INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 86-91

e-ISSN : 2583-1062

Impact Factor: 5.725

NETWORK TRAFFIC ANOMALY DETECTION USING ML

Oviya G[1], Preethi J[2], Thoufeeq. A[3]

[1,2,3]Department of Computer Science Rathinam college of arts and science, Coimbatore-0009, India.

DOI: https://www.doi.org/10.58257/IJPREMS33135

ABSTRACT

Networked computer systems are deeply integrated in every aspect of our information-overloaded modern society. The mechanisms that keep our modern society flowing smoothly, with activities such as efficient execution of government and commercial transactions and services, or consistent facilitation of social transactions among billions of users, are all dependent on large networked computer systems. Today, every aspect of our lives is influenced by networked computer systems. The Internet, which provides transportation to all types of information in- including complex real-time multi-media data, is the universal network of millions of interconnected computer systems, organized as a network of thousands of distinct smaller networks. The recent growth of the Internet has been phenomenal and consequently, the computers and the networks that make the Internet hum have become the targets of enemies and criminals. Intrusions into a computer or network system are activities that destabilize them by compromising security in terms of confidentiality, availability or integrity, the three main characteristics of a secure and stable system.

Machine learning is used to extract valid, novel, potentially useful and meaningful patterns from a dataset, usually large, in a domain of interest by using non-trivial mechanisms. A machine learning algorithm attempts to recognize complex patterns in datasets to help make intelligent decisions or predictions when it encounters new or previously unseen data instances. To deal with unseen examples, a machine learning algorithm must be cognizant of this necessity and thus, when it learns it must make conscious and diligent efforts to generalize from examples it has seen. Good generalization from data is a prime activity a learner program must perform.

Keywords: machine learning, algorithm, networking, anomaly, anomaly detection, attack recognition.

1. INTRODUCTION

An advanced network security technology involves anomaly detection of network traffic that monitors various characteristic parameters to detect unusual network behaviors.

Research indicates that these characteristic parameters are distributed differently between normal and abnormal network traffic, which makes it possible to detect anomalies by analyzing the change law of entropy value in real time. Information entropy for description of network traffic behavior features and a linear auto-regressive algorithm for anomaly detection may provide higher efficiency of network traffic analysis and better precision of anomaly discovery.

Additionally, it helps in identifying real-time network anomaly traffic and potential anomaly types. A real-time measurement of network behaviour anomaly through formulating normal network behaviour trend, comparison of such behaviour trend with historic patterns of normal network behaviour, and determination as to what extent the current behaviour eigenvalue overshoots that of the normal network behaviour eigenvalues trend. Examples of such attacks include distributed denial of service (DDoS) whereby a surge in norma3l network traffic occurs. Bandwidth exhaustion of the target network and resource exhaustion of the target system is what leads them to DDOS attacks. This allows them to send massive amounts of spoofed data from various control points at once producing many million records within a brief period. Consequently, this way consumes all available system resource up to 100%, leading to slurring response to normal users' queries and increased network load caused by invalid packets completely occupying network communications bandwidth.

In practice, DDoS attacks upon a system can be gauged using the twin indicators of actual network bandwidth and CPU/memory usage. DDOS attack against this network exhibits behavioral traits like numerous source IP addresses per single target IP address and very low success percentage in three-way handshake, comparing with normal network conduct.

The network anomaly detection model includes defining the patterns of normal network behavior and detecting deviating activity. Network behavior is considered normal by creating baseline data, digitizing data that depicts the characterization of normal network traffic, training enormous amounts of data with data mining-relevant techniques, and guaranteeing network behavior stability. Real-time detection module uses current network traffic data to extract characteristic values of current network behavior and compares them against historical normal patterns to produce a new normal pattern.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 86-91

e-ISSN : 2583-1062

Impact Factor: 5.725

2. LITERATURE SURVEY

2.1 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Research of Network Traffic Anomaly Detection Model Based on Multilevel Autoregression:

This paper presents a real-time anomaly detection model of network traffic based on information entropy. The model is zero-averaged, established using third-order linear auto regression, and the residual ratio of information entropy is proposed as the measure value of anomaly detection. The model is tested in a simulation network environment, and experiments show that it can detect more than 95% of network anomaly traffic. This approach is crucial in network security as it helps differentiate between normal and abnormal network behaviors. The model can transform the anomaly detection problem into a classification problem based on traffic characteristic entropy.

2.2 Network Traffic Analysis using Machine Learning: an unsupervised approach to understand and slice your network

The rapid growth of smart devices has led to a surge in data generation and heterogeneity, necessitating new network solutions for better traffic analysis. High-performance computing (HPC) has made it easier to deploy machine learning (ML) to solve complex problems, with its efficiency validated in various domains. Network slicing (NS) has gained significant attention from industry and academia due to its ability to address diverse service requirements. This paper focuses on analyzing network data to define network slices based on traffic flow behaviors. A feature selection method was used to select relevant features from a dataset of over 3 million instances. K-Means clustering was then applied to understand and distinguish traffic behaviors. The results showed good correlation among instances in the same cluster, which can be further integrated in real environments using network function virtualization.

3 Proposed system:

The Network Traffic Anomaly Detection System is a novel approach that uses machine learning to enhance network security and efficiency. The system aims to identify and mitigate anomalies in network traffic patterns, providing real-time threat detection and proactive response mechanisms. The system's core objective is to protect network infrastructure against malicious activities, unauthorized access, and performance disruptions. Key components of the system include data collection and preprocessing, feature engineering, machine learning models, training and validation, real-time monitoring, alerting mechanism, adaptive learning, and seamless integration with existing security infrastructure. Data is collected from various sources, including routers, switches, and monitoring devices, and preprocessed for ML algorithms. Feature engineering involves selecting relevant features to capture the distinctive patterns of normal and anomalous network behavior. Machine learning models are employed, including supervised models like Support Vector Machines (SVM), Random Forest, or Neural Networks, and unsupervised models like K-means clustering or Isolation Forests. The trained models are then integrated into the network infrastructure for continuous monitoring of live traffic, enabling prompt detection of anomalies and prompt response to potential threats.

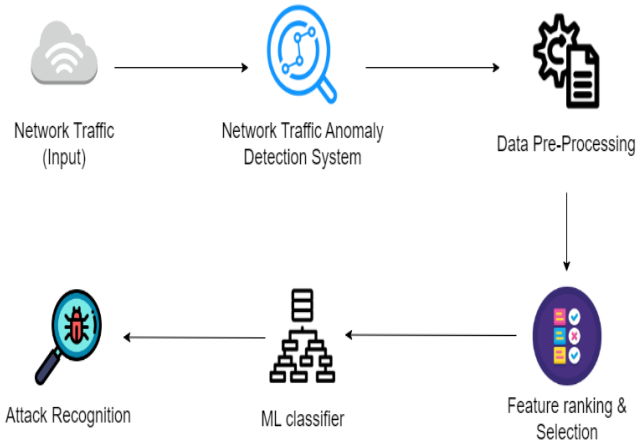

Figure 1 Architecture

Alerting mechanisms are also implemented, triggering alerts to network administrators or a Security Information and Event Management (SIEM) system when an anomaly is detected. The system incorporates adaptive learning mechanisms to continuously update and improve the ML models based on evolving network patterns.

In conclusion, the proposed Network Traffic Anomaly Detection System uses machine learning to enhance network security, providing a robust defense against cyber threats by identifying and responding to anomalous behavior in real-time.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 86-91

3. METHODOLOGY

Machine learning-based network anomaly detection is a structured approach that consists of problem definition, collection of relevant network data, data preprocessing, labelling2 of the data, setting up of training, validation, and test datasets, selection of appropriate anomaly detecting models, training of the model Project scope covers identification of the types of network anomalies to be detected and establish particular targets and objectives. In order to collect data, one will require network information such as flow records, logs, packets, and other telemetry information. Handling missing data, normalisation of numerical characteristics, encoding categorical ones, feature extraction/engineering. This includes labelling instances as normal or anomaly and splitting the dataset into training, validation, and test sets.

The process of feature selection may involve feature importance analysis or dimensionality reduction to choose the most important features for a preferred machine learning model. This means that modelling involves training of the selected models with the labelled training data coupled with performance evaluation using such metrics as precisions, recall, F-1 scores, ROC AUC, or accuracy. There are a number of ways in which false positives can be mitigated ranging from setting a threshold for an anomaly scores, through to post processing techniques such as combing multiple models. This research paper applies DOS attack tools to send improper datagrams on the experimental network that are recorded as TCP Dump data file in raw format by Wireshark. Through the use of Wireshark commands, the raw TCP dump data file is transformed into a common network flow format which contains individual network packet snapshots sent at various intervals. Wireshark is a live monitoring tool that records data flow details per time in a packet like source IP address, source port, destination IP address, destination port and packets incoming, outgoing and total. At this stage, the algorithm of anomaly using the residual ratio of information entropy as a measurement value is implemented at the simulation test experiments where it determines whether the abnormal action occurs or not when the residual ratio of information entropy is below the acceptable threshold.

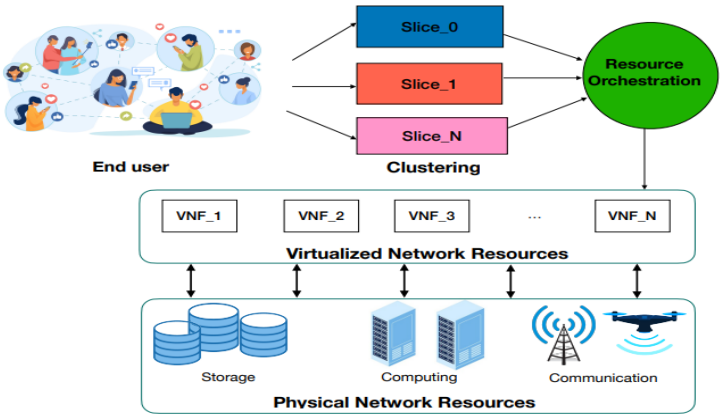

Figure 2 Technological chart

Machine Learning Model: LSTM and RNN model has been selected in this study as the deep learning model. LSTMs have a particular advantage when it comes to sequential data processes, which makes them ideal for solving problems relating to time-series analysis such as network traffic flows.

Certainly, let's delve into the technical steps for implementing a network traffic anomaly detection project:

1. Data Collection:

- Use packet capture tools like Wireshark, tcpdump, or Bro IDS to capture network traffic data. Store this data in a structured format, such as PCAP files.

2. Data Preprocessing:

- Convert the PCAP data into a format suitable for analysis, such as CSV or JSON.
- Remove duplicates and handle missing values if necessary.

3. Feature Extraction:

- Extract relevant features from the network data. Common features include source and destination IP addresses, source and destination ports, packet sizes, protocols, and timestamps.
- Calculate additional features like packet rates, connection durations, or byte counts.

4. Labeling:

- Annotate the data with labels indicating normal or anomalous traffic. You can use historical data or intrusion detection systems for labeling.

- Create a binary classification label (0 for normal, 1 for anomaly) for each data point.

5. Data Splitting:

- Split the dataset into training, validation, and testing sets. A typical split might be 70% training, 15% validation, and 15% testing.

6. Feature Engineering:

- Calculate the Mutual Information (MI) between features and the binary classification label to identify the most informative features for anomaly detection.
- Select the top N features with the highest MI scores.

7. Model Selection:

- Choose a machine learning or deep learning model. Random Forests, Support Vector Machines (SVMs), or autoencoders are common choices for anomaly detection.
- Implement the selected model using libraries like scikit-learn or TensorFlow/Keras.

8. Model Training:

- Train the model using the training dataset. Fine-tune hyperparameters using the validation dataset.
- Experiment with different algorithms and model architectures to find the best-performing one.

9. Model Evaluation:

- Evaluate the model's performance on the testing dataset using metrics like precision, recall, F1-score, ROC curves, and confusion matrices.
- Adjust the model threshold to balance false positives and false negatives based on your network security requirements.

10. Deployment:

- Deploy the trained model in your network infrastructure. This may involve integrating it with network monitoring tools or using it for real-time analysis.
- Implement batch processing for offline analysis of historical data.

11. Monitoring and Maintenance:

- Continuously monitor the model's performance in a production environment. Set up alerts for unusual model behavior.
- Regularly update the model with new data to adapt to changing network patterns and threats.

12. Documentation:

- Document all technical aspects of your project, including data sources, preprocessing steps, feature engineering, model details, and deployment procedures.

13. Security Measures:

- Implement security measures to protect your model and data. Ensure access controls and encryption are in place.

14. Scaling:

- Consider scalability if your network traffic data grows. Implement distributed computing solutions if necessary.

15. Reporting and Visualization:

- Create dashboards and reports to provide insights into network traffic and detected anomalies for network administrators and security analysts.

16. Feedback Loop:

- Establish a feedback loop with cybersecurity experts to continually improve the model's performance based on real-world threats and incidents.

Python Libraries Used:

The project leverages several Python libraries for various tasks:

Pandas (v1.3.4): Used for data manipulation and analysis.

NumPy (v1.21.4): Utilized for numerical computing.

TensorFlow (v2.6.0): Chosen as the machine learning framework for implementing deep learning models.

Matplotlib (v3.4.3): Used for creating plots and visualizations.

Scikit-learn (v1.0): Employed for machine learning tasks.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 86-91

e-ISSN : 2583-1062

Impact Factor: 5.725

4. EXISTING SYSTEM

Distributed Denial of Service (DDOS) attack is one of the attacks that lead to abnormal behaviour of network traffic. Large-scale DDOS results in bandwidth exhaustion of target network and resource exhaustion of target system. DDOS uses a large number of widely distributed controlled hosts to send a large number of short and huge data packets or numerous invalid connections to the target computer at the same time, thus generating a large number of network communications in a relatively short time, making millions of data messages flow into the target computer and the target network. This method not only causes the utilization rate of system resources (CPU and memory) of the attacked host to be as high as 100%, resulting in slow or no response to normal user requests, but also makes the network traffic increase sharply, and the bandwidth resources are occupied by 100% of the invalid traffic, resulting in the network unable to communicate normally. Ultimately it leads to the paralysis of the target system and network. Therefore, from the analysis of representation characteristics of DDOS attack, DDOS attack has the dual characteristics of resource exhaustion of target system and bandwidth exhaustion of target network [4]. In order to improve the accuracy of anomaly detection of network traffic, it is usually feasible to judge whether the system is attacked by DDOS by using the dual indicators of the actual utilization rate of network bandwidth and the utilization rate of system resources.

5. FLOWCHART

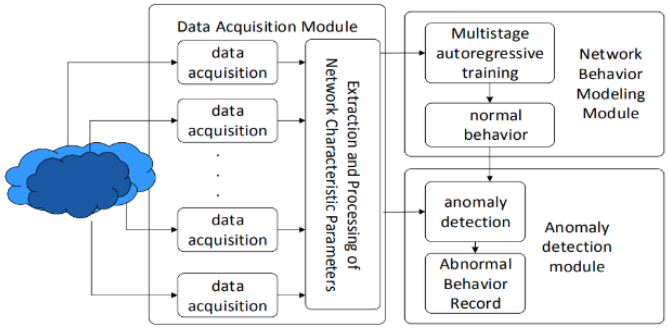

Figure 3 Flowchart

6. SAMPLE GRAPH

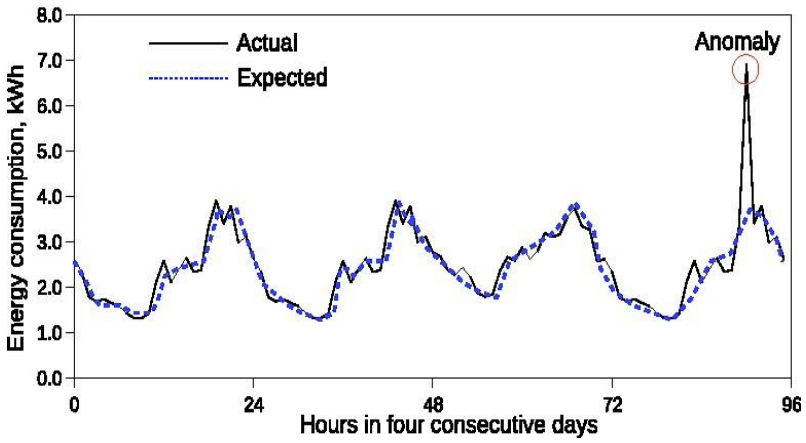

Figure 4 Sample reference graph

7. FUTURE WORK

Network Security comprises a significant area of research in detecting real time anomalies in the network traffic. The distribution of various characteristic parameters defines the nature of network behavior, and these natures vary significantly for normal and anomalous network activities. The distribution characteristics of information entropy can characterize the network behavior, and turn the anomaly detection of network traffic into an entropy-based pattern recognition problem. This study develops an information-theoretical based multilevers autoregressive anomaly detection model of network traffic. Initially, the average information entropy is eliminated before the anomaly traffic detection model that uses third-order linear auto regression is built. It also suggests the residual ratio of information entropy in order to measure the detections' anomaly. Finally, network traffic tests model of multilevel autoregressive anomaly detection in simulation network environment. A multilevel model of multilevel autoregressive anomaly detection will be able to discover more than 95 percent of traffic attack.

8. CONCLUSION

Machine Learning in detecting network anomalies is gaining popularity towards increased security, better reliance, and improved efficiency of networks. The detection of anomalous patterns in network traffic as well as infrastructure is done through this technology. It helps to respond preventively to the possible threats that are identified. These advantages contain improved security, real time detection, lesser downtime, flexibility, effectiveness and lower costs, decrease of false positives, compliance with laws, continuous studying and improvements.

ML-based anomaly detection system can recognize unknown threats and intrusions, allowing timely remediation measures during the initial stages of a threat. This also helps improve availability and reduces downtime, as well as minimize the impact that network failures may have on continuous business operations. This is why machine learning models are good at identifying dynamic attacks because they can adjust to new conditions and learn from fresh information.

Deep learning-based models, explainable AI, real time decision support, and adaption of future network technologies such as 5G or Internet of Things should be explored as future research directions in network anomaly detection using machine learning. With the advancement in technology and the threat landscape, cybersecurity will be very important in protecting digital assets and network reliability.

9. REFERENCES

[1] Zuyun Fu. Information Theory - Basic Theory and Application [M]. Beijing: Publishing House of Electronics Industry.

[2] Mengli Dong, Geng Yang. Network Traffic Forecasting Method [J]. Computer Engineering, 37 (16): 98-100.

[3] Wei Liu. Traffic Anomaly Detection Based on Information Entropy [D]. Nanjing: Southeast University.

[4] Xiangkun Mu, Jinsong Wang, et al. Network Anomaly Traffic Detection Method Based on Active Entropy [J]. Journal on Communications, 34 (Z2): 51-57 minutes

[5] Jiuqiang Xu, Yangyang Zhou, et al. Network Traffic Anomaly Detection Based on Flow Time Domain [J]. Journal of Northeast University, 40 (1): 27-31. [6] Yanyun Cheng, Shouchao Zhang. Time Series Outlier Detection Based on Big Data [J]. Computer Technology and Development, 26 (5): 139-144.

[6] Xuesong Qiu, Xun Zhang, et al. Two-stage Flow Anomaly Detection Method Based on Entropy and Linear Relationship [J]. Journal of Beijing University of Posts and Telecommunications, 41 (4): 56-62.

[7] Xiaoshi Fan, Chenghai Li. Application of Weighted Conditional Entropy in Anomaly Detection [J]. Computer Applied Research, 31 (1): 203-205.

[8] Xixin Cui, Wei Su, et al. Research and Implementation of Flow Analysis and Anomaly Detection Technology Based on Entropy [J]. Computer Technology and Development, 23 (5): 121-124.

[9] Rahmani H, Sahli N. Distributed denial-of-service attack detection scheme-based joint-entropy[J]. Security and Networks, 5(9): 1049-1061.