RADemics

Supervised Learning Techniques for Malware Detection and Classification

Rajesh M, Supriya R K, Dr Syed Naimatullah Hussain

Vellore Institute of Technology, Dayananda Sagar Academy of Technology and Management, Nagarjuna college of engineering and technology

Supervised Learning Techniques for Malware Detection and Classification

[1]Rajesh M, Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600 127. rajesh.mvr@vit.ac.in

[2]Supriya R K, Assistant professor, Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore, Supriya-ise@vkvtm.edu.in

[3]Syed Naimatullah Hussain, professor & HoD, CSE (Data Science), Nagarjuna college of engineering and technology, near Devanahalli Bengaluru, Mail ID: syed.hussain@gmail.com

Abstract

Malware detection has emerged as a critical challenge in the domain of cybersecurity due to the increasing sophistication and volume of cyber threats. Traditional methods, such as signature-based detection, often fail to identify novel and polymorphic malware, leading to a growing interest in machine learning techniques, particularly supervised learning, for more robust and adaptive detection systems. This chapter explores the role of supervised learning in the detection and classification of malware, with a particular focus on addressing the challenges posed by imbalanced datasets. The impact of data imbalance on classifier performance is thoroughly examined, highlighting the importance of advanced metrics like balanced accuracy, and the use of resampling techniques such as oversampling, undersampling, and hybrid methods to mitigate bias. Additionally, the chapter discusses the integration of cost-sensitive learning approaches to prioritize error minimization, emphasizing the trade-off between accuracy and risk management in the context of cybersecurity. Through a comprehensive analysis of current methodologies, challenges, and emerging trends, this chapter provides a detailed overview of the state-of-the-art in malware detection using supervised learning, offering insights into future directions for improving detection accuracy and generalization. Key concepts such as malware classification, data imbalance, supervised learning, resampling techniques, cost-sensitive learning, and balanced accuracy are explored, providing readers with a thorough understanding of the strategies employed in modern malware detection systems.

Keywords: malware classification, data imbalance, supervised learning, resampling techniques, cost-sensitive learning, balanced accuracy

Introduction

Malware detection has become one of the most critical aspects of cybersecurity in recent years [1]. As cyber threats continue to grow in sophistication and scale, traditional detection methods such as signature-based approaches have shown limitations in identifying new or unknown malware variants [2]. Signature-based systems rely on predefined patterns and known threats, making them ineffective against novel or polymorphic malware [3]. This challenge has led to a paradigm shift toward the application of machine learning, particularly supervised learning models, which offer a more dynamic and adaptive approach to detecting malicious software [4].

Supervised learning models learn from labeled datasets, identifying patterns and features that distinguish benign files from malicious ones, providing a more effective means to identify evolving threats [5].

One of the most significant challenges in developing supervised learning-based malware detection systems is the issue of data imbalance [6]. In typical malware detection datasets, the number of benign files far exceeds the number of malware instances, creating a class imbalance that significantly skews model performance [7]. When trained on such imbalanced data, models tend to develop a bias toward the majority class (benign files), often misclassifying malicious samples as benign [8]. This issue results in a significant reduction in the sensitivity of the model to detect malware [9]. The imbalance leads to a situation where even a model with high overall accuracy might fail to identify the minority class effectively, which is crucial in real-world malware detection systems where missing a malicious file can have severe consequences [10].

To address this challenge, various resampling techniques have been proposed to modify the distribution of the dataset during model training [11]. Oversampling techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), generate synthetic samples from the minority class (malware) to balance the dataset, thereby allowing the model to focus more on detecting malware [12]. Conversely, undersampling methods reduce the number of benign samples to create a more balanced dataset [13]. While these techniques have proven effective in reducing bias, they are not without limitations [14]. For example, oversampling can introduce overfitting by generating redundant samples, while undersampling may result in the loss of valuable benign data, potentially impacting the model's ability to generalize. Hybrid approaches, which combine both oversampling and undersampling, are also explored to mitigate these challenges [15].