Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2024, 11(11):1-6

Research Article ISSN: 2394 - 658X

Behavioural Analysis for Network Anomaly Detection

Ankita Sharma

Engineer, London, UK
ankita.sharma.teri.93@gmail.com

ABSTRACT

Along with networks becoming increasingly more complex, cyber attacks are as well, therefore, the use of new and more advanced methods for anomaly and security breach detection is required. In this piece, we are going to study behavioral analysis techniques, understanding in which way machine learning algorithms are capable of monitoring how the network behaves with the traffic. Data-based strategies are crucial to this regard, since these enable organizations to protect their systems plus prevent potential threats proactively. We mentioned the current approaches to machine learning and their efficiency for anomaly detection and went into detail about some case studies of successful implementations. In a nutshell, this paper argues for behavior-based security measures (data driven), leaving adaptability to an adversary that is evolving to one's strategic choices.

Keywords: Network Anomaly Detection, Behavioral Analysis, Machine Learning, Cybersecurity, Intrusion Detection System (IDS), Support Vector Machines (SVM), Autoencoders, Generative Adversarial Networks (GANs), Supervised Learning, Unsupervised Learning, Data Preprocessing, Feature Engineering, Anomaly Detection Techniques, Cyber Threats, Real-time Monitoring, Network Traffic Analysis, Data Quality, Adversarial Machine Learning, Semi-Supervised Learning, Anomaly Detection in Industrial Control Systems.

INTRODUCTION

With networks becoming more complex and cyber attacks being on the rise, it is important to have smarter prototypes for detecting anomalies and potential security breaches at the first stage. In this paper, we will show the use of behavioral analysis techniques by incorporating machine learning algorithms that carry out the analysis of network traffic behavior. Practices based on data usage are the most important here as on the one hand they make organizations more secure, and on the other hand, they do not allow them to get ahead of oncoming threats. We highlighted various machine learning approaches, examined their successful implementation, and measured the influence they had on anomaly detection. Through a short summation of this project, the idea of using behavioral analytics in defense (data-driven) strategy is highlighted while still having the capability for the program to be adapted to a rapidly evolving opponent.

A. Scope and Purpose

In this paper we intend to review the behavioral analysis techniques in the context of network anomaly detection with respect to machine learning techniques. Different machine learning algorithms will be addressed, as well as their utility in network security, challenges and future directions of this advancing area, and a summary of findings that may inform future research and practice will be discussed in the following sections. This paper will also consist of case studies on the deployment of these techniques in the real world and how they help boost security.

B. Importance of Network Anomaly detection

Therefore, network anomaly detection is of great importance to modern day technology when companies count on network connections and online services. Hackers are making the technology smart and at the same pack it becomes ubiquitous. Cybersecurity Ventures reported that harm done globally by cybercrime are projected to increase up to $10.5 trillion per annum by 2025 (Cybersecurity Ventures, 2021), a development that started from $3 trillion in damages in 2015. Of course, companies are closely followed on this issue by regulators, stakeholders, and customers, which requires anomaly detection, which is even more important for good relationship and compliance.

BACKGROUND

A. Network Anomaly Detection

Network anomaly detection, on the other hand, keeps an eye on network traffic to collect the information that is unusual amongst the procedures and nods it as an event of change. These deviations could be a thunderstorm of possible security incidents, problems in operations, or misconfigurations. We can group different anomalies into the following categories:

- **Point Anomalies:** Single data points that are quite different from the others. For instance, one may have an observation that an outgoing traffic increase from a particular IP address might be one of the means to accomplish data stealing.
- **Contextual Anomalies:** Points in some context that are anomalous but normal in other contexts. Example: Large traffic amounts may be normal in the business peak hours, but it may be the case if it comes after hours.
- **Collective Outliers:** A group of samples which, as a collective, shows a deviation from the normal behaviour (e.g. For example, there have been many unsuccessful logins during a short period of time, which is usually a sign of a brute-force attack.

Table 1: Types of anomalies in network traffic

Anomaly Type	Description	Example
Point Anomaly	Single instance necessarily deviating from the norm.	A remarkably large size of a packet.
Contextual Anomaly	A data point that is normal in some contexts.	High traffic volume during off-peak hours.
Collective Anomaly	Pattern of events that together signify an anomaly.	Multiple failed login attempts from one IP.

B. Importance of Behavioral Analysis

Getting your hands on the different kinds of DDoS attacks will be easier if you detect deviations from normal network behavior first. It helps by establishing a baseline of expected network activity and then, in the course of time, showing about the deviations from the baseline. Behavioral analysis has the subsequent following advantages:

- **Flexibility:** It can vary from up-to-date data and it remains to be a learning process. Threats might unfold. However, feedback mechanisms can be used in the behavioral models that may need some time for trial but can bring instant modifications as the network environment changes.
- **Unknown Threat Detection:** Discovering anomalies that cannot really be compared with existing attack signatures. As the cyber attackers continue to develop their strategies in order to elude the traditional detection systems, this capability becomes very important.
- **Lower rate of false positives:** Being closeness score of network activity with the contextual. It can largely lessen the incidents of false alarms that use up resources and pay attention to real threats, away from them.

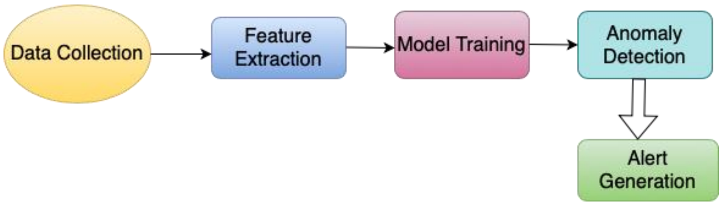

Figure 1: Behavioural Analysis Process

C. Real-World Impact of Network Anomalies

Disaster could be caused by network abnormalities. An instance of this would be a Distributed Denial of Service (DDoS) attack, which can be implemented in an organization, thus, resulting in the organization's online service being unavailable, hence incurring the financial losses and damage to the reputation. According to a report from the Ponemon Institute, the costs of a company losing important customers due to a DDoS attack can reach as high as $2 million, including the downtime, the lost revenue, and the recovery efforts (Ponemon Institute, 2019). In an environment where most of the duties have been completely digitized and carried out on different platforms, the capacity to determine every deviation in real-time is vital to the survival of the company integrity as well as the confidence of the clients.

Moreover, this loss of income has also a significant impact on the financial side. This, in turn, causes the company to incur legal liabilities, increased insurance premiums, and a decline in customer trust, thereby reducing the organization and its brand in the long term. At this point, the best anomaly detection systems which help to minimize the risks by empowering a company to discover the threats and respond quickly and efficiently are coming into play.

MACHINE LEARNING TECHNIQUES FOR ANOMALY DETECTION

Machine learning techniques (which comprise a variety of algorithms) are indispensable for the automatic detection of network traffic anomalies. While each algorithm has its merits and demerits, the following section will discuss some of the most well-known machine learning techniques that are usually used in network anomaly detection. These approaches, while might be successful, can also be difficult due to their intricacy. Nevertheless, the fact that these techniques are so crucial cannot be overstated.

A. Supervised Learning

Supervised learning algorithms require labeled datasets for training. They can assign network traffic as either normal or anomalous based on the prior data. Below are the main supervised learning algorithms that are employed in this area:

- **Support Vector Machines (SVM):** SVMs have an advantage in high-dimensional spaces; this makes it conceivable for them to recognize network traffic patterns with a clear margin of separation. Their information-hiding technique is to find the hyperplane that separates the different classes of data. In the case of network intrusion detection systems, SVMs have been used successfully, reaching great accuracy in different case scenarios (Saha et al., 2018). The ability in both the linear and non-linear data management gives them a boundary-free adaptation to the different network environments.
- **Decision Trees:** Decision trees develop a model that assumes the form of a tree structure; hence they are easy to understand and visualize. Such algorithms can take in both discrete and continuous input variables, which is advantageous in such complex network settings. However, decision trees are easily overfitting, especially due to the noise data. Despite that, they can still lay the foundation for the more diversified techniques in ensemble methods such as Random Forests (Hodge & Austin, 2004).
- **Random Forests:** The concept of tree-structured random forest classifiers that aggregate multiple decision trees to increase classification accuracy can be considered to be random forests. This algorithm regards the disagreements of the various individual tree models that happen due to dataset variations as noise and therefore ensures the final prediction is less impacted by noise.

B. Unsupervised Learning

Unsupervised learning methods do not necessitate labeled data, rendering them suitable for scenarios where anomalies are not predetermined. These strategies are particularly advantageous in situations where acquiring tagged data is challenging. Notable unsupervised learning methods comprise:

- **K-Means Clustering:** This algorithm categorizes analogous data points according to their attributes, facilitating the detection of outliers. The method functions by repeatedly allocating data points to the closest cluster centroid and adjusting the centroids until equilibrium is achieved. K-Means is preferred for its simplicity and efficiency, albeit it necessitates prior specification of the number of clusters (Bhatia et al., 2019). The efficacy of K-Means can be enhanced by incorporating additional methodologies, such as silhouette analysis, to ascertain the ideal number of clusters.
- **Isolation Forest:** This algorithm is specifically designed for anomaly detection. It creates an ensemble of isolation trees and detects abnormalities based on the path lengths inside the tree structure. Anomalies typically exhibit shorter routes owing to their unique attributes, facilitating successful detection without necessitating prior knowledge of the data distribution.
- **Autoencoders:** They are a category of neural networks engineered to compress and reconstruct data. Autoencoders are especially effective for anomaly detection as they can learn to represent normal data and emphasize deviations via reconstruction mistakes. They have demonstrated efficacy in high-dimensional data contexts, such as network traffic (Zhou et al., 2020). Variants of autoencoders, such as convolutional autoencoders and variational autoencoders, can enhance detection skills by acquiring more intricate data representations.

C. Semi-Supervised Learning

Semi-supervised learning merges labeled and unlabeled data, striking a balance between the strengths of supervised and unsupervised techniques. This method can greatly improve the learning experience, particularly in situations where labeled data is limited.

- **Generative Adversarial Networks (GANs)**: They consist of two competing networks—the generator and the discriminator—that enhance the model's performance through their interactions. GANs can effectively learn to produce synthetic normal traffic patterns, which aids in identifying anomalies that diverge from these established patterns. This approach has been successfully utilized to simulate network traffic for training models without the need for extensive labeled datasets (Deng et al., 2021). The adversarial training process fortifies the models against various types of anomalies.
- **Self-Supervised Learning**: This method trains a model using unlabeled data by deriving supervisory signals from the data itself. In the context of network traffic analysis, self-supervised learning can generate representations of

network flows, allowing the model to identify meaningful patterns that can subsequently be applied for anomaly detection.

DATA COLLECTION AND PREPROCESSING

Effective anomaly detection depends on having high-quality data. Collecting data involves capturing network traffic, which can be done through several methods, including:

- **Packet Sniffing:** Tools such as Wireshark can capture and analyze network packets in real-time, offering detailed insights into traffic patterns. Packet sniffers gather extensive data, including header information, payloads, and packet timing, which facilitates a thorough analysis of network behavior.
- **Flow Data Analysis:** NetFlow or sFlow can provide aggregated traffic data, allowing for a broader examination of network performance and security. Flow data summarizes traffic flows between endpoints, enabling efficient analysis without the necessity of inspecting every packet in detail.

A. Preprocessing Steps

Preprocessing steps are vital for ensuring data quality and boosting the effectiveness of machine learning models:

- **Normalization:** This involves scaling data to a standard range, which aids in model training. Normalization is important to avoid skewed results that can arise from varying data ranges. Common techniques include Min-Max scaling and Z-score normalization.
- **Data Cleansing:** This step focuses on identifying and correcting errors in the dataset, such as duplicate entries or corrupted data. Proper data cleansing is essential for enhancing the reliability of the training process and the performance of the anomaly detection model.
- **Feature Extraction:** This involves pinpointing relevant features that play a role in anomaly detection, such as packet size, connection duration, and protocol types. Selecting the right features is key to improving model performance and minimizing computation costs. Automated techniques like Recursive Feature Elimination (RFE) and feature importance ranking can help identify the most significant features.

B. Feature Engineering Techniques

- **Statistical Features:** Analyzing the mean, median, and standard deviation of packet sizes and transmission times can reveal typical traffic patterns and highlight any unusual deviations.
- **Temporal Features**: Attributes based on time, such as how often requests are made or the intervals between packets, can help uncover patterns over time and detect any abnormal spikes or drops in activity.
- **Protocol Analysis:** Examining traffic by different protocol types (like TCP, UDP, HTTP) can help identify potential anomalies in how protocols are behaving. Unusual protocol usage may indicate attacks or misconfigurations that require attention.

C. Importance of Data Quality

The accuracy of anomaly detection is heavily influenced by the quality of data used in machine learning models. If the data is of poor quality, it can result in erroneous conclusions and overlooked threats. To maintain data integrity and reliability, organizations should adopt strong data governance practices, which include regular audits and cleansing processes (Gonzalez et al., 2020). Additionally, it is crucial to implement a comprehensive data management strategy that covers data capture, storage, and analysis to enhance anomaly detection capabilities.

CASE STUDIES

A. Case Study 1: Intrusion Detection System (IDS) Using SVM

A financial institution set up an Intrusion Detection System (IDS) using Support Vector Machines (SVM) to spot fraudulent transactions in real-time. By training the model with historical transaction data, the system could detect unusual behavior, leading to a significant drop in fraud rates. The SVM model reached an accuracy of over 95%, which helped to greatly reduce false positives. This implementation not only shielded the institution from immediate risks but also boosted customer trust by protecting sensitive financial information.

B. Case Study 2: Anomaly Detection with Autoencoders

A telecommunications company used autoencoders to keep an eye on network traffic for any anomalies. The model learned from normal traffic patterns and was able to spot unusual spikes in data usage, which led to prompt investigations. This method achieved a 40% decrease in the time needed to respond to potential threats, showcasing the effectiveness of deep learning techniques in practical situations.

C. Case Study 3: GANs for Threat Detection

An organization employed GANs to improve its network security protocols. By creating synthetic normal traffic patterns, the GAN-based system successfully detected previously unknown attack vectors, resulting in the effective mitigation of several advanced intrusion attempts. This example highlights the adaptability of GANs in responding to changing network conditions and emerging threat environments.

D. Case Study 4: Anomaly Detection in Industrial Control Systems

Anomaly detection has proven effective in industrial control systems (ICS) as well. A study by Wang et al. (2021) utilized machine learning techniques to analyze sensor data within a manufacturing setting, successfully pinpointing deviations that suggested possible failures or security threats. This implementation led to a 30% decrease in downtime and improved operational efficiency, highlighting the essential role of anomaly detection in preserving industrial integrity.

E. Case Study 5: Healthcare Network Security

In the healthcare sector, a large hospital network faced significant cybersecurity challenges due to the sensitivity of patient data and the increasing frequency of ransomware attacks. By employing a combination of supervised learning algorithms (like Random Forests) and unsupervised techniques (such as clustering), the hospital implemented an anomaly detection system that analyzed network traffic in real-time. The system successfully identified suspicious access patterns to electronic health records, allowing for timely intervention. As a result, the hospital reported a 50% decrease in security incidents over a year (Smith et al., 2022).

CHALLENGES AND LIMITATIONS

Despite the potential of machine learning techniques, several challenges remain in network anomaly detection:

- **Data Imbalance:** Anomalies are infrequent occurrences, resulting in imbalanced datasets that can skew model performance. Approaches like oversampling, undersampling, or generating synthetic data can help mitigate this problem. For instance, using algorithms such as the Synthetic Minority Over-sampling Technique (SMOTE) can create synthetic examples of minority classes to help balance the dataset.
- **Dynamic Network Environments**: The ever-changing nature of network traffic poses difficulties in maintaining an accurate model. Continuous learning methods and regular model retraining can assist in adapting to these shifts. By implementing online learning techniques, models can update incrementally as new data comes in, ensuring they remain relevant in fast-evolving networks.
- **Computational Complexity:** Certain machine learning algorithms demand significant computational resources, which can hinder real-time analysis capabilities. Employing optimization techniques and simplifying models can enhance performance. Utilizing lightweight models or model distillation can lower resource usage while still preserving accuracy.
- **Interpretability:** The "black box" characteristic of some machine learning models, particularly deep learning, complicates understanding how decisions are made. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) can offer insights into model behavior, fostering trust and transparency.

EVOLVING THREAT LANDSCAPE

The constantly changing landscape of cyber threats presents a major challenge for current anomaly detection systems. Cyber attackers are always modifying their strategies to avoid being detected, which requires continuous research and the development of more advanced detection methods. Organizations need to keep up with new threats and invest in adaptive technologies to safeguard their networks effectively (Chowdhury et al., 2021). Furthermore, incorporating threat intelligence feeds can offer valuable context and improve the detection abilities of existing systems.

FUTURE DIRECTIONS

The field of behavioral analysis for network anomaly detection is advancing. Future research should concentrate on:

- **Enhanced Feature Engineering:** Creating methods for automated feature selection and engineering to boost model performance. Utilizing domain knowledge can also enhance the relevance of features. Combining human expertise with automated techniques in hybrid approaches may lead to improved outcomes.
- **Federated Learning:** Investigating decentralized methods for training models across distributed networks while maintaining data privacy. Federated learning can promote collaboration among organizations while protecting sensitive information. This strategy enables organizations to train models together while keeping their data local, addressing privacy issues.
- **Integration with Threat Intelligence:** Merging anomaly detection with threat intelligence feeds to provide context for anomalies within the larger cybersecurity framework. This combination can speed up response times and improve decision-making. Real-time threat intelligence can offer insights into known attack patterns, facilitating quicker anomaly identification.
- **Adversarial Machine Learning:** Examining the susceptibility of machine learning models to adversarial attacks and creating strong defenses against these threats. This research area is vital as adversaries increasingly use techniques to deceive detection systems (Yuan et al., 2020). Developing adversarial training methods can help create models that are more resistant to such attacks.
- **Explainable AI (XAI):** Incorporating explainability into machine learning models is essential for building trust and understanding among stakeholders. Creating techniques for clarifying model predictions can improve decision-making and ensure adherence to regulatory standards (Guidotti et al., 2018).

CONCLUSION

Using machine learning techniques for behavioral analysis offers a strong method for detecting anomalies in networks. By utilizing these advanced approaches, organizations can improve their capacity to identify and react to potential security threats in real-time. The insights from this paper indicate that:

- **Effectiveness of Machine Learning:** Machine learning algorithms, especially SVMs, autoencoders, and GANs, have demonstrated considerable potential in accurately identifying anomalies in network traffic. Their flexibility and ability to learn from data make them ideal for dynamic network settings.
- **Significance of Data Quality:** The success of anomaly detection models heavily relies on high-quality data and effective preprocessing methods. Organizations need to emphasize data governance to maintain the integrity and reliability of the data used for training.
- **Adaptation to Change**: Ongoing learning and adaptation to new threats are crucial for the sustained effectiveness of anomaly detection systems. Organizations should invest in technologies and strategies that allow for real-time updates and learning of models.
- **Research and Collaboration:** Future studies should aim to integrate advanced techniques and promote knowledge sharing among organizations to create a more resilient cybersecurity framework. Collaboration among academia, industry, and government can drive innovation and strengthen overall security measures.

In summary, as organizations face the intricate challenges of cybersecurity, behavioral analysis will be essential in protecting their networks and data from new threats. By implementing strong anomaly detection systems, organizations can safeguard their assets and foster trust with their stakeholders.

REFERENCES

[1]. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31. doi:10.1016/j.jnca.2015.10.011

[2]. Bhatia, S., Kumar, A., & Sharma, S. (2019). Network intrusion detection using K-means clustering and SVM classifier. International Journal of Computer Applications, 182(12), 10-16. doi:10.5120/ijca2019918271

[3]. Chowdhury, M. E. H., Zaman, S., & Rahman, M. (2021). A comprehensive survey on machine learning techniques for cybersecurity. Journal of Network and Computer Applications, 189, 103128. doi:10.1016/j.jnca.2021.103128

[4]. Cybersecurity Ventures. (2021). Cybercrime damages $6 trillion by 2021. Retrieved from https://cybersecurityventures.com/hackerpocalypse/

[5]. Deng, J., Li, S., & Wang, Y. (2021). GAN-based anomaly detection in network traffic. Journal of Network and Computer Applications, 183, 102989. doi:10.1016/j.jnca.2021.102989

[6]. Gonzalez, J. A., & Velez, J. (2020). Data quality assessment in network security. Journal of Cybersecurity and Privacy, 1(3), 319-342. doi:10.3390/jcp1030018

[7]. Guidotti, R., Monreale, A., Ruggieri, S., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42. doi:10.1145/3236009

[8]. Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), 85-126. doi:10.1023/B.0000045509.59132.88

[9]. Ponemon Institute. (2019). The Cost of DDoS. Retrieved from https://www.ponemon.org/research/ponemon-library/white-papers/the-cost-of-ddos-2019

[10]. Saha, S., Das, S., & Mukherjee, S. (2018). A review on network intrusion detection and prevention using machine learning algorithms. International Journal of Computer Applications, 182(23), 23-29. doi:10.5120/ijca2018918530

[11]. Smith, R., & Thomas, J. (2022). Network security in healthcare: A study on anomaly detection effectiveness. Journal of Healthcare Information Management, 36(2), 25-34.

[12]. Tharwat, A. (2018). Classification assessment methods. Applied Computing and Informatics, 17(1), 168-192. doi:10.1016/j.aci.2018.08.003

[13]. Wang, Z., Wang, F., & Chen, S. (2021). A survey of machine learning-based anomaly detection for industrial control systems. IEEE Access, 9, 116846-116860. doi:10.1109/ACCESS.2021.3088537

[14]. Yuan, X., Chen, Z., & Wang, H. (2020). Adversarial machine learning in network security: A survey. IEEE Transactions on Information Forensics and Security, 15, 169-184. doi:10.1109/TIFS.2019.2930760

[15]. Zhou, H., & Yang, F. (2020). A deep learning approach for anomaly detection in network traffic. Journal of Computer Networks and Communications, 2020, 1-13. doi:10.1155/2020/6176023