Discrimination Prevention in Data Mining for Intrusion and Crime Detection

Sara Hajian, Josep Domingo-Ferrer and Antoni Martínez-Ballesté
Universitat Rovira i Virgili
Dept. of Computer Engineering and Maths, UNESCO Chair in Data Privacy
Av. Països Catalans 26 - E-43007 Tarragona, Catalonia
Email {sara.hajian,josep.domingo,antoni.martinez}@urv.cat

***Abstract*—Automated data collection has fostered the use of data mining for intrusion and crime detection. Indeed, banks, large corporations, insurance companies, casinos, etc. are increasingly mining data about their customers or employees in view of detecting potential intrusion, fraud or even crime. Mining algorithms are trained from datasets which may be biased in what regards gender, race, religion or other attributes. Furthermore, mining is often outsourced or carried out in cooperation by several entities. For those reasons, discrimination concerns arise. Potential intrusion, fraud or crime should be inferred from objective misbehavior, rather than from sensitive attributes like gender, race or religion. This paper discusses how to clean training datasets and outsourced datasets in such a way that legitimate classification rules can still be extracted but discriminating rules based on sensitive attributes cannot.**

I. Introduction

Discrimination can be viewed as the act of unfairly treating people on the basis of their belonging to a specific group. For instance, individuals may be discriminated because of their race, ideology, gender, etc. In economics and social sciences, discrimination has been studied for over half a century. There are several decision-making tasks which lend themselves to discrimination, *e.g.* loan granting and staff selection. In the last decades, anti-discrimination laws have been adopted by many democratic governments. Some examples are the US Equal Pay Act [1], the UK Sex Discrimination Act [2], the UK Race Relations Act [3] and the EU Directive 2000/43/EC on Anti-discrimination [4].

Surprisingly, discrimination discovery in information processing did not receive much attention until 2008 [5], even if the use of information systems in decision making is widely deployed. Indeed, decision models are created from real data (training data) in order to facilitate decisions in a variety of environments, such as medicine, banking or network security. In these cases, if the training data are biased for or against a particular community (*e.g.* foreigners), the learned model may show unlawfully prejudiced behavior. Discovering such potential biases and sanitizing the training data without harming their decision-making utility is therefore highly desirable. Information technologies could play an important role in discrimination discovery and prevention (*i.e.* anti-discrimination [5], [6]). In this respect, several data mining techniques have been adapted with the purpose of detecting discriminatory decisions.

Anti-discrimination also plays an important role in cyber security where computational intelligence technologies such as data mining may be used for different decision making scenarios. To the best of our knowledge, this is the first work that considers anti-discrimination for cyber security. Clearly, here the challenge is to avoid discrimination while maintaining data usefulness for cyber security applications relying on data mining, *e.g.* intrusion detection systems (IDS) or crime predictors.

The main contributions of this paper are as follows: (1) introducing anti-discrimination in the context of cyber security; (2) proposing a new discrimination prevention method based on data transformation that can consider several discriminatory attributes and their combinations; (3) proposing some measures for evaluating the proposed method in terms of its success in discrimination prevention and its impact on data quality.

In this paper, Section II discusses related work; Section III introduces anti-discrimination for cyber security applications based on data mining; Section IV reviews discrimination discovery; Section V presents the method for discrimination prevention and its evaluation; a discussion is given in Section VI; conclusions are drawn in Section VII.

978-1-4244-9906-9/11/$26.00 ©2011 IEEE

TABLE I
TOY EXAMPLE: SUBSCRIBER INFORMATION COLLECTED BY A TELECOMMUNICATIONS OPERATOR

SubsNum	Gender	Race	Age	Zip	DownProf	P2P	PortScan	Intruder
1	Female	White	Young	43799	Low	Yes	Yes	NO
2	Male	Black	Young	43700	High	No	Yes	YES
2	Male	White	Aged	84341	Normal	Yes	No	NO
2	Male	Black	Young	72424	Low	No	Yes	YES
1	Female	White	Aged	43743	High	Yes	Yes	YES
3	Female	White	Young	43251	High	No	No	No

II. RELATED WORK

The existing literature on anti-discrimination in computer science mainly elaborates on data mining models and related techniques. Some proposals are oriented to the discovery and measure of discrimination. Others deal with the prevention of discrimination.

- *Discrimination discovery* is based on formalizing legal definitions of discrimination[1] and proposing quantitative measures for it. These measures were proposed by Pedreschi in 2008 [5], [7]. This approach has been extended to encompass statistical significance of the extracted patterns of discrimination in [8], and it has been implemented as reported in [9]. Data mining is a powerful aid for discrimination analysis, capable of discovering the patterns of discrimination that emerge from the data.
- *Discrimination prevention* consists of inducing patterns that do not lead to discriminatory decisions even if trained from a dataset containing them. Three approaches are conceivable: (i) adapting the preprocessing approaches of data transformation and hierarchy-based generalization from the privacy preservation literature [6], [10]; (ii) changing the data mining algorithms (in-processing) by integrating discrimination measure evaluations within them [11]; and (iii) post-processing the data mining model to reduce the possibility of discriminatory decisions [8]. Although some methods have been proposed, discrimination prevention stays a largely unexplored research avenue.

Clearly, a straightforward way to handle discrimination prevention would consist of removing discriminatory attributes from the dataset. However, as stated in [5], [6] there may be other attributes that are highly correlated with the sensitive one. Hence, one might decide to remove also those highly correlated attributes as well. Although this would solve the discrimination problem, in this process much useful information would be lost. Hence, another challenge regarding discrimination prevention is to find an optimal trade-off between anti-discrimination and usefulness of the training data.

[1]For instance, the U.S. Equal Pay Act [1] states that: "a selection rate for any race, sex, or ethnic group which is less than four-fifths of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact".

III. ANTI-DISCRIMINATION AND CYBER SECURITY

In this paper, we use as a running example the training dataset shown in Table I. It corresponds to the data collected by an Internet provider to detect subscribers possibly acting as intruders. The dataset consists of nine attributes, the last one (Intruder) being the class attribute. Each record corresponds to a subscriber of a telecommunication company determined by SubsNum attribute. Other than personal attributes (Gender, Age, Zip, Race), the dataset also includes the following attributes:

- *DownProf*: stands for downloading profile and measures the average quantity of data the subscriber downloads monthly. Its possible values are *High*, *Normal*, *Low*, *Very low*.
- *P2P*: indicates whether the subscriber makes use of peer-to-peer software, such as eMule.
- *PortScan*: indicates whether the subscriber makes use of a port scanning utility, such as Nmap.

Anti-discrimination techniques should be used in the above example. If the training data are biased towards a certain group of users (*e.g.* young people), the learned model will show discriminatory behavior towards that group and most requests from young people will be incorrectly classified as intruders.

Additionally, anti-discrimination techniques could also be useful in the context of data sharing between IDS. Assume that various IDS share their IDS reports (that contain intruder information) in order to improve their respective intruder detection models. Before an IDS shares its report, this report should be sanitized to avoid inducing biased discriminatory decisions in other IDS.

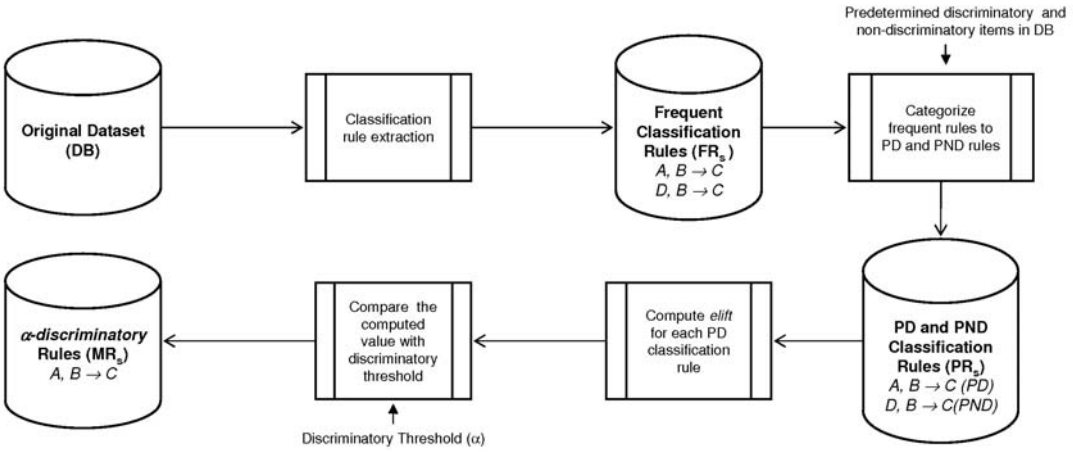

Fig. 1. Discrimination discovery process

IV. Discovering Discrimination

Discrimination discovery is about finding out discriminatory decisions hidden in a dataset of historical decision records. The basic problem in the analysis of discrimination, given a dataset of historical decision records, is to quantify the degree of discrimination suffered by a given group (*e.g.* an ethnic group) in a given context with respect to the classification decision (*e.g.* intruder yes or no). Figure 1 shows the process of discrimination discovery, based on approaches and measures described in this section.

A. Basic Definitions

- An *item* is an attribute along with its value, *e.g.*{Gender=Female}.
- *Association/classification rule mining* attempts, given a set of transactions (records), to predict the occurrence of an item based on the occurrences of other items in the transaction.
- An *itemset* is a collection of one or more items, *e.g.* {Gender=Male, Zip=54341}.
- A *classification rule* is an expression $X \rightarrow C$, where X is an itemset, containing no class items, and C is a class item, *e.g.* {Gender=Female, Zip=54341} $\rightarrow$ Intruder=YES. X is called the premise (or the body) of the rule.
- The *support* of an itemset, $supp(X)$, is the fraction of records that contain the itemset X. We say that a rule $X \rightarrow C$ is completely supported by a record if both X and C appear in the record.
- The *confidence* of a classification rule, $conf(X \rightarrow C)$, measures how often the class item C appears in records that contain X. Hence, if $supp(X) > 0$

$$conf(X \rightarrow C) = \frac{supp(X, C)}{supp(X)}$$

Support and confidence range over $[0, 1]$. In addition, the notation also extends to negated itemsets, *i.e.* $\neg X$.

- A *frequent classification rule* is a classification rule with a support or confidence greater than a specified lower bound. Let $\mathcal{DB}$ be a database of original data records and $\mathcal{FR}_s$ be the database of frequent classification rules.

B. Potentially Discriminatory and Non-Discriminatory Classification Rules

With the assumption that discriminatory items in $\mathcal{DB}$ are predetermined (*e.g.* Race=Black, Age = Young), rules fall into one of the following two classes with respect to discriminatory and non-discriminatory items in $\mathcal{DB}$.

1) A classification rule $X \rightarrow C$ is *potentially discriminatory* (PD) when $X = A, B$ with A a non-empty discriminatory itemset and B a non-discriminatory itemset. For example {Race=Black, Zip=43700} $\rightarrow$ Intruder=Yes.
2) A classification rule $X \rightarrow C$ is *potentially non-discriminatory* (PND) when X is a non-discriminatory itemset. For example {PortScan=Yes, Zip=43700} $\rightarrow$ Intruder=YES.

The word "potentially" means that a PD rule could probably lead to discriminatory decisions, so some measures are needed to quantify the discrimination potential.

Also, a PND rule could lead to discriminatory decisions if combined with some background knowledge, *e.g.* if in the above example one knows that zip 43700 is mostly inhabited by black people (indirect discrimination).

C. Discrimination Measures

Pedreschi *et al.*[5], [8] translated the qualitative statements in existing laws, regulations and legal cases into quantitative formal counterparts over classification rules and they introduced a family of measures of the degree of discrimination of a PD rule. In our contribution we use their *extended lift* measure ($elift$), which is recalled next.

Definition 1: Let $A, B \rightarrow C$ be a classification rule with $conf(B \rightarrow C) > 0$. The extended lift of the rule is

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)}$$

The idea here is to evaluate the discrimination of a rule by the gain of confidence due to the presence of the discriminatory items (*i.e.* A) in the premise of the rule. Indeed, $elift$ is defined as the ratio of the confidence of the two rules: *with* and *without* the discriminatory items. Whether the rule is to be considered discriminatory can be assessed by thresholding[2] $elift$ as follows [8].

Definition 2: Let $\alpha \in R$ be a fixed threshold. A PD classification rule $c = A, B \rightarrow C$ is α-protective w.r.t. $elift$ if $elift(c) < \alpha$. Otherwise, c is α-discriminatory.

Consider rule

$$c = \{\text{Race=Black}, \text{Zip=43700}\} \rightarrow \text{Intruder=YES}$$

from Table I. If $\alpha = 1.4$ and $elift(c) = 1.46$, rule c is 1.4-discriminatory.

In terms of indirect discrimination, the combination of PND rules with background knowledge probably could generate α-discriminatory rules. If a PND rule c with respect to background knowledge generates an α-discriminatory rule, c is an α-discriminatory PND rule and, if not, c is an α-protective PND rule. However, in our proposal we concentrate on direct discrimination and thus consider only α-discriminatory rules and assume that all the PND rules in $\mathcal{PR}_s$ are α-protective PND. According to Figure 1, let $\mathcal{MR}_s$ be the database of α-discriminatory rules extracted from database $\mathcal{DB}$.

[2]Note that α is a fixed threshold stating an acceptable level of discrimination according to laws and regulations.

V. A Proposal for Discrimination Prevention

In this section we present a new discrimination prevention method which follows the preprocessing approach mentioned in Section I above. The method transforms the source data by removing discriminatory biases so that no unfair decision rule can be mined from the transformed data. The proposed solution is based on the fact that the dataset of decision rules would be free of discriminatory accusation if for each α-discriminatory rule r' there would be at least one PND rule r leading to the same classification result as r'.

Our method makes use of the p-instance concept, formalized in [12] in the following way.

Definition 3: Let $p \in [0, 1]$. A classification rule $r' : A, B \rightarrow C$ is a p-instance of $r : D, B \rightarrow C$ if

1) $conf(r) \geq p \cdot conf(r')$ and
2) $conf(r'' : A, B \rightarrow D) \geq p$.

If each r' in $\mathcal{MR}_s$ was a p-instance (where p is 1 or a value near 1) of a PND rule r in $\mathcal{PR}_s$, the dataset of decision rules would be free of discriminatory accusation.

Consider rules r and r' extracted from the dataset in Table I:

$$r' = \{\text{Race=Black}, \text{Zip=43700}\} \rightarrow \text{Intruder=YES}$$

$$r = \{\text{PortScan=Yes}, \text{Zip=43700}\} \rightarrow \text{Intruder=YES}$$

With $p = 0.8$, rule r' is 0.8-instance of rule r if:

1) $conf(r) \geq 0.8 \cdot conf(r')$
2) $conf(r'') \geq 0.8$

where rule r'' is:

$$r'' = \{\text{Race=Black}, \text{Zip=43700}\} \rightarrow \text{PortScan=Yes}$$

Although r' is α-discriminatory based on the $elift$ measure, the existence of a PND rule r that leads to the same result as rule r' and satisfies both Conditions (1) and (2) of Definition 3 demonstrates that the subscriber is classified as intruder not because of race but because of using port scanning. Hence, rule r' is free of discriminatory accusation, because the IDS could argue that r' is an instance of a more general non-discriminatory rule r. Clearly, r is legitimate, because port scanning can be considered an unbiased indicator of a suspect intruder.

Our solution for discrimination prevention is based on the above idea. We transform data by removing all evidence of discrimination appeared in form of α-discriminatory rules. These α-discriminatory rules are

divided into two groups: α-discriminatory rules such that there is at least one PND rule leading to same result and α-discriminatory rules such that there is no such PND rule. For the first group a suitable data transformation with minimum information loss should be applied for ensuring Conditions (1) or (2) of Definition 3 in case they are not satisfied. For the second group, also a suitable data transformation with minimum information loss should be applied in such a way that those α-discriminatory rules are converted to α-protective rules based on the definition of the discriminatory measure (*i.e.* $elift$). The detailed process of our solution is described by means of the following phases:

- *Phase 1*. Use Pedreschi's measures on each rule to discover the patterns of discrimination emerged from the available data. Figure 1 details the steps of this phase.
- *Phase 2*. Based on Definition 3, find the relationship between α-discriminatory rules and PND rules extracted in the first phase and determine the transformation requirement for each rule.
- *Phase 3*. Transform the original data to provide the transformation requirement for each respective α-discriminatory rule without seriously affecting the data or other rules.
- *Phase 4*. Evaluate the transformed dataset with the discrimination prevention and information loss measures of Section V-B below, to check whether they are free of discrimination and useful enough.

The first phase, depicted in Figure 1, consists of the following steps. In the first step, frequent classification rules are extracted from $\mathcal{DB}$ by well-known frequent rule extraction algorithms such as Apriori [17]. In the second step, with respect to the predetermined discriminatory items in the dataset, the extracted rules are divided into two categories: PD and PND rules. In the third step, for each PD rule, the $elift$ measure is computed to determine the collection of α-discriminatory rules saved in $\mathcal{MR}_s$.

The second phase is summarized next. In the first step of this phase, for each α-discriminatory rule in $\mathcal{MR}_s$ of type $r' : A, B \rightarrow C$, a collection of PND rules in $\mathcal{PR}_s$ of type $r : D, B \rightarrow C$ is found. Call D_{pn} the set of these PND rules. Then the conditions of Definition 3, for a value of p at least 0.8, are checked for each rule in D_{pn}. Three cases arise depending on whether Conditions (1) and (2) hold:

1) There is at least one rule in D_{pn} such that both Conditions (1) and (2) of Definition 3 hold;
2) There is no rule in D_{pn} satisfying both Conditions (1) and (2) of Definition 3, but there is at least one rule satisfying one of those two conditions;
3) No rule in D_{pn} satisfies any of Conditions (1) or (2).

In the first case, it is obvious that currently there is at least one rule r in D_{pn} such that r' is p-instance of r for $p \geq 0.8$. In this case no transformation is required. In the second case, the PND rule r_b in D_{pn} should be selected which requires the minimum data transformation to fulfill both Conditions (1) and (2). A smaller difference between the values of the two sides of Conditions (1) or (2) for each r in D_{pn} indicates a smaller required data transformation. In this case, Conditions (1) and (2) in r_b determine the transformation requirement of r'. The third case happens when there is no r rule in D_{pn} satisfying any of Conditions (1) or (2). In this case, the transformation requirement of r' determines that this α-discriminatory rule should be converted to an α-protective rule based on the definition of the respective discriminatory measure (*i.e.* $elift$). The output of the second phase is a database $\mathcal{TR}_s$ with all $r' \in \mathcal{MR}_s$, their respective transformed rule r_b and their respective transformation requirements (see below).

The following list shows the first, second and third transformation requirements that can be generated for each $r' \in \mathcal{MR}_s$ according to the above cases:

1) $conf(r' : A, B \rightarrow C) \leq conf(r : D, B \rightarrow C)/p$
2) $conf(r" : A, B \rightarrow D) \geq p$
3) If $f() = elift$, $conf(r' : A, B \rightarrow C) < \alpha \cdot conf(B \rightarrow C)$

For the α-discriminatory rules with the first and second transformation requirements, it is possible that the cost of satisfying these requirements would be more than the cost of the third transformation requirement. In other words, satisfying the third transformation requirement could lead to a smaller data transformation than satisfying the first or second requirements. So for these rules the method should also do this comparison and select the transformation requirement with minimum cost. We consider all possible cases to achieve minimum data transformation.

Finally, we have a database of α-discriminatory rules with their respective transformation requirements. An appropriate data transformation method (Phase 3) should be run to satisfy these requirements with minimum degree of information loss and maximum degree of discrimination removal.

A. Data Transformation Method

As mentioned above, an appropriate data transformation method is required to modify original data in such a way that the transformation requirement for each α-discriminatory rule is satisfied without seriously affecting the data or the non α-discriminatory rules. Based on these objectives, the data transformation method should increase or decrease the confidence of the rules to the target values with minimum impact on data quality, that is, maximize the disclosure prevention measures and minimize the information loss measures of Section V-B below. It is worth mentioning that decreasing the confidence of special rules (sensitive rules) by data transformation was previously used for knowledge hiding [13], [14], [15] in privacy-preserving data mining (PPDM).

We assume that the class item C is a binary attribute. The details of our proposed data transformation method are summarized as follows:

1) For the α-discriminatory rules with the first transformation requirement (inequality $conf(A, B \rightarrow C) \leq conf(D, B \rightarrow C)/p$), the values of both sides of the inequality are independent, so the value of the left-hand side could be decreased without any impact on the value of the right-hand side. A possible solution for decreasing

$$conf(A, B \rightarrow C) = \frac{supp(A, B, C)}{supp(A, B)} \qquad (1)$$

to any target value is to perturb the class item from C to $\neg C$ in the subset $\mathcal{DB}_c$ of all records in the original dataset which completely support the rule $A, B \rightarrow C$ and have minimum impact on other rules to decrease the numerator of Expression (1) while keeping the denominator fixed. (Removing the records of the original dataset which completely support the rule $A, B \rightarrow C$ would not help because it would decrease both the numerator and the denominator of Expression (1).)

2) For the α-discriminatory rules with the second transformation requirement (inequality $conf(A, B \rightarrow D) \geq p$), the value of the right-hand side of the inequality is fixed so the value of the left-hand side could be increased independently. A possible solution for increasing

$$conf(A, B \rightarrow D) = \frac{supp(A, B, D)}{supp(A, B)} \qquad (2)$$

above p is to perturb item D from $\neg D$ to D in the subset $\mathcal{DB}_c$ of all records in the original dataset which completely support the rule $A, B \rightarrow \neg D$ and have minimum impact on other rules to increase the numerator of Expression (2) while keeping the denominator fixed.

3) For the α-discriminatory rules with the third transformation requirement (inequality $conf(A, B \rightarrow C) < \alpha \cdot conf(B \rightarrow C)$), unlike the above cases, both inequality sides are dependent; hence, a transformation is required that decreases the left-hand side of the inequality without any impact on the right-hand side. A possible solution for decreasing

$$conf(A, B \rightarrow C) = \frac{supp(A, B, C)}{supp(A, B)} \qquad (3)$$

is to perturb item A from $\neg A$ to A in the subset $\mathcal{DB}_c$ of all records of the original dataset which completely support the rule $\neg A, B \rightarrow \neg C$ and have minimum impact on other rules to increase the denominator of Expression (3) while keeping the numerator and $conf(B \rightarrow C)$ fixed. (Removing the records of the original dataset which completely support the rule $A, B \rightarrow C$ would not help because it would decrease both the numerator and the denominator of Expression (3) and also $conf(B \rightarrow C)$. Changing the class item C would not help either because it would impact on $conf(B \rightarrow C)$.)

Records in $\mathcal{DB}_c$ should be changed until the transformation requirement is met for each α-discriminatory rule. Among the records of $\mathcal{DB}_c$, one should change those with lowest impact on the other rules. Hence, for each record $db_c \in \mathcal{DB}_c$, the number of rules whose premise is supported by db_c is taken as the impact of db_c, that is $impact(db_c)$; the rationale is that changing db_c impacts on the confidence of those rules. Then the records db_c with minimum $impact(db_c)$ are selected for change, with the aim of scoring well in terms of the four utility measures proposed below. It means that transforming db_c with minimum $impact(db_c)$ could reduce the impact of this transformation on turning the α-protective rules to α-discriminatory rules and on generating the extractable rules from original dataset in the transformed dataset.

B. Utility measures

The proposed solution should be evaluated based on two aspects:

- The success of the proposed solution in removing all evidence of discrimination from the original dataset (degree of discrimination prevention).

- The impact of the proposed solution on data quality (degree of information loss).

A discrimination prevention method should provide a good trade-off between both aspects above. The following measures are proposed for evaluating our solution:

- *Discrimination Prevention Degree* (DPD). This measure quantifies the percentage of α-discriminatory rules that are no longer α-discriminatory in the transformed dataset.
- *Discrimination Protection Preservation* (DPP). This measure quantifies the percentage of the α-protective rules in the original dataset that remain α-protective rules in the transformed dataset (DPP may not be 100% as a side-effect of the transformation process).
- *Misses Cost* (MC). This measure quantifies the percentage of rules among those extractable from the original dataset that cannot be extracted from the transformed dataset (side-effect of the transformation process).
- *Ghost Cost* (GC). This measure quantifies the percentage of the rules among those extractable from the transformed dataset that could not be extracted from the original dataset (side-effect of the transformation process).

The DPD and DPP measures are used to evaluate the success of proposed method in discrimination prevention; ideally they should be 100%. The MC and GC measures are used for evaluating the degree of information loss (impact on data quality); ideally they should be 0%. MC and GC were previously proposed as information loss measures for knowledge hiding in PPDM [16].

VI. Discussion

Although there are some works about anti-discrimination in the literature, in this paper we introduced anti-discrimination for cyber security applications based on data mining. Pedreschi *et al.* in [5], [7], [8], [9], [12] concentrated on discrimination discovery, by considering each rule individually for measuring discrimination without considering other rules or the relation between them. However in this work, we also take into account the PND rules and their relation with α-discriminatory rules in discrimination discovery. Then we propose a new preprocessing discrimination prevention method. Kamiran *et al.* in [6], [10] also proposed a preprocessing discrimination prevention method. However, their works try to detect discrimination in the original data for only one discriminatory item based on a simple measure and then they transform data to remove discrimination. Their approach cannot guarantee that the transformed dataset is really discrimination-free, because it is known that discriminatory behaviors can often be hidden behind several items, and even behind combinations of them. Our discrimination prevention method takes into account several items and their combinations; moreover, we propose some measures to evaluate the transformed data in degree of discrimination and information loss.

VII. Conclusions

We have 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. Our contribution concentrates on 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. In the future, we want to run our method on real datasets, improve our methods and also consider background knowledge (indirect discrimination).

Disclaimer and Acknowledgments

The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS", CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES" and TSI-020100-2009-720 "everification", by the Government of Catalonia under grant 2009 SGR 01135, and by the European Commission under FP7 project "DwB". The second author is partly supported as an ICREA Acadèmia Researcher by the Government of Catalonia.

References

[1] United States Congress, *US Equal Pay Act*, 1963. http://archive.eeoc.gov/epa/anniversary/epa-40.html

[2] Parliament of the United Kingdom, *Sex Discrimination Act*, 1975. http://www.opsi.gov.uk/acts/acts1975/PDF/ukpga_19750065_en.pdf

[3] Parliament of the United Kingdom, *Race Relations Act*, 1976. http://www.statutelaw.gov.uk/content.aspx?activeTextDocId=2059995

[4] European Commission, *EU Directive 2000/43/EC on Anti-discrimination*, 2000. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:180:0022:0026:EN:PDF

[5] D. Pedreschi, S. Ruggieri and F. Turini, "Discrimination-aware data mining". *Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pp. 560-568. ACM, 2008.

[6] F. Kamiran and T. Calders, "Classification without discrimination". *Proc. of the 2nd IEEE International Conference on Computer, Control and Communication (IC4 2009)*. IEEE, 2009.

[7] S. Ruggieri, D. Pedreschi and F. Turini, "Data mining for discrimination discovery". *ACM Transactions on Knowledge Discovery from Data*, 4(2) Article 9, ACM, 2010.

[8] D. Pedreschi, S. Ruggieri and F. Turini, "Measuring discrimination in socially-sensitive decision records". *Proc. of the 9th SIAM Data Mining Conference (SDM 2009)*, pp. 581-592. SIAM, 2009.

[9] S. Ruggieri, D. Pedreschi and F. Turini, "DCUBE: Discrimination Discovery in Databases". *Proc. of the ACM International Conference on Management of Data (SIGMOD 2010)*, pp. 1127-1130. ACM, 2010.

[10] F. Kamiran and T. Calders, "Classification with No Discrimination by Preferential Sampling". *Proc. of the 19th Machine Learning conference of Belgium and The Netherlands*, 2010.

[11] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification", *Data Mining and Knowledge Discovery*, 21(2):277-292. 2010

[12] D. Pedreschi, S. Ruggieri and F. Turini, "Integrating induction and deduction for finding evidence of discrimination". *Proc. of the 12th ACM International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pp. 157-166. ACM, 2009.

[13] V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, "Association rule hiding". *IEEE Trans. on Knowledge and Data Engineering*,16(4):434-447, 2004.

[14] Y. Saygin, V. Verykios and C. Clifton, "Using unknowns to prevent discovery of association rules". *ACM SIGMOD Record*, 30(4):45-54, 2001.

[15] J. Natwichai, M. E. Orlowska and X. Sun, "Hiding sensitive associative classification rule by data reduction". *Advanced Data Mining and Applications (ADMA 2007)*, LNCS 4632, pp: 310-322. 2007.

[16] S. R. M. Oliveira and O. R. Zaiane. "A unified framework for protecting sensitive association rules in business collaboration". *International Journal of Business Intelligence and Data Mining*, 1(3):247287, 2006.

[17] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases". *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499. VLDB, 1994.