Detection of anomalous actions in the network based on machine learning

Kulia Vladyslav
[1]Kharkiv National University of Radio Electronics, 14 Nauky Ave, Kharkiv UA-61166, Ukraine, vladyslav.kulia@nure.ua

Petrenko Olha
[2]Kharkiv National University of Radio Electronics, 14 Nauky Ave, Kharkiv UA-61166, Ukraine, olha.petrenko@nure.ua

Abstract. *As cyber threats become more sophisticated, the need for effective anomaly detection systems has never been more critical. The usage of machine learning techniques to detect anomalous behavior in network traffic in comparison with traditional detection methods has been researched in this paper. Key challenges, benefits, and real-world applications of machine learning in this area were discussed.*

Keywords: *Anomaly Detection, Machine Learning, Network Security, Cybersecurity, Intrusion Detection Systems (IDS), Supervised Learning, Unsupervised Learning, Semi-Supervised Learning*

I. Introduction and Problem Statement

Detecting anomalous activities in networks has become a main concern of modern cyber security strategies. Traditional signature-based detection systems struggle to keep up with rapidly evolving threats. Machine learning (ML) provides an effective solution for this problem by detecting anomalies that deviate from normal network behavior patterns. The methods and techniques of using machine learning to detect anomalies in network security are researched in this paper.

II. Problem Solution and Results

Anomaly detection focuses on identifying activities that deviate from established baselines, often indicating malicious intent or the presence of network intrusions. Unlike rule-based systems that rely on predefined signatures, anomaly detection systems can detect previously unknown threats by learning what constitutes "normal" behavior. To perform the specified action, machine learning is used. The paper analyzes the methods of machine learning to detect anomalies.

Among them, supervised learning in intrusion detection systems (IDS) plays a crucial role, providing powerful tools for analyzing and classifying network data [1]. In supervised learning, ML algorithms use a pre-trained data set consisting of input data (such as network packets) and corresponding labels that determine whether the data is normal or contains traces of an intrusion. This approach requires careful preparation and creation of a database where every element must be correct. Such annotation may include labels that identify specific types of attacks or assert secure behavior. The use of well-prepared data allows the algorithm to accurately "learn" from cases of malicious and safe behavior, which subsequently improves the accuracy of classification in real conditions. Supervised learning algorithms can include a variety of techniques, such as logistic regression, neural networks, and decision trees [2]. Each of these methods has its own characteristics and advantages in different intrusion detection scenarios. Algorithms such as Random Forest, SVM and neural networks are commonly used in this context.

One of the key advantages of supervised learning is its ability to effectively detect known types of attacks for which pre-trained data exists. This makes this approach particularly valuable in combating common and well-documented threats. However, one of the limitations is the dependence on the quality and coverage of the data set used for training, which can limit the algorithm's ability to detect new, previously unknown types of attacks.

Unsupervised learning of successors is a type of machine learning. Unsupervised machine learning is used with unlabeled data. The model will improve itself as it discovers patterns and information from the data set it is fed. Usually, the algorithm groups different data into categories that have similarities or differences. Unsupervised learning is useful for big data analysis. As shown in fig. 1, unsupervised learning can be divided into three types of problems: clustering, association, and dimensionality reduction [3].

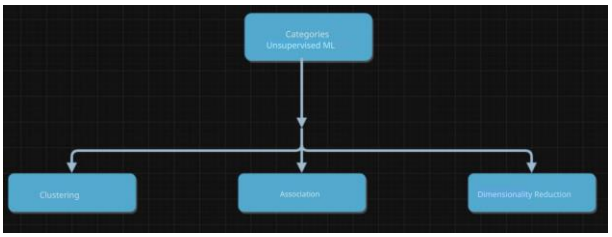

Figure 1 Unsupervised learning

These techniques are widely used in various fields, including IDS, where they are used to identify hidden patterns and relationships in data. Clustering consists in grouping a set of objects in such a way that objects in one cluster (group) are similar to each other, and objects from different clusters are different.

In the context of IDS, clustering can be used to detect anomalous patterns of network behavior by grouping together similar types of network traffic and highlighting unusual activity that may indicate malicious activity [4]. Association methods focus on discovering rules or relationships between different elements in data sets. This can be particularly useful for detecting complex threats and attacks that have multiple phases or use different attack vectors. For example, identifying the relationship between certain types of network requests and a sequence of malicious actions. Dimensionality reduction is another key element of unsupervised machine learning, which allows us to simplify data by reducing the number of random variables that describe it. This can be done using methods like principal component analysis (PCA) or t-SNE. In the context of IDS, it allows to highlight the main factors or attributes that characterize normal or abnormal network traffic, facilitating the process of analysis and detection of intrusions [5].

Overall, unsupervised machine learning provides important tools for analyzing and understanding complex network data. It

allows you to discover hidden relationships and anomalies that may not be obvious with a traditional approach, and is indispensable in detecting and responding to sophisticated and high-tech cyber-attacks.

However, due to the large amount of data required for training, it requires a lot of computing power and a lot of time. In addition, unsupervised learning has a higher risk of producing inaccurate results compared to supervised learning. Thus, at the end of training, human intervention is often required to verify that the output variables are correct. This check also takes a lot of time.

Semi-supervised and reinforcement learning is one of the key approaches in the field of machine learning, which finds application in many areas, including cybersecurity and IDS. This type of learning differs from supervised and unsupervised learning in that it focuses on the agent's interaction with a dynamic environment where it tries to maximize some measure of reward through its actions.

In the context of IDS, reinforcement learning can be used to design systems that adapt to changing attack patterns and the hostile environment. For example, a reinforcement algorithm can learn to determine the most effective strategies for detecting new types of intrusions or even adapt to changes in the network infrastructure. The learning process involves defining strategies (policies) that tell the agent which actions to choose in certain states of the environment in order to maximize the expected reward. In the context of IDS, this "reward" may be related to successful intrusion detection or effective attack prevention. One important aspect of reinforcement learning is its ability to learn in multistage environments where the effects of certain actions may not be immediately obvious. This allows the system to develop more complex strategies that take into account both the immediate and long-term consequences of decisions. However, reinforcement learning also has its challenges, especially in the context of IDS. First, it requires a well-defined reward mechanism, which is sometimes difficult to develop in the dynamic cyber threat environment. Second, reinforcement models may require a significant amount of interaction with the environment for effective learning, which may be difficult to provide in real-world settings [6]. Due to these features, reinforcement learning opens up new perspectives for the development of IDSs capable of adapting and evolving in response to changing cyber threat scenarios, although it requires a careful approach to development and validation.

The main problems in the application of machine learning are associated with false positives.

False positives are one of the most important problems in anomaly detection systems. The main task is to minimize false positives. Machine learning models, especially in unsupervised learning, can flag benign behavior as malicious, overwhelming the security team with alerts. Attacks can also be a problem, where the machine learning models themselves can be attacked when attackers create inputs that fool the system. Research on robust ML models is critical to mitigating such risks.

III. CONCLUSIONS

Anomaly detection systems based on machine learning provide a reliable solution to today's cyber security challenges. By analyzing large amounts of network data and detecting anomalous patterns of behavior, these systems improve the detection of both known and unknown threats. However, continued advances in machine learning techniques and solutions to current challenges such as data quality and adversarial resilience will determine the future effectiveness of these systems in the field of cybersecurity**.**

REFERENCES

[1] ThreatStack. The History of Intrusion Detection Systems (IDS)—Part 1.[Electronic resource]. – Access mode: https://www.threatstack.com/blog/the-history-of-intrusion-detectionsystems-ids-part-1,12/18/2023.

[2] IBM Cloud Education. Supervised Learning. [Electronic resource]. – Access mode: www: https://www.ibm.com/cloud/learn/supervisedlearning, 12/18/2023.

[3] Seldon Machine Learning Regression Explained. [Electronic resource].– Access mode: https://www.seldon.io/machine-learning-regressionexplained, 12/18/2023.

[4] Terence S. All Machine Learning Models Explained in 6 Minutes. [Electronic resource]. – Access mode: https://www.ibm.com/cloud/learn/unsupervised-learning, 12/18/2023.

[5] Thapa S., Mailewa A. The role of intrusion detection/prevention systems in modern computer networks. In Conference: Midwest Instruction and Computing Symposium (MICS). 2020. Vol. 53.pp. 1-14. URL: https://www.researchgate.net/publication/340581541_THE_ROLE_OF_INTRUSION_DETECTIONPREVENTION_SYSTEMS_IN_MODERN_COMPUTER_NETWORKS_A_REVIEW.

[6] Adeleke O. (2020). Intrusion detection: issues, problems and solutions. In 3rd International Conference on Information and Computer Technologies (ICICT). IEEE. 2020 pp. 397-402. URL: https://www.researchgate.net/publication/341400652_Intrusion_Detection_Issues_Problems_and_Solutions.