Master's thesis abstract: "Bias quantification measures based on fuzzy rough sets"

Lisa Koutsoviti Koumeri[1,2]
Supervisor: Dr. Gonzalo Nápoles[2]

[1] Hasselt University, Martelarenlaan 42, 3500 Hasselt, Belgium
[2] Tilburg University, Warandelaan 2, 5037 AB Tilburg, Netherlands

Motivation. Artificial Intelligence (AI) systems are widely employed to solve pattern classification problems. These often include classifying which people can get a loan, receive medical treatment, or commit a crime. These life-changing decisions should be fair, i.e. not be based on protected features like race or gender. However, research revealed that this is not always the case due to biased labels or imbalanced data. Aiming to tackle this issue, numerous bias measures have been proposed, but despite these efforts, there is still great need to introduce new measures [1] for the following reasons. Existing approaches depend on different and often conflicting notions of fairness [4] or might consider part of the information available in the dataset (only sensitive and target features) [2]. Moreover, they often depend on black-box Machine Learning (ML) models whose outputs are sensitive to data preprocessing or training-test splits, and are not intuitively explainable. Finally, users need to make assumptions regarding the discriminated feature-category. We attempt to offer a remedy to these challenges.

This thesis proposes five measures based on the fuzzy-rough set (FRS) theory to quantify bias related to sensitive features of pattern classification datasets. This mathematical theory allows analyzing inconsistency in decision making systems [7], can define similarity thresholds when handling continuous features [6] while offering an explainable semantic background.

Methods. The measures are computed in a two step process. As a first step, we build information granules describing each decision class following the FRS formalization as introduced by [8]. Three information granules are computed per decision class: a positive, negative and boundary fuzzy-rough region. The membership value of an instance to a certain positive region indicates the extent to which the instance belongs to a decision class, does not belong to that class or the extent to which the instance belongs to the boundary region. This fuzzy granulation process is repeated twice: first, the three fuzzy-rough regions are calculated using all features in the data and, next, they are calculated again *excluding one of the protected features.* The intuition is that removing features from the decision making process should not cause large changes in the fuzzy-rough regions. The extent to which this happens is a proxy for bias.

As a second step, the five measures are calculated. These measures quantify the change in the membership values characterizing fuzzy-rough regions after the suppression of a protected feature. The first two measures quantify the change locally (between decision classes and information granules) and the rest globally (between information granules). Note that these values are not absolute but should be interpreted relatively to the respective values reported when we suppress a different protected feature. If the measures report relatively larger values regarding a certain protected feature, then that means that the exclusion of this same feature has a greater impact on the classification process, which can be understood as evidence for explicit bias.

Numerical simulations. The proposed measures are tested on *German Credit* and *Compas* datasets [5]. Protected features are *age* and *gender* for the former and *race* and *gender* for the latter. Decision classes are *creditworthy* or the opposite and *likely* or *unlikely to re-offend* respectively. The outputs of the proposed fuzzy-rough measures are compared to four popular bias measures [3] that fall under the category of group fairness and are computed using the AIF360 open source toolkit [5]. Results showed that almost all proposed measures differ from the literature measures both in direction and magnitude (a sample of the results is shown in Table 1). Such a disagreement raises concerns regarding the consistency of measures for bias quantification.

Table 1: Results of baseline and global fuzzy-rough measures tested on *German Credit* dataset. Ideal value of the former is 0.

Protected att.	**Baseline measures**			**Proposed global measures**		
	Statistical Parity	Equal Opportunity	Average Odds	Positive regions	Negative regions	Boundary regions
Age (young)	-0.28	-0.3	-0.25	0.01	0.01	0.04
Sex (female)	-0.002	0.04	-0.01	0.02	0.02	0.08

Conclusions. The proposed measures rely on an intuitive notion of explicit bias related to the uncertainty in decision-making as expressed by changes in the fuzzy-rough boundary regions. Our measures have several advantages that can be summarized as follows. First, the measures do not depend on any ML model. Second, the measures consider all features and feature-groups at once. This means that all available information is being leveraged and that arbitrary assumptions regarding the discriminated groups are avoided. Third, no discretization is needed during pre-processing to handle numeric features. Finally, the measures are not affected by data imbalances. Potential limitations of our approach include the limited number of considered literature measures, lack of experimentation with respect to bias that is implicitly encoded in non-sensitive features and dependence on the distance function and fuzzy operators.

As for the ramification of this thesis, we developed a stronger measure [9]. The corresponding paper received the Best Paper Award at the 25th Iberoamerican Congress on Pattern Recognition. An extended version of this work is currently under review for publication [10] at the Pattern Recognition Letters journal. Finally, we have recently submitted a journal contribution to the Neurocomputing journal where a neural model using a different approach confirmed the patterns found by the five proposed measures.

References

1. Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 329–338 (2019)
2. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806 (2017)
3. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), **54**(6), 1–35 (2021)
4. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (fairware), pp. 1–7 (2018)
5. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, **63**(4/5), pp. 4–1. (2019)
6. Inuiguchi, M., Wu, W., Cornelis, C., Verbiest, N.: Fuzzy-Rough Hybridization. Handbook of Computational Intelligence (2015)
7. Pawlak, Z.: Rough sets. International Journal of Computer & Information sciences, **11**(5), pp. 341–356 (1982)
8. Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. In: International Journal of General System, **17**(2-3), pp. 191–209 (1990)
9. Koutsoviti Koumeri, L., Nápoles, G.: Bias Quantification for Protected Featuresin Pattern Classification Problems, In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 25th Iberoamerican Congress, Lecture Notes in Computer Science, Springer (2021)
10. Nápoles, G., Koutsoviti Koumeri, L.: A fuzzy-rough uncertainty measure to discover bias encoded explicitly or implicitly in features of structured pattern classification datasets, arXiv (2021)