Discrimination Prevention using Privacy Preserving Techniques

Asmita Kashid
Post Graduate Student
Department of Computer Engineering,
Maharashtra Institute of Technology,
Savitribai Phule Pune University,
Maharashtra, India

Vrushali Kulkarni
Head of Computer Engineering Department
Department of Computer Engineering,
Maharashtra Institute of Technology,
Savitribai Phule Pune University,
Maharashtra, India

Ruhi Patankar
Assistant Professor
Department of Computer Engineering,
Maharashtra Institute of Technology,
Savitribai Phule Pune University,
Maharashtra, India

ABSTRACT
Recently, it is observed that data mining technique may come across two problems- potential discrimination and potential privacy violation. Discrimination occurs as a result of use of discriminatory datasets for data mining tasks. Privacy violation occurs if a person's sensitive information is displayed to an unauthorized entity as a result of data mining tasks. Use of privacy preserving techniques to make data privacy protected can affect the amount of discrimination caused. It is important to study the relation of privacy and discrimination in the context of data mining. In this paper, we are trying to propose a method in which privacy preserving technique can be used to prevent discrimination and we can make the original data both privacy protected and discrimination-free.

General Terms
Data Mining

Keywords
Discrimination discovery, discrimination prevention, privacy preserving techniques

1. INTRODUCTION
Although, data mining is a technique to extract knowledge from a large volume of raw data, it may suffer from two major problems- potential discrimination and potential privacy violation. The meaning of discrimination is, without considering individual merit of a person, treating him/her unequally, only because he/she belongs to a minor community. E.g. an institute does not grant admission to foreign students. Discrimination can happen by directly mentioning the minority group or without directly mentioning the minority group. Data mining faces the problem of potential discrimination if it uses a dataset biased towards a particular group. As a result of this biased dataset, the results obtained from data mining tasks will also become biased. Discrimination is wrong because it is not a proper way to treat people. There are many anti-discriminatory laws exist, which state that people should not be discriminated according to their age, gender, race, nationality, etc. Discrimination Aware Data Mining (DADM) deals with finding methods to measure, discover or prevent discrimination using data mining techniques. The research in this area has been started from 2008.

Privacy means the right of a person to decide how to use his/her sensitive information. E.g. a person wants to hide sensitive information such as salary, job etc. Data mining faces the problem of potential privacy violation if some of the data mining tasks result in accessing sensitive information of a person. Main aim of Privacy Preserving Data Mining (PPDM) is to develop techniques for modifying the original data, so that the private data remain private even after the data mining process. Privacy preserving methods hide the sensitive information of user in such a way that the hidden information must be retrieved back for data analysis. This is a well explored area. Many privacy preserving techniques and models are already created by many researchers.

Privacy protection and anti-discrimination are dependent on each other. Hiding discriminatory attribute for privacy protection may affect the discrimination caused. E.g. in case of employee hiring, the results of hiring a candidate is different in the two cases: if the employer knows the religion of the candidate and if the employer does not know the religion of the candidate. So it is important to explore the relationship between discrimination prevention & privacy preservation. As privacy preservation is well explored area, many of the privacy preservation methods can be used for discrimination prevention.

Main aim of this paper is to propose architecture to develop a method to make the original data both privacy protected and discrimination-free. Rest of the paper is organized as follows: section 2 presents the overview of related work in both these fields. Section 3 presents architecture of our proposed work. Section 4 presents experiments, dataset used for the experiments and evaluation of results. Section 5 presents conclusions and future research directions.

2. RELATED WORK
Researchers are doing research in DADM field since the year 2008[1]. The research in this area has been started by the researcher D. Pedreschi. Methods are proposed to discover direct and indirect discrimination in [2]. Discrimination can be prevented in one of the three ways [3]: pre-processing, in-processing and post-processing. Pre-processing means the transformations are done on original discriminatory dataset to remove discrimination. In case of in-processing approach, standard data mining algorithms are changed in such a way that no discriminatory decisions are taken. Post-processing approach deals with changing the final results of data mining tasks to remove discrimination.

Research in DADM deals with developing different discrimination prevention methods using any of the above

three approaches. Different methods for discrimination prevention using preprocessing approach are shown in [3] [4] [5]. Modifications are done to standard decision tree techniques to prevent discrimination in [6]. Here both in-processing and post-processing discrimination prevention approaches are used. Standard Naïve Bayes algorithms are modified to prevent discrimination in [7]. It also uses both in-processing and post-processing approaches. Different metrics to measure amount of discrimination is given in [8].

Different tools are developed to discover discrimination. A reference model, named LP2DD, to detect discrimination in automatic Decision Support System is presented in [9]. The tool DCUBE has been generated to discover discrimination [10]. The improvements in DCUBE have been done by developing DCUBE-GUI tool [11]. Further enhancements are done in [12]. The case studies to discover and prevent discrimination using the existing DADM methods are presented in [13] [14]. There are some other issues related to DADM such as imbalanced datasets [15] and conditional discrimination [16]. DADM is done using regression methods in [17].

Some research also exists to identify relation between PPDM and DADM. [18] specifies the survey of different privacy preserving technique. The effect of data anonymization techniques (e.g. generalization and suppression) on anti-discrimination is given in [19]. The method to make data discrimination free using privacy preserving model (e.g. t-closeness) is shown in [20]. The impact of knowledge publishing on anti-discrimination is shown in [21] [22]. Privacy attack strategies are used to discover indirect discrimination in [23]. Slicing [24] is a data anonymization method, which provides privacy protection by making horizontal grouping of records and combining most co-related columns.

3. PROPOSED WORK

Our proposed work is to identify the relation between privacy preserving technique called slicing on anti-discrimination and developing a new technique to make data both discrimination-free and privacy-protected i.e. to develop a preprocessing technique which will handle both the problems faced by data mining- potential privacy violation and potential discrimination. Although some work is already done to identify effect of data anonymization on anti-discrimination [20] [21] [22], there are many other privacy preserving techniques such as permutation, perturbation, bucketization, slicing[24] which can be used for discrimination prevention. Slicing allows less information loss than that of generalization; it preserves both privacy and correlation in the original data. Because of this advantage, it is decided to use slicing in the proposed work. It may provide better results with anti-discrimination than that of generalization and suppression. As slicing preserves relation between most correlated attributes and breaks relationship between unrelated attributes, it may be possible to generate a method to handle conditional discrimination also.

Architecture of the proposed work is as depicted in Fig 1.

The hypothesis for the proposed work is - Slicing may impact the discrimination. So it is necessary to consider both privacy and discrimination while creating discrimination prevention method.

Inputs to our proposed system are mentioned below-

Sensitive attribute (s): attributes in the input dataset which contain sensitive information (e.g. salary).

Quasi-identifier attributes (QI): set of attributes in the input dataset which can be used to re-identify an individual (e.g. age, gender which can be used to re-identify a person's salary).

Discriminatory attribute (DA): set of attributes in the input dataset which are specified as discriminatory by law (e.g. gender, race etc.)

Discrimination threshold (α): fixed threshold which states an acceptable level of discrimination according to laws and regulations.

The modules in the proposed system are as follows:

1) Module 1: It identifies discrimination in the given discriminatory dataset. Inputs to this module are discrimination threshold, discriminatory dataset and discriminatory attributes. This module calculates amount of discrimination w.r.t. given discriminatory attributes.
2) Module 2: Slicing method is applied in this module. Slicing partitions the highly co-related attributes in a same column. The co-relation between attributes is found using mean-square contingency coefficient. Then tuples are partitions into buckets. Most correlated attributes are combined with sensitive attribute. Swap values among each bucket. Split and put swapped values in the original table.
3) Module 3: Effect of slicing is checked on anti-discrimination. For this, variation in number of α-discriminatory rules is calculated. If rules in transformed dataset are less than that of original dataset, then the method reduces discrimination, otherwise the method increases discrimination.
4) Module 4: Depending on output of Module 3, the discrimination prevention technique is created.

With the proposed work we are planning to achieve the following objectives: 1) improving the existing approach 2) getting better results 3) incorporating conditional discrimination 4) creating a new method 5) comparing the proposed method with the existing methods. We are proposing to handle only direct discrimination.

4. EXPERIMENTS

4.1 Dataset

German Credit [25] dataset is used for the proposed work German Credit data set contains 1000 records, 13 categorical and 7 numeric attributes, with credit as a class label. This dataset is frequently used in anti-discrimination literature. This can be used combined for DADM and PPDM researches. We are using categorical attribute for proposed dissertation work. Prediction task associated with this dataset is to determine whether a person is granted a credit (good) or denied a credit (bad). Foreign Worker = Yes and Personal Status = Female and not single are considered as discriminatory attributes. Job can be considered as a sensitive attribute. Personal_status, Age, Foreign_worker, Property_magnitude, Own_Telephone can be considered as QIs.

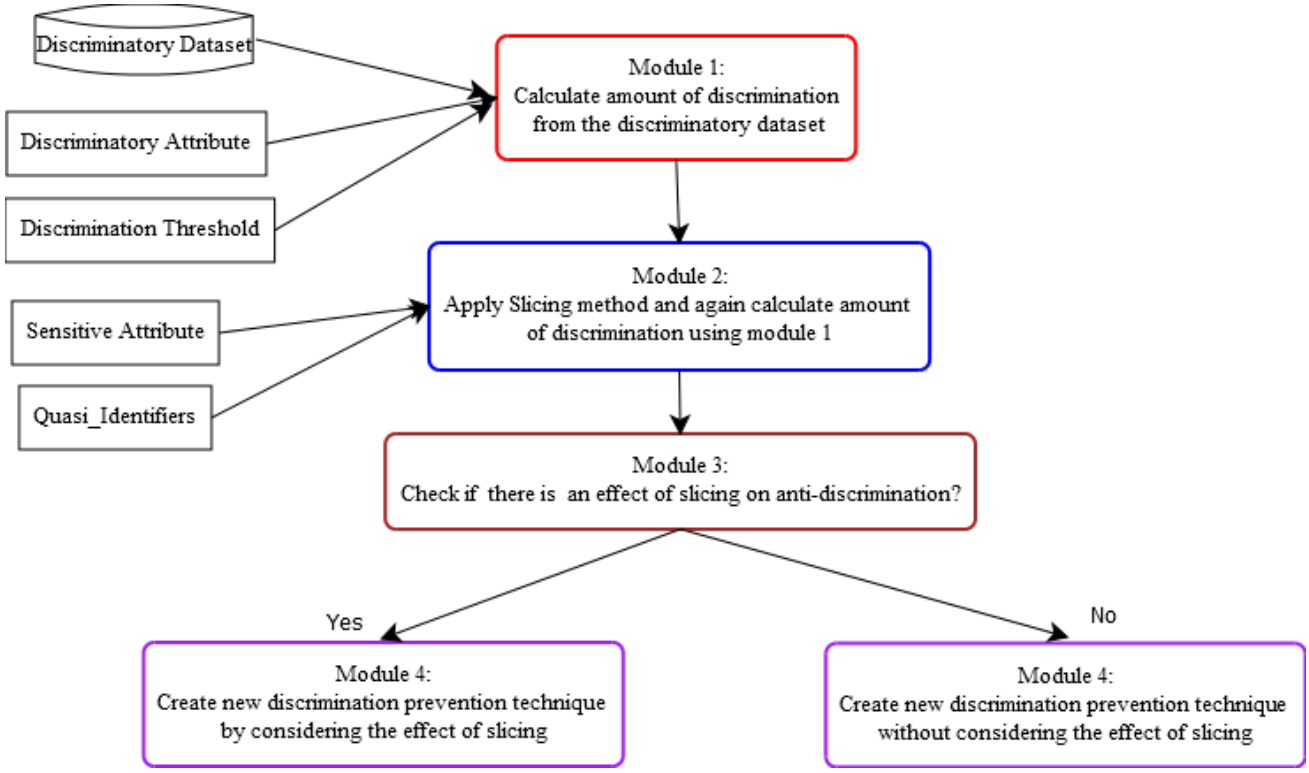

Fig 1: Architecture of the proposed system

4.2. Evaluation of Results

From partial implementation, it can be concluded that, the input dataset is discriminatory according to the given discriminatory attributes. Partial results of the system are α-discriminatory and α-protective rules. System identifies number of PD-classification rules for German credit dataset. E.g.[Purpose=Radio/TV,Foreign_Worker=Yes]➔[Class=Bad].Base rule for this rule is [Purpose=Radio/TV]➔[Class=Bad]

Classification rule is the association rule with right hand side equals to class item. PD-classification rule is a classification rule which contains Potentially Discriminatory (PD) item set [2]. PD item sets are specified by human rights laws. E.g. gender, age, marital_status etc. Base rule is a classification rule with all items same as PD-classification rule, except without PD item set. Each rule has confidence which states the probability of denying benefit. The confidence of the rule [Purpose=Radio/TV]➔[Class=Bad] (conf: 0.22143) states the probability of denying credit to the people with purpose radio/TV. Confidence of the rule [Purpose=Radio/TV, Foreign_Worker=Yes]➔{Class=Bad] (conf: 0.22545) states probability of denying credit to foreign workers who have asked credit for radio/TV.

Elift of PD-classification rule is the ratio of confidences of PD-classification rule and corresponding base rule [2]. Elift of the above PD-classification rule is:

Elift = conf (PD-classification rule) / conf (Base rule) = 1.01818

This elift states that being foreign worker increases the probability of denying credit w.r.t. all other people who have denied credit and who have asked credit for radio/TV. [Purpose=Radio/TV] is called context of the rule.

α is a fixed threshold, which states an acceptable level of discrimination according to rules and regulations[8]. The PD-classification rule is called α-protective if elift < α and if elift >= α, the rule is called α-discriminatory [8]. The above rule is α-discriminatory as elift > α. Given α = 1.

System discovers such kind of α-discriminatory rules in different contexts, so the german credit dataset is discriminatory w.r.t. Foreign_Worker.

Table 1 depicts values of number of α-discriminatory and α-protective rules for different values of α for german credit dataset. Graph in the Fig. 2 represents variations in the values of α-discriminatory and α-protective rules with different values of α. X-axis represents different values of α and Y-axis represents number of α-discriminatory and α-protective rules. It can be concluded from the graph that, as value of α increases, number of α-discriminatory rules decreases and number of α-protective rules increases. I.e. discrimination reduces as value of α increases. This behavior is quite similar to the behavior in the literature.

Table 2. German Credit Dataset: α-discriminatory and α-protective rules

Values of α	Number of α-discriminatory rules	Number of α-protective rules
1	216	25
1.1	39	202
1.2	15	226
1.25	3	238

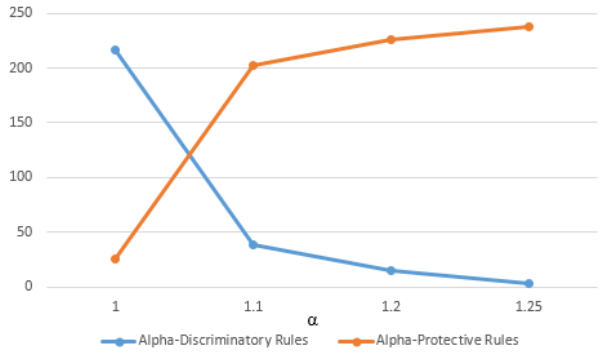

Fig 2: German Credit dataset: *α* versus *α*-discriminatory and *α*-protective rules

5. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

DADM and PPDM are related to each other. PPDM is a well explored area, so many privacy preserving techniques can be used to study impact of PPDM on discrimination prevention. This opens a large research avenue in the interdisciplinary approach of discrimination and other areas such as privacy protection. It is needed to study privacy literature in detail to identify different privacy preserving techniques to be incorporated with discrimination. Different data anonymization methods can have different impact on discrimination. Some techniques may increase discrimination, some may decrease it or some may not have any effect on discrimination. It is possible to analyze impact of other data anonymization techniques than this paper.

The proposed system will be useful in discrimination prevention data mining. The domains where this system is applicable are: credit and insurance, public health agency, health care, medical institutions, financial organization. The proposed system is very useful to make data discrimination free and privacy protected.

Future research directions include studying the relation between privacy and anti-discrimination by using methods other than used in existing literature. Privacy preserving techniques can be used to handle only bad/explainable discrimination. Existing discrimination discovery and prevention techniques can be used to solve real life discrimination problem. E.g. identifying discrimination caused at an institute while granting admission. There is still a scope to generate new discrimination prevention techniques and new discrimination discovery measures. It is also possible to modify discrimination prevention techniques from the existing literature.

It is a good research avenue to identify severity of discrimination and study discrimination in terms of cause and effect.

6. 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.

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

[3] S. Hajian & J. Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination prevention in data mining," IEEE transaction on knowledge & data engg. pp. 1445-1459, July 2013.

[4] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," Intl' Journal of Knowledge & Information Systems, Springer, Vol.33, no.1, pp. 1-33, 2012.

[5] F. Kamiran, T. Calders, "Classifying without discriminating," In: Proceedings of IEEE IC4 international conference on computer, Control & Communication, pp. 1-6, 2009.

[6] 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.

[7] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.

[8] 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.

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

[10] 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.

[11] B.Gao and B. Berendt, "Visual Data Mining for higher-level patterns: discrimination aware data mining and beyond," in Proc. 20th Benelearn, http://www.benelearn2011.org/, 2011.

[12] B.Berendt and S.Preibusch, "Exploring Discrimination: A user-centric evaluation of discrimination-aware data mining," in Proc. IEEE 12th Int'l Conf. on Data Mining Workshops (ICDMW), pp. 344-351, 2012.

[13] S. Hajian, J. Donimgo-Ferrer and 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, 2011.

[14] A. Romei, S.Ruggieri and F.Turini, "Discovering gender discrimination in project funding," ICDM Workshops, pp. 394-401, 2012.

[15] G. Ristanoski, W. Liu, J. Bailey, "Discrimination Aware Classification for Imbalanced Dataset," Proc. 22nd ACM Int'l conference on information & knowledge management, pp. 1529-1532, 2013.

[16] I. Zliobaite, F. Kamiran, T. Calders, "Handling Conditional Discrimination," 11th IEEE International Conference on Data Mining, pp. 992-1001, 2011.

[17] T.Calders, A. Karim, F. Kamiran, W.Ali, and X. Zhang, "Controlling Attribute Effect in Linear Regression," Proc. IEEE 13th Int'l Conf. on Data Mining (ICDM), pp. 71-80, 2013.

[18] B.C.M Fung, K. Wang, R. Chen, P.S.Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.* 42(4), Article 14, 2010.

[19] S Hajian and J. Domingo-Ferrer, "A Study on the Impact of Data Anonymization on Anti-Discrimination," Proc. I.EEE 12th International Conference on Data Mining Workshops, pp. 352-359, 2012.

[20] S.Ruggieri, "Data Anonymity Meets Non-Discrimination," IEEE 13th International Conference on Data Mining Workshops (ICDMW), pp. 875-882, 2013.

[21] S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer and F. Ginnotti, "Injecting Discrimination and Privacy Awareness into Pattern Discovery," Proc. IEEE 12th International Conference on Data Mining Workshops, pp. 360-369, 2012.

[22] S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer and F. Ginnotti, "Fair Pattern Discovery," Proc. 29th Annual ACM Symposium on Applied Computing, pp. 113-120, 2014.

[23] S.Ruggieri, S. Hajian, F. Kamiran, and X. Zhang, "Anti-discrimination Analysis using Privacy Attack Strategies," Machine Learning and Knowledge Discovery in databases, Springer, pp.694-710, 2014.

[24] Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy, "Slicing: A new approach for privacy preserving data publishing", IEEE transactions on knowledge and data engineering, Vol. 24, no.3. 2012.

[25] D.J. Newman, S.Hettich, C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases," http://archive.ics.uci.edu/ml, 1998.