Investigating the Stability and Evaluation Challenges of Sample Selection Bias in Supervised Learning

Authors:

Eyad Salama – 120250025
Noor Alrahel – 220252393
Hamed Musallam – 120241667

Supervised Machine Learning , Dr. Ayman Maliha , Master of Computer Engineering
The Islamic University of Gaza , December 2025

1. Introduction

Supervised learning traditionally operates under the assumption that training and test data are drawn from the same distribution, commonly referred to as the independent and identically distributed (i.i.d.) assumption. This assumption simplifies model training and evaluation; however, it rarely holds in real-world scenarios. In practice, data collection is often influenced by non-random processes such as accessibility, cost, and human behavior, particularly in domains like healthcare and social sciences.

These non-random sampling mechanisms give rise to *sample selection bias*, where the training data fails to accurately represent the target population on which the model will ultimately be deployed. As a result, supervised learning models may exhibit strong performance during training and validation while performing poorly in real-world deployment **[5]**. More critically, evaluation metrics computed on biased datasets can lead to overly optimistic and misleading performance estimates, masking the true reliability and generalization capability of the model. Consequently, sample selection bias represents a fundamental challenge to the trustworthiness and practical applicability of supervised learning systems.

2. Problem Definition

Sample selection bias occurs when the data available for training a model does not represent the target population it is intended to serve **[4]**. This mismatch leads to biased parameter estimation, where the model may exhibit high accuracy on the biased training set but fails significantly when applied to the broader, unbiased population. The core problem is how to maintain predictive performance when the "sampling mechanism" is not independent of the outcome variables **[4]**.

3. Research Gap and Motivation

While various correction techniques like "Importance Weighting" have been proposed, a critical gap exists regarding their stability in finite-sample settings **[1]**. Current methods often suffer from high variance, where a few samples are given extreme weights, leading to unstable gradients and poor generalization **[2]**. The motivation for this research is to address the lack of robust evaluation protocols and the failure of existing models under severe bias, especially in high-stakes domains like healthcare where unrealistic assumptions about data distributions can lead to harmful decisions **[5]**.

4. Research Question

How does severe sample selection bias impact the stability and generalization of supervised learning models, and what are the fundamental limitations of current mitigation strategies in realistic, finite-sample settings?

5. Research Objectives

The primary objectives of this research are:

- To analyze how varying degrees of sample selection bias influence model stability **[1]**.
- To investigate the limitations of existing mitigation strategies under severe bias **[2]**.
- To assess the impact of selection bias on performance evaluation **[3]**.
- To identify open research challenges in real-world biased learning settings.

6. Conclusion

Understanding the intricate dynamics of sample selection bias is crucial for developing trustworthy AI systems. This research focuses on diagnosing the instability and evaluation pitfalls inherent in current methods. By identifying these limitations, we lay the groundwork for more robust learning frameworks that can perform reliably in the face of complex, real-world data distributions.

7. References

· **[1]** C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, "Sample Selection Bias Correction Theory," 2008.

· **[2]** F. Liu and B. Ziebart, "Robust Classification under Sample Selection Bias," 2014.

· **[3]** G. C. Cawley and N. L. C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," 2010.

· **[4]** C. Winship and R. D. Mare, "Models for Sample Selection Bias," 1992.

· **[5]** V. K. Chauhan, et al., "Sample Selection Bias in Machine Learning for Healthcare," 2025.