GEORGE WASHINGTON UNIVERSITY

ADA UNIVERSITY

COMPUTER SCIENCE AND DATA ANALYTICS

GUIDED RESEARCH I

Report 4

Asiman Mammadzada

Title: Network Intrusion Detection System using Machine Learning

Instructors: Prof. Dr. Stephen Kaisler, Assoc. Prof Dr. Jamaladdin Hasanov

INTRODUCTION

In contemporary years, the dramatic growth of cyber threats and attacks has caused significant challenges to the security of computer systems. Network Intrusion Detection Systems (NIDS) play a crucial role in safety of these networks by identifying and responding to suspicious activities and potential security issues. Machine Learning (ML) techniques have shown great promise in enhancing the effectiveness of NIDS by enabling automated and real-time intrusion detection. This report presents the current progress of the research on NIDS using ML and shortly mentions the work done along with the planned future work.

CURRENT PROGRESS

The current progress of the research can be divided into the following sections:

1. Data Preprocessing and Labeling

To create a supervised learning framework for NIDS, I collected a comprehensive dataset containing various types of network activities from KDD dataset, including both normal and malicious behaviors. I merged different types of crimes and categorized them into a single attack label, simplifying the problem into a binary classification task.

2. Binary Classification

The first step was to train ML models to predict whether a given network instance represents an attack or not. I utilized a set of independent features, including 'duration', 'protocol_type', 'service', 'flag', 'src_bytes', 'dst_bytes', and several others, to train the model. This initial binary classification provides a foundation for identifying potential threats within network traffic.

3. Multiclass Classification

For instances classified as attacks in the binary classification, I continue the analysis to perform multiclass classification. This step involved predicting the specific attack type from a set of predefined categories, such as back, satan, ipsweep, portsweep, warezclient, and teardrop. This detailed classification allows for a better understanding of the attack patterns and can help in specifying which attack type it is.

4. Feature Selection

To enhance the performance of the ML models, I identified a subset of critical features that significantly contribute to the accurate detection of attacks. The selected features include 'protocol_type', 'service', 'src_bytes', 'dst_bytes', 'wrong_fragment', 'hot', 'logged_in', 'lnum_compromised', 'count', 'rerror_rate', 'same_srv_rate', 'dst_host_count', 'dst_host_same_srv_rate', 'dst_host_same_src_port_rate', and 'dst_host_srv_diff_host_rate' (Details in Report 3).

WHAT I HAVE DONE

At this stage of the research, I have successfully implemented and trained ML models for binary and multiclass classification of the network instances. I have identified the most relevant features for accurate detection and classification of the attacks. The performance of the models has been evaluated using standard metrics, and initial results show promising levels of accuracy.

WHAT I WILL DO

My research on NIDS using ML is ongoing, and I have outlined the following steps for future work:

1. Detection of Specific Intrusion Types

To improve the coverage of the NIDS, I plan to build additional ML models to detect specific types of intrusions, such as smurf, Neptune, pod, and nmap. These attacks has significant threats to network security and require specific detection techniques.

2. Model Optimization

I will focus on optimizing the existing ML models to achieve better accuracy and minimize false positives. This involves fine-tuning hyperparameters, exploring different ML algorithms, and applying feature engineering techniques.

3. Real-time Implementation

Efficient real-time detection is critical for the NIDS. Therefore, I will work on streamlining the ML models to operate in real-time environments with minimal time latency.

4. Dataset Expansion

To enhance the robustness and generalization capabilities of the models, I will search for augmenting the dataset with new instances by ensuring a comprehensive representation of the network activities.