Comparison of Effective Features Selection Method in Intrusion Detection System with Testbed

Htay Htay Yi
Information and Communication Technology Training Institute (ICTTI)
Information and Communication Technology Research Center (ICTRC)
Yangon, Myanmar
htayhtayyee@ictresearch.edu.mm

Khaing Khaing Wai
Department of Information Technology Support and Maintenance
University of Computer Studies, Yangon (UCSY)
Yangon, Myanmar
khaingkhaingwai@ucsy.edu.mm

***Abstract*—The feature selection method is curial to enhance the performance of the system by reducing the unnecessary features in preprocessing of data. The system implements a network testbed with a firewall, and Intrusion Detection System (IDS) and creates a dataset from it using Normal traffic, DoS attack, and Portscan attack traffics. The effective features in the proposed dataset apply feature selection methods as Gain Ratio (GR) and correlation-based feature selection (CFS) that improve the detection of intrusion system. The aim of the system is to compare the false positive rate with an existing dataset CICIDS2017 by machine learning classifiers. It is to prove that the features are superior by reducing the false positive rate and saving the time in the proposed system.**

Keywords— Intrusion Detection System, feature selection methods, false positive rate, features

I. INTRODUCTION

In the real world, the number of internet users increases at that time security on the network becomes more important. So, IDS and firewall are made use of security technology. In this work, the system implements a network testbed that includes a firewall, and IDS that analyzes DoS and Portscan traffic and applies machine learning to detect intrusion. The machine learning better version is proposed to improve results of Intrusion Detection and false positive rate or false alarm [1]. Therefore, the detection rate and false positive rate are important for IDS. In comparison with 26 features of CICIDS2017 [5], the best features or attributes method is not used, and its requirement is considered in this paper. The main research areas of this paper are: 1) Creating the firewall policies and apply on the firewall each interface. 2) Providing IDS predefined rules and proving with machine learning. 3) Proposed dataset implemented to improve the performance an especially false positive rate of the system. 4) Compare the proposed with an existing dataset to prove the effective features of a proposed dataset.

II. IMPLEMENTATION AND RESULT

In the proposed system, a software-based firewalls as IPCoP and IDS as a snort in the testbed environment that takes from the traffic of firewall, IDS, web server, and public attacks for creation of a proposed dataset.

A. Proposed Design for the System

The software-based firewall is configured for External Network, Local Area Network (LAN) and De-Militarized Zone (DMZ) for public and local user's access in figure 1. The firewall specifies the policies in each of IPCoP's five interfaces for the security of the organization [4]. The IDS of Snort operates with two NIC cards for the external networks and LAN networks and it applies predefined rules related to the firewall.

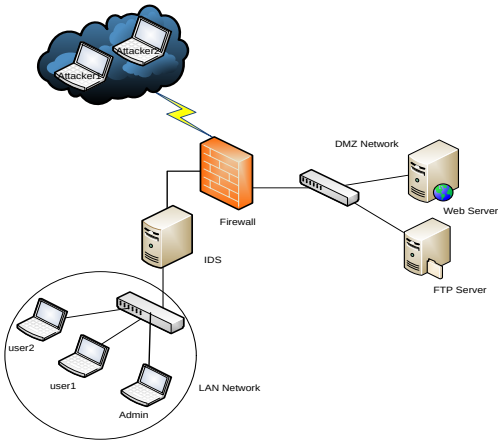

Fig 1. Proposed System Design

B. Selected Feature from Network Traffic

All the network traffics use the **tcpDump** tool to capture and create a pcap file. Each of the packet ranges applies the **Wireshark** tool from the pcap file. The **hping3** tool is used to capture the attack traffics and is used together with the relevant attack options. The proposed dataset is composed of the main 16 selected features in Table 1. The package range from normal and attacks traffic based on time [3]. When choosing the values of the features that calculate in detail depending on the inbound/outbound of the destination host according to the package range in the proposed system.

C. Proposed Dataset on Features Selection

The proposed dataset uses Correlation-based feature selection (CFS) and Gain ration (GD) to prove the effective features with the best first method and ranks evaluator by using WEKA (Waikato Environment for Knowledge Analysis) data mining tool. The increasing number of false positive rates impacts on the detection rate of IDS. In this work, machine learning classifiers are used on DoS and Portscan (PScan) attacks to measure the false positive rate of the system.

The CFS selects the proposed system of the subset of six features 1,8,12,13,14,15 for DoS and five features 9,11,12,13,15 for Portscan. And then Gain Ratio selects 15 features with each feature rank except the class feature. The minimize of false positive rate, correctly, and incorrectly classified instances of the proposed dataset determine with two attacks by using six machine learning classifiers in Table 2 and Table 3.

TABLE 1. DATASET FEATURES APPLIED IN SYSTEM

No	Feature	Description
1	Dst_port	Destination port
2	Dst_IP	Target IP address
3	Total_Inpkt	Total Inbound packages to destination host
4	Total_Outpkt	Total Outbound packages from destination host
5	Inpkt_bytes	Inbound packages bytes to Destination
6	Outpkt_bytes	Outbound packages bytes from destination
7	Total_InOut_pkt	Total packages to/from destination host
8	Inpkt_bits/s	Inbound packet bits/s
9	Outpkt_bit/s	Outbound packet bits/s
10	Protocol	Protocol as TCP or UDP
11	Service	Service types as http, ftp
12	Min_pktlen	Minimum packet length in the packet range
13	Max_pktlen	Maximum packet length in the packet range
14	Avg_pktlen	Average length of packet that fall in the range
15	Inout_count	Number of packets count with source and destination IP in this range
16	Class	Describe normal or attack

TABLE 2. CFS SELECTED FEATURES OF PROPOSED DATASET

Classifiers	False Positive Rate		Correctly Classified Instances		Incorrectly Classified Instances	
	DoS	PScan	DoS	PScan	DoS	PScan
Logistic Regression	0.001	0.023	99.9003	99.5432	0.0997	0.4577
Naïve Bayes	0.004	0.000	99.5015	99.5423	0.4985	0.4577
Bayes Net	0.000	0.000	100	100	0	0
J48	0.001	0.000	99.9003	100	0.0997	0
Random Tree	0.000	0.000	100	100	0	0
Random Forest	0.001	0.000	99.9003	100	0.0997	0

D. Comparison of Proposed and Existing Dataset

Among the 78 features in CICIDS2017 [2], flag features are not considered, and the remaining features are considered. The CFS features a selection that chooses four features as Destination Port, Total Length of Bwd Packets, Init_Win_bytes_forward, and Idle Max for DoS attacks and also takes four features as Bwd Packet Length Mean, Init_Win_bytes_backward, act_data_pkt_fwd, and min_seg_size_forward for Portscan attack respectively. Table 4, it shows the results with machine learning classifiers.

The comparison of good features can be clearly seen in Table 2 and Table 4. CICIDS's DoS attack uses full features including the flag features and the result is the same as not using the flag features in Correlation based feature selection. So, the flag features are not considered in both the proposed dataset and CICIDS2017. In this work, CFS selected affective attributes or features from both datasets were considered.

TABLE 3. GAIN RATIO SELECTED FEATURES OF PROPOSED DATASET

Classifiers	False Positive Rate		Correctly Classified Instances		Incorrectly Classified Instances	
	DoS	PScan	DoS	PScan	DoS	PScan
Logistic Regression	0.005	0.000	99.5015	99.5423	0.4985	0.4577
Naïve Bayes	0.004	0.001	99.6012	98.8558	0.3988	1.1442
Bayes Net	0.001	0.000	99.9003	99.5432	0.0997	0.4577
J48	0.001	0.023	99.9003	99.7712	0.0997	0.2288
Random Tree	0.013	0.023	98.8036	99.0847	1.1936	0.9153
Random Forest	0.002	0.000	99.8006	1.000	0.1994	100

TABLE 4. CFS SELECTED FEATURES OF CICIDS2017 DATASET

Classifiers	False Positive Rate		Correctly Classified Instances		Incorrectly Classified Instances	
	DoS	PScan	DoS	PScan	DoS	PScan
Logistic Regression	0.273	0.109	82.4848	91.2074	17.5152	8.7926
Naïve Bayes	0.026	0.429	88.5795	65.6009	11.4205	34.3991
Bayes Net	0.009	0.004	98.2586	99.5064	1.7414	0.4936
J48	0.008	0.002	99.2144	99.7693	0.7856	0.2307
Random Tree	0.008	0.002	99.2226	99.7734	0.7774	0.2266
Random Forest	0.008	0.002	99.2344	99.7738	0.77656	0.2262

III. CONCLUSION AND FUTURE WORK

The system proposed the network traffic by using machine learning and chose the effective features with correlation based feature subset attribute and gain ratio. Comparing of the proposed system and existing dataset CICIDS2017 is to measure the good features of the system. By reducing the unnecessary features and their values, the false positive rate and processing time will be reduced and the system will be effective. In future work, the users can add not only the false positive rates but also other accuracies to prove the good performance of the proposed system.

REFERENCES

[1] K. Kumar, and J. S. Batth, "Network Intrusion Detection with Feature Selection Techniques using Machin-Learning Algorithms", International Journal of Computer Applications, Vol 150, No. 12, 2016.

[2] Kurniabudi, D. Stiawan, and et al. "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection", IEEE, July, 2019.

[3] A. H. Lashkari, G. D. Gil, M. S. I. Mamun, and Ali A. Ghorbani ,"Characterization of Tor Traffic using Time based Features", In Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP 2017), pp. 253-262, 2017.

[4] H. H. Yi, Z. M. Aye, "Security Awareness of Network Infrastructure: Real-time Intrusion Detection and Prevention System with Storage Log Server", The 16th International Conference on Computer Application, 2018, pp. 678-686.

[5] H. H. Yi, Z. M. Aye, "Performance Analysis of Traffic Classification with Machine Learning", International Conference on Information Technology and Electrical Engineering (ICITEE 2021), 2021, pp. 33-38.