Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Cyber-Attack Detection in a Network

using Logistic Regression

Syed Mohammed Kareemuddin 1, Bilal Ahmed Siddique 2, Syed Fahad Quadri 3,

Mrs Heena Yasmin 4

1, BE Student – Dept. Of CSE, ISL Engineering College, TS, India

2, BE Student – Dept. Of CSE, ISL Engineering College, TS, India

3, BE Student – Dept. Of CSE, ISL Engineering College, TS, India

4, Assistant Professor – Dept. Of CSE, ISL Engineering College, TS, India

Article Info
Page Number: 1455-1461
Publication Issue:
Vol. 72 No. 1 (2023)

Article History
Article Received: 15 October 2022
Revised: 24 November 2022
Accepted: 18 December 2022

ABSTRACT – Cyber-attack detection can identify unknown attacks from network traffics and has been an effective means of network security. Nowadays, existing methods for network anomaly detection are usually based on traditional machine learning models, such as KNN, RF, etc. Although these methods can obtain some outstanding features, they get a relatively low accuracy and rely heavily on manual design of traffic features. To solve the problems of low accuracy and feature engineering in intrusion detection, a traffic anomaly detection model is proposed. The cyber-attack detection model uses Logistic Regression classifier and Multi-Layer Perceptron algorithms for better efficiency in prediction.

Keywords: Cyber-attack, network anomaly detection, Logistic Regression, Multi-Layer Perceptron.

1. INTRODUCTION

With the development and improvement of Internet technology, the Internet is providing various convenient services for people. However, we are also facing various security threats. Network viruses, eavesdropping and malicious attacks are on the rise, causing network security to become the focus of attention of the society and government departments. Fortunately, these problems can be well solved via intrusion detection. Intrusion detection plays an important part in ensuring network information security. However, with the explosive growth of Internet business, traffic types in the network are increasing day by day, and network behaviour characteristics are becoming increasingly complex, which brings great challenges to intrusion detection [1], [2]. How to identify various malicious network traffics, especially unexpected malicious network traffics, is a key problem that cannot be avoided. In fact, network traffic can be divided into two categories (normal traffics and malicious traffics). Furthermore, network traffic can also be divided into five categories: Normal, DoS (Denial of Service attacks), R2L (Root to Local attacks), U2R (User to Root attack) and Probe (Probing attacks). Hence, intrusion detection can be considered as a classification problem. By improving the performance of classifiers in effectively identifying malicious traffics, intrusion detection accuracy can be largely improved. In [10], the authors provide an analysis of the viability of Recurrent Neural Networks (RNN) to detect the behaviour of network traffic by modelling it as a sequence of states that change over time. In [10], the authors verify the performance of

Long Short-term memory (LSTM) network in classifying intrusion traffics. Experimental results show that LSTM can learn all the attack classes hidden in the training data. All the above methods treat the entire network traffic as a whole consisting of a sequence of traffic bytes. They don't make full use of domain knowledge of network traffics. For example, CNN converts continuous network traffic into images for processing, which is equivalent to treating traffics as independent and ignore the internal relations of network traffics. Firstly, network traffic is a hierarchical structure. Specifically, network traffic is a traffic unit composed of multiple data packets. Data packet is a traffic unit composed of multiple bytes. Secondly, traffic features in the same and different packets are significantly different. Sequential features between different packets need to be extracted independently. In other words, not all traffic features are equally important for traffic classification in the process of extracting features on a certain network traffic.

2. LITERATURE REVIEW

In the research of network intrusion detection based on machine learning, scholars mainly distinguish normal network traffic from abnormal network traffic by dimensionality reduction, clustering, and classification, to realize the identity fiction of malicious attacks [9], [10]. Pervez proposed a new method for feature selection and classification merging of multi-class NSL-KDD Cup99 dataset using and discussed the classification accuracy of classifiers under different dimension features [10]. Shiraz studied some new technologies to improve CANN intrusion detection methods' classification performance and evaluated their performance on the NSL-KDD Cup99 dataset [3]. He used the K Farthest Neighbour (KFN) and the K Nearest Neighbour (KNN) to classify the data and used the Second Nearest Neighbour (SNN) of the data when the nearest and farthest neighbours have the same class label. The result shows the CANN detection rate and reduces the failure the alert rate is improved or provides the same performance. Bhattacharya proposed a machine learning model based on hybrid Principal Component Analysis (PCA)-Firefly [4]. The dataset used was the open dataset collected from Kaggle. Firstly, the model performs one key coding for transforming the IDS dataset, then uses the hybrid PCA-Firefly algorithm to reduce the dimension, and the algorithm classifies the reduced dataset. In recent years, with the powerful ability of automatic feature extraction, deep learning has made remarkable achievements in the fields of Computer Vision (CV), Autonomous driving (AD), Natural Language Processing (NLP). Many scholars apply deep learning to intrusion detection for traffic classification, which has become a hot spot of current research. The method of deep learning is to mine the potential characteristics of high-dimensional data through a training model and transform network traffic anomaly detection into classification problem [5]. Through a large number of sample data training, adaptive learning 12 between normal network traffic and abnormal network traffic effectively enhances real-time intrusion processing. Torres et al. [6] first converted network traffic characteristics into a series of characters and then used Recurrent Neural Network (RNN) to learn their temporal characteristics, which were further used to detect malicious network traffic. Wang et al. [7] proposed a malicious software traffic classification algorithm based on Convolutional Neural Network (CNN). By mapping the traffic characteristics to pixels, the network traffic image is generated, and the image is used as the input of the CNN to realize traffic

classification. Staudemeyer and Shamsinejad [9] proposed an intrusion detection algorithm based on Long Short-Term Memory (LSTM), which detects DoS attacks and probe attacks with unique time series in the KDD Cup99 dataset. Kwon et al. [5] has carried out relevant research on the deep learning model, focusing on data simplification, dimension reduction, classification, and other technologies, and proposes a Fully Convolutional Network (FCN) model. By comparing with the traditional machine learning technology, it is proved that the FCN model is useful for network traffic analysis. Tama et al. [9] proposed an anomaly-based IDS based on a two-stage meta-classifier, which uses a hybrid feature selection method to obtain accurate feature representations.

3. PROPOSED SYSTEM

We propose an end-to-end Machine Learning based cyber-attack detection method that is composed of Linear Regression and other Machine Learning Algorithms. Linear Regression can well solve the problem of intrusion detection and provide a new research method for intrusion detection.

We train model using linear regression model and predict packet is malicious or not based on user input.

We evaluate our proposed network with a real NSL-KDD dataset.

We compare accuracy of multiple machine learning models and find accuracy of each model. In proposed system along with Linear Regression MLP multi-layer perception method is used to train dataset.

4. SYSTEM ARCHITECTURE

This project will be built in Python, HTML and MySQL. We use the classic NSL-KDD and the up-to-date CSECIC-IDS2018 as benchmark datasets and conduct detailed analysis and data cleaning. (2) This work proposes a machine learning algorithm, reducing the majority samples and augmenting the minority samples in the difficult set, tackling the class imbalance problem in intrusion detection so

that the classifier learns the differences better in training. (3) The classification model uses Random Forest (RF), Linear Regression & NLP with other methods.

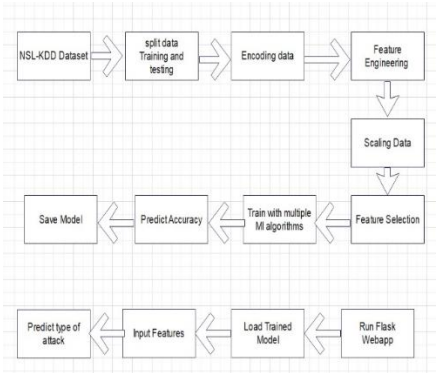


Fig 1: SYSTEM ARCHITECTURE

5. IMPLEMENTATION

5.1 Data Collection:

There are three symbolic data types in NSL-KDD data features: protocol type, flag and service. We use one-hot encoder mapping these features into binary vectors. One-Hot Processing: NSL-KDD dataset is processed by one-hot method to transform symbolic features into numerical features. For example, the second feature of the NSL-KDD data sample is protocol type. The protocol type has three values: tcp, udp, and icmp. One-hot method is processed into a binary code that can be recognized by a computer, where tcp is [1, 0, 0], udp is [0, 1, 0], and icmp is [0, 0, 1].

5.2 Train-Test Split and Model Fitting:

Now, we divide our dataset into training and testing data. Our objective for doing this split is to assess the performance of our model on unseen data and to determine how well our model has generalized on training data. This is followed by a model fitting which is an essential step in the model building process.

Model Evaluation and Predictions:

This is the final step, in which we assess how well our model has performed on testing data using certain scoring metrics, I have used 'accuracy score' to evaluate my model. First, we create a model instance, this is followed by fitting the training data on the model using a fit method and then we will use the predict method to make predictions and store it in a variable called test accuracy, a variable that will hold the testing accuracy of our model. We followed these steps for a variety of classification algorithm models and obtained corresponding test accuracy scores.

Logistic Regression (LR):

N. V. Chawla reported a study on forecasting the downtime of a printing machine based on real time predictions of imminent failures. In their study, they utilized unstructured historical machine data to train the ML classification algorithms including RF and LR in predicting the machine failures. Various metrics were analysed to determine the goodness of fit of the models. These metrics include empirical cross-entropy, area under the receiver operating characteristic curve (AUC), receiver operating characteristic curve itself (ROC), precision-recall curve (PRC), number of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) at various decision thresholds, and calibration curves of the estimated probabilities. Based on the results obtained, in terms of ROC, all the algorithms performed significantly better and almost similar. Logistic regression is used to estimate the categorical contingent variations. Graph of the linear regression model and logistics regression model are shown in Figure 9.

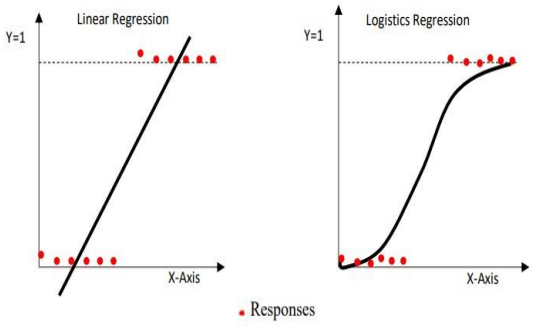


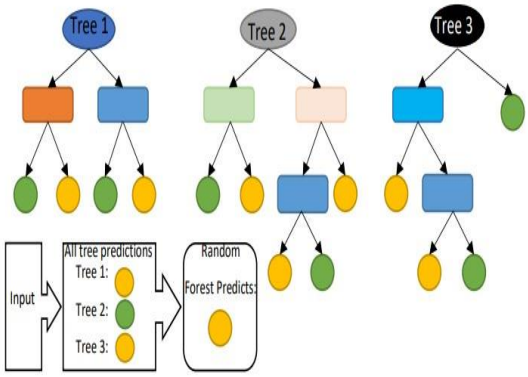
Figure 9. Logistic regression.

Random Forest (RF):

The Random Forest classification model is made up of several decision trees. In simple terms, it combines the results from numerous decision trees to reach a single result. The main difference between decision trees and random forests is that decision trees consider all the possible feature splits, however, random forests will only select a subset of those features.

RF was developed by Breiman, L. [60]. This is an ensemble learning algorithm made up of several DT classifiers, and the output category is determined collectively by these individual trees. When the number of trees in the forest increases, the fallacy in generalization error for forests converges. There are also important benefits of the RF. For example, it can manage high-dimensional data without choosing a feature; trees are independent of each other during the training process, and implementation is fairly simple; however, the training speed is generally fast and, at the same time, the generalization functionality is good enough [4].

Random forest algorithm for machine learning has tree predictions, and based on tree predictions, the RF provides random forest predictions [6]. The RF model is visualized in Figure.



Decision Tree (DT):

Decision Tree is a network system composed primarily of nodes and branches, and nodes comprising root nodes and intermediate nodes. The intermediate nodes are used to represent a feature, and the leaf nodes are used to represent a class label [52]. DT can be used for feature selection [57]. DT algorithm is presented in Figure

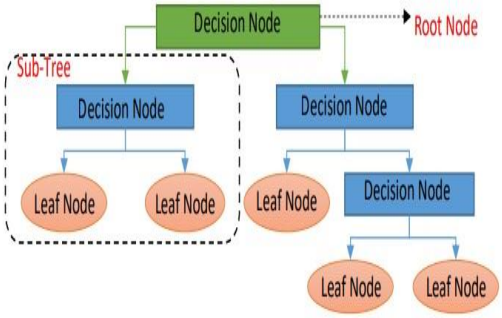


Figure 7. Decision tree algorithm, adapted from.

DT classifiers have gained considerable popularity in a number of areas, such as character identification, medical diagnosis, and voice recognition. More notably, the DT model has the potential to decompose a complicated decision-making mechanism into a series of simplified decisions by recursively splitting covariate space into subspaces, thereby offering a solution that is sensitive to interpretation.

6. CONCLUSION

This paper proposed a novel Difficult Set Sampling Technique, which enables the classification model to strengthen imbalanced network data learning. A targeted increase in the number of minority samples that need to be learned can reduce the imbalance of network traffic and strengthen the minority's learning under challenging samples to improve the classification accuracy. We used six classical classification methods in machine learning and deep learning and combined them with other sampling techniques. Experiments show that our method can accurately determine the samples that need to be expanded in the imbalanced network traffic and improve the attack recognition more effectively. In the experiment, we found that deep learning performance is better than machine learning after sampling the imbalanced training set samples through the MLP algorithm. Although the neural networks strengthen data expression, the current public datasets have already extracted the data features in advance, which is more limited for deep learning to learn the pre-processed features and cannot take advantage of its automatic feature extraction. Therefore, in the next step, we plan to directly use the deep learning model for feature extraction and model training on the original network traffic data, performance the advantages of deep learning in feature extraction, reduce the impact of imbalanced data and achieve more accurate classification.

7. REFERENCES

1. D. E. Denning, "An intrusion-detection model," IEEE Trans. Softw. Eng., vol. SE-13, no. 2, pp. 222–232, Feb. 1987.
2. N. B. Amor, S. Benferhat, and Z. Elouedi, "Naive Bayes vs decision trees in intrusion detection systems," in Proc. ACM Symp. Appl. Comput. (SAC), 2004, pp. 420–424. Page 35
3. M. Panda and M. R. Patra, "Network intrusion detection using Naive Bayes," Int. J. Comput. Sci. Netw. Secur., vol. 7, no. 12, pp. 258–263, 2007.

4. M. A. M. Hasan, M. Nasser, B. Pal, and S. Ahmad, ‘‘Support vector machine and random forest modeling for intrusion detection system (IDS),’’ J. Intell. Learn. Syst. Appl., vol. 6, no. 1, pp. 45–52, 2014.
5. N. Japkowicz, ‘‘The class imbalance problem: Significance and strategies,’’ in Proc. Int. Conf. Artif. Intell., vol. 56, 2000, pp. 111–117.
6. Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521, no. 7553, pp. 436–444, 2015.
7. Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, ‘‘Deep learning for visual understanding: A review,’’ Neurocomputing, vol. 187, pp. 27–48, Apr. 2016.
8. T. Young, D. Hazarika, S. Poria, and E. Cambria, ‘‘Recent trends in deep learning based natural language processing [review article],’’ IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55–75, Aug. 2018.
9. Sharma, R., & Dhabliya, D. (2019). Attacks on transport layer and multi-layer attacks on manet. International Journal of Control and Automation, 12(6 Special Issue), 5-11. Retrieved from www.scopus.com
10. Sherje, N. P., Agrawal, S. A., Umbarkar, A. M., Dharme, A. M., & Dhabliya, D. (2021). Experimental evaluation of mechatronics based cushioning performance in hydraulic cylinder. Materials Today: Proceedings, doi:10.1016/j.matpr.2020.12.1021
11. Sherje, N. P., Agrawal, S. A., Umbarkar, A. M., Kharche, P. P., & Dhabliya, D. (2021). Machinability study and optimization of CNC drilling process parameters for HSLA steel with coated and uncoated drill bit. Materials Today: Proceedings, doi:10.1016/j.matpr.2020.12.1070
12. Shukla, A., Almal, S., Gupta, A., Jain, R., Mishra, R., & Dhabliya, D. (2022). DL based system for on-board image classification in real time, applied to disaster mitigation. Paper presented at the PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing, 663-668. doi:10.1109/PDGC56933.2022.10053139 Retrieved from www.scopus.com
13. N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, ‘‘A deep learning approach to network intrusion detection,’’ IEEE Trans. Emerg. Topics Comput. Intell., vol. 2, no. 1, pp. 41–50, Feb. 2018.
14. D. A. Cieslak, N. V. Chawla, and A. Striegel, ‘‘Combating imbalance in network intrusion datasets,’’ in Proc. IEEE Int. Conf. Granular Comput., May 2006, pp. 732–737.
15. Mohammed Abdul Bari, Shahanawaj Ahamad, Mohammed Rahmat Ali,” Smartphone Security and Protection Practices”, IJEACS; ISBN: 9798799755577 Volume: 03, Issue: 01, December 2021 (International Journal,U K) Pages 1-6
16. Baig, M. S., Bari, D. R. M. A., & Khan, P. A. (n.d.). Weapon detection using artificial intelligence and deep learning for security applications. Ijarst.In. Retrieved May 10, 2023.
17. Khan, 1. Pathan Ahmed, & Waheed Farooqi, 2. M. R. M. (n.d.). Functional outsourcing of linear programming in secured cloud computing. Pen2print.org. Retrieved May 10, 2023.
18. Rahmat, M., & Pathan, A. (n.d.). Byod. A systematic approach for analyzing and visualizing the type of data and information breaches with cyber security.
19. Khan, P. A., Mir, M., Zakir, M. D., & Khan, A. (n.d.). Crop yield prediction using machine learning algorithms.