A System to Analyze an Effect of Privacy Protection on Direct Discrimination

Asmita Gorave
Assistant Professor
Department of Computer Engineering
MIT College of Engineering, Pune

Vrushali Kulkarni
Head of Department
Department of Computer Engineering
Maharashtra Institute of Tech, Pune

ABSTRACT
Recently, it is observed that data mining technique comes across two major potential risks from social perspective: discrimination and privacy violation. Discrimination means treating people unfairly, just because they belong to minority group, without taking into account their individual qualification. Data mining technique undergoes risk of discrimination, if data mining tasks are performed using discriminatory dataset. Discrimination Prevention Data Mining is an area, which deals with discovering, preventing and measuring discrimination. Privacy provides right to a person to decide whether to disclose or not to disclose his/her sensitive information. Privacy violation occurs if a person's sensitive information is disclosed as a result of data mining tasks. Privacy Preserving Data Publishing is an area, which provides methods for publishing useful information while preserving data privacy. Recently, it is identified that these two areas are dependent on each other. So it is important to bridge the research gap between these areas. In this paper, our implemented system is described, which is useful to analyze effect of privacy protection methods on discrimination. Results of our system provide effect of different privacy protection methods on direct discrimination.

General Terms
Data Mining

Keywords
discrimination prevention, privacy protection, discrimination discovery, data anonymization methods

1. INTRODUCTION
Data mining is a technique to extract knowledge from raw data. Recently, it is observed that data mining technique comes across two major potential risks from social perspective: discrimination and privacy violation.

Unequal treatment given to people belonging to minority group, without taking into consideration their individual qualification, is called discrimination. E.g. denial of loan to a person, because the person belongs to a minority group. Data mining techniques extracts knowledge from raw data in terms of classification or association rules. If the rules are learnt from the training dataset which is discriminatory towards particular community, then learnt rules/decisions become discriminatory. This puts data mining at the risk of discrimination. To avoid this risk, a research is going on in an area, called Discrimination Prevention Data Mining (DPDM) and it deals with discovering, preventing and measuring discrimination. Discrimination can happen directly by mentioning discriminatory attributes, e.g. gender, age, color, religion, and ethnicity etc., specified by human rights laws. Discrimination can happen indirectly without mentioning discriminatory attributes. Discrimination prevention can be performed in three ways:

- Pre-processing: It involves transforming the original discriminatory dataset, such that discriminatory decisions are not made.
- In-processing: It involves changing standard data mining algorithms, in order to remove discrimination.
- Post-processing: It deals with removing discrimination from the final results of data mining tasks.

Privacy provides right to a person to decide whether to disclose or not to disclose his/her sensitive information e.g. a person may not want to disclose her/his disease. Privacy violation occurs if a person's sensitive information is exposed as a side effect of data mining tasks. One way to avoid privacy violation is to remove explicit identifier (e.g. name) of a person while publishing person specific data. Even if such explicit identifier is removed, there are other attributes, called Quasi Identifiers (QIs), which can be identified from external sources and combined to identify the person and his/her sensitive information. To avoid this risk, the research is going on in an area called, Privacy Preserving Data Publishing (PPDP) and it deals with developing techniques to modify the original data in some way, so that private data remain private even after data mining process. Anonymous version of QIs are created, so that even if the attacker identifies QIs, the sensitive information of a person cannot be exposed. PPDP deals with privacy attacks, privacy models and anonymization techniques.

Recently, it is observed that DPDM and PPDP are dependent on each other, as they have common methodological problems to be solved and they have common challenges. It is important to provide simultaneous protection against both these risks. Even though these areas are dependent on each other and have some commonalities, there is a significant gap between researches going on in these two areas. So it is an important research avenue, to bridge the gap between these two areas. Our system is an effort towards bridging the gap between these two areas by analyzing effect of privacy protection i.e. data anonymization methods on discrimination.

Main aim of this paper is to describe the system to analyze the effect of privacy protection on direct discrimination. The rest of the paper is organized as follows: section 2 shows the literature survey related to these two fields under the heading related work. Section 3 defines basic terminology in DPDM and PPDP. Section 4 presents problem statement and architecture of the system. Section 5 discusses results of performed experiments on different datasets. Section 6 presents conclusions and future work.

2. RELATED WORK

The research in DPDM area is started in 2008 [1]. Discrimination discovery method is explained in [2]. Three different approaches for discrimination prevention are specified in [3]: pre-processing, in-processing and post-processing. Research in DPDM deals with developing different discrimination prevention methods using one of the above three approaches. Discrimination prevention using pre-processing approach is described in [3] [4]. Methods for discrimination prevention using decision tree technique are described in [5]. Decision tree methods consist of both in-processing and post-processing approaches. Naïve Bayes model is used for discrimination prevention in [6]. This model uses both in-processing and post- processing approaches. Different metrics to measure amount of discrimination are specified in [7].

Research of PPDP is started in 2000 [8]. Many algorithms and techniques have been developed to preserve user's privacy. Data anonymization method, called Permutation [9] permutes values of QIs in order to break the relationship between QIs and sensitive attributes. Bucketization [10] separates QIs and sensitive attributes, makes horizontal group of tuples and then permutes values of sensitive attribute within the horizontal group in order to break the relation between QIs and sensitive attributes. Slicing [10] combines the most co-related QI with sensitive attribute, makes horizontal group of tuples. Then swaps values of QIs, in order to preserve relation between the most co-related attributes and to break relation between uncorrelated attributes. Generalization [11] replaces the values of QIs with generalized values using generalization taxonomy tree of QIs. Suppression [11] suppresses some/all of the values of the QIs. [12] specifies survey of different data anonymization methods, privacy models and privacy attacks.

Research is going on to identify relationship between PPDP and DPDM. Study of impact of data anonymization methods (e.g. generalization and suppression) on anti-discrimination is specified in [11]. The method to make data discrimination free using privacy preserving model (e.g. t-closeness) is depicted in [13]. The effect of knowledge publishing on antidiscrimination is shown in [14] [15].

3. BASIC TERMINOLOGY

3.1 Basic Terminology in DPDM

Terminology related to rule-based discovery and discrimination prevention [2] [3] are mentioned below. This is useful to understand the developed system.

- A data item is said to be Potentially Discriminatory (PD) if it is decided as discriminatory according to laws and regulations.
- A classification rule A, B → C is potentially discriminatory (PD) when A is a discriminatory item set and B is a non-discriminatory item set.
- elift is the metric to measure discrimination, which states the ratio of confidence of two rules, with and without the PD item.
- Discrimination Threshold (α), is a fixed threshold stating an acceptable level of discrimination according to laws and regulations [16].
- A PD classification rule c = A, B → C is α-protective w.r.t. elift, if elift < α. Otherwise, c is α-discriminatory.

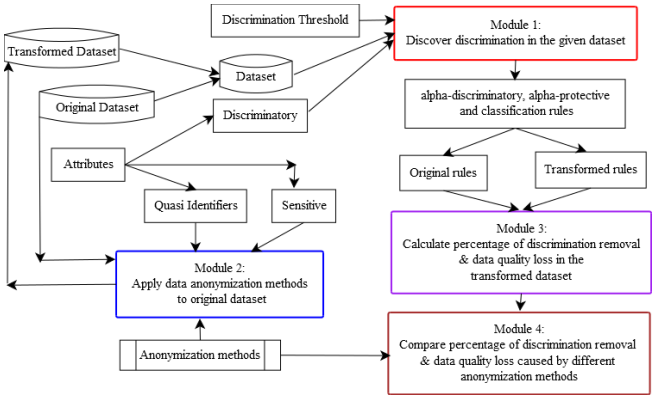

Fig 1: High-level architecture of the system

3.2. Basic Terminology in PPDP

Basic terminology in PPDP [12] are described below. This is useful to understand the developed system.

- Explicit identifier is a set of attributes that explicitly/uniquely identifies record owners.
- Quasi Identifier is a set of attributes that, in combination can be linked with external information to re-identify the record owner to whom the information refers.
- Sensitive attributes contain sensitive person specific information such as disease, salary or disability status.
- Non-Sensitive attributes contain all the attributes which do not belong to other three categories.
- Data Anonymization is an approach of PPDP that hides the identity and/or sensitive data of record owners, assuming sensitive data must be retained for data analysis.

4. PROBLEM DEFINITION AND ARCHITECTURE

The problem statement for our work is, to analyze effect of different privacy protection (data anonymization) techniques on direct discrimination. The aim is to compare the percentage of discrimination removal and data quality loss by different data anonymization methods. The basic idea is to measure amount of discrimination in the original input dataset, apply anonymization methods, again measure amount of discrimination from the transformed dataset. Then calculate percentage of discrimination removal and percentage of data quality loss in the transformed dataset. The architecture of the system is as shown in Fig. 1. System consists of four modules:

- Module 1: Amount of discrimination is calculated in terms of α-discriminatory and α-protective rules. Inputs are discriminatory dataset, discriminatory attributes and discrimination threshold (α). Output is amount of discrimination in the dataset.
- Module 2: Different anonymization methods are applied in this module and dataset gets transformed. Anonymization methods are generalization and suppression [11], permutation [9], bucketization [10] and slicing [10].
- Module 3: Percentage of discrimination removal and percentage of data quality loss is calculated using measures DDPD, DDPP and GC, MC.
- Module 4: Different data anonymization methods are compared in this module.

5. RESULTS

In [11], impact of generalization and suppression methods is checked on discrimination using a small table and single PD classification rule. Our developed system is validated using same table.

5.1 Datasets

Two data sets are used for performing experiments using the system: Adult [17] and German Credit [18]. They can be used for combined research in DPDM and PPDP. Their brief information is given below:

Adult dataset: This dataset consists of 48,842 records. The data set has 14 attributes (without class attribute). Prediction task associated with this data set is to determine whether a person makes more than 50K$ a year based on census and demographic information about people. Sex= Female and Age = Young are considered as discriminatory attributes. Occupation can be considered as a sensitive attribute. Age, Workclass, Marital status, race, sex can be considered as QIs.

German Credit dataset: This dataset consists of 1000 records. It has 20 attributes (without class attribute). 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.

5.2 Utility Measures

Results are evaluated (i.e. impact of data anonymization methods on direct discrimination) based on two aspects: direct discrimination removal and data quality loss. The method which provides more discrimination removal and less data quality loss is a better method. To measure discrimination removal, two metrics are used [3]:

- Direct Discrimination Prevention Degree (DDPD): It is the percentage of α-discriminatory rules that are no longer α-discriminatory in the transformed dataset.
- Direct Discrimination Protection Preservation (DDPP): It is the percentage of α-protective rules that remain α-protective in the transformed dataset.

To measure data quality loss, two metrics are used [3]:

- Ghost Cost (GC): It is the percentage of the rules those are extractable from the transformed data set, but those were not extractable from the original data set.
- Misses Cost (MC): It is the percentage of rules those are extractable from the original data set that cannot be extracted from the transformed data set.

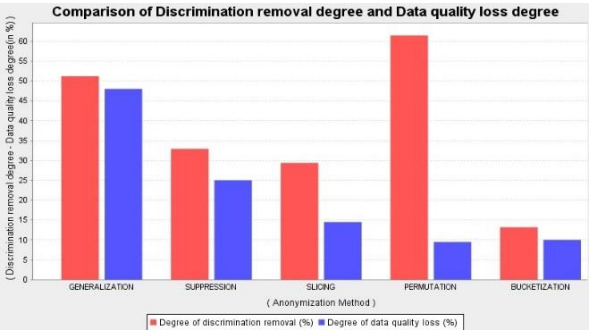

Fig 2: Adult Dataset: Discrimination removal degree & data quality loss degree vs anonymization methods (α=1 & DA=QI)

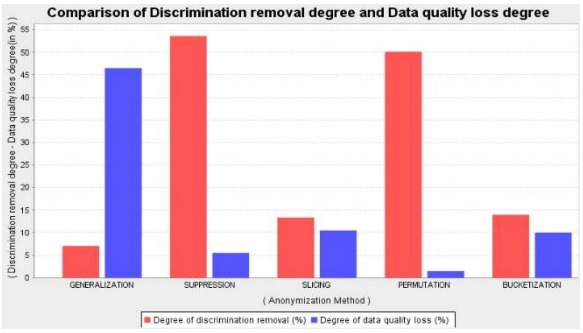

Fig 3: Adult Dataset: Discrimination removal degree & data quality loss degree vs anonymization methods (α=1 & DA!=QI)

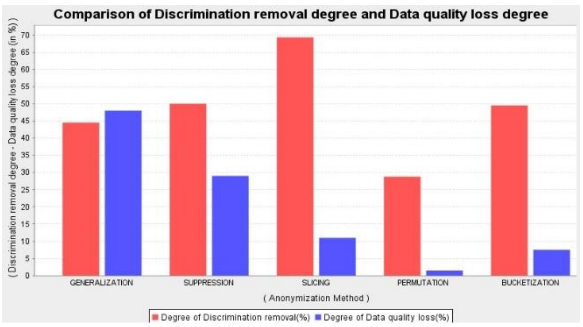

Fig 4: German Credit Dataset: Discrimination removal degree & data quality loss degree vs anonymization methods (α=1 & DA=QI)

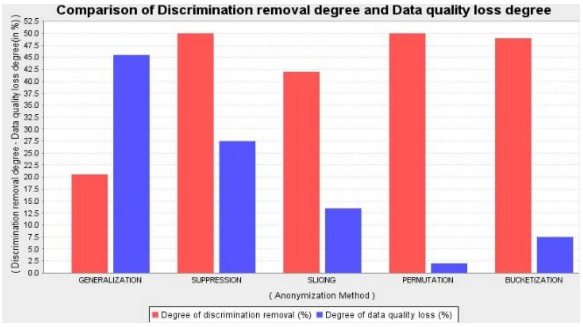

Fig 5: German Credit Dataset: Discrimination removal degree & data quality loss degree vs anonymization methods (α=1 & DA!=QI)

5.3 Evaluation of Results

After applying anonymization methods, either of the following cases occur:

- α-discriminatory rule becomes α-protective
- α-protective rule becomes α-discriminatory
- α-protective rule remains α-protective or α-discriminatory rule remains α-discriminatory
- Either α-discriminatory or α-protective rule no longer remain PD classification rule.
- Rules are deleted as discriminatory attribute records are suppressed
- Classification rules are selected as PD classification rules.

All these cases are handled by the discrimination removal and data quality loss measures.

The graph in Fig 2 specifies values of discrimination removal degree (as mean of DDPP and DDPD) and data quality loss

degree (as mean of MC and GC) for adult dataset, when selected DA and QI is same. The graph in Fig 3 specifies values of discrimination removal degree and data quality loss degree for adult dataset, when selected DA and QI is different. For Adult dataset, if DA and QI is same (Fig 2), then generalization and permutation are better in discrimination removal than that of suppression, slicing and bucketization. However, generalization shows more data quality loss. Permutation is a better method with more discrimination removal and less data quality loss. Suppression and slicing provides moderate discrimination removal and moderate data quality loss. In bucketization, less discrimination removal and less data quality loss occurs. Generalization performs lowest discrimination removal, if selected discriminatory attribute and QI are different (Fig 3).

Graph in Fig 4 specifies values of discrimination removal degree (as mean of DDPP and DDPD) and data quality loss degree (as mean of MC and GC) for German credit dataset, when selected DA and QI is same. Graph in Fig 5 specifies values of discrimination removal degree and data quality loss degree for German credit dataset, when selected DA and QI is different. For German Credit dataset, if DA and QI is same (Fig 4), then slicing and suppression provides more discrimination removal. However, suppression provides more data quality loss. Slicing is a better method with high discrimination removal and less data quality loss. Generalization performs lowest discrimination removal, if selected DA and QI are different (Fig 5).

X-axis of all these graphs represents different data anonymization methods used in the experiments viz. generalization, suppression, slicing, permutation and bucketization. Y-axis of the graphs shows degree of discrimination removal (as average of DDPD and DDPP) and degree of Data quality loss (as average of GC and MC). It can be concluded from the graphs that effect of anonymization methods depend on dataset and discriminatory behavior of data.

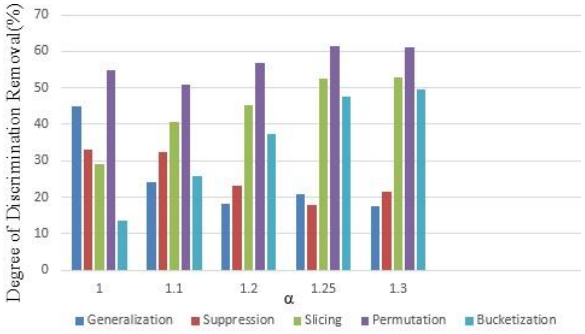

Fig 6: Adult Dataset: Degree of discrimination removal for anonymization methods vs values of α

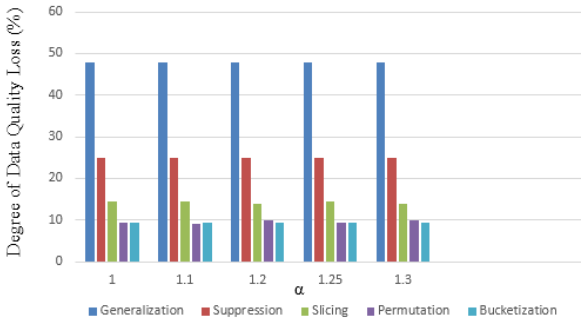

Fig 7: Adult Dataset: Degree of data quality loss for anonymization methods vs values of α

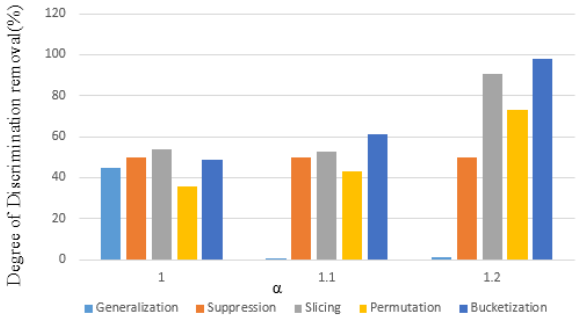

Fig 8: German Credit Dataset: Degree of discrimination removal for anonymization methods vs values of α

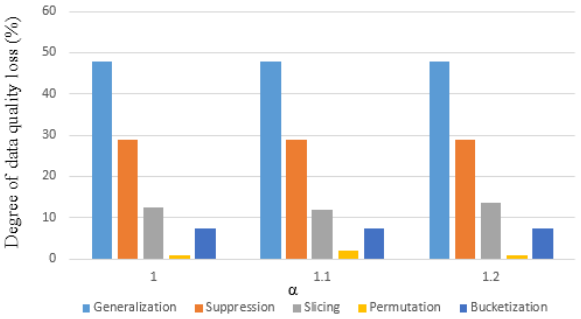

Fig 9: German Credit Dataset: Degree of data quality loss for anonymization methods vs values of α

Fig 6 shows variation in the values of degree of discrimination removal with different values of α for Adult dataset. X-axis represents different values of α and y-axis represents degree of discrimination removal. It is clear from the graph that, increase in value of α has different effect on anonymization methods. Fig 7 shows variation in the values of degree of data quality loss with different values of α for Adult dataset. Fig 8 and Fig 9 show variation in values of degree of discrimination removal and degree of data quality loss resp. with different values of α for German Credit dataset. It is clear from graphs in Fig 6 to Fig 9 that, with increasing value of α, discrimination removal by permutation based methods increases and discrimination removal by generalization and suppression decreases. However, variation in value of α does not affect data quality loss. The reason is, α is a discrimination threshold, so it will not have effect on data quality loss.

In general, the anonymization methods used in the system can be categorized in three categories: permutation-based, generalization, suppression. By doing various experiments, we can infer from results that, permutation-based methods have varying/undefined effect on discrimination removal and data quality loss on small datasets (e.g. test). They have fairly consistent effect on large datasets (e.g. German credit and Adult). Behavior of Permutation-based methods depends on number of records in the dataset (as number of records increases, permutations are normalized and method becomes consistent).Suppression and generalization have consistent effect on discrimination (moderate discrimination removal). Permutation-based methods have low data quality loss than generalization and suppression. The reason is, in generalization actual values of attributes are changed and in suppression, values are suppressed. However, in permutation based methods, actual values are not changed, but only permuted. Discrimination removal and data quality loss by permutation and slicing depend upon which/or how QI is permuted/ which QI is sliced. Discrimination removal and

data quality loss by bucketization depends upon which SA is permuted.

6. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Data anonymization methods affect discriminatory biases in the data set. They may increase discrimination, may decrease discrimination or may not have any effect on discrimination. So it is important to find relation between privacy preservation and anti-discrimination. It is impossible to protect original data against privacy attacks without taking into account anti-discrimination requirement. The knowledge of this relationship, can help in making the original data protected against both privacy and discrimination risks.

Our system can work as a tool for analyzing effect of privacy preserving techniques on discrimination. Our system provides a proper methodology to analyze effect of privacy preserving techniques on discrimination. Number of experiments can be performed using the system by changing different parameters. Different cases can be evaluated using the system and it will be a promising step towards bridging the gap between DPDM and PPDP.

In future, the system can be extended to other data anonymization techniques in the privacy literature. The immediate next future research direction includes, fine tuning the methods to achieve privacy protection, discrimination removal and data quality loss. E.g. if any method is achieving privacy, but lacking in discrimination removal and data quality loss, then we can fine-tune the method to achieve nearly 100% discrimination removal and nearly 0% data quality loss. In this way, a method can be created which makes the data both privacy protected, discrimination free, with less data quality loss. Currently, we have reduced scope of our system to direct discrimination discovery, it can be extended to indirect and/or conditional discrimination discovery in future.

7. 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, 2013.

[4] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Int'l Journal of Knowledge and Information Systems, Springer*, Vol. 33, Issue 1, pp. 1-33, 2011.

[5] 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.

[6] 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.

[7] 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.

[8] R.Agrawal and R.Srikant, "Privacy-preserving Data Mining", *In Proc. Of the ACM SIGMOD*, pp. 439-450, 2000.

[9] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables", *in ICDE*, 2007.\

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] R. Kohavi and B.Becker, "UCI Repository of Machine Learning Databases," http://archive.ics.uci.edu/ml/ datasets/Adult,1996.

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