

# STUDY OF PRIVACY DISCRIMINATION AND PREVENTION IN DATA MINING

<sup>1</sup>Ms Madhuri P. Gole, <sup>2</sup>Prof Sachin Vyawahare

<sup>1</sup>ME Scholer Sanmati Engineering College, Washim."

<sup>2</sup>Assistant Professor, Sanmati Engineering College Washim (MH)"

\*\*\*

**Abstract** - Day by day new technologies are developed and simultaneously huge amount of data is collected, integrated and analyzed in data warehouses using data mining techniques. However, the major problem arises in data mining is privacy and discrimination. Classification is only one technique used in data mining for making decision. In data mining discrimination is of two types direct and indirect. Direct discrimination works only when decisions are made based on sensitive attributes. Indirect discrimination occurs when decisions are made on non sensitive attributes which are strongly related with sensitive. Therefore, in this paper going to study the privacy discrimination methods in data mining and also focuses on discrimination measurement and prevention in data mining

**Key Words:** classification, Discrimination, privacy discrimination, prevention etc

## 1. INTRODUCTION

In data mining, discrimination is one of the major problems discussed in the recent literature. Discrimination denies the members of one group with others. A law is designed to prevent discrimination in data mining. Discrimination can be done on attributes viz. religion, nationality, marital status and age. A large amount of data is collected by organization credit card companies, bank and insurance agencies. Thus, these collected data are auxiliary utilized by companies for decision making purpose in data mining techniques. The association and or classification rules can be used in making the decision for loan granting and insurance computation. Discrimination can be direct and indirect. Direct discrimination consists of rules or procedures that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership. Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or unintentionally could generate discriminatory decisions.

## 2. LITURATURE REVIEW

Rupanjali T.Dive et. al. [1] reviews the recent the approaches for antidiscrimination techniques and also focuses on discrimination discovery and prevention in data mining. On the other hand, they also study a theoretical approach for enhancing the results of the data quality and also discuss how to clean training data sets and outsourced

data sets in such a way that direct discrimination decision rules are change into the legitimate classification rules.

Miss. Melanie Ann Thomas et. al. [2] proposed the system that tackles discrimination in data mining by using methods like Rule protection, Rule generalization and eligibility of the client is done by using the Preferential algorithm. This is applicable for both direct and indirect discrimination. The cleaning training data sets in such a way that direct and/or indirect discriminatory decision rules are converted to nondiscriminatory classification rules is discussed.

Krishna Kumar Tripathi [3] proposed is the web-based framework for discrimination prevention with classification and privacy preservation. Here registered user can upload the training dataset, perform rule generalization (RG) and rule protection, perform data discretization, data preprocessing and classification.

Sara Hajian et. al. [4] examined how discrimination could impact on cyber security applications, especially IDSs. IDSs use computational intelligence technologies such as data mining. It is obvious that the training data of these systems could be discriminatory, which would cause them to make discriminatory decisions when predicting intrusion or, more generally, crime. In proposed work producing training data which are free or nearly free from discrimination while preserving their usefulness to detect real intrusion or crime. In order to control discrimination in a dataset, a first step consists in 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.

Nickesh Rochlani et. al. [5] develop a novel pre-processing discrimination prevention methodology including different data transformation methods that can prevent direct discrimination, indirect discrimination or both of them at the same time. To attain this objective, the first step is to measure discrimination and identify categories and groups of individuals that have been directly and/or indirectly discriminated in the decision making processes; the second step is to transform data in the proper way to remove all those discriminatory biases. Finally, discrimination-free data models can be produced from the transformed data set without seriously damaging data quality.

### 3. BASICS OF DISCRIMINATION

Discrimination can be classified into direct and indirect (also called systematic). Direct discrimination consists of rules or procedures that provide internally mention minority or disadvantaged groups based on sensitive discriminatory data sets related to group membership. Indirect discrimination existing of rules or procedures that provide while not internally view the discriminatory datasets. Which is intentionally or un-intentionally could deriving discriminatory decisions? Financial institutions are redlining (refusing to grant mortgages or insurances in urban areas they consider as deteriorating) is an example of indirect discrimination, although certainly not the only one. The sake of compactness for a slight abuse of language. [6]

### 4. DISCRIMINATION PREVENTION AND PRIVACY PRESERVATION

Firstly the agent has been authorized service provider or a person. The agent requests for the database as per his need from the distributor, which is discrimination free and privacy preserved. The resultant database is received by the agent in text format. Then he has to retrieve the original database. The distributor collects all the databases as far as possible.[7]

When an agent's request for a database is received, the distributor searches for the database in its collection. After finding out the required database, the distributor performs the necessary processes to remove the discrimination and performs privacy preserving based on the level of user. The resultant database is send to the agent in the text format due to security reasons. The process for discrimination prevention consists of two phases such as discrimination measurement and data transformation. On this discrimination free database, the slicing algorithm is used for privacy preservation.

### 5. DISCRIMINATION MEASUREMENT

Direct and indirect discrimination discovery includes identifying  $\alpha$ -discriminatory rules and red lining rules. Based on the predetermined discriminatory items in database, frequent classification rules are classified in to 2 groups such as Potentially Discriminatory classification rules (PD) and Potentially Non-Discriminatory (PND).If we have a set of classification rules and discriminatory items in database, then a classification rule is said to be PD, when it contains a nonempty discriminatory item set and a non-discriminatory item set. A classification rule is PND, when an item set is nondiscriminatory. A PND rule could lead to discriminatory decisions in combination with some background knowledge. Direct discrimination is measured by identifying  $\alpha$ -discriminatory rules among the PD rules using a direct discrimination measure (elift) and a discriminatory threshold ( $\alpha$ ). The purpose of direct discrimination discovery is to identify  $\alpha$ -discriminatory rules. They indicate

biased rules that are directly inferred from discriminatory items and are called as  $\alpha$ -discriminatory direct rules. Indirect discrimination is measured by identifying redlining rules among the PND rules combined with background knowledge; using an indirect discriminatory measure (elb) and discriminatory threshold ( $\alpha$ ).The purpose of indirect discrimination discovery is to identify redlining rules. Redlining rules indicate biased rules that are directly inferred from nondiscriminatory items because of their correlation with discriminatory ones.[7]

The data transformation transform the original database in such a way to remove direct and/ or indirect discriminatory biases, with minimum impact on the data and on legitimate decision rules, so that no unfair decision rule can be mined from transformed data. [7]

### 6. CONCLUSIONS

Along with privacy, discrimination is a very important problem when considering the legal and ethical aspects of data mining. It is more than obvious that most people do not want to be discriminated because of their gender, religion, nationality, age and so on, Those attributes are used for making decisions about them like giving the m a job, loan, insurance, etc. In this paper going to study the privacy discrimination methods in data mining and also focuses on discrimination measurement and prevention in data mining.



### REFERENCES

- [1] Rupanjali T. Dive, , Anagha P. Khedkar, "An Approach for Discrimination Prevention in Data Mining", International Journal of Application or Innovation in Engineering & Management (IJAIEEM), Volume 3, Issue 6, June 2014.
- [2] Miss. Melanie Ann Thomas, Mrs. Joshila Grace, "Prevention of Discrimination in Data Mining", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727Volume 16, Issue 1, Ver. VII (Feb. 2014), PP 99-101.
- [3] Krishna Kumar Tripathi, "Discrimination Prevention with Classification and Privacy Preservation in Data mining", 7th International Conference on Communication, Computing and Virtualization, Procedia Computer Science 79 ( 2016 ) 244 – 253, 2016.
- [4] Sara Hajian, Josep Domingo-Ferrer and Antoni Martinez-Balleste, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection", IEEE, 978-1-4244-9906-9/11/\$26.00 ©2011.
- [5] Nickesh Rochlani, Prof. A.D.Chokhat, "Direct and Indirect Discrimination Prevention in Data Mining", International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 4, April 2015, pg.823 – 828.

[6] K. Madhavi, , Krishna Naik Mudavath, "Prevention Methods for Discrimination in Data Mining ", International Journal of Science and Research (IJSR), Volume 3 Issue 11, November 2014.

[7] Anjali P S, , Renji S, "Discrimination Prevention in Data mining with Privacy Preservation", International Journal of Science and Research (IJSR), Volume 4 Issue 3, March 2015.

## BIOGRAPHIES

	<p>Miss Madhuri P. Gole completed Bachelor of Engineering in Computer Science and Engineering from Rajashri Shahu College of Engineering Buldhana, and pursuing Master of Engineering in Computer Science and Information Technology from Sanmati Engineering College, Washim.</p>
	<p>Prof. Sachin Vyawahare is working as Asst. Professor and HOD of Computer Department at Sanmati Engineering College Washim (MH). He received BE degree and ME degree from S.G.B.A.U. Amaravati. His research interest includes networking, Operating System and Image Processing.</p>