RADemics

Hybrid Supervised and Unsupervised Learning Models for Identifying Network Anomalies

D.Kanchana, Hubert Mary.L, A.Thilagavathy
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY,
JEPPIAAR INSTITUTE OF TECHNOLOGY, R.M.K.
ENGINEERING COLLEGE

4. Hybrid Supervised and Unsupervised Learning Models for Identifying Network Anomalies

[1]D.Kanchana,Assistant Professor,Department of Computer Science and Applications,SRM Institute of Science and Technology,Ramapuram campus, Chennai, Tamil Nadu, India. kanchand@srmist.edu.in

[2].Hubert Mary.L, Assistant Professor , Department of ECE, Jeppiaar Institute of Technology hubertmaryl@gmail.com .

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

Abstract

Anomaly detection in network systems was a critical task for ensuring security and stability in the face of ever-evolving threats. Hybrid models that combine clustering and classification techniques have emerged as effective solutions for identifying anomalies with higher accuracy and efficiency. This chapter explores the integration of supervised and unsupervised learning methods to address the complexities of detecting network anomalies. The synergy between clustering algorithms, such as K-means and DBSCAN, and classification models, including support vector machines (SVM) and random forests, enhances the model's capability to detect both known and novel threats. Key challenges, such as handling data imbalance, optimizing model parameters, and feature engineering, are discussed in the context of hybrid models. Additionally, the chapter examines the use of statistical and domain-specific features to improve detection accuracy and reduce false positives. Practical case studies highlight the application of hybrid models in real-world network environments, demonstrating their effectiveness in scenarios such as intrusion detection and DDoS attack identification. The balance between detection accuracy and computational efficiency is also critically evaluated, providing insight into the practical deployment of hybrid models in large-scale network systems. This chapter offers a comprehensive framework for researchers and practitioners aiming to develop robust, scalable, and efficient hybrid models for network anomaly detection.

Keywords: Hybrid Models, Anomaly Detection, Clustering, Classification, Network Security, Data Imbalance.

Introduction

Anomaly detection was a crucial element in modern network security, as it helps identify suspicious activities that deviate from normal behavior, which could indicate potential threats such as cyberattacks, intrusions, or fraud [1]. With the increasing complexity of network systems and the diversity of attacks targeting them, traditional detection methods often fall short in accurately identifying unknown or evolving threats [2]. This limitation has driven the development of advanced techniques that combine both supervised and unsupervised learning methods, known as

hybrid models [3]. These models offer a promising solution by enhancing the capabilities of individual algorithms through synergy, thereby improving the detection of network anomalies with higher precision and efficiency [4].

The fusion of clustering and classification techniques is particularly effective in hybrid models [5]. Clustering algorithms like K-means and DBSCAN excel in grouping data based on inherent similarities, allowing the identification of previously unknown anomalies that might not align with predefined attack signatures [6]. On the other hand, classification models, such as support vector machines (SVM) or decision trees, are well-suited for distinguishing between normal and anomalous data, based on labeled training data [7]. The integration of these methods allows for a multi-faceted approach, where clustering can uncover potential outliers, and classification helps refine the decision-making process to ensure accurate anomaly detection [8]. This complementary relationship between clustering and classification provides a more robust and adaptable solution for anomaly detection in dynamic network environments [9].

In most real-world network datasets, the majority of traffic consists of normal behavior, while anomalies are rare and often difficult to identify [10]. This imbalance can lead to biased models, where the algorithm is more likely to classify data as normal, thus overlooking anomalies [11]. To mitigate this issue, various strategies, such as resampling techniques (e.g., oversampling the minority class or undersampling the majority class) and cost-sensitive learning approaches, are employed [12]. These methods ensure that the minority class (anomalies) is adequately represented during the training process, improving the model's ability to detect and classify rare events without compromising accuracy [13].

Another challenge associated with hybrid models is the optimization of model parameters and algorithm selection [14]. The performance of clustering and classification techniques can be highly sensitive to the parameters chosen, such as the number of clusters in K-means or the regularization parameter in SVM [15]. The interplay between the clustering and classification phases requires careful consideration to ensure that the model is both accurate and efficient [16]. For instance, the clustering phase should effectively identify potential anomalies, while the classification phase should refine this identification to minimize false positives and false negatives [17]. Optimizing these parameters is critical for achieving a balance between detection accuracy and computational efficiency, particularly in large-scale network systems where both speed and precision are essential [18].

Feature engineering is another crucial aspect of hybrid anomaly detection models [19]. The choice of features used for clustering and classification can significantly impact the model's performance [20]. While traditional features such as packet size, inter-arrival time, and protocol type are commonly used in network anomaly detection, domain-specific features tailored to particular network environments or attack types can enhance the model's ability to detect subtle or complex anomalies [21-22]. Additionally, statistical features, such as mean, standard deviation, and skewness, can help capture important patterns in the data, improving the clustering process and providing valuable inputs for the classification phase. By selecting and engineering the right features, hybrid models can be better equipped to handle diverse types of anomalies in network traffic [23].

The evaluation of hybrid models requires a comprehensive analysis of both detection performance and computational efficiency [24]. It is essential to assess the model's ability to accurately identify anomalies (through metrics like precision, recall, and F1-score) while

considering the computational cost of running these models in real-time or at scale. Detection accuracy ensures that the system reliably identifies malicious activity without generating excessive false alarms, while computational efficiency ensures that the model can be deployed in dynamic environments without causing delays or straining system resources. Achieving a balance between these two factors was key to the success of hybrid models in network anomaly detection, especially as networks grow in size and complexity, requiring scalable solutions that remain effective over time [25].

, 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.