Empirical Performance Evaluation Of Feature Selection Approach For Network Intrusion Detection

Vishwa Pratap Singh, Rameshwar Lal Ujjwal

Abstract: Internet use has dramatically increased in the last decade with the advent of efficient and affordable technology. The organizations and the companies are using the Internet to boost up their efficiency and for effective communication but there is always a threat of security breaches that can be disastrous for them. The number of security attacks has been exponentially increased in the last decade and there is a need for efficient intrusion detection systems. The classical rule-based or behavior-based intrusion detection system can identify known attacks but they very inefficient in detecting unknown attacks. Researchers are applying machine learning techniques to detect unknown attacks by using various clustering and classification techniques. In this paper, we have evaluated the performance of feature selection (filter and wrapper method) with Naïve Bayes, OneR, Adaboost, J48 decision tree for detection the attacks. We have used UNSW_NB15 data set and performance evaluations are performed on the basis of Precision, accuracy, recall, MCC and sensitivity. The outcomes of experiments show that the feature selection methods enhance the performance of Bayes, OneR, Adaboost, J48.

Index Terms: Intrusion detection system, Feature selection, Machine learning, NIDS.

1 INTRODUCTION

The use of information technology has been increased by many folds in the last two decades. New efficient and affordable technology lead computer networks to widely used by the general public for personal use as well as industry for professional use. This rapid growth also poses several challenges for researchers like making computer networks more scalable. Security is one of the major challenges as it was not considered from the beginning of computer networks and later added as patches. The numbers of security attacks on networks are also increased by many folds due to easy availability of various hacking tools and involvement of commercial organizations. Attackers are using new technologies to attack networks and classical rule-based and behavior-based techniques are not able to thwart the attacker. Attackers are easily managing to bypass these intrusion detection systems. Intrusion detection systems are basically of two types: Anomaly-based and misuse based. Anomaly-based intrusion detection system[1] detect security attacks on the basis of previously stored (recorded) data but these types of systems lead to high rate false positives makes the system inefficient. Misuse based NIDS uses attacks signatures to detect attack packets. They are low in the false positive rate but also not efficient in detecting attacks. Machine learning has been applied in various fields including NIDS and it is able to give better results than classical systems. Researchers have used various machine learning algorithms to identify attack data.

- *Vishwa Pratap Singhr name is currently pursuing Ph.D. program in computer Science and engineering in GGSIP University, India. E-mail: vishwapratap.phd@gmail.com*
- *R.L. Ujjwal is currently working as Assistant professor in GGSIP University. E-mail: Ujjwal@ipu.ac.in*

In this paper, our main motive is studied and evaluates feature selection method with the use of various parameters. The remaining paper is organized as follows: Section 2 is about previous work done in this field. In section 3 we have discussed various machine learning and feature selection algorithms. Section 4 describes research methodology and section 6 discussed about experimental setup and results. We have analyzed results in greater details in section 7.

2 LITERATURE REVIEW

Many techniques are available in the today's literature for detecting network intrusion. Intrusion detection has gained a lot of interest among the researchers as, in the recent times it is applied for preserving network security. Researchers have carried out lot of works in applying machine learning techniques in various fields. Owens and Levary[2] have proposed to include expert system technology in intrusion detection systems. The solutions that have been presented are within the scope of network anomalies. The solutions related to attack detection focused on examining raw traffic. Few researchers have used support vector machine (SVM) that overtakes other typical machine learning approaches in many areas. Horng et al. [3] has proposed SVM based IDS in which they have combined feature selection; clustering algorithms (hierarchical clustering) used SVM. Three categories have been used mainly: filter, wrapper, and hybrid methods. Other solutions that are based on SNMP were used to detect network intrusion.

3 FEATURE SELECTION METHOD AND MACHINE LEARNING TECHNIQUES

In this section, we have discussed various machine learning and feature selection algorithm proposed by researchers in the past. Machine learning algorithms can be categorized into two categories supervised learning and non-supervised learning. Supervised learning is used to classify the pre-labeled classes

and unsupervised learning does not have such labels. Unsupervised learning is able to identify new types of attacks. Random tree Classifier: It is one among the classifiers. Before implementing this classifier, we should fix the number of trees. In this, a single decision tree is represented by each individual tree. Each tree has selected attributes. That's why this classifier can be treated as a finite set of decision trees. The technique for predicting the dataset is to migrate the outputs from decision trees and select the winner class on the basis of a number of votes J48. Random Forest Classifier: As it is a classification algorithm, its aim is to use the base of the forest and to enhance trees classifiers. The classifiers generated here have a high accuracy rate. This eliminated the re-modification process as well. Navie Bayes Classifier. This classifier represents the group of probabilistic classifiers. It uses Bayes theorem for further problems. Initially, we find the total number of classes and then we calculate the conditional probability. After calculating the conditional probability for each attribute, the standard formula can be applied. This technique has the tendency to work with discreet and continuous attributes also. This can also be represented as a Bayesian network. Using these networks, it was successful to generate the training models for all the selected learning Ross Quilan implemented C4.5[6] algorithm and this classifier improved the implementation. The output expected by this classifier is decision binary trees. These trees are more stable Feature selection and data cleaning are the most important part of the learning process. Feature selection automatically or manually selected features which or correlated and removing the redundant part. Efficient feature selection and reduce the overfitting and increase the efficiency of algorithms. Filter based feature selection method used correction to identify the features which are less useful for the learning process. For correlations, we can use Pearson correlations, LDA, ANOVA or chi-square method [4]. The way for feature selection is a wrapper method in which we use the subset of various features to train the model. Forward feature selection, backward feature eliminations, and recursive feature eliminations are some examples.

4 RESEARCH METHODOLOGY

Intrusion detection is a binary technique which classifies data into attack data and normal data. Intrusion detection using machine learning is a three-step process in which the first step includes acquiring dataset. Data acquisition can be performed by capturing real-time data and then arranged data into categories. The second step is to perform pre-processing of data, which includes the removal of redundant data and analysis of missing values. The third category is to identify whether particular data is attack data or not. The dataset contains many 49 features and effective feature selection can result in better threat identification. We have used the filter-based and wrapper method for feature selection and then applied classification algorithms

4.1 UNSW-NB15 DATASET

This section describes the dataset used for the analysis in this paper. We have used UNSW-NB15 [5] dataset that was generated by IXIA. The dataset contains a total of 49 features. 47 features out of 49 are related to IDS and the remaining 2 attacks category related. The features of the dataset are divided into five categories namely; flow feature, basic feature, content feature, time feature, and additional generated features. This dataset includes nine categories of security attack data (analysis, backdoor, DoS, exploits, fuzzes, generic, probe, shellcode, worm) with normal data. The brief description of the dataset is given in Table I.

4.2 EVALUATION OF MATRIX PARAMETERS

We have used five parameters (Sensitivity, Precision, Recall, MCC, Accuracy (%)) for evaluating learning techniques which are evaluated with and without using feature selection method.

TABLE 2
USED ACRONYMS

Positive	***-***	***P***
Negative	***-***	***N***
True Positive	***-***	***TP***
True Negative	***-***	***TN***
False Positive	***-***	***FP***
False Negative	***-***	***FN***
False-positive rate	***-***	***FPR***

Precision: It measures the accuracy of the system. It is also known as a positive predictive value which defines how much prediction is correct.

$$\Pr ecision = \frac{TP}{TP + FP}$$

Recall (true positive rate): Measurement of True positive that are actually identified correctly.

$$\mathrm{Re}\, call = \frac{TP}{TP + FN}$$

Specificity: Measurement of Actual negative correctly identified

$$Specificity = \frac{TP}{FN + TN}$$

False Positive Rate: It is the rate of negative events identified as true.

$$FPR = \frac{FP}{FP + TN}$$

Mathews Correlation Coefficient: MCC is widely used in various filed as a measure of the quality of multiclass or binary classification. MCC is regarded as a balanced measure which takes into account TP and FP and N. It is basically a correlation between target and prediction.

$$MCC = \frac{(TP + TN) - (FP + FN)}{\sqrt{((TP + FP)(TP + FN)) + (FN + TN)}}$$

Accuracy: Accuracy tells about how much predictions are

correctly identified.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

TABLE II *FEATURES AND CATEGORIES OF UNSW-NB15 DATASET*

#	Feature	Description	Category
1.	Srcip	Source IP address	Flow Features
2.	Sport	Source port address	
3.	Dstip	Destination IP address	
4.	Dsport	Destination port address	
5.	Proto	Transaction protocol	
6.	State	The state and its dependent protocol, e.g.,	Basic Feature
7.	Dur	Record total duration	
8.	Sbytes	Source to destination bytes	
9.	Dbytes	Destination to source bytes	
10.	Sttl	Source to destination time to live	
11.	Dttl	Destination to source time to live	
12.	Sloss	Source packets retransmitted or dropped	
13.	Dloss	Destination packets retransmitted or dropped	
14.	Service	http, ftp, smtp	
15.	Sload	Source bits transfer rate	
16.	dload	Destination bits transfer rate	
17.	spkts	Number of packets transferred from source	
18.	Dpkts	Number of packets transferred from destination to source	
19.	Swin	Value of window advertisement of tcp at Content features	
19.	Swin	Value of window advertisement of tcp at Content features	Content Feature
20.	Dwin	Value of window advertisement of tcp at destination	
21.	Stcpb	tcp sequence number at source	
22.	Dtcpb	tcp sequence number at Destination	
23.	Smeanz	Mean value of packet size transferred by source	
24.	24 Dmeanz	Mean value of packet size transferred by destination	
25.	25 Trans_depth	Pipelined depth of http request/response transaction in a connection	
26.	Res_bdy_len	Size of data transferred by servers http	
27.	Sjit	Jitter produced at source Time feature	Time Feature
28.	Djit	Jitter produced at destination	
29.	Stime	Start timestamp	Additional Generated Features
30.	Ltime	Ltime Last timestamp	
31.	Sinpkt	ime gap between two consecutive incoming packet at source	
32.	Dintpkt	Time gap between two consecutive incoming packet at destination	
33.	Tcprtt	Round trip time taken for tcp connection establishment	
34.	Synack	Time gap between SYN and SYN_ACK	
35.	Ackdat	Time gap between the syn_ack and ack packets	
36.	Is_sm_ips_ports	If source and destination are same and port numbers are equal	
37.	Ct_state_ttl	No. of connection for every state based on the specific range	
38.	Ct_ftw_http_mthd	No. of flows having methods such as Get and Post in http service	
39.	Is_ftp_login	If the ftp session is accessed by user and password then this feature is set as 1 else 0	
40.	Ct_ftp_cmd	No. of flows that has a command in ftp	
41.	Ct_srv_	No. of records out of 100 records having same service and source based on the last time stamp	
42.	Ct_srv_dst	No. of records out of 100 records, having same service and destination based on the last times stamp	
43.	Ct_dst	No. of records out of 100 records, having same destination based on the last time stamp	
44.	Ct_src_itm	No. of records out of 100 records having same source based on the last time stamp	
45.	Ct_src_dsport_ltm	No. of records out of 100 records having same source and the destination port based on the last time stamp	
46.	Ct_dst_sport_ltm	No. of records out of 100 records having same destination and the source port based on the last time stamp	
47.	Ct_dst_src_ltm	No. of records out of 100 records having	

5 EXPERIMENT SETUP

The number of attack cases is small in number in comparison to the whole data set so we have used 10 fold cross-validation approaches. The algorithms classify the attack in classes as mentioned but we have used gross values after classification for the sake of simplicity. We have used Naïve Bayes[7], OneR[8], Adaboost[9], J48[10] for classification purposes and then we use filter method for the feature selection. Information gain approach has been applied in the experiments and we have removed the attributes that have very low or no correlation. After features selection, we used the same learning algorithms and stored the results in table 3. The outcome of this method shows that the accuracy and performance of the system are increased after using feature selection and removing non related attributes.

6 RESULT ANALYSIS

The dataset consists of 45 features. Dataset was complete with no missing values and errors, so it does not require data cleaning. In our experiment with 4 attributes, the J48 algorithms perform better than others. And then have analyzed the result after selecting features and remove them before performing the learning process again. The overall efficiency of all algorithms got better after feature selection and again j48 works best for this experiment.

7 CONCLUSIONS

In this study, we have applied machine learning algorithms on UNSW_NB15 dataset to identify various attacks. The result we got after our experiments shows that the feature selection method is better than learning algorithms without removing non correlated features.

8 REFERENCES

[1] Kaushik, Sapna S., and P. R. Deshmukh. "Detection of attacks in an intrusion detection system." International Journal of Computer Science and Information Technologies (IJCSIT) 2, no. 3 (2011): 982-986.

[2] Alkasassbeh, Mouhammd. "An empirical evaluation for the intrusion detection features based on machine learning and feature selection methods." arXiv preprint arXiv:1712.09623(2017).

[3] Horng, Shi-Jinn, Ming-Yang Su, Yuan-Hsin Chen, Tzong-Wann Kao, Rong-Jian Chen, Jui-Lin Lai, and Citra Dwi Perkasa. "A novel intrusion detection system based on hierarchical clustering and support vector machines." Expert systems with Applications 38, no. 1 (2011): 306-313.

[4] Yitzhaky, Yitzhak, and Eli Peli. "A method for objective edge detection evaluation and detector parameter selection." IEEE Transactions on pattern analysis and machine intelligence 25, no. 8 (2003): 1027-1033.

[5] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." In 2015 military communications and information systems conference (MilCIS), pp. 1-6. IEEE, 2015.

TABLE 2

UNITS FOR MAGENTIC PROPERTIES

Parameters	*Classifiers*			
	Naïve Bayes	*Adaboost*	*J48*	*OneR*
Sensitivity	*0.452*	*0.633*	*0.893*	*0.763*
Specificity	*0.971*	*0.72*	*0.989*	*0.959*
Precision	*0.743*	*0.668*	*0.893*	*0.739*
MCC	*0.485*	*0.774*	*0.881*	*0.401*
Accuracy	*45.2121 %*	*63.2731 %*	*82.23 %*	*76.268 %*
Kappa statistic	*0.3451*	*0.3957*	*0.8501*	*0.6632*

TABLE 3

EXPERIMENT OUTCOME AFTER FEATURE SELECTION USING FILTER METHOD

Parameters	*Classifiers*			
	Naïve Bayes	*Adaboost*	*J48*	*OneR*
Sensitivity	*0.458*	*0.633*	*0.893*	*0.763*
Specificity	*0.971*	*0.72*	*0.989*	*0.959*
Precision	*0.743*	*0.722*	*0.893*	*0.723*
MCC	*0.488*	*0.498*	*0.881*	*0.401*
Accuracy	*45.8352*	*63.273*	*89.267*	*76.268*
Kappa statistic	*0.3516*	*0.3957*	*0.8503*	*0.6632*

[7] Chen, Jingnian, Houkuan Huang, Shengfeng Tian, and Youli Qu. "Feature selection for text classification with Naïve Bayes." Expert Systems with

[8] Applications 36, no. 3 (2009): 5432-5435.

[9] Soman, Thara, and Patrick O. Bobbie. "Classification of arrhythmia using machine learning techniques." WSEAS Transactions on computers 4, no. 6 (2005): 548-552.

[10] Schapire, Robert E. "The boosting approach to machine learning: An overview." In Nonlinear estimation and classification, pp. 149-171. Springer, New York, NY, 2003.

[11] Firdausi, Ivan, Alva Erwin, and Anto Satriyo Nugroho. "Analysis of machine learning techniques used in behavior-based malware detection." In 2010 second international conference on advances in computing, control, and telecommunication technologies, pp. 201-203. IEEE, 2010.