Techniques To Reduce Bias in Training Datasets: A Survey in Fairness in Artificial Intelligence

VIVEK SANTHOSH RAI
Department of information technology, Ramnarain Ruia Autonomous College

Abstract- Artificial Intelligence (AI) and Machine Learning (ML) systems are increasingly used in sensitive areas like healthcare, recruitment, finance, and law enforcement. However, people often question the fairness of these systems due to biases in training datasets. These biases come from issues like sampling errors and historical prejudice. They can carry through algorithms and cause unfair or discriminatory outcomes. This survey reviews current methods to reduce dataset bias in ML models. The study divides these methods into pre-processing, in-processing, and post-processing approaches and compares their strengths and weaknesses. The paper notes recent developments in fairness-focused ML and offers insights into the trade-offs between model performance and fairness. It wraps up with a discussion on future research opportunities to create more fair and transparent AI systems.

Index Terms- Artificial Intelligence, Bias Mitigation, Dataset Fairness, Ethical Machine Learning, Responsible AI

I. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have become key technologies in today's world, driving changes in finance, healthcare, education, and governance. Unfortunately, biased training datasets can create flawed models that continue discrimination, eroding trust in AI systems. Examples, like gender bias in hiring algorithms or racial bias in judicial risk assessments, highlight the urgent need to tackle fairness in AI.

Bias often stems from historical prejudice, unbalanced datasets, or poor data collection processes. As AI systems become more common in decision-making, fairness in models becomes a societal and ethical issue, not just a technical one. This survey explores leading techniques to reduce bias in training datasets, grouping them into three main strategies: pre-processing, in-processing, and post-processing.

II. RESEARCH ELABORATION

A. Sources of Bias in Datasets

1. Sampling Bias: Not enough representation of minority groups.
2. Historical Bias: Existing inequalities in past data.
3. Measurement Bias: Problems in data labeling and collection.
4. Algorithmic Bias: Exaggeration of bias during model training.

B. Techniques to Reduce Bias

Bias reduction methods generally fit into three categories:

1. Pre-processing Techniques

- Re-sampling: Either increasing underrepresented groups or decreasing dominant groups.
- Re-weighting: Assigning weights to balance the contributions of different groups.
- Data Augmentation: Creating synthetic samples for minority classes.
- Fair Data Representation: Changing data into a new form that lessens bias.

2. In-processing Techniques

- Fairness Constraints: Adding rules during model optimization to ensure fairness.
- Adversarial Debiasing: Training models with adversarial networks that penalize biased predictions.
- Regularization: Using fairness-focused loss functions.

3. Post-processing Techniques

- Threshold Adjustment: Changing decision boundaries to achieve equal outcomes.
- Equalized Odds and Demographic Parity Adjustments: Balancing fairness metrics among groups.
- Calibrated Equal Odds: Finding a balance between predictive performance and fairness.

C. Tools and Frameworks

Several open-source tools have been created to help ensure fairness in ML, including:

- AI Fairness 360 (IBM): Offers metrics and debiasing algorithms.
- Fairlearn (Microsoft): Focuses on fairness evaluation and reduction.
- What-If Tool (Google): Visual tool for fairness assessment.

III. RESULTS

From reviewing the literature and applications, we gather these insights:

- Pre-processing techniques work well when dataset bias is obvious but can lessen data richness.
- In-processing techniques give more direct fairness guarantees but add computational costs and complexity.
- Post-processing techniques are easy to implement but might compromise model interpretability.
- Fairness vs. Accuracy Trade-off: Striving for perfect fairness often reduces accuracy, requiring careful balancing based on the specific application. Case studies show that combining different techniques usually leads to better results than depending on a single method.

CONCLUSION

Bias in training datasets is a major challenge in the field of Artificial Intelligence. This survey categorizes and compares existing mitigation techniques, showing that no single method fits all. The choice of technique depends on the application context, fairness needs, and performance limits. Future research should focus on:

1. Creating combined debiasing approaches that use multiple techniques.
2. Developing clear fairness models for better transparency.
3. Examining how debiasing methods can work across different domains.

By tackling dataset bias, AI can move closer to being more inclusive, ethical, and socially responsible.

ACKNOWLEDGEMENT

The author thanks Ramnarain Ruia Autonomous College and the faculty of the Department of Information Technology for their guidance in preparing this paper.

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 (CSUR), 54(6), 1-35.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. Cambridge: fairmlbook.org.

[3] Bellamy, R. K., et al. (2019). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4-1

[4] Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

[5] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of Innovations in Theoretical Computer Science (ITCS).