A Survey on Machine Learning based Intrusion Detection System on NSL-KDD Dataset

Surbhi Solanki
M.Tech Scholar, PG Dept. of CSE
SIRTS, Bhopal

Chetan Gupta
Asst. Prof., PG Dept. of CSE
SIRTS, Bhopal

Kalpana Rai, PhD
Asso. Prof., PG Dept. of CSE
SIRTS, Bhopal

ABSTRACT

Nowadays, Intrusion detection system is the most emerging trend in our society. Intrusion detection system act as a defensive tool to detect the security attacks on the web. It is a device or software application that monitor network for malicious activity and alert to the administrator. Intrusion Detection System work by either looking for signatures of known attacks or deviations of normal activity. In this paper we have survey various type of intrusion detection system and techniques which are based on Support Vector Machine (SVM), machine learning, fuzzy logic, supervised learning. Also we have compared various techniques on the basis of their accuracy on NSL-KDD Datasets. We have also suggested that if we use hybrid combination of SVM and Machine learning then the accuracy can be improved.

Keywords

SVM, KDD, IDS, Dos, Probe, R2L, U2R.

1. INTRODUCTION

Security is an important issue in today's environments. Nowadays, with the development of internet technologies services in the world, the intruders have been increased rapidly [1][2][3]. Therefore, the need of intrusion detection system in the security of network field prevents intruders from having access to the information. Intrusion detection system is a device or software application that monitors the network or system for malicious activity or policy violations and sends alerts to system administrators at the proper time. Intrusion detection systems monitor both inbound and outbound traffic and activities to detect possible intrusions.

Machine learning algorithm is also used in the field of intrusion detection system. Machine learning is the method of data analyzing that automates analytical model building [4][5]. It provides the powerful tools and technologies for various fields and experienced rapid development over the last two decades [6][7].

In this work the use of NSL-KDD Dataset is suggested which is a network dataset and a refined version of its predecessor KDD CUP 99. NSL-KDD advent to solve the inherent problems of KDD CUP 99 [12]. For the classification of data support vector machine is used [13]. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs [14][15]. The classification of various techniques and their computed accuracy is shown in Table 3. The follow table shows the attacks category and their names where attacks fall in 4 categories-

Table 1: NSL KDD Dataset Attack Categories

S.NO	ATTACK CATEGORY	ATTACK NAME
1	U2R (user to root attack) Unauthorized access to local super-user or root	Buffer Overflow, Httptuneel, Rootkit, LoadModule, Perl, Xtern, Ps, SQL attacks
2	R2L(remote to local attack) Unauthorized access to a victim machine in the entire network.	WarezMaster, Guess Password, Imap, Spy, Sendmail, Xsnoop, Multihop, Phf
3	DOS (Denial of Service attack) Deny legitimate requests to a system	Apache2, Neptune, Pod, Land, Smurf, Mailbomb
4	Probe attack Information gathering attack	Satan, Saint, Ipsweep, Portsweep, Nmap, Mscan

There are some protocol used in NSL-KDD dataset and below table shows the protocol wise distribution of data in the NSL-KDD [16][17][18].

Table 2: KDD Dataset Training and Testing Dataset Sample

DATA SET	TCP	UDP	ICMP
KDDTrain +20%	20526	3011	1655
KDDTrain+	102689	14993	8291
KDDTest+	18880	2621	1043

2. LITERATURE SURVEY

In [1], this paper author proposed Supervised Learning K-NN and Unsupervised Learning K-Means machine learning algorithm for feature selection and weighting of the data packets to make the algorithm more effective. He also indicate that the proposed method show comparable performance for detecting all attack categories, which improve the performance of U2R classification, which is considered a challenge in intrusion detection.

In [2], researcher proposed a new network Intrusion Detection System based on anomaly detection approach. This proposed scheme includes: data transformation, normalization, relevant attribute selection and novelty detection model based on SSPV-SSVD i.e.; (Satisfiability Solvers and Program Verification-Sparse Single Value Decomposition) as classifiers and SMD (Short Message Delivery) as a solver to decide whether the network traffic is normal or attack. In this work they have tested our Intrusion Detection System with the whole NSL-KDD which contains nodes for training and testing.

In [3], researcher proposed the art for utilizing data mining in network security setting like intrusion discovery frameworks where, HFSA is then joined with the Naïve Byes multinomial strategy. He also compare classification performance in terms of classification accuracy, precision, recall then a portion of the current identification approaches. These keep the distribution execution as well as amazingly decrease the computational time and cost moreover.

In [4], author proposed the techniques secure enabled virtual quality routing (SEVQ) for the network nodes to estimate and characterize the impact of congestion and for a source node to incorporating these estimates into its traffic allocation. He also proposed an efficient way of using intrusion detection systems that sits on every node of a mobile ad hoc network. And to check the evaluation scheme then it is to be done by comparing the performances of the intrusion detection systems under 2 scenarios- a) keeping intrusion detection systems running throughout the simulation time b) and using proposed scheme to reduce the intrusion detection systems active time at each node in the network.

In [5], researcher proposed an (GA-IFS) genetic algorithm with improved feature selection to efficiently identify anomalies in the network by providing first an effective dataset preprocessing procedure. In his paper the dataset preprocessing technique was able to achieve a 79.07% reduction rate for the training data and 80.47% for test data.

In [6], author proposed a discretization feature process which is performed just to simplify the decision making procession continuous valued features. There are several machine learning algorithm and also feature selection techniques which is used to analyze their influence to the intrusion detection accuracy and their detection speed. Though his result shows that the uses of feature selection in this algorithm has succeeded increasing the speed of the testing but will slightly reducing the accuracy.

In [7], author proposed hybrid approaches which combine some techniques like J48 Decision Tree, Support Vector machine and Naïve Bayesian for detection of different types of attacks and also contains different types of accuracy according to algorithms. These all tests took place on NSL-KDD Dataset.

In [8], researcher uses different classifiers like DCNN (Deep Convolution Neural Network), RF (Random Forest), and NAÏVE BAYESIAN for the detection of different types of attacks. In his paper he differentiates classifiers in minority and majority base which conclude that the false detection of minority class will lead to many discrepancies in IDES (Intrusion Detection Expert System).

In [9], author proposed an anomaly based intrusion detection system based on OPSO-PNN (Oppositional Particle Swarm Optimization- Probabilistic Neural Network) model. This model compared with PSO-PNN, PSO-RB and PSO-PNN with the standard NSL-KDD dataset. The developed Oppositional Particle Swarm Optimization – Probabilistic Neural Network has better classification abilities to compare to PSO-PNN. It obtained the higher accuracy and higher detection rate and the false positive rate was so slow.

In [10], researcher proposed two types of algorithm i.e.; DBN (Deep Belief Network) and SPELM (State Preserving Extreme Machine Learning) where DBN is used to analyze and extract attack signature from dynamic data and huge amount of volume of network data. SPELM enhances the attack detection accuracy and able to differentiate the normal or attack node. He also concludes that State Preserving Extreme Machine Learning is better than Deep Belief Network.

In [11], author proposed a hybrid intrusion detection method based on improved FCM (Firebase Cloud Messaging) and SVM (Support Vector Machine). In his paper the method reduces the complexity of large scale datasets and helps to improve the performance of the Support Vector Machine classifier which is firstly uses Firebase Cloud Messaging incorporating feature information gain ratio to cluster the pre-processed training datasets. Although, a Support Vector Machine classifier is constructed for each cluster whose entropy exceeds a specified threshold to locate the attack type further.

Table 3: Literature Analysis

S. NO	Authors Name	ALGORITHMS	ACCURACY%			
			U2R	R2L	DOS	PROBE
1	XIAO YAN WANG [1] 2018	K-NN & K-MEANS	70.83	97.56	94.34	98.17
2	ELMER C.MATEL [5] 2019	GA-IFS(Genetic algorithm with improved feature selection	98.22	94.89	97.17	95.11
3	AFREEN BHUMGARA[7] 2019	J48 DECISION TREE	97.5	97.7	98.1	97.6
		SUPPORT VECTOR	93.4	93.7	97.5	97.1

		MACHINE				
		NAÏVE BAYESIAN	**71.1**	**69.9**	**74.2**	**73.9**
4	**RITU MBHRA UIKEY [8] 2019**	**NAÏVE BAYES**	**45**	**31**	**61**	**59**
		DCNN	**67.31**	**86.55**	**93.20**	**97.44**
		RF	**95.48**	**52.17**	**99.98**	**98.43**
5	**KUNAL SINGH [10] 2019**	**DBN (Deep Belief Network)**	**53**	**52**	**53**	**48**
		SPELM (State preserving extreme learning machine)	**93**	**92**	**95**	**92**
6	**ZHIYOU ZHANG[11] 2019**	**Hybrid intrusion detection based on FCM $ SVM**	**65**	**90**	**98.4**	**95**
7	**ZAKARIA EI MRABET[12] 2019**	**NAÏVE BAYES**	**92**	**91**	**93**	**98.8**
		DECISION TREE	**98**	**93**	**92**	**82**
		SUPPORT VECTOR MACHINE	**99.9**	**98**	**98.9**	**95**
		RANDOMN FOREST	**99.9**	**92**	**98**	**97**

3. PROBLEM DOMAIN

Following are some concluded problems taken from various papers:

1. Low Performance for detecting attacks.
2. Feature selection reducing the accuracy.
3. Less accuracy of U2R and R2L attacks.
4. Execution of various classifiers algorithms on diminished datasets
5. Rate of false positives and false negatives are still problematic.

4. PROPOSE WORK

In many papers researchers introduce various types of algorithm/techniques for IDS. But in this paper hybrid approach is suggested that can combine both Support Vector Machine and Machine learning algorithm on NSL-KDD Datasets. The propose work steps are shown in the below flowchart.

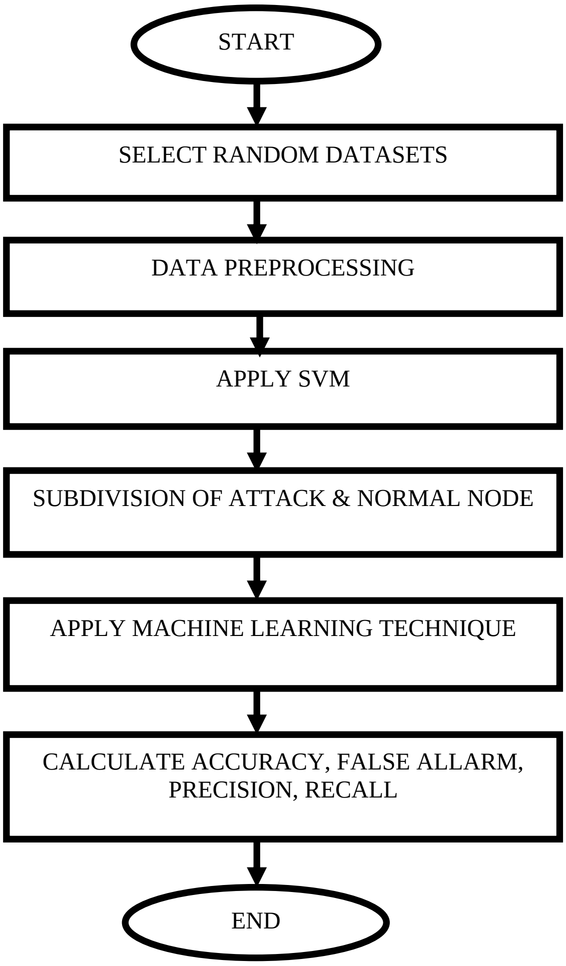
START
SELECT RANDOM DATASETS
DATA PREPROCESSING
APPLY SVM
SUBDIVISION OF ATTACK & NORMAL NODE
APPLY MACHINE LEARNING TECHNIQUE
CALCULATE ACCURACY, FALSE ALLARM, PRECISION, RECALL
END

Figure1: Flowchart of Propose Work

In this survey it is suggested that the algorithm first randomly select the data from the datasets then data preprocessing is performed which is used for feature selection. Then Support Vector Machine is used to remove the unnecessary or redundant record by classify them in normal or attacks node. Finally the machine learning technique will apply which classify the data and calculate accuracy, false alarms, precision, and recall.

5. CONCLUSION

After analyzing several research works we suggest a hybrid approach on NSL-KDD datasets and also compared the accuracy of different algorithm in all kind of attack, and concluded that intrusion detection security needs are not for the corporate world but also for network users. For future we can improve the multiple classifiers to improve the accuracy.

6. REFERENCES

[1] XIAOYAN WANG, HANWEN WANG “A High Performance Intrusion Detection Method Based on Combining Supervised and Unsupervised Learning” at IEEE Smart World, Ubiquitous Intelligence $ Computing Advanced $ Trusted Computing, Scalable Computing, Internet of People and Smart City Innovations in 2018.

[2] MOHAMMAD EI BOUJNOUNI $ MOHAMED JEDRA “New Intrusion Detection System Based on Support Vector Domain Description with Information Metric” at International Journal of Network Security in 2018.

[3] KARUNA S.BHOSALE, Assoc. Prof. MARIA, “ Data Mining Based Advanced algorithm for intrusion detection in Communication Networks” at international conference on Computational Techniques, Electronics & Mechanical System (CTEMS) in 2018.

[4] P.AMALA, G. GAYATHRI, S.DINESH “Effective Intrusion Detection System Using Support Vector Machine Learning” at International Journal of Advanced Science and Engineering Research” in 2018.

[5] ELMER C. MATEL, ARIEL M.SISAN “Optimization of Network Intrusion Detection System using Genetic Algorithm with Improved Feature Selection Technique” at Technological Institute of the Pilippines Quezon City, Phillipines2019.

[6] LUKMAN HAKIM, RAHILLA FATMA NOVRIANDI “Influence Analysis of Feature Selection to Network Intrusion Detection System Performance Using NSL-KDD Dataset” at ICOMITEE 2019, October 16th-17th 2019, Jember, Indonesia in 2019.

[7] AFREEN BHUMGARA, ANAND PITALE, “Detection of Network Intrusions Using Hybrid Intelligent System” at International Conferences on Advances in Information Technology in 2019.

[8] RITUMBHRA UIKEY, Dr. MANARI CYANCHANDANI “ Survey on Classification Techniques Applied to Intrusion Detection System and its Comparative Analysis” at 4th International Conference on Communication $ Electronics System (ICCES 2019) IEEE Conference Record #45898; IEEE Xplore ISBN; 978-1-7281-1261-9 in 2019.

[9] T.SREE KALA, A.CHRISTY, “An Intrusion Detection System Using Opposition Based Particle Swarm Optimization Algorithm and PNN” at International conference on Machine Learning, Big Data, Cloud and Parallel Computing, India 14th-16th feb 2019.

[10] KUNAL SINGH, Dr. K.JAMES MATHAI, “Performance Comparison of Intrusion Detection System between DBN and SPELM Algorithm” at National Institute of Technical Teacher Training $ Research, Bhopal India in 2019.

[11] ZHIYOU ZHANG, PEISHANG PAN “A Hybrid Intrusion Detection Method Based on Improved Fuzzy C-Means and SVM” at International Conference on Communication Information System and Computer Engineer [CISCE] in 2019.

[12] ZAKARIA EI MRABET “A Performance Comparison of Data Mining Algorithms Based Intrusion Detection System for Smart Grid” at National Institute of Posts and Telecommunication Rabat, Morocco in 2019

[13] ADITYA PHADKE, MOHIT KULKARNI, PRANAV BHAWALKAR AND RASHMI BHATTAD “ A Review of Machine Learning Methodologies for Network Intrusion Detection” at 3rd National Conference on Computing Methodologies and Communication (ICCMC 2019) IEEE Xplore Part Number: cfp19k25-art; isbn; 978-1-5386-7807-4 in 2019.

[14] S.SIVANTHAM, R.ABIRAMI, R.GOWSALYA “Comparing the Performance of Adaptive Boosted Classifiers in Anomaly Based Intrusion Detection System for Networks” at International Conference on Vision towards Emerging Trends in Communication and Networking (ViTECoN) in 2019.

[15] RAJESH THOMAS, DEEPA PAVITHRAN “A Survey of Intrusion Detection Models Based on NSL-KDD Data Sets” at the 5th HCT INFORMATION TECHNOLOGY TRENDS (ITT 2018), Dubai, UAE, Nov, 2018.

[16] HASSAN AZWAR, MUHMMAD MURTAZ, MEHWISH SIDDIQUIE, SAAD REHMAN “ Intrusion Detection in Secure Network for Cyber security Systems Using Machine Learning and Data Mining” at IEEE 5th International Conference on Engineering Technologies $ Applied Sciences, 22-23 Nov 2018, Bangkok Thailand in 2018.

[17] AZAR ABID SALIH, MAIWAN BAHJAT ABDULRAZAQ “Combining Best Features Selection Using Three Classifiers in Intrusion Detection System” at International Conference on Advanced Science and Engineering (ICOASE), University of Zakho, Duhok Polytechnic University, Kurdistan Region, Iraq in 2019.

[18] Dr. UMA KUMARI, UMA SONI “ A Review of Intrusion Detection using Anomaly Based Detection” at 2nd International Conference on Communication and Electronics Systems (ICCES 2017) IEEE Xplore Compliant – Part Number: CFP17AWO-ART,ISBN:978-1-5090-5013-0 IN 2017.