ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20240305001375

Original scientific paper

Automated Network Attack Detection Techniques Based on Improved Random Forests

Ke XIANG*, Xing YANG

Abstract: The current network environment needs to face massive data and information; to ensure the network security, the automated detection technology of network attacks needs to be strengthened urgently. Therefore, the study adopts an improved random forest model based on Spark big data platform to optimize the detection process through online analysis and offline feedback mechanism. The focus is on the machine learning detector constructed by utilizing the MLlib library, and the semi-voting mechanism is used to speed up the model prediction. The experiments involve weight calculation methods weighted by correlation coefficients and enhance the model generalization ability by means of combinatorial optimization problems. The study also calculates decision tree similarity for random forest algorithm optimization in conjunction with packet characterization. In the experiments, on the UNSW-NB15 dataset, the improved model achieved a detection accuracy of 0.68 when using correlation coefficient weighting and tended to be stable with 16 decision trees. On the CICIDS2017 dataset, the detection accuracy obtained by this weighting method was 0.73 and stabilized with 12 decision trees. The Relative-RF-50% model using the semi-voting mechanism improved the prediction accuracy to 0.93844 on the CICIDS2017 dataset and obtained a substantial improvement in the execution time. Results show that the improved random forest model enhances the performance of automated cyber-attack detection, especially in terms of accuracy, recall, and efficiency showing obvious advantages.

Keywords: cyber-attack; correlation coefficient; detection; random forest; semi-voting

1 INTRODUCTION

As network technology rapidly advances and Internet applications become more widespread, ensuring network security has emerged as a significant global challenge. The means of network attacks have become increasingly cunning and advanced, making the traditional network security protection mechanism face unprecedented challenges. Currently, cyber-attacks are not only diversified and complex, but also highly covert, which makes the traditional rule- and signature-based detection methods difficult to adapt to the new attack patterns, especially the accuracy and efficiency problems when dealing with large-scale and high-dimensional data [1, 2]. In order to effectively address these challenges, automated detection of cyber-attacks using machine learning and big data techniques has become particularly critical [3, 4]. Machine learning methods, especially the random forest model in integrated learning, are increasingly used in cybersecurity due to their excellent classification performance, processing power, and good generalization performance [5]. However, with the increase of data dimensions and the update of attack methods, the traditional random forest model still has some limitations in terms of its effectiveness and efficiency in dealing with complex network environments [6]. To address this problem, the study proposes an automated detection technique for network attacks based on improved random forests. The core of the study is to improve the model accuracy in network attack detection by optimizing the weight calculation and voting mechanism of the random forest model. Specifically, this study adopts the weight calculation method of correlation coefficient for comprehensive analysis and experimental validation and introduces the semi-voting mechanism to further improve the detection speed and accuracy in different cyber-attack scenarios. The innovation of this research is that the weight calculation method of correlation coefficient is invoked, and a semi-voting mechanism is implemented to improve the prediction speed of detection. This method accelerates decision making by terminating voting early in the model decision making process, which is an innovative point to improve efficiency while maintaining high accuracy. The research is significant for ability improvement of network security protection and lays a solid foundation for future research on network security defense technology under the conditions of high-dimensional feature space. At the same time, the research also provides new perspectives and methods for the application of machine learning in network security.

2 RELATED WORKS

Automated detection of cyber-attacks is a hot issue in information security, especially in the data-driven era. Random forest, as an effective machine learning method, is applied in cyber security, especially in cyber-attack detection. It trains and classifies data by constructing multiple decision trees to achieve the identification of potential threats [7]. Some of the related researches by scientists and academicians are presented below. Mishra A. K. et al. proposed a machine learning based approach to successfully counter modern cyber-attacks by analyzing the patterns of encrypted network traffic. The strategy utilizes labeled datasets and combines synthetic attack and normal traffic features to improve detection efficiency. Classifiers such as light gradient enhancer, random forest and random gradient descent were applied for data analysis. Results showed that all these classifiers achieved the highest prediction accuracy [8]. Hussein A. Y. et al. proposed an intrusion detection system with an improved random forest algorithm that efficiently predicted security vulnerabilities in network traffic without compromising the efficiency of the network infrastructure. The system reported suspicious activities by analyzing copies of network traffic and sending alerts to administrators. The results showed that the system exhibited significant advantages in terms of accuracy [9]. Subbiah S. et al. proposed a novel intrusion detection system to cope with the growing security needs in Internet and computer applications. The system used feature selection and grid search random forest algorithms to effectively detect network intrusions. In comparison with other machine learning algorithms, the proposed model achieved 99% in attack detection accuracy, which was significantly better than other algorithms [10]. Niandong L. et al. proposed a network anomaly detection method that combines information entropy and random forest classification. The method first captured and extracted the metadata of network probe streams in real time, then performed feature selection and accurate classification by incremental learning. The results showed that the method efficiently located anomalies, significantly reduced the workload, and improved the reliability and early warning capability [11].

Srivani P. et al. proposed the use of Random Forest Classifier to identify and mitigate botnet attacks in IoT. As IoT devices became part of critical infrastructure, the potential risks of these attacks were increasing, including business disruption, data leakage, and physical risks. The classifier was widely used in areas such as cybersecurity by integrating multiple decision trees to improve prediction accuracy [12]. Anwer M. et al. proposed a framework for detecting malicious network traffic that combines three popular classification-based methods such as support vector machines, gradient boosting decision trees, and random forests. The algorithm was compared for performance, including training and prediction time, specificity and accuracy demonstrated a higher accuracy of 85.34% [13]. Gaur V. et al. proposed a system for early detection of DDoS attacks on IoT devices. Feature selection techniques such as Cardinality, Extra Tree and ANOVA were applied on four classifiers such as Random Forest, Decision Tree, k Nearest Neighbor and XGBoost. Experimental results showed that the method demonstrated excellent performance and effectively facilitated early detection [14]. Kayode-Ajala O. et al. proposed a method to effectively categorize network traffic into normal and abnormal categories using Random Forest classifier and its principal component analysis variant. The results showed that the random forest classifier performed well in achieving high accuracy, precision, and recall. Integrating PCA into Random Forest brought only minimal performance degradation, ensuring that the dimensionality reduction process did not affect the model effectiveness. Random forest analysis revealed that login attempts were the most critical feature in network classification, followed by the contact ratio of different target hosts for the same service [15].

In summary, these studies suffer from low efficiency in handling large-scale and high-dimensional data, lack of sufficient flexibility to adapt to novel and complex attack patterns, and sacrifice prediction speed while improving detection accuracy. Therefore, the study improves the ability to handle massive data by using the Spark big data platform and MLlib library and enhances the model's adaptability and response speed to novel attack patterns by introducing new weight calculation methods and semi-voting mechanisms.

3 AUTOMATED DETECTION OF NETWORK ATTACKS BASED ON IMPROVED RANDOM FORESTS

In the face of growing data size and increasingly complex cyber-attack patterns, the traditional random forest model still has some limitations, such as the ability to handle high-dimensional data, prediction efficiency, and adaptability to novel attack patterns. Therefore, the core goal of the research is to overcome these limitations through algorithmic improvement to enhance the model performance in the field of automatic cyber-attack detection.

3.1 Analysis of Automatic Network Attack Detection Technology

The network security analysis algorithm based on big data technology is mainly divided into three modules: online, offline and feedback. For the online security analysis module, it first calls the application programming interface that captures packets to capture network traffic data. Then removes redundant information through feature selection engineering on the Spark big data platform to retain key traffic features. Finally, key traffic features are load-balancedly assigned to cluster nodes that contain the MLlib library, which can automatically analyze the traffic packets online through the machine learning method and can detect most of the network attacks and abnormal traffic through supervised learning of the features [16]. Online analysis of traffic packets, and most of the network attacks and abnormal traffic can be detected by supervised learning of the features [16]. Its framework structure is shown in Fig. 1.

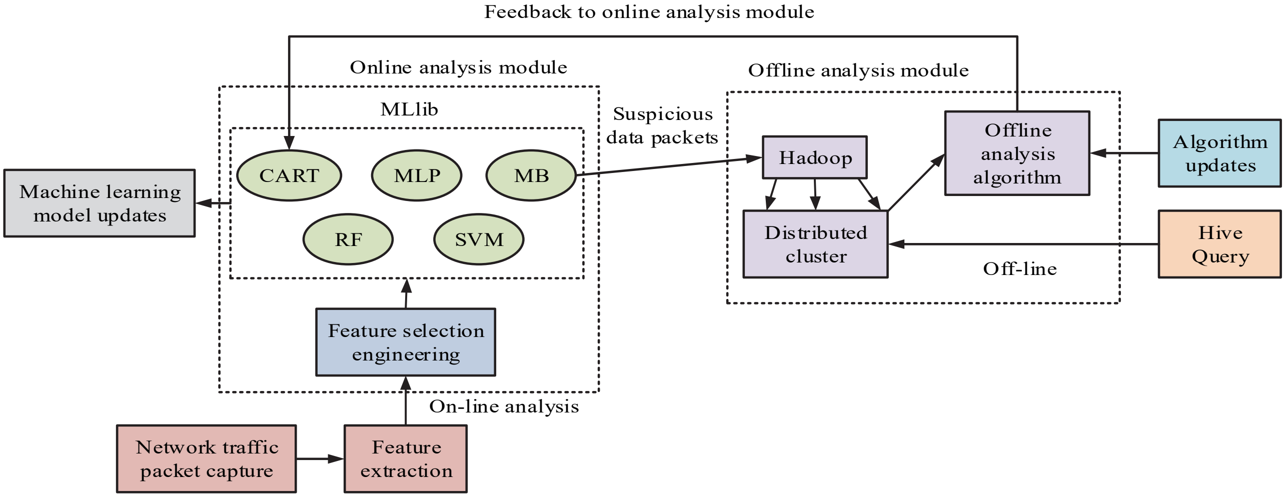


Figure 1 Framework for big data-based cybersecurity analysis

In order to realize most of the cyber-attacks and anomalous traffic detection, the machine learning models in the machine learning library MLlib provided on the big data platform Spark are selected to build detectors for online detection of anomalous traffic data. The MLlib library contains CART Decision Trees (CRAT), Random Forests (RF), Support Vector Machines (SVMs), Plain Bayes (NB) and Neural Networks (MLP), five machine learning models [17]. In terms of data characterization in online detection, in addition to the use of traditional packet traffic characteristics, some distributional features in the data and characteristics such as the dispersion of the samples in the response time slot and the full-fledged value of the distribution should also be taken into account. The study optimizes the Random Forest algorithm by calculating the similarity of decision trees in Random Forest, turning the above problem into a combined optimization problem, thus improving the Random Forest generalization capability. The workflow of random forest is shown in Fig. 2.

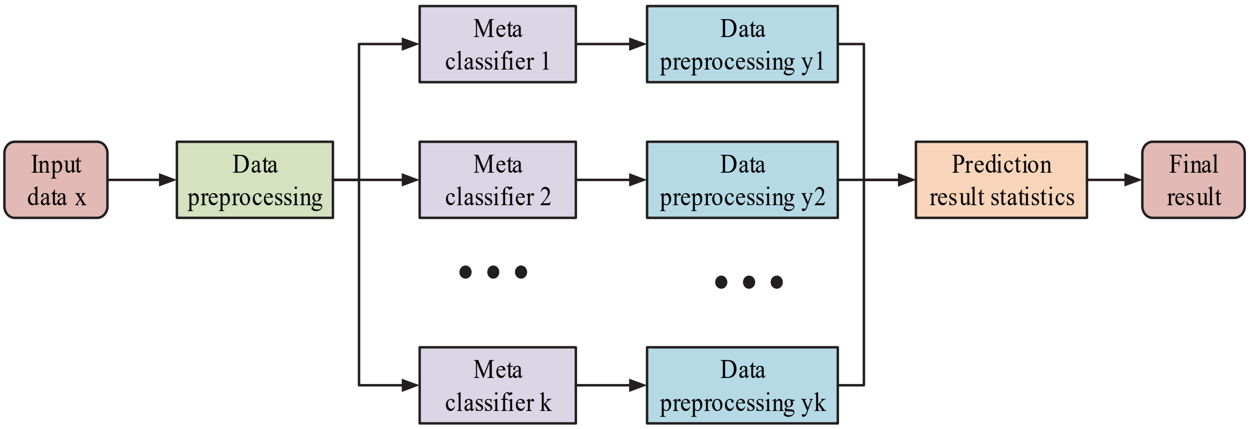


Figure 2 Random forest model workflow diagram

Assuming that the representation of a random forest is, the degree of similarity between a decision tree and a random forest is defined as shown in Eq. (1) [18].

$$\Gamma_i = \frac{1}{n-2}\left(\sum_{i=1}^{n} sim(tree_j, tree_i) - sim_{\max} - sim_{\min}\right) \tag{1}$$

When $i = j$, the similarity is 0. Subtracting both the maximum similarity and the minimum similarity removes the interference from the special decision tree model. The larger Γ_i is, the lower the influence factor and lower the decision tree's weight in the random forest. This similarity calculation can reduce the decision tree similarity and improve the RF generalization ability. In model optimization, the construction of loss function is very important to achieve the optimization objective through a reasonable loss measure. In a random forest model problem with a known sample/label set (x, y) distribution, the goal is to learn a series of classifiers that minimize the loss. If each decision tree is classified on a per decision tree basis, the final optimization objective can be derived as shown in Eq. (2).

$$\min_{f \in \theta} \frac{1}{n}\sum_{i=1}^{n} \varphi(\Gamma_i) \cdot L(y_i, f(x_i)) \tag{2}$$

In Eq. (2), φ denotes the influence factor function, L denotes the loss function, and f denotes the classifier. In the random forest model, the randomness of its sampling will lead to some differences in the predictive ability of the decision tree model obtained by the construction, and there is some irrationality in this voting mechanism. In order to make the model effective in detecting the real attacks in the network, the study uses the correlation coefficient for decision number voting weight assessment method for prediction.

3.2 Weighted Random Forest Model

To address the limitations of using correlation tables and graphs to accurately represent the degree of correlation between two variables, as well as their inability to serve as a criterion for evaluating the correlation measure, this study utilizes the correlation coefficient as a basis for determining the linear correlation between two variables. The correlation coefficient of two variables can be expressed by Eq. (3) [19].

$$r = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{k}(x_i - \bar{x})^2 \sum_{i=1}^{k}(y_i - \bar{y})^2}} \tag{3}$$

In Eq. (3), r represents the correlation coefficient and $\bar{x}$ and $\bar{y}$ represent the mean values of the variables. Through the calculation of Eq. (3), it can be found that the value of correlation coefficient is distributed from –1 to 1. When r is greater than 0, the two variables are positively correlated; when the value of r is less than 0, the two variables are negatively correlated; when the value of r is 1 or –1, the variables are absolutely correlated; when r is 0, there is no correlation between the variables. The correlation coefficient reflects the change influence of one variable on another variable, so the absolute value of the correlation coefficient should be taken when using the correlation coefficient of cyber-attack features to calculate the weight of cyber-attack features. For the multidimensional set of cyber-attack features, the corresponding set of correlation coefficients is obtained using the training samples, denoted as $\{r_1, r_2, ..., r_n\}$, which is constructed using this feature set and the training dataset to obtain the random forest model, for the decision tree definition of the voting weighted value is shown in Eq. (4).

$$p_r = \alpha \sum_{j=1}^{m} \left| r_j \right| \tag{4}$$

In Eq. (4), the voting weighting value of the decision tree is appropriately adjusted by adding the tuning factor because the value of the correlation coefficient is small. The random forest model possesses the weighted values, and for any input vector, the prediction results are represented as shown in Eq. (5).

$$\max\{c \left| c_i = \sum_{k}^{j=1} p_{r\_i} I(h_j(x) = l), l \in C \right.\} \tag{5}$$

In Eq. (5), C denotes the set of all category labels, l denotes the categorized labels, I denotes the schematic function, which takes the value of 1 or 0 when the prediction result of the decision tree is consistent with the category labels, p_j denotes the weighted value of the voting process, and c_i denotes the weighted result of the votes obtained by the category labels, the number of weighted votes of the individual categorized labels are counted, and the categorized labels with the largest number of weighted votes are taken as the output of the final prediction, and the data prediction is shown in Fig. 3.

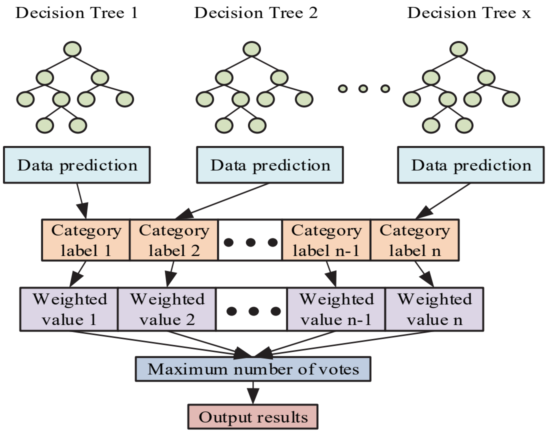


Figure 3 Schematic diagram of random forest result prediction weighted by correlation coefficient

From Fig. 3, the model has the same likelihood of each categorical label receiving votes during the voting process, and the final result of the model depends on the categorical label that receives the most number of votes, which is the one that gives the most number of votes as the prediction of the model, [20]. The study improves this voting pattern by introducing the pattern of half-voting quantity with the aim of improving the speed of model prediction without affecting the accuracy of the random forest algorithm. The total votes received by all decision trees in a traditional random forest model for the same categorical label, the total votes for the categorical label, for a random forest model without the introduction of voting weights, is equal to the number of decision trees included in the combined model, and the total votes can be expressed by Eq. (6).

$$S = \sum_{i=1}^{k} p_i \tag{6}$$

In Eq. (6), denotes the total voting volume. Therefore, the expression for the half-vote volume is shown in Eq. (7).

$$S_h = S / 2 \, (7)$$

In the process of the random forest model for the half-vote volume model, while the model calls a single decision tree one by one for outcome prediction and weighted voting, the model detects the current voting results of each categorical label, and if the number of votes obtained by a categorical label has already reached half of the full vote volume, it terminates the subsequent decision tree prediction and voting, and outputs this categorical label as the final prediction result. The flowchart of the half-vote volume algorithm is shown in Fig. 4.

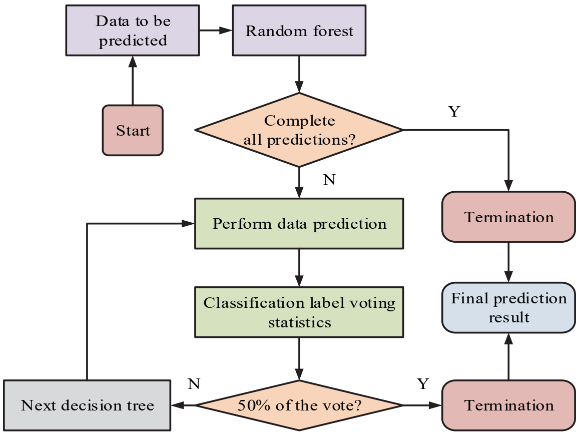


Figure 4 Flowchart of semi-voting volume model

As can be seen in Fig. 4, the semi-voting volume model will terminate the prediction process to obtain the prediction results before the traditional random forest model completes the voting of all the decision trees, but if the number of categorical labels with half of the full number of votes never occurs during the whole voting process, the termination condition of the whole semi-voting volume model will not be triggered, and it is necessary to give the model prediction results after all the decision tree classifiers have completed the data prediction and voting.

4 ANALYSIS OF AUTOMATED NETWORK ATTACK DETECTION TECHNIQUES BASED ON IMPROVED RANDOM FORESTS

The study designed a series of related experiments to verify the performance of the proposed improved random forest in the automatic detection of network attacks. For the improved random forest model, the study used four weight calculation methods to conduct comparison experiments, analyzed the prediction accuracy before and after the algorithm improvement, and counted the performance difference between the random forest algorithm with voting weights and the traditional algorithm when the decision tree was at 100. For the automatic detection of network attacks, the study used different datasets for validation and analysis, and evaluated and analyzed by four indicators: precision rate, recall rate, F1 score and execution time.

4.1 Performance Analysis of Random Forest Based on Weight Calculation

The study chose Spark 3.0, Hadoop 3.2, and Hive 3.1 to complete the simulation experiments, and selected UNSW-NB15 and CICIDS2017 datasets for experimental analysis and divided the dataset into training and testing sets according to the ratio of 8:2. The study used out of bag data (OOB), correlation coefficient (Relative), chi-square (Chi), and mutual information (Mltu) for comparative analysis. The test results of the four weighting methods in the UNSW-NB15 dataset are shown in Fig. 5.

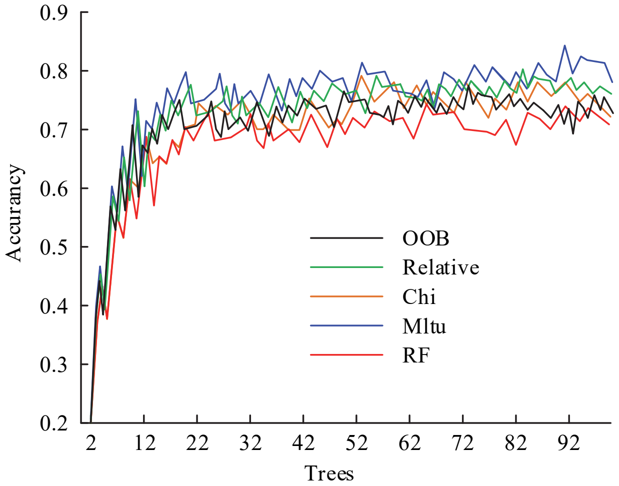


Figure 5 Accuracy results in the UNSW-NB15 dataset

In Fig. 5, the detection accuracy of RF in the dataset is about 0.63 as a whole, and when the number of decision trees is small, its detection accuracy is low. As the number of decision trees increased, its accuracy increased along with it. When the number of decision trees was about 20, its detection accuracy was gradually stabilized, but it still had a little fluctuation effect. Compared with the improved method of weight calculation, the accuracy and stability of the improved method were improved, among which the correlation coefficient weight calculation method used in the study was the most excellent. Its detection accuracy was around 0.68 overall, and the algorithm detection accuracy was gradually stabilized when the decision tree was around 16. The detection results in the CICIDS2017 dataset are shown in Fig. 6.

In Fig. 6, the detection accuracy of RF in the CICIDS2017 dataset is around 0.68 as a whole, and when the number of decision trees is around 18, its detection accuracy no longer shows obvious changes, but still has certain fluctuations. Compared with the improved method of weight calculation, the accuracy and stability of the improved method were also improved in the CICIDS2017 dataset, in which the detection accuracy of the correlation coefficient weight calculation method used in the study was about 0.73, and its detection accuracy was gradually stabilized when the number of decision trees was about 12. By analyzing the results in Fig. 5 and 6, it is shown that the method of calculating the weights of decision trees in the data prediction process of the random forest model by evaluating the magnitude of the classification ability of a single decision tree was reasonable, and that the overall generalization ability of the random forest model was improved by introducing voting weights. The study employed 100 decision trees in the random forest for the prediction of different feature spaces, and the detection results in the UNSW-NB15 dataset are shown in Fig. 7.

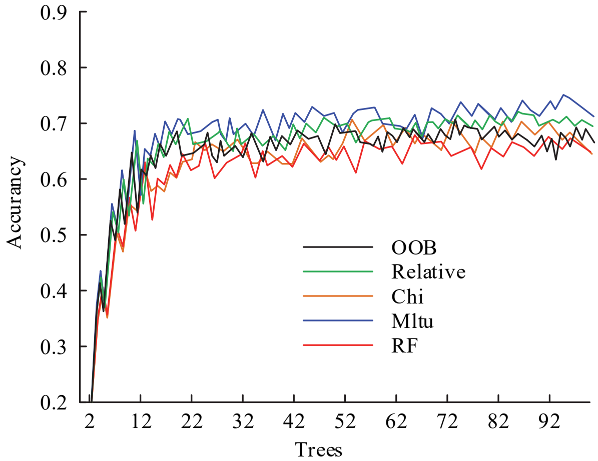


Figure 6 Accuracy results in the CICIDS2017 dataset

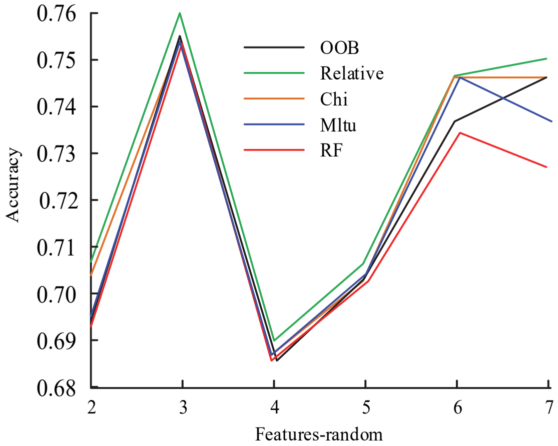


Figure 7 Detection results of different feature spaces in the UNSW-NB15 dataset

In Fig. 7, when the number of feature spaces is 2, the detection accuracy of the research-use method is 0.78. When the number of feature spaces is 3, its detection accuracy reaches 0.76. When the number of feature spaces is 4, its detection accuracy shows a significant decrease and is only 0.69. When the features continue to increase, the detection accuracy of the request increases along with it. The rest of the methods had the same trend as the research use method, but the detection accuracy at all stages was not as good as the research use method. Fig. 7 Results from the feature space dimension, as the feature space increased, the higher the dimension, the more data volume was required as the space becomes sparser and a small training set could not adequately cover the feature space. This in turn led to a situation where there was a sudden drop in accuracy, and subsequently after a certain number of features, the model may adapt to the increased features and show improved performance. The detection results in the CICIDS2017 dataset are shown in Fig. 8.

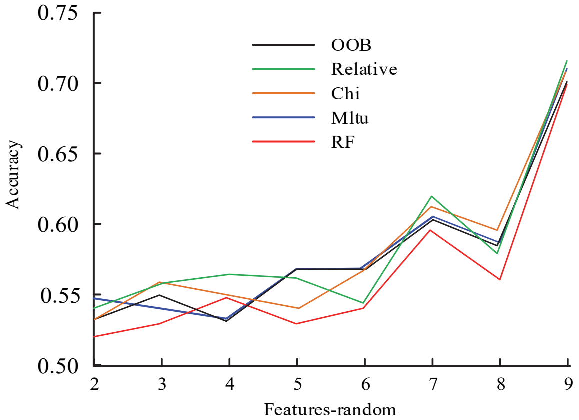


Figure 8 Detection results of different feature spaces in the CICIDS2017 dataset

In Fig. 8, the feature detection accuracy increases with the number of feature spaces, but the detection accuracy of the research-use method dips when the number of feature spaces is 6, at which point the accuracy is 0.56, while the accuracy of the other methods shows a steady increase. The number of feature spaces continues to grow, and the detection accuracy of the research use method returns to its highest value and is 0.73 at a feature space number of 9.

4.2 Characterization of Automatic Network Attack Detection

The study also proposed a semi-voting model to improve the random forest model, which aimed to improve the detection speed while maintaining the detection accuracy. The study used RF, RF-50%, Relative-RF, and Relative-RF-50% algorithms for comparative analysis and the results are shown in Tab 1.

In the results of Tab. 1, the prediction accuracies of RF and RF-50% are the same in both datasets, and the prediction accuracies of Relative-RF and Relative-RF-50% are the same. However, the prediction time of RF-50% is shorter in comparison to RF, and the prediction time of Relative-RF-50% is shorter in comparison to Relative-RF. The results indicated that the model using the semi-voting method had better prediction results. The study examined the correlation between features and network attacks using Relative-RF-50% method and the results are shown in Fig. 9.

Table 1 Results of different methods run on the dataset

UNSW-NB15 dataset			
Method	Prediction accuracy	Prediction total time / ms	Prediction data volume
RF	0.52946	696123.32	1024
RF-50%	0.52946	614625.01	1024
Relative-RF	0.63658	698043.39	1024
Relative-RF-50%	0.63658	452071.67	1024
CICIDS2017 dataset			
Method	Prediction accuracy	Prediction total time / ms	Prediction data volume
RF	0.90778	59662.49	844
RF-50%	0.90778	52307.11	844
Relative-RF	0.93844	59464.55	844
Relative-RF-50%	0.93844	50216.47	844

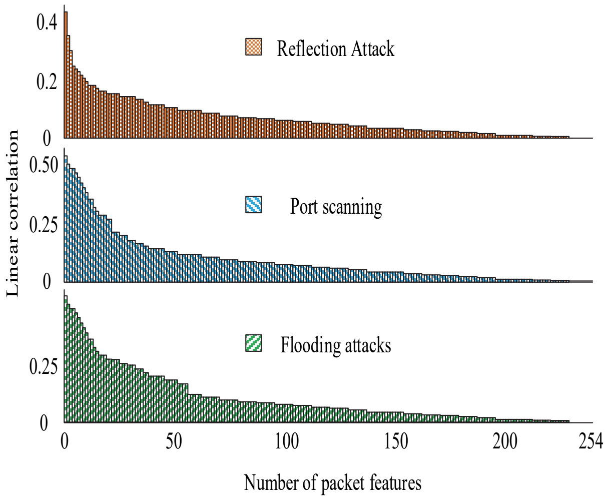


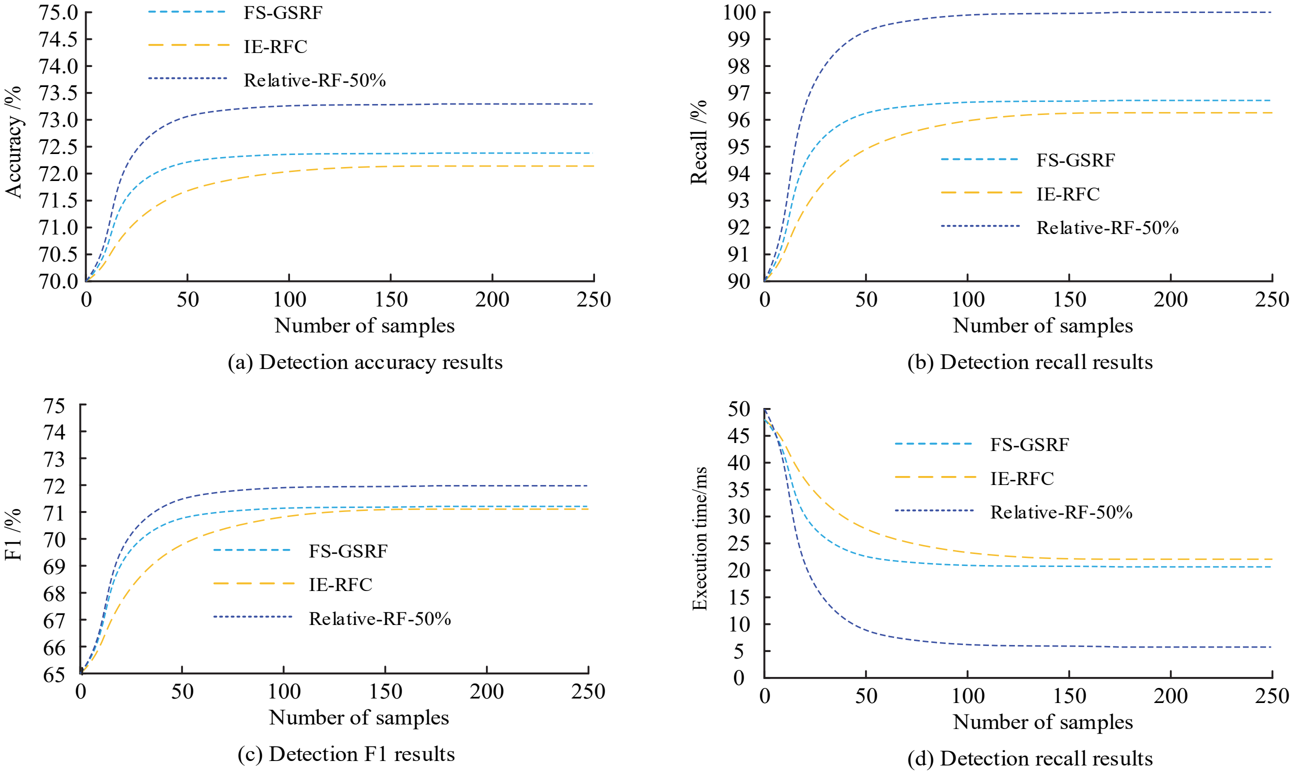
Figure 9 Linear correlation analysis between network characteristics and attack

From the results in Fig. 9, the research uses the method for flooding attack, port scanning, and reflection attack are 51,29,11 features, respectively, to realize the initial feature dimensionality reduction. Then, a subset of features that were weakly correlated with each other but highly correlated with the target were selected as features in order to remove some more redundant features, and the number of features after selection was 19, 19, and 13, respectively. The results showed that the elative-RF-50% method effectively reduced the number of features, which was conducive to the simplification of the model, reduced the computational resource requirements, and may improve the generalization ability. To further examine the difference between the performance of the algorithm in offline and online states, the comparison results are shown in Tab. 2.

Tab. 2 shows that the performance of inspection precision, recall, F1 score and execution time in the offline state is better than the performance of online inspection in both datasets. And there is a significant improvement in inspection precision and execution time. The longer execution time of online inspection may be due to the need to process data in real time and respond with limited resources. Although offline inspection provided better performance metrics, online inspection was necessary for real-time network attack detection.

Table 2 Performance comparison between offline and online states

Performance index RF	UNSW-NB15 dataset		CICIDS2017 dataset	
	Online detection	Offline detection	Online detection	Offline detection
Accuracy / %	73.3	77.2	63.2	72.4
Recall rate / %	69.7	65.8	66.9	71.5
F1 score / %	71.5	71.1	64.9	71.9
Execution time / ms	23.1	6.4	32.2	11.4



(a) Detection accuracy results

(b) Detection recall results

(c) Detection F1 results

(d) Detection recall results

Figure 10 Detection performance of different algorithms

The research compared feature selection and grid search random forest algorithm (FS-GSRF) proposed in reference [10], information entropy and random forest classification (IE-RFC) proposed in reference [11] to verify the progressiveness of the research algorithm. The results are shown in Fig. 10.

The results in Fig. 10 show that the detection accuracy of the research method is 1.02% and 1.24% higher than FS-GSRF and IE-RFC, respectively. The recall of the research method is 0.88% and 0.86% higher than FS-GSRF and IE-RFC, respectively. The F1 score of the research method is 1.05% and 0.99% higher than FS-GSRF and IE-RFC, respectively. The execution time of the research method is less than 15.4 ms and 16.8 ms compared to FS-GSRF and IE-RFC, respectively. and 0.99%. The execution time of the research method is lower than 15.4 ms and 16.8 ms compared to FS-GSRF and IE-RFC, respectively. These results overall showed that the algorithms used in the research were more effective than previous methods in detecting cyber-attacks, demonstrating higher precision and recall, as well as better execution efficiency.

5 CONCLUSION

For the problem of improving the performance of automated network attack detection technology, the study adopts the weight calculation method to improve the voting rule of random forest, and in this way analyzes the detection accuracy and the weights of the decision tree in the process of model data prediction. In addition, a semi-voting mechanism is introduced to further improve the efficiency of the algorithm operation. In the experiment. The improved random forest model using the correlation coefficient as weights exhibited a high detection accuracy of 0.68 on the UNSW-NB15 dataset and 0.73 on the CICIDS2017 dataset. The prediction accuracy of the Relative-RF-50% model with the half-voting mechanism was 0.63658 on the UNSW-NB15 dataset, and the prediction accuracy of the model with the half-voting mechanism was 0.63658 on the CICIDS2017 dataset was 0.93844. In the state-of-the-art analysis, the execution time of the research method was reduced by about 15-17 ms compared to the reference [10] and the reference [11]. The improved Random Forest model is competitive in automated detection of cyber-attacks, and especially excels in precision, recall and efficiency. The semi-voting mechanism provides a significant improvement in the model's ability to handle real-time network traffic, which helps to identify cyber-attacks quickly and efficiently in real-time scenarios. Despite the positive results of the study, the generalization ability and robustness of the model need to be further optimized in the case of performance fluctuations due to the increase in feature space. Future research can focus on deep learning or more sophisticated integrated learning techniques to deal with high-dimensional feature spaces and explore generalized models for various network environments and attack types.

Acknowledgements

Sichuan Provincial Department of Science and Technology 2022 Soft Science Research Project (2022JDR0039) - Modernization of Big Data-Driven Vocational Education Governance System; Sichuan Provincial Department of Education 2022-2024 Vocational Education Talent Cultivation and Educational Reform Project (GZJG2022-644) - Practice and Research on the Construction of Internal Quality Assurance System in Vocational Colleges Based on School-Based Big Data.

6 REFERENCES

[1] Han, F. X., Liu, S. Y., Zhang, T. Z., F. Xin Lü, & Li, F. Y. (2021). Sparse auto-encoder combined with kernel for network attack detection. *Computer Communications*, *173*(1), 14-20. https://doi.org/10.1016/j.comcom.2021.03.004

[2] Wu, Y., Ru, Y., Shi, Z., Xu, J., Liu, J., Zhang, F., Ni, M., & Li, M. (2022). A cyber - attack detection method for load control system based on cyber and physical layer crosscheck mechanism. *IET Generation, Transmission & Distribution*, *16*(14), 2805-2815. https://doi.org/10.1049/gtd2.12338

[3] Yamamoto, Y., Nakano, S., & Shigeta, Y. (2022). Dynamical Interaction Analysis of Proteins by a Random Forest-Fragment Molecular Orbital (RF-FMO) Method and Application to Src Tyrosine Kinase. *Bulletin of the Chemical Society of Japan*, *96*(1), 42-47. https://doi.org/10.1246/bcsj.20220304

[4] Elamin, O. A. (2023). The causal effect of informal job search on wage and job satisfaction: Evidence from Egypt and Jordan using random forest method. *International Journal of Social Economics*, *50*(4), 522-536. https://doi.org/10.1108/IJSE-05-2022-0318

[5] Gu, T., Han, Y., & Duan, R. (2023). A transfer learning approach based on random forest with application to breast cancer prediction in underrepresented populations. Pacific Symposium on Biocomputing. *Pacific Symposium on Biocomputing*, *28*(3), 186-197. https://doi.org/10.1142/9789811270611_0018

[6] Blumhagen, R. Z., Schwartz, D. A., Langefeld, C. D., & Fingerlin, T. E. (2023). Identification of influential rare variants in aggregate testing using random forest importance measures. *Annals of Human Genetics*, *87*(4), 184-195. https://doi.org/10.1111/ahg.12509

[7] Garai, S., Paul, R. K., Kumar, M., & Choudhury, A. (2023). Intra-Annual National Statistical Accounts Based on Machine Learning Algorithm. *Journal of Data Science and Intelligent Systems*, *2*(2), 12-15. https://doi.org/10.47852/bonviewJDSIS3202870

[8] Mishra, A. K., & Paliwal, S. (2022). Mitigating cyber threats through integration of feature selection and stacking ensemble learning: The LGBM and random forest intrusion detection perspective. *Cluster Computing*, *26*(4), 2339-2350. https://doi.org/10.1007/s10586-022-03735-8

[9] Hussein, A. Y., Falcarin, P., & Sadiq, A. T. (2021). Enhancement performance of random forest algorithm via one hot encoding for IoT IDS. *International University of Sarajevo*, *9*(3), 579-591. https://doi.org/10.21533/PEN.V9I3.2204

[10] Subbiah, S., Anbananthen, K. S. M., Thangaraj, S., Kannan, S., & Chelliah, D. (2022). Intrusion detection technique in wireless sensor network using grid search random forest with Boruta feature selection algorithm. *Journal of Communications and Networks*, *24*(2), 264-273. https://doi.org/10.23919/JCN.2022.000002

[11] Niandong, L., Yanqi, S., Sheng, S., Xianshen, H., & Haoliang, M. (2020). Detection of probe flow anomalies using information entropy and random forest method. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, *39*(1), 433-447. https://doi.org/10.3233/JIFS-191448

[12] Srivani, P., Vandana, C., Vinusha, P., Sathvika, S., & Manimanognya, A. (2023). Random Forest Classifier for the Detection of IoT Botnet Attacks from Data of Provision PT 37E Security Camera. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *14*(3), 290-301.

[13] Anwer, M., Farooq, M. U., Khan, S. M., & Waseemullah, W. (2021). Attack Detection in IoT using Machine Learning. Engineering, *Technology and Applied Science Research*, *11*(3), 7273-7278. https://doi.org/10.48084/etasr.4202

[14] Gaur, V. & Kumar, R. (2022). Analysis of Machine Learning Classifiers for Early Detection of DDoS Attacks on IoT Devices. *Arabian Journal for Science and Engineering*, *47*(2), 1353-1374. https://doi.org/10.1007/s13369-021-05947-3

[15] Kayode-Ajala, O. (2021). Anomaly Detection in Network Intrusion Detection Systems Using Machine Learning and Dimensionality Reduction. *Sage Science Review of Applied Machine Learning*, *4*(1), 12-26.

[16] Gheisari, M., Hamidpour, H., Liu, Y., Saedi, P., Raza, A., Jalili, A., Rokhsati, H., & Amin, R. (2022). Data Mining Techniques for Web Mining: A Survey. *Artificial Intelligence and Applications*, *1*(1), 3-10. https://doi.org/10.47852/bonviewAIA2202290

[17] Wu, Z. & Wu, Z. (2023). Urban garden spatial environment layout method based on random forest. *International Journal of Environmental Technology and Management*, *26*(3/4/5), 263-274. https://doi.org/10.1504/IJETM.2023.130798

[18] Jordi, V. C., Winterauer, D. J., Niels, K. L., Sascha, R., Fan, L., Lüttjohann Stephan, Roland, H., & Jes, V. (2023). Random forest microplastic classification using spectral subsamples of FT-IR hyperspectral images. *Analytical Methods*, 15(18), 2226-2233. https://doi.org/10.1039/d3ay00514c

[19] Chen, J., Zhu, W., & Yu, Q. (2021). Estimating half-hourly solar radiation over the Continental United States using GOES-16 data with iterative random forest. *Renewable Energy*, *178*(34), 916-929.

[20] Riss, G., Romano, M., Memon, F. A., & Kapelan, Z. (2021). Detection of water quality failure events at treatment works using a hybrid two-stage method with CUSUM and random forest algorithms. *Water Supply*, *21*(6), 3011-3026. https://doi.org/10.2166/ws.2021.062

Contact information:

Ke XIANG
(Corresponding author)
Sichuan Post and Telecommunication College,
Chengdu, Sichuan, 610000, China
E-mail: xiangke@sptc.edu.cn

Xing YANG
Geely University of China,
Chengdu, Sichuan, 610000, China
E-mail: rryangxing@163.com