AN ENHANCED CLASSIFICATION APPROACH FOR NETWORK INTRUSION DETECTION USING HOEFFIDING INDUCTION TREE ALGORITHM

M.Deepa[1], Dr.P. Sumitra[2],
[1]Ph.D Research Scholar, Department of Computer Science,
legithasai2010@gmail.com
[2]Professor, Department of Computer Science,
drsumitra@vicas.org
[1,2] Vivekananda College of Arts and Sciences for Women (Autonomous),
Elayampalayam

ABSTRACT

Data mining is now used by many institutions widely and generally. Intrusion detection for network operators & security specialists is one of the top priorities and challenges. Sensitive data, anonymity and device availability from attacks are protected by the Intrusion detection system. In order to describe resources from those in the database through a network, IDS uses data mining techniques. A robust algorithm must also be built to produce successful rules for the detection of attacks. In this paper, optimization algorithms focused on classification were used to detect attacks over the NSL KDD dataset. Depending on this stranglehold, the current method is explained an improved Hoeffiding Induction Tree algorithm to resolve the drawbacks. The results demonstrate that the proposed HITNB algorithm has improved precision, a lower alarm rate and the ability to detect a new type efficiently.

Keywords: Intrusion Detection System, Data mining, optimization techniques, classification algorithm, NSL-KDD dataset

1. INTRODUCTION

The use of the computer system now spreads rapidly and gives details quickly. Safety offences such as hackers, viruses, worms etc., spread even more quickly throughout the network. Since the Attacker will threaten the security, credibility and resources of the network. Firewall has been a gateway for protective initiatives for the last few years. However, the network traffic that the certain port or valid user port has done is not detected. Intrusion detection[1] is therefore important in order to control the manipulation of data by ransom ware. Intrusion detection is carried out to determine and supervise the tasks performed either by individual device or by the network. Our primary incentive is to defend the device from attacks only through the intrusion detection system. The device was originally coined in a scientific paper by Anderson (1980) [2]. The monitoring device serves as an effective instrument against attacks and threats. It tracks the user’s actions & discovers consequent masqueraders who dishonestly accessed the system. In the intrusion detection method, the three significant phases are,

- Track and analyze traffic on networks
- Analyze the irregular behaviors if any occur.
- Raising alarms to deal with the situation.

A passive approach is the intrusion detection system. It basically tracks network or hosts details which generates warnings when there is an intrusion. However, data mining techniques may classify these data as they appear and predict them alone by acquiring a successful approach feature. Data mining was generally accepted as a valuable way to evaluate valuable data in vast quantities of noisy, fluffy and random data [3]. Generated patterns may be used to strengthen business strategy such as sales, promotions and clients. Two types of IDS are Host based IDS and Network based IDS depending on the approach of network. In HIDS [4] the data comes from the audit reports, device logs, application software, etc. The encrypted messages get it from the system files over the network and then decode them in a host. There are no data compromised and no particular hardware needs to be mounted on a particular host other than a monitoring device. Usually one intrusion detection device is appropriate for the entire LAN in networked ids [5]. It is inexpensive and willing to afford. Many attacks like DoS, DDoS etc. are analyzed but HIDS doesn't analyze them. Primarily, the intrusion detection system has two categories: anomaly detection and abuse / signature detection. The breach monitoring relates the incoming network activity to the known attack database by means of intrusion detection signatures. It functions efficiently in the study of identified attacks contained in the database. But new attacks that are not predefined cannot be detected. The anomaly detection method, on the other hand, generates the network-based profile and hosting and raises alarms and notifies the administrator in order to control the situation. Yet new and rare attacks can be observed. In evaluating any discrepancies from regular pattern, there are two kinds of false alarms that are false positive and false negative. The key objective is to keep warnings to a minimum. Intrusion detection was carried out using data mining techniques including association, grouping, clustering and neural networks [6, 7].

This article is structured as follows. Section 2 offers similar work on an intrusion detection data mining approach. The data set used and its characteristics are explained in Section 3. The classification algorithm is compared to the proposed one based on optimization techniques. The design of the work proposed is explained in Section 5. Paragraph 6 deals with measuring results. Paragraph 7 contains the experiment-based findings & discussion. The conclusion & future improvement is described in Section 8.

2. RELATED WORKS

The group of seven non - parametric selection techniques was presented by Jundong et al.[8]; named first order statistic selection (FOS). Three classifiers are used in their work: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). SVM is a common classifier of linear discriminates, LR is a simple and efficient regression model, and RF is a powerful classifier based on an ensemble. RF was introduced by the Weka data mining method. The classifiers measure the classification ability of 7 FOS techniques. Major contribution of their work that links the related techniques of selection into one family to analyze them.

In this research, an algorithm was proposed in the Novel Feature Selection (NFS) [9]. In order to ensure optimum accuracy in classification, the NFS algorithm extracts more important fitments.

In datasets with 13 attributes, the NFS algorithm is implemented and selects 6 best features. To test their methodology they conducted various experiments. The neural network predicts 93% of the accuracy of all studies, with SMO predicting 89% by using NFS. Therefore, their analysis showed that this method leads to a superior selection process in order to reduce the sum of the required variable and improve the accuracy of classifying in order to better predict.

Another NFS method called SIP-FS was suggested in [10]. Consistency and error rate have been enhanced by the proposed approach without losing predictability. First, the two key contributions to their work were generalized correlations rather than shared knowledge. Second, in SIP-FS, stability restriction was introduced to pick consistent ranking results in the case of data variation. Therefore, for data representation, the SIP-FS algorithm effectively selects a rational and compact function subset. Their approaches concentrate on filtering techniques based on various assessment metrics, such as distance criterion, entropy criterion, coefficient of determination, accuracy and mutual knowledge.

Miao et al . [11] suggested MD (maximising the difference) to evaluate datasets for customer feedback. As the new type of word-of - mouth (WOM) knowledge developed rapidly, online customer reviews have played an important role in consumer buying decisions.A new method of feature selection called Peculiar Genes Selection (PGS) was recently proposed [12]. The proposed approach has improved the performance classification of imbalanced data sets. In the proposed process, the choice of features is carried out in three steps. First, the detection of differentially expressed genes according to experimental conditions. Secondly, the low discriminative power features are filtered out. Select a good role for each class in the third stage. Using SVM as a classifier, they proposed a supervised approach.

The primary goal of this study is the study of the earlier and well-known problem of network intrusion and its approaches to present application of systems in the sense of network intruding detection systems by seeking the improved classification approach via HITNB. By organizing the data into attack and natural the HITNB algorithm reduces data size with a standard time interval, increased precision and lower false alarm rate. Also it shows that, due to its features in terms of reduction by inductive tree, the HITNB algorithm is of great importance in the intrusion detection region of the network.

3. NSL-KDD DATASET

Several datasets are used for identifying Intrusions on the network. some of which are self-created datasets and some of which are publicly accessible. Some of the publicly accessible datasets are

- The MIT Lincon Laboratories developed DARPA datasets (1998, 1999 and 2000). By adding manually generated network-based attacks, the dataset is generated. The dataset is completed in 5 weeks and data is used for the purpose of research for the last two weeks.
- The intrusion data for KDD 99 is extracted from the DARPA 98 dataset. The dataset includes 41 attributes and one additional class attribute.
- The NSL-KDD dataset is offline network data based on the dataset KDD 99[9].

The suggested technique was employed to the NSL-KDD dataset with 41 attributes and one feature class. The size of NSL-KDD dataset is smaller than KDD99 because KDD99 includes more redundant records. The NSL-KDD training set does not contain redundant records and thus reduces the level of complexity [13]. Training is carried out on NSL-KDD Train data containing 22 types of attacks, and testing is carried out on NSL-KDD Test data containing 17 additional types of attacks. These attacks may be categories with certain common characteristics in four different forms, as shown for training and testing in Table I. The four attack groups are:

- Denial of Service (DoS): A malevolent action to prevent computer network systems and services.
- Probe : This attack gathers information on the target system 's possible weaknesses that can be used to initiate attacks against the target system later.
- Remote to Local (R2L): Unauthorized ability to dump data packets to a remote device over a network and obtain access either as a user or as a root to conduct their unauthorised operation.
- User to Root (U2R): In this situation, attackers enter the device as a regular user and crack vulnerabilities to obtain administrative privileges.

Table 1: TRAIN & TEST DATA SET WITH 4 CLASS OF ATTACKS

DOS	PROBE	U2R	R2L
apache2	Ipsweep	bufferoverflow	*Spy*
Back	**mscan**	loadmodule	*Warezclient*
Land	nmap	perl	ftp_write
Mailbomb	**saint**	**ps**	guesspasswd
Neptune	satan	rootkit	**httptunnel**
Pod		**snmpguess**	imap
Processtable		**sqlattack**	multihop
Smurf		**worm**	phf
Teardrop		**xterm**	**sendmail**
Udpstorm			**snmpgetattack**
			warezmaster
			xlock
			xsnoop

Table 1 describes different types of attacks in the NSL-KDD dataset. In this table all attacks appear in both training and testing dataset, but the attacks in bold letter only present in testing dataset not in training set and attacks in italic present only in train set not in testing set. The NSL-KDD dataset includes 148517 total instances further it can be divided into 125973 training models and 22544 test sets.

4. ARCHITECTURE OF PROPOSED WORK

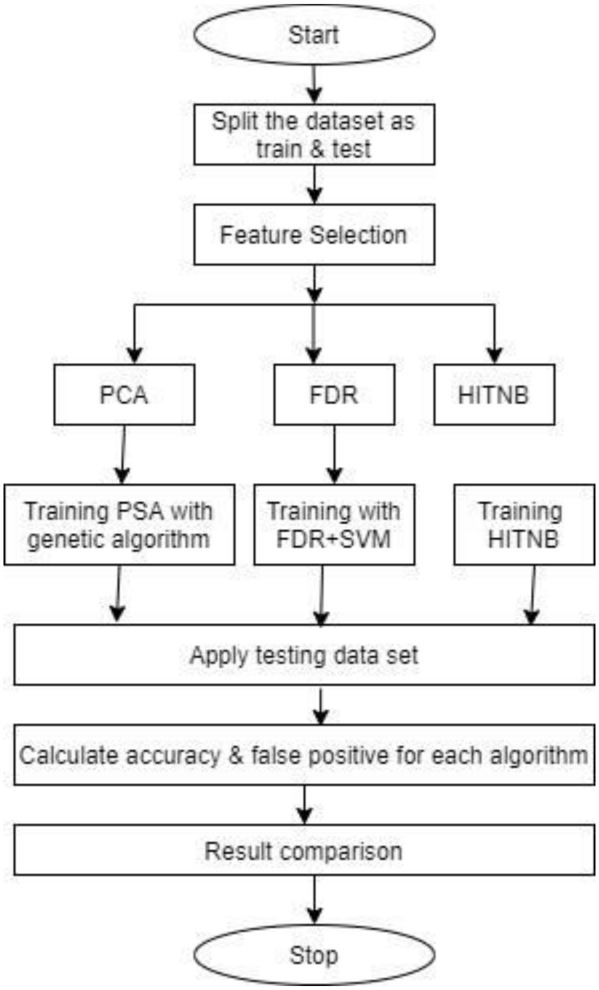

Start
Split the dataset as train & test
Feature Selection
PCA
FDR
HITNB
Training PSA with genetic algorithm
Training with FDR+SVM
Training HITNB
Apply testing data set
Calculate accuracy & false positive for each algorithm
Result comparison
Stop

Figure 1 : System flow of proposed HITNB algorithm

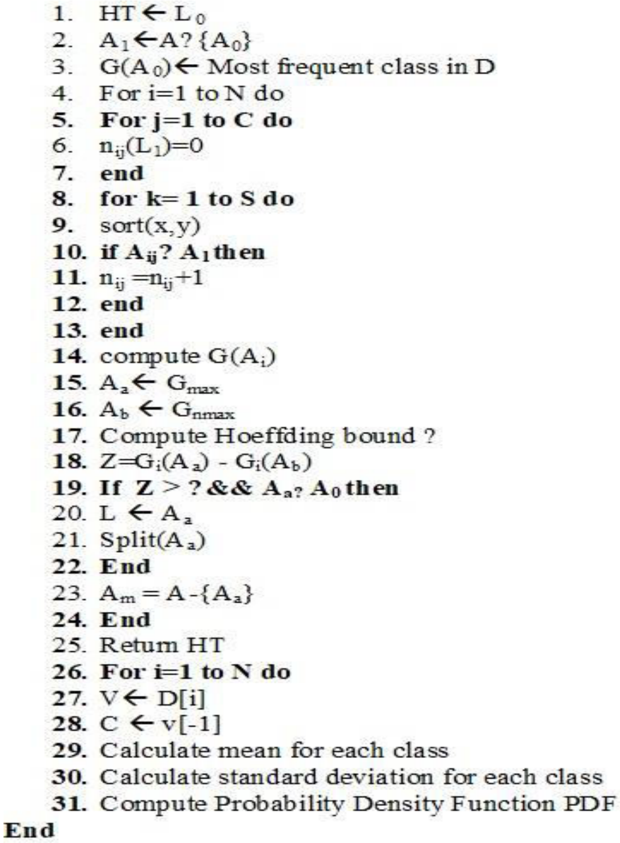
As shown by Figure 1, the NSL KDD dataset input will compare the output of different existing intrusion detection systems algorithms. To enhance the accuracy and performance of the intrusion detection system, use information gain only by removing the significant aspects from the dataset. Next, pick the data set for training and test data. To categories the data set as regular or suspicious and train the model, use one of the classification methods like: PCA, VFDT, GHSOM, Adaboost and Improved HITNB. To determine detection rate (DR) and false alarm rate (FAR) values, implement it on the test dataset. Compare the values reached in the last stage and then display the results by calculation of the model's efficiency.

4. PROPOSED ENHANCED HOEFFIDING BASED NAÏVE BAYES ALGORITHM

As there are a range of data mining techniques developed, IDS detecting attacks over large networks is an ambitious and productive activity. This algorithm reduces the data set space. For this purpose a delay between the arrival and the detection time of attacks will be useful for the network superintendent. The warning rate and computing time in real time are also less false. This algorithm includes the steps:

1. Extract the successful features for the given dataset
2. Return to the right option
3. Shape VFDT input into the best features
4. Activate the pheromone to get the best solution
5. Initialize the tree with best root node
6. Predict the sufficient statistics for each attribute X_i
7. Sort the instance into HT
8. Select the attribute X_a with highest statistical result G
9. Predict the next attribute X_b with next highest statistical result for evaluating all instances
10. Determine the difference between the hit count of two selected attributes X_a,X_b is greater than the Hoeffding bound and the selected attribute is not present the list .
11. If the result is yes then replace existing node l with X_a.
12. Add subsequent nodes with same statistics result.
13. Repeat steps 10 to 12 till all instances.
14. For each attribute A, calculate mean and standard deviation for each class
15. By using the gauss density equation calculate the probability of each predictor variables
16. Repeat step 15 until all data instances should be covered
17. Finally train the HITNB model

4.1 PSEUDOCODE FOR HITNB ALGORITHM5



```
HT ← $L_0$
$A_1$ ← A? {$A_0$}
G($A_0$) ← Most frequent class in D
For i=1 to N do
    For j=1 to C do
        $n_{ij}(L_1)$=0
    end
    for k= 1 to S do
        sort(x,y)
        if $A_{ij}$? $A_1$ then
            $n_{ij}$ =$n_{ij}$+1
        end
    end
    compute G($A_i$)
    $A_a$ ← $G_{max}$
    $A_b$ ← $G_{nmax}$
    Compute Hoeffding bound ?
    Z=$G_i(A_a)$ - $G_i(A_b)$
    If Z > ? && $A_a$? $A_0$ then
        L ← $A_a$
        Split($A_a$)
    End
    $A_m$ = A-{$A_a$}
End
Return HT
For i=1 to N do
    V ← D[i]
    C ← v[-1]
    Calculate mean for each class
    Calculate standard deviation for each class
    Compute Probability Density Function PDF
End
```

5. RESULTS AND DISCUSSION

Using data mining methods for intruder detection systems, new and unknown attacks can be easily reduced, which is the greatest aid for the complex security of the intricate detection system. This decision are made with only a machine learning platform, with 20000 NSL KDD dataset records, in order to evaluate the efficacy of our proposed approach with conventional algorithms. The reliability of the different algorithms is reliable, sensitive, precise and false alarm rates. For C4.5, SVM, PCA, VFDT VFDT+SVM,C4.5+PCA, SVM+VFDT and HITNB algorithms, the precision, sensitiveness and specificity of Table 1. The ROC curve is here a susceptibility graph and attributes specificity. The proposed Efficient Hoeffding Induction Tree algorithm Naive Bayes shows that, when compared with an existing algorithm, the sensitivity is 89.36%, specificity is 95.68% and false alarm is 0.81%, followed by PCA+VFDT 92.15, specialty 85.38 and fake alarm 0.92%. A proposed EDADT algorithm detects the attack effectively less False Alarm Rate for computing.

6. CONCLUSION AND FUTURE SCOPE

For intrusion detection method, a novel Effective Hoeffding Induction Tree based Naive Bayes algorithm for NIDS has been proposed in this paper. This system monitors network and device layers packets and status. The results of NSL KDD data set experiments show that the improved HITNB algorithm can easily find classification rules that are roughly competitive in predicative precision and simplicity. In terms of precision, the proposed HITNB algorithm is 19.4% better than C4.5, 18.8% better than SVM, 19.6% better than PCA, 19.2% better than SVM+VFDT, 19.7% better than C4.5+ VFDT, 19.3% better than VFDT + PSA. The proposed HITNB algorithm thus decreases the actual size of the dataset and allows the administrator to accurately evaluate the ongoing attacks with a lower false alarm rate.

In the future, it is possible to build a hybrid intrusion detection system based on the statistical mining algorithms that would be efficient and robust in detecting the enormous range of new and unprecedented attacks

.

7. REFERENCES

1. The Intrusion-Detection [online]. Available from: http://en.wikipedia.org /wiki/intrusion detection [last cited on 2012 June 15].

2. Anderson. J.P, “Computer Security Threat Monitoring & surveillance “Technical Report, James P Anderson Co., Fort Washington, Pennsylvania, 1980.

3. Jiawei Han and Kamber,” Data Mining: Concepts and Techniques”, 2nd Edition, Morgan Kaufman Publishers, Elsevier Inc, 2006.

4. Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Srivastava, J., Kumar, et al,” The MINDSMinnesota intrusion detection system”, Next generation data mining, MIT Press, 2009.

5. Bace, Rebecca G, NIST, special publication on “intrusion detection systems”, 2002.

6. Ben Sujatha. B, Kavitha .V, “Survey on intrusion detection approaches”, International Journal of Advanced Research in Computer Science, vol. 3, no. 1, pp.363-371, 2012.

7. Alok Ranjan, Ravindra S. Hegadi, “Emerging Trends in Data Mining for Intrusion Detection”, International Journal of Advanced Research in Computer Science, vol. 3, no. 2, pp.279-281, 2012.

8. Li, Jundong, and Huan Liu, "Challenges of feature selection for big data analytics." *IEEE Intelligent Systems* 32, no. 2 (2017): 9-15.

9. R. Suganya, S. Rajaram, A. Sheik Abdullah and V. Rajendran, 2016. A Novel Feature Selection Method for Predicting Heart Diseases with Data Mining Techniques. Asian Journal of Information Technology, vol 1, issue 8 p.no 1314-1321.

10. T Parlar, SA Özel, F Song ,“QER: a new feature selection method for sentiment analysis” - Human-centric Computing and Information, 2018 – Springer

11. Y Wang, L Feng , “A new feature selection method for handling redundant information in text classification.”, Frontiers of Information Technology & Electronic …, 2018 – Springer

12. YY Wang, H Zhang, CH Qiu, SR Xia , “A Novel Feature Selection Method Based on Extreme Learning Machine and Fractional-Order Darwinian PSO” ,Computational intelligence and …, 2018 - hindawi.com

13. NSL-KDD dataset [online] available: http://nsl.cs.unb.ca/nsl-kdd/. Accessed on 7/21/2014.