An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms

Rohit Kumar Singh Gautam
CSE, NITTTR
Chandigarh, India
2rsingh@gmail.com

Er.Amit Doegar
CSE, NITTTR
Chandigarh, India
Amit@nitttrchd.ac.in

Abstract— Countering network threats, especially intrusion detection (ID), is an exigent field of research in the area of data security. The primary research problem of IDS from the research concerns is optimizing its efficiency that receives increasingly attention. The chance from spammers, attackers & crook organizations has grown up with the enlargement of net, hence, IDS grew to be a core part of digital network for the reason that of incidence of such threats. This research performs three arrangements of examinations. From the major investigation, the frameworks are ready using the entire 41 highlights. The second trial where we perform feature selection through making use of Entropy based analysis as Filter Method to decide upon the satisfactory factors (Rank) as opposed to utilizing all the 41 entails and play out the trial with Naive Bayes, Adaptive Boost and PART(Partial Decision Tree) and believe concerning the results [11]. The third analysis where we perform Ensemble Approach by means of making use of Information gain as to opt for the fine components as opposed to using all the 41 entails and play out the trial with Naive Bayes, Adaptive Boost and PART, and examine the effects [11, 23].

Keywords: IDS (Intrusion Detection System); Machine learning; KDDcup99; Dataset; Feature Selection; Ensemble Approach.

I. INTRODUCTION

IDS are defined as a software utility that detects system activities for hazardous movements and generates experiences to management. An Intrusion Detection System is security counter step to recognize set of intrusion that compromises the acquaintance, availability, and integrity of data sources [1]. The groups are using IDS with aim to identify problems with security policies and documenting existing threats.

Due to the advancement of internet technologies, application, and protocol, traffic analysis of network have become more immense since it conception with the maximum amount of network traffic data. In HIDS (Host based Intrusion Detection System), anti-hazard application software such as antivirus software, firewall & spyware detection programme installed on each computer which is connected over a network that has two way access to the outward environment such as the internet. A snapshot of process documents is taken by it and compared it with the earlier taken snapshot. If we when put next it with firewall, nonetheless they both establish with protection, IDS framework varies from firewall. Firewall constrains approach between techniques to prevent interception and do not flag an assault from throughout the procedure. An IDS, analyze a suspected interference as soon as it has happened and flags a warning. A substructure that ends associations is called an interference counteractive action substructure. Today, IDS [4, 6, 9] has become the need of nearly every organization. It helps to record information associated with detected actions, and alert security administrator and produce reports [12]. This system constantly monitors network for any abnormal activity.

An ID is mainly of two types i.e. Network-IDS and Host-IDS. In NIDS, anti-threat software is installed only at specific instance such as servers that provide communication between
System attack is as a rule characterised as an interruption to your method base a good way to first break down your surroundings and acquire knowledge with a exact finish goal to abuse the present open ports or vulnerabilities - this may increasingly comprise unapproved access to your assets too [13].
Passive attacks are in nature of roof dropping on, or checking of transmission. Inactive assaults contain exercise examination, checking of unprotected correspondences, unscrambling feebly encoded action, and catching confirmation information, for example, passwords [18].
Active attack includes some alteration of the information stream or formation of the false stream.

II. PROPOSED ALGORITHM

The methodology of designing the proposed scheme is divided into three phases: Normalization, Feature Selection and Ensemble Method.

A. Normalization

In the primary analysis, the frameworks are prepared utilizing all the 41 features and reduced the dimensionality of dataset, the following steps takes place as shown in Figure 1.

Step1: KDDcup-99 dataset with 41 features.

Step 2: Normalization of Data set with help of formula.

$$Z' = \frac{Z- min_A}{max_A - min_A} \tag{1}$$

978-1-5386-1719-9/18/$31.00 ©2018 IEEE

Where z’ is normalized value and z is initial value. Max and min value for attribute A before normalization.

Step3: Feature Selection by Information Gain.

Step4: Input the feature & labels into Naïve Bayes, PART & Adaptive boost & make three models.

Step5: Perform the test on these models and calculate the precision, recall & accuracy.

B. *Feature Selection*

The Filter Method uses statistical approaches to provide a ranking to the features. After providing the ranking user can choose best k feature for the experiment like top 24or 41 according to their requirement. third analysis where we perform feature selection by means of making use of Information Gain as to opt for the fine components as opposed to using all the 41 entails and play out the trial with straight Naïve Bayes, PART and Adaptive Boost and examine the effects. We use correlation method because it represents the distribution based similarity of features with reducing the unbalancing of features in form of distribution. In this feature selection phase, the following steps take place as shown in the Figure 1.

Step1: KDD-99 dataset with 41 features.

Step2: Normalization of Data set with help of formula.

Step3: Feature Selection by Information Gain.

Step4: Input the feature & labels into Naïve Bayes, PART & Adaptive Boost and make three models.

Step5: Perform the test on these models and calculate the precision, recall & accuracy.

C. *Ensemble Approach*

Ensemble Approach is performed in the third analysis by utilizing to choose some best components as opposed to utilizing all the 41 include and plays out the trial with Naïve Bayes, PART and Adaptive boost and analyze the outcomes [2]. We use Bagging method because it represents the distribution based similarity of features with reducing the imbalancing or biasing and variance of features in form of distribution. An ensemble approach we combined the result of multiple classifier using average or majority of voting. The Bagging helps to reduce the variance error. The Bagging is also work as bootstrap aggregation. In bootstrapping we choose ‘n’ observations or n-row of original dataset. But main key factor is replacement of each row is selected original dataset so that each row is equally likely to be selected in the every iteration. It is similar to sampling Technique. In the ensemble approach phase the following steps takes place as shown in Figure 1.

Step1: KDD-99 dataset with 41 features.

Step2: Normalization of Data set with help of formula.

Step3: Feature Selection by Information Gain.

Step4: Input the feature & labels into Naïve Bayes, PART &Adaptive Boost and make Ensemble Model.

Step5: Perform the test on Ensemble models & calculate the precision, recall & accuracy.

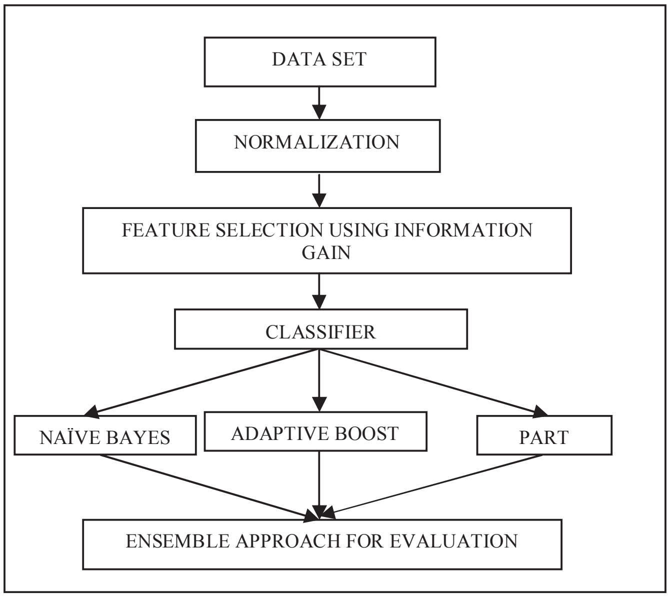
DATA SET
NORMALIZATION
FEATURE SELECTION USING INFORMATION GAIN
CLASSIFIER
NAÏVE BAYES
ADAPTIVE BOOST
PART
ENSEMBLE APPROACH FOR EVALUATION

Figure 1. General Methodology

III. EXPERIMENT AND RESULT

A. *Discription of Dataset*

Data set KDDcup 99 are used to carry out the experiment. It was once created headquartered on the Defence developed research undertaking company DARPA based on intrusion detection analysis software [19]. They simulated computer network operated as associate usual setting that used to be contaminated by using quite a lot of varieties of attacks. The uncooked facts set turned into processed into connection files. For every connection, forty one more than a few features had been extracted. Each and every connection was labelled as traditional or below exact kind of assault. There are 24 attacker forms that could be labelled into 4 important categories which summarized in Table 1. There are four important classes of assaults described below:-

1) Denial of Service Attack (DOS) : It prohibits the users by closing the network or process else by rejecting the accessibility of resources. DoS attack is categorized into two parts: the First one is operating system attacks which focus on bugs and fix them with patches [3]. The second type of attack is network attacks which put the basic restriction on networking protocol e.g. Smurf, Teardrop, SYN Flood.

2) User to Root Attack (U2R): This attack occurs when an attacker tries to get illegal eruption to the root level of the target machine. The most common User to root attack is buffer overflow where attacker capitalized on the program fault and assembled additional data into a buffer that is supposed to be kept on an execution stack e.g. Buffer_overflow.

TABLE -1 ATTACKS TYPE IN KDDcup99 DATASET

Types of Attacks	Attacks Pattern
Denial of Attacks	*Back, Land, Neptune, Pod, Smurf, Teardrop*
Probe	*Ipsweep, nmap, portsweep, satan*
Root to Local	*Fp_write, guess_password, imap, multihop, phf, spy, warezclient, warezmaster*
User to Remote	*Buffer_overflow, loadmodule, perl, rootkit*

3) Remote to Local Attack (R2L): At the point, when an attacker gets the access as a user of the system or as the root from the remote system through the network that type of attack known as the Root to Local attacks. The guessing Password is most common Remote to local attacks.(e.g. guest and dictionary attack).

4) Probing Attack (PROBE): In this type of attacks, the attacker tries to collect all information by analyzing the network traffic to bypass the security management. An intruder tries to fetch the valid IP address to find out different services used and which platform is used such as operating system e.g. Ipsweep, nmap, portsweep, satan.

B. Performance Matrics

The important performance parameters chosen to analyze the results are:

1. Precision
2. Recall
3. Accuracy

Precision

Precision is positive predicted value. Precision gives ratio of true positive and predicted condition positive. It is calculated as proportion of true positive from all positives and is calculated as:

$$Precision = \frac{TP}{TP+FN} \tag{2}$$

Recall

Recall determine how much useful data is access from any machine learning algorithm. It target on the important information.

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Accuracy

Accuracy is one of the primary assess for evaluating the performance of any algorithm. Accuracy is the propinquity of analysis results to the true value and is calculated by following formula:

$$Accuracy = \frac{TN+TP}{TP+TP+FN+FP} \tag{4}$$

Where TN (true negative), TP (true positive), FP (false positive) and FN (false negative)

C. Experimental Setup

The experiment used to be performed KDDcup99 [20]. Linux machine is used to carry out experiment having configuration 8GB RAM and Processor Intel® Core™ i7-4790CPU@3.60GHz*8. Now, we have used R programming language and Weka tool.

D. Results

In this section, we reward a performance evaluation for some supervised learning methods. Here we used good identified KDD Cup99 data [20] to make important investigations for network anomaly. In this we perform three arrangements of trials. In the primary analysis, the frameworks are prepared utilizing all forty one features. Feature selection is performed in second investigation by utilizing Information Gain as to choose the best components as opposed to utilizing all the 41 includes and play out the trial with Naïve Bayes, PART and Adaptive Boost and analyze the outcomes. Ensemble method is performed in the third analysis by voting to choose some best components as opposed to utilizing all the 41 includes and play out the trial with Naïve Bayes, PART and Adaptive Boost and analyze the outcomes.

1) Performance Measure with all 41 Features Results : We first compared the performance of three classification schemes, namely Naïve Bayes, PART and Adaptive Boost. Table 2 illustrates the performance of Naïve Bayes, PART and Adaptive Boost algorithms for KDDcup99 data set for all 41 features without applying feature selection. The result showed by PART has significant by 41 features. Also Adaptive Boost showed better results in comparison Naïve Bayes.

TABLE-2 PERFORMANCE EVALUATION WITH ALL 41 FEATURES FOR KDDcup99 DATA SET [20]

Classifier	Accuracy	Precision	Recall
Naïve Bayes	*92.7840*	*98.90*	*92.70*
PART	*99.9601*	*99.96*	*99.90*
Adaptive Boost	*97.8597*	*96.20*	*97.90*

2) Performance Measure for Feature using Information Gain : Outcome of feature selection utilizing Information Gain confirmed that PART and Adaboost are most efficient for detecting assaults than Naïve Bayes as proven in Table 3.

TABLE-3 PERFORMANCE EVALUATION WITH APPLYING FEATURE SELECTION USING INFO GAIN FOR KDDcup99 DATA SET [20]

Classifier	Accuracy	Precision	Recall
Naïve Bayes	*91.9818*	*98.90*	*92.00*
PART	*99.9589*	*99.90*	*99.60*
Adaptive Boost	*97.9073*	*96.20*	*97.90*

3) Performance Measure for Ensemble Approach using Bagging method : Outcome of Ensemble Method utilizing

Information Gain showed that Ensemble Approach has accuracy rate higher than PART, Adaptive boost and Naïve Bayes as proven in Table 4.

TABLE-4 EFFICIENCY ANALYSIS WITH MAKING USE OF ENSEMBLE APPROACH USING BAGGING FOR KDDcup99 DATA SET [20]

Classifier	Accuracy	Precision	Recall
Naïve Bayes	***91.9818***	***98.90***	***92.00***
PART	***99.9589***	***99.90***	***99.60***
Adaptive Boost	***97.8597***	***96.20***	***97.90***
Ensemble Approach	***99.9732***	***99.99***	***99.98***

IV. CONCLUSION & FUTURE WORK

This paper emphasis on make a optimize classifier model for two classes attack and not attack but problem is data imbalance, which is improved by Ensemble Approach. We perform three arrangements of examinations. From the major investigation, the frameworks are ready using the entire 41 highlights. The second trial where we perform feature selection through making use of Information Gain as to decide upon the satisfactory factors as opposed to utilizing all the 41 entails and play out the trial with Naïve Bayes, PART and Adaptive boost and believe concerning the results. The third analysis where we perform Ensemble Approach by means of making use of bootstrapping as to opt for the fine components as opposed to using other classifier entails and play out the trial with Naïve Bayes, PART and Adaptive Boost and examine the effects [23]. These results conclude average performance of Ensemble Approach better than other classifiers.

There is still scope of improvements to propose systems which are able to detect all types of attacks & can reduce the feature set by feature selection and Ensemble Method with the help of different classifier.

References

[1] C. Guo, Y. Ping, N. Liu, and S. S. Luo, “A two-level hybrid approach for intrusion detection,” Neurocomputing, vol. 214, pp. 391–400, 2016.

[2] A. A. Aburomman and M. Bin Ibne Reaz, “A novel SVM-kNN-PSO ensemble method for intrusion detection system,” Applied Soft Computing Journal, vol. 38, pp. 360–372, 2016.

[3] S. O. Al-mamory and F. S. Jassim, “On the Designing of Two Grains Levels Network Intrusion Detection System,” Karbala International Journal of Modern Science, Elsevier, vol. 1, pp. 15–25, 2015.

[4] W. Bul’ajoul, A. James, and M. Pannu, “Improving network intrusion detection system performance through quality of service configuration and parallel technology,” Journal of Computer and System Sciences, vol. 81, pp. 981–999, 2015.

[5] K. Zheng, Z. Cai, X. Zhang, Z. Wang, and B. Yang, “Algorithms to speedup pattern matching for network intrusion detection systems,” Computer Communications, vol. 62, pp. 47–58, 2015.

[6] S. Rastegari, P. Hingston, and C. P. Lam, “Evolving statistical rulesets for network intrusion detection,” Applied Soft Computing Journal, vol. 33, pp. 348–359, 2015.

[7] J. Cervantes, F. García Lamont, A. López-Chau, L. Rodríguez Mazahua, and J. Sergio Ruíz, “Data selection based on decision tree for SVM classification on large data sets,” Applied Soft Computing, vol. 37, pp. 787–798, 2015.

[8] M. S. Gondal, A. J. Malik, and F. A. Khan, “Network Intrusion Detection using Diversity-based Centroid Mechanism,” 12th International Conference on Information Technology - New Generations, pp. 224–228, 2015.

[9] A. Dastanpour and A. Selamat, “Comparison of Genetic Algorithm Optimization on Artificial Neural Network and Support Vector Machine in Intrusion Detection System,” IEEE Conference on Open Systems (ICOS), pp. 72–77, 2014.

[10] Y. Choi, D. H. Kim, K. N. Plataniotis, and Y. M. Ro, “Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography,” Expert Systems with Applications, vol. 46, pp. 106–121, 2016.

[11] C. A. Ronao and S. B. Cho, “Anomalous query access detection in RBAC-administered databases with PART and PCA,” Information Sciences, vol. 369, pp. 238–250, 2016.

[12] R. A. R. Ashfaq, X. Z. Wang, J. Z. Huang, H. Abbas, and Y. L. He, “Fuzziness based semi-supervised learning approach for intrusion detection system,” Information Sciences, vol. 378, pp. 484–497, 2017.

[13] M. Stevanovic and J. M. Pedersen, “An efficient flow-based botnet detection using supervised machine learning,” 2014 International Conference on Computing, Networking and Communication, pp. 797–801, 2014.

[14] A. Karim, R. Salleh, M. Shiraz, S. Shah, I. Awan, and N. Anuar, “Botnet detection techniques: review, future trends, and issues,” Computer and Electronics, vol. 15, pp. 943–983, 2014.

[15] A. Feizollah, N. B. Anuar, R. Salleh, and A. W. A. Wahab, “A review on feature selection in mobile malware detection,” Digital Investigation, vol. 13, pp. 22–37, 2015.

[16] A. Karim, S. Adeel, A. Shah, R. Bin Salleh, M. Arif, and R. Noor, “Mobile Botnet Attacks – an Emerging Threat : Classification , Review and Open Issues,” TIIS 9, vol. 9, pp. 1471–1492, 2015.

[17] S. Garasia, D. Rana, and R. Mehta, “Http Botnet Detection Using Frequent Patternset Mining,” Intl. Journal of Engineering Science and Advanced Technology (IJESAT), pp. 619–624, 2012.

[18] C. Livadas, R. Walsh, D. Lapsley, and W. T. Strayer, “Using Machine Learning Techniques to Identify Botnet Traffic,” Local Computer Networks, Proceeding. 2006 31st IEEE Conference, pp. 967–974, 2006.

[19] I. Mohammad, R. Pandey and A. Khatoon, “A Review of types of Security Attacks and Malicious Software in Network Security,” International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol. 4, pp. 413–415, 2014.

[20] KDD Cup 1999 Intrusion Detection Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html[Accessed on: Feburary 2017]

[21] M. Xu, N. Ye, “Probabilistic networks with undirected links for anomaly detection”, In Proceedings of IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop, pp. 175–179, 2000.

[22] E. B. Beigi, H. H. Jazi, N. Stakhanova, and A. A. Ghorbani, “Towards effective feature selection in machinelearning-based botnet detection approaches,” in IEEE Conference on Communications and Network Security (CNS), pp. 247–255, 2014.

[23] K. Singh, S. C. Guntuku, A. Thakur, and C. Hota, "Big data analytics framework for peer-to-peer botnet detection using PARTs," Information Sciences, vol. 278, pp. 488-497, 2014.