Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241032

Detection of malicious encrypted communication based on feature engineering and machine learning

Zhiting OuYang

School of Information Science and Engineering, University of Jinan, Jinan, China

1849586560@qq.com

Abstract. With the advancement of network technology, malicious encrypted communication has become a covert network threat and has drawn significant attention in the field of cybersecurity. Network threats are increasingly severe, and traditional detection methods struggle to cope with the intricate changes in malicious encrypted communication. Therefore, in order to find effective approaches to detect malicious encrypted communication, this research focuses on exploring the detection methods of malicious encrypted communication based on feature engineering and machine learning. The crucial role of feature engineering and the application of machine learning methods in this context are extensively discussed. The research conclusions indicate that the proper design of feature engineering and method selection can improve the detection accuracy and efficiency. However, practical applications still face challenges such as data scarcity and limited computational resources. Therefore, future research directions are proposed, including further optimizing feature engineering methods, developing feature representations more suitable for detecting malicious encrypted communication, and exploring more efficient deep learning models. The significance of this research lies in providing theoretical guidance and practical application advice for professionals and researchers in the field of cybersecurity, contributing actively to building a safer and more stable network environment, and jointly safeguarding the security and stability of the digital world.

Keywords: Malicious Encrypted Communication Detection, Feature Engineering, Machine Learning, Network Security.

1. Introduction

In today's digital age, with the continuous increase in threats from malicious encrypted communication, the field of cybersecurity faces unprecedented challenges. Malicious encrypted communication refers to the network communication method used by malicious actors to conceal their malicious activities using encryption techniques. In this scenario, malicious actors convert communication content into ciphertext using encryption algorithms to prevent third parties or network defense systems from easily understanding or intercepting the communication. Examples of such activities include the spread of malware, data theft and leakage, phishing, and social engineering. This encrypted communication makes malicious activities more covert and difficult to detect, thereby increasing the complexity of cybersecurity threats. Traditional security measures are no longer sufficient to effectively counter the increasingly sophisticated malicious activities. The detection of malicious encrypted communication is a critical component of cybersecurity, and its accuracy and efficiency directly impact the security of user information and systems. This study explores the integration of feature engineering and machine learning, providing new insights and solutions to the field of cybersecurity. This research will offer important references and guidance for cybersecurity practitioners and decision-makers, thereby promoting further development of cybersecurity defense technologies.

This study aims to explore malicious encrypted communication detection methods based on feature engineering and machine learning. By carefully designing feature engineering and selecting appropriate machine learning models, the goal is to achieve rapid and accurate detection of malicious encrypted communication. Specifically, this paper will analyze the role of feature engineering in malicious encrypted communication detection, investigate the application of different machine learning methods to this problem, conduct method comparisons and selections. The objective is to provide practitioners and researchers in the field of network security with theoretical guidance and practical recommendations regarding the detection of malicious encrypted communication.

Through this research and experimentation, we aim to provide new ideas and approaches for the field of malicious encrypted communication detection, both in terms of research and practical implementation. Effective malicious encrypted communication detection technology will contribute to raising the level of cybersecurity and protecting individuals and organizations from the impact of malicious communication.

2. Literature Review

Traditional methods of detecting malicious encrypted communication primarily rely on rules and manually defined features, which often exhibit certain effectiveness in specific scenarios. For example, statistical features based on network traffic, such as traffic size, latency, and frequency, are utilized to detect abnormal encrypted communication. Kumar, for instance, accurately identifies malicious encrypted traffic by analyzing traffic interaction patterns, indicating that traditional methods require more analysis of the interaction patterns of malicious encrypted traffic [1]. However, traditional methods often perform inadequately when confronted with complex malicious attacks and encryption techniques, as they struggle to adapt to new attack patterns and diverse malicious behaviors.

Existing methods for detecting malicious encrypted communication encompass machine learning, deep learning, feature engineering, and traffic analysis, among others. Yang combines deep learning to detect novel malicious SSL traffic, reconstructing SSL records from captured packets and generating a series of unencrypted data from consecutive SSL records for classification [2]. Pastor designs, trains, and tests a set of deep learning models to detect cryptocurrency mining activities [3]. From the literature, it is evident that deep learning can automatically learn features and patterns from raw data without the need for manual rule definition or feature engineering. However, deep learning models often require a substantial amount of training data to achieve satisfactory performance when dealing with complex problems. Cabaj employs HTTP traffic features (HTTP message sequences and their corresponding sizes) to detect malicious communication between infected hosts and attacker C&C servers [4]. Traffic analysis methods capture fine-grained features of communication traffic, demonstrating real-time capabilities and timeliness. Zhang proposes a novel and cost-effective feature extraction method and an efficient deep neural network architecture for accurate and rapid malware detection [5]. Approaches like these, which utilize feature-based detection of malicious encrypted communication, can extract and enhance crucial information about malicious encrypted communication from raw data, assisting models in better understanding and distinguishing malicious behaviors. Furthermore, by selecting and transforming features to reduce data dimensions and minimize redundancy, training and predictive efficiency of the model are enhanced. However, when excessive features are selected, the model may overfit the training data, affecting performance on new data, thus necessitating the integration of machine learning.

3. Feature Engineering in Malicious Encrypted Communication Detection

Feature engineering is a crucial component in the process of detecting malicious encrypted communications. It involves extracting, transforming, and selecting the most informative features from raw data, enabling machine learning algorithms to better classify and detect patterns. The following outlines the specific roles of feature engineering in the detection of malicious encrypted communications.

In a dataset, there may be a large number of features, but not all of them are useful for detecting malicious encrypted communications. Feature selection is a method for excluding irrelevant features by assessing their importance and relevance. In this step, methods such as correlation coefficients and mutual information can be used to evaluate the correlation between features and malicious encrypted communications, thus identifying features that contribute significantly to the classification task.

During the feature extraction process, original network communication data is transformed into more representative and interpretable feature representations. Features related to malicious encrypted communications can be categorized into three classes:Data element statistical features: Packet size, arrival time sequence, and byte distribution.TLS features: Encrypted suite and TLS extensions offered by the client, client public key length, selected encryption suite by the server, certificate information (whether self-signed, number of entries in SAN X.509 extension, validity period, etc.).Contextual data features: Can be further divided into DNS data stream and HTTP data stream features. DNS features focus on domain name length in DNS responses, ratio of numeric to non-numeric characters in domain names, TTL value, number of IP addresses returned by DNS responses, domain ranking in Alexa website, etc. HTTP features focus on various fields of inbound and outbound HTTP (Set-Cookie, Location, Expires, Content-Type, Server, etc.) as well as HTTP response codes.

In the detection of malicious encrypted communications, data is often high-dimensional, which increases the complexity of model training and resource consumption. Therefore, feature dimensionality reduction is an important step. Techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) can map high-dimensional features to a lower-dimensional space while retaining the most critical information. Feature dimensionality reduction also helps reduce the risk of overfitting and improves model generalization.

Feature engineering holds significant importance and positive impact in the detection of malicious encrypted communications. Through appropriate feature selection, extraction, and dimensionality reduction, raw data can be transformed into more representative and interpretable features, aiding machine learning models in accurately learning patterns and behaviors of malicious encrypted communications. Moreover, effective feature engineering can reduce the complexity of model training, enhance detection performance, and provide feasible solutions for practical applications. However, the challenge of feature engineering lies in fully harnessing the latent information within the data while avoiding overdesign and overfitting issues. Therefore, careful selection of appropriate feature extraction methods and dimensionality reduction strategies is necessary to obtain efficient and robust models for detecting malicious encrypted communications.

4. Machine Learning in Malicious Encrypted Communication Detection

A Support Vector Machine (SVM) is a supervised learning algorithm used to solve classification and regression problems. Its core idea is to find an optimal hyperplane in feature space that separates different classes of data points as much as possible while maximizing the margin between the classification boundaries. Selecting appropriate features contributes to enhancing the classification performance of SVM. By analyzing the correlation between different features and malicious encrypted communications, the most informative features can be chosen, reducing dimensionality while improving model efficiency. Converting time-series data into statistical features helps extract global characteristics of the data. Dimensionality reduction techniques like Principal Component Analysis (PCA) can reduce the number of features while retaining crucial information. Malicious encrypted communications are often less common than normal communications, resulting in an imbalanced dataset. This may lead to good performance of SVM on the normal class but poor performance on the malicious class. Adequate methods need to be employed to address the imbalance issue, such as undersampling, oversampling, or cost-sensitive learning.

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines them to classify and regress data. During training, each decision tree randomly selects a subset of features, thereby reducing model variance and overfitting risk. The application of Random Forest in malicious encrypted communication detection possesses the following characteristics: The ensemble strategy of Random Forest effectively reduces model variance and enhances generalization, making it suitable for problems with limited samples and high noise in detecting malicious encrypted communications. Random Forest assigns an importance weight to each feature, aiding in evaluating their contribution to classification. This assists in feature selection and analysis, enabling the identification of crucial features for detecting malicious encrypted communications. Feature engineering plays a vital role in Random Forest: Feature selection can influence the subset of features used by each decision tree during training. Proper feature selection can decrease decision tree correlation, thus enhancing the performance of the ensemble model. Feature extraction and transformation help convert raw data into feature representations suitable for decision trees. For instance, transforming continuous communication frequency data into discrete frequency distribution features helps capture patterns of malicious encrypted communications. Random Forest can assess the importance of each feature, aiding in selecting influential features. In cases of high feature dimensionality, dimensionality reduction and selection contribute to reducing model complexity and improving performance.

Convolutional Neural Networks (CNNs) are a specialized form of deep learning model tailored for processing grid-structured data, such as images and sequences. In the context of detecting malicious encrypted communication, CNNs exhibit distinct characteristics: Utilizing convolutional layers, they capture local features within the data, such as sensitive packet patterns in communication data. Pooling layers aid in diminishing feature dimensions and extracting crucial data insights. A standard CNN configuration comprises multiple convolutional layers, pooling layers, and fully connected layers, thereby facilitating the hierarchical extraction of varied abstract features from the data. This hierarchical approach assists in discerning between malicious and normal communication patterns. When employing deep learning methodologies, the task of feature extraction and representation learning is typically delegated to the model itself. Through the training process, the model autonomously assimilates valuable features from the data, negating the need for manual feature engineering. Deep learning techniques enable end-to-end learning, directly deriving the ultimate results of malicious encrypted communication detection from raw data. This approach streamlines the intricacies of intermediate steps. However, it's important to underscore that deep learning methods often necessitate a substantial dataset to effectively train complex models. The challenge may arise from the potential difficulty in acquiring an adequate quantity of malicious encrypted communication data. In scenarios where data is limited, methods like data augmentation are employed to expand the dataset while striking a balance between model complexity and performance. Deep learning models encompass multiple hyperparameters, including the learning rate and the number of hidden units. Optimizing these hyperparameters through techniques like cross-validation is pivotal for enhancing model performance. By carefully designing the structure and parameters of deep learning models, the full potential of convolutional neural networks can be harnessed to more accurately capture the features and patterns of malicious encrypted communication. While deep learning methods offer substantial promise in the domain of malicious encrypted communication detection, they must also surmount challenges like data scarcity and model optimization.

In the context of detecting malicious encrypted communication, the choice of machine learning methods directly impacts detection performance. Contrasting the advantages and disadvantages of different methods aids in understanding their applicability and limitations in the detection of malicious encrypted communication: Support Vector Machine (SVM) is suitable for small-sample, high-dimensional data and can handle both linear and nonlinear problems. However, it may exhibit higher computational complexity when dealing with large-scale data and might not perform well with imbalanced class data. Random Forest is capable of handling high-dimensional data and addressing class imbalance issues, demonstrating good robustness and generalization capability. However, its performance may be suboptimal when handling continuous and sequential data. Deep learning methods have the ability to learn complex feature representations, making them suitable for large-scale data and intricate patterns. They can handle various types of data, such as images, text, and sequences. However, they require a substantial amount of data and computational resources, and optimizing hyperparameters can be more complex. They might also exhibit subpar performance in scenarios with limited data samples.

When selecting a suitable method, it is crucial to consider the advantages and disadvantages of different approaches and the impact of feature engineering. The following factors are essential for method selection: Sample quantity and class distribution: The number of samples and the distribution of classes influence method selection. For example, SVM can be chosen for small-sample data, while Random Forest is suitable for dealing with class-imbalanced data. Feature type (continuous, discrete) and dimension (high-dimensional, low-dimensional): The type and dimensionality of features determine the choice of feature engineering. Deep learning is well-suited for large-scale high-dimensional data. Computational resources: Deep learning methods require significant computational resources, while SVM and Random Forest are relatively faster. Choose a method based on the availability of computational resources. By comprehensively considering the pros and cons of different methods and the impact of feature engineering, one can select an appropriate approach to improve the detection performance and efficiency of malicious encrypted communication. In practical applications, combining multiple methods through ensemble learning can further enhance the accuracy and robustness of malicious encrypted communication detection.

5. Conclusion

Malicious encrypted communication, as a covert network threat, has been receiving increasing attention in the field of cybersecurity. This paper focuses on the problem of detecting malicious encrypted communication and conducts research and exploration based on feature engineering and machine learning methods. It provides a detailed analysis of the role of feature engineering, the application of different machine learning methods, and the comparison and selection of these methods. This study draws the following conclusions and conclusions.

Feature engineering plays a crucial role in the detection of malicious encrypted communication. Properly designing feature engineering allows the extraction of key features from malicious encrypted communication, enhancing the classification performance and generalization ability of the model. Feature engineering includes steps such as feature selection and filtering, feature extraction and transformation, and feature dimensionality reduction and selection. The choice of different methods directly impacts the performance of the model. By appropriately selecting and transforming features, the accuracy and efficiency of malicious encrypted communication detection can be effectively improved.

In addressing the problem of detecting malicious encrypted communication, this paper investigates the application of machine learning methods such as Support Vector Machine (SVM), Random Forest, and Deep Learning. These methods each have different advantages, disadvantages, and applicable scenarios, requiring comprehensive consideration of factors such as data size, feature type, and computational resources. By flexibly combining multiple methods and forming ensemble learning, the accuracy and robustness of malicious encrypted communication detection can be further improved.

When choosing appropriate methods, we must take into account the specificity and complexity of detecting malicious encrypted communication. Although deep learning methods have advantages in handling complex patterns and large-scale data, they may not perform well when data is insufficient or computational resources are limited. Therefore, the design and selection of feature engineering are also crucial. By properly processing and transforming the data, the strengths of different machine learning methods can be fully utilized, thereby improving the detection performance.

When selecting appropriate methods, we must consider the uniqueness and complexity of detecting malicious encrypted communication. Although deep learning methods have advantages in handling complex patterns and large-scale data, they may not perform well when data is scarce or computational resources are limited. Therefore, the design and selection of feature engineering are also crucial. By properly processing and transforming the data, we can fully leverage the strengths of different machine learning methods and improve the detection performance.

In conclusion, detecting malicious encrypted communication is a complex and crucial cybersecurity issue. Feature engineering and machine learning methods play significant roles in achieving effective detection and classification of malicious encrypted communication. By employing well-designed feature engineering and appropriate method selection, it is possible to achieve accurate and efficient detection of malicious encrypted communication. However, in practical applications, challenges such as insufficient data and limited computational resources still exist, necessitating further research and exploration of more effective solutions.

Future research directions may include further optimizing feature engineering methods, developing feature representations more suitable for detecting malicious encrypted communication, and exploring more advanced and efficient deep learning models. Through continuous improvement and innovation, we can jointly build a safer and more stable network environment, better safeguarding the information security of users and businesses.

Professionals and researchers in the field of cybersecurity can utilize the theoretical guidance and practical advice provided in this paper to choose suitable feature engineering methods and machine learning models, enhancing the accuracy and efficiency of detecting malicious encrypted communication. Furthermore, we should closely monitor the latest developments and technological advancements in the field of network security, continuously improving and optimizing malicious encrypted communication detection techniques, and actively contributing to building a more secure network environment. Through collaborative efforts, we can establish a more robust and reliable network security defense, safeguarding the safety and stability of the digital world.

References

[1] kumar, M.N., Rao, D.S., & Sravanthi, D. (2011). A Novel Approach for Cheating Prevention through Visual Cryptographic Analysis. International Journal of Computer Science & Engineering Survey, 2, 123-131.

[2] Yang, J., & Lim, H. (2021). Deep Learning Approach for Detecting Malicious Activities Over Encrypted Secure Channels. IEEE Access, 9, 39229-39244.

[3] Pastor, A., Mozo, A., Vakaruk, S., Canavese, D., López, D.R., Regano, L., Gómez-Canaval, S., & Lioy, A. (2020). Detection of Encrypted Cryptomining Malware Connections With Machine and Deep Learning. IEEE Access, 8, 158036-158055.

[4] Cabaj, K., Gregorczyk, M., & Mazurczyk, W. (2016). Software-Defined Networking-based Crypto Ransomware Detection Using HTTP Traffic Characteristics. Comput. Electr. Eng., 66, 353-368.

[5] Zhang, Z., Qi, P., & Wang, W. (2019). Dynamic Malware Analysis with Feature Engineering and Feature Learning. ArXiv, abs/1907.07352.