International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 422 - 433

Classification of Malevolent and Benevolent Network Traffic

Aayushi Jain[1], Vimal Kumar[2]

[1]*M. techStudent,Meerut Institute of Engineering and Technology, Meerut*

[2]*Assistant Professor, Meerut Institute of Engineering and Technology, Meerut*

[1]*aayushi.jain19@gmail.com,*[2]*vimal.kumar@miet.ac.in*

Abstract

Technological advancement has inclined people towards digital world. According to Global Digital Report, around 4.39 billion people are using internet and sharing their vital information over it. Due to such high usage, digital world has become the hub of information thereby, leading to increase in number of intruders. All vital information traverses through network and it is very important to protect our significant data by classifying our network traffic.We need to apply various techniquesto differentiate between malicious or non-malicious data packets. Due to continuous evolution of technology and dynamic nature of internet, traditional models for network traffic classification like port number and payload classification are not competent enough.The main challenge for the researchers is to classify encrypted and encapsulated traffic.This gap has provided us scope to try and use machine learning for classification of network traffic effectively. Machine learning approach helps us in extracting knowledge from the encrypted traffic.

Keywords: *Machine Learning Approach, Network Traffic Classification*

1. Introduction

The demand of an intrusion detection system (IDS) in various applications has increased in the recent years since huge amount of data is available to be stored and processed every day. The networking systems are generating huge amount of data by monitoring the surroundings of applications in which they are deployed. Any kinds of suspecting behaviors are detected by the devices. Any kinds of vulnerabilities in any computer network can be found by an intruder that aims to harm the users using that device. For preventing the entry of intrusions, the best solution is to protect the system or its resources [1]. Following are the important security perspectives to be considered to secure a computer system:

- **Confidentiality**: The information can be accessed only by an authorized user.
- **Integrity**: The information must not be affected by any vulnerability of system.
- **Availability**: By ensuring that the functioning of system is not degraded, authorized users must be provided access to the systems and its resources.

Any activity that attempts to trigger an event due to which the system's security is compromised is called as an intrusion. Either internally or externally, the intrusions might occur in any system. Any kind of illegal activity or fraud information that makes a computer

ISSN: 2233-7857 IJFGCN

hazardous can be considered as intrusion. An IDS is the method in which the events being performed in a computer network can be monitored and analyzed so that it becomes easy to recognize the security problems. An IDS acts as an alarm and any kinds of violations in the system are identified by it. Even if there are false messages in messages, mails or video sounds, they can be alerted by the systems. A tool that acts as a guard such that the system can be secured against any kinds of intrusions or attacks is called intrusion detection system. To check the attack scenarios and provide required support for defense management are the important objectives of IDS. Today, in networking, almost all the applications are using IDS systems.

IDS can be used to detect any malicious activities which cannot be detected by a common firewall. Against the sensitive services, computer applications and other regions, attacks are possible in the computer systems. Data driven attacks are possible in computer applications, network attacks in sensitive services and unauthorized logins in case of sensitive files are faced due to intrusions.

Basic IDS architecture is shown in Figure1

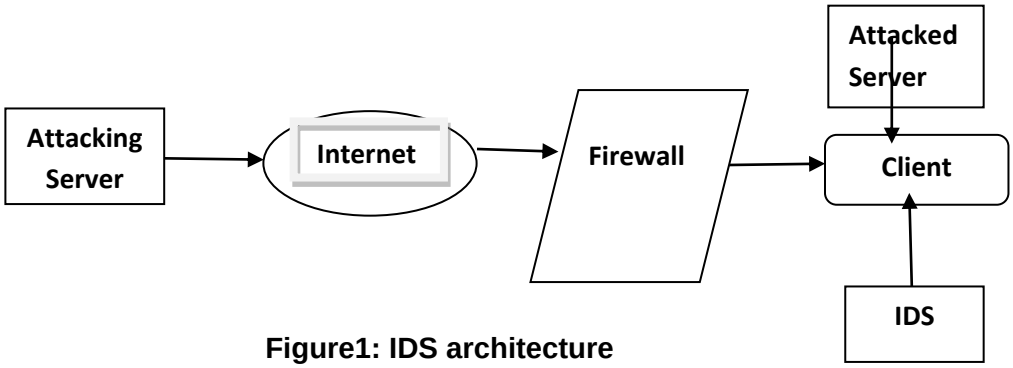


Figure1: IDS architecture

IDSes observe the network traffic in order to find out whether there is any intrusion is done by unauthorized entities [2]. IDSes passes on these functions partially or fully to the professionals managing the security:

- It monitors the functioning of firewalls, routers, keymanagement servers and files that are used by other security controls for detecting, preventing or recovering data.
- It providesunique approach to understand and maintain respectiveOS audit trails/logs to the administrators which are otherwise neglected.
- It provides easy to use interfaceto that even a non-technical member can help with management of system security.
- It includesa huge database covering extensive information of attack signature which can be match with the system information.
- It recognized and report when the IDS detects alteration in the data files.
- It notifies whenever security has been compromised by trigger an alarm.
- It gave reply to the hackers by blocking them.

ISSN: 2233-7857 IJFGCN

An IDS can be deployed on customer hardware as a software. For cloud deployment, cloud base IDS are also available.

1.1 Host-Based Intrusion Detection System: The systems which monitor the device on which they are installed are known as host-based IDSs. The states of main system through audit logs to the program execution are monitored by this approach to execute the monitoring program.

There are two broader categorizations of HIDS based on the source of data that is needed to be examined, which are:

- **The HIDS Based Application:** The data present in the application is received by the IDS of this type. For instance, the management software, server web or firewalls that generate the log files can be included here. The layer application includes the vulnerability of this technique [3].
- **The HIDS Based Host*:*** The information related to the activity of supervised system is included in this type of IDS. In the form of audit traces of operating system, the information is received. The logs system of separate logs generated by the processes of operating system is included here. Within the standard audit of operating system and the logging methods, the contents of object system do not reflect. The results of other IDS of the Based Application type are used by these types of IDS.

1.2. Network-Based Intrusion Detection System: For monitoring the traffic being received from devices to certain other devices of a network, the NIDS is deployed at certain points of network. Wiretapping and these systems have similar concepts. In a network, the intrusions tap and hear the transmissions being held in the networks. The risk needs to be reduced by reducing the network activity of intruder. In comparison to HIDS, the portability of NIDS is higher. Across the network, the traffic is monitored by them and the operating system on which they run is not their dependent factor. For determining if the data is malicious or not, multiple techniques are used which analyze the traffic. For network data analysis, two methods are applied:

- **Packet-based analysis:** The complete packet that includes payload and headers is used here. A packet-based NIDS is the system in which the packet-based analysis is performed. Huge amount of data is included for processing which is the benefit of performing this kind of analysis. To check if the packet is malicious or not, every single byte of packet is utilized.
- **Flow-based analysis:** The general aggregated data related to network flows is utilized instead of individual packets in this process. A flow-based NIDS is the system in which flow-based analysis is applied. In between the host and another device, a single connection is used to define a flow [4].

1.3. Intrusion Prevention System: For preventing the attacks another method known as intrusion prevention system is introduced. Even if it is preferred, IDS does not detect the

attacks at the exact movement of occurrence. Attacks are detected at real-time by the IPS since the prevention of attacks is also provided through them. The connections can be closed, IPs areblocked, or limited data throughput is provided such that these kinds of attacks can be prevented.

The contribution of this paper is structured as follows: Section 2 comprises of traditional approach. Section 3 gives the brief descriptionof machine learning. Section 4 give the step of ML algorithm. Section 5 explain the performance matrix. Section 6 describes the related work. Section 7 contains the brief introduction of traditional model used to classify traffic and it explains the amount of work done to classify network traffic using machine learning technology.

2. Traditional Approach

Various techniques used to classify network traffic in the past are as follows:

2.1.Classification Based on Port Number: Initially network traffic was classified on the basis of port number where each application has their port number registered in IANA For Eg: FTP file has 21 port number but, there was some limitation of using this technique like some application didn't not registered themselves in IANA , some application uses another port number other than the registered and because of dynamically allotment of port number this technique was unreliable.

2.2.Classification Based on Payload: This technique is also known as DPI(Deep Packet Inspection). In this technique packet payload are analyzed to check whether they contain the characteristics nature of registered or known application. This technique has been works well for Peer to Peer(P2P) Traffic. This technique also has some limitation that it breaches security and privacy policy as it analyzed the data, this technique also increases the computational cost and lastly this approach does not work for encrypted data.

2.3.Transport Layer Heuristics: Karagiannis et al proposes an approach that uses the unique behavior of Peer to Peer application when they transfer data or making connection to identify the network traffic.This technique performed better than the port-based classification technique.

2.4.Machine Learning Approach: ML provide a suitable solution not only for the traffic classification but also for the prediction and new knowledge discovery. In this, statistical features of IP flows are extracted from network traces and they are saved to generate historic data. In this way different Machine learning models are trained with the historic data.

3. Basics of Machine Learning Approach

Machine Learning is concerned with the system that can be learned from data. In 1959,Arthur Samuel defined it as the "Field of study that gives computers the ability to learn without being

explicitly programmed". Machine learning consist of two main parts first is the model building and second is the classification. A model is build using trained data and then this model work as input for classifier which further classify it into dataset as shown in Figure.2

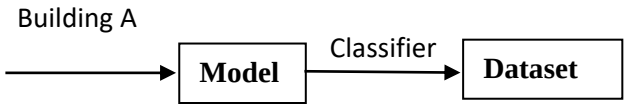


Figure2: Basic Structure

3.1.Different Types of Machine Learning Algorithm

Machine Learning algorithm can be categorized based on desired outcome of the algorithm or based on input available.

3.1.1.Supervised Learning: In supervised learning, model parameters are adjusted in such a way to minimize the error between the model outcome and the expected outcome of the machine.

Supposewe have input variable =x

Output variable=y

We apply algorithm for mapping from the input function to output

Such that,y=f(x),the main objective of this algorithm is that for every new input x,we can predict the output y of that new input.

Supervised learning problem is again categorized into two categories:

a. Regression: When the output variable is a real value. For E.g.Weight Height

b. Classification: When the output is category. For E.g. Male or Female ','Red or green'.

3.1.2.Unsupervised Learning: In this, we only know the input variable, but we do not know the corresponding output to that input.

The main objective of unsupervised learning is to model the distribution of data so that we can analyze data and draw some more results.

Unsupervised Learning problems are further categorized into two categories:

a.Clustering: Grouping of similar data into one category.

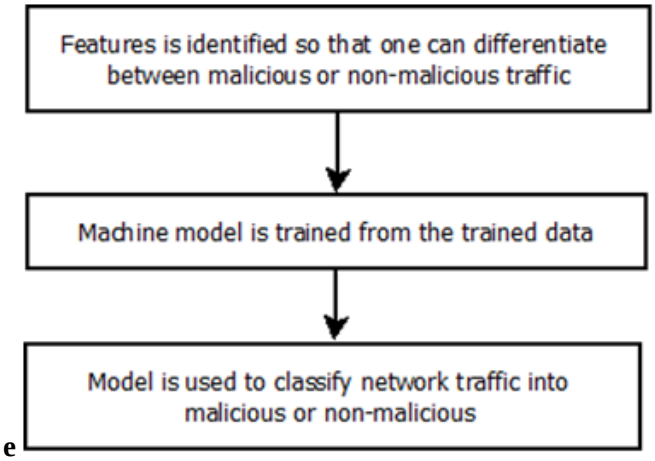
b. Association: We create rules to describe large amount of data.

3.1.3.Reinforcement Learning: RL is to take actions in an environment tomaximize the reward.

Table 1: Types of ML

Types	**Supervised Learning**	**Unsupervised Learning**	**Reinforcement Learning**
Data	Training Data	Untrained Data	-
Known data	Both input and Output	Only Input	Reward/Penalty
Technique	Classification	Clustering	Q-Learning
Example	Naïve Bayes	K-Mean	

4. ML Model for Network Traffic Classification



e

Figure3: ML Model

5. Performance Evaluation Matrix

A summary of the classification results which are obtained by using the machine language classifier for a binary classification problem is shown in the Table 2below:

Table 2: Evaluation matrix

Metrics		**Actual Class**	
		X	**Not X**
Predicted Class	X	TP	FP

	Not X	FN	TN

The performance metrics for binary classification are as follows:

1.Accuracy: It is the ratio of total no.true prediction with the total no. of predictions done. Accuracy=$(TP + TN)/(TP + FP + FN + TN)$

2. Recall: It is the ratio of true positive with the total no. of truly classified items.

Recall=$TP/(TP + FN)$

3.Precision: It is the ratio of true positive classification from the total no. of the truly predicted classification.

Precision=$TP/(TP + FP)$

6. Network Traffic Classification Using Machine Learning Approach

Mane et.al (2019)proposed for traffic classification uses two machine learning algorithm Bayesian Algorithm and SVM (Support Vector Machine) Algorithm [5]. In the initial stage they create a dummy website using eclipse software. The main purpose for this website is for Ddos attack i.e. denial of service attack for stopping the services of the website. As the Ddos attack occur they capture the network data and then the classifier is used to distinguish between the normal or abnormal data. The result shows that 95% accuracy is been achieved.

Zhong Fan & Ran Liu(2017) uses SDN(Software Defined Networking) for network traffic classification.In this paper researcher used two machine learning algorithm K-Mean and SVM(Support Vector Machine)[6].It is found that 95% accuracy is been achieved. It is also proved in this paper that SVM gave higher accuracy than any other algorithms.

ZiedAouini, AbdesselemKortebi,YacineGhamri and Iyad Lahsen(2017)proposed the classification algorithm for residential traffic. The algorithm they used for classification is C5.0 ML algorithm[7].

AltyebAltaher, (2017) proposed a hybrid mechanism through which the websites as Legitimate, Suspicious, or Phishing websites can be categorized. There are two stages used in this proposed algorithm to generate this hybrid method in which the KNN and SVM classifiers are combined [8]. The KNN is applied in the initial stage which is effective and robust to the noisy data. In the second stage, another powerful classifier is applied which is known as SVM. The effectiveness of SVM algorithm is improved by the proposed algorithm when the simplicity of KNN is integrated. Evaluations are performed by conducting simulation experiments and it is seen through the outcomes that in comparison to other approaches, the accuracy of proposed approach is highest which is 90.04%.

ISSN: 2233-7857 IJFGCN

Jayshree Jha, et.al (2013) proposed a research that was based on two important contributions. In the initial contribution, the intrusion detection performed using SVM was reviewed in this paper along with the other technologies proposed by different authors [9]. Further, to detect intrusion, the best feature was chosen by proposing a novel method in the second contribution. To select the relevant features, a hybrid approach was proposed in which the filter and wrapper models were combined. The performance and detection accuracy of SVM based detection model were increased by reducing the dataset. Furthermore, with the reduction in the set of feature, it is also possible to reduce the training and testing time.

L.Dhanabal, et.al (2016) performed an analysis of the KSL-KDD dataset. The anomalies in network traffic patterns were detected by studying the effectiveness of different classification algorithms [10]. For generating anomalous network traffic, the relationship of protocols available in commonly used network protocol stack was analyzed with the attacks used by intruders that generated the anomalous network traffic. The data mining tool WEKA was used to perform analysis using the classification algorithms. Several facts that bonded between the protocols and network attacks were exposed in this study.

WathiqLaftah Al-Yaseen, et.al (2015) proposed a multi-level hybrid IDS model. Here, the efficiency to detect known and unknown attacks was improved using the support vector machine and extreme learning machine [11]. For improving the performance of classifiers, a modified k-means algorithm was also proposed which built a high-quality training dataset. The new small training datasets which represented the complete original training dataset were generated using the modified k-means algorithm. Thus, the training time of classifiers was reduced and the performance of IDS was improved by this proposed method. The proposed model was evaluated using the popular KDD Cup 1999 dataset. In terms of attack detection, high efficiency was achieved by the proposed model in comparison to other methods designed and implemented on similar dataset. Also, the accuracy achieved was also better than all the algorithms studied so far on this research.

Amol Borkar, et.al (2017) presented a survey of the Internal-IDS and IDS in which the real time based data mining and forensic techniques algorithms were applied [12]. In support of intrusion detection, different data mining techniques were proposed for cyber analytics. Based on the studies of different techniques presented by different authors, this research presented the manners in which the intruder could be detected. The survey presented in this paper helped in drawing the conclusion. The accuracy and detection rate were improved up to 95% by applying the proposed technique in comparison to the existing techniques which provided around 90% of accuracy and detection rate.

Jianguo Yu, et.al (2018) studied that in the information security related field, huge focus has been shown on the intrusion detection in the rail transit field. For the intrusion detection and mis-operation of subway environment control subsystem, the BAS intrusion detection expert system was designed on the basis of expert system. Further, in the expert system, the knowledge base and inference engine design were introduced [13]. For the mis-operation and mis-utilization of intrusion detection, expert systems were used by the system. Further, for

preventing anomalous intrusion, the black and white list rules were added which helped in protecting the information security of subway environment control system to the utmost level. Further, to provide information security of multiple subsystems of subway, foundation was laid by this method. Currently, due to certain imperfections, this system is only being used in exploratory state. However, the IDSs can be applied to the complete metro region with the advent of era of big data.

Marin E. Pamukov, et.al (2017) studied about an algorithm through which a huge percentage of possible intrusions could be categorized as being true or false without requiring the presence of any operator input [14]. The co-stimulation principles of Immunology and Negative Selection algorithm were used as base to design this method. The co-stimulation mechanism which aimed to reduce the number of detection errors without requiring the operator input was implemented here using a two-tier negative selection method. Around 34% of all intrusions could be detected using MNSA algorithm without needing the knowledge about the non-self. Further, more than 90% of those detections were confirmed by the proposed method without requiring the additional information or operator unit.Table 3 shows the evolution of ML technologies

Table3: Evolution of ML Technologies

Author /Year	Paper Published	Methodology Used	Outcome
Mane/2019	Traffic Classification usingMachine Learning	SVM Naïve Bayesian Decision Tree	The result shows that 95% accuracy is been achieved.
Zhong Fan & Ran Liu/2017	Investigation of machine learning based Network Traffic classification	Used-SDN(Software Defined Networking) ML Algorithm-SVM K-Mean	The result shows that 95% accuracy is been achieved and SVM gave higher accuracy than any other unsupervised algorithm.
ZiedAouini /2017	Early classification of residential network traffic using C5.0 machine learning algorithm	C5.0	The result shows that 98.8% accuracy is been achieved.
AltyebAltaher/ (2017)	Phishing Websites Classification using Hybrid SVM and KNN Approach	KNN SVM	The result shows that 90.04% accuracy is been achieved.
Hardeep Singh		Unsupervised K-Mean Expectation	The result shows that K-mean was better than expectation

		Maximization Algorithm	maximization. Small no.of cluster K-mean accuracy=65% EM=55% Larger no. of cluster K-mean=88% EM=84%
L.Dhanabal/ 2016	A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms	KDD dataset used WEKA Tool	Facts that bonded between the protocols and network attacks were exposed .
WathiqLaftah Al-Yasee 2015	Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System	SVM Extreme Learning Machine K-Mean	High efficiency was achieved by the proposed model in comparison to other methods designed and implemented on similar dataset.

7. Conclusion

This paper contains the brief introduction of traditional model used to classify traffic and it explains the amount of work done to classify network traffic using machine learning technology.It is clear that machine learning approach overcome the drawback of port based and payload- based classification approach but so far ML has not been utilized to its potential to classify the network traffic. And there is still lot of scope left in this area.

Nevertheless, Machine learning based network traffic classification opens the new doors for the researchers to try their hand in various areas like intrusion detection, anomaly detection, routing traffic and many others.

Acknowledgement

I would like to express my sincere gratitude to my guide Dr. Vimal Kumar for providing their invaluable guidance, comments and suggestions. Secondly,I would like to thank my parents and my brother who helped me a lot.

References

[1] Amrita, Kiran Kumar Ravulakollu, "A Hybrid Intrusion Detection System: Integrating Hybrid Feature Selection Approach with Heterogeneous Ensemble of Intelligent Classifiers", International Journal of Network Security, Vol.20, No.1, PP.41-55, Jan. 2018

[2] Bayu Adhi Tama and Kyung-Hyune Rhee, "Performance evaluation of intrusion detection system using classifier ensembles", Int. J. Internet Protocol Technology, Vol. 10, No. 1, 2017

[3] Nanak Chand, Preeti Mishra, C. Rama Krishna, Emmanuel Shubhakar Pilli and Mahesh Chandra Govil, "A Comparative Analysis of SVM and its Stacking with other Classification Algorithm for Intrusion Detection", 2016, IEEE

[4] M. Mazhar, U. Rathore, "Threshold-based generic scheme for encrypted and tunneled Voice Flows Detection over IP Networks", Journal of King Saud University Computer and Information Sciences, vol. 27, pp. 305–314, 2015.

[5] Prof.PranitaMane,"Traffic Classification Using Machine Learning",2019 2nd International Conference on Advances in Science & Technology(ICAST-2019).

[6] Zhong Fan & Ran Liu,Invetigation of machine learning based Network Traffic Classification

[7] ZiedAouini, AbdesselemKortebi, YacineGhamri and Iyad Lahsen (2017) Early classification of residential network traffic using C5.0 machine learning algorithm.

[8] AltyebAltaher, "Phishing Websites Classification using Hybrid SVM and KNN Approach", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017.

[9] Jayshree Jha, Leena Ragha, "Intrusion Detection System using Support Vector Machine", 2013, International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA International Conference & workshop on Advanced Computing

[10] L.Dhanabal, Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", 2016, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6

[11] WathiqLaftah Al-Yaseen, Zulaiha Ali Othmana, MohdZakree Ahmad Nazri, "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", 2015, Expert Systems With Applications.

[12] Amol Borkar, Akshay Donode, Anjali Kumari, "A survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and protection system (IIDPS)", 2017 International Conference on Inventive Computing and Informatics (ICICI), Pages: 949 – 953

[13] Jianguo Yu, Pei Tian, Haonan Feng, Yan Xiao, "Research and Design of Subway BAS Intrusion Detection Expert System", 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Pages: 152 – 156

[14] Marin E. Pamukov, Vladimir K. Poulkov, "Multiple negative selection algorithm: Improving detection error rates in IoT intrusion detection.

ISSN: 2233-7857 IJFGCN

Authors

Aayushi Jain, currently undertaking a master's degree in computer science at Meerut Institute of Engineering and Technology, Meerut. Studied B. tech in Information Technology from Radha Govind Engineering college, Meerut, Uttar Pradesh,India.

Dr. Vimal Kumar is a Associate Professor in the Department of Computer Science & Engineering at MIET, Meerut,(U.P),India.He received his B.tech Degree in 2007 from Uttar Pradesh Technical University, Lucknow and M.tech degree in Information Securityfrom Motilal Nehru National Institute of Technology, Allahabad, India in 2011. He did his Ph.D in Computer Science and Engineering from MMMEC Gorakhpur (AKTU, Lucknow), India in 2017. He has published a large number of various research papers in International and National journals and conferences of high repute. His research interests lie inMobile Ad hoc Network, Network Security and Network Forensics.

ISSN: 2233-7857 IJFGCN