A proposal for a new definition of discrimination for personalization algorithms

Duguépéroux Joris supervised by Gambs Sébastien and Tapp Alain

UQAM, Montréal, Date of the internship: 06/06/2016 to 29/07/2016

Abstract. In recent years, many discriminations and their consequences have been detected and recognized in both past and modern societies. Furthermore, with the growing use of machine learning, the risk of sustaining such discriminations by learning from biased data is all the more performing. Many studies have been dealing with this issue, by designing techniques to make classifiers more fair, using various methods. However, despite these studies, there is no real consensus on what definition to give of "discrimination". In this context, we propose a new definition to discrimination, which aims at tackling some of the main drawbacks and limits of previous definitions.

1 Introduction

The study of discrimination in personalization algorithms, first introduced by [2] and [11] for classifiers, is motivated by the fact that some unwanted discrimination may appear such algorithms. Indeed, as discrimination is present in many datasets, it is likely to be reproduced by algorithms that learn from these datasets, producing discriminatory classifiers. Since these learning algorithms are more and more used and studied, there is a real need to ensure that no discrimination will be drawn from the classifier they produce.

In the existing literature, many studies try to correct discrimination, with different methods (see Section 5). However, there is no real consensus on how to define discrimination. Indeed, current definitions all have one or many issues, concerning either intuitive aspects of discriminations that are not dealt with, or heavy cost of computation.

To describe these discriminations, we will use the expressions "protected features" to refer to features that should not impact the result (such as sex, gender, ethnic origins, religion...) and "regular features" for the other.

This paper is organised as follows. First, we expose the subtleties that require attention when defining discrimination in Section 2. Then, we present the models in which discrimination is studied, and survey some of the definitions with their limits in Section 3. In Section 4, we propose a new definition of discrimination which relies on two distinct measures : fairness and stability. Finally, after a review of related work about methods to limit discrimination in Section 5, we conclude and expose some possible extension of our work in Section 6.

2 Types of Discrimination

2.1 Indirect Discrimination

Indirect discrimination, as defined in [19], describes the fact that discrimination is possible even without knowing directly the protected features.

For instance, external knowledge may give us relations between regular features and protected features, so that we could infer one from the other. Thus, it becomes interesting to study discrimination upon the concerned regular features.

Even if we do not have access to the protected feature, external knowledge could still used to attack them, and learning algorithm could deduce them from regular features. As illustrated in [3], discrimination is still possible even without any access to protected features.

2.2 Explanatory Discrimination

On the other hand, there might also be some features which could, in some way, explain discrimination. This notion of "explainable features" is developed in [22]. If in a dataset, some regular features are correlated with protected features, there might be situations in which the output is also correlated with the protected features, without leading to discrimination. On the contrary, if the regular feature is important (for instance, suppose that a high education is correlated with a protected feature, and that the decision is hiring in engineering), forcing the parity on protected features may be qualified as affirmative action. It can be an objective goal but it is out of the scope of our work to study it.

To be completely fair relatively to this concept, this condition might be formulated as a wishful thinking to have *equal treatment independent from protected features, provided that the regular features are fixed.*

However, we have to be careful not to apply this type of rule blindly for every regular feature, since it could also become a case of indirect discrimination. For instance, if the ability to carry heavy weights is correlated with some protected feature, it may be regarded as a way to explain discrimination or on the opposite a way to discriminate indirectly, according to the context.

2.3 Multi-criteria Discrimination

In many cases, we may want to avoid discrimination against multiple features simultaneously: for instance sex, religion, ethnic origins, age or any other depending on the context.

A first naive approach would be to consider protected features separately, and to ensure the absence of discrimination for each of them. However, this approach is insufficient in general. For example, for a dataset such as the one shown in Table 1, we can see that there is almost no discrimination against sex or age when they are analyzed independently, while when looking at the combination of these two criteria, the discrimination against elderly women is obvious.

Sex \Age	Young	Old	Total
Male	$\frac{1050}{2000} \simeq 53\%$	$\frac{1000}{1900} \simeq 53\%$	$\frac{2050}{3900} \simeq 53\%$
Female	$\frac{1000}{1900} \simeq 53\%$	$\frac{0}{100} = 0\%$	$\frac{1000}{2000} \simeq 50\%$
Total	$\frac{2050}{3900} \simeq 53\%$	$\frac{1000}{2000} \simeq 50\%$	$\frac{3050}{5900} \simeq 52\%$

Table 1. In this example, each ratio represents the number of hired persons over the total number of applicants. For simplicity, we here consider Sex and Age as binary, although there could be more than two possible values

Hence, there is a need to address multi-criteria discrimination and not only binary discrimination. A naive approach consisting into dealing with each potentially concerned subclass independently, with a binary method, would lead to a computation exponential with respect to the number of protected features, which can be an issue if they are to numerous.

Another issue is to determine what to measure precisely when we have several protected features. Indeed, since there are potentially many values to compare, or even many distinct discrimination, it is not trivial to determine which subset of the population is discriminated against.

2.4 Comparing Similar Individuals

Finally, a certain kind of discrimination can also exist that is not directly deductible from the number of individuals hired. Indeed, if we do not compare similar individuals, we might as well compare different people we do not want to distinguish -for instance male and female-, with completely different criteria -for instance education and physical appearance-.

Another way to see this problem, as expressed in [5], is to ensure that there is no discrimination in the global population, but also in every subset of this population. This last type of discrimination is hard to formalize properly, and literature dealing with this issue is currently relatively poor ([5], [6])

3 Discrimination: Models and definitions

We here propose to present the main models in which discrimination is studied, and the formal definitions that have been developed to take them into account.

3.1 Models

Datasets. The first case study is when we only have records of past decisions, and we want to determine whether or not there has been any discrimination in it. For instance, these situations could be analyzed in order to help justice to make decisions for either potential incriminations or compensations.

In this model, we consider a dataset A and set of features describing individuals F. We also consider a non-empty set of *protected* features $S \subseteq F$ for

which we want to measure discrimination. Based on some prior knowledge, we can also consider protected features that are not included in F. For this model, we distinguish two main types of studies: first, studies which focus on detecting discrimination such as [17], and then, research whose aim is to modify the data, so as to produce a fair version of it, which could be used for learning without discrimination (see more details on section 5).

When focusing in detection, it is not always interesting to take indirect discrimination into account:

- If we know the protected feature and detect both discrimination against this feature and whether or not similar individuals are treated equally, there is no need to look for indirect discrimination against this feature. For instance, if we detect discrimination against a protected feature such as religion, and ensure that similar individuals are treated equally no matter their religion, there is no need to study indirect discrimination.
- If we have neither information about the protected feature nor exterior knowledge about any correlation to known features (or combination of features), no detection can reasonably be made. For instance, if we do not have prior knowledge nor access to the protected feature, no indirect discrimination can be studied.

Hence, if the problem of treating equally similar individuals is taken care of, the only case in which indirect discrimination has to be studied is the one in which we do not know the protected feature, but we have background knowledge about correlation to other known features (or combination of features).

Personalisation system as a box. Discrimination can also be studied by analysing the set of decisions taken by a personalisation system. For instance, the personalisation system could be a classifier built thanks to a learning algorithm trained on previous decision records, and used to provide advice on hiring. According to the study, we either know or not the parameters used by the classifier. If we know the description of the classifier, this one is called a *white box*, otherwise, we use the term *black box* to refer to it. These models can be viewed as follows: we have a classifier to which we can provide entries so as to obtain the corresponding outputs, and we want to either detect discrimination it may introduce, correct it, or both.

In this case, having access to the training data matters less than having the distribution of the data the algorithm will be applied to. Indeed, knowing the distribution of data is important to determine how biased the algorithm might be, and to what extent we have to correct it.

In contrast to the previous model, in some scenarios, indirect discrimination is very important to take into account, especially when we want to correct the algorithm. Retrieving the protected feature is not sufficient to train a fair classifier (as in [3]), and detecting indirect discrimination might be useful to test the box on more specific entries.

Note that a valid definition for datasets can be extended to fit this new model. With a definition of discrimination for datasets, if we want to define

discrimination for a black-box while knowing the distribution, we can build a dataset by measuring the outputs of the black-box for entries selected according to the distribution. Then, measuring the discrimination of the dataset produced is equivalent to measuring the discrimination of the black-box for the given distribution of profiles. However, the opposite reasoning is wrong since any profile can be provided to a classifier, so that we can get the corresponding ouput, but we have no guarantee that all profiles are represented in a dataset.

3.2 Definitions

According to the model which is used, and the objective we want to achieve, various definition of "discrimination" have been proposed. Hereafter, we review some of them.

Discrimination Score and Disparate Impact. Using dataset model, a first definition of discrimination is given by the US law [1] as the Disparate Impact.

The original objective of disparate impact is to determine whether a particular population (for instance $X = 1$ in the table 2 could be $Sex = Female$) is discriminated, based on a binary criterion (for instance, $C = YES$ could be "has been hired") which is either positive or negative. Disparate impact is defined based on the probability that a category of the population gets an advantage divided by the same probability for the complementary category.

Hence, using the notation of the Table 2, we get

$$\mathsf{DisparateImpact} = \frac{P(C = YES|X = 1)}{P(C = YES|X \neq 1)} = \frac{d/(b+d)}{c/(a+c)}$$

	$X \neq 1$	$X = 1$
$C = NO$	a	b
$C = YES$	c	d

Table 2. Confusion matrix, $X = 1$ indicates the protected subset, C is the outcome and a, b, c and d are the number of individuals fitting in the associated categories

This measure gives a value strictly between 0 and $+\infty$, is not defined if $c = 0$ (among other criteria) and unstable for low values of c or d. Hence, in the literature, another measure is used that is closely related: the Discrimination Score, as defined in [3] (also called the Statistical Parity).

This measure is defined as

$$DiscriminationScore = P(C = YES|X = 1) - P(C = YES|X \neq 1) = \frac{d}{b+d} - \frac{c}{a+c}$$

In contrast to the disparate impact, this measure is symmetrical, gives a value between -1 and 1 and keep being defined if $c = 0$.

These two measures have both their advantages and drawbacks. In the following, we will focus on the Discrimination Score instead of the Disparate Impact, since both have similar properties. First, this measure is easy to compute for every dataset, intuitively logical, and easy to understand.

However, there are also many concerns with respect to this measure, which are mentioned in [5]. Indeed, the first concern of this measure is that it does not deal with multi-criteria discrimination. For the same reason, it cannot take indirect discrimination into account. But even within this restriction, this definition has some flaws. First, the strict application of this measure can lead to some positive discrimination. More precisely, it does not take explanatory discrimination into account. Finally, as underlined in [5], it cannot necessarily be used to compare similar individuals. Hence, if statistical parity is respected for a given set, it does not mean that it is respected for its subsets.

α − protection. Another definition is given in [17] that tries to determine whether a dataset is discriminatory. The main idea is to determine inference rules based on the dataset considered (in particular whether the value of a protected feature implies a specific result). This definition relies on the following notions, with A and C referring to specific values of two different features:

- The support of A and C, denoted by $supp(A, C)$, is the number of occurrences of A and C for an individual within the dataset.
- The confidence of the rule $A \to C$ denoted by $conf(A \to C)$ is equal to $supp(A, C)/supp(A)$.

Let A be a set non-empty of values for protected features, B a set (potentially empty) of values for regular features, and C the output, we denote by $\delta = conf(A, B \to C)$ and $\gamma = conf(B \to C)$

The "extended lift" is defined as $elift(\gamma, \delta) = (\gamma/\delta) = \frac{conf(A,B\to C)}{conf(B\to C)} < \alpha$.

The definition of $\alpha - protection$ given by [17] is the following: for a given α, a rule is $\alpha - protective$ if $elift(\gamma, \delta) < \alpha$.

This definition is interesting because it focuses on inference rules and not on the data itself. Hence, if a rule is determined as being discriminatory, we can still study it, and eventually, if the discrimination corresponds to explanatory discrimination, accept it. Furthermore, methods are also mentioned in [17] to determine indirect discrimination, by studying features that are often associated with protected ones.

However, there are still flaws in this method. The first one is that features absent from A and B can take any possible value. which means that we can compare very different profiles, especially if A is empty. The second one is the number of rules generated. Indeed, if a high number of rules is detected, which is often the case with this method, it is both hard to interpret them correctly, and subject to statistic bias. More precisely, statistically, even with completely random datasets, the more numerous are the rules we study, the higher the probability to find correlation with the output, although it is not necessarily discriminatory. Hence, false positive might be a real issue with this definition.

k-nearest-neighbor-consistency. The measure designed in [5], to solve the main limits of the disparate impact, aims at treating similar individuals equally.

To realize this, it assumes a distance between individuals d, a mapping from the individuals to the output M, and a distance D between the possible outputs. From that, the idea is to set a Lipschitz condition to respect: for V a set of individuals, $\forall x, y \in V, D(M(x), M(y)) \leqslant d(x, y)$.

Remark that both the main advantages and the main limits of this definition come from the metric that rules the distance between individuals. Indeed, with such a definition, it becomes possible to make sure that we only compare similar individuals. We can also ensure the absence of multi-criteria discrimination by fixing a very low or even null distance for these criteria so as to guarantee similar treatment. Similarly, with a relevant choice of distance, neither indirect nor explanatory discrimination are real issues : for instance, low distance for features that are likely to cause indirect discrimination, and higher ones for explanatory features.

However, the main limit of the method precisely comes from this "relevant choice of distance". We believe that it is possible to design distances that are relevant for specific domains, but these distance are not easy to adapt to other domains since features generally do not have the same weights in all domains. In particular, a feature can be seen as explaining discrimination in some cases and as causes for indirect discrimination in others. Thus, the main question is how to create good distances. Both relying on experts or committees, and relying on machine learning have their limits, since the first is subject to the bias of the persons concerned and the second relies on potentially discriminating data, which we precisely want to analyse.

Hence, although this definition addresses many of the problems mentioned previously, it appears to be very hard to scale in a general case.

Quantitative Input Influence. Another definition, proposed in [4], requires to use directly a classifier as a black-box. To apply this definition, a classifier is seen as a black box and the distribution of the data is known. Given a set of protected features, and a supposedly discriminated group (*i.e.* a set of feature), the paper [4] proposes to study the difference between the average classification of that group, against the classification of the same group, by changing protected features according to the distribution. We refer the interested reader to [4] for more more details.

Although this method is interesting in theory since it allows to take many criteria into account at once, and only them (it has no problem with explanatory discrimination), it does have many shortcomings. First, it still requires either to make one test for each kind of discrimination, which means in the worst case 2^n tests in which n is the number of protected features, or to know exactly which combination of criteria to assess.

Even for one test, computation is still heavy, and even more if the number of protected features is large. In this case, heuristics are mandatory for high number of protected features to obtain reasonable computation time.

Furthermore, this definition also requires the distribution of the outputs, and not only the distribution of the data to be used by the classifier. This distribution could be obtained either directly in the model, or inferred by a sufficient amount of trials on the black-box. In the first case, it adds another assumption on the model while otherwise it increases the computational time.

4 A proposal for a new definition

Based on the limits of the previous definitions, we propose a new one, which aims at solving the following shortcomings:

- Handling correctly multi-criteria discrimination.
- Dealing with indirect discrimination when it occurs.
- Being able to handle explanatory discrimination.
- Being able to compare similar individuals.
- To have a reasonable computational complexity.

To capture both the global discrimination and the comparison of similar profiles, we propose to use two complementary measures: fairness and stability.

The fairness estimates the global discrimination against a certain subset of individuals (which all have similar characteristics on protected features), while the stability quantifies the consistency with which similar decisions are taken for similar individuals, without taking protected features into account, even if these variations are balanced at a global scale.

To do so, we propose to study classifiers (given either as a white or a black boxes) with known distribution of the data it will be used on. The method that we propose is parametrized by a variable k, which impacts its precision as well as its complexity.

We first define "regular profile" as all the combination of values for the regular features, "protected profile" a combination of values for every protected features.

By combining an regular profile with a protected profile, we obtain a complete individual, artificially created. The method works as follows. First, we select k regular profiles at random according to the given distribution. Then, for each of these profiles, we add each possible protected profile, and check the output of the classifier for corresponding individuals.

We define the gain of a protected class as the sum of the positive outputs for the k regular profiles, divided by k so as to get a value between 0 and 1. Then, we define fairness as the difference between the maximum and the minimum gains for a protected profile: the higher the value, the less fair it is.

With the example given in Table 3, the fairness would be computed as

$$\begin{aligned}\mathsf{Fairness} &= max_{x \in ProtectedClass}(gain(x)) - min_{y \in ProtectedClass}(gain(y)) \\ &= gain(Sex = Male, Age = Young) - gain(Sex = Female, Age = Old) \\ &= 42/50 - 17/50 \\ &= 0.5\end{aligned} \tag{1}$$

Sex Age	Young	Old
Male	42	35
Female	37	17

Table 3. In this example, each number represent the number of times that the black box as given a positive result for the associated protected profile for $k = 50$ regular profiles tested. For simplicity, we consider Sex and Age as binary, although there could be more than two possible values

We also define the stability of an regular profile for binary output as the minimum between the number of positive outputs and the number of negative outputs when we provide the classifier with all possible individuals that have these regular profile (we combine the regular profile with each possible protected profile), multiplied by 2 and divided by the number of regular profiles (this value is normalized between 0 and 1).

Our global measure of stability is defined as the sum of the stability of all regular profiles, divided by k (*i.e.* the number of regular profiles). With this definition, the higher this value is, the less stable the classifier is.

With the example given in Table 4, the stability would be computed as

$$\begin{aligned} \mathsf{Stability} &= \frac{1}{k} * \sum_{i \in RegularProfile} \frac{2 * min(Output(i))}{2 * k * |ProtectedProfile|} \\ &= \frac{2 * (3 + 2 + 7)}{3 * 16} \\ &= 0.5 \end{aligned} \tag{2}$$

Regular profile Output	Positive	Negative
Profile A	3	13
Profile B	14	2
Profile C	7	9

Table 4. In this example, each number represent the number of apparition of the corresponding output for each regular profile. We here consider $k = 3$ and 4 binary protected features, which leads to $2^4 = 16$ protected profiles and outputs by regular profile

We have summarized our method in Algorithm 1, which computes these values in a generic case.

Note that our method need to compute all possible combinations of protected features. Thus, this method has both advantages and drawbacks. First, the computational cost is very heavy, being exponential in the number of protected features. For that reason, in practice, we recommend not to use a large

Algorithm 1: Fairness and Stability

Input: S a set of protected features, R the set of regular features, C a classifier that gives a binary output for a given individual, D the distribution of the population to be tested D, $k$the parameter controlling the precision and the complexity

Output: $fairness$, $stability$

$fairness \leftarrow 0$;
$stability \leftarrow 0$;
$regular\_profiles \leftarrow create\_regular\_profiles(R, D, k)$;
$protected\_profiles \leftarrow create\_protected\_profiles(S)$;
$nb\_protected \leftarrow |protected\_profiles|$;
$table\_protected \leftarrow void\_table$ **for** $r \in regular\_profiles$ **do**
    $count\_pos \leftarrow 0$;
    $count\_neg \leftarrow 0$;
    for p *in* $protected\_profiles$ **do**
        $profile \leftarrow fusion\_profile(r, p)$;
        if $classifier(indiv) == pos$ **then**
            $count\_pos \leftarrow count\_pos + 1$;
            $table\_protected(p) \leftarrow table\_protected(p) + 1$;
        else
            $count\_neg \leftarrow count\_neg + 1$;
        $stability \leftarrow stability + \frac{2*min(count\_pos, count\_neg)}{nb\_protected}$;
$stability = \frac{stability}{k}$;
$fairness = \frac{max(table\_protected) - min(table\_protected)}{nbre\_protected}$;
return $fairness$ *and* $stability$;

number of protected features. However, even if heuristics are possible, any non-exhaustive approach may lead to forget some specific type of discrimination.

This exhaustive aspect may also induce the classifier to be tested on individuals that are not likely to exist in reality. For instance, it can be tested on profiles which do not exist neither on the training dataset, nor on the one used for the testing. Thus it is possible that outliers are generated, which can lead to aberrant values. However, we believe that it is still a good idea to take into account unlikely profiles: hence, the method can detect flaws in the learning algorithm which could damage even very unlikely candidates.

Furthermore, the larger is the number of protected profiles, the most likely it is to find what could be called "artificial" correlation. Indeed, by multiplying the number of combinations studied, we statistically increase the probability of finding correlation between the input and the output. Even with completely random datasets, if we increase the number of protected profiles, it is likely that we shall find correlation between the output and one of them, even if it is not significant. In that case, the measure of fairness will worsen. Hence, a user of this definition has to be aware of this potential issue for a high number of protected features.

Finally, the selection of regular profiles, at random following the targeted distribution, is a design choice. Indeed, other sampling methods are possible, such as uniform random selection, or a generation of profiles that would progressively focus on profiles similar to those for which unfairness is detected. Intuitively uniform random selection might cause to focus on marginal discrimination instead of discrimination that really occurs, and focusing on problematic profiles might cause to bias the measure since it would voluntarily and artificially amplify the discrimination detected.

5 Related Work

In the existing literature, many works focus on the correction of bias in either datasets or algorithms. This work can be globally classified into three main families of approach.

5.1 Pre-processing

A first way to deal with the correction of bias is to correct the data itself, so as to make it free of discrimination, and easily used as training data afterwards. There are two main issues to this method: first, the data used later on will not be the original data, so that it can impact the quality of the prediction. Second, as we do not know which algorithm will be applied on the corrected data, it is always hard to ensure formally that no discrimination can be drawn from it.

Among the techniques developed, some propose to add an artificial feature, so as to make the training free from discrimination, when it is done with specific algorithms ([3]). Other propose to modify individuals considered as uncertain after training from a first algorithm. The modification performed can be change

in values, deletion or a change in the number of times a particular profile appears in the dataset ([11,12,2,9,22,8,18]). Finally, some modify the dataset so as to avoid discriminative rules detected. In that case, the definition of $\alpha - protection$ is used ([10]).

Some of these methods also take privacy in consideration such as [18], or [8], since some techniques are close (deletion of profiles for instance). Indeed, these two topic are quite close, and managing to reach both in order to release datasets publicly would be very interesting.

5.2 Modification of the algorithm

Another family of approach modifies the algorithm itself so as to avoid discrimination. Such an approach is interesting since it can be used on any data, but it is often specific to a certain kind of learning algorithm.

For instance, this approach is developed in [20] to ensure absence of correlation with protected features for some classic algorithms (logistic regression, hinge loss, Support Vector Machines), in [3] for Bayesian methods, in [13] for decision tree learning by using an alternative value instead of direct entropy, in [15], [14] and [21] by proposing regulizers to take discrimination into account in the objective function or and in [6] that presents a method for boosting by adding weak learners dedicated to correction of bias.

5.3 Post-processing

Finally, the last approach is to modify the classifier itself, after its training. Although this approach has theoretical limits when the classifier only gives the result and no confidence ([13]), this generic approach remains a very practical way to obtain an unbiased classifier without modifying the original data.

Some work have instantiated this approach, such as [3] and [17] which modify the output for either unsure individuals (when the classifier gives a confidence score), or individuals on which distinct classifiers disagree. The work proposed in [13] also study this approach, but focuses on relabelling tree classifiers.

6 Conclusion and future work

Currently, many definitions of discrimination still co-exist in current studies: even if Statistical Parity and α-protection are the most used, they still have limits and many propositions aim at outpassing them.

Our contribution advocates for a dual measure of discrimination, in order to take into account both the "fairness" to quantify the global discrimination, and the "stability" to describe to what extent similar individuals are treated equally. We have proposed a first way to compute both of these measures. Even if this proposition is still restricted to some specific cases and not perfect, we think that the combination of these measures tackles many of the issues found in other definitions.

Furthermore, the duality of these aspects is also promising for a good adaptation to more general cases, or even for good heuristics if the computation were to be too heavy. We propose here a non-exhaustive list of points which still require further study.

6.1 Refining our definition

The first limit of our measures relies in the fact that we only study *complete* protected profiles: combination of every protected feature. This measure is necessary not to forget some combinations, but it may obscure some more general discrimination. For instance, in the table 1, our definition would focus on the discrimination of the class characterized by values 1 and 1 for criteria A and B against the class characterized by values 2 and 2 for the same criteria, while in practical situations, we may prefer to detect the more global discrimination on criterion A alone.

Criterion A	Criterion B	Ratio of positive results
1	1	100
1	2	50
2	1	98
2	2	48

Fig. 1. Example of problematic case, where Criterion 1 and 2 are protected features

Even with this issue, we think that our measure of fairness remains valid: the value still quantify the maximum gap between to complete protected profiles. However, any interpretation to find which population is discriminated against has to be done cautiously: with multi-criteria discrimination, it is hard to determine which feature, or combination of features, is the most discriminating one. The stability measure does not suffer from this issue because it precisely aims at detecting disparities between these complete classes.

We notice that for non-binary outputs, these definitions also have to be adapted in a proper way: for stability, an proposition would be to count, for each unprotected profile, the number of output which differ from the most common one. For the fairness, we could also have distinct counters for each possible value of the output. But if the values of the output are ordered (categories of wages for instance), some other measures taking into account the global distribution on the output may be preferable.

Finally, our definition only consider simple classifiers that give as only output a binary classification. In reality, we could get confronted to more elaborate classifiers which also give confidence score. In this case, our definition work, but it does not fully exploit the information. Taking the trust score into account, could contribute to improve our measures: intuitively, with a low trust score, the weight of a measure could be lightened in both fairness and stability measures.

6.2 Discovering indirect discrimination

An other issue in discrimination is indirect discrimination discovery. Even if some method exist, such as given in [16], this method is not perfect and focuses on a dataset, as opposed to a classifier.

Hence, a general scheme of discovery for indirect discrimination, depending on a classifier, could be very interesting to elaborate. To this aim, the use of feature selection (various method of such selections are surveyed for instance in [7]) to detect what feature could be used to infer the protected features might be a promising way to begin such a study, with either black or white boxes.

This method of feature selection could also be used to measure discrimination of a dataset with high number of protected features, in order to reduce the study field.

References

1. Supreme court of the united states. watson v. fort worth bank & trust. 487 u.s. 977, 995, 1988.
2. Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. technical report, 2009.
3. Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.
4. Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. 2016.
5. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM.
6. Benjamin Fish, Jeremy Kun, and Ádám Dániel Lelkes. A confidence-based approach for balancing fairness and accuracy. *CoRR*, abs/1601.05764, 2016.
7. Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.
8. Sara Hajian and Josep Domingo-Ferrer. A study on the impact of data anonymization on anti-discrimination. In Jilles Vreeken, Charles Ling, Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, editors, *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012*, pages 352–359. IEEE Computer Society, 2012.
9. Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. on Knowl. and Data Eng.*, 25(7):1445–1459, July 2013.
10. Sara Hajian, Josep Domingo-Ferrer, and Antoni Martínez-Ballesté. *Rule Protection for Indirect Discrimination Prevention in Data Mining*, pages 211–222. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
11. Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling.
12. Faisal Kamiran and Toon Calders. Classifying without discriminating.

13. Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Proceeding of ICDM 2010*, pages 869–874. IEEE Computer Society, 2010.
14. T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, Dec 2011.
15. Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. *Fairness-Aware Classifier with Prejudice Remover Regularizer*, pages 35–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
16. Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pages 581–592. SIAM, 2009.
17. Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 560–568, New York, NY, USA, 2008. ACM.
18. Salvatore Ruggieri. Data anonymity meets non-discrimination. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE, IEEE, 2013.
19. Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. Data mining for discrimination discovery. *ACM Trans. Knowl. Discov. Data*, 4(2):9:1–9:40, May 2010.
20. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Fairness constraints: A mechanism for fair classification. In *FAT ML: Fairness, Accountability and Transparency in Machine Learning Workshop (ICML 15)*, 2015.
21. Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 325–333. JMLR Workshop and Conference Proceedings, May 2013.
22. I. Žliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *2011 IEEE 11th International Conference on Data Mining*, pages 992–1001, Dec 2011.