A Survey on 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

Abstract

Supervised learning methods are usually developed under the assumption that training and test data originate from the same distribution. In practical applications, however, this assumption often does not hold. One common reason is selection bias, where the available training data fails to reflect the characteristics of the actual target population. In such cases, models may appear to perform well during evaluation while encountering difficulties when applied in real-world settings.

Motivated by these practical challenges, this survey explores selection bias as a central issue in supervised learning. We discuss its main causes and commonly observed forms, examine how it affects model generalization and evaluation, and review several mitigation strategies proposed in the literature, including sample reweighting and robust learning techniques. We also highlight limitations of existing approaches and outline open challenges that suggest why selection bias remains an important topic for ongoing research.

1. Introduction

Supervised learning plays a major role in modern machine learning and forms the basis of many real-world systems, including medical diagnosis, recommendation platforms, and decision-support applications. In such settings, models are trained on labeled data with the expectation that learned patterns will remain valid when new data is encountered. This expectation, however, strongly depends on how representative the training data is of the environment in which the model will eventually be used.

However, in practical settings, data is rarely collected through fully random sampling processes. Constraints related to data availability, cost, accessibility, and human behavior often influence which samples are included in the training dataset. These non-random data collection mechanisms give rise to selection bias, where certain groups or patterns are overrepresented while others are underrepresented or entirely missing.

The presence of selection bias can significantly reduce the reliability of supervised learning models. Even when common evaluation metrics such as accuracy or loss indicate strong performance, the learned model may fail to generalize beyond the biased training distribution. This gap between training conditions and real-world deployment is particularly concerning in high-stakes domains such as healthcare, finance, and policy-making.

Selection bias has been extensively studied in statistics and machine learning, and numerous techniques have been proposed to detect and mitigate its effects. Common approaches include sample reweighting, importance weighting, and robust learning frameworks. Despite these efforts, existing methods often rely on strong assumptions about the data generation or selection process, which are difficult to satisfy in real-world scenarios.

In this survey, we systematically examine selection bias in supervised learning. We review its underlying causes, categorize its main types, analyze its impact on model performance and evaluation, and discuss existing mitigation strategies along with their limitations. By synthesizing findings from prior research, this survey aims to clarify why selection bias remains a persistent challenge and to identify directions for future research.

2. Problem Definition: Selection Bias in Supervised Learning

Selection bias generally occurs when the data available for training does not fully reflect the characteristics of the population on which a supervised learning model is intended to operate. In many practical scenarios, this issue emerges not because of modeling choices, but as a consequence of how data is collected and filtered before training even begins. This phenomenon occurs due to non-random data collection processes, where the probability of a data point being included in the training set depends on certain characteristics of the sample rather than being uniformly random [1].

From a probabilistic perspective, most supervised learning algorithms assume that training and test data are independently and identically distributed (i.i.d.). However, under selection bias, the distribution of the input features in the training data differs from that of the deployment environment, even if the conditional distribution of labels given features remains unchanged. This violation of the i.i.d. assumption can significantly affect the reliability of learned models [2].

Selection bias should be distinguished from related issues such as class imbalance and overfitting. While class imbalance refers to unequal class frequencies and overfitting arises from excessive model complexity relative to the data size, selection bias is fundamentally a data acquisition problem. It originates before model training begins and cannot be resolved solely through algorithmic adjustments or increased model capacity [3].

Prior research has shown that selection bias can lead to overly optimistic performance estimates during evaluation, particularly when training and validation data are drawn from the same biased source. As a result, models may appear to perform well under standard evaluation protocols while failing catastrophically when applied to real-world data [4].

3. Types of Selection Bias

Selection bias is not a single, uniform phenomenon but rather appears in multiple forms depending on how the data collection process deviates from random sampling. Understanding these different types is essential for analyzing their impact on supervised learning models and for selecting appropriate mitigation strategies. In this section, we outline the most common types of selection bias discussed in the machine learning literature.

While several forms of selection bias are discussed for completeness, this survey primarily focuses on sample selection bias, as it represents the most commonly studied and practically relevant form in supervised learning literature.

Sample Selection Bias occurs when the probability of including a data instance in the training set depends directly on the input features. In this setting, some regions of the input space are overrepresented while others are underrepresented or completely absent. This type of bias is commonly observed in real-world applications where data is collected based on availability or accessibility rather than representativeness [1].

Underrepresentation Bias arises when certain subpopulations are systematically missing or sparsely represented in the training data. Unlike general sample selection bias, underrepresentation specifically affects minority groups or rare conditions. Models trained under such conditions may perform adequately for majority groups but fail to generalize to underrepresented populations, leading to unreliable and potentially unfair predictions [4].

Another important form of selection bias is related to **Missing Not At Random (MNAR)** data. In this case, the absence of data is itself dependent on unobserved variables or the target outcome. For example, instances with extreme outcomes may be less likely to be recorded. Standard learning algorithms typically assume missing data to be random, and violations of this assumption can introduce significant bias that cannot be easily corrected through simple imputation techniques [2].

These different types of selection bias share a common characteristic: they distort the empirical data distribution observed during training, causing a mismatch between training and deployment environments. As a result, even well-regularized models may exhibit poor generalization when faced with unbiased real-world data.

4. Impact of Selection Bias on Supervised Learning

Selection bias can influence supervised learning models in ways that are not always immediately visible during evaluation. While performance metrics may appear satisfactory, underlying biases in the training data can lead to unexpected behavior once the model is exposed to real-world conditions. One of its most critical consequences is the degradation of model generalization, where a model trained on biased data performs poorly when applied to data drawn from the true target distribution. This occurs because the learned decision boundaries reflect patterns present in the biased training data rather than the underlying structure of the real-world population [1].

A particularly challenging aspect of selection bias is that it often leads to misleading evaluation results. When both training and validation datasets are drawn from the same biased source, standard evaluation metrics such as accuracy, precision, and loss may indicate strong performance. However, these metrics fail to capture the discrepancy between the biased evaluation setting and the unbiased deployment environment, resulting in overly optimistic performance estimates [3].

Selection bias can also amplify unfairness in supervised learning systems. Models trained on data that underrepresent certain subpopulations tend to favor majority groups, producing systematic errors for minority or rare cases. Such behavior is especially problematic in sensitive domains such as medical diagnosis or decision support systems, where biased predictions can have serious real-world consequences [4].

Furthermore, selection bias complicates the interpretation of model behavior and limits the effectiveness of traditional regularization techniques. While regularization methods are designed to prevent overfitting to noise, they do not address distributional mismatches caused by biased data collection. As a result, even well-regularized models may fail under selection bias, highlighting the need for bias-aware learning and evaluation strategies [2].

5. Mitigation Strategies for Selection Bias

A wide range of methods has been proposed to mitigate the effects of selection bias in supervised learning. Most of these approaches aim to reduce the mismatch between the biased training distribution and the target deployment distribution, either by reweighting training samples or by modifying the learning objective to account for uncertainty caused by biased data.

Several existing approaches, such as covariate shift correction and domain adaptation, attempt to mitigate selection bias by aligning the distributions of the training data and the target population. While these methods are theoretically well-motivated, recent studies have suggested that enforcing such alignment may unintentionally lead to a loss in predictive performance, particularly in practical deployment scenarios where the target distribution is only partially observed [5].

An alternative line of research addresses selection bias by explicitly modeling the data selection process as a separate learning task. In this setting, the learning framework distinguishes between identifying whether an instance belongs to the study population and performing the primary prediction task. By accounting for the selection mechanism directly, these approaches aim to reduce bias without fully forcing distribution alignment, which may help preserve predictive performance in certain real-world applications [5].

One of the most commonly used approaches is **sample reweighting**, also known as importance weighting. In this method, training instances are assigned weights proportional to the ratio between the target distribution and the training distribution. By emphasizing underrepresented samples and down-weighting overrepresented ones, reweighting attempts to approximate unbiased learning despite biased data collection [1].

Despite its theoretical appeal, importance weighting suffers from practical limitations. When selection bias is severe, estimated weights can have high variance, causing learning algorithms to overemphasize a small number of samples. This often leads to unstable models and poor generalization performance, particularly in finite-sample settings [2].

Another class of approaches focuses on **robust learning**, where models are explicitly designed to perform well under uncertainty about the data distribution. Rather than assuming that the reweighted training data accurately reflects the target distribution, robust methods aim to minimize worst-case loss under plausible distribution shifts. These approaches have shown improved stability under selection bias but typically require strong modeling assumptions or additional unlabeled data [2].

More recently, researchers have emphasized the importance of **bias-aware evaluation protocols**. Techniques such as nested cross-validation and careful separation of model selection and evaluation stages have been proposed to avoid overly optimistic performance estimates caused by biased data reuse. While these methods do not eliminate selection bias itself, they provide more reliable assessments of model performance under realistic conditions [3].

6. Limitations and Open Challenges

Despite extensive research on mitigating selection bias, existing approaches suffer from several fundamental limitations. A primary challenge is that many mitigation techniques rely on strong assumptions about the data generation or selection process. In practice, these assumptions are often violated, as the mechanisms governing data collection are complex, partially unknown, or influenced by external human and social factors [1].

Another significant limitation lies in the estimation of importance weights. Reweighting-based methods require accurate estimates of the ratio between the target distribution and the biased training distribution. However, when this ratio is estimated from finite samples, errors in estimation can introduce high variance and instability, ultimately degrading model performance rather than improving it [2].

Selection bias also poses substantial challenges for model evaluation. In many real-world scenarios, unbiased validation data is unavailable, making it difficult to assess true model performance. As a result, researchers and practitioners may unknowingly deploy models whose reported accuracy does not reflect their behavior in deployment environments [3].

These limitations highlight several open research challenges. Developing methods that can effectively address selection bias under minimal assumptions, designing evaluation protocols that remain reliable in the presence of biased data, and integrating causal reasoning into supervised learning frameworks remain open and important directions for future research. Addressing these challenges is essential for building robust and trustworthy machine learning systems.

7. Conclusion

Selection bias remains a persistent challenge in supervised learning, largely because it originates from practical data collection decisions rather than purely technical modeling errors. As highlighted throughout this survey, such biases can subtly affect training, evaluation, and deployment, even when models appear to perform well under standard testing procedures.

As demonstrated throughout this survey, selection bias can significantly distort model training, evaluation, and deployment, leading to unreliable and potentially harmful outcomes despite seemingly strong empirical performance.

This survey reviewed the primary causes and types of selection bias, examined its impact on model generalization and fairness, and discussed commonly used mitigation strategies along with their inherent limitations. While existing methods provide partial solutions, they often depend on assumptions that are difficult to satisfy in real-world settings.

Ultimately, addressing selection bias requires a combination of careful data collection practices, bias-aware evaluation protocols, and learning methods that explicitly account for uncertainty and distributional mismatch. Continued research in this area is critical to ensuring that supervised learning systems remain robust, fair, and trustworthy when deployed in practical applications.

8. References

[1] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh,
"Sample Selection Bias Correction Theory,"
in Proceedings of the 19th International Conference on Algorithmic Learning Theory (ALT),
Lecture Notes in Artificial Intelligence, vol. 5254, pp. 38–53, 2008.

[2] F. Liu and B. Ziebart,
"Robust Classification under Sample Selection Bias,"
in Advances in Neural Information Processing Systems (NeurIPS),
2014.

[3] G. C. Cawley and N. L. C. Talbot,
"On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation,"
Journal of Machine Learning Research, vol. 11, pp. 2079–2107, 2010.

[4] A. M. S. Rahman, T. Ahmed, and M. R. Amin,
"Revisiting Sampling Bias in Machine Learning,"
IEEE Access, vol. 7, pp. 134123–134135, 2019.

[5] V. K. Chauhan, L. Clifton, A. Salaün, H. Y. Lu, K. Branson,
P. Schwab, G. Nigam, and D. A. Clifton,
"Sample Selection Bias in Machine Learning for Healthcare,"
ACM Transactions on Computing for Healthcare,
vol. 6, no. 4, Article 52, Oct. 2025.