Journal of Industrial Engineering and Applied Science
ISSN 3005-6071 (Print) | ISSN 3005-608X (Online) | Vol. 2, No. 1, 2024

The Application of Machine Learning in Network Security Research

LIU, Ting [1*] JU, Hang [2]

[1] Guangzhou Institute of Technology, China
[2] Universiti Utara Malaysia, Malaysia

**** LIU, Ting is the corresponding author, E-mail: 429423693@qq.com***

Abstract: In the era of rapid development of the Internet, cybersecurity issues have become more important. Traditional network security defense technologies and methods can only meet some needs, so it is necessary to keep up with the times and use emerging technologies and methods to cope with constantly changing network threats. Machine learning, as an emerging technology in the context of artificial intelligence, is widely applied in the field of cybersecurity. This article will explore the current application status and future development trends of machine learning in cybersecurity research, and analyze its role and challenges in network intrusion detection, malicious code recognition, threat intelligence analysis, and other areas.

Keywords: Machine Learning, Cyberspace security, Network Intrusion Detection, Malicious Code Identification, Threat Intelligence Analysis

DOI: https://doi.org/10.5281/zenodo.10521643

1 Overview of Machine Learning Technologies

Machine learning, as the core field of artificial intelligence, involves the intersection of multiple disciplines. It improves the performance of algorithms and models through computer simulation or implementation of human learning processes, through data and experience. The goal of machine learning is to guide computers in acquiring knowledge and continuously improving their skills through data. Through continuous learning and training, machine learning systems can automatically adapt and improve their performance.

Learning algorithms can generate predictive models that help us classify, regress, and make decisions about data. With the increasing amount of data, the accuracy of machine learning applications will also become higher. Nowadays, machine learning technology has a wide range of applications, such as home life, shopping, entertainment media, and healthcare.

In the field of network security, old-fashioned threat detection systems detect threats and anomalies by using heuristic and static signatures. For example, antivirus software will generate and maintain a virus signature library (continuously updated) based on the characteristics of the virus program, and identify the virus program by comparing it with the signatures in the library during killing. Although signature based threat detection technology is easy to understand, it is not robust. One of its biggest problems is how to ensure that the signature comparison process matches the data inflow speed when the data size and flow rate significantly increase. Each package needs to be compared with each signature in the database. If synchronization cannot be maintained, it means that only a portion of the data packets can be discarded, resulting in a "missed fish".

Nowadays, signature based systems are gradually being replaced by intelligent cybersecurity agents. Machine learning has made positive progress in identifying new types of malware, zero day attacks, and advanced persistent threats (APT). Machine learning can identify attacks in the early stages and prevent them from spreading to the entire system. Many cybersecurity companies use advanced analysis methods such as user behavior analysis and predictive analysis to identify APT attacks in the early stages of the threat lifecycle, which can effectively prevent identity data leakage and internal threats. Prescriptive analysis is more responsive and can analyze what response measures should be taken to minimize losses after a network attack occurs [1].

2 The application process of machine learning in network security research

The application process of machine learning in cyberspace security research mainly includes the following steps:

(1) Security abstraction: Firstly, data preprocessing is a crucial step. This includes cleaning, denoising, and standardizing the original data to ensure the quality and consistency of the data. In addition, it is necessary to handle imbalanced data distribution to make model training more accurate and stable. Secondly, feature engineering is the process of extracting and selecting appropriate features. This requires extracting meaningful and representative features of security issues from the raw data. Feature engineering may involve techniques such as text processing, image analysis, and time series analysis to obtain informative feature representations. Finally, the problem definition is to clearly define the goals and measurement indicators of security issues. This helps to identify suitable machine learning algorithms and evaluation methods for problem-solving. The problem definition needs to consider recognition and detection performance, such as accuracy, recall, and error rate, to ensure the usability and reliability of the algorithm in practical applications.

① Classification problem: In the field of network security, classification problem is commonly used to identify and distinguish between normal behavior and malicious behavior. For example, by analyzing network traffic or user behavior data, it can be classified as normal traffic and potential malicious behavior. Machine learning algorithms can automatically learn and extract key features from data through training and learning, and classify based on these features. This classification model can help network security teams detect and respond to potential security threats early, and improve the security of network systems. By performing machine learning on network traffic or user behavior data, the model can learn to distinguish between normal traffic and potential malicious traffic. Meanwhile, the detection of malware can also be achieved through machine learning models. Machine learning can analyze large amounts of data, discover hidden patterns, and continuously optimize and adapt to new threats. Such applications can improve network security, detect and prevent potential attack behaviors early.

② Clustering problem: By clustering data, different attack sources or malicious behavior patterns can be identified. Clustering algorithms can gather similar data points together to reveal potential attack patterns or patterns of malicious behavior. This analysis can help security teams better understand and combat network threats, and improve the security of the system.

③ Dimensionality reduction problem: When dealing with high-dimensional data, dimensionality reduction techniques can help reduce the complexity of the data while retaining key information. This is particularly important in network security, as a large number of features (such as various metadata in network traffic) are usually collected, but most of them may be redundant or irrelevant. By reducing dimensionality, key information can be extracted more effectively, thereby improving the performance and efficiency of the model.

The abstraction of security issues is a crucial first step in using machine learning, which determines the direction of subsequent data collection, feature selection, and model construction. By accurately defining the type of problem, the effectiveness and accuracy of machine learning models can be ensured, thereby improving the security of the cyberspace.

(2) Data collection: In the data collection stage, it is necessary to collect data related to security issues from multiple levels such as the system layer, network layer, and application layer. These data will serve as inputs for subsequent machine learning models to analyze and detect potential security threats.

① System level data collection: This level of data mainly focuses on the information of the underlying hardware and software. For example, chip information, device information, system log information, and real-time operating status information are crucial for studying system security issues. By analyzing this data, we can understand the security of the system, detect hardware and software vulnerabilities, and evaluate the security under real-time operating conditions.

② Network layer data collection: Network traffic and network logs are the main sources of data at this level. By collecting and analyzing network traffic data, we can understand the flow of data packets in the network and detect potential network attack behaviors, such as malware propagation and denial of service attacks. Meanwhile, network log data can also provide valuable information about network activities, such as user behavior, access patterns, etc.

③ Application layer data collection: This level of data mainly focuses on the operation of specific applications or services. For example, user behavior data, input and output of the application, and other information related to application security. These data can help analyze the security of applications, detect potential vulnerabilities and malicious behavior.

By collecting relevant data from different levels, comprehensive and accurate information can be obtained, providing a strong data foundation for subsequent security analysis. When collecting data, attention should be paid to the completeness, accuracy, and timeliness of the data. In addition, it is necessary to choose appropriate collection methods and techniques based on specific security issues to ensure that the collected data can provide strong support for subsequent machine learning models.

(3) Data preprocessing and security feature extraction: The first step is to preprocess the collected data, mainly identifying and processing missing values, outliers, and data unrelated to security issues. For unstructured or semi-structured data, it is necessary to convert it into a structured data lattice to make it suitable for machine learning algorithms. In order to make all features have the same scale, the data also needs to be standardized or

normalized.

Secondly, extract features related to security issues based on security features. Security feature extraction includes static features, dynamic features, contextual features, deep learning features, etc. It is particularly noteworthy that in recent years, deep learning methods such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have also been used to automatically extract complex and non-linear features. When extracting features, attention should be paid to avoiding overfitting and bias in feature selection. The use of appropriate feature selection methods and evaluation metrics (such as mutual information, variance thresholds, etc.) can help determine which features are most predictive of security issues.

(4) Model construction: Choosing the appropriate machine learning algorithm is a key step in building a model. In the field of network security, there are many common machine learning algorithms, such as decision trees, support vector machines, random forests, K-nearest neighbors, neural networks, etc. These algorithms have different characteristics and applicable scenarios. According to the specific requirements of the task, we need to carefully evaluate the advantages and limitations of each algorithm, and choose the most suitable algorithm to build the model. Choosing the appropriate feature set is also an important step in the model construction process. The purpose of feature selection is to extract the most representative and relevant features from the raw data. This can be achieved through techniques such as statistical methods, information gain, and correlation analysis. A suitable feature set can improve the performance and generalization ability of the model. During the model construction process, we also need to tune and optimize the model parameters. By adjusting the parameters of the model, we can achieve better model performance. Common optimization methods include grid search, random search, and Bayesian optimization.

(5) Model validation: Using known security data to validate the model, this step aims to ensure the accuracy and reliability of the model, while identifying potential overfitting or underfitting issues to ensure the accuracy and reliability of the model. Cross validation techniques, such as k-fold cross validation, are commonly used to evaluate the performance of models. Through multiple repeated experiments, the performance of the model can be comprehensively evaluated and more reliable results can be obtained.

(6) Model performance evaluation: Model performance evaluation and actual deployment are key steps in machine learning projects, which help to understand and validate the actual performance of the model and whether it is ready for production deployment. By collecting actual operational data, testing and comparing the model, one can understand performance indicators such as accuracy, recall, F1 score, etc. In addition, factors such as the real-time performance, interpretability, and robustness to unknown attacks of the model need to be considered.

The application of machine learning in cybersecurity research requires continuous iteration and optimization. By continuously updating data, optimizing features, adjusting models, and monitoring feedback, the performance of models can be continuously improved and the security of the cyberspace can be enhanced [2].

3 The Important Role of Machine Learning in Studying Cyberspace Security

The application of machine learning in intrusion detection systems is a typical scenario. IDS is used to monitor network traffic and system logs, detect and alert potential attack behaviors. Traditional IDS typically relies on rules and feature libraries to detect attacks, but is often powerless against unknown attacks. By using machine learning algorithms, IDS can automatically classify network traffic and achieve detection of unknown attacks. This is mainly due to the autonomous learning ability of machine learning algorithms. They can extract useful features from a large amount of data and classify or cluster based on these features. For example, classification algorithms based on support vector machines (SVM), neural networks, or decision trees can be used to detect malicious traffic patterns; Clustering algorithms such as K-means can be used to discover abnormal behavior or unknown threats.

Deep learning models, such as convolutional neural networks or recurrent neural networks, can extract higher-level feature representations from the original network traffic, thereby more accurately identifying abnormal behavior. These methods usually require a large amount of annotated data for training, but they have good application prospects in large-scale network environments. In addition to classification and anomaly detection of network traffic, machine learning can also be applied to user behavior analysis, malware detection, vulnerability assessment, and other aspects. By analyzing the behavior patterns of users in the network, abnormal login and malicious operations can be detected; By analyzing the code and behavioral characteristics of malicious software, new variants can be quickly identified and addressed; By analyzing the security logs and configuration information of the system, the security vulnerabilities and risks of the system can be evaluated.

In addition, machine learning can also help security teams deal with vulnerabilities caused by improper permissions. Computer devices can scan millions of folders on the network and look for warning signs, such as certain special permissions or special users. In addition, they can also search for the login credentials of those who have resigned. Once intrusion behavior is detected, the security team needs to minimize losses as quickly as possible and drive the attacker out of their network system. The urgent

task at this time is to understand the root cause of data leakage events, understand the degree of infection, and determine the scope of impact. With the help of machine learning, security teams can quickly establish a "knowledge graph" to describe the scope of attacks. They can accurately locate IP addresses, devices, and individual users, and are faster and more accurate than manual analysis. This enables the team to quickly remove all infected elements and achieve automated responses to security events, including isolating intruders in separate subnets, closing ports, isolating devices, or encrypting data, among others [4]. Overall, the application of machine learning in the field of network security has enormous potential and can significantly improve the level of network security protection.

4 The Application of Machine Learning in Network Security

4.1 Application of Machine Learning in Network Intrusion Detection

Machine learning plays an important role in network intrusion detection. Traditional intrusion detection systems mainly rely on rules and feature libraries, often unable to effectively detect unknown attacks. Machine learning technology, through autonomous learning, can extract useful features from a large amount of data and automatically identify network attacks and abnormal behavior.

Machine learning technology can be applied to multiple aspects of intrusion detection systems. Firstly, it can be used for anomaly detection. By analyzing network traffic or system logs, machine learning algorithms can identify abnormal behaviors that do not match normal behavior patterns, thereby detecting potential attacks. For example, classification algorithms based on support vector machines, neural networks, or decision trees can be used to detect malicious traffic patterns, while clustering algorithms such as K-means can be used to discover abnormal behavior or unknown threats.

Secondly, machine learning can be used to automatically generate intrusion detection rules. By learning and analyzing historical attack data, machine learning algorithms can automatically extract attack features and patterns, thereby generating effective intrusion detection rules. These rules can be used to filter known attack traffic or behavior, improving the system's defense capability [7].

In addition, machine learning can also be used for predicting and warning intrusion threats. By long-term monitoring and analysis of network traffic and system logs, machine learning models can predict future threat trends and attack patterns. This predictive ability can help security teams take proactive measures to respond to potential attacks and improve the overall security of the network.

4.2 Application of Machine Learning in Malicious Code Recognition

Malicious code is one of the important forms of network security threats, and its constant mutation and concealment pose challenges to traditional feature matching based malicious code detection methods. Machine learning technology can effectively identify unknown variants of malicious code through feature extraction and classification of malicious code samples, and continuously update recognition models to adapt to new threats. In practical applications, machine learning has achieved significant results in identifying malicious code, providing strong support for network security defense [5].

4.3 Application of Machine Learning in Threat Intelligence Analysis

Threat intelligence analysis is an important research direction in the field of cybersecurity, with the core task of mining useful threat information from massive network data. Machine learning technology can model and analyze threat data, discover patterns and trends hidden behind the data, and help security experts timely warn and respond to potential threat events. In the field of threat intelligence analysis, the application of machine learning not only improves analysis efficiency, but also provides more scientific basis for security decision-making [6].

5 Conclusion

Machine learning, as an emerging technology, has achieved significant results in the application of network security research. Its automation and intelligence characteristics make the detection and defense of network attacks more efficient. Through machine learning technology, unknown attack behaviors can be quickly identified and responded to, improving the overall security of the network.

However, there are still some challenges in the application of machine learning in the field of network security. Firstly, the quality and quantity of data have a significant impact on the accuracy and generalization ability of the model. How to utilize unsupervised and semi supervised learning to handle security issues without sufficient labeled data is a key issue. Secondly, achieving a balance between accuracy and real-time performance is also an important issue. When facing large-scale network traffic, it is necessary to ensure that the model can make quick judgments. In addition, as network attacks continue to evolve, continuously updating and optimizing models to address new threats is also an important research direction.

With the continuous development of machine learning technology, its application in the field of network security will be more extensive and in-depth. On the one hand, advanced machine learning methods such as deep learning

can be utilized to conduct more detailed analysis and modeling of network traffic and user behavior, further improving the detection and defense capabilities against unknown attacks. On the other hand, combining other technologies of artificial intelligence, such as natural language processing and computer vision, can expand the application scope of machine learning in the field of network security, such as monitoring and analyzing online public opinion, and detecting malicious information on social media. In addition, with the popularization of technologies such as edge computing and the Internet of Things, network security issues will become more complex and diverse. How to combine machine learning technology with these new technologies to achieve more intelligent security protection will be an important research direction in the future.

Acknowledgments

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author Contributions

Not applicable.

About the Authors

LIU, Ting

Affiliation: Department of Robotics and Artificial Intelligence, Guangzhou Institute of Technology

JU, Hang

Affiliation: Universiti Utara Malaysia

References

[1] Hu, S., Li, Y., Gao, Y., & Wu, L. (2023). Research on 5G Network Security under Artificial Intelligence and Machine Learning. Information Technology.

[2] Liu, H., Rong, X. , Li, C., & Liao, J. (2020). Research on Cloud Computing Network Security Guarantee Method Based on Security Domain. Information Security Research.

[3] Li, X., Wu, G., Yao, L., Zhang, W., & Zhang, B. (2021). Research Progress and Future Challenges in Machine Learning Security Attacks and Defense Mechanisms. Journal of Software.

[4] Lei, D. (2021). Discussion on the Application of Machine Learning in Network Security. Network Security Technology and Application.

[5] Zhu, Q. (2021). Research on Computer Network Security in Cloud Computing Environment. Information Recording Materials.

[6] Li, M., et al. (2020). Research on the Application of Machine Learning in Network Security. Journal of Communications, 41(12), 88-95.

[7] Zhang, H., et al. (2019). Research on Machine Learning-based Network Intrusion Detection Technology. Computer Applications, 39(6), 172-176.