RADemics

Role of Feature Selection and Dimensionality Reduction in Hybrid Learning Systems for Threat Detection

V.Samuthira Pandi, Shobana D, R Geetha
Chennai Institute of Technology, Rajalakshmi engineering college, R.M.K. Engineering College.

3. Role of Feature Selection and Dimensionality Reduction in Hybrid Learning Systems for Threat Detection

[1]V. Samuthira Pandi, Department of ECE, Centre for Advanced Wireless Integrated Technology,Chennai Institute of Technology, Chennai. samuthirapandiv@citchennai.net

[2]Shobana D, Department of Mechatronics, Rajalakshmi engineering college, shobana.d@rajalakshmi.edu.in

[3]R Geetha, Associate Professor, Department of Computer Science and Engineering, R.M.K. Engineering College RSM Nagar, Kavaraipettai, Thiruvallur District, 601206. rga.cse@rmkec.ac.in

Abstract

In the rapidly evolving landscape of cybersecurity, the increasing volume and complexity of data present significant challenges in detecting novel threats and anomalies. Feature selection and dimensionality reduction techniques play a pivotal role in enhancing the performance and efficiency of machine learning models used for threat detection. This chapter explores the critical importance of these techniques in hybrid learning systems, focusing on their application in anomaly detection, intrusion detection, and advanced persistent threat identification. The synergy between feature selection methods, including filter, wrapper, and embedded approaches, with dimensionality reduction techniques such as autoencoders and neural networks, was discussed to improve model accuracy and reduce computational costs. The chapter addresses the challenges faced by these techniques in high-dimensional data spaces, including overfitting, class imbalance, and computational complexity. By examining the integration of feature selection and dimensionality reduction, this work provides insights into the development of more robust, scalable, and interpretable threat detection systems. The chapter emphasizes the need for novel approaches to optimize feature subsets and effectively reduce dimensionality, enabling more efficient and accurate cybersecurity solutions in real-world applications.

Keywords: Feature selection, dimensionality reduction, threat detection, hybrid learning systems, anomaly detection, autoencoders.

Introduction

The rapid expansion of digital technologies and the increasing volume of data generated by organizations present significant challenges to cybersecurity [1]. Traditional methods of threat detection, such as signature-based approaches, are increasingly ineffective in identifying sophisticated, novel, or previously unseen threats [2]. Attackers are constantly developing new tactics, making it difficult for conventional detection systems to keep pace with evolving cyber risks [3]. As a result, there is a pressing need for more advanced and adaptive threat detection systems that can effectively identify and mitigate these evolving threats [4]. One of the key obstacles in this domain is the high dimensionality of the data, which can obscure important patterns, leading to reduced model performance and increased computational complexity [5]. Therefore, reducing the dimensionality of the data while preserving critical information becomes essential for improving the efficiency and effectiveness of threat detection systems [6].

Feature selection was an essential step in the process of building robust machine learning models for threat detection [7]. In the context of cybersecurity, feature selection helps identify the most relevant variables from large and complex datasets, ensuring that only the most informative features are utilized for model training [8]. The primary objective of feature selection is to reduce the dimensionality of the data by eliminating irrelevant or redundant features, which not only enhances computational efficiency but also improves model accuracy [9]. Feature selection techniques can be broadly classified into three categories: filter methods, wrapper methods, and embedded methods [10]. Each of these approaches offers distinct advantages and challenges, depending on the nature of the data and the type of threat detection task being addressed [11]. By selecting the optimal subset of features, feature selection techniques allow machine learning models to focus on the most important information, reducing the risk of overfitting and improving generalization to new data [12].

While feature selection is crucial for identifying the most relevant features in a dataset, dimensionality reduction techniques provide a complementary approach for compressing high-dimensional data into lower-dimensional spaces [13]. Unlike feature selection, which involves the removal of individual features, dimensionality reduction techniques aim to transform the data into a more compact form, retaining as much meaningful information as possible [14]. Non-linear dimensionality reduction methods, such as autoencoders and neural networks, are particularly well-suited for complex datasets commonly encountered in threat detection [15]. These techniques are capable of capturing non-linear relationships and intricate patterns within the data, making them more effective for cybersecurity tasks where the data is often non-linear and highly variable [16]. Autoencoders, for example, learn to encode input data into a lower-dimensional representation and then decode it back into the original space, enabling the model to extract the most significant features while discarding irrelevant information [17]. By leveraging such non-linear dimensionality reduction techniques, threat detection systems can better handle high-dimensional, noisy data and improve their ability to detect both known and novel cyber threats [18].

Incorporating both feature selection and dimensionality reduction techniques into a hybrid learning system offers a promising solution for enhancing the performance of threat detection models [19]. Hybrid models combine the strengths of multiple approaches, allowing for more efficient processing and better generalization to new, unseen data [20]. Combining filter-based feature selection methods with neural network-based dimensionality reduction can improve the robustness and scalability of the threat detection system [21]. Feature selection can be used to eliminate irrelevant or redundant features before applying dimensionality reduction techniques, which helps reduce the computational burden and ensures that the model focuses on the most informative data [22]. By integrating these techniques within a hybrid framework, the resulting model can achieve superior performance in identifying a wide range of cyber threats, including both known attacks and emerging anomalies. Hybrid models can be adapted to different types of data, from network traffic logs to user behavior analytics, offering a flexible solution for various threat detection tasks [23].

High-dimensional data is often noisy, sparse, and imbalanced, which can complicate the feature selection process and diminish the effectiveness of dimensionality reduction [24]. Selecting the optimal subset of features and the appropriate dimensionality reduction technique for a given task can be a complex and computationally expensive process. Overfitting remains a significant concern, particularly when dealing with small or unbalanced datasets. Careful consideration must be given to the choice of techniques and their integration into the overall threat detection pipeline. Advances in hybrid learning systems, including the use of ensemble methods and advanced neural network architectures, offer promising solutions to address these challenges [25]. By leveraging the strengths of multiple techniques, researchers and practitioners can develop more efficient, accurate, and scalable threat detection models that are capable of adapting to the ever-changing cybersecurity landscape.