Information Security Education Journal (ISEJ)

Print ISSN: 2349-8161
Online ISSN: 2349-817X

ISEJ 2024;11 (2)
https://doi.org/10.6025/ISEJ/2024/11/2/37-43

Network Security Education with Machine Learning Optimization

Xiaoyan Yao[1*] Wei Liu[2]
[1]Department of Information Engineering
Shandong Foreign Trade Vocational College
266100, Qingdao, Shandong, China

[2]Qingdao Gaoxin Vocational School
Qingdao, Shandong, China
yanziyao2000@163.com

ABSTRACT

The current network security environment is increasingly severe, and more than static network protection mechanisms are needed to meet network security demands. Dynamic intrusion detection methods can provide real-time protection and detection for the network security environment. Therefore, this paper builds an intrusion detection education based on the machine learning algorithm and random forest and optimizes the model using the FSCA and TPE algorithms. Experimental results show that the intrusion detection model in this study has higher detection efficiency than other models and demonstrates strong adaptability, enabling rapid identification of intrusion behaviors and timely response in different intrusion detection environments, thereby improving the stability and practicality of the model.

Received: 10 June 2024
Revised: 1 September 2024
Accepted: 12 September 2024

Keywords: Machine Learning, Random Forest, Network Education, Network Security Optimization

1. Introduction

With the widespread application of Internet technology, cyberspace has become an indispensable part of people's lives, studies, and work. Since entering the information age, the development of big data, the Internet of Things, cloud computing, and other technologies has promoted the growth of cyberspace. The scale of Internet users is rapidly increasing, and network traffic is also growing exponentially. According to relevant survey results, normal human-generated traffic accounts for only about 60% of the generated network traffic, while malicious traffic generated by machines accounts for close to one-fourth of the total traffic [1]. This indicates that while the development and application of the Internet and information technology have facilitated various aspects of people's lives, they have also brought new challenges and threats. Although traditional network security issues such as malicious programs and vulnerability threats have been effectively controlled, their corresponding threats have declined. However, network attack techniques are constantly being updated,

and compared to the past, they are more organized, targeted, and time-sensitive. The number of event-based and high-risk zero-day vulnerabilities is on the rise, making the security situation in cyberspace increasingly complex [2].

Education for network space security refers to the monitoring and protection of software, hardware, and data information in computer information systems and the uninterrupted maintenance of network services. To achieve network space security, it is necessary to ensure confidentiality, integrity, availability, controllability, and non-repudiation of information systems. Currently, common network attacks are based on these characteristics and involve interruptions, interceptions, tampering, and counterfeiting of network data [3]. Various security mechanisms are commonly used to enhance network security, including data encryption, identity authentication, access control, firewalls, and intrusion detection. With the expansion of network information system scale, the first four static security mechanisms can no longer provide strong security guarantees in the current severe network security environment, gradually revealing limitations. In contrast, as a dynamic security mechanism, intrusion detection technology can identify intrusion behavior and take proactive intervention measures in real-time data information situations before intrusions or potential harm [4]. However, intrusion detection technology usually requires security personnel to manually set up matching conditions, which may increase the capture rate by lowering the detection rule matching conditions. However, this also increases the false alarm and missed detection rates, thus increasing the cost of handling warning events. Therefore, this paper introduces machine learning technology to optimize network security and builds a random forest intrusion detection model combined with FSCA and TPE optimizations, which allows it to flexibly adapt to different network attack scenarios while maintaining a high capture rate without reducing matching conditions, reducing false alarms and missed detection, and improving network security.

2. Current Research Status of Computer Network Security Optimization

In the history of network optimization research, the development of network intrusions and attacks has always driven the study of network optimization. In this passive state, some scholars proposed the concept of real-time intrusion detection models. These models are based on existing data to construct corresponding behavioral databases and match system logs, network node logs, and other data. This approach effectively identifies various systems intrusions and timely warnings [5]. Other scholars introduced the concept of network monitors, which monitor the traffic data in the network and serve as the basis for detecting intrusion behaviors [6]. Coupling theory has also been applied to intrusion detection, and scholars have defined the security detection framework based on these research results, determining the data format and interaction process of IDS. With the widespread application of machine learning, many researchers have incorporated it into the study of network security optimization. Some scholars pointed out that classic machine learning algorithms like fuzzy clustering and Bayesian optimization have issues such as multi-valued and non-linear optimal solutions with low accuracy. However, by using swarm intelligence algorithms, machine learning algorithms can be effectively improved, enhancing the performance of intrusion detection models [7]. Other scholars proposed a support vector machine model for high-dimensional and small-sample network datasets, and experimental results showed that the model had clear advantages, avoiding overfitting and reducing errors, with good generalization ability to adapt well to other datasets [8]. Additionally, some researchers pointed out that clustering forests and neural networks perform well in pattern recognition problems, effectively shortening the model training time with less prior knowledge [9]. There are also studies combining LSTM models and hyperparameter optimization techniques to improve the accuracy of intrusion detection models [10]. Machine learning algorithms are widely used in network security optimization, and many studies have achieved good results. However, there are still issues, such as data redundancy, inaccurate parameter selection, and some low indicator values, that require further research and resolution.

3. Machine Learning-Based Network Security Optimization Model

3.1. Logistic Regression and Random Forest in Machine Learning Algorithms

A typical computer network intrusion detection system consists of data collection, analysis, and response mechanisms. The response mechanism can be divided into active response and passive response, and based on the information acquisition method, it can be categorized into HIDS, NIDS, or a hybrid IDS. A machine learning-based IDS consists of a data collection module, an analysis module that extracts intrusion behavior feature information and reduces dataset redundancy, and a

model establishment and evaluation module. In this paper, logistic regression and random forest were chosen to construct a machine learning-based intrusion detection model, considering the characteristics of intrusion detection technology and the security needs of cyberspace.

Logistic regression is one of the classification algorithms in machine learning, commonly used for binary classification problems. This study will be used in the detection scenario to predict whether corresponding data indicates intrusion behavior. The simplified formula is shown as (1):

$$\begin{cases} P(y=1|x,w) == \dfrac{1}{1+e^{-w^T x}} \\ P(y=0|x,w) == \dfrac{1}{1+e^{w^T x}} \end{cases} \quad (1)$$

In this context, the input features of the traffic data are denoted as and, and the final output is denoted as, representing benign or malicious traffic. The weight vector is represented as. The logistic regression model is a nonlinear model that estimates through maximum likelihood estimation. Before estimation, the likelihood function of model parameters must be completed to convert the probability of observed data. Then, the function's maximum value is obtained based on the estimated parameter values.

The random forest algorithm's base classifier is the CART classification tree, which includes multiple classification trees. Each classification tree corresponds to a training subset formed by independent sampling and maintains a consistent probability of feature selection. This process is repeated several times. These classification trees maintain independence during the growth process, but they are influenced by the random vector of the random distribution, resulting in differences between the results. A simple vote on the results of the classification trees obtains the final classification result. The CART classification tree consists of three parts: root, intermediate nodes, and leaf nodes. The root node is the entry point for all input sample data, and each input sample has the same number of feature attributes that form the candidate feature set for splitting. Each feature attribute is traversed, and there is only one opportunity for selection. In the intermediate nodes, the optimal classification feature and corresponding value are selected from the sample based on the Gini coefficient principle, and the process is repeated until all samples contained in the leaf node belong to the same class. At this point, the candidate feature set is empty, and the splitting stops. When the Gini index is at its minimum, the splitting node contains only samples of the same type; when the Gini index is at its maximum, the splitting node contains samples of different types. The independent sampling of the CART classification tree is performed by random sampling with replacement. Therefore, the probability of being selected for each sample in each sampling is equal. The probability of not being selected in the sampling is shown in (2):

$$P = (1-\frac{1}{i})^i \quad (2)$$

Suppose the number of original samples in the above equation is sufficiently large. In that case, the probability value will gradually converge to a convergence value, which means that only about 63.2% of the samples in the sample set may exist in the training subset, and the rest belong to "out-of-bag data," namely OOB.

3.2. Random Forest Intrusion Detection Model Optimized with FSCA and TPE Algorithms

In the intrusion detection model, if the dimensionality of the dataset is high and there is a large amount of data redundancy, which leads to increased computational complexity of the model and inefficient response mechanisms, these issues can be addressed through model optimization. One of the essential aspects of machine learning is model optimization, and hyperparameter optimization can help determine the optimal hyperparameters of the machine learning model when evaluating its performance on the validation dataset. This study chooses an automatic hyperparameter tuning method to optimize the intrusion detection method. The automatic hyperparameter tuning method requires the formation of knowledge regarding the relationship between hyperparameter values and

model performance. Based on this prior knowledge, the next set of hyperparameters is inferred, reducing the number of experiments and shortening the experimental time to some extent during the search process. Automatic hyperparameter tuning is mainly achieved through the Sequential Model-Based Optimization (SMBO) algorithm, where the target function generated in the evaluation domain often uses Gaussian process methods and the TPE (Tree-structured Parzen Estimator) algorithm for modeling, which have certain differences in modeling. Considering the needs and actual conditions of the intrusion detection method in this paper, the TPE algorithm is selected for modeling. In the TPE algorithm's modelling, the non-parametric density is based on changes in the feature space and the observed values are replaced during this process, allowing the TPE hyperparameter optimization algorithm to choose the density freely. Formula (3) shows the description of under two density conditions:

$$p(x|y)=\begin{cases} l(x) & if \ \ y < y^* \\ g(x) & if \ \ y \geq y^* \end{cases} \tag{3}$$

Where the target function of the observed value is less than, it forms the density. Otherwise, it forms the density. This algorithm uses the upper part of the observations above the optimal observed result and the lower part below the observed result as the observations. When the TPE observations in the observation domain remain sorted, the variation between the iteration duration and the number of improved features is linear, as shown in formula (4):

$$EI_{y^*}(x)=\int_{-\infty}^{\infty}(y^*-y)p(x|y)dy=\int_{-\infty}^{y^*}(y^*-y)\frac{p(x|y)p(y)}{p(x)}dy \tag{4}$$

Based on the above formula, the maximum and minimum probability are taken as the maximum improvement points for the observation. After each iteration, the point with the maximum probability is returned, as shown in formula (5):

$$EI_{y^*}(x)=(\lambda+\frac{g(x)}{l(x)}(1-\lambda))^{-1} \tag{5}$$

Figure 1 shows the flowchart of the FSCA and TPE-optimized intrusion detection method based on random forests. From the diagram, it can be seen that the optimization process can be divided into three parts. First, the correlation matrix of feature data and the t-SNE algorithm are calculated based on the Pearson coefficient. Then, the feature relationships in the dataset are analyzed in depth, and the dataset is dimensionally reduced according to three feature selection methods. At the same time, the hyperparameters of the random forest are improved using the TPE algorithm based on SMBO, and the corresponding sample quantity adjustment is performed based on the obtained data weights.

4. Machine Learning-Based Network Security Optimization Model

To validate the machine learning-based network security optimization model proposed in this paper, a comparative experiment was conducted to compare the model's training time and testing time. The comparative experiment included Naive-Bayes, LR (Logistic Regression), AdaBoost, and RF (Random Forest) algorithms. The results are shown in Figure 2. From the training time perspective, the LR algorithm model has the longest training time, followed by the AdaBoost algorithm. Both of their training times are higher than the sum of the training times of the other three algorithms. The shortest training time for the Naive-Bayes algorithm model is less than 6 seconds. The traditional RF algorithm model's training time is 171.26 seconds, and the proposed model's training time is 37.6% lower than that of the RF algorithm model. Regarding the testing time, the AdaBoost algorithm model has the highest testing time, which is 46.68 seconds. The RF algorithm model's testing time is 15.28 seconds. Although the LR algorithm model has the longest training time, its testing time is in the middle range compared to the five

algorithms, slightly lower than that of the Naive-Bayes algorithm model. The proposed machine learning-based network security optimization model significantly reduces the training time compared to the RF algorithm model, achieving a reduction of about 62.2%. The results indicate that the Naive-Bayes algorithm model among the five intrusion

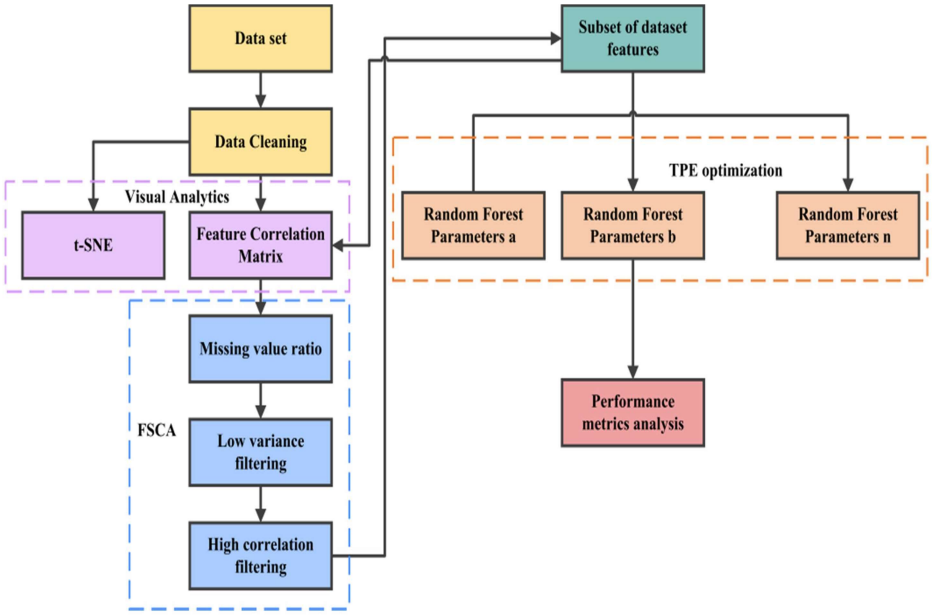

Figure 1. Flowchart of the FSCA and TPE optimized intrusion detection method based on random forests

detection models inherently assumes the independence between the features in the dataset, and its logic and algorithm simplicity and stability are relatively high, resulting in relatively low training and testing times. On the other hand, compared to the RF algorithm model, the proposed algorithm model effectively and significantly reduces both the training time and detection time, demonstrating its good adaptability in large-scale intrusion scenarios with high network concurrency and effectively improving the model's detection efficiency.

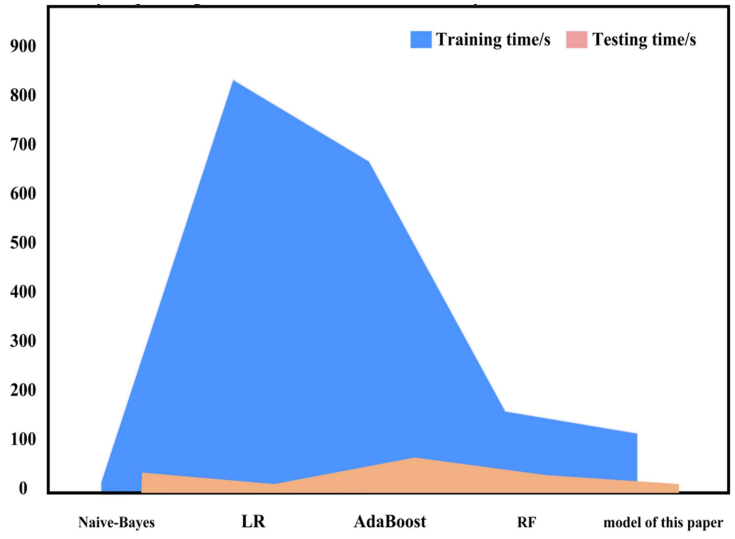

Figure 2. Comparison of training time and testing time results for five intrusion detection models

Network security education faces a constantly changing environment, and intrusion detection models need to adapt to different detection scenarios. Therefore, this paper compares the robustness of the five intrusion detection models under different environments, as shown in Figure 3. In this experiment, the ratio of the training set to the validation set is 1:1, and the Dummy Classifier function is selected as the baseline classifier for the dataset. The Naive-Bayes algorithm model shows consistency with the baseline in terms of accuracy, recall rate, and F1 score, and its precision is nearly twice that of the baseline. The LR and AdaBoost algorithm models perform better than the Naive-Bayes algorithm model in all four metrics, with the AdaBoost algorithm model having slightly higher values for the four metrics. All four metrics for the proposed algorithm model are significantly higher than the baseline and noticeably higher than those of the other three algorithm models. This indicates that the proposed intrusion detection model demonstrates better robustness and stability than other algorithm models, showing strong adaptability. Additionally, it can meet the requirements with fewer labeled data samples in different intrusion detection scenarios, enhancing its practicality.

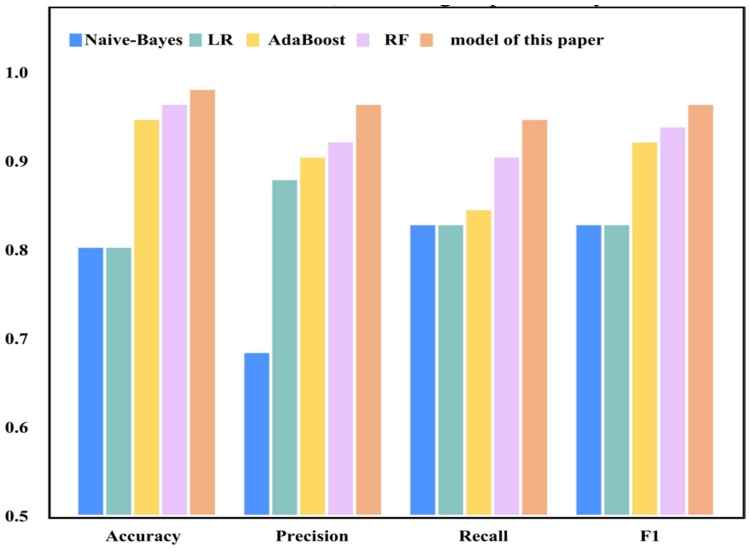

Figure 3. Comparison of the robustness of the five intrusion detection models in different environments

In conclusion, the random forest intrusion detection model optimized with FSCA and TPE effectively reduces the model's training time and testing time and demonstrates good robustness, making it suitable for various intrusion detection scenarios with stable performance. Compared to other algorithm-based intrusion detection models, this model exhibits higher precision and effectiveness and improves the efficiency and performance of the detection process.

5. Conclusions

With the development of information technology and internet technology, network security environments have become increasingly complex, and network attacks and intrusions continue to evolve. Traditional static network protection methods are no longer sufficient to meet the current demands for network security. Instead, dynamic intrusion detection methods can provide real-time monitoring and timely response to ensure effective network security protection. Therefore, this paper constructed an intrusion detection method based on the random forest algorithm in machine learning and optimized the algorithm model through FSCA and TPE algorithms. The experimental results show that the optimized random forest intrusion detection method significantly reduces the model's training and testing time, improving the detection efficiency. Moreover, compared to the baseline and other algorithm models, it substantially improves accuracy, precision, recall rate, and F1 score, showing better robustness. This indicates that the proposed machine learning-based network security optimization method possesses strong adaptability, effectively detecting intrusion behaviour in different scenarios and responding promptly with efficient measures.

References

[1] Najafabadi, Maryam M., Villanustre, Flavio, Khoshgoftaar, Taghi M., Seliya, Naeem, Wald, Randall, Muharemagic, Edin. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data 2*(1) 1–21.

[2] Shah, Hitarth, Kakkad, Vishruti, Patel, Reema, Doshi, Nishant. (2019). A survey on game theoretic approaches for privacy preservation in data mining and network security. *Procedia Computer Science 155* 686–691.

[3] Liao, H. J., Lin, C. H. R., Lin, Y. C., Others. (2013). Intrusion detection system: A comprehensive review. *Journal of Network & Computer Applications 36*(1) 16–24.

[4] Nguyen, Hai T. (2012). Reliable machine learning algorithms for intrusion detection systems: Machine learning for information security and digital forensics. *Acta Physiologica Scandinavica 142*(2) 191–199.

[5] Kim, J., Kim, J., Kim, H. (2020). CNN-based network intrusion detection against denial-of-service attacks. *Electronics 9*(6) 916.

[6] Tealab, Ahmed. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal 3*(2) 334–340.

[7] Giang Nguyen., Stefan Dlugolinsky., Martin Bobák., Viet Tran., Alvaro López García., Ignacio Heredia., Peter Malík., Ladislav Hluch. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey. *Artificial Intelligence Review* 52(1) 77–124.

[8] Boutaba, R., Salahuddin, M. A., Limam, N., et al. (2018). A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities. *Journal of Internet Services and Applications* 9(1) 1–99.

[9] Kim, Jaeseok., Cauli, Nino, Vicente., Pedro, Damas., Bruno, Bernardino., Alexandre, Santos-Victor, José., Cavallo, Filippo. (2020). Cleaning tasks knowledge transfer between heterogeneous robots: A deep learning approach. *Journal of Intelligent & Robotic Systems: With a Special Section on Unmanned Systems 98*(1) 191–205.

[10] Chouhan, Siddharth Singh., Singh, Uday Pratap., Jain, Sanjeev. (2020). Applications of computer vision in plant pathology: A survey. *Archives of Computational Methods in Engineering: State of the Art Reviews 27*(12) 611–632.