POLITECNICO DI TORINO
Repository ISTITUZIONALE

Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data

Original
Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data / Mecati, Mariachiara; Vetro, Antonio; Torchiano, Marco. - (2021), pp. 4287-4296. (Intervento presentato al convegno 2021 IEEE International Conference on Big Data (IEEE BigData 2021) - First International Workshop on Data Science for equality, inclusion and well-being challenges (DS4EIW 2021) tenutosi a Orlando, FL, USA nel 15-18 Dec. 2021) [10.1109/BigData52589.2021.9671443].

Availability:
This version is available at: 11583/2950474 since: 2022-01-20T11:15:23Z

Publisher:
IEEE

Published
DOI:10.1109/BigData52589.2021.9671443

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

04 May 2024

Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data

Mariachiara Mecati
Department of Control and Computer Engineering (DAUIN)
Politecnico di Torino
Torino, Italy
mariachiara.mecati@polito.it

Antonio Vetrò
Department of Control and Computer Engineering (DAUIN)
Politecnico di Torino
Torino, Italy
antonio.vetro@polito.it

Marco Torchiano
Department of Control and Computer Engineering (DAUIN)
Politecnico di Torino
Torino, Italy
marco.torchiano@polito.it

***Abstract*—Bias in the data used to train decision-making systems is a relevant socio-technical issue that emerged in recent years, and it still lacks a commonly accepted solution. Indeed, the "bias in-bias out" problem represents one of the most significant risks of discrimination, which encompasses technical fields, as well as ethical and social perspectives. We contribute to the current studies of the issue by proposing a data quality measurement approach combined with risk management, both defined in ISO/IEC standards. For this purpose, we investigate imbalance in a given dataset as a potential risk factor for detecting discrimination in the classification outcome: specifically, we aim to evaluate whether it is possible to identify the risk of bias in a classification output by measuring the level of (im)balance in the input data. We select four balance measures (the Gini, Shannon, Simpson, and Imbalance ratio indexes) and we test their capability to identify discriminatory classification outputs by applying such measures to protected attributes in the training set. The results of this analysis show that the proposed approach is suitable for the goal highlighted above: the balance measures properly detect unfairness of software output, even though the choice of the index has a relevant impact on the detection of discriminatory outcomes, therefore further work is required to test more in-depth the reliability of the balance measures as risk indicators. We believe that our approach for assessing the risk of discrimination should encourage to take more conscious and appropriate actions, as well as to prevent adverse effects caused by the "bias in-bias out" problem.**

***Index Terms*—Data quality, Data bias, Data ethics, Algorithm fairness, Automated decision-making**

I. Introduction

Decision-making processes are rapidly turning into automated decision-making (ADM) systems in a variety of sectors of our society, both in private and public organizations, leveraging the large availability of data and classification/prediction algorithms [1]. This new phase of automation is supposed to increase economic efficiency as well as remove human subjectivity and errors. However, alongside these possible benefits, indisputable harms are now evident: when trained on biased data, automated data-driven processes replicate or even amplify the same bias of our society [2] [3]. From a technical perspective, one form of biased data is imbalanced data, which is an unequal distribution of the occurrences between the classes of a given attribute (e.g. gender, ethnic group, etc.) [4]. Causes of imbalanced data include errors or limitations in the data collection process (including design and operations), or merely because the reality that the data reproduce is itself imbalanced in given characteristics (e.g., data about nurses can be easily imbalanced with respect to gender). Specifically, the imbalance is between-class when only two classes are taken into consideration and one class is over-represented with respect to the other, or multiclass when imbalances exist between multiple classes. Herein we focus on the more general case, i.e., multiclass imbalance. Imbalanced data is known for a long time to be a problematic aspect in the machine learning domain [4]: this issue is still relevant [5] because it can give rise to very heterogeneous accuracy across the classes of data and, consequently, to relevant social, ethical, and legal issues. In this paper we study how an imbalance in the training data can be used as a predictor of possible unfair software output, combining concepts from data quality measurement and risk management.

We provide the theoretical foundations of the approach in section II; then, we describe the experimental design in section III, while results are reported and discussed in section IV. After that, we position our work in the literature in section V and we take into consideration the limitations of the approach in section VI. Finally, we highlight conclusions and potential future work in section VII.

II. Data imbalance as risk indicator

Our measurement approach is derived from the series of standards ISO/IEC 25000:2014 Software Engineering — Software Product Quality Requirements and Evaluation (SQuaRE) [6], which describes quality models and measurements of software products, data and services. In this family of standards, quality is composed of quantifiable characteristics and sub-characteristics. Data imbalance (or its dual concept of data

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

balance) is not a characteristic of data quality in ISO/IEC 25012:2008, however it can be seen as a possible extension because it is a key element in the chain of effects and dependencies described in SQuaRE: according to this principle, data quality has an effect on the system quality in use and as a consequence on the users of a software system. In our context, imbalanced datasets may lead to imbalanced software outputs, which means –in the context of ADM systems– differentiation of products, information and services based on personal protected characteristics, and thus discrimination. Therefore, according to this line of reasoning, we treat data imbalance as an extension of the data quality model formalized in ISO/IEC 25012:2008, and we quantify it with proper measures, extending those already defined in ISO/IEC 25024:2015.

The second pillar behind our approach is represented by the ISO 31000:2018 standard on risk management [7] that provides the guiding principles for the management of risks. Because of its role in systematic and unjustified unequal treatment – and even unlawful discrimination– made by ADM systems, we propose that data imbalance shall be also considered as a risk factor in all those systems that rely on historical data and that automate decisions on important aspects of the lives of individuals, which concern the exercise of their rights and freedoms (think to automated decisions on wages, education, working positions, social benefits, etc.).

For reasons of space we cannot analytically report on all the relations between our proposed approach and the two ISO standards (which can be found in [8]), however, following the line of reasoning exposed above, should clearly emerge our hypothesis: by measuring the level of (im)balance of specific attributes in a dataset, it is possible to detect the risk of bias in the classification output from ADM systems. More in detail, we refer to the following specifications:

- software systems are biased when they "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others [by denying] an opportunity for a good or [assigning] an undesirable outcome to an individual or groups of individuals on grounds that are unreasonable or inappropriate [9];
- we identify as social groups of category object of possible discrimination those identified by the characteristics provided in "Article 21 - Non- discrimination" of the EU Charter of Fundamental Rights [10]:

  1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

  2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited.

III. Experimental design

Our goal consists in understanding how the balance of protected attributes in training data can be used to assess the risk of algorithmic unfairness. For this purpose, we conducted a study aimed at answering the following *Research Question*:

> Is it possible to identify the risk of bias in a classification output by detecting the level of (im)balance in the input data?

In order to conduct this analysis, we selected a set of indexes that are able to measure balance in the data –and thus its absence, i.e. imbalance–, and we assessed how well such balance measures applied to a given dataset reflect a discrimination risk. Specifically, we followed this procedure:

1. we selected a large *dataset* (available in the literature and described in section III-A) and a multiclass *protected attribute*[1] with cardinality "m" ;
2. using a *mutation technique*, we generated a number of derived synthetic datasets having different levels of balance; specifically, we adopted a pre-processing method as mutation technique (see section III-B) and we mutate the distribution of the occurrences between the classes of a certain attribute by adjusting the specific parameter `C.perc`;
3. we implemented a *binomial logistic regression* model in order to predict the *score variable* for each synthetic dataset; particularly, we trained a binary classifier on a training set composed by the 70% (randomly selected) of the data and we ran it on the remaining 30%, which represents the test set;
4. we measured the level of (im)balance of the protected attribute in the training set through four different widely used *balance measures* (described in section III-C);
5. we applied two distinct *fairness criteria* (see section III-D) to the protected attribute in the test set –i.e. to the classifications obtained from the model– for a total of three unfairness measures on each output.
6. we analyzed first the behavior of both balance and unfairness measures in response to mutations, and then we examined the relationship between balance measures and fairness criteria, by checking in the end whether a negative correlation holds, that is, whether a lower level of balance corresponds to a higher level of unfairness, and vice-versa.

Therefore, by examining the *balance* features of the protected attributes in the training data, we aim at analyzing if the indexes of balance taken into account are able to reveal a risk of bias in the test set, for the purpose of evaluating the reliability of such balance measures as risk indicators of distorted recommendations or biased decisions –i.e., discrimination risks– in the context of ADM systems.

[1]For identifying an attribute as *protected*, we take as reference the definition provided in "Article 21 - Non-discrimination" of the EU Charter of Fundamental Rights [10], as already mentioned in section II.

A. *Data*

With a view to exploring the potential of our approach in one of the prominent fields of application of ADM systems, we examined a dataset belonging to the application domain of financial services: **Default of Credit Card Clients**, whose properties have been summarized in table I.
This dataset –referred to as *Dccc* hereinafter– has been retrieved by the Kaggle platform[2] and was chosen because of the high impact of using ADM systems in this domain and that particular dataset because of popularity: at the time of the research, it was ranked as the fourth most voted dataset on credit cards on Kaggle[3] and it fits better our study than the one ranked first (Credit Card Fraud Detection), which is based on transactions, while we are interested in datasets that collect data on persons.

Dccc is composed of 25 variables: it contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005 [11]. Moreover, since this dataset does *not* contain a pre-computed classification, so we built a *binomial logistic regression* model in order to predict the *score* variable: in particular, we trained a binary classifier on a $training$ set composed by the 70% (randomly selected) of the original dataset and we ran it on the remaining 30%, which represents the $test$ set. Finally, note that in real datasets we can often find missing values (NA), so we decided *not* to exclude missing values from the analysis and to consider them as a separate "NA" category.

TABLE I
SUMMARY OF THE DATASET'S PROMINENT PROPERTIES.

Dataset	Size	Domain	Target variable	Protected attribute	m
Default of credit cards clients (Dccc)	30000×25	Financial	default payment next month	education	6

B. *Mutation technique*

We adopted a specific *pre-processing* method as mutation technique in order to create several variations of the distribution of the occurrences between the classes of a given protected attribute. To this end, we used the `UBL-package` [4]:

"*The package provides a diversity of pre-processing functions to deal with both classification (binary and multi-class) and regression problems that encompass non-uniform costs and/or benefits.*".

In particular, we assumed the `SmoteClassif` function [5] as mutation technique:

"*This function handles unbalanced classification problems using the SMOTE method. Namely, it can generate a new 'SMOTEd' data set that addresses the class unbalance problem.*".

This method has been applied with the following settings:

- "`education~`" is the multi-class protected attribute chosen as formula; as a consequence of our decision to include missing values in the analysis and count them as a separate category, this attribute consists of six classes: "NA", "graduate school", "university", "high school", "others", "unknown".
- "`C.perc`" is a list containing the percentages of under-sampling or/and over-sampling to apply to each class of the protected attribute in the formula: an over-sampling percentage is a number above 1, while an under-sampling percentage should be a number below 1; in particular, a class remains unchanged if the number 1 is provided for that class; note that there exists an infinite number of possible combinations of the percentages of the classes. Alternatively, `C.perc` may be set to the two values "balance" (the default) or "extreme", cases where the sampling percentages are automatically estimated either to balance the examples between the minority and majority classes, or to invert the distribution of examples across the existing classes transforming the majority classes into the minority, and vice-versa.
- "`repl=FALSE`" is a boolean value controlling the possibility of having (or not, as in this case) repetition of examples when performing under-sampling by selecting among the majority class(es) examples.

In our study we decided to examine five different cases for the parameter `C.perc`: first, we set the parameter to the pre-established value "balance" –which is the perfect uniform distribution, with all the occurrences equally distributed between the classes–, then we assigned four different lists of percentages for the classes of the protected attribute, corresponding to the exemplar distributions "Power2", "HalfHigh", "OneOff" and "QuasiBalance" already analyzed in [12]. In particular, such exemplar distributions are described as follows:

- *Power 2*: occurrences are distributed according to a power-law with base 2, i.e., distributions among the classes increase like the powers of 2;
- *Half High*: occurrences are distributed mostly among half of the classes while the remaining have a very low frequency –specifically, a ratio of 1:9 has been chosen for the frequencies of the two halves;
- *One Off*: occurrences are distributed among all classes but one;
- *Quasi Balance*: half of the classes are 10% higher w.r.t. max balance and the other half is 10% lower.

In addition, for each exemplar distribution we considered 6 permutations of the values of the percentages assigned to the different classes of the protected attribute. Finally, in order to increase the variability –and thus the reliability– of our method, we decided to vary a `seed` (an integer recommended

[2]https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
[3]https://www.kaggle.com/datasets?search=credit+card&sort=votes, last visited on November 9, 2021.
[4]https://rdocumentation.org/packages/UBL/versions/0.0.6/topics/UBL-package, last visited on November 9, 2021
[5]https://www.rdocumentation.org/packages/UBL/versions/0.0.6/topics/SmoteClassif, last visited on November 9, 2021

for reproducibility purposes to keep track of the samples) by setting 100 randomly sampled values between 1 and 1000.

Therefore, for the discussion of the results we kept track of the outcomes for each value of the `seed` in the case of the mutation with `C.perc`="balance", for a total of 1×100=100 values for each measurement –both *balance measures* and *fairness criteria*–; whereas in the case of the mutations corresponding to the four different lists of percentages, we collected a total of
4 (exemplar distributions) × 6 (permutations) × 100 (seed) = 2400 values for each measurement, leading to a grand total of 100+2400=2500 values for each *balance measure* and 2500 values for each *unfairness measure*.

C. Balance measures

In this study, we limited our attention to *categorical* attributes and we selected four indexes of data balance (summarized in table II) that are widely used in the literature. The measures have been normalized in order to meet two criteria:

- range in the interval $[0, 1]$;
- share the same interpretation: the closer the measure to 1 and the higher the balance (i.e. categories have similar frequencies); vice-versa, values closer to 0 means more concentration of frequencies in few categories, thus an imbalanced distribution.

a) ***Gini index****:* it is a measure of heterogeneity that reflects how many different types are represented and it is used in many disciplines with different designations: examples are political polarization, market competition, ecological diversity as well as racial discrimination. In statistics, the heterogeneity of a discrete random variable can vary between a degenerate case (= minimum value of heterogeneity) and an equiprobable case (= maximum value of heterogeneity, since categories are all equally represented). Thus, for a given number of categories the heterogeneity increases if probabilities become as equal as possible, i.e. the different classes have similar representations.

b) ***Shannon index****:* it is a measure of species diversity in a community, which is a widely employed concept in biology, phylogenetics and ecology. Indeed, diversity indexes represent a useful tool to measure imbalance providing information about community composition taking the relative amounts of different species (classes) into account.

c) ***Simpson index****:* it is another indicator of diversity that measures the probability that two individuals randomly selected from a sample belong to the same species, i.e., the same category. It is employed in social and economic sciences for measuring wealth, uniformity and equity, as well as in ecology for measuring the diversity of living beings in a given location.

d) ***Imbalance Ratio****:* the Imbalance Ratio (IR) is a widely used measure made of the ratio between the highest and the lowest frequency. We take the inverse in order to normalize it in the range $[0, 1]$ and to make it comparable to the previous balance measure, i.e., the higher the values and the higher the balance.

TABLE II
THE *balance measures* WITH THE RESPECTIVE FORMULA, WHERE WE CONSIDER A DISCRETE RANDOM VARIABLE WITH m CLASSES, EACH WITH FREQUENCY f_i (= PROPORTION OF THE CLASS i W.R.T. THE TOTAL) WHERE $i = 1, ..., m$:

Gini	$G = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^{m} f_i^2\right)$
Simpson	$D = \frac{1}{m-1} \cdot \left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$
Shannon	$S = -\left(\frac{1}{\ln m}\right) \sum_{i=1}^{m} f_i \ln f_i$
Imbalance Ratio	$IR = \frac{\min(\{f_{1..m}\})}{\max(\{f_{1..m}\})}$

D. Fairness assessment

We assessed the *unfairness* of automated classifications relying on two criteria formalized in [13]. Note that hereinafter we call indistinctly "Fairness criteria" and "Unfairness measures", as we assume the fairness criteria as measures of unfairness in a classification output.
In general, to evaluate unfairness we consider a sensitive categorical attribute A that can assume different values $(a_1, a_2, ...)$, a target variable Y and a predicted class R where Y is binary (i.e., $Y = 0$ or $Y = 1$ and thus $R = 0$ or $R = 1$). In practice, we aim to check whether the ADM system, which assigned a predicted class, behaved fairly w.r.t. the different values of a sensitive attribute.

Independence criterion*:* this criterion requires the acceptance rate to be the same in all groups, where acceptance corresponds to the event $R = 1$, and it has been explored through many equivalent terms or variants referred to as, for instance, demographic parity or statistical parity, since it enforces groups to have equal selection rates. Thus –in terms of probability– it corresponds to the following constraint:

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b) = ...$$

If A is binary (that is, A = a_1 or a_2), then we can compute the Independence unfairness measure as:

$$\mathfrak{U}_I(a_1, a_2) = |P(R = 1 \mid A = a_1) - P(R = 1 \mid A = a_2)|$$

Separation criterion*:* roughly speaking, since in many scenarios the sensitive characteristic may be correlated with the target variable, the separation criterion allows correlation between the score and the sensitive attribute to the extent that it is justified by the target variable, reason why it is also said equalized odds, equality of opportunity, or even conditional procedure accuracy. Specifically, the separation criterion requires the equivalence of true positive rate and false positive rate for each level of the protected attributed under analysis:

$$P(R = 1 \mid Y = 1, A = a_1) = P(R = 1 \mid Y = 1, A = a_2) = ...$$

$$P(R = 1 \mid Y = 0, A = a_1) = P(R = 1 \mid Y = 0, A = a_2) = ...$$

Therefore, if A is binary we can compute two Separation unfairness measures ($\mathfrak{U}$) as follows:

- $\mathfrak{U}_{S\_TPR}(a_1, a_2) =$

$$|P(R = 1 \mid Y = 1, A = a_1) - P(R = 1 \mid Y = 1, A = a_2)|$$

- $\mathfrak{U}_{S\_FPR}(a_1, a_2) =$

$$|P(R = 1 \mid Y = 0, A = a_1) - P(R = 1 \mid Y = 0, A = a_2)|$$

Both the definition of the Independence criterion and of the Separation criterion can be easily extended to the case of non-binary attributes – i.e. $m > 2$ – by taking the mean of indexes computed considering all the possible pairs of levels in A:

$$\mathfrak{U}(a_1, ..., a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathfrak{U}(a_i, a_j)$$

Note that all the unfairness measures described herein range in the interval $[0, 1]$: they assume values equal to zero for a perfectly fair classification and higher values for unfair behavior.

Finally, since the score variable was not included in the original datasets, after defining a logistic regression model we assessed the unfairness on the test set, whereas we computed the balance measures on the training set of the model.

IV. Results and Discussion

First of all, we remark that hereinafter the values of both balance measures and fairness criteria have been multiplied by 100, so that all measures range in the interval [0,100], in order to simplify the readability of the results. Hence, we remind that:

- in the case of *balance measures*, values close to 0 indicate a high imbalance, vice-versa the closer the measure to 100 and the higher the balance;
- in the case of *unfairness measures*, values close to 0 reveal a fair classification, on the contrary, high values indicate unfair behavior.

Before addressing the *Research Question*, we observe the behavior of both balance measures and fairness criteria as the permutation of a specific mutation varies –for each of the four mutations corresponding to the exemplar distributions.

Given a certain mutation, we note that the values of the balance measures remain substantially unchanged for all six permutations, suggesting that permutations have a very weak effect or no effect at all on the balance measures.

On the contrary, regarding the fairness criteria, we observe an irregular behavior particularly in the case of mutations that lead to more imbalanced distributions –Power2, HalfHigh and OneOff–, while the values tend to be more stable for QuasiBalance; thus, permutations result to have some effect on the measures of unfairness.

A. Analysis of the Balance measures in response to mutations.

With a view to analyzing more in-depth the behavior of the indexes, we report in Fig. 1 the box plots of the whole distributions for each balance measure with respect to mutations. Keeping in mind the description of the exemplar distributions in section III-B, we expect balance measures to increase as the mutation tends to be increasingly balanced. For this aim, we remind that the most imbalanced distribution is represented by Power2, followed by HalfHigh (which is slightly more balanced with respect to Power2), OneOff (slightly more balanced again), QuasiBalance and Balance –which is the best case, with all the occurrences equally distributed between the classes. Indeed, we note an overall absence of variance and we observe that balance measures increase as the mutations become increasingly balanced, with the lowest values in correspondence of the case Power2, respectively followed by HalfHigh (which presents higher values with respect to the previous, indicating a more balanced distribution) and OneOff (with even higher values); then, we observe the highest outcomes corresponding to the cases QuasiBalance and Balance, confirming our general expectations.

Looking at the individual measures, Gini and Shannon indexes present a similar behavior, with values in the range between 75 and 100, and apparently no difference in detecting QuasiBalance and Balance, both with values close to 100. The Simpson index covers a larger range, about 38-100, with a slight difference between the cases QuasiBalance and Balance. Finally, the IR index appears to be spanned over the whole range [0,100], with well distinct values for the two most balanced cases, and the uncommon presence of zero values in correspondence of the mutation OneOff: indeed, by definition of IR, in the special case of one or more *empty* classes[6] the value of the IR index results to be zero, that is the reason for which we observe null values in the case of OneOff.

Therefore, we can confirm the ability of the mutation approach to generate synthetic datasets that spread the whole range of conventional balance measures.

B. Analysis of the Fairness criteria in response to mutations.

An analogous analysis has been performed for the unfairness measures and is reported in Fig. 2, which presents the box plots of the whole distributions for each fairness criterion in correspondence of the five mutations. First of all, we observe that the variance decreases as the mutations tend to be more and more balanced, with a very large variance in the cases Power2 and HalfHigh; then, the variance tends to drop in the intermediate case OneOff, and becomes much smaller for QuasiBalance and Balance. Such variance trend is substantially the same for all the unfairness measures, but looking at the individual measures, we observe that the Separation criterion in the case of TP rate assumes values in the range [0,23], while it assumes values between 0 and 4 in the case of FP rate; finally, we observe that the Independence criterion ranges in the interval [0,7].

Despite the different ranges of values, we note that all the unfairness measures present overall very similar distributions with respect to mutations: we observe the highest values in correspondence of Power2 (thus indicating the most unfair

[6]We define as *empty class* a class with null frequency as there are no occurrences, i.e., a class that *exists* because potentially there could be occurrences, but is *not* represented in the dataset.

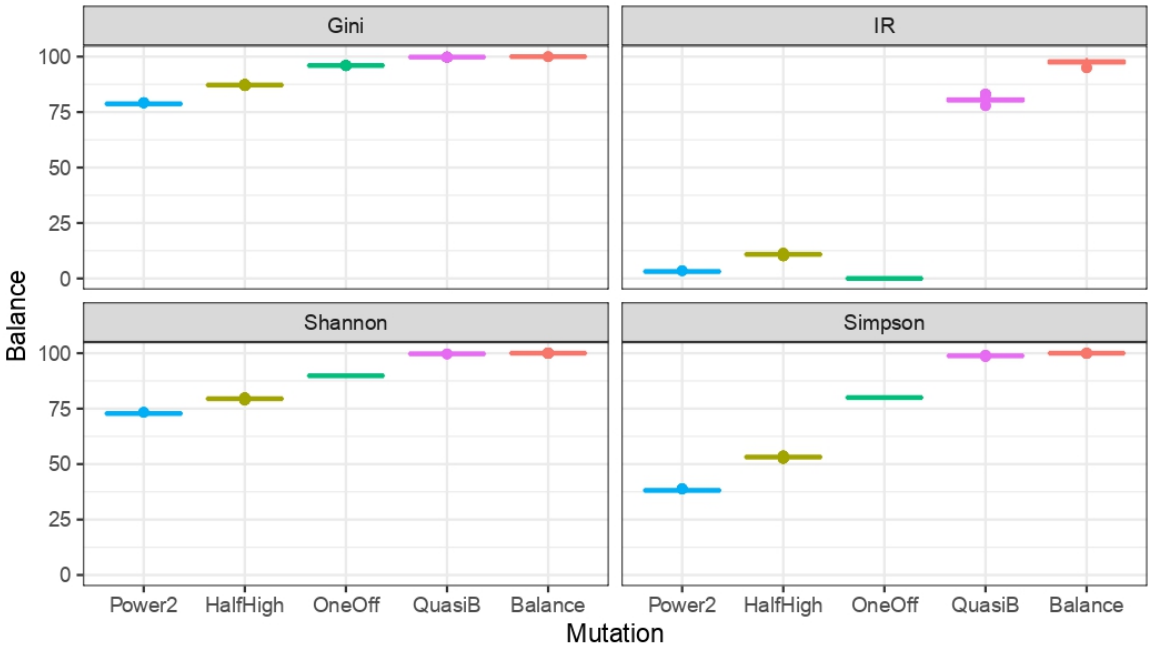


Fig. 1. Distributions of the Balance measures with respect to mutations.

classification output), followed by HalfHigh, OneOff, QuasiBalance and Balance which all present lower values compared to the case Power2, revealing a fairer classification output, but substantially no difference between the mean values. In conclusion, only on the condition of considering the highest extreme values of each mutation, the general trend of the unfairness measures seems to decrease (thus indicating an increasingly fair classification) as the mutations tend to be increasingly balanced.

C. Analysis of the Fairness criteria in response to the Balance measures.

In the subsequent analysis we examine the trends of the *fairness criteria* in response to the *balance measures* with respect to the different mutations, by considering (for each mutation) first the mean value of *Unfairness*, as reported in Fig. 3(a), and then the maximum value, which is represented in Fig. 3(b). Particularly, we aggregate data for mutation and we plot the distributions of the three fairness criteria (Y axis) with respect to the increase of the balance measures (X axis); therefore, the dashed lines trace the trend of the unfairness measures as the balance measures increase. We also specify that regarding the balance measures we always consider the mean values for each mutation (as in the previous analysis of Fig. 1 we observed an absence of variance); whereas, concerning the unfairness measures, after aggregating data for mutation, we first compute the mean values ("Mean case") and then we take the maximum values ("Worst case", which corresponds to the most unfair output for that given mutation), since we previously observed in Fig. 2 a large variance, above all in correspondence of highly imbalanced distributions. Overall, we note *a decrease in the unfairness measures as the balance measures increase*.

This trend is confirmed in both the Mean and the Worst cases, but looking at the individual indexes of balance we observe an irregular behavior for the IR index: indeed, we already explained the special case of one or more classes with null frequency that make the IR index drop to zero, therefore in correspondence of the mutation OneOff the IR index results to be zero, while the unfairness level assumes an intermediate value between the mutation HalfHigh and QuasiBalance, thus reflecting the same order of the *unfairness levels* in response to the other balance measures. In turn, we observe that the unfairness measures decrease –thus indicating an increasingly fair classification– as the mutations tend to be increasingly balanced, with the highest values in the case of Power2, respectively followed by HalfHigh (which presents lower values compared to the previous, revealing a fairer classification output) and OneOff (with even lower values); then, the lowest values are obtained in the cases QuasiBalance and Balance, thus indicating the fairest output.

To analyze results more in depth, we integrate our study with the computation of the Spearman correlation coefficient between *balance* and *unfairness* measures. Specifically, we expect the coefficient to be negative, as we expect the balance measures to be high (meaning low imbalance) if the unfairness values are low (indicating higher fairness). Thus, the stronger the negative correlation, the stronger is the relationship between balance and unfairness measures.
As we can observe from table III, all the balance measures present a negative correlation with the fairness criteria, meaning that the higher the indexes of balance, the lower the

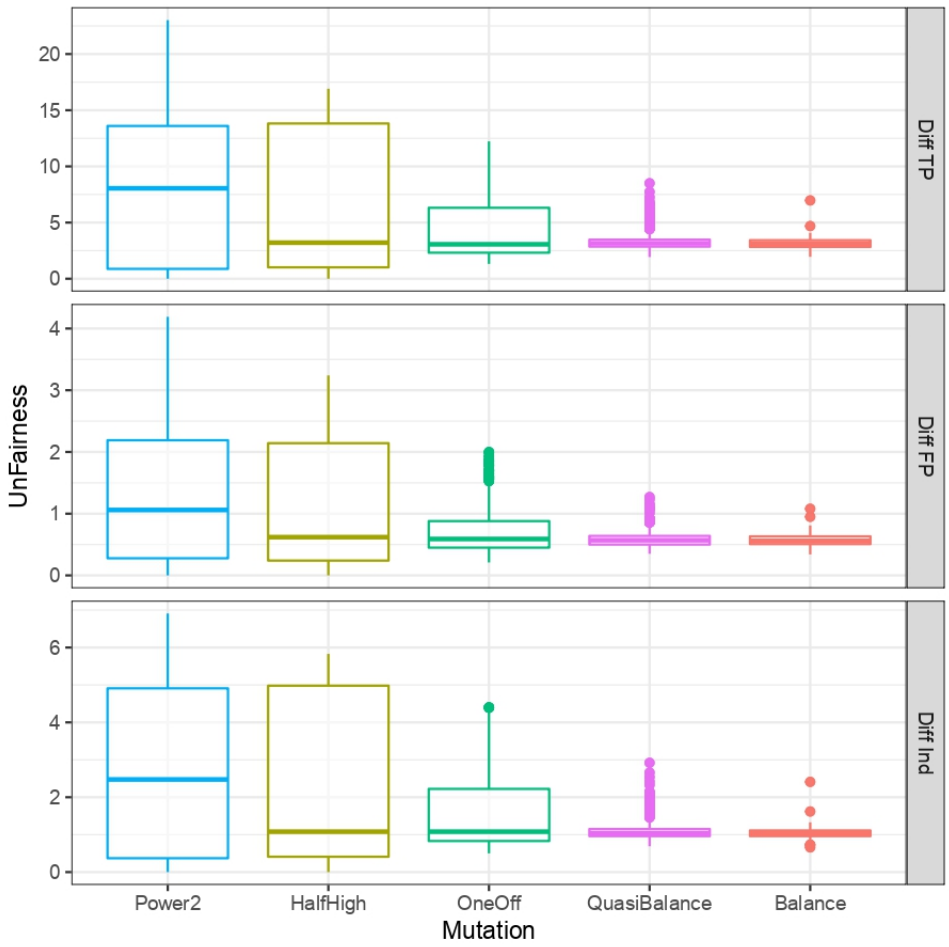


Fig. 2. Distributions of the Unfairness measures with respect to mutations.

unfairness measures; in addition, the computations reveal that such values are all significant (p-value$<$0.05) except for the IR index in correspondence of Separation TPR. More in detail, we notice that the Imbalance Ratio index always presents a weaker negative correlation (between -0.018 and -0.049) with respect to the other three balance measures, which seem to reflect very similarly the different unfairness measures; specifically, the more accurate balance measure is the Shannon index, followed by Gini and Simpson indexes respectively, each one with correlation values between -0.08 and -0.1.

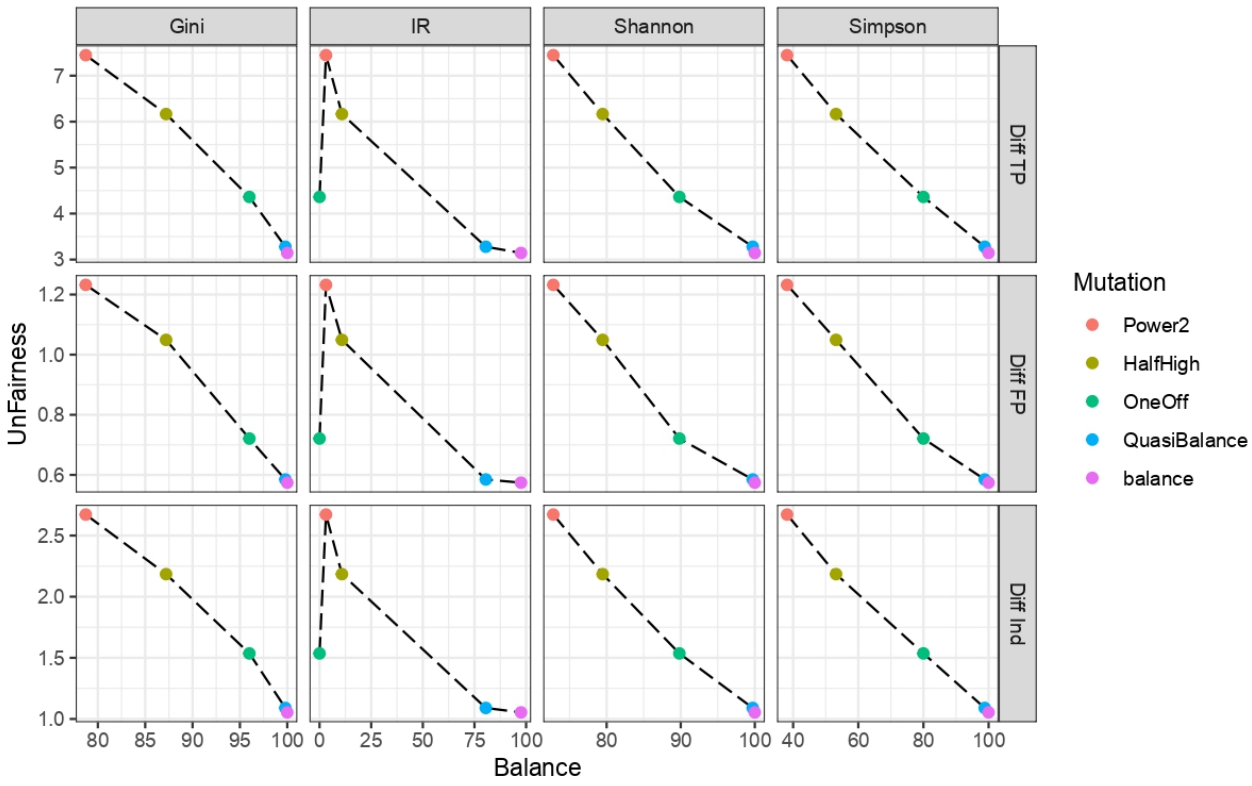
From the perspective of the unfairness measures, the Separation criterion in the case of True Positive rate results to be the most difficult to detect (with correlation values around -0.08, and even -0.018 in correspondence of the IR index), followed by the Independence criterion –which presents a slightly stronger negative correlation–, while the Separation criterion in the case of False Positive rate appears to be the best to detect, showing a stronger negative correlation above all with the Gini, Shannon and Simpson indexes (with correlation values around -0.1).

Therefore, although correlations are weak, they are always negative (and significant, i.e. p-value$<$0.05, except for IR w.r.t. Separation TPR): we can assert that overall the correlation analysis do not reject the hypothesis that the balance measures are capable of revealing unfairness of software output, with some variation among the balance measures (e.g., we observed halved correlation values in correspondence of the IR index, which is highly sensitive to extreme values of balance and imbalance).

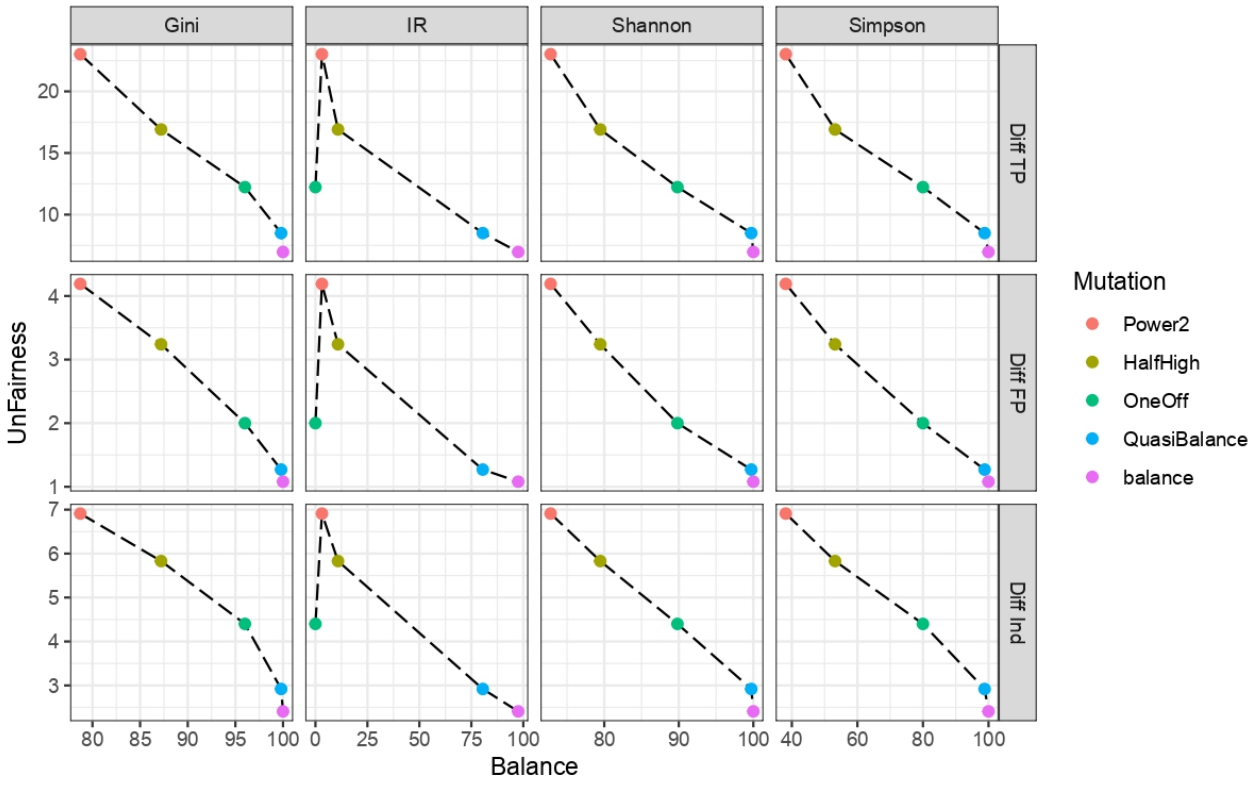
In conclusion, on the basis of all the highlighted observations and within the limits of this study, we positively answer our initial research question: it is possible to identify the risk of unfairness in a classification output by detecting the level of (im)balance in the input data.

V. Related work

We contribute to the main corpus of researches on algorithmic bias and fairness by moving the focus from the outcomes of ADM systems (where most of the literature concentrate) to their inputs and processes, as indicated as necessary in several recent studies (.e.g, [14], [15] and [16]). Our proposal differentiates from the reference literature for two additional important aspects: i) it is built upon a series of international standards, which incorporate by design a multi-



(a) Mean case of unfairness.



(b) Worst case of unfairness.

Fig. 3. Trends of the *Fairness criteria* in response to the *Balance measures* with respect to the different mutations, by considering (for each mutation) the Mean value of *unfairness* (a), and then the maximum value –which corresponds to the most *unfair* output– i.e., the Worst case (b).

TABLE III
CORRELATION BETWEEN BALANCE MEASURES AND UNFAIRNESS MEASURES.

Fairness criteria \ Balance Measures	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.088	-0.089	-0.087	-0.046
Separation (TPR)	-0.083	-0.084	-0.082	-0.018
(FPR)	-0.102	-0.103	-0.100	-0.049

stakeholder perspective; ii) we examine the balance features of input datasets by looking at data imbalance as a risk factor, and not as a technical fix. Indeed, we firmly believe that a risk approach is more advisable because it keeps the ultimate responsibility in the realm of human agency, rather than delegating the mitigation of the issue to yet another algorithm, with a very low probability of success given the socio-technical nature of the problem.

Noteworthy it is the work of Takashi Matsumoto and Arisa Ema [17]: they adopted an approach similar to ours but with a wider scope, by proposing a risk chain model for risk reduction in Artificial Intelligence (AI) services, named RCM, where they consider both data quality and data imbalance as risk factors. Even though our work is not as wide in scope, we believe that it can be easily plugged into the RCM framework, due to the fact that we propose a quantitative way to measure balance, backed by a structural relation to the ISO/IEC standards on software quality requirements and risk management. In addition, our work is complementary to the existing toolkits for bias detection and mitigation [18], since the balance measures proposed herein have not been taken into account yet.

Other approaches that can be related to ours are in the direction of labeling datasets. The "The Dataset Nutrition Label Project" [7] has been an inspiring work for us. Similar to nutrition labels on food, this initiative aims to identify the "key ingredients" in a dataset such as provenance, population, missing data.

A similar goal was declared by authors of the "Ethically and socially-aware labeling" (EASAL) [19], who identified three types of data input properties that could lead to downstream potential risks of discrimination: data quality, correlations and collinearity, and disproportions in datasets. The last property coincides with imbalanced data: indeed, the same authors lately published a data annotation and visualization schema based on Bayesian statistical inference [20], always for the purpose of warning about the risk of discriminatory outcomes of a given dataset.

VI. LIMITATIONS

As limitations of our approach, first of all we highlight the limited amount of data that has been taken into account, as we tested the mutation technique on just one multiclass protected attribute belonging to a specific dataset: it would be advisable to retrieve a larger number of datasets with all the concerning information, with the aim of further assessing the reliability of this approach. Secondly, it would be recommended to take into consideration additional measures of balance (also for non-categorical data) as well as other fairness criteria. Both directions would help to generalize the validity of the findings of this study.

The binomial logistic regression used for the classification task assumes linearity between the dependent variable and the independent variables, and limited or no multi-collinearity between independent variables. These requirements were not taken into account and verified in our analyses. In addition, other classification algorithms (each with different parameters) could be applied to understand if they propagate imbalance differently from the logistic regression, and in general to improve the external validity of this study.

Eventually, other types of mutation techniques should be taken into account, for instance by adopting different pre-processing methods to reproduce several distributions of the occurrences between the classes of the protected attributes.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we assess imbalance in a given dataset as a potential risk factor for detecting discrimination in the classification outcome of automated decision-making systems. The approach combines aspects of data quality and risk management (backed by ISO/IEC standards). For this purpose, we selected four balance measures (the Gini, Shannon, Simpson, and Imbalance Ratio indexes, normalized to share the same meaning and the same range of values) and we tested their ability to anticipate discrimination occurring in the classification output. Overall, the results reveal that the proposed approach is suitable for the given goal, however further work ought to be devoted to testing more systematically both balance measures and fairness criteria on a larger number of datasets and protected attributes, to derive usage guidelines for each index, given the different behaviors observed in the study. In addition, the use of different classification algorithms and mutation techniques could further enrich this study.

We hope that these preliminary results will encourage researchers and policy-makers to assess the risk of discrimination in ADM systems by measuring the imbalance of the protected attributes in training sets, hopefully adopting the approach we proposed here and improving it with additional measures and techniques.

[7]It is the result of a joint initiative of MIT Media Lab and Berkman Klein Center at Harvard University: https://datanutrition.org/

REFERENCES

[1] F. Chiusi, S. Fischer, N. Kayser-Bril, and M. Spielkamp, "Automating Society Report 2020," https://automatingsociety.algorithmwatch.org, Berlin, Oct. 2020.

[2] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," https://papers.ssrn.com/abstract=2477899, Rochester, NY, 2016.

[3] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, Jan. 2018.

[4] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[5] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[6] International Organization for Standardization, "ISO/IEC 25000:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE," https://www.iso.org/standard/64764.html, 2014.

[7] ——, "ISO 31000:2018 Risk management — Guidelines," https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/56/65694.html, 2018.

[8] A. Vetrò, "Imbalanced data as risk factor of discriminating automated decisions: a measurement-based approach," *Journal of Intellectual Property, Information Technology and Electronic Commerce Law - Special issue on Impact of Technological Advances on Individuals: Interaction of Law & Informatics*, vol. 4, pp. 331–347, 2021.

[9] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Trans. Inf. Syst.*, vol. 14, no. 3, pp. 330–347, Jul. 1996.

[10] E. U. A. for Fundamental Rights, "EU Charter of Fundamental Rights - Article 21 - Non-discrimination," https://fra.europa.eu/en/eu-charter/article/21-non-discrimination, December 2007.

[11] U. M. Learning, "Default of Credit Card Clients Dataset," https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset, 2016.

[12] A. Vetrò, M. Torchiano, and M. Mecati, "A data quality approach to the identification of discrimination risk in automated decision making systems," *Government Information Quarterly*, vol. 38, no. 4, 2021.

[13] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[14] D. Firmani, L. Tanca, and R. Torlone, "Ethical dimensions for data quality," *Journal of Data and Information Quality (JDIQ)*, vol. 12, no. 1, pp. 1–5, 2019.

[15] E. Pitoura, "Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias," *Journal of Data and Information Quality*, vol. 12, no. 3, pp. 12:1–12:8, Jul. 2020.

[16] B. Hutchinson and M. Mitchell, "50 Years of Test (Un)fairness: Lessons for Machine Learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 49–58.

[17] T. Matsumoto and A. Ema, "RCModel, a Risk Chain Model for Risk Reduction in AI Services," http://arxiv.org/abs/2007.03215, Jul. 2020.

[18] M. S. A. Lee and J. Singh, "The landscape and gaps in open source fairness toolkits," 2020.

[19] E. Beretta, A. Vetrò, B. Lepri, and J. C. De Martin, "Ethical and Socially-Aware Data Labels," in *Information Management and Big Data*, J. A. Lossio-Ventura, D. Muñante, and H. Alatrista-Salas, Eds. Cham: Springer International Publishing, 2019, pp. 320–327.

[20] E. Beretta, A. Vetrò, B. Lepri, and J. C. De Martin, "Detecting discriminatory risk through data annotation based on bayesian inferences," 2020.