CERIAS

The Center for Education and Research in Information Assurance and Security

Achieving Algorithmic Fairness through Label Flipping

Shashank Thandri, Romila Pradhan

Abstract

As machine learning (ML) and artificial intelligence (AI) become increasingly prevalent in high-stake decision making, fairness has emerged as a critical societal issue. Individuals belonging to diverse groups receive different algorithmic outcomes largely due to the inherent errors and biases in the underlying training data, thus resulting in violations of *group fairness* or *bias.*

We address the problem of resolving group fairness by flipping the labels of instances in the training data. We propose solutions to obtain an ordering in which the labels of training data instances should be flipped to reduce the bias in predictions of a model trained over the modified data. We experimentally evaluate our solutions on several real-world datasets and demonstrate that bias is reduced by flipping a small fraction of training data labels.

Algorithmic Fairness

Fairness is measured in two broad categories: *individual* fairness and *group* fairness [1].

- **Individual fairness:** individuals that are *similar* should be treated the same
- **Group fairness:** individuals belonging to different *sensitive* groups (according to e.g., race, gender, age etc.) should receive the same treatment

The core idea of group fairness is for the outcome probabilities of the *privileged* and *unprivileged* groups to satisfy certain statistical properties. Below are three popular methods to quantify group fairness:

- **Statistical parity:** privileged and unprivileged group have equal probability of having a favorable outcome
- **Equalized odds:** privileged and unprivileged groups have equal true positive rate (TPR) and false positive rate(FPR)
- **Predictive parity:** privileged and unprivileged groups have equal precision

Bias mitigation techniques can be broadly categorized as pre-processing, in-processing, and post-processing techniques. Of these, pre-processing techniques have been shown to be effective, model agnostic and easy to implement. Within pre-processing, Kamiran and Calders (2009) introduced the concept of **label flipping** to change the labels of a few training data instances (that might have erroneous labels as a result of data errors or annotation errors) and mitigate bias. Recently, Zhang et al (2023) have demonstrated that label flipping is effective for achieving individual fairness.

References

1. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35.
2. Kamiran, F., & Calders, T. (2009). Classifying without discriminating. In IEEE Conference Publication | IEEE Xplore. 2009 2nd International Conference on Computer, Control and Communication.
3. Zhang, H., Tae, K. H., Park, J., Chu, X., & Whang, S. E. (2023). iFlipper: Label flipping for individual fairness. Proceedings of the ACM on Management Data, 1(1), 8.

Problem statement

Label flipping for group fairness: In what order should we flip the labels of training data instances such that the bias in model predictions is reduced the most?

Proposed solutions

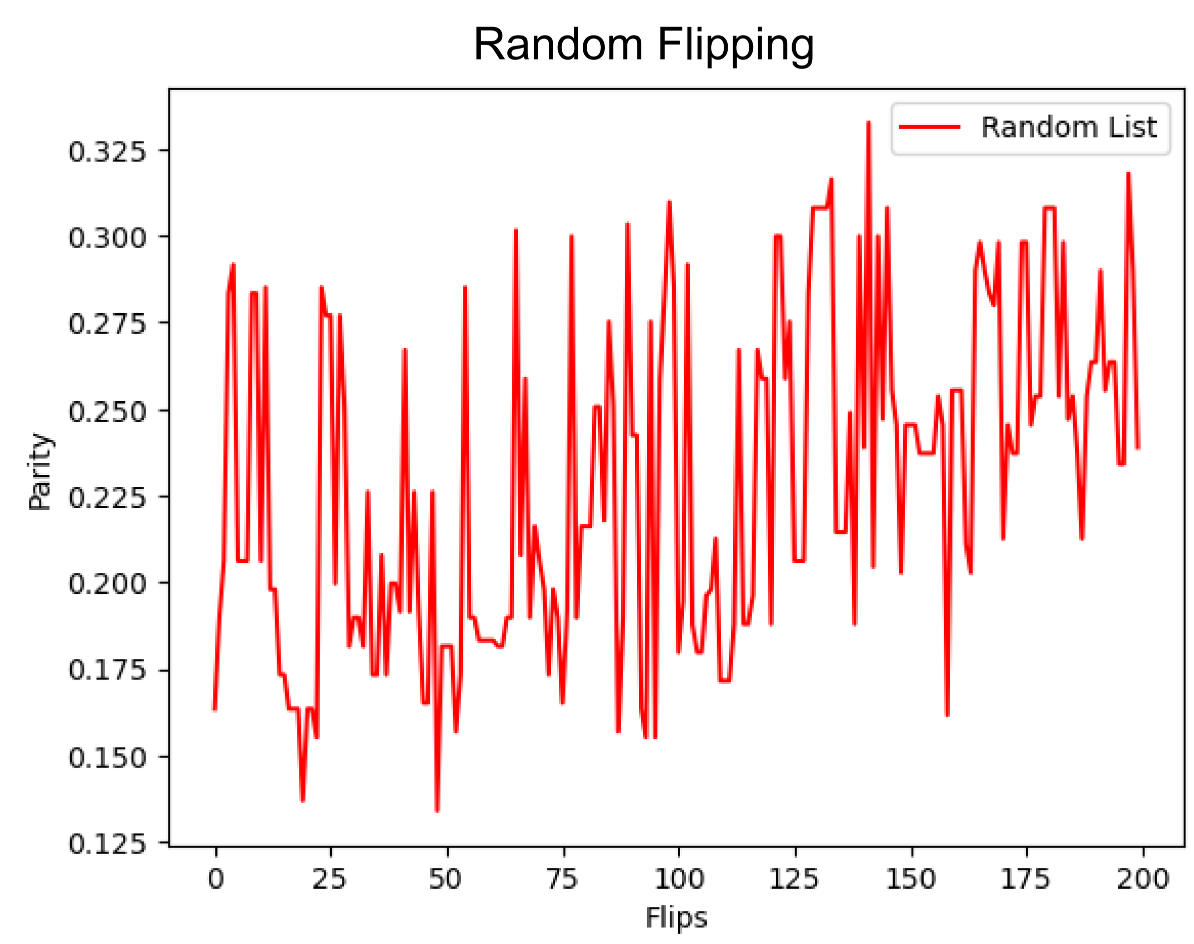
Random: Randomly rank training data instances for label flipping (baseline)

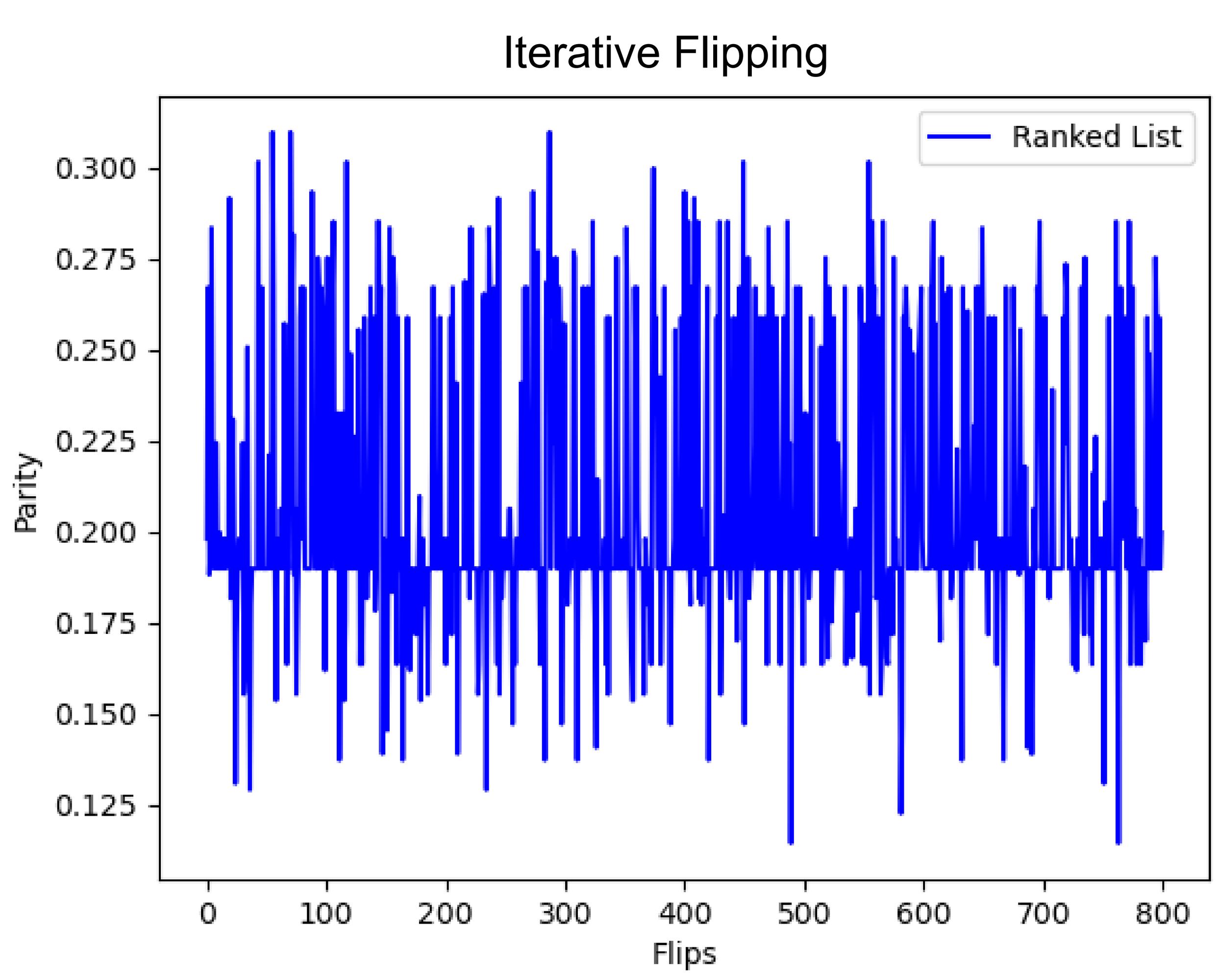
Iterative: For each training data instance, flip its label and measure change in fairness. Rank instances in decreasing order of their change

Uncertainty reduction: Using probabilities output by the model, compute Shannon entropy of each training data instance. Rank instances in decreasing order of their entropy

Maximum expected utility: Use the concept of expected utility to determine the resulting fairness of the system in case the label is flipped for an instance weighted by the probability of a flip and in case the label is not flipped weighted by the probability of not flipping.

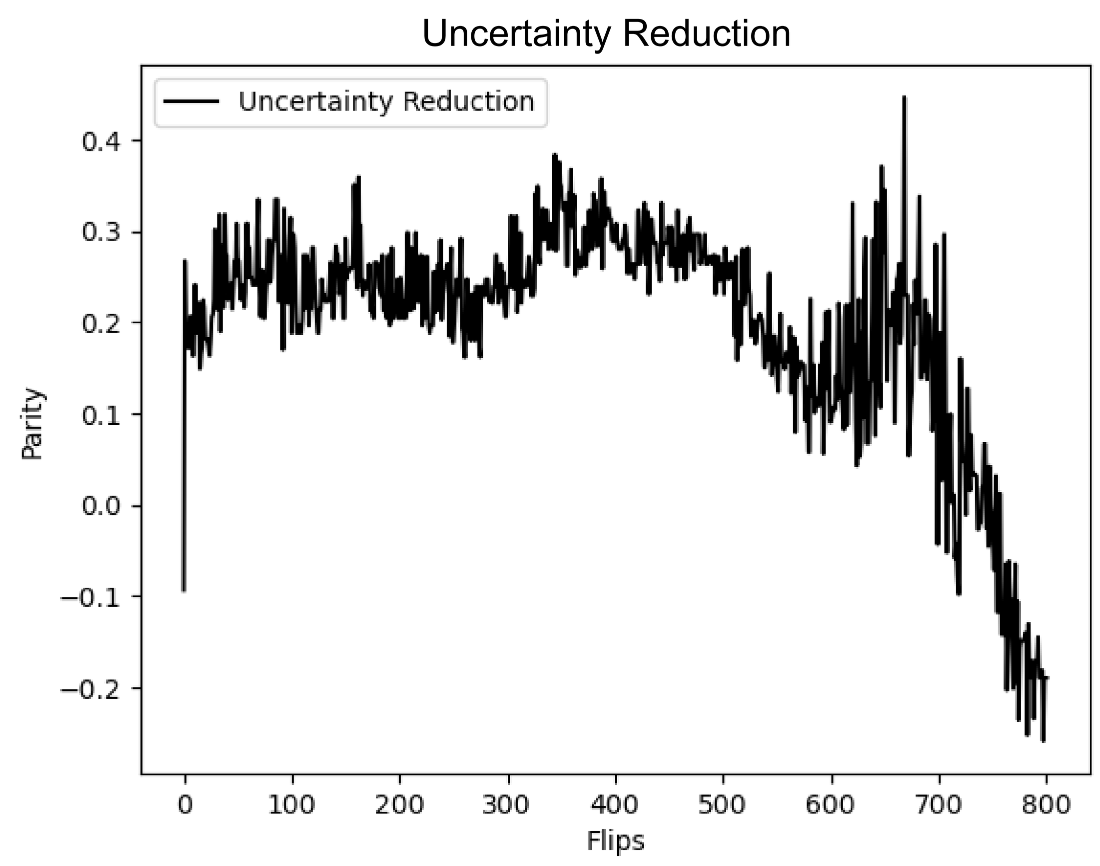
- F_result = (P(*y*) * *F*) + (P(*y_S*) * *F_S*)
- Rank instances in increasing order of expected utility

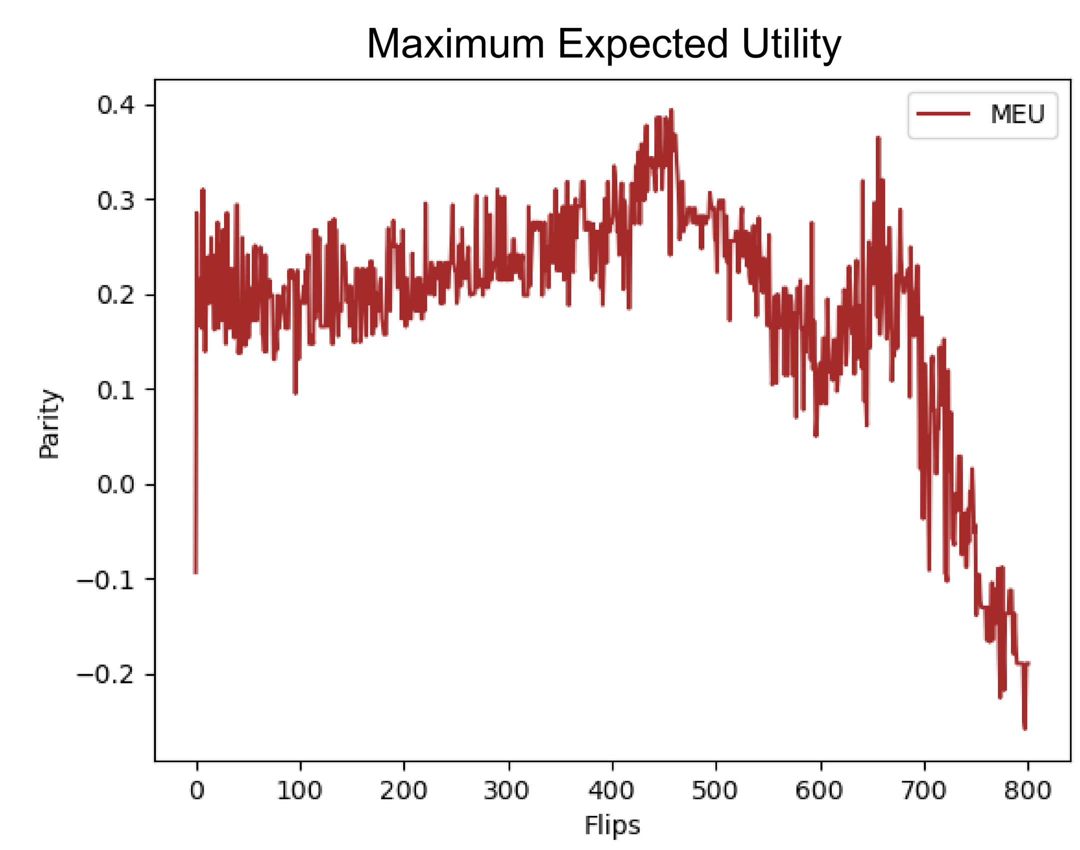




Experimental Evaluation

Dataset	Size	Classification Task	Sensitive Attribute
German Credit	1K	Is an individual a good or bad credit risk?	Age
ACSIncome	1.67M	Does an individual earn ≥ 50K annually?	Sex
COMPAS	7.2K	Is a convict at a low or high risk to recidivate?	Race





Conclusion

- The proposed solutions effectively identify a minimal fraction of training data instances whose label should be flipped to mitigate bias in the learned model's predictions.
- Methods based on entropy and expected utility are the most effective in determining the order in which the labels of training data instances should be flipped.
- Label flipping is effective in mitigating model bias and is a relatively less intrusive pre-processing bias mitigation technique.

CENTER FOR ADVANCING SAFETY OF MACHINE INTELLIGENCE

A collaboration between Northwestern University and UL Research Institutes