Fairbeat: Assessing and Mitigating Bias with the Composite Balance Score

Pierre-Antoine Lequeu[1], Sofiane Lagraa[2] (✉), Geoffroy Robin[2], and Moussa Ouedraogo[2]

[1] Sorbonne University, Paris, France
pierre-antoine.lequeu@sorbonne-universite.fr
[2] Fujitsu Technology Solutions S.A., Capellen, Luxembourg
firstname.lastname@fujitsu.com

Abstract. *Fairbeat*, a novel Fairness Assessment tool for Resampling-based Bias Elimination and Algorithm Training, addresses the critical challenge of fairness in machine learning. Machine learning models often exhibit biases stemming from imbalances in training data concerning protected attributes, leading to discriminatory outcomes. *Fairbeat*leverages the Composite Balance Score (CBS), a comprehensive metric that evaluates the balance of the dataset by integrating the imbalance of attributes, the imbalance of labels and the association of attributes and labels into a single normalized score. This tool facilitates proactive bias assessment prior to model training, supports multi-class attributes, and provides a user-friendly environment for exploring and visualizing the impact of various bias mitigation techniques, including resampling methods, thereby promoting the development of more equitable and ethically sound AI systems. The demonstration video can be found at https://youtu.be/9aHKfZgtXKg.

Keywords: Composite Balance Score · Fairness · Bias Mitigation · Machine Learning.

1 Introduction

Ensuring fairness in machine learning models is crucial for ethical compliance and societal impact. Models often exhibit biases due to imbalances in training data, particularly concerning protected attributes like gender, age, and ethnicity. Addressing these biases is essential to prevent discrimination and achieve equitable outcomes.

Problem Statement. Machine learning models are increasingly used in human-centric decision-making areas such as judiciary systems, human resources, credit assessment, and healthcare. However, these models often show unfair behavior towards social groups based on protected attributes, leading to discrimination and ethical concerns. The challenge lies in predicting the fairness of these models by analyzing their training data and implementing bias mitigation strategies without compromising performance.

Existing Works. Previous studies have explored various bias mitigation techniques, including pre-processing [3], in-processing [4], and post-processing [5] methods. These approaches often focus on single, binary protected attributes, neglecting the complexities of multi-class attributes. While some methods have shown promise in improving fairness, they frequently lead to a reduction in utility, known as the fairness-utility trade-off [1]. Moreover, handling missing data and proxy attributes remains a challenge, influencing both fairness and model performance [2]. Moreover, FairnessEval [9] is a Python framework for evaluating and comparing fairness in ML models, streamlining data preparation, evaluation, and result presentation to aid in model selection and validation.
Novelty and Contribution. Our paper introduces *Fairbeat*, a Fairness Assessment Interface for Resampling-based Bias Elimination and Algorithm Training. It is based on the Composite Balance Score (CBS), a novel metric designed to evaluate the balance of datasets with respect to protected attributes and predict model fairness by analyzing the training data. *Fairbeat* informs the decision of whether or not to apply bias mitigation. It offers several key advantages, including a **comprehensive balance measure** that combines attribute imbalance, label imbalance, and attribute-label association into a single, normalized score ranging from 0 to 1. This easy-to-interpret metric assesses overall dataset balance and **predicts model fairness** by focusing on dataset balance as an indicator of potential bias. Furthermore, CBS supports multi-class attributes, extending beyond binary categories, and enables **proactive bias assessment and mitigation**, allowing for bias evaluation before model training begins. Finally, *Fairbeat* has a friendly user interface for bias assessment and mitigation. The video of the demonstration is available at https://youtu.be/9aHKfZgtXKg.

2 *Fairbeat*: Assessing and Mitigating Bias with the Composite Balance Score

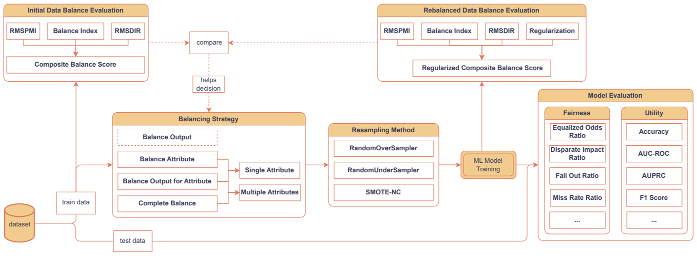


Fig. 1: *Fairbeat*: balancing strategies and resampling methods workflow.

Fairbeat is a tool designed to simplify the evaluation and mitigation of bias in machine learning datasets. It provides an intuitive tool for users to assess dataset fairness, explore bias mitigation techniques, and visualize their impact.

The tool democratizes fairness-aware machine learning, enabling practitioners to build equitable AI systems. Figure 1 outlines the workflow for evaluating and mitigating bias using the Composite Balance Score (CBS) and related metrics. If bias mitigation is needed, a balancing strategy and a resampling method are selected. The balance of the rebalanced data is then evaluated, and a machine learning model is trained and tested for fairness and utility. The CBS value helps determine the success of the balancing, comparing results to the initial data balance.

2.1 Assessing dataset balance using the Composite Balance Score (CBS)

We introduce the *Composite Balance Score*, a new metric for evaluating the balance of protected attributes. It uses three measures: the *Balance Index*, *RMSDIR*, and *RMSPMI*.

Balance Index (Bal) is introduced as a novel metric to quantify the balance of classes within a protected attribute, addressing a critical aspect of dataset fairness. $\mathrm{Bal}(A) = 1 - \frac{\mathrm{imb}(A)}{\sqrt{\frac{n-1}{n}}}$, where $\mathrm{imb}(A)$ is the imbalance index, n is the number of classes, and A is the protected attribute. Unlike prior works [6] that often rely on arithmetic means to assess imbalance, the Balance Index employs a quadratic mean of the distribution deviation, providing a more sensitive measure to variations in class representation. Furthermore, it's normalized to a [0, 1] scale, offering intuitive interpretability where 1 signifies perfect balance and 0 indicates extreme imbalance. The Balance Index offers a unique combination of sensitivity and interpretability, making it a valuable tool for evaluating and addressing attribute imbalances in fairness-aware machine learning.

Root Mean Squared Disparate Impact Ratio (RMSDIR): To quantify label imbalance across protected attribute classes, this paper introduces the Root Mean Squared Disparate Impact Ratio (RMSDIR): $\mathrm{RMSDIR}(A) = \sqrt{\frac{\sum_{c \neq c_{\mathrm{priv}}} \mathrm{DIR}_{\mathrm{nor}}(c)^2}{|\{c \neq c_{\mathrm{priv}}\}|}}$, where $\mathrm{DIR}_{\mathrm{nor}}(c)$ is the normalized disparate impact ratio for class c proposed in [7], [8]. Building upon the concept of Disparate Impact Ratio (DIR): $DIR(c_i) = \frac{P(Y=1 \mid A=c_i)}{P(Y=1 \mid A=c_{privi})}$ with $c_i \neq c_{privi}$, commonly used to compare favorable outcome rates between groups, RMSDIR offers a crucial normalization step. Unlike traditional DIR, which lacks an upper bound and can be challenging to interpret [7], RMSDIR leverages the normalized disparate impact introduced by Badran et al. to ensure a [0, 1] scale. This normalization allows for a more intuitive understanding of label imbalance, where 1 signifies perfect balance and 0 indicates significant disparity, regardless of whether it favors the privileged or unprivileged class. By aggregating these normalized values using a root mean square, RMSDIR provides a single, robust measure of label imbalance for the entire protected attribute, offering a more comprehensive assessment than individual pairwise comparisons.

Root Mean Squared Pointwise Mutual Information (RMSPMI) is introduced as a novel measure to capture the information shared between classes of a protected attribute and the target variable's labels, offering a unique perspective on dataset bias. RMSPMI$(A) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{y=0}^{1} \text{PMI}_{\text{nor}}(c_i,y)^2}{2n}}$, where PMI$_{\text{nor}}(c_i, y)$ is the normalized pointwise mutual information for class c_i and label y. Unlike traditional fairness metrics that focus solely on outcome disparities [7], [8], RMSPMI leverages the normalized Pointwise Mutual Information (PMI): $PMI(c_i, y) = \log \frac{P(A=c_i,\, Y=y)}{P(A=c_i)P(Y=y)}$ to quantify the degree of association between each class and each label. While in [8], the authors used PMI to measure unwarranted associations, RMSPMI aggregates these individual PMI values using a root mean square, providing a single, comprehensive measure of the overall dependency between the protected attribute and the target variable. This approach allows for a more nuanced understanding of how a protected attribute might be influencing predictions beyond simple outcome disparities, capturing subtle biases that could be missed by other metrics. By focusing on information sharing, RMSPMI complements existing fairness measures and provides valuable insights for bias mitigation strategies.

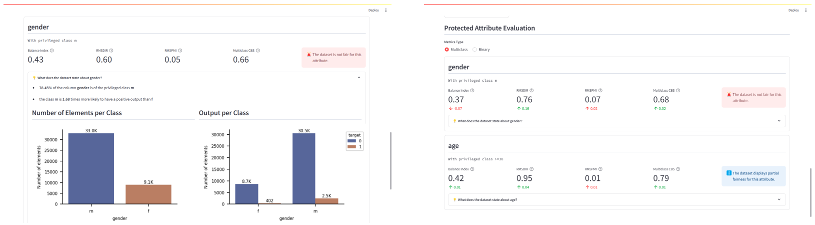
Composite Balance Score (CBS) is a new metric designed to evaluate the balance of a dataset concerning a protected attribute, as shown in Figure 2a. CBS is calculated as: CBS$(A) = \frac{\text{Bal}(A)+\text{RMSDIR}(A)+(1-\text{RMSPMI}(A))}{3}$. CBS captures attribute and label imbalances and the statistical dependence between the attribute and the target variable. Normalized to a [0, 1] scale, CBS helps assess dataset fairness, guiding bias mitigation strategies and tracking their effectiveness. By calculating CBS for each protected attribute, users can identify attributes with scores below a threshold (e.g., 0.80) that may need bias mitigation. CBS guides the application of bias mitigation techniques, such as resampling methods, to improve dataset balance and model fairness. Integrating CBS into workflows enables organizations to proactively address biases, resulting in fairer and more equitable machine learning models.

2.2 Resampling techniques

Resampling techniques are integral tools in data preprocessing after fairness assessment using the CBS score, enabling modification of datasets through the addition or removal of rows for bias mitigation, as shown in Figure 2b. These techniques are used predominantly to rectify imbalanced labels in classification tasks. In the realm of fairness, prior research has investigated resampling methods to equilibrate protected attributes. The strategies for balancing include: no balance, balancing labels, balancing classes, balancing labels across all classes/attributes, and achieving complete balance. Resampling methods to implement these strategies are classified into over-sampling (Random Over-Sampling (ROS), SMOTE-NC) and under-sampling (Random Under-Sampling (RUS)).

3 Conclusion

Fairbeat underscores the pivotal importance of dataset balance in reducing bias within machine learning models, particularly in binary classification scenarios



gender

With privileged class m

Balance Index	RMSDIR	RMSPMI	Multiclass CBS
0.43	0.60	0.05	0.66

The dataset is not fair for this attribute.

What does the dataset state about gender?

- **78.45%** of the column **gender** is of the privileged class **m**
- the class **m** is **1.68** times more likely to have a positive output than **f**

(a) Fairness evaluation.

Protected Attribute Evaluation

Metrics Type

● Multiclass ○ Binary

gender

With privileged class m

Balance Index	RMSDIR	RMSPMI	Multiclass CBS
0.37 ↓ -0.07	0.76 ↑ 0.16	0.07 ↑ 0.02	0.68 ↑ 0.02

The dataset is not fair for this attribute.

What does the dataset state about gender?

age

With privileged class >=38

Balance Index	RMSDIR	RMSPMI	Multiclass CBS
0.42 ↑ 0.01	0.95 ↑ 0.04	0.01 ↑ 0.01	0.79 ↑ 0.01

The dataset displays partial fairness for this attribute.

What does the dataset state about age?

(b) Bias mitigation.

Fig. 2: *Fairbeat* dashboard.

involving multi-class protected attributes. The introduced Composite Balance Score (CBS) serves as a robust predictor of model fairness. Implementing balancing strategies, notably the equalization of labels within classes, markedly enhances fairness while incurring minimal utility loss. Although the efficacy of CBS wanes with intersectional attributes, maintaining balanced datasets is essential for fostering fairer and more equitable machine learning outcomes.

4 Acknowledgments

This work was supported by the European Union's HE RAIDO project under the grant agreement number 101135800.

References

1. Bertsimas, D., Farias, V. F., and Trichakis, N. (2012). On the efficiency-fairness trade-off. *Management Science*, 58:2234–2250.
2. Caton, S., Malisetty, S., and Haas, C. (2022). Impact of imputation strategies on fairness in machine learning. *J. Artif. Intell. Res.*, 74:1011–1035.
3. Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740.
4. Wadsworth, C., Vera, F., and Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction. *CoRR*, abs/1807.00199.
5. Mishler, A., Kennedy, E. H., and Chouldechova, A. (2021). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *ACM FAccT*, pages 386–400.
6. Gong, Y., Liu, G., Xue, Y., Li, R., and Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268.
7. Badran, et al. (2023). Can ensembling preprocessing algorithms lead to better machine learning fairness? *Computer*, 56:71–79.
8. Tramèr, F., Atlidakis, V., Geambasu, R., Hsu, D. J., Hubaux, J.-P., Humbert, M., Juels, A., and Lin, H. (2015). Discovering unwarranted associations in data-driven applications with the FairTest testing toolkit. *CoRR*, abs/1510.02377.
9. Baraldi, A., Brucato, M., Dudík, M., Guerra, F., and Interlandi, M. (2025). FairnessEval: a framework for evaluating fairness of machine learning models. In *EDBT*, pages 123–134. ACM.