BiMi sheets for bias mitigation methods

MaryBeth Defrance, Maarten Buyl, Guillaume Bied, Jefrey Lijffijt, and Tijl De Bie

Ghent University, Ghent 9000, Belgium {first-name}.{last-name}@ugent.be

Over the past 15 years, hundreds of bias mitigation methods have been proposed in the pursuit of fairness in machine learning (ML) [3]. However, fairness cannot be reduced to a single concept. This diversity stems from the impossibility of reducing fairness to a single concept, and, given a selected fairness definition, from different possible locations of interventions in the model pipeline (pre/in/post-processing) and algorithmic strategies [6]. However, this proliferation makes it unclear when, where, and how a method is applicable in practice.

We propose *BiMi sheets* as a portable, uniform guide to the design choices of any bias mitigation method. These complement (and were inspired by) *datasheets* for datasets [9] and *model cards* for models [12]. Datasheets and model cards focus on the biases present in the resource. BiMi sheets focus on the capabilities of the bias mitigation method to handle certain types of bias.

Figure 1 provides an example of a BiMi sheet. The sheet uses tags which provide a quick overview of the main design choices that are often made in fairness. Each section enriched with a description providing additional details. The sheet is structured as follows:

1. **Metadata**: Basic information on the method
2. **Method Description**: Main properties and description of the method
   tag: Method type based on existing taxonomies such as Caton et al. [3, 4]
   tag: Envisioned ML task of the method. tag: Compatible dataset types
3. **Pipeline Architecture**: Pipeline architecture compatible with the method
   tag: Location of method in the pipeline tag: Compatible ML model
4. **Fairness Type**: The fairness goal or the specific biases mitigated
   tag: Compatible sensitive attribute type tag: Guaranteed fairness level
   tag: Fairness kind tag: Specific fairness definitions
5. **Implementation Constraints**: Specific programming environment
   tag: Specific environment tag: Known compatible packages
6. **Tested Use Cases**: Use cases on which the method is already tested
   tag: Tested datasets

BiMi sheets will be designed through an iterative process. Building on systematic surveys of the fairness literature [3, 13, 11, 8, 4, 10], covering a variety of ML tasks, we will identify and document cited methods for which an implementation is readily available, ensuring the breadth and representativity of the sheets' coverage. As this project requires significant input from the research community, the authors of the covered methods will be contacted with requests for feedback on their method's BiMi sheets. In a later stage, we also intend to solicit fairness practitioners for qualitative feedback.

We aim to provide templates in multiple formats for authors to create their own BiMi sheet. These templates include a guide on the appropriate tags and guidance on the content of each section. Further, we plan to host a website containing the BiMi sheets, linking to the method's repository.

Metadata

Name: Error-parity
Authors: André Cruz and Moritz Hardt
Version: 0.3.11 License: MIT License
Proposed in *Unprocessing Seven Years of Algorithmic Fairness* [5]

Method Description

Thresholding | Binary prediction | Hard labels | Dataset independent

Error-parity sets groupspecific acceptance thresholds so as to **minimize risk while achieving an equality in error rates** across a desired set of groups. It is both simple and computationally efficient. Error-parity achieves exact error rate equality, unlike many preprocessing and inprocessing, which achieve some relaxation of the constraint. The method uses the output scores and returns hard prediction labels.

Pipeline Architecture

Post-processing | Probabilistic Classifier

Error-parity is compatible with any underlying learner that can produce scores of predicted probabilities.

Fairness Type

Categorical Attributes | Guaranteed Equality Level
Group Fairness | Demographic Parity | Equalized Odds
Equal Opportunity | Predictive Equality

Fairness is achieved when the absolute difference of a specific statistical property is smaller than a predetermined threshold.

Implementation constraints

Python 3.8-3.12 | Scikit-learn [14] | fairlearn [2]

The implementation requires a trained score predictor that takes in samples, X, in shape (num_samples, num_features), and outputs real-valued scores, R, in shape (num_samples,) as the model that feeds into error-parity.

Tested Use Cases

Synthetic dataset | Adult dataset [1] | Folktables datasets [7]

Fig. 1: Example of a BiMi sheet for the Error-parity [5] package

Bibliography

[1] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[2] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

[3] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, July 2024.

[4] Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in large language models: A taxonomic survey. 26(1):34–48, July 2024.

[5] André Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] MaryBeth Defrance, Maarten Buyl, and Tijl De Bie. ABCFair: an adaptable benchmark approach for comparing fairness methods. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[7] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024.

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. (arXiv:1803.09010), December 2021. arXiv:1803.09010 [cs].

[10] Charlotte Laclau, Christine Largeron, and Manvi Choudhary. A survey on fairness for machine learning on graphs, 2024.

[11] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, July 2022.

[12] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 220–229, January 2019. arXiv:1810.03993 [cs].

[13] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and unfairness in machine learning models:

A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 2023.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.