Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 12
December 2023

Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method

[1] Shindhe Sai Kiran, [2] Mr.N Sandeep Chaitanya

[1]M.Tech Student, Department of CSE, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology, Hyderabad, India -500090

[2]Assistant Professor, Department of CSE, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technologyy, Hyderabad, India -500090

Abstract:
Network intrusion is a critical challenge in information and communication systems amongst other forms of fraud perpetrated over the Internet. Despite the various traditional techniques proposed to prevent this intrusion, the threat persists. These days, intrusion detection systems (IDS) are faced with detecting attacks in large streams of connections due to the sporadic increase in network traffics. Although machine learning (ML) has been introduced in IDS to deal with finding patterns in big data, the irrelevant features in the data tend to degrade both the speed and accuracy of detection of attacks. Also, it increases the computational resource needed during training and testing of IDS models. Over the past years, IDSs and IPSs using different approaches have been developed and implemented to ensure that computer networks within enterprises are secure, reliable and available. In this paper, we focus on IDSs that are built using machine learning (ML) techniques. IDSs based on ML methods are effective and accurate in detecting networks attacks. However, the performance of these systems decreases for high dimensional data spaces. Therefore, it is crucial to implement an appropriate feature extraction method that can prune some of the features that do not possess a great impact in the classification process. Moreover, many of the ML based IDSs suffer from an increase in false positive rate and a low detection accuracy when the models are trained on highly imbalanced datasets. In this paper, we present an analysis the UNSW-NB15 intrusion detection dataset that will be used for training and testing our models.

Keywords: ML, IDS, IPS, UNSW-NB 15, attack, speed, accuracy.

I Introduction

Nowadays, it is challenging to protect confidential data from the eye of attackers. The traditional methods like firewall and antivirus failed to handle all types of attacks. So, there is a need for additional security along with traditional methods. IDS play a significant role in this regard. It carefully keeps a track on the network traffic data and differentiates the data as normal or attack. Dependency on networked computers for all walks of life has lured the attackers to launch attack on them.Protecting real-time high-speed networked computers gains high priority in the agenda of security specialists. Robust NIDS are required to thwart the intrusions launched by professional intruders. Data sets are used to train and test NIDS. A few of the key features in the benchmark dataset are sufficient to build NIDS models [1]. Inclusion of feature selection / reduction process have resulted in implementation of light-weight and robust NIDS. In the last three decades computer networks have grown in size and complexity drastically. This tremendous growth has posed challenging issues in network and information security, and detection of security threats, commonly referred to as intrusion, has become a very important and critical issue in network, data and information security. The security attacks can cause severe disruption to data and networks. Therefore, Intrusion Detection System (IDS) becomes an important part of every computer or network system. An IDS can monitor computer or network traffic and identify malicious activities that compromise the integrity, confidentiality, and availability of information resources and alerts the system or network administrator against malicious attacks. Since, an IDS needs to examine very large data with high dimension even for small network. Due to this, IDS has to meet the challenges of low detection rate and large computation. Therefore, Feature selection is a very important issue and plays

a key role in intrusion detection in order to achieve maximal performance. It is one of the important and frequently used techniques in data preprocessing for selecting a subset of relevant features to build robust IDS. Feature selection is the selection of that minimal cardinality feature subset of original feature set that retains the high detection accuracy as the original feature set [1]. The efficient feature subset can improve the training and testing time that helps to build lightweight IDS guaranteeing high detection rates and makes IDS suitable for real time and on-line detection of attacks.

Ii Survey Of Research

This section provides the literature survey on the ML algorithms. This section's main motive is to give an overview of the research work done in the field of intrusion detection. It is found in the literature that researchers have put a lot of efforts into ML algorithms, and some of their contributions are described below:

Narudin et al. (2014) [1] described an evaluation of ML classifiers, namely RF, J-48, MLP, NB, and KNN, to detect mobile malware using MalGenome and private datasets using Weka Tool. The performance metrics such as TPR, FPR, precision, recall and F-measure were used to validate ML algorithms' performance. The accuracy obtained using RF Classifier is 99.99% during experimental work on MalGenome dataset. The author has suggested the use of feature selection methods for improving the results in their future work.

Belavagi & Muniyal, (2016) [2] designed a NIDS with the various supervised machine learning classifiers. NSL-KDD dataset was used to check the performance of various classifiers. The result shows that RF classifier outperforms other classifiers. It results in the lowest FPR and the highest TPR and accuracy obtained is 99%. But still, there is a need for classifiers that can be used for the multiclass classification.

This refined dataset has been validated using 7 ML algorithms on WEKA tool [3] without tuning the parameters of the learners. The resulting classification accuracy per classifier are as follows; NB-Tree: 82.67%, J48: 81.05%, Random Forest (RF): 81.59%, Multi-Layer Perceptron (MLP): 77.41%, Naive Bayes (NB): 76.56% and Support vector machine (SVM): 69.52%. Using this dataset, a various number approach is being taken to either increase the classification accuracy or reduce the time needed during training and testing IDS. In some case, both objectives are achieved. One approach is to find optimal parameter setting of the classifier, and another is to tactically reduce the feature used to achieve faster training and testing time.

Garg and Kumar [4] reviewed various selection and classification techniques. The work tested the performance of combining two to three feature selection methods using Boolean AND operation. Out of 10 techniques tested, the combination of Symmetric and Gain Ratio for feature selection using 15 features and IBK classifier yielded the highest accuracy. However, no reason was given on why and how random data was selected from the dataset. Hence, the result can not be replicated.

Roshan et al. (2018) [5] discussed an adaptive design of IDS based on Extreme Learning Machines (ELM). The NSL-KDD dataset was applied for the evaluation. It was found that it can detect novel attacks and known attacks with an acceptable rate of detection and false positives.

Ali et al. (2018) [6] proposed a PSO-FLN classifier for intrusion detection. The benchmark dataset KDD99 was used to validate the results. PSO-FLN has outperformed ELM and FLN classifiers in terms of accuracy. But for some classes like R2L, it does not show accurate results.

Anwer et al. (2018) [7] proposed a feature selection framework for efficient network intrusion detection. The UNSW-NB15 data set was used to evaluate five different strategies. J48 and NB methods were used for evaluation. The best strategy was to use a filter ranking method to select 18 features, and then J48 was applied as a classifier. By using this strategy, 88% accuracy and a speedup factor of 2 was achieved.

Iii Existing System

The increasing growth of machine learning, computer techniques divided into traditional methods and machine learning methods. This section describes the related works of performance of Intrusion Detection systems and how machine learning methods are better than traditional methods. The existing method in this project have a certain flow is used for model development. SVM

and ANN are used algorithms in existing system. But it requires large memory and result is not accurate

Iv Proposed System

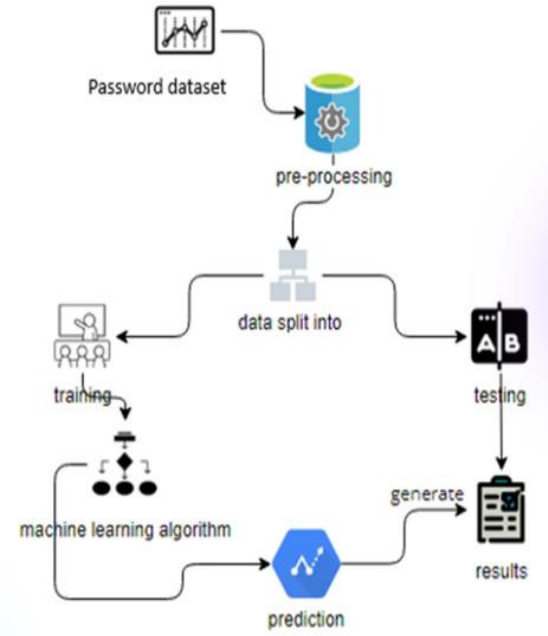
Much research has been done on studying performance analysis of intrusion detection of systems. In early days if have to check the system performance it's a time taking process. We can detect the system performance by using Decision tree, KNN (K-Nearest Neighbour), LG (Logistic Regression), Decision Tree, XGBoost. To get better accuracy we used Decision Tree algorithm. Bu using these machine learning techniques we can easily detect the system performance like weather the system is attacked or not.

Intrusion Detection System

An intrusion is defined as an attempt to compromise the confidentiality, integrity, availability, unauthorized use of resources, or to bypass the security mechanisms of a computer system or network and James P. Anderson introduced Intrusion Detection (ID) early in 1980s [2]. Dorothy Denning proposed several models for IDS in 1987 [3]. Ideally, Intrusions Detection (ID) should be an intelligent monitoring process of events occurring in system and analyzing them for security violations policies. An IDS is required to have a high attack Detection Rate (DR) with a low False Alarm Rate (FAR). Refer [4] for the organization of a generalized IDS. Approaches of IDS based on detection are anomaly based and misuse based intrusion detection approach. In anomaly based intrusion detection approach [5], the system first learns the normal behavior or activity of the system or network to detect the intrusion. In misuse or signature based intrusion detection approach [6], the system first define the attack and the characteristics of the attack that distinguish this attack from normal data or traffic to detect the intrusion. Approaches of IDS based on location of monitoring are Network based intrusion detection system (NIDS) [7] and Host-based intrusion detection system (HIDS)[8]. NIDS detects intrusion by monitoring network traffic in terms of IP packet. HIDS are installed locally on host machines and detects intrusions by examining system calls, application logs, file system modification and other host activities made by each user on a particular machine.

V Methodology

Though some IDS adopts the working principle of ML for classification of connections in a network, it still faces a critical challenge which limits its use in some real life environments. Some systems proposed are either slow or raise false alarms, e.g., classification of normal connection as intrusive which could frustrate the experience of client or classification of intrusive connection as normal, leading to a significant loss. Also, the fundamental issues about UNSW_NB15 dataset give doubt on the accuracy presented by most works. For an IDS to be effective, the classification accuracy must be high, and detection rate must be fast without the usage of excessive computational resources. To achieve this goal, there is the need for removing irrelevant features from network connections during classification. Hence, the major problem is how to select the best set of features for the right ML algorithm without loss of relevant information.



Vi Implementation

The benchmark datasets used in the literature are older datasets and contain repeated records due to which ML algorithms give unfair results. So, selected ML algorithms are tested using UNSW NB15 dataset, which is novel dataset. This dataset comprises 49 attributes, including a class label and 25, 40, 044 labelled instances, each being labelled either normal or attack. The pre-processing steps are shown in Fig, shows the methodology used. In preprocessing, first of all, the null values present in the dataset are

handled. Then the categorical data is converted into the numerical form using label encoder. After this, one hot encoder is used to break the relation between the values obtained through label encoder. The standard scaler is used to standardize the values at one scale.

After performing pre-processing, significant features are selected using chi-square, and after this, the pre-processed data is separated into training and testing data. 80% of data is used for training, and 20% of data is used for testing. The KNN, LR, NB, SGD and RF classifiers are used to construct the models. Then the prediction of labels of test data is made using these models. A comparison is carried out between actual labels and predicted labels.

The procedural steps to construct the models are given below:

1. Start with pre-processing of the dataset.
2. Select the significant features using the chi-square method.
3. Divide dataset into two parts, i.e. training and testing.
4. Construct the classifier model using training data for KNN, LR, NB, RF and SGD.
5. Take the test data.
6. Testing of classifier models using training data.
7. Calculate and compare Accuracy, Recall, Precision, F1-Score and MSE for the selected models.

Conclusion

This research explored the application of the XGBoost algorithm for feature selection in conjunction with multiple ML techniques including ANN, kNN, DT, LR and SVM in order to implement accurate IDSs. The UNSW-NB15 dataset was utilized to assess the performance of these methods. In this work, the binary and the multiclass classification settings were considered. Furthermore, the XGBoost based attribute selection method was applied over the UNSW-NB15 and as a result, 19 optimal features were selected. To put our research into perspective, we carried out a thorough literature review whereby various techniques for feature selection applied to the UNSW-NB15 dataset were reviewed. Moreover, we compiled a summary of the performance results obtained by the various classifiers in the literature and we compared them to those obtained in our proposed methodology. Initially, we carried out the experiments using the proposed ML approaches over the full feature space of the UNSW-NB15 dataset. Thereafter, we ran the experiments using the reduced feature vector that was generated by the XGBoost feature extraction algorithm proposed in this work. The experimental results demonstrated that using a reduced (optimal) feature vector has its merits in terms of reducing the models complexity as well as increasing the detection accuracy on test data. A notable instance is the XGBoost-ANN. It managed to reduce number of neurons in the hidden layer by 50 percent the number of nodes in the stand alone ANN.

Referances

[1] Wang Z: Deep learning-based intrusion detection with adversaries. IEEE Access. 2018;6:38367–384.
[2] Kasongo SM, Sun Y. A deep gated recurrent unit based model for wireless intrusion detection system. Cakovec: ICT Express; 2020.
[3] Ribeiro J, Saghezchi FB, Mantas G, Rodriguez J, Abd-Alhameed RA. Hidroid: prototyping a behavioral host-based intrusion detection and prevention system for android. IEEE Access. 2020;8:23154–168.
[4] Van NTT, Thinh TN. Accelerating anomaly-based IDS using neural network on GPU. In: 2015 international conference on advanced computing and applications (ACOMP). IEEE; 2015. pp. 67–74.
[5] Jabez J, Muthukumar B. Intrusion detection system (IDS): anomaly detection using outlier

detection approach. Procedia Comput Sci. 2015;48:338–46.

[6] Neelakantan S, Rao S. A threat-aware anomaly-based intrusion-detection approach for obtaining network-specific useful alarms. In: International conference on distributed computing and networking. Springer. 2009; pp. 175–180.

[7] Kasongo SM, Sun Y. A deep learning method with filter based feature engineering for wireless intrusion detection system. IEEE Access. 2019; 7:38597–607.

[8] El Naqa I, Murphy MJ. What is machine learning? In: Machine learning in radiation oncology. Berlin: Springer; 2015. p. 3–11.

[9] Khatri S, Arora A, Agrawal AP. Supervised machine learning algorithms for credit card fraud detection: a comparison. In: 2020 10th international conference on cloud computing, data science & engineering (confluence), IEEE; 2020. pp. 680–83.

[10] Singh P. Supervised machine learning. In: Learn PySpark. Springer; 2019. pp. 117–59.

[11] Harrington P. Machine learning in action. New York: Manning Publications Co.; 2012.

[12] Dong G, Liu H. Feature engineering for machine learning and data analytics. Boca Raton: CRC Press; 2018.

[13] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD interna- tional conference on knowledge discovery and data mining, 2016; pp. 785–94.

[14] Zhu Z, Ong Y-S, Dash M. Wrapper-filter feature selection algorithm using a memetic framework. IEEE Trans Sys Man Cybern Part B (Cybern). 2007;37(1):70–6.

[15] Moustafa N, Turnbull B, Choo K-KR. An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of internet of things. IEEE Internet Things J. 2018;6(3):4815–830