Ahsanullah Univesity of Science and Technology

CSE 4262

Data Analytics Sessional

Project Proposal

Title : Identifying Network Anomalies: Supervised & Unsupervised Learning

Name	ID
Mahmudul Haque	**20200204006**
Shadman Chowdhury	**20200204019**
MD Iftekhar Hossain	**20200204047**

Introduction:

The rapid expansion of digital technologies and interconnected systems has brought unprecedented convenience and efficiency to various sectors of society. However, this increased connectivity has also introduced significant risks, with cyber-attacks posing a major threat to the stability and security of networked environments. Traditional intrusion detection systems (IDS) often rely on static and signature-based methods, which are limited in their ability to detect novel and sophisticated attacks. As cyber threats evolve in complexity and frequency, there is an urgent need for advanced solutions that can proactively identify anomalies and mitigate potential risks.

Anomaly detection, a critical aspect of network security, focuses on identifying patterns in data that deviate from established norms. These deviations often indicate unauthorized activities, such as Denial of Service (DoS) attacks, data breaches, or other malicious behaviors. The challenge lies in effectively detecting both known and unknown attack types while maintaining a low false-positive rate.

This proposal aims to explore the integration of machine learning techniques, both supervised and unsupervised, for anomaly detection in network traffic. Leveraging the KDD Cup 1999 dataset as a benchmark, the study will evaluate the potential of clustering algorithms and classification models to identify anomalous behavior. The motivation for this work stems from the growing demand for scalable, efficient, and intelligent network security solutions capable of adapting to dynamic threat landscapes. By employing advanced data analytics and machine learning methodologies, the proposed approach seeks to enhance the robustness and responsiveness of intrusion detection systems, contributing to safer digital ecosystems.

Releted Works :

Numerous studies have explored various techniques for anomaly detection in network security, focusing on both supervised and unsupervised learning methods. Here are some related works we have reviews:

"A Detailed Analysis of the KDD CUP 99 Data Set" by Tavallaee et al. [1] analyzes the KDDCUP’99 dataset, widely used for evaluating anomaly-based intrusion detection systems (IDSs). The authors identify two major shortcomings: a high volume of redundant records, biasing learning algorithms, and an uneven attack distribution, rendering cross-validation challenging. They propose the NSL-KDD dataset, a refined version addressing these issues by removing redundant records and balancing attack representation**.** NSL-KDD allows consistent evaluation and comparison of IDSs, providing a more reliable benchmark for research. Experimental results demonstrate that diverse machine learning methods exhibit a broader performance range on NSL-KDD, highlighting its effectiveness in assessing learning techniques.

"Data Clustering for Anomaly Detection in Network Intrusion Detection" by Nieves and Jiao [2] focuses on unsupervised anomaly detection using data clustering. They highlight the advantages of data clustering for anomaly detection such as the ability to identify deviations from normal behavior and its potential to detect new and emerging threats. They use the k-means algorithm for

grouping unlabeled network data into clusters based on similarities, labeling smaller clusters as potential attacks. The authors argue that this approach can detect emerging threats without requiring labeled training data. Their evaluation using the KDD'99 dataset shows promising results with a high detection rate but acknowledges the potential for a high false alarm rate.

"A hybrid network intrusion detection framework based on random forests and weighted k-means" by Elbasiony et al. [3] introduces a hybrid intrusion detection system combining misuse and anomaly detection techniques. The authors employ the random forest algorithm for misuse detection to build intrusion patterns from labeled training data. For anomaly detection, they use a weighted k-means algorithm, incorporating feature importance derived from the random forest to cluster network connection data. They propose a method for identifying anomalous clusters by injecting known attacks into the dataset, enhancing the accuracy of anomaly detection. Their evaluation on the KDD'99 dataset demonstrates that the hybrid framework achieves a higher detection rate with a lower false positive rate compared to individual misuse and anomaly detection methods.

"Network Intrusion Visualization with NIVA, an Intrusion Detection Visual Analyzer with Haptic Integration" by Nyarko et al. [4] presents NIVA, a tool for visualizing network intrusion data with haptic feedback. NIVA utilizes different node placement algorithms, including IP-Space, spring technique, and the helix technique to effectively represent network intrusion data in 3D space. The system addresses the challenges of visualizing large networks by employing aggregation techniques to manage scale and complexity. Haptic integration adds a tactile dimension to the visualization, allowing users to perceive data properties like attack frequency and intensity through force feedback. An experimental evaluation shows that haptic feedback enhances the accuracy of attack detection and improves the understanding of relationships within a network.

"Detecting Denial of Service Attacks Using Emergent Self-Organizing Maps" by Mitrokotsa and Douligeris [5] explores anomaly-based detection of Denial of Service (DoS) attacks using Emergent Self-Organizing Maps (ESOMs). Recognizing the limitations of traditional Kohonen's SOMs in handling diverse network traffic, the authors leverage ESOMs to achieve a more reliable classification. They use the U-Matrix method to visualize network traffic patterns, distinguishing normal traffic from DoS attacks based on the location of data points in the U-Matrix. Evaluation using the KDD'99 dataset with carefully selected features, including duration, source bytes, and connection error rates, demonstrates the effectiveness of their approach. The results show high detection rates ranging from 98.3% to 99.81% with low false alarm rates between 0.1% and 2.9%, outperforming previous SOM-based methods.

Methodology :

In this study, we will use machine learning methods to find anomalies in network traffic. The methodology will include the following steps:

Dataset Collection and Preparation

1. **Dataset Overview:** We will use the KDD Cup 1999 dataset, which contains labeled instances of normal traffic and various types of attacks, such as Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R).
2. **Data Cleaning:** Rows with missing or incomplete data will be removed to ensure data consistency and quality.
3. **Encoding Categorical Features:** The dataset contains categorical columns, such as "protocol type," "service," and "flag." These will be converted into numerical values using label encoding to make them compatible with machine learning models.
4. **Feature Scaling:** Numerical features will be normalized using Min-Max scaling to bring all values into a range between 0 and 1. This step ensures fair treatment of features during model training, especially for distance-based methods like K-Means.
5. **Correlation Analysis:** Features will be analyzed for multicollinearity. Highly correlated features will be identified using a correlation matrix, and redundant features will be removed to reduce dimensionality and improve model performance.
6. **Dimensionality Reduction:** Techniques such as Principal Component Analysis (PCA) will be applied to reduce the number of features while retaining the most significant information. This will make the dataset easier to process and improve computational efficiency.
7. **Exploratory Data Analysis (EDA):** We will visualize the data distributions, identify patterns, and detect potential outliers to gain insights into the dataset. EDA will guide feature selection and preprocessing strategies.
8. **Train-Test Split:** The dataset will be divided into training and testing sets, typically in an 80:20 ratio. Stratified sampling will be applied to maintain the proportion of attack categories in both sets.

Supervised Learning

We will train models such as Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, and Stochastic Gradient Descent (SGD) using the labeled training data.

Model performance will be evaluated using metrics like accuracy, precision, recall, and F1-score to assess their ability to correctly classify anomalies.

Unsupervised Learning

Clustering methods like K-Means and K-Medoids will be applied to unlabeled data to identify patterns and group similar data points. Anomalies will be detected as outliers that do not fit well into any cluster.

We will use Silhouette Scores to evaluate the quality of the clustering results and adjust the number of clusters (k) for optimal performance.

Implementation and Optimization

Apache Spark’s MLlib library will be utilized to handle the computational complexity of processing large-scale data efficiently.

Hyperparameter tuning, such as the number of estimators in Random Forest or the value of k in K-Means, will be performed to improve model accuracy and clustering quality.

Evaluation and Validation

The models will be tested on unseen data from the testing set to validate their performance and robustness. A comparative analysis of supervised and unsupervised learning methods will be conducted to highlight their strengths and limitations in anomaly detection.

Conclusion :

In conclusion, this study highlights the potential of machine learning in advancing network security through effective anomaly detection. By utilizing the KDD Cup 1999 dataset, we applied both supervised and unsupervised learning techniques to identify unusual patterns in network traffic that may signal cyber-attacks. The integration of data preprocessing methods, such as feature scaling, dimensionality reduction, and correlation analysis, ensures the dataset's quality and enhances model performance.

Supervised learning models, such as Random Forest and K-Nearest Neighbors, demonstrated their ability to accurately classify anomalies with high precision, recall, and F1-scores, while unsupervised clustering techniques, like K-Means, proved effective in detecting previously unknown threats by identifying outliers. Leveraging tools like Apache Spark enabled efficient processing of large-scale data, and hyperparameter tuning further optimized the models for robust anomaly detection.

This approach not only underscores the importance of machine learning in handling large datasets but also demonstrates its adaptability to the dynamic nature of cyber threats. Future work could explore the integration of deep learning techniques and real-time data streams to enhance detection capabilities further. By addressing challenges like false alarms and evolving attack strategies, this study contributes to the development of advanced and scalable intrusion detection systems that strengthen overall network security.

Reference:

[1] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009, pp. 1–6. DOI: 10.1109/CISDA.2009.5356528.

[2] J. Nieves, “Data clustering for anomaly detection in network intrusion detection,” Jan. 2009.

[3] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted kmeans," Ain Shams Engineering Journal, vol. 4, no. 4, pp. 753–762, 2013, ISSN: 2090-4479. DOI: https://doi. org/10.1016/j.asej.2013.01.003. [Online]. Available: https : / / www. sciencedirect . com / science / article / pii / S2090447913000105.

[4] K. Nyarko, T. Capers, C. Scott, and K. Ladeji-Osias, "Network intrusion visualization with niva, an intrusion detection visual analyzer with haptic integration," in Proceedings 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. HAPTICS 2002, 2002, pp. 277–284. DOI: 10.1109/HAPTIC. 2002.998969.

[5] A. Mitrokotsa and C. Douligeris, "Detecting denial of service attacks using emergent self-organizing maps," in Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005., 2005, pp. 375–380. DOI: 10.1109/ISSPIT.2005. 1577126.