A CORRELATIVE STUDY OF MACHINE LEARNING CLASSIFICATION TECHNIQUES FOR ANOMALY BASED INTRUSION DETECTION SYSTEM

N.Sai Lohitha[1], Dr.M.Pounambal[2]
[1]Dept.of C.S.E, Assistant Professor, SPMVV, Tirupati, India.
[2]SITE, Associate Professor, VIT University, Vellore, India.

Abstract

Data explosion is the most significant problem being faced in today's security world. This explosion is due to increased number of internet users, deployment of sensors etc. This data comes up with high velocity, variety and volume. As the data size increases, information security plays a major role. Intrusion detection system is software that is deployed in the perimeter of a network, computer and firewalls to monitor and analyze the data to detect any suspicious or anomalous behavior in the incoming traffic. To detect malware and various attacks in this big data, traditional techniques may not be fruitful to yield accurate results. Machine learning techniques overcome the limitations of existing techniques and present high rate of efficiency for large datasets by reducing false negatives and increasing accuracy. IDS are categorized as Signature-based detection and Anomaly-based detection. J48, Support Vector Machine (SVM) and Extreme Learning Machine (ELM) techniques are applied as they are prominent for classification. NSL-KDD dataset is used for the application of above mentioned algorithms for efficient anomaly based detection. The results show that J48 perform better in detecting anomalies compared to SVM and extreme learning techniques.

Keywords: Machine learning, Intrusion detection system, Support vector machine, Extreme learning machine.

I. INTRODUCTION

Information and network security has become predominant in the current emerging computer areas. Intrusions over the network have targeted availability, confidentiality and integrity of an organization. Intrusion detection system portrays a key job to suppress the intrusions targeting an organization. Intrusion Detection System has been split into two categories: Host based IDS and Network based IDS [1][13]. HIDS is setup on the host like computer, router which keeps track of log files. Any alterations in these tracks lead to a suspicious behavior. It also depends on the operating system of the target, so if the operating system is not updated periodically, the target is prone to attack. NIDS is deployed at several points over the network to detect any malicious activity and filter the packets. It is transparent to other devices in the network. It is further branched into signature based IDS and Anomaly based IDS [3][14]. Signature based IDS manages a database which consist of several signatures of already existing attacks. The drawback is it can't detect new attack as the signature for the respective is not present in the database. Anomaly based IDS maintains a threshold value for normal behavior of a network. Any activity gets deviated from this threshold value is considered as malicious. This way it can detect even novel attacks as well. The disadvantage is that the threshold value relies upon the size of the system and number of network devices associated [1][15] which prompts to false alert rate.

Machine learning, which is a broad category of artificial intelligence which helps the physical devices to learn it-self based on the past behavioral patterns and experiences [16]. It provides several classification algorithms to learn and behave accordingly. Support vector machine, J48 and Extreme leaning machine are considered as the most performance efficient supervised learning techniques. The work introduced in the paper utilizes the NSL–KDD(Knowledge Discovery and Data mining dataset [4], a refined type of KDD99 dataset[8] and it is treated as a standard to assess intrusion detection techniques.

II. RELATED WORK

Wang and Gu [5] recommended that "an intrusion detection model based on SVM and checked that method on NSL–KDD dataset and claimed that their method has given 99.92% effectiveness rate, which was far better than other approaches; the drawback is that they did not mention dataset statistics, percentage of training and testing samples"[8].

Tahir Mehmood et al. [2] compared SVM, J.48 and Naive Bayes techniques and calculated true positive rates, false positive rates, accuracy and misclassification rate which showed that J.48 has low misclassification rate. The disadvantage is that the techniques are applied on KDD99 dataset.

Mohammed Almseidin et al. [6] conducted a vast research by using several machine learning techniques like random tree, random forest classifier, MLP, decision tree classifier, J.48 and Naive bayes using 60000 random samples from KDD99 dataset and concluded that no single machine learning technique can handle all types of attacks. To maintain availability and confidentiality of resources over network only true positives and the average accuracy rates alone are not enough to detect the intrusion. False positives and false negative rates are also required [17].

Ahmed et al. [8] performed a detailed comparative study on extreme learning machine, support vector machine and random forest classifiers for IDS on NSL-KDD dataset which showed that extreme learning classifier technique is suitable for analyzing huge amount of data.

Jabbar and Farnaaz designed a novel model for intrusion detection system based on random forest. They used NSL-KDD set and the concluded that their model detection rate is 99.7% better than J48 classification techniques [6].

Revathi and Malathi [4], focused on a detailed study of NSL-KDD dataset that contain only chosen records [15]. Five machine learning classification algorithms namely Naive Bayesian classifier, CART, J48, Random Forest and Support Vector Machine techniques were tested. The outcome showed that SVM and Random forest has given most promising results.

III. PROPOSED WORK

The main phases of the work include dataset preparation, pre-processing, application of classification techniques, Results and visualization.

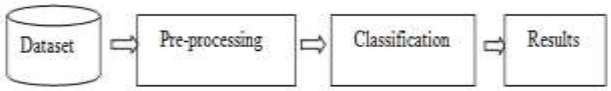

Fig. 1. Model for Anomaly-based Intrusion detection system

A. Dataset

This study utilizes NSL KDD informational collection rather than KDD99 dataset. The advantage of utilizing NSL KDD dataset is that it doesn't comprise of repetitive records in the train set, so the classifiers won't produce any one-sided result. Due to non-repetitive information in the test set it gives a superior or decrease rate. The quantity of chose records from each troublesome level gathering is conversely relative to the level of records in the first KDD informational collection. The preparation dataset comprise of 21 distinct attack categories out of the 37 present in the test dataset [18]. The known attack types are those present in the preparation dataset while the new attack patterns are extra in the test dataset which are not accessible in the preparation datasets. The attack categories are gathered into four classifications: DoS, Probe, U2R and R2L [4]. The dataset comprises of 65,535 samples.

B. Pre-processing

The classifiers can't process the raw dataset as a result of few symbolic features [8]. Thus, pre-preparing is significant, in which non-numeric data or symbolic features are restored or excluded as they don't have vital role. The mentioned procedure creates overhead including all the more preparing time; the classifier's design gets unpredictable and memory wastage occurs. In this manner, the non-numeric data is separated from the raw dataset for better execution and approximate result [19].

C. Classification

Supervised learning techniques are classification and regression. The classification algorithms applied are support vector machine, J48 and Extreme learning machine [8]. These algorithms are chosen as they give the highest accuracy. Algorithms and its importance is explained as

Support Vector machine

SVM is a process that best support binary classification.SVM targets distinguishing the capacity of the best characterization that separates individuals in preparing information of the two classes. For a straightly divisible dataset, a direct characterization work agrees to a separating hyper plane f(X) that goes through the focal point of the two classes, isolating it into two planes [20]. The capacity f(Xn), Xn has a place with the positive class if f(Xn) > 0[21]. The explanation for utilizing SVM for finding the most extreme edge hyper planes is that it gives the best classification [22] in terms of accuracy on the dataset selected [7]. It very well may be utilized for multi-classification also.

J48

J48 Classifier algorithm is mainly used for statistical classification to compare and create by making use of the notation of information entropy, a decision tree identified from a group of the training dataset. This algorithm makes use of the top down approach to indicate the decision tree for classification [9]. J48 classifier algorithms are also called as a simple C4.5 decision tree for classifications. This decision tree is considered the most appropriate supervised classification technique that is simple and faster in learning and classification. This technique can be applied in several domains.

Extreme Learning

ELM is also known as single or multiple hidden layer feed forward neural networks. ELM can provide solutions for various real-time problems by using regression, classification and clustering techniques [8]. It consists of input layer, output layer and several intermediate hidden layers. The alteration of weights in intermediate is costly and tedious. It requires increasingly number of rounds to merge to overcome this drawback. To beat this disadvantage Huang et al. [10] proposed new system SLFN where weights are balanced consequently to decrease cost and time.

IV. RESULTS AND DISCUSSION

The number of samples used is 65535. 10% of test data and 90% of train data is considered. Accuracy is used as evaluation metrics. The NSL-KDD intrusion data is divided into four categories:

1. **Denial of Service Attack (DoS):** The intruder right now utilizes flooding procedure by utilizing some memory or systems administration assets keeping them occupied with the goal that real clients get to is denied to their machine or assets.

2. **User to Root Attack (U2R):** In this sort of attack[23], the intruder accesses real clients account by utilizing sniffing, mocking strategies or secret key taking and endeavors to pick up root access[24] of the framework.

3. **Remote to Local Attack (R2L):** In this category, the masquerader sends a particular packet to a particular machine over a system without the information on track machine by sniffing procedure. This procedure encourages intruders to increase neighborhood access to the machine.

4. **Probing Attack:** The attack is done by acquiring the data of target host or target network using port filtering or scanning for vulnerabilities over a network

Comparison results for anomaly intrusion detection system using machine learning supervised learning techniques

Accuracy: Accuracy can be termed as "the total number of two correct predictions, True Positive (TP) + True Negative (TN) divided by the total number of a dataset Positive (P) + Negative (N)" [8].

$$Accuracy = \frac{TP + TN}{P + N} \quad (1)$$

Table 1: Real-time analysis results of NSL-KDD dataset on four types of attacks

Classification Technique	Accuracy
SVM	98.83%
J48	99.92%
ELM	99.64%

Table 1 results show the accuracy ratio for three classification methods within general four attacks (DoS, Probe, U2R and R2L [26]) for training NSL-KDD [25] dataset.

The figure 2 shows the comparison outline for three classification strategies on four attack classes of NSL - KDD dataset. The outcomes show that the accuracy of J48 (99.92%) classifier method performs better contrasted with SVM and ELM techniques.

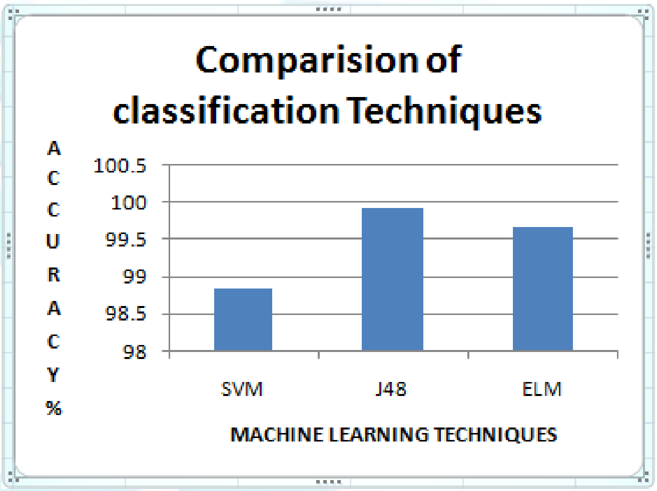

Fig 2 **:** Accuracy rate of SVM,J48 and ELM for NSL-KDD full samples

V. CONCLUSION

Anomaly based intrusion detection systems is necessary for the present and future information systems since today's world regular activities are heavily dependent on web. Several machine learning strategies are being utilized for intrusion detection systems of which some are more suitable for analyzing huge amount of memory for intrusion detection. To address this issue, three machine learning methods, to be explicit, SVM, J48 and ELM are explored and looked at. The outcomes indicated that J48 beats different methodologies in accuracy. The above specified work can be additionally extended to think about in terms of precision and recall too.

REFERENCES

[1] Ganapathy S, Kulothungan K, Muthurajkumar S, Vijayalakshmi, M, Yogesh P and Kannan, “An Intelligent feature selection and classification techniques for intrusion detection in networks a survey”. In: EURASIP Journal on Wireless Communications and Networking, vol. 2013, pp. 1-16, (2013).

[2] Tahir Mehmood, Helmi B Md Rais, “ Machine Learning Algorithms In Context Of Intrusion Detection”. In: 3rd International Conference On Computer And Information Sciences (2016).

[3] Host- vs. Network-Based Intrusion Detection Systems, Global Information Assurance Certification paper.

[4] S. Revathi, Dr. A. Malathi, “ A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection”. In: International Journal of Engineering Research & Technology (IJERT), Volume. 02, Issue. 12 (December 2013).

[5] Wang H, Gu J, Wang S,” An effective intrusion detection framework based on SVM with feature augmentation, Knowledge-Based Systems”. Volume.36, Pages. 130-139, ISSN 0950-7051,(2017).

[6] N Farnaaz, M. A. Jabbar, “Random Forest Modeling for Network Intrusion Detection System”. In: Procedia Computer Science, Volume. 89, pages. 213-217, ISSN 1877-0509 (2016).

[7] Xingdong Wu, Vipin Kumar, et.al, “Top 10 Algorithms in Data Mining”. In: Knowl edge Information System Vol. 14, pp. 1-37 (2007).

[8] I. Ahmad, M. Basheri, M.J Iqbal, A. Raheem, “Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection”. In: IEEE Access, (2018).

[9] Himadri Chauhan, Vipin Kumar, Sumit Pundir and Emmanuel S. Pilli, “A Comparative Study of Classification Techniques for Intrusion Detection”. In: International Sym- posium on Computational and Business Intelligence pp. 40-43(2013).

[10] Huang G.B, Zhu Q.Y and Siew C.K, “ Extreme learning machine: A new learning scheme of feed forward neural networks”. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541),pp.985-990 vol.2. doi: 10.1109/IJCNN.2004.1380068(2014).

[11] Elbasiony R.M, Sallam E.A, Eltobely T.E, and Fahmy M.M,” A hybrid network intrusion detection framework based on random forests and weighted k-means”. In: AinShams Eng. Journal, vol. 4,no. 4,pp. 753–762(2013).

[12] L. Dhanabal and S.P. Shantharajah,,”A study on NSL-KDD dataset for intrusion detection system based on classification algorithms”. In: International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446–452, (2015).

[13] www.inderscience.com

[14] Tahir Mehmood, Helmi B Md Rais, “SVM for network anomaly detection using ACO feature subset”. In: International Symposium on Mathematical Sciences and Computing Research (iSMSC), (2015)

[15] www.i-scholar.in

[16] Yong Zhang, Yuting Zhang, Jianying Wang, Xiaowei Zheng, “Comparison of classifi- cation methods on EEG signals based on wavelet packet decomposition”. In: Neural Computing and Applications, (2014).

[17] Suresh L, Dash S, Panigrahi B, “Artificial Intelligence and Evolutionary Algorithms in Engineering Systems”. In: Springer Science and Business Media LLC, (2015).

[18] www.ijert.org

[19] Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, Aneel Rahim, “ Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection”. In: IEEE Access (2018).

[20] docplayer.net

[21] neforecast.net

[22] Xindong Wu, “Top 10 algorithms in data mining. In: Knowledge and Information Systems”,(2008).

[23] Asish Kumar Dalai, Sanjay Kumar Jena, Chapter 35- Hybrid Network Intrusion Detection Systems: A Decade’s Perspective. In: Springer Science and Business Media LLC (2017).

[24] Prasant Kumar Pattnaik, Siddharth Swarup Rautaray, Himansu Das, Progress in Computing, Analytics and Networking. In: Springer Science and Business Media LLC, (2018).

[25] Iftikhar Ahmad., “Feature Selection Using Particle Swarm Optimization in Intrusion Detection”. In: International Journal of Distributed Sensor Networks (2015).

[26] Preeti Mishra, Vijay Varadharajan, Uday Tupakula, Emmanuel S. Pilli, “A Detailed Investigation and Analysis of using Machine Learning Techniques for Intrusion Detection”. In: IEEE Communications Surveys & Tutorials (2018).