Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2022.030302
ISSN 2790-0983 Vol. 3, Issue 3: 10-17

Intrusion Detection Classification Based on Random Forest

Baiming Liu*

Beijing Potential Big Data Research Institute (PRI), Beijing 10095, China

Lait_liu@pri.ac.cn

**corresponding author*

Keywords: Random Forest, Intrusion Detection, Detection Classification, K-means++ Clustering

Abstract: In recent years, network security has become increasingly severe and network intrusions are frequent. On the one hand, it is difficult to capture intrusions because of the increasingly large volume of computer networks and complex network topology, and on the other hand, it is more difficult to capture intrusions because the characteristics of intrusions are becoming more diverse and complex. The purpose of this paper is to study intrusion detection classification based on random forest. After constructing a high-precision decision tree, the double failure metric is used as the distance between two decision trees, and the k-means++ algorithm is used to select high-precision decision trees with a certain degree of independence to form the final random forest. The improved random forest algorithm for setting up clustering and the random forest algorithm were experimented as algorithm comparisons. The KDD-NSL dataset was selected and the experimental results were compared to demonstrate the effectiveness of the random forest algorithm based on gradient boosting. The experimental results show that the improved random forest model has a significant reduction in decision tree size and a significant increase in diversity due to the k-means++ algorithm for clustering and extracting cluster centers.

1. Introduction

The increasingly complex security conditions and diverse network intrusion methods in cyberspace pose greater challenges to the deployment of cyberspace security infrastructure [1-2]. Therefore, it is crucial to investigate more effective intrusion detection techniques. Intrusion detection is a proactive defense technique that detects ongoing and existing malicious attacks in a timely manner and has always been an access point for cybersecurity research. However, as

network attacks become more sophisticated, traditional techniques such as user authentication, access control, data encryption, signatures, and firewalls are becoming increasingly difficult to handle complex network attacks [3-4]. Among them, identifying various cyber attacks is an important and unavoidable technical challenge. In addition, continuous network communication brings a large amount of data problems, which brings new challenges to intrusion detection, and establishing a stable, reliable and efficient intrusion detection model to improve network security has broad application prospects [5].

Intrusion detection, as the second security gate after firewall, is still a key research in network security maintenance [6].V. Sandeep studied bayesian decision tree aggregation in integration. The focus is on multi category classification, and the sample size is obviously biased towards one category. The algorithm uses ready-made datasets to estimate the prediction errors of a single tree, and then uses these errors according to Bayesian rules to optimize the comprehensive decision. The goal is to improve the detection capability of operating malware detection systems. Although we can keep the accuracy of the system above 94%, which means that only 6 of the 100 detections displayed to the network administrator are error messages, we can increase the number of detections by about 7% [7]. Dzelila Mehanovic constructed a random forest classifier based on features extracted from the token smart contract code of the DeFi project. The final classifier obtained an F1 score of 98.6%. Through further feature-level analysis, they found that individual features make this a highly detectable problem [8]. Therefore, it is of pivotal importance to work on system intrusion detection to protect network security.

In this paper, we study intrusion detection techniques based on random forest algorithm in machine learning, firstly, we analyze the intrusion detection framework based on machine learning, design a simple and easy to implement, low overhead clustering improved random forest algorithm, and finally, for the objective situation of unbalanced network intrusion categories, we propose detection experiments for all types of intrusion data, improve the intrusion detection classification, and improve the fine-grained detection of intrusion categories accuracy.

2. A Study of Intrusion Detection Classification Based on Random Forest

2.1. Intrusion Detection Framework Based on Machine Learning

Machine learning enables classification and detection of unknown network traffic data samples by training classification models from the obtained network traffic data and constructing a knowledge base to achieve intelligent discrimination of intrusion categories. The intrusion detection system framework based on machine learning consists of four main parts: data acquisition, data preprocessing, model construction, and decision analysis [8].

(1) Data acquisition, from the monitored network, to obtain network traffic data.

(2) Data preprocessing, which processes the already acquired data.

(3) Model learning, where specific machine learning algorithms are applied in this phase. The model performance is then evaluated by training classifiers on the preprocessed data and implementing evaluation metrics on the classification results [9-10].

(4) Decision discrimination, in this stage, the intrusion detection system discriminates the decision on the detection results. The trained one is deployed on the test dataset to be detected or deployed in the network environment to discriminate the intrusion detection results, identify the intrusion events, and respond [11-12].

2.2. Random Forest

The essence of the RF model is an improvement of the decision tree algorithm, which combines

multiple decision trees, each based on individually collected samples. The distribution of each tree in the forest is the same [13]. The generalization error depends on the classification and mapping ability of each tree in the forest. The selection operation randomly divides each node and compares the errors that occur in different cases. The inherent estimation errors, classification and detectable dependencies determine the number of selected features [14].

2.3. Problems of the Random Forest Algorithm

Random forest itself has many advantages and features, but there are some problems within a specific application scenario:

(1) Since bagging is located in the original training set, there is independence between different rounds of training sets within the original set, which reduces the classification accuracy, although it can reduce the adjustment parameters.

(2) The classification result of random forest depends on the voting of multiple meta-classifiers, and the classification accuracy of the algorithm is mainly determined by two aspects, one is the correlation of any two meta-classifiers, the greater the correlation, the lower the classification accuracy; the second is the lack of processing ability for continuity fields, and the classification accuracy is also limited [15].

2.4. Improved Random Forest Algorithm for Clustering

The double failure metric mentioned above is used as the distance criterion between two trees to group similar decision trees into a cluster, and the sample centers of each cluster are selected to form a new high-precision low-similarity random forest [16-17].

The specific procedure of the algorithm is as follows:

(1) Randomly select one data in the dataset N (each sample in the dataset is a high-precision decision tree generated in the previous stage) as a cluster center;

(2) Calculate the shortest distance between each data x in the dataset and the existing clustering center, and the distance between two data is expressed as the reciprocal of the double failure measure, which is denoted as D(x), and calculate the probability of its selection according to the value of D(x) min, and then use the roulette wheel method to select the next clustering center. The formula for calculating the probability of data x being selected is shown in equation (1):

$$p = \frac{D(x)_{\min}^2}{\sum_{x \in N} D(x)_{\min}^2} \tag{1}$$

(3) Repeat the second step to select K clustering centers;

(4) Calculate the distance from the data di to all cluster centers and divide di to the cluster center with the closest distance to form a cluster;

(5) Calculate the average distance of all sample points to other points within the cluster, and use the point with the smallest average distance as the new cluster center of the cluster;

(6) Repeat the above two steps until all (or most) of the cluster centers are not updated or all (or most) of the samples are not re-clustered;

(7) Change the number of clustering centers K and repeat the above steps to calculate the contour coefficient after each clustering. The contour coefficient can measure the effect of clustering and is calculated as shown in equation (2):

$$SC = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{\max(a_i, b_i)} \quad (2)$$

Where n denotes all data in the dataset, and ai and bi denote the average distance of sample i from other samples in its same and neighboring clusters, respectively. The profile coefficients take values between -1 and 1, with -1 indicating the worst effect of cluster class classification and 1 indicating the best effect [18].

(8) The cluster centers of each cluster, or the decision trees near the cluster centers, are selected to form a random forest.

3. A Survey and Study of Random Forest Based Intrusion Detection Classification

3.1. Dataset

The KDD-NSL dataset is an improved dataset of the classical intrusion detection dataset KDD99, which solves the problem of having more identical data samples and makes a more reasonable dataset sampling, and has become a standard dataset in the field of intrusion detection. The details of the dataset are shown in Table 1:

Table 1. KDD-NSL data set

	Training set	Test set
Normal	16052	8143
DOS	2481	6770
Probing	6200	3281
R2L	2815	4460
U2R	8445	3864

3.2. Test Evaluation Metrics

Recall (R) is another metric that characterizes the performance of an intrusion detection model. It indicates the proportion of samples correctly predicted as positive by the classifier among all the actual positive class samples, i.e., how many are detected compared to the actual entire positive class. Recall is defined as follows:

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (3)$$

The correctness (P) is the proportion of all positive class samples predicted by the classifier that are actually positive, i.e., how many of the predicted results are correct, and the correctness directly affects the performance. Accuracy is defined as follows:

$$\mathrm{Pr}ecision = \frac{TP}{TP + FN} \quad (4)$$

TP (true case) is the actual number of attack samples classified as attack samples by the classifier; FP (false positive case) is the number of samples that are actually attack samples but are classified as normal samples by the classifier.

4. Analysis and Research of Random Forest Based Intrusion Detection Classification

We obtained the comparison results of correctness, accuracy and recall obtained for traditional RF and the improved random forest algorithm of clustering in this paper under five different network attacks.

The experimental results of intrusion detection classification for Normal class are shown in Table 2.

Table 2. Performance comparison of two algorithms on Normal data

Algorithm	Precision /%	Accuracy /%	Recall /%
Traditional rf	66	70	68
This paper improves the algorithm	85	88	89

The experimental results of intrusion detection classification for the DOS class are shown in Table 3.

Table 3. Performance comparison of the two algorithms on DOS data

Algorithm	Precision /%	Accuracy /%	Recall /%
Traditional rf	78	70	77
This paper improves the algorithm	80	81	80

The experimental results of intrusion detection classification for the Probing class are shown in Table 4.

Table 4. Performance comparison of the two algorithms on Probing data

Algorithm	Precision /%	Accuracy /%	Recall /%
Traditional rf	79	77	78
This paper improves the algorithm	81	85	82

The experimental results of intrusion detection classification of R2L class are shown in Figure 1.

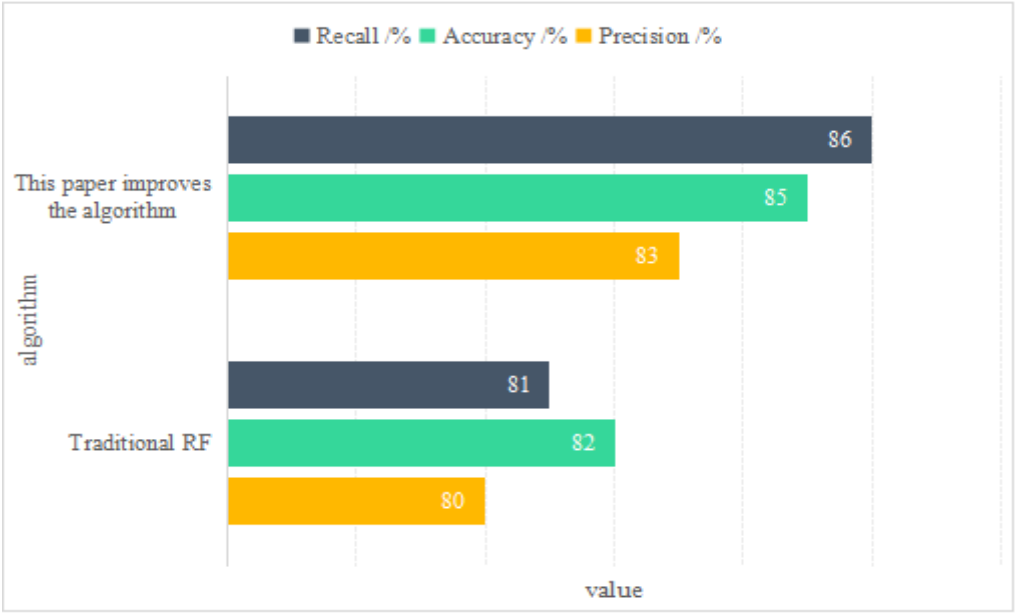

Figure 1. Performance comparison of the two algorithms on R2L data

The experimental results of intrusion detection classification of U2R class are shown in Figure 2.

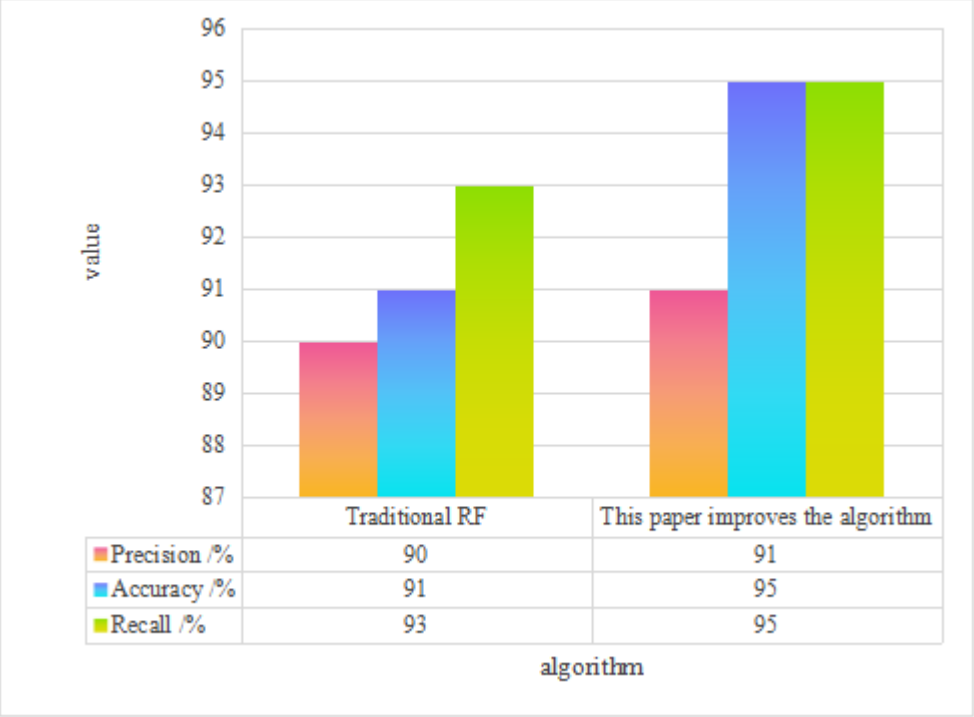

Figure 2. Performance comparison of two algorithms on U2R data

It can be seen that the improved random forest algorithm for clustering in this paper outperforms the traditional RF under each metric. In Normal class data, the improvement of the improved random forest algorithm is especially obvious in each index, the correct rate is improved from 66% to 85%, the accuracy rate is improved from 70% to 88%, and the recall rate is also improved from 68% to 89%; the data volume in DOS data is small, and the improved random forest algorithm of clustering has a certain improvement on its detection rate, in which the recall rate is improved by Probing data and R2L data are similar, the feature dimension is not high and the distribution is average, so the algorithm has limited improvement on its characteristics; U2R data because its data itself has undergone some processing, the data is more balanced so the original RF on its classification efficiency is very high, after using the improved algorithm the model correct rate and accuracy rate has a small increase, the recall rate is more obvious, which indicates that the algorithm on The classification rate on the inverse class has been improved and the model can better identify harmful attacks.

5. Conclusion

Intrusion, as the main behavior that undermines network security, is an important problem of today's Internet. In this paper, we analyze and study the existing intrusion detection models, integrate a variety of learning strategies, and propose a new and improved intrusion detection algorithm model based on the random forest algorithm. Due to my limited level, some technology, experimental environment and equipment, etc. are not enough, it is inevitable that there are certain problems in the research, which need to be further improved and strengthened in the process of future research. The research content of this paper mainly has the following problems: due to the rapid development of today's network technology, the network environment is increasingly complex, many industries have their own characteristics of the special network, such as the intelligent

network of the power system, the current overall analysis method may have been inappropriate for the network environment, how to conduct independent research on this part of the network.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Ankit Rajeshkumar Kharwar, Devendra V. Thakor: An Ensemble Approach for Feature Selection and Classification in Intrusion Detection Using Extra-Tree Algorithm. Int. J. Inf. Secur. Priv. 16(1): 1-21 (2022) https://doi.org/10.4018/IJISP.2022010113

[2] Kapil Kumar, Arvind Kumar, Vimal Kumar, Sunil Kumar: A Hybrid Classification Technique for Enhancing the Effectiveness of Intrusion Detection Systems Using Machine Learning. Int. J. Organ. Collect. Intell. 12(1): 1-18 (2022) https://doi.org/10.4018/IJOCI.2022010102

[3] P. Manjula, S. Baghavathi Priya: An effective network intrusion detection and classification system for securing WSN using VGG-19 and hybrid deep neural network techniques. J. Intell. Fuzzy Syst. 43(5): 6419-6432 (2022) https://doi.org/10.3233/JIFS-220444

[4] Padideh Choobdar, Marjan Naderan, Mahmood Naderan: Detection and Multi-Class Classification of Intrusion in Software Defined Networks Using Stacked Auto-Encoders and CICIDS2017 Dataset. Wirel. Pers. Commun. 123(1): 437-471 (2022) https://doi.org/10.1007/s11277-021-09139-y

[5] Sivamohan Krishnaveni, Sivanandam Sivamohan, S. S. Sridhar, S. Prabakaran: Efficient feature selection and classification through ensemble method for network intrusion detection on cloud computing. Clust. Comput. 24(3): 1761-1779 (2021) https://doi.org/10.1007/s10586-020-03222-y

[6] Zahra Asghari Varzaneh, Marjan Kuchaki Rafsanjani: Intrusion detection system using a new fuzzy rule-based classification system based on genetic algorithm. Intell. Decis. Technol. 15(2): 231-237 (2021) https://doi.org/10.3233/IDT-200036

[7] V. Sandeep, Saravanan Kondappan, A. Amir Anton Jone, S. Raj Barath: Anomaly Intrusion Detection Using SVM and C4.5 Classification with an Improved Particle Swarm Optimization (I-PSO). Int. J. Inf. Secur. Priv. 15(2): 113-130 (2021) https://doi.org/10.4018/IJISP.2021040106

[8] Dzelila Mehanovic, Dino Keco, Jasmin Kevric, Samed Jukic, Adnan Miljkovic, Zerina Masetic: Feature selection using cloud-based parallel genetic algorithm for intrusion detection data classification. Neural Comput. Appl. 33(18): 11861-11873 (2021) https://doi.org/10.1007/s00521-021-05871-5

[9] Maryam Yousefnezhad, Javad Hamidzadeh, Mohammad Aliannejadi: Ensemble classification for intrusion detection via feature extraction based on deep Learning. Soft Comput. 25(20): 12667-12683 (2021) https://doi.org/10.1007/s00500-021-06067-8

[10] Allen Yang, Boxiang Dong, Dawei Li, Weifeng Sun, Bharath K. Samanthula: DeepICU: imbalanced classification by using deep neural networks for network intrusion detection. Int. J. Big Data Intell. 7(3): 137-147 (2020) https://doi.org/10.1504/IJBDI.2020.10031966

[11] Imad Bouteraa, Makhlouf Derdour, Ahmed Ahmim: Intrusion detection using classification techniques: a comparative study. Int. J. Data Min. Model. Manag. 12(1): 65-86 (2020) https://doi.org/10.1504/IJDMMM.2020.105596

[12] Shen Kejia, Hamid Parvin, Sultan Noman Qasem, Bui Anh Tuan, Kim-Hung Pho: A classification model based on svm and fuzzy rough set for network intrusion detection. J. Intell. Fuzzy Syst. 39(5): 6801-6817 (2020) https://doi.org/10.3233/JIFS-191621

[13] Preethi Devan, Neelu Khare: An efficient XGBoost-DNN-based classification model for network intrusion detection system. Neural Comput. Appl. 32(16): 12499-12514 (2020) https://doi.org/10.1007/s00521-020-04708-x

[14] Julio Lamas Piñeiro, Lenis Wong Portillo: Web architecture for URL-based phishing detection based on Random Forest, Classification Trees, and Support Vector Machine. Inteligencia Artif. 25(69): 107-121 (2022) https://doi.org/10.4114/intartif.vol25iss69pp107-121

[15] María Guadalupe Bedolla-Ibarra, María del Cármen Cabrera-Hernández, Marco Antonio Aceves-Fernández, Saúl Tovar-Arriaga: Classification of attention levels using a Random Forest algorithm optimized with Particle Swarm Optimization. Evol. Syst. 13(5): 687-702 (2022) https://doi.org/10.1007/s12530-022-09444-2

[16] Tiebo Sun, Jinhao Liu, Jiangming Kan, Tingting Sui: Research on target classification method for dense matching point cloud based on improved random forest algorithm. Int. J. Inf. Commun. Technol. 21(3): 290-303 (2022) https://doi.org/10.1504/IJICT.2022.125541

[17] Nasrin Amini, Ahmad Shalbaf: Automatic classification of severity of COVID-19 patients using texture feature and random forest based on computed tomography images. Int. J. Imaging Syst. Technol. 32(1): 102-110 (2022) https://doi.org/10.1002/ima.22679

[18] C. Venkata Narasimhulu: An automatic feature selection and classification framework for analyzing ultrasound kidney images using dragonfly algorithm and random forest classifier. IET Image Process. 15(9): 2080-2096 (2021) https://doi.org/10.1049/ipr2.12179