SIGMOD Response Letter, Paper ID 770

We thank the reviewers for their comments. Please find below individual comments to the issues raised.

- R1.1: The paper considers two pre-processing steps, which do not simulate any practical scenario (most practical scenarios consider tens or hundreds of processing steps). The steps mentioned in Section 2.1 like data transformation, discretization, scaling, etc are quite interesting and should be considered.

  R1.1. We fully agree with the reviewer that there are numerous steps that could be analyzed, which we acknowledge in the paper. We opted to focus on the two steps as these are most directly related to fairness approaches, i.e., steps that "counteract" often perceived shortcomings of existing approaches wrt number of protected attributes / groups supported and difficult parameterization. Further, note that the "small" pipeline already requires a considerable number of experiments. We believe that the open source code of our evaluation framework will foster the study of further pre-processing steps that are out of the scope of a single paper, that is, our framework serves as a blue-print where the pre-processing step can be easily extended / overwritten for future evaluations in a similar way.

  Out of the two steps that are considered, the first one is redundant because of two reasons. i) Most techniques internally throw the sensitive attribute to not bias the results (See Table 1) ii) All techniques guarantee fairness even when the sensitive attribute is given as an input. Therefore, throwing it would not affect the model's performance at all.

  Concerning (i), we agree that most in-processing techniques ignore sensitive attributes, however, this does not hold for pre-and post-processing approaches. Among all algorithms that run with / without ignoring these sensitive attributes, we observe what is summarized in Tab. 6 and leads to Takeaway 4, which indicates that removing the sensitive attribute is not necessarily the method of choice to avoid bias, which is the opposite of what is generally assumed. Concerning (ii), we agree that all algorithms aim to reduce bias (but without a formal guarantee in general), irrespective of the consideration of the sensitive attribute. However, our experiments showcase that the relative improvement compared to a baseline "bias-agnostic" classifier like Logistic Regression is not always significant.

- R1.2: The paper claims to have used all techniques after 2019, which is not entirely correct. There is a lot of research on feature selection methods and causal fairness in SIGMOD and VLDB, which is neither discussed nor considered in the experimental study. (Also, half of the methods used are older than 2019 too).

  R1.2. It was not our intention to claim that we have exhaustively covered all techniques since 2019. If the review can point out where we raised this impression, we are happy to rephrase it. Clearly, Since there have been $\approx$ 200 papers published between 2020 and 2022 only (see [40]), we could not cover all techniques. However, our selection of algorithms included in the evaluation makes sure to cover all subcategories defined in the survey by Hort et al. [40], i.e., we selected representatives for each subcategory. Concerning the cited related papers, we observe that they fall in the class of causal fairness algorithms. While this is a separate class in some surveys (e.g. by Caton), in the categories we consider, "Capuchin: Causal database repair for algorithmic fairness" fits the category of Sampling and "Causal feature selection for algorithmic fairness" fits the category of Representation Learning, which are already covered. More importantly, we opted not to include causal fairness methods as a special class of algorithms to consider, because measuring the total causal effect requires a causal graph (as seen in the code supplementary paper by Islam et al.). This requires additional domain knowledge on the datasets. Due to the huge amount of work in fair classifications, we opted to focus on fairness research that does not require additional domain knowledge. Although some of the methods are designed to work without a causal graph as input, they still require the admissible attributes. In a revision, we can clarify the reasons why we did not choose to evaluate causal fairness approaches. If the reviewer thinks this class is essential, we can try to integrate additional algorithms as pointed out by the reviewer (preferably those with available code is available) and evaluate those analogously to the other algorithms.

- R1.3: The experiments consider logistic regression model for the analysis. It is unclear why the authors made this choice as I expect random forests, neural networks, boosting methods to perform better than logistic regression.

  R1.3. We chose Logistic Regression for two main reasons (as outlined in Sec. 2.1): (1) The experimental paper by Islam et al. uses Logistic Regression, relying on their evaluation where, when using different types of classifiers as input, logistic regression consistently finished among the top performing classifiers; (2) All approaches that additionally need a given classifier as input support logistic regression (at least as one possible option) and many of them are solely tested when opting for logistic regression. Given that this was not the focus of our evaluation, we thus kept this standard choice, which in turn guided our choice of using logistic regression as baseline algorithm for a "fair" comparison. In a revision, we

could add further baseline classifiers for comparison. While using other classifier types as input for the pre- and post-processing algorithms is too time-consuming for a revision, we can still use the mentioned classifiers as baseline for the analysis (e.g. by choosing the optimal "classic" classifier for each conducted run).

- R1.4: The approach to pick parameter selection is to run grid search. This approach can work with 1-2 components, but does not scale well. Please consider state-of-the-art methods for parameter selection.

  R1.4. Our choice of optimization algorithm was primarily guided by RQ1, where the brute force approach of grid search allowed us to make the observation summarized in Sec. 5.2 with confidence. However, we agree that the discussion of overall runtime is negatively affected by this choice of optimization algorithm (hence our consideration of both overall time and iteration time in Sec. 5.6). We can clarify this in a revision.

- R1.5: Experiments:
  a. The first takeaway is not conclusive. Only one of the considered pre-processing algorithms has a continuous parameter, which is highly sensitive to choice of parameters. Also, the takeaway is not very interesting because if the data has high bias, choosing the best parameter means more fairness and hence high deviation from the untuned pipeline.
  Also, the presentation of results is unclear. Unfortunately, I do not have a concrete suggestion on how to plot, but anything that shows the takeaway clearly and is not spread across 5 plots would be good.

  R1.5a. We clarify that the untuned pipeline still opts to mitigate bias, it just uses the default parameter settings. Hence, some of the algorithms are not as heavily influenced by parameter tuning (see e.g. [43], [45]) although they also have continuous variables. However, you are correct that many of the pre-processing algorithms (except for two) only have few parameters to tune, which can be an explanation for why most in-processing algorithms are typically more affected by the parameter optimization component. We can rephrase our discussion/takeaway to clarify this. Maybe the more interesting takeaway is that through the effect of optimization (current focus of Takeaway 1), we actually revisit a takeaway of Islam et al., stating that the performance of the (non-optimized) algorithms on other fairness metrics, for which they are not designed for, is arbitrary. Indeed, we show, that hyperparameter tuning actually can close the gap, which leads to the conclusion that (highly parameterized) approaches that are designed for one metric can also perform similarly well when we opt for another fairness metric. This increases the applicability of several approaches. We are happy to elaborate and restructure main conclusions / takeaways.

  b. Binarizing the group of protected attributes, considers male and white as privileged and all others as minority groups. This is significantly different from treating gender and race as the sensitive attributes where the

fairness requirement is with respect to both race and gender. The takeaway 4 does not discuss this information loss due to the preprocessing step and can lead to misleading outcomes.
R1.5b. There seems to be a misunderstanding here. For this set of experiments, we still evaluate the classifier output with respect to both race and gender, thus we apply the evaluation metrics over all groups, irrespective of whether the binarization flag is set to True or False. This is briefly mentioned in the setup in 5.4, but we can make this more clear in this section.

c. Section 5.5 uses default parameters, which is completely opposite of what takeaway 1 mentioned (parameter tuning is important). Please tune parameters in this experiment.

R1.5c. Our rationale for not using the tuned version is to give the combination of approaches its best shot in complementing one another by leaving the room for improvement as large as default parameter settings allow (analogous to our observation in the second part of Takeaway 1). We can of course run the experiments when using the individually optimized algorithms in combination (where we expect even less of a valuable impact).

- R1.6: Many of the points do not agree with Islam et al. [44] and these should be discussed in more detail. Discussing these issues of generalization in these benchmarking papers would help the reader make an accurate opinion from takeaways.

  R1.6. We can go further into details in a possible revision.

- R1.7: The conclusion section should make concrete recommendations about each algorithm. The takeaways are not specific enough to understand how to use any of the considered methods.

  R1.7. Based on our experiments, we have actually compiled a decision tree on appropriately selecting/setting pipelines based on application requirements/settings. Due to space constraints, it does not fit into the paper, but we are happy to elaborate on concrete recommendations in a revision.

- R2.1: Although there is a detailed interpretation of the plots, the reasons behind some takeaways remain unclear. For example, why does the effect of hyperparameter tuning on pre- and post-processing techniques appear to be less pronounced than in-processing? More analysis and explanation about all the key takeaways might be needed.

  R2.1. As correctly pointed out by reviewer 1, one factor for Takeaway 1 is the number and types of parameters to tune. In a revision, we can refocus the analysis and explanation to main / most interesting takeaways to discuss them in more detail.

- R2.2: In this paper, the hyperparameter tuning only uses accuracy as the objective. However, since the application scenario is fair classification, I

am curious about what we will get if the objective considers both fairness and accuracy. Would that change the conclusions about hyperparameter tuning?

R2.2. There seems to be a misunderstanding here. We do NOT only use accuracy but use a combination of both fairness and accuracy for the hyperparameter tuning part (see Sec. 2.2). That is why a large set of classifiers actually has lower accuracy when applying hyperparameter tuning (as a tradeoff for the improvement in fairness), similar to the results shown in "Promoting fairness through hyperparameter optimization". In that work, fair hyperparameter tuning has been applied on base classifiers optimized for accuracy. Thus, there has been a consistent improvement and fairness with accuracy as a tradeoff. Since we apply it on approaches that already try to mitigate bias (while not ignoring accuracy), there are also approaches for which we could increase the accuracy using parameter optimization. We suspect that part of the misunderstanding may stem from the consideration of multiple fairness metrics. Essentially, we run the experiments multiple times for each fairness metric, i.e. in the first set of experiments we consider accuracy (or error rate to be more precise) and demographic parity as objective functions, in the second set accuracy and equalized odds, etc. We can make the optimization goals and procedure clearer in a revision.

- R2.3: The generation of synthetic data seems unclear. More descriptions of how two biases differ in the synthetic data will help readers understand the related key findings.

  R2.3. Essentially, social bias indicates direct bias and implicit bias is indirect bias. If direct bias is induced, it means that the protected attributes solely affect the label, thus there is a correlation between protected attribute and label but no correlation between protected and unprotected attributes. For implicit bias, there is a high correlation between the protected attributes to the attributes that were used to influence the label. We can clarify this in a revision and provide some details on the generation of the datasets (with examples, possibly relegated to the GitHub repository though due to space constraints)

- R2.4: There are a few fairness-inducing algorithms missing (e.g. [2, 3]) compared to a highly related recent survey paper [4].

  R2.4. See comments to R1.2, explaining our restriction to non-causal fairness methods. Furthermore, in Islam et al., both approaches did not reach outstanding performances on the non-causal metrics compared to the other evaluated algorithms. As they have been implemented to optimize causal fairness, this is expected, and would not result in a "fair" evaluation in our experimental paper. As replied to R1, we are open to including some causal fairness algorithms as well.

- R4.1: Fairness in Machine Learning is heavily studied in past 5 years and

many papers, surveys, and experimental evaluations (including the ones mentioned by the authors) have been published across various, mostly ML, conferences and in ACM Facct. I did not find the addition by this paper novel or significant enough.

R4.1. While we agree that several surveys and experimental evaluations have been published in the past, we are not aware of any experimental evaluation papers focusing on the questions studied in this paper that arise in a (previously not considered) larger context of a pipeline. This goes significantly beyond repeating the same experiments on more datasets and newer algorithms, as outlined in the contributions paragraph Section 1. We would appreciate a more differentiated comment, pointing out prior work that "reduces" specific experimental contributions of this paper or a clearer justification of why they are considered not significant.

- R4.2: I did not find the "takeaways" from the experiments very interesting as (at least most of) those are already well-known within the fairness community. For example, a few of the takeaways are about the importance of hyperparameters. This is not a new fact for the FairML community since, not only they know this but there also have been work on how to improve fairness with hyper-parameter tuning (using techniques such as Bayesian optimization)

  R4.2. We would appreciate slightly more constructive feedback pointing out more clearly what previous results we overlooked with respect to the detailed study of hyperparameter tuning in FairML (citations?) in particular, and analogously for the other takeaways that the reviewer all qualifies as not interesting.
  Focusing on the specific takeaway about hyperparameter optimization brought up by the reviewer, we agree that the statement that optimization actually optimizes results is not surprising/interesting. However, our evaluation and conclusions are more nuanced. For instance, we are not aware of a larger experimental evaluation that studies the actual effect in detail, revealing, e.g., that some algorithms, like [43] or [45] are barely affected by tuning the parameters, while for others, like [60] or [35], the actual effect is bigger than one has probably imagined. More interestingly perhaps, we showed that hyperparameter tuning levels the playing field among competing approaches by making algorithms work competitively in settings they are not initially designed for (i.e., for other fairness definitions, see also reply to R1.5a). This has some very practical implications as it makes the choice of an algorithm / metric easier and less error-prone. Generalizing the above comment, we point out that many, more detailed insights did not "make it" into the general Takeaway boxes and we would appreciate some comment on whether some of these would be more interesting to the reviewer (see also replies to R1). Such rebalancing / refocusing would be in the scope of a revision.