Computing, Performance and Communication Systems (2025)
Clausius Scientific Press, Canada
DOI: 10.23977/cpcs.2025.090105
ISSN 2371-8889 Vol. 9 Num. 1

Research on Computer Network Virus Defense Method Based on Data Mining Technology

Wang Zeding

North China Electric Power University, Baoding, Hebei, China

Keywords: Data Mining Technology; Computer Network Viruses; Defense Methods

Abstract: With the rapid development of computer network technology, virus attack methods are becoming increasingly complex and concealed, and traditional defense methods relying on signature database matching have difficulty coping with unknown threats. Data mining technology, with its powerful data processing and pattern discovery capabilities, provides new ideas for network virus defense. By analyzing massive amounts of data such as network traffic and system logs, the potential patterns of virus behavior can be mined, and dynamic detection models can be constructed to achieve a transformation from passive response to active defense. The research combines classification, clustering, and association rule algorithms to explore key technologies for virus feature extraction and behavior prediction, aiming to break through the limitations of traditional methods, improve detection efficiency and accuracy, and lay a foundation for building an intelligent and adaptive network security system.

1. Introduction

Computer network viruses have become a significant factor threatening the stability of the digital society, and their spread speed and destructive power increase exponentially with technological iterations. Traditional defense mechanisms rely on static signature database updates, and they respond sluggishly when facing zero - day attacks and variant viruses, resulting in frequent protection gaps. Data mining technology can identify abnormal patterns from complex noise through in - depth analysis of network behavior data, providing a dynamic solution for virus detection. For instance, machine - learning - based classification algorithms can accurately distinguish normal traffic from malicious code, and cluster analysis can reveal the group characteristics of virus transmission paths. However, how to balance algorithm efficiency and detection accuracy and reduce the false - alarm rate remains an urgent problem to be solved. This research focuses on optimizing data pre - processing and feature engineering and explores a defense model with multi - algorithm collaboration, aiming to provide technical references with both theoretical value and practical significance for the field of network security [1].

2. Overview of Computer Network Virus

2.1. Definition and Characteristics of Network Virus

Network viruses are program entities composed of malicious code, and their core objective is to achieve illegal intrusion, self - replication, and systematic destruction through the network environment. The spread mechanism usually relies on vulnerability exploitation or social engineering means. For example, it can induce users to execute by disguising as legitimate files or spread horizontally to other hosts through system service vulnerabilities. The structure of virus code shows highly dynamic characteristics. Some variants can modify their own signatures in real - time to bypass static detection rules, while others gradually penetrate the defense system through multi - stage payload delivery. From a technical perspective, virus behavior patterns often include key steps such as concealed residency, privilege escalation, and lateral movement. In the residency stage, persistent control is often achieved by modifying the registry or injecting processes, and privilege escalation relies on undisclosed system vulnerabilities or configuration defects to obtain higher operation permissions. Lateral movement often uses automated scripts to scan internal network weaknesses and combines credential theft technology to break through security boundaries. The destructiveness of such malicious entities is not only manifested in data theft or system paralysis, but the deeper threat lies in building botnets or extortion ecosystems to form sustainable attack chains.

2.2. Types and Hazards of Common Network Viruses

Computer network viruses can be divided into three categories according to their behavior patterns and destruction targets: extortion - type, control - type, and information - stealing type. Ransomware hijacks users' data with high - intensity encryption algorithms and demands ransoms. Its encryption mechanism often uses asymmetric algorithms to block conventional recovery methods, forcing victims into a payment dilemma. Botnet viruses infect a large number of terminal devices to build a group of remotely controllable nodes. Attackers use command - and - control servers to launch distributed denial - of - service attacks or spam storms, leading to the paralysis of critical services and the exhaustion of network resources. Spyware focuses on long - term lurking and information stealing, using technologies such as keylogging, screen capture, and memory scraping to steal sensitive information like identity credentials and financial data. Some variants can even bypass sandbox detection to establish covert communication channels. These three types of viruses often mix social engineering and zero - day vulnerabilities during the spread process to form a composite attack chain [2]. Their harmfulness increases geometrically as the dependence on network infrastructure deepens, directly threatening the core assets and operational security of the digital economy.

2.3. Current situation and challenges of network virus defense

The current network virus defense system is mainly built on signature matching and behavior rule libraries. The static detection mechanism is prone to create protection gaps when dealing with polymorphic code and zero - day attacks. Traditional signature database updates rely on manual analysis to extract virus samples. The lag causes new malicious programs to spread rampantly during the window period. Although dynamic detection based on behavior analysis can identify some unknown threats, it is difficult to balance security and availability due to the high false - alarm rate. Attackers use obfuscation techniques to dynamically deform virus payloads and combine legitimate digital certificates with whitelist hijacking methods, making the boundary between

malicious behavior and normal operations increasingly blurred. The development of defense technology still needs to break through the computing power bottleneck of real - time traffic in - depth analysis and solve the technical blind spots in identifying malicious traffic in encrypted communication scenarios. Some advanced persistent threats even use supply - chain pollution to pre - position the attack chain, forcing the defense system to transform from terminal detection to full - lifecycle monitoring.

3. Basics of Data Mining Technology

3.1. The concept and process of data mining

Data mining technology, as a key branch in the field of information processing, focuses on extracting implicit and non - obvious knowledge from massive heterogeneous data. Its technical framework consists of three major modules: data cleaning, feature engineering, and pattern recognition. In data cleaning, the problems of noise interference and missing value filling in the original data need to be solved. Feature engineering transforms high - dimensional information into interpretable low - dimensional vectors through dimensionality reduction or transformation operations. In the pattern recognition stage, classification, clustering, or association rule algorithms are used to establish mapping relationships between data. For example, the decision tree algorithm divides the feature space based on information gain, and the support vector machine uses the kernel function to project non - linear problems into a high - dimensional linear space for solution. The actual operation process usually follows the CRISP - DM cross - industry standard, successively completing the stages of business understanding, data preparation, modeling, evaluation, and deployment. In the modeling stage, supervised or unsupervised learning paradigms need to be selected according to the target scenario. The key difference between data mining and traditional statistical analysis lies in its ability to process unstructured data, such as the time - series features in network traffic or the sentiment tendency in log texts. Its value is particularly significant in the field of network security, as it can identify abnormal patterns and potential threats from seemingly disordered communication behaviors [3].

3.2. Commonly Used Data Mining Algorithms

The theoretical framework of data mining algorithms provides multi - dimensional analysis tools for identifying network virus behaviors. The decision tree algorithm recursively divides the feature space based on information gain or Gini coefficient to generate a tree - like structure to distinguish normal traffic from potential threats, and its rule - generation process is highly interpretable, which facilitates security personnel to understand the classification logic. The support vector machine algorithm searches for the optimal hyperplane in the high - dimensional feature space and uses kernel functions to handle non - linearly separable data, separating virus attack patterns from regular operations. The Bayesian network algorithm constructs a directed acyclic graph to represent the probabilistic dependencies between feature variables, infers the possibility of abnormal behaviors based on prior probabilities and conditional probabilities, and is suitable for dealing with uncertain association information in network logs. The random forest algorithm improves the generalization ability by integrating the prediction results of multiple decision trees, and each tree is trained on a subset of features and samples to reduce the risk of overfitting. The isolation forest algorithm uses a random partitioning strategy to measure the path length of data points, quickly detects high - dimensional abnormal points deviating from the main distribution, and is suitable for rapid threat location in real - time monitoring scenarios. The latent Dirichlet allocation model extracts implicit themes from text - based logs, captures the semantic associations of virus attack

behaviors, and assists in discovering hidden malicious code propagation patterns.

4. Network Virus Defense Methods Based on Data Mining Technology

4.1. Data Acquisition and Preprocessing

Data acquisition and preprocessing, as the fundamental part of the network virus defense system, have the core task of building a highly reliable original dataset. Data sources usually cover network traffic mirrors, terminal log records, and virus sample libraries. Among them, network traffic mirrors need to capture all communication packets with the help of switch port mirroring technology, and terminal log records rely on agent programs to collect process behavior and registry change information in real - time. The virus sample library integrates the dynamic analysis results of sandboxes and the malicious family labels provided by threat intelligence platforms, providing annotation basis for subsequent feature engineering. The preprocessing process includes three parts: data cleaning, format standardization, and feature extraction. Data cleaning requires removing duplicate packets or log breakpoints caused by network jitter. Format standardization unifies multi - source heterogeneous data into structured time - series storage. Feature extraction focuses on parsing protocol header fields, payload hash values, and session behavior patterns. For example, it extracts User - Agent fingerprints from HTTP traffic and counts domain name access frequencies for DNS requests. Normalization processing eliminates the interference of different dimensions on model training. For example, it converts timestamps into relative time differences and encodes IP addresses in segments. Data dimensionality reduction uses principal component analysis or information gain algorithms to select key features, avoiding excessive consumption of computing resources caused by high - dimensional sparsity. The preprocessing results directly affect the generalization ability and detection accuracy of subsequent models, and a balance needs to be sought between data completeness and computing efficiency [4].

4.2. Virus Feature Extraction

Virus feature extraction focuses on capturing the essential identifiers of malicious behaviors from preprocessed data, and its technical routes cover three categories: static features, dynamic features, and behavior sequence features. Static features are centered on the structure of the virus itself. Through disassembly technology, opcode sequences and imported function tables are extracted, and the N - gram algorithm is used to count the frequency of instruction combinations. Some studies introduce entropy calculation to evaluate the randomness of code segments for identifying packed samples. Dynamic features are based on capturing process memory operations, registry modifications, and network connection behaviors in a sandbox environment. API call chains are used to reconstruct the execution path of malicious code, and the abnormal fluctuations of system call intervals are analyzed in combination with the time dimension. Behavior sequence features emphasize mining the correlations of multi - stage attacks. For example, the heartbeat packet features of botnets are matched through the time - series patterns of session traffic, or the hidden Markov model is used to describe the stage - transition rules of ransomware encryption behaviors. In the feature fusion stage, graph embedding technology is often used to map heterogeneous features into a unified vector space. The nodes in the graph structure can represent entities such as processes, files, and registries, and the edge weights quantify the interaction intensity between entities. To defend against adversarial samples, a noise - tolerance mechanism needs to be embedded in feature engineering. For example, wavelet transform is applied to network traffic features to filter protocol camouflage interference, or adversarial training is used to enhance the robustness of the model against code obfuscation attacks. Feature dimension reduction often

uses mutual information theory to select key indicators strongly related to virus categories, avoiding the distortion of the decision boundary of the classifier caused by redundant features.

4.3. Constructing Virus Defense Model

The construction of virus defense models centers around the core logic of feature space and classifier design. Classifier design needs to take into account both the dimensions of the feature space and the characteristics of sample distribution. The ensemble learning framework often combines the random forest and gradient - boosting tree algorithms to improve generalization ability. Supervised learning models rely on labeled samples to train classification boundaries. The logistic regression model fits the probability distribution of malicious samples through the sigmoid function, and the support vector machine uses the kernel trick to solve the problem of linear inseparability in the feature space. Unsupervised models are suitable for zero - day attack detection scenarios. Autoencoders capture abnormal residual patterns when reconstructing input data, and clustering algorithms divide potential threat groups according to sample similarity. Graph neural network models are good at processing network topology - related features. They encode the relationships between host nodes and communication edges into embedding vectors and use the message - passing mechanism to capture collaborative attack signals from multi - hop neighbors. The model update mechanism needs to design an incremental learning strategy to cope with the evolution of virus variants. A sliding time window is used to select new samples to optimize the decision boundary, and the online learning module adjusts model parameters according to real - time detection results. Defending against adversarial attacks requires the model to embed robustness constraints. An adversarial sample gradient penalty term is introduced into the loss function, and the Monte Carlo dropout technique is used to evaluate the uncertainty of prediction results to identify potential adversarial perturbations [5]. The model interpretability module uses methods such as LIME or SHAP to visualize the contribution of key features, assisting security analysts in verifying the credibility of detection logic. Transfer learning technology uses public threat intelligence libraries to pre - train model parameters, alleviating the over - fitting risk in small - sample scenarios.

4.4. Model Training and Optimization

Model training and optimization focus on parameter space search and generalization performance improvement. The Xavier method is often used in the parameter initialization strategy to balance the distribution of activation values of neurons in each layer and avoid the phenomena of gradient vanishing or explosion. The design of the loss function needs to match the task characteristics. The cross - entropy loss is suitable for the multi - classification task of virus family recognition, and the contrastive loss function enhances the aggregation of similar samples in the feature space. The choice of optimization algorithm needs to balance the convergence speed and the risk of local optimum. The Adam algorithm combines the momentum term and the adaptive learning rate mechanism to maintain the stability of parameter updates in the scenario of sparse gradients. Regularization techniques use the Dropout layer to randomly mask neuron connection paths to suppress overfitting, and the L2 norm constraint incorporates weight decay into the loss calculation process. The learning rate scheduling strategy uses the cosine annealing or hot restart mechanism to dynamically adjust the update step size and conduct a fine - grained search in the flat area of the loss surface in the later stage of training [6]. The batch normalization layer standardizes the input distribution of the intermediate layer to alleviate the negative impact of internal covariate shift on the training speed. The ensemble learning framework integrates the differences in decision boundaries of multiple base models, and the Stacking method uses the output probabilities of

primary classifiers as secondary meta - features for training. The cross - validation strategy divides the training set and the validation set to evaluate the generalization ability of the model, and the early - stopping method monitors the loss curve of the validation set to terminate ineffective training epochs. Hyperparameter tuning relies on Bayesian optimization to construct a surrogate model of the objective function to approximate the globally optimal parameter combination within a limited number of trials. Knowledge distillation technology transfers the knowledge of a complex teacher model to a lightweight student model, taking both detection accuracy and deployment efficiency into account [7].

5. Conclusion

The application of data mining technology in network virus defense marks a paradigm shift of security protection from rule - driven to data - driven. Experiments show that the defense model based on multi - dimensional feature analysis can effectively identify known viruses and intercept some unknown attacks, and its detection accuracy is significantly improved compared with traditional methods. However, the generalization ability of the model is limited by data quality and algorithm complexity, and the real - time requirement poses challenges to computing resources. Future research needs to further integrate deep learning and edge computing technologies, optimize the automation level of feature extraction, and build a lightweight and scalable defense framework. Network security is a long - term battle of attack and defense. Only by continuously innovating technological means can we gain an advantage in the spiral evolution of viruses and defenses.

References

[1] Yin H. Computer Network Virus Defense Technology Based on Data Mining Technology[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2018, 382(5): 052019.

[2] Hou J, Ding F .Design and Analysis of Network Virus Defense System Based on Multimodal Data Mining Technology [J].Scientific Programming, 2022, 2022(000):11.

[3] Sheng J .Research on SQL Injection Attack and Defense Technology of Power Dispatching Data Network: Based on Data Mining [J].Mobile Information Systems, 2022, 2022(000):8.

[4] Zuo C .Defense of Computer Network Viruses Based on Data Mining Technology [J].Int. J. Netw. Secur. 2018, 20:805-810.

[5] Yin H .Computer Network Virus Defense Technology Based on Data Mining Technology[C]. IOP Conference Series: Materials Science and Engineering, 2018:18. DOI 10.1088/1757-899X/382/5/052019

[6] Lan C .Data Mining Technology Based on Granular Computing in Computer Network Virus Defense [C]//International Conference on Frontier Computing. Springer, Singapore, 2021:12-14.

[7] Jingzi D, Xuangong L .Design and Implementation of Network Virus Defense System Based on Data Mining Technology[J].China Computer & Communication, 2018:13-15.