NETWORK INTRUSION DETECTION USING HYBRID NB ALGORITHM ON NSL-KDD DATASET

Asmitha Shree R[1] , Nivedha S[2] ,KishoorKumar M[3], KaranBabu B[4], Naveen V[5]
Department of Computer science and engineering
[1,2] Assistant Professor, Sri Krishna college of Technology, Coimbatore, India- 641042
[3,4,5] Research Scholar, Sri Krishna college of Technology, Coimbatore, India- 641042

Abstract- *As the communication industry has connected distant corners of the globe using advances in network technology, intruders or attackers have also increased attacks on networking infrastructure commensurately. System administrators can attempt to prevent such attacks by using intrusion detection tools and systems. In recent years Machine Learning (ML) algorithms has been gaining popularity in Intrusion Detection system (IDS). Swarm intelligence and naive bayes has become one of the popular ML algorithm used for intrusion detection due to their good generalization nature and the ability to overcome the curse of dimensionality. As quoted by different researchers number of dimensions still affects the performance of Swarm intelligence based IDS and another issue quoted is that naive bayes treats every feature of data equally. In real intrusion detection datasets, many features are redundant or less important. Proposed system hybridized swarm intelligence and naïve bayes, would be better if consider feature weights during hybrid swarm-naive bayes training. This paper presents a study that incorporates Information Gain Ratio (IGR) and swarm intelligence to naïve bayes for intrusion detection. In purposed framework NSL-KDD dataset is ranked using precision, accuracy and later feature subset selection is done using proposed algorithm.*

Keywords- KDD99, efficient learning, ensembles, feature selection, machine-learning (ML).

I. INTRODUCTION

An Intrusion Detection System is an important part of the Security Management system for computers and networks that tries to detect break-in attempts. There is no disputing fact that the number of hacking and intrusion incidents is increasing year to year as technology rolls out, unfortunately in todays interconnected Ecommerce world there is no hiding place [1]. The impetus could also be a gain, intellectual challenge, espionage, political, or just troublemaking and it exposed to a variety of intruder threats. The first important deficiency in the KDD [8] data set is the huge number of redundant record for about 78% and 75% are duplicated in the train and test set, respectively. Which makes the learning algorithm biased, that makes U2R more harmful to network. To solve these issues a new version of KDD dataset, NSL-KDD is publicly available for researchers through our website. Although, the data set still suffers from some of the problems discussed by McHugh [2] and may not be a perfect representative of existing real networks, because of the lack of open data sets for networkbased IDSs, we believe it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods.

The NSL- KDD dataset used for intrusion detection is a raw data which highly susceptible to noise, missing values and inconsistency [3]. To improve data efficiency feature reduction and filtering technique is needed, As a result the paper proposed a novel simplified swarm optimization incorporates with Random forest classifier for preprocessing, to mine raw data. Data mining provide decision support for intrusion management, and also help IDS for detecting new vulnerabilities and intrusions by discovering unknown patterns of attacks or intrusions.

The statistical analysis showed that there are important issues in the KDD data set [19] which highly affects the performance of the systems, and results in a very poor estimation of anomaly detection approaches. To solve these issues, a new data set as, NSL-KDD [7] is proposed, which consists of selected records of the complete KDD data set. The advantage of NSL KDD dataset are

1. No redundant records in the train set, so the classifier will not produce any biased result
2. No duplicate record in the test set which have better reduction rates.
3. The number of selected records from each difficult level group is inversely proportional to the percentage of records in the original KDD data set.

The training dataset is made up of 21 different attacks out of the 37 present in the test dataset. The known attack types are those present in the training dataset while the novel attacks

are the additional attacks in the test dataset i.e. not available in the training datasets. The attack types are grouped into four categories: DoS, Probe, U2R and R2L.

During last decade, KDDCup 1999 intrusion detection benchmark dataset is used by many researchers in order to build an efficient network intrusion detection system. However, recent study shows that there are some inherent problems present in

KDDCup 1999 dataset [27].The first important limitation in the KDDCup 1999 dataset is the huge number of redundant records in the sense that almost 78% training and 75% testing records are duplicated, which cause the learning algorithm to be biased towards the most frequent records, thus prevent it from recognizing rare attack records that fall under U2R and R2L categories.

At the same time, it causes the evaluation results to be biased by the methods which have better detection rates on the frequent records. This new dataset, NSL-KDD provided in [27] is used for our experimentation and is now publicly available for research in intrusion detection. It is also stated that though the NSL-KDD dataset still suffers from some of the problems discussed in [14] and may not be a perfect representative of existing real networks, it can be applied an effective benchmark dataset to detect network intrusions. More details about the inherent problems found in KDDCup dataset can be obtained from [27]. In this NSLKDD dataset, the simulated attacks can fall in any one of the following four categories.

The remainder of this paper is organized as follows: Section II introduces the common attacks in WSN and the analytic tools of intrusion detection. In Section III, the proposed methods and architecture of our research are introduced. The simulation results used to evaluate the performance of the proposed system are presented in Section IV. Finally, the conclusion and future work is discussed in Section V.

II. LITERATURE REVIEW

Intrusion Detection Systems gross raw network information or audit records as input that ends up in a large network traffic data size and the invisibility of intrusive patterns which are normally hidden among the irrelevant and redundant features to identify it as normal or attack. A new collaborating filtering technique for pre-processing the probe type of attacks is proposed by G. Sunil Kumar, [4] based on hybrid classifiers on binary particle swarm optimization and random forests algorithm for the classification of probe attacks in a network. Fernando [5] Used n-gram theory to identify redundant subsequence and proposed Hidden Markov Model for service selection to reduce audit data significantly. Wei-Chang yeh et.al [6] proposed new method by combining SSO with weighted exchange local search method for intrusion detection.

Mahoney and Chan introduced a set of instruction called learning algorithm which structures design of usual nature from anomalies free network traffic. Nature that bifurcates from the known normal design signals possible novel attacks. Their intrusion detection system is special in two aspects. In first, the nonstationary model is presented in which the designing chances based on the span of time from the time when the occurrence of last event instead of the rate. Now in the second, the intrusion detection system monitors the protocol collection in order to identify the unknown attacks that try to harm design faults in poorly monitored characteristics of the target software. On the 1999 DARPA intrusion detection system evaluation information set, they identified 70 of 180 attacks, and portioned among user behavioural anomalies and protocol anomalies. As their ways are alternative, there is a symbolic non-overlap of their intrusion detection system with the genuine DARPA members, which symbolise that they can be taken overall to enhance the coverage [10].

Mahoney and Chan introduced a set of instruction called LERAD that operates principles for identifying few occurrences in normal time series information with long order reliance. They used LERAD to identifying anomalies in network traffic packets and TCP sessions to identify novel intrusions. LERAD results the actual participants in the DARPA dataset, and identified almost all attacks that arise a firewall. LERAD is wellorganized for three causes. First, only a small part of the traffic has been tested. Second, the principles using only a little sample of the training information has been generated. Third, for building a small collection of principles that mostly covers the information, a coverage test has been used [11]. Aydin et al proposed a hybrid intrusion detection system which is the combination of misuse and anomaly based intrusion detection. In this paper they took snort as misuse based with PHAD and NETAD as anomaly based intrusion detection. PHAD and NETAD are the anomaly based statistical method. Firstly, snort is tasted on IDEVAL dataset then it detects 27 attacks out of 201 attacks, secondly PHAD is added to the snort as a preprocessor (Snort + PHAD) is tested on same dataset then the number of attacks detected is increases up to 51 out of 201 attacks, finally NETAD is added to the snort and PHAD as a preprocessor (Snort + PHAD + NETAD) is tested on same dataset then the number of attacks detected is increases up to 146 out of 201 attacks [1].

The inherent problem of KDD dataset leads to new version of NSL KDD dataset that are mentioned in [7, 8]. It is very difficult to signify existing original networks, but still it can be applied as an effective benchmark data set for researchers to compare different intrusion detection methods [2]. In [8] they have conducted a statistical analysis on this data set and found two important issues which highly affect the performance of evaluated system, and results in very poor evaluation of anomaly detection approaches. To solve these issues, they proposed a new dataset, NSL-KDD, which consists of only selected records form the complete KDD dataset and does not suffer from any of the mentioned shortcomings. Data mining [14] and machine learning technology has been extensively applied in network intrusion detection and prevention system by discovering user behavior patterns from the network traffic data.

For semi supervised approach 5000 dataset are taken, in that 2500 taken as training phase and least is taken as testing phase. Training phase includes both the labeled and unlabeled data together. The result of semi supervised approach shows 98.88 % detection rate and 0.5529 % false alarm rate [12].

Nandiammai and Hemalatha proposed an intrusion detection system which is the combination of four approaches such as classification of data named as EDADT (combination of hybrid PSO with C4.5), snort based processing named as hybrid IDS (combination of snort which is misuse based IDS with ALAD and LERAD which are anomaly based statistical algorithm), semi-supervised approach, migrating DDoS attacks named as Varying HOPERAA. Firstly EDADT algorithm gives result as 92.51% sensitivity, 88.39% specificity, 95.37% accuracy, 0.72% false alarm rate. Secondly hybrid IDS gives result as discussed above and Third semi supervised gives result as also discussed above. Finally in HOPERAA algorithm a variable clock drift method is proposed to avoid the client waiting time for server and at the same time message loss is avoided greatly. Thus HOPERAA can minimize the message transfer delay as well as execution time [13].

III. PROPOSED SYSTEM DESIGN

hybrid swarm-naive bayes, is the common name for various groups of insect species. It includes Elateridae, Lampyridae and several members of the families Phengodidae. Krishnanand and Ghose (2009) proposed Swarm Optimization (SO) as a new SI-based technique with an objective to optimize multi-modal functions. This optimization employs with physical agents called swarm-naive bayes. The swarm-naive bayes (m), at time (t) has three main parameters. It is based on the position in the search space (xm(t)), a luciferin level (lm(t)) and a neighbourhood range (rm(t)). They stated that these three parameters may vary with respect to the time. In Ant colony optimization, the finite regions being randomly placed in the search area but SO have an advantage to distribute the swarm-naive bayes randomly in the workspace. After this process, other parameters are initialized with predefined constants. This approach consists of Machine learning algorithm (Wei, 2005) to identify the attacker. The objective is to select the innermost data theft points by calculating the position of each attributes by weightage concept. After completing all the process, the Swarm behaviour is realised with machine learning to perform metrics.

Naïve Bayes (NB) is called as Idiot's Bayes, Simple Bayes and Independent Bayes, which is popular for its simplicity, elegance and robustness in building a classifier. Naïve Bayes takes a small amount of training data to estimate the means and variances of each class under consideration rather than the entire covariance matrix [18]. Further, the assumption of considering all attributes conditionally, independent for a given class provides some relaxation to improve its classification performance.

In distributed network, the data propagation will be different, based on the location and its internal characteristics. It enables the automatic detection of proper kernel at each location. The main advantage of this method is, it can locate the node even if different types of nodes are in same location. Hence, the trilateration is noticed by formulating the unknown nodes as x, y and z. Next process is indentifying the location. For this process, the swarm intelligence optimization based algorithm is considered. Here, the swarm-naive bayes based algorithm considered,. The algorithm reflects the behaviour of fireflies and lightning bugs.

Algorithm for Proposed swarm-naive bayes based optimization technique is given as follows:

Step 1: Initialize the dataset with each attributes.

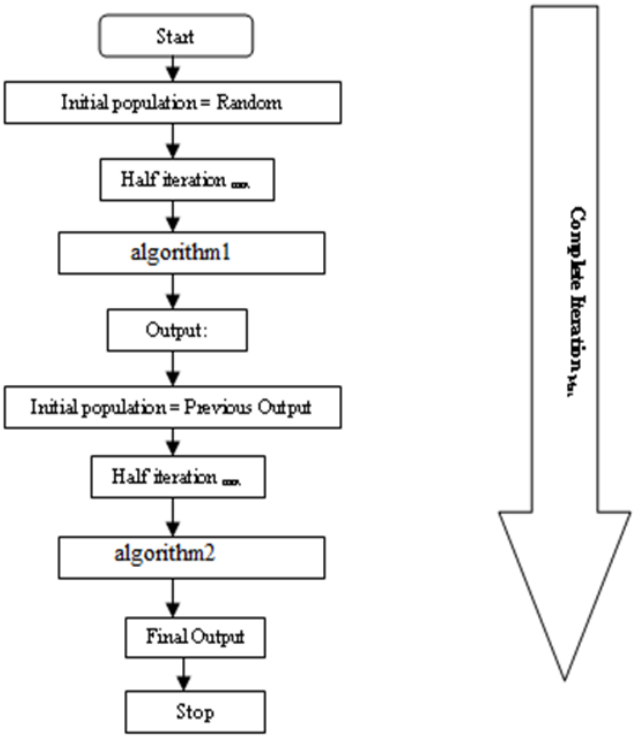
Step 2: select random samples.

Step 3: Identify the particular affect dataset.

Step 4: Calculate the initial conditions and its attributes state with machine learning concepts.

Step 5: With respect to the user details, calculate the weight and utilize it for finding the attack with least cost.

Step 6: Find the attack node in a particular location based on following SO process.



- Represent the luciferin level of swarm-naive bayes (i) and time (t), it is mentioned by li(t).
- Find the nearest nodes that have higher intensity.
- Calculate the value li(t). For example, if i=c, then it results in lc(t). In this case, if d has the nearest location, then lc(t) moves towards ld(t). Here, c and d are glow-worms.
- Updation process: It is given by

$$l_i(t+1) = (1-\rho)l_i(t) + \gamma J(x_i(t+1))$$

Based on the node variation we need to update the process.

where,

J(xi(t)) represents the objective function at sensor node position or location.

$\rho$is the delay constant, it may vary from$(0 < \rho < 1)$.

γis the luciferin enhancement constant, and

J(xi (t+1)) represents the value of the objective function at agent i's location at time t .

Step 7: Assessment the exact attack type and find the accuracy. Else go back to Step 5 and proceed until it locates the position.

IV. RESULT AND DISCUSSION

The collection of data is completely determined by each user, we collected the user profile, and updated in network, 32 attributes. The goal of intrusion classification over the network data is to predict whether a profile is either spammer, or genuine. Accuracy – It is the measures of classifier to generate an accurate classification of

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

$$= \frac{Total\ number\ of\ correctly\ classified\ cases}{Total\ number\ of\ cases}$$

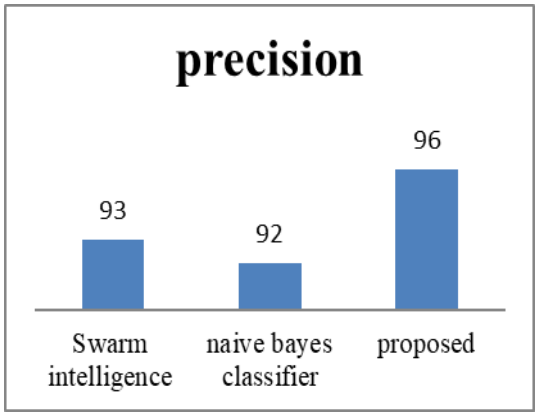
Where True Positive (TP) denotes positive result of liver disease classification

True Negative (TN) denotes negative result of liver disease classification

False Positive (FP) shows the positive result for negative liver disease classification

False Negative (FN) shows the negative result for positive liver disease classification.

It is tested and implemented with the JAVA simulation to measure the perception power between spammers and genuine users. The dataset KDD99 is considered for training/ testing sets and made an analyze between random selected datasets. The Receiver Operating Characteristics (ROC) is determined with the false positive rate on the X axis and true positive rate on the Y axis. The accuracy of the system classifier is determined by this ROC. The corner left of the ROC curve is determines the maximum accuracy. The ideal ROC curve includes the coordinate (0,1), indicating no false positives and a 100% true positive rate. From Table 2, it is clear that the results of proposed relevance vector machine based hybrid model are effective with an accuracy of 87.27% with a least precision id 96.0%. To obtain the discrimination power and the feature set of each are noticed with respect to the generated correct prediction as shown in the figure 3. It contains the account created age, average tweets per day, URL per tweet, similarity and the percentage of bidirectional. The combined false positive rates of the three algorithms are mentioned in the figure 4.



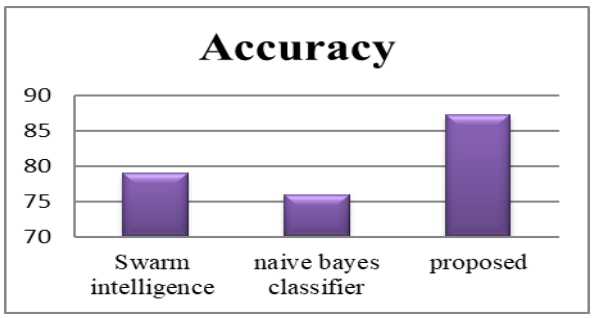


Figure 4 a) Experimental results in terms of precision 5 b) Experimental results in terms of re-call 5 c) Experimental results in terms of accuracy

Figure 4 shows the evaluation results of the sample test set. The bar chart shows the detection of the spam based on advertisement concepts. It is noticed that the classified terms and deceptive spam profile of displayed as least performance when comparing it with the other method. Here, two test results such as with postfilter and without postfilter are considered to find the prediction accurately. The substance comparability in tweets of every spammer is the biggest contrasted with alternate classes since some of them post nearly a similar substance or even attacks.

V. CONCLUSION

In this paper, analyzed the NSL-KDD dataset that solves some of the snags of KDD99 dataset. Proposed analysis shows that NSL-KDD dataset is very ideal for comparing different intrusion detection models. Using all the 41 features in the network to evaluate the intrusive patterns may leads to time consuming detection and also the performance degradation of the system. Some of the features in this are redundant and irrelevant for the process. We have used the proposed SI-Naïve bayes technique for reduce the dimensionality of the data. Our experiment has been carried out with three different optimization algorithms for the dataset with and without feature reduction and in that proposed SI-Naïve bayes shows a high test accuracy compared to all other algorithms in both the cases. So in the case of reduced feature set this analysis shows that SI-Naïve bayes is speeding up the training and the testing methods for intrusion detection that is very essential for the network application with a high speed and even providing utmost testing accuracy. In future we can try to use IoT with optimization technique to build an efficient intrusion detection system.

REFERENCES

[1] M. Ali. Aydin, A. Halim Zaim and K. Gokhan Celyan, "A hybrid intrusion detection system design for computer network security", Computer and Electrical Engineering 35(2009) 517-526.

[2] Qingqing Zhang, Hongbian Yang, kai Li and Qian Zhang, "Research on the intrusion detection technology with hybrid model", 2nd Conference on environmental science and information application technology, IEEE, 2010.

[3] Sumaiya Thaseen and Aswani Kumar, "Intrusion detection model using fusion of PCA and optimized SVM", IEEE, 2014.

[4] Divya and Surendra Lakra, "HSNORT: A Hybrid intrusion detection system using artificial intelligence with snort", International journal computer technology & application, Vol 4(3), 466-470, 2013.

[5] Vinod Kumar and Dr. Om Prakash Sangwan, "Signature based intrusion detection system using SNORT", International Journal of computer application & information technology, 2012.

[6] Natttawat Khamphakdee, Nunnapus Benjamas and Saiyan Saiyod, "Improving intrusion detection system based on snort rules for network probe attack detection", International conference on information and communication technology, IEEE, 2014.

[7] Kapil Wankhade, Sadia Patka and Ravinndra Thool, "An efficient approach for intrusion detection using data mining methods", IEEE, 2013.

[8] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi and Lilly Suriani Affendey, "Intrusion detection using data mining techniques", IEEE, 2010.

[9] Matthew V. Mahoney, "Network traffic anomaly detection based on packet bytes", ACM, 2003.

[10] Matthew V. Mahoney and Philip K. Chan, "Learning nonstationary models of normal network traffic for detecting novel attacks", ACM, 2002.

[11] Matthew V. Mahoney and Philip K. Chan, "Learning Rules for Anomaly Detection of Hostile Network Traffic", Florida Institute of Technology, Melbourne, FL 32901.

[12] G. V. Nadiammai and M. Hemalatha, "Handling intrusion detection system using snort based statistical algorithm and semi-supervised approach", Research Journal of Applied Sciences, Engineering and Technology 6(16): 2914-2922, 2013.

[13] G. V. Nadiammai and M. Hemalatha, "Effective approach toward intrusion detection system using data mining techniques", Egyptian Informatics Journal (2014) 15, 37–50.

[14] Matthew V. Mahoney and Philip K. Chan, "PHAD: Packet Header Anomaly Detection for Identifying Hostile Network Traffic, Florida Institute of Technology", Melbourne, FL 32901.

[15] M. Feily, A. Shahrestani, and S. Ramadass, "A survey of botnet and botnet detection," in Proc. 3rd Int. Conf. Emerg.

Security Inf., Syst. Technol., SECURWARE, Athens, Greece, 2009, pp. 268–273.

[16] G. Gu, P. A. Porras, V. Ye gneswaran, M. W. Fong, and W. Lee, "BotHunter: Detecting malware infection through IDS-driven dialog correlation," in Proc. USENIX Security, Boston, MA, USA, 2007, pp. 167–182.

[17] A. Karasaridis, B. Rexroad, and D. Hoeflin, "Wide-scale botnet detection and characterization," in Proc. 1st Conf. First Workshop Hot Topics Und. Botnets, Cambridge, MA, USA, 2007, p. 7.

[18] G. Gu, J. Zhang, and W. Lee, "BotSniffer: Detecting botnet command and control channels in network traffic," 2008.

[19] S. Arshad, M. Abbaspour, M. Kharrazi, and H. Sanatkar, "An anomalybased botnet detection approach for identifying stealthy botnets," in Proc. IEEE Int. Conf. Comput. Appl. Ind. Electron. (ICCAIE), Penang, Malaysia, 2011, pp. 564–569.

[20] L. Cao and X. Qiu, "Defence against botnets: A formal definition and a general framework," in Proc. IEEE 8th Int. Conf. Netw., Archit. Stor. (NAS), Xi'an, China, 2013, pp. 237–241.

[21] R. Villamarin-Salomon and J. C. Brustoloni, "Identifying botnets using anomaly detection techniques applied to DNS traffic," in Proc. IEEE 5th Consum. Commun. Netw. Conf. CCNC, Las Vegas, NV, USA, 2008, pp. 476–481.

[22] H. Choi, H. Lee, H. Lee, and H. Kim, "Botnet detection by monitoring group activities in DNS traffic," in Proc. 7th IEEE Int. Conf. Comput. Inf. Technol., CIT, Aizuwakamatsu, Japan, 2007, pp. 715–720.