Pattern Recognition Letters 154 (2022) 29–36

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A fuzzy-rough uncertainty measure to discover bias encoded explicitly or implicitly in features of structured pattern classification datasets

Gonzalo Nápoles [a,*], Lisa Koutsoviti Koumeri [b]

[a] Department of Cognitive Science & Artificial Intelligence, Tilburg University, The Netherlands
[b] Business Informatics Research Group, Hasselt University, Belgium

ARTICLE INFO

Article history:
Received 1 August 2021
Revised 4 January 2022
Accepted 7 January 2022
Available online 10 January 2022

Edited by Maria De Marsico

Keywords:
Bias
Fairness
Explainable machine learning
Fuzzy-rough sets

ABSTRACT

The need to measure bias encoded in tabular data that are used to solve pattern recognition problems is widely recognized by academia, legislators and enterprises alike. In previous work, we proposed a bias quantification measure, called fuzzy-rough uncertainty, which relies on the fuzzy-rough set theory. The intuition dictates that protected features should not change the fuzzy-rough boundary regions of a decision class significantly. The extent to which this happens is a proxy for bias expressed as uncertainty in a decision-making context. Our measure's main advantage is that it does not depend on any machine learning prediction model but a distance function. In this paper, we extend our study by exploring the existence of bias encoded implicitly in non-protected features as defined by the correlation between protected and unprotected attributes. This analysis leads to four scenarios that domain experts should evaluate before deciding how to tackle bias. In addition, we conduct a sensitivity analysis to determine the fuzzy operators and distance function that best capture change in the boundary regions.

1. Introduction

Data-driven decision support systems have been accused of being a fertile ground to produce biased results, thus leading to discriminatory decisions [1]. As historical data often encode biases [2] explicitly or implicitly [3], pattern recognition algorithms inevitably relate their predictions with protected characteristics such as race or gender. The Equality Act 2010 of government of the United Kingdom defines protected attributes as personal characteristics (such as gender or race) that should not put a person at a substantial disadvantage compared to people with different personal characteristics. In the literature, more than 20 definitions of fairness [4] and respective bias metrics have been proposed. However, existing metrics express different and often contradictory notions of fairness [5–7] depending on local legal and cultural conventions [8] or on the type of decision-support system [1]. Deciding which metric is most appropriate for the task at hand is difficult [9] as several parameters need to be considered such as causal influences among features, mis-representation of groups and different modalities of data [4]. Therefore, the need for introducing general-purpose, direct and indirect bias measures is evident [8].

The literature on this subject relies on two dominating notions of fairness: group-based and individual-based fairness measures [10]. Group-based measures have been criticized for leading to inverse discrimination [11] and being oblivious to features other than the sensitive feature [6,12]. Moreover, they often require discretization of numeric sensitive features such as age, which can alter bias measures' outputs [7]. Individual-based fairness measures require strong assumptions such as the availability of an agreed-upon similarity metric, or knowledge of the underlying data generating process [13]. These measures act as bias proxies as they do not measure bias directly. For example, they can rely on the consistency in classification or the redundancy in data. Finally, both groups of measures are often applied on predictions generated by black-box machine learning models for fairness assessment [14]. However, most successful prediction models are not intuitively explainable [15] and tend to be sensitive to variations in the input arising from variations in training-test splits [7].

Another sensitive issue refers to implicit bias or indirect discrimination, which occurs when decisions are made based on nonsensitive features that strongly correlate with biased sensitive ones [8]. This means that even if protected features are excluded from the decision making process, a classification algorithm might still produce biased results. Existing implicit bias measures are found in [8,16] where background knowledge is used to manually set classification rules combined with discriminatory thresh-

* Corresponding author.
E-mail addresses: g.r.napoles@uvt.nl (G. Nápoles), lisa.koutsoviti@uhasselt.be (L. Koutsoviti Koumeri).

https://doi.org/10.1016/j.patrec.2022.01.005

olds. The possible pitfall in such an approach is that human experts might misjudge the impact of feature categories on the decision outcomes [17].

Recently, we proposed a measure called fuzzy-rough uncertainty (FRU) to quantify explicit bias of protected features in pattern classification problems [18]. Our measure quantifies the changes in the fuzzy-rough boundary regions after removing a protected feature as a proxy for measuring fairness. To that end, we use the advantages of rough sets [19] for analyzing inconsistency in decision systems. Measuring the distance or the change between the regions of fuzzy-rough sets has been examined in the literature [20,21], but not in the context of bias quantification, as far as we know. To cope with the issue of defining similarity thresholds when handling problems involving continuous features, we use fuzzy-rough sets as defined by [22]. This mathematical theory allows computing membership values that express the extent to which instances belong to each information granule [23]. The intuition behind FRU is that, in fair decision-making scenarios, removing a protected feature should not cause big changes in the decision boundaries. The extent to which that happens can be used to quantify the explicit bias attached to a given protected feature.

While the FRU measure brings the added value that it does not rely on any prediction model but information granules derived from the data, it cannot capture implicit bias. For example, if a protected feature is correlated with an unprotected one, its removal might not cause significant changes to the boundary regions. This suggests that we should analyze the FRU values together with existing correlation/association patterns between protected and unprotected features. Another issue that cries for further research is the impact of fuzzy operators and distance functions on the performance of our measure.

Motivated by these two research gaps, our paper brings three main contributions. Firstly, we illustrate how the FRU is able to capture explicit bias while state-of-the-art individual-based measures struggle to capture the effect of removing a protected feature. For simulation purposes, we use the *German Credit* data set [24], which classifies loan applicants in terms of creditworthiness and is widely used in the context of AI Fairness [14]. Secondly, we conduct a sensitivity analysis to study the impact of fuzzy operators and distance functions on the FRU results. Such a study led to recommended parametric settings that can be adopted for other datasets (as reported in the supplementary materials). Finally, we discuss four scenarios that relate the changes in the boundary regions (after removing a protected feature) with the correlation/association between protected and unprotected features [25] as a way to detect implicit bias.

The remainder of the paper is organized as follows. The next section introduces the mathematical formalism behind the computation of the fuzzy-rough regions from data. Section 3 describes the similarity function we deployed and the proposed bias quantification measure. Section 4 presents the experimental setup and analyzes the measures' outputs. Finally, Section 5 discusses possible implications to the field.

2. Fuzzy-rough set theory

This section presents the FRS theory as described by [22]. This theory is used to transform tabular data into information granules characterizing each decision class. The output of this fuzzy granulation process is membership values, which will be used to define our bias quantification measure.

Let us assume that we have a universe of discourse U, a fuzzy set $X \in U$ and a fuzzy binary relation $R \in Q(U \times U)$ such that $\mu_X(x)$ and $\mu_R(y,x)$ are their membership functions, respectively. The membership function $\mu_R : U \to [0,1]$ determines the degree to which $x \in U$ is a member of X, whereas $\mu_R : U \times U \to [0,1]$ denotes the degree to which y is considered to be a member of X from the fact that x is a member of the fuzzy set X. Whenever opportune, $R(x)$ is denoted with its membership function $\mu_{R(x)}(y) = \mu_R(y,x)$.

Firstly, let us build a partition of U according to the decision classes. The X_k set contains all objects associated with the k-th decision class. The membership degree of $x \in U$ to a subset X_k was computed using the following hard membership function: $\mu_{X_k}(x) = 1$ for $x \in X_k$ and $\mu_{X_k}(x) = 0$ for $x \notin X_k$, as we assume that all problem instances are correctly labeled.

Secondly, we need to define a fuzzy binary relation $\mu_R(y,x)$ to determine the fuzzy similarity between instances x and y. This function should combine the membership degree $\mu_{X_k}(x)$ with the similarity degree $\phi(x,y)$ between two objects $x, y \in U$. Overall, we define $\mu_R(y,x) = \mu_{X_k}(x)\phi(x,y)$. In the next section, we will give more details about the similarity function, which is expressed in terms of a distance function.

Aiming at defining the lower approximations, we use the degree of x being a member of X_k under the knowledge R. This can be measured by the truth value of the statement '$y \in R(x)$ implies $y \in X_k$' under fuzzy sets $R(X)$ and X_k. We use a necessity measure $inf_{y \in U} \mathcal{I}(\mu_R(y,x), \mu_{X_k}(y))$ with a fuzzy implication function $\mathcal{I} : [0,1] \times [0,1] \to [0,1]$ such that $\mathcal{I}(0,0) = \mathcal{I}(0,1) = \mathcal{I}(1,1) = 1$ and $\mathcal{I}(1,0) = 0$. It also holds that $\mathcal{I}(.,a)$ decreases and $\mathcal{I}(a,.)$ increases, $\forall a \in [0,1]$. Equation (1) displays the membership function for the lower approximation $R_*(X_k)$ associated with the k-th decision class,

$$\mu_{R_*(X_k)}(x) = min\{\mu_{X_k}(x), inf_{y \in U} \mathcal{I}(\mu_R(y,x), \mu_{X_k}(y))\}. \quad (1)$$

To derive the upper approximations, we measure the truth value of the statement '$\exists y \in U$ such that $x \in R(y)$' under fuzzy sets $R(x)$ and X_k. The true value of this statement can be obtained by a possibility measure $sup_{y \in U} \mathcal{T}(\mu_R(x,y), \mu_{X_k}(y))$ with a conjunction function $\mathcal{T} : [0,1] \times [0,1] \to [0,1]$ such that $\mathcal{T}(0,0) = \mathcal{T}(0,1) = \mathcal{T}(1,0) = 0$ and $\mathcal{T}(1,1) = 1$, where both $\mathcal{T}(.,a)$ and $\mathcal{T}(a,.)$ increase, $\forall a \in [0,1]$. Equation (2) displays the membership function for the upper approximation $R^*(X_k)$ associated with the k-th decision class,

$$\mu_{R^*(X_k)}(x) = max\{\mu_{X_k}(x), sup_{y \in U} \mathcal{T}(\mu_R(x,y), \mu_{X_k}(y))\}. \quad (2)$$

This model takes the minimum between $\mu_{X_k}(x)$ and $inf_{y \in U} \mathcal{I}(\mu_R(y,x), \mu_{X_k}(y))$ when calculating $\mu_{R_*(X_k)}(x)$, and the maximum between $\mu_{X_k}(x)$ and $sup_{y \in U} \mathcal{T}(\mu_R(x,y), \mu_{X_k}(y))$ when calculating $\mu_{R^*(X_k)}(x)$ to preserve the inclusiveness of $R_*(X_k)$ in X_k and the inclusiveness of X_k in $R^*(X_k)$.

Finally, we define the fuzzy-rough regions using the upper and lower approximations. The membership functions for the fuzzy-rough positive, negative and boundary regions can be defined as $\mu_{POS(X_k)}(x) = \mu_{R_*(X_k)}(x)$, $\mu_{NEG(X_k)}(x) = 1 - \mu_{R^*(X_k)}(x)$ and $\mu_{BND(X_k)}(x) = \mu_{R^*(X_k)}(x) - \mu_{R_*(X_k)}(x)$, respectively. Membership values to positive regions indicate the extent to which the instances belong to a decision class, membership values to negative regions indicate the extent to which the instances do not belong to a decision class, whereas membership values to boundary regions indicate the extent to which the instances are uncertain to the problem at hand.

3. Fuzzy-rough uncertainty measure

This section introduces our measure to quantify bias in tabular datasets used for pattern classification. This measure assumes that experts can determine the set of protected features (i.e., those likely related to bias) beforehand. The intuition of our measure is that a protected feature should not have a leading role on the decision process. For example, let us assume that we have a problem described by several features where *Gender* is deemed a protected feature. If we remove that feature and there is an increase in the

misclassifications, then one could conclude that *Gender* is relevant to separate the decision classes. The extent to which the decision boundaries become less separate can be understood as a bias indicator.

Before presenting our measure, let us describe the similarity function [26] used to compare the instances. Such a function will be derived from a normalized heterogeneous distance function. In particular, we will employ two distance functions: the Heterogeneous Manhattan-Overlap Metric (HMOM) [27] and the Heterogeneous Euclidean-Overlap Metric (HEOM) [27] because of their ability to deal with instances having mixed-type features. Equation (3) portrays the similarity function, which produces values in the (0,1) interval,

$$\phi(x,y) = e^{-\lambda\left(d(x,y)\right)} \tag{3}$$

where $\lambda > 0$ is a user-specified smoothing parameter to avoid saturation problems in which similarity values have low variability even for quite dissimilar instances.

The *fuzzy-rough uncertainty* (FRU) measure [18] quantifies how much the absence of the protected feature f_i modifies the fuzzy-rough boundary regions. If the difference is positive, we can conclude that the boundary regions became bigger after removing the protected feature, so there is more uncertainty (i.e., the feature was important for the classification). If the difference is negative, we can conclude that the boundary regions became smaller after removing the protected feature, so there is less uncertainty (i.e., the feature was causing uncertainty and its removal might be convenient).

To quantify these differences, we use the membership values of instances in U to the boundary regions using (i) the full set of features, and (ii) the set of features without including the protected feature f_i (denoted by $\neg f_i$). Equation (4) shows how to compute the FRU value associated with the k-th decision class and the i-th protected feature,

$$\Omega_k(f_i) = \frac{\sqrt{\Sigma_{x\in U}(\Delta^+_{B_k\neg f_i}(x))^2}}{\sqrt{\Sigma_{x\in U}(\mu_{B_k}(x))^2}} \tag{4}$$

such that $\Delta^+_{B_k\neg f_i}(x) = \mu_{B_k}(x) - \mu_{B_k\neg f_i}(x)$ when the removal of the i-th feature increases the uncertainty. Otherwise, we will assume that $\Delta^+_{B_k\neg f_i}(x) = 0$. To lighten the notation, we denote the k-th boundary region $\mu_{BND(X_k)}(x)$ with $\mu_{B_k}(x)$. Notice that the FRU measure is normalized by dividing by the fuzzy cardinality of the fuzzy-rough boundary region, thus leading to relative values that are not likely to be affected by class imbalance. Overall, the proposed granular measure is similar to computing the relevance of the protected feature to preserving the decision boundaries attached to the problem. Recall that in multiclass classification problems, the final FRU measure is the average FRU value of all decision classes.

4. Numerical simulations and discussion

The case study used in our experiments is the *German Credit* dataset, which is used for classifying loan applicants at a bank as credit worthy or the opposite. Based on the literature, protected features are *Age* and *Gender* [14].

Data preprocessing included (1) normalizing numeric features such that their minimum and maximum values are 0.0 and 1.0, respectively, (2) encoding target classes as integer identifiers starting at zero, and (3) re-coding the nominal protected feature *sex&marital status* to include only gender-related information.

Our experiments consist of four parts. The first part involves calculating the FRU values for protected features and comparing them to individual state-of-the-art measures as in [18]. Although our paper studies the bias towards protected features, we also calculate the FRU values of unprotected features for reference. The second part is a sensitivity analysis where we examine variations in the FRU values when changing the parametric settings. In an effort to further explore the behavior of our measure, we test it on three additional datasets. The results are included in the supplementary material due to space limitations. The third part attempts to discover whether bias encoded in protected features might also be encoded in unprotected features implicitly. Finally, the last part compares our FRU measure with group-fairness measures. A separate section for such a comparison is deemed necessary as individual and group fairness measures should not be directly compared.

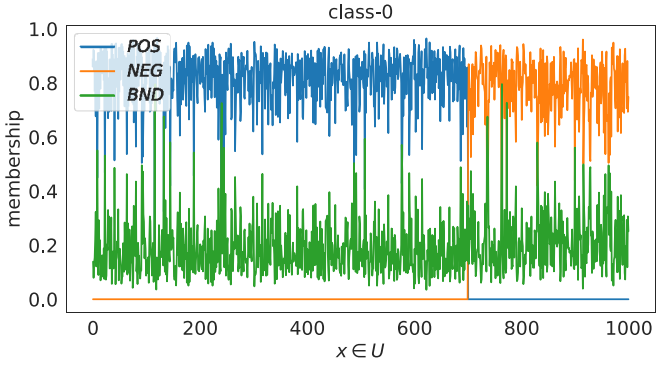

Fig. 1. Membership values to the negative, boundary and positive regions using the complete feature set. The x axis represents the instances and the y axis their respective membership values.

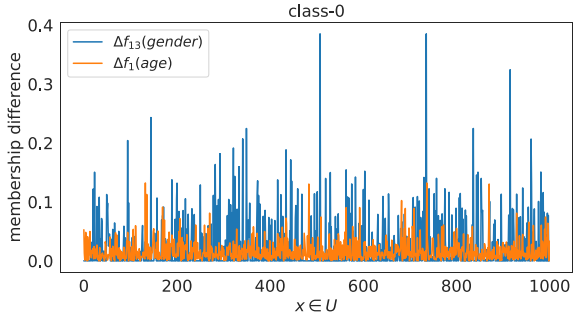

Fig. 2. Difference between membership values to the boundary regions after removing *Age* (Δf_1) and after removing *Gender* (Δf_{13}) per instance.

4.1. Individual fairness metrics and FRU values

The first part starts with the calculation of the FRU values for protected features as in [18]. To do that, we follow a two-step process as mentioned in Section 3. First, the membership values to the positive, negative and boundary fuzzy-rough regions per decision class are computed using the full set of features. Fig. 1 shows these membership values.

The graphs show that the fuzzy-rough regions are relatively distinct from one another while involving dissimilar membership values. Second, the membership values to the regions are computed once again excluding one protected feature from the dataset. In all simulations in this sub-section, we used $\lambda = 0.5$, the Łukasiewics implicator and an arbitrary t-norm. Recall that we are only interested in the positive changes that occur in the membership values to the boundary regions after suppressing a protected feature. These are used to compute the FRU values as in Equation (4). Fig. 2 shows the changes in the membership values the boundary regions per instance.

Table 1
Results of proposed and state-of-the-art measures. The ideal value of CON is one, while for the remaining ones is zero.

Individual fairness metrics			
Feature set	CON	GEI	FRU
F	0.746	0.093	n/a
$F/\{f_1\}$	0.746	0.095	0.107
$F/\{f_{13}\}$	0.743	0.093	0.224

Next, we compute two individual state-of-the-art measures using the aif360.sklearn package [14] and our preprocessed dataset. The first individual fairness metric is the *consistency score* (CON) coupled with a logistic regression model as the underlying predictor. The second individual metric is the *generalized entropy index* (GEI) that relies on a k-nearest neighbors algorithm. Let F denote the set of protected and unprotected features. We are interested in exploring three settings when calculating these metrics: (i) using all features in F, (ii) excluding *Gender* (f_{13}) and (iii) excluding *Age* (f_1). Table 1 shows the outputs of all measures for these settings.

It can be noticed that both CON and GEI measures report roughly the same values in all three settings. The fact that the outputs of the individual fairness measures report very small changes when protected features are removed would suggest that they failed to quantify the bias issue in this problem. In contrast, our FRU measure reports larger changes in the boundary regions when the protected feature *Gender* is excluded compared to *Age*. In other words, greater uncertainty in classification is reported when *Gender* is suppressed. This indicates that *Gender* encodes more bias than *Age*.

We continue with calculating the FRU values for all features to study the protected ones in a wider perspective. We designate this step as level-1 analysis since one feature is suppressed at a time and denote it as $\Omega(f_i)$ where f_i is either protected or unprotected. Observe that these values should not be interpreted as the absolute relevance of each feature in the classification process since we do not analyse the relationships for all possible feature combinations as needed in a feature selection context. Instead, we aim to investigate the extent to which protected features behave similarly to the unprotected ones. This is made possible by dividing each FRU value by the greatest FRU value among all problem features and is defined mathematically in Section 4.2. Table 2 portrays these results. Moreover, we report a correlation measure between each protected feature and the unprotected ones and whether or not the correlation is significant. Further details about the correlation measure will be disclosed in the last sub-section since it will be the tool we will use to detect implicit bias in the dataset.

The results reveal that the feature having the largest FRU value is *Checking account*, which will serve as a reference feature. We notice that *Age*'s FRU value is among the five smallest FRU values and at least three times lower than *Checking account*'s FRU value. *Gender*'s FRU value is among the medium-ranked FRU values and about half of *Checking account*'s FRU value. In this research, the change occurring in the boundary regions (as quantified by the FRU measure) after removing a protected feature is defined as *explicit bias*.

4.2. Sensitivity analysis

Next, we conduct a sensitivity analysis to measure the effect of variations in the following parameters on the measure's outputs: (1) fuzzy operators (the fuzzy implicator and the t-norm) taken from [28,37], (2) the smoothing parameter in the similarity function and (3) the distance function. Tables 3 and 4 list the different combinations to be explored.

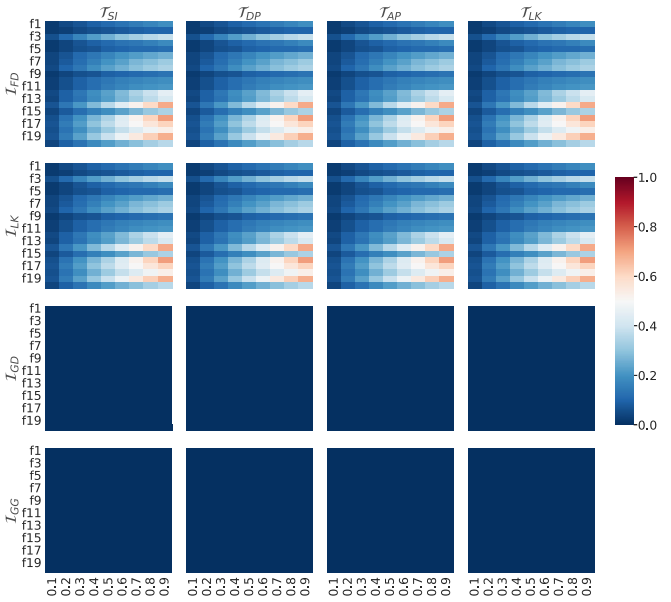

Fig. 3. Effect of the smoothing parameter (x axis), fuzzy conjunction and fuzzy implicator on the FRU values. In these simulations, we use the HMOM distance function. The y axis represents the problem features.

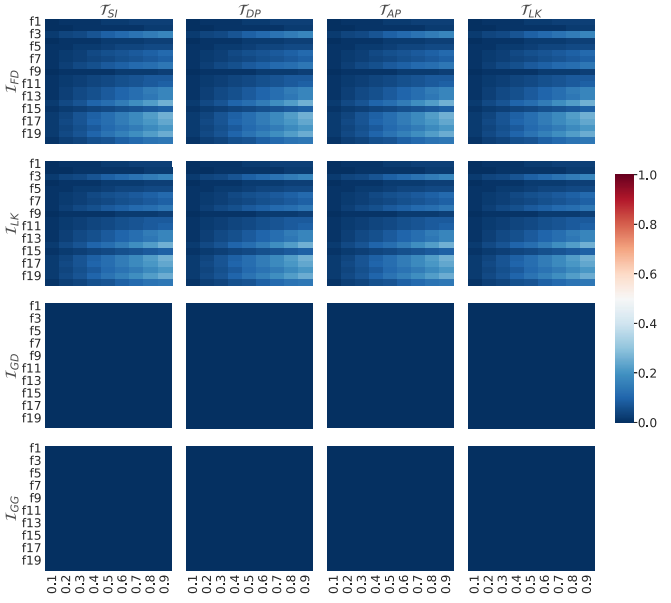

Fig. 4. Effect of the smoothing parameter (x axis), fuzzy conjunction and fuzzy implicator on the FRU values. In these simulations, we use the HEOM distance function. The y axis represents the problem features.

The simulation results displayed in Figs. 3 and 4 show that the choice of the fuzzy implicator has a significant impact on the measure's behavior. However, the FRU measure seems to be invariant to the choice of the fuzzy conjunction operator. Moreover, using Łukasiewics and Fodor as the fuzzy implicator produces the same FRU values, while the rest of the implicators are unable to measure any FRU change at all (as illustrated in the last two rows of Figs. 3 and 4). The same patterns emerge if we use the HEOM distance function, but the changes in the FRU values reported by Łukasiewics and Fodor implicators are much more subtle ranging between 0.0 to 0.03. This confirms our finding that HMOM better captures changes in boundary regions. Łukasiewics is therefore chosen as the fuzzy implicator for the next round of simulations.

Table 2
Correlation/association coefficients between protected and unprotected features, FRU values, ratio between FRU and FRU of reference.

Idx	Features	Corr. with Gender[a]	Corr. with Age[a]	FRU	FRU ratio[b]
f1	Age	0.03*	1.0*	0.11	0.28
f2	Credit amount	0.01*	0.03	0.07	0.18
f3	Credit history	0.12*	0.03*	0.26	0.67
f4	Months	0.01	−0.04	0.11	0.28
f5	Foreign worker	0.04	0.0	0.09	0.23
f6	Housing	0.23*	0.09*	0.17	0.44
f7	Installment rate	0.01	0.06	0.2	0.51
f8	Job	0.09*	0.03*	0.23	0.59
f9	Existing credits	0.01*	0.15*	0.09	0.23
f10	People liable	0.2*	0.01*	0.14	0.36
f11	Other debtors	0.01	0.0	0.15	0.38
f12	Other installment	0.05	0.0	0.23	0.59
f13	Gender	1.0*	0.03*	0.22	0.56
f14	Employment since	0.22*	0.17*	0.36	0.92
f15	Residence since	0.0	0.27*	0.19	0.49
f16	Property	0.09*	0.05*	0.34	0.87
f17	Purpose	0.15*	0.03*	0.37	0.95
f18	Savings account	0.07	0.01	0.32	0.82
f19	Checking account[c]	0.03	0.01	0.39	1.00
f20	Telephone	0.07*	0.02*	0.22	0.56

[a] Asterisks indicate significant p-value ($p < .05$) or F-statistic (larger than the critical value).
[b] Divide FR-Uncertainty value with FR-Uncertainty value of the reference feature.
[c] Reference feature.

Table 3
Fuzzy implicators explored in this paper.

Implicator	Formulation
Fodor	$\mathcal{I}_{FD}(x,y) = \begin{cases} 1 & , x \leq y \\ \max(1-x,y) & , x > y \end{cases}$
Gödel	$\mathcal{I}_{GD}(x,y) = \begin{cases} 1 & , x \leq y \\ y & , x > y \end{cases}$
Goguen	$\mathcal{I}_{GG}(x,y) = \begin{cases} 1 & , x \leq y \\ y/x & , x > y \end{cases}$
Łukasiewicz	$\mathcal{I}_{LK}(x,y) = \min\{1-x+y, 1\}$

Table 4
T-norms explored in this paper.

T-norm	Formulation
Standard intersection	$\mathcal{T}_{SI}(x,y) = \min\{x,y\}$
Algebraic product	$\mathcal{T}_{AP}(x,y) = xy$
Łukasiewicz	$\mathcal{T}_{LK}(x,y) = \max\{0, x+y-1\}$
Drastic product	$\mathcal{T}_{DP}(x,y) = \begin{cases} x & , y = 1 \\ y & , x = 1 \\ 0 & , otherwise \end{cases}$

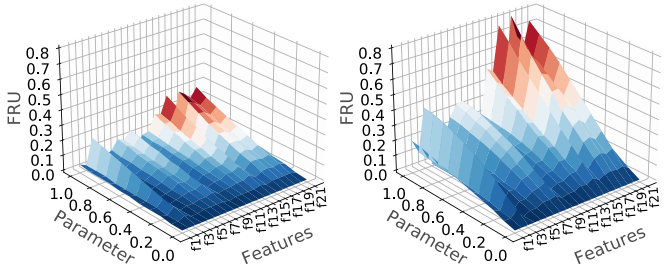

Fig. 5. Effect of the smoothing parameter and distance function on the FRU values. The measure produces larger values when using the HMOM distance function, which aligns well with the expected behavior of these distance functions. Moreover, the larger the smoothing parameter value, the larger the values produced by our measure. $\mathcal{I}_{LK}$ and $\mathcal{T}_{LK}$ are used here.

Fig. 5 offers a three-dimensional view of the FRU values at different smoothing parameter levels (from 0 to 1 with a step of 0.1) per similarity function, using Łukasiewicz both as implication and conjunction operator.

Overall, we observe that the FRU values increase when increasing the smoothing parameters. This means that increasing the smoothing parameter better separates the boundary regions. One should be careful not to confuse those variations with the absolute amount of bias measured. Therefore, we recommend computing FRU values that are relative to the size of the boundary regions instead of using the absolute ones. These relative FRU values can be computed as $\hat{\Omega}_k(f_i) = \Omega_k(f_i)/\Omega_k(f_j)$ with f_j being a reference feature provided that $\Omega_k(f_j) > \Omega_k(f_i)$. This ratio is reported in the last column of Table 2. If no reference feature is available, we can compute the relative FRU values as $\hat{\Omega}_k(f_i) = \Omega_k(f_i)/\sum_j \Omega_k(f_j)$.

4.3. Recommendations to detect implicit bias

In this subsection, we explore whether the bias encoded in the protected features explicitly, might also be implicitly encoded in unprotected features. The intuition is that implicit bias demonstrates itself when pairing two seemingly unrelated concepts [8,29], one of them being a protected feature. Overall, we define implicit bias as the maximal absolute correlation between the protected feature being processed and an unprotected one. Hence, we need to compute the correlation/association between each unprotected and protected feature. For the sake of simplicity, we will refer to both correlation and association as correlation unless specified otherwise.

The correlation patterns are quantified using three different but conceptually sound statistical tools [30,31]. In our study, the Pearson correlation coefficient [32] is used to measure correlation between the numeric protected feature *Age* and the rest of the numeric unprotected features. To do that, we adopt the SciPy Python package [33]. The Cramér's V [34] is used to capture the association strength between the nominal protected feature *Gender* and the unprotected nominal features. Finally, we use the R-squared coefficient of determination [31] to measure the percentage of variation in the numeric unprotected features that is explained by the

Table 5
Scenarios relating correlation and FRU values.

	Large FRU value	Small FRU value
Strong correlation	Explicit & Implicit bias	Implicit bias
Weak correlation	Explicit bias	Safe scenario

protected nominal feature *Gender* coupled with an F-test of joint significance [35]. This measure is computed using the ordinary least squares method from the statsmodels Python package [36]. The selected measures of association are chosen to preserve consistency since they are related to the Pearson's correlation coefficient (even though features do not meet the assumptions of normality, linear dependence or homoscedasticity) [30]. The resulting values are reported in Table 2. The asterisk accompanying each value represents either a p-value lower than 0.05 or an F-statistic larger than the critical value. In short, it refers to the confidence to which the presence or absence of correlation is observed. For example, a correlation coefficient of 0.03* should be understood as no correlation with high confidence.

Table 5 depicts four scenarios that can be derived from the analysis of FRU and correlation values.

The scenario "*weak correlation* and *large FRU value*" means that suppressing the feature causes alterations in the boundary regions. This behavior is defined as explicit bias. The scenario "*strong correlation*" and "*small FRU value*" would indicate implicit bias. In other words, the removal of the protected feature did not change the regions significantly, but the strong correlation suggests that at least an unprotected feature encodes the protected one. The scenario "*strong correlation* and *large FRU value*" might imply both types of bias. Removing a protected feature that is strongly correlated with another might still cause changes in the fuzzy-rough boundary regions. It has not escaped our notice that these scenarios involve rather subjective linguistic terms such as "*strong*" or "*weak*" that should ideally be defined by domain experts.

Let us analyze a potential situation encoding implicit bias. A close inspection at the results in Table 5 reveals that the unprotected features *Residence since* and *Employment since* show the strongest correlation with *Age*. While the correlation values might not be categorized as strong, it is concerning that the largest ones appear associated with unprotected features from which we can roughly infer *Age*. Moreover, we notice that the feature *Employment since* has the second-largest FRU value. When coupling all pieces, we can conclude that Age moderately correlates with unprotected features whose removal causes alterations in the boundary regions.

In order to complement the analysis above, we measure the changes in the fuzzy-rough boundary regions after pairs of protected and unprotected features are excluded simultaneously. We designate this step as level-2 analysis and denote it as $\Omega(f_i, f_j)$ such that f_i is a protected feature and f_j is a unprotected one. As the level-1 analysis might not be enough to discover the role of a protected feature for the problem, we should investigate whether that same feature might become important when combined with an unprotected one. Fig. 6 shows that changes in boundary regions when a single feature is excluded are relatively proportional to the changes occurring when excluded together with a protected feature.

This simulation shows that changes caused by combinations involving *Gender* are larger than those involving *Age*. That confirms the main finding that the results are more biased toward the former than the latter. The results also indicate that the correlation between the excluded features is not strong enough for the boundary regions to remain unchanged. However, the main conclusion from this analysis is that the binary categorization of explicit and implicit bias is too narrow: a protected feature can be important

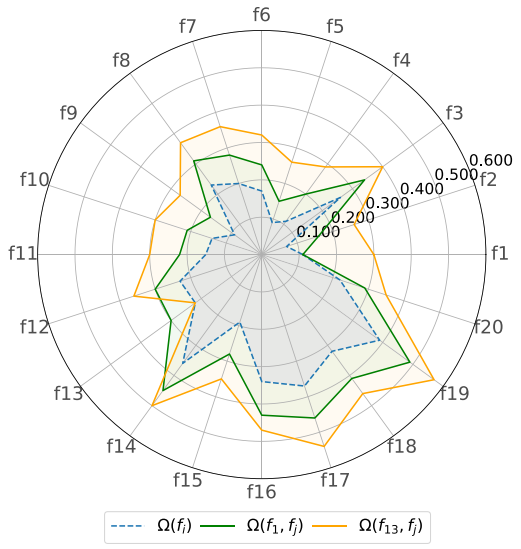

Fig. 6. FRU values when (i) suppressing each feature, (ii) suppressing each feature and *Age* (f_1), and (iii) suppressing each feature and *Gender* (f_{13}).

to some extent by itself when it comes to the boundary regions while also being partially encoded into unprotected features. Such a conclusion paves the road for a new research direction in which explicit and implicit biases are quantified within the fuzzy logic formalism.

Next, we compute the individual baseline measures in same the way as in the previous simulation. Let CON(F) and GEI(F) denote the values of the CON and GEI measures using the whole set of features F describing the problem. Similarly, let CON($F/\{f_i\}$) and GEI($F/\{f_i\}$) denote the values of these measures after suppressing the protected feature f_i from F. Finally, let CON($F/\{f_i, f_j\}$) and GEI($F/\{f_i, f_j\}$) be the values of these measures after removing the protected feature f_i and the unprotected feature f_j from F. The values for the level-1 analysis are computed as $\Delta\text{CON}(f_i) = |\text{ CON}(F) - \text{CON}(F/\{f_i\})|$ and $\Delta\text{GEI}(f_i) = |\text{ GEI}(F) - \text{GEI}(F/\{f_i\})|$. The values for the level-2 analysis are computed as $\Delta\text{CON}(f_i, f_j) = |\text{ CON}(F) - \text{CON}(F/\{f_i, f_j\})|$ and $\Delta\text{GEI}(f_i, f_j) = |\text{ GEI}(F) - \text{GEI}(F/\{f_i, f_j\})|$. We quantify the absolute difference because (i) they better illustrate the different scenarios and (ii) our FRU measure itself is the difference between the fuzzy-rough boundary regions. Figs. 7 and 8 compare the ΔCON and ΔGEI values respectively for both level-1 and level-2 analyses.

The state-of-the-art individual fairness measures report infinitesimal changes as problem features are suppressed, while the changes captured by our FRU measure vary between 0.1 and 0.6. Overall, these figures support our conclusion that literature measures do not capture bias in the same manner as our fuzzy-rough granulation approach.

4.4. Comparison with group-based measures

In an effort to examine bias from different perspectives, we also calculate the state-of-the-art group fairness measures using the aif360.sklearn package [14] and our preprocessed dataset. Prerequisite for computing these measures is discretizing *Age* into people younger and older than 25 years old [14]. Table 6 summarizes the outputs for the state-of-the-art measures along with the results of the FRU measure.

Group fairness measures report slightly larger bias towards *Age* than *Gender*. On the contrary, our FRU measure captures the exact opposite trend which means that it is fundamentally different from existing group fairness metrics. This apparent contradiction

Table 6
Results of proposed and state-of-the-art measures. The ideal value of *Disparate Impact* is one, while for the remaining ones is zero.

Group fairness metrics					
Protected group	Statistical parity	Disparate Impact	Equal Opportunity	Average odds	FRU
Gender/Female	−0.135	0.834	−0.056	−0.132	0.224
Age/Young	−0.202	0.752	−0.124	−0.149	0.107

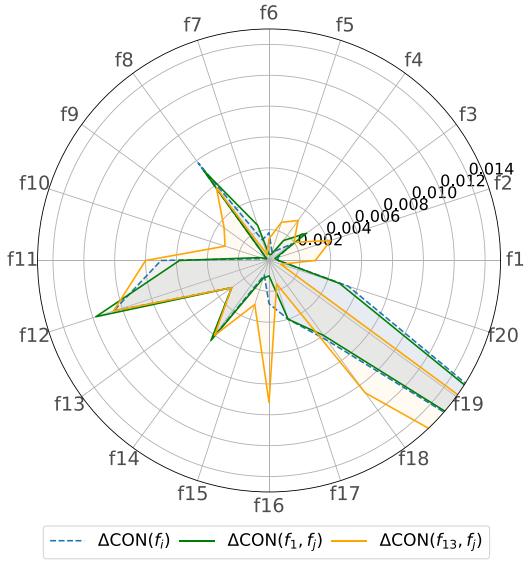

Fig. 7. ΔCON values when (i) suppressing each feature, (ii) suppressing each feature and *Age* (f_1), and (iii) suppressing each feature and *Gender* (f_{13}).

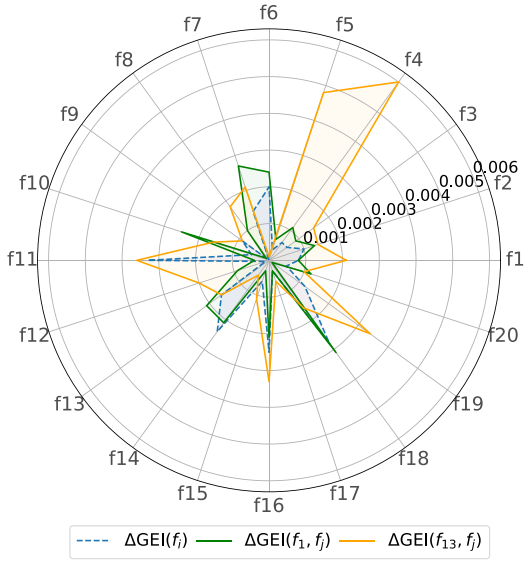

Fig. 8. ΔGEI values when (i) suppressing each feature, (ii) suppressing each feature and *Age* (f_1), and (iii) suppressing each feature and *Gender* (f_{13}).

only tells us that results might be impacted by the granularity of the bias analysis. Therefore, broader studies are often needed when analyzing bias.

5. Concluding remarks

This paper builds upon our recent work [18] where we propose a measure termed *fuzzy-rough uncertainty* that quantifies bias encoded in protected features. Applicable in pattern classification settings, our FRU measure quantifies the change occurring in the fuzzy-rough boundary regions after removing a protected feature. In other words, we use the change in the decision boundaries as a proxy for explicit bias. Advantages of our measure are that (i) it takes into account all features and feature categories at once, (ii) it can handle both numeric and nominal data, so no discretization is needed, (iii) it does not depend on any machine learning model to compute its outcomes but on a solid mathematical foundation, and (iv) it is less likely to be influenced by class imbalance.

The simulation results, using the *German Credit* dataset, allow us to draw interesting conclusions. First, our measure suggests that the dataset is more biased toward *Gender* than *Age* when it comes to explicit bias. When contrasting our finding against the state-of-the-art measures, we observe that individual fairness measures report a barely noticeable change under the same setting. In contrast, group fairness measures show the exact opposite trend. This suggests that focusing on a particular feature-category pair instead of analyzing the protected feature as a whole might give rise to misleading results. Second, even though FRU values depend on the choice of parameters, all configurations in our sensitivity analysis consistently report greater bias against *Gender*. Third, we recommend normalizing the FRU values of protected features using a relevant unprotected feature as reference. Moreover, we suggest using either Łukasiewicz or Fodor as implicators and the HMOM distance function since they report the largest changes. Finally, we found evidence of implicit bias in protected features (such as *Age*) encoded via the unprotected features.

There are several directions to be explored in future research endeavours. Firstly, reducing the computational complexity of our algorithm is vital as it is rooted in a lazy approach that can hardly be applied if instances exceed thirty thousand. Secondly, we suggest framing the concepts of *explicit bias* and *implicit bias* into a multi-valued logic approach such as the fuzzy set theory. Finally, it would be convenient to analyze implicit bias taking into consideration all associations/correlations between protected and unprotected features.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2022.01.005

References

[1] A. Balayn, C. Lofi, G.J. Houben, Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems, VLDB J (2021) 1–30.
[2] D.J. Fuchs, The dangers of human-like bias in machine-learning algorithms, Missouri S&T's Peer to Peer 2 (1) (2018) 1.
[3] K.D. Elsbach, I. Stigliani, New information technology and implicit bias, Acad Manage Perspect 33 (2) (2019) 185–206.

[4] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al., Bias in data-driven artificial intelligence systemsan introductory survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (3) (2020) e1356.
[5] S. Verma, J. Rubin, Fairness definitions explained, 2018 IEEE/ACM International Workshop on Software Fairness (fairware), IEEE, 2018.
[6] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 797–806.
[7] S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E.P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 329–338.
[8] S. Hajian, J. Domingo-Ferrer, A methodology for direct and indirect discrimination prevention in data mining, IEEE Trans Knowl Data Eng 25 (7) (2012) 1445–1459.
[9] M.J. Kusner, J.R. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017.
[10] R. Binns, On the apparent conflict between individual and group fairness, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 514–524.
[11] A. Chouldechova, Fair prediction with disparate impact: a study of bias in recidivism prediction instruments, Big Data 5 (2) (2017) 153–163.
[12] Y. Choi, G. Farnadi, B. Babaki, G.V.d. Broeck, Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 10077–10084.
[13] T. Kehrenberg, Z. Chen, N. Quadrianto, Tuning fairness by balancing target labels, Frontiers in Artificial Intelligence 3 (2020) 33.
[14] R.K. Bellamy, K. Dey, M. Hind, S.C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al., Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018, ArXiv preprint arXiv:1810.01943.
[15] K. Holstein, J.W. Vaughan, H. Daumé III, M. Dudik, H. Wallach, Improving fairness in machine learning systems: What do industry practitioners need? in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–16.
[16] D. Pedreshi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 560–568.
[17] N. Grgić-Hlača, M.B. Zafar, K.P. Gummadi, A. Weller, Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[18] L.K. Koumeri, G. Nápoles, Bias quantification for protected featuresin pattern classification problems, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 25th Iberoamerican Congress, CIARP 2021, Porto, Portugal, Springer, 2021. Proceedings, Lecture Notes in Computer Science
[19] Z. Pawlak, Rough sets, Int J Comp Inform Sci 11 (5) (1982) 341–356.
[20] J. Yang, T. Xu, F. Zhao, Modified uncertainty measure of rough fuzzy sets from the perspective of fuzzy distance, Math Problems Eng 2018 (2018).
[21] M. Bello, G. Nápoles, R. Morera, K. Vanhoof, R. Bello, Outliers detection in multi-label datasets, in: Mexican International Conference on Artificial Intelligence, Springer, 2020, pp. 65–75.
[22] M. Inuiguchi, W. Wu, C. Cornelis, N. Verbiest, Fuzzy-rough hybridization, in: Springer Handbook of Computational Intelligence, Springer, 2015, pp. 425–451.
[23] W. Pedrycz, G. Vukovich, Feature analysis through information granulation and fuzzy sets, Pattern Recognit 35 (4) (2002) 825–834.
[24] D. Dua, C. Graff, UCI machine learning repository, 2017,
[25] R.K. Prematunga, Correlational analysis, Aust Crit Care 25 (3) (2012) 195–199.
[26] S. Vluymans, L. D'eer, Y. Saeys, C. Cornelis, Applications of fuzzy rough set theory in machine learning: a survey, Fundam Inform 142 (1–4) (2015) 53–86.
[27] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J Art Int Res 6 (1997) 1–34.
[28] G. Nápoles, C. Mosquera, R. Falcon, I. Grau, R. Bello, K. Vanhoof, Fuzzy-rough cognitive networks, Neural Networks 97 (2018) 19–27.
[29] G.D. Pinal, S. Spaulding, Conceptual centrality and implicit bias, Mind Language 33 (1) (2018) 95–111.
[30] H. Akoglu, User's guide to correlation coefficients, Turkish Journal of Emergency Medicine 18 (3) (2018) 91–93.
[31] N.J. Nagelkerke, et al., A note on a general definition of the coefficient of determination, Biometrika 78 (3) (1991) 691–692.
[32] M.J. Rovine, A.V. Eye, A 14th way to look at a correlation coefficient: correlation as the proportion of matches, Am Stat 51 (1) (1997) 42–46.
[33] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., Scipy 1.0: fundamental algorithms for scientific computing in python, Nat Methods 17 (3) (2020) 261–272.
[34] H. Cramér, Mathematical Methods of Statistics (PMS-9), volume 9, Princeton University Press, 2016.
[35] M. Kramer, R2 statistics for mixed models, in: Proceedings of the Conference on Applied Statistics in Agriculture, volume 17, 2005, pp. 148–160.
[36] S. Seabold, J. Perktold, Statsmodels: Econometric and statistical modeling with Python, in: Proceedings of the 9th Python in Science Conference, Fuzzy Sets and Systems, volume 57, 2010, p. 61. Austin, TX
[37] B. Jayaram, R. Mesiar, On special fuzzy implications, Fuzzy Sets Syst 160 (14) (2009) 2063–2085.