, nationality, age and so on, especially when these attributes are used for making decisions about our jobs, loans, insurance and many more which effect human life. Discrimination of any form must be detected and removed to get the unbiased results.

REFERENCES

[1] C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? *In Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (KDD 2003), Tutorial, Washington, DC (USA), 2003.

[2] D. Pedreschi, S. Ruggieri, and F. Turini, Discrimination-aware data mining. In Y. Li, B. Liu, and S. Sarawagi, editors, *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (KDD 2008), pages 560–568.*ACM, 2008*.

[3] D. Pedreschi, S. Ruggieri, and F. Turini, Measuring discrimination in socially-sensitive decision records, *In Proc. Of the SIAM Int. Conf. on Data Mining* (SDM 2009), pages 581–592. *SIAM, 2009*.

[4] F Kamiran, T Calders, Classifying without discriminating, *Proceedings of IEEE IC4 International conference on Computer, Control and Communication. (2009a) IEEE Press.*

[5] S. Ruggieri, D. Pedreschi, and F. Turini, DCUBE: Discrimination discovery in databases In A. K. Elmagarmid and D. Agrawal, editors, *Proc. of the ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2010)*, pages, 1127–1130. *ACM, 2010c*.

[6] S. Ruggieri, D. Pedreschi, and F. Turini, Integrating induction and deduction for finding evidence of discrimination, *Artificial Intelligence and Law, 18(1)*:1–43, 2010.

[7] I. Zliobaitye, F. Kamiran, and T. Calders, Handling conditional discrimination, *In Proc. of the IEEE Int. Conf. on Data Mining* (ICDM 2011), pages 992–1001. I*EEE Computer Society*, 2011.

[8] B Luong, S Ruggieri, F Turini, K-nn as an implementation of situation testing for discrimination discovery and prevention, *Technical Report TR-11-04*: (2011), Dipartimento di Informatica, Universita di Pisa.

[9] F. Kamiran and T. Calders, Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst (2012)* 33:1–33.

[10] S. Hajian and J. Domingo-Ferrer, A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, page to appear, 2012.

[11] A. Romei, S. Ruggieri, A multidisciplinary survey on discrimination analysis, *The Knowledge Engineering Review*, 2013.

AUTHOR BIBLOGRAPHY

Jagriti Singh is Post Graduate Student of Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, University of Pune, Maharashtra, India. She received her B.Tech Degree in Information Technology from Uttar Pradesh Technical University, Uttar Pradesh.

Prof. Dr. S.S. Sane is the Head of Computer Engineering Department, K. K. Wagh Institute of Engineering Education and Research, Nashik, University of Pune, Maharashtra, India.