Network Intrusion Detection System: Classification, Techniques and Datasets to Implement

Nilesh B. Nanda
Student - Research Scholar (Computer Science)
Gujarat Vidyapith, Ahmadabad-Gujarat (INDIA)
e-mail: nilideas@gmail.com

Dr. Ajay Parikh
Head Department of Computer Science
Gujarat Vidyapith Ahmedabad-Gujarat (INDIA)
email:ajayparikh.gvp@gmail.com

Abstract—The Network Intrusion Detection System (NIDS) is a useful security utility that helps to prevent unauthorized and unwanted access to network resources by observing the network traffic and identify the records as either normal or abnormal. In this paper, compare three algorithms for network intrusion detection SVM, KNN and Decision Tree over Dos, Normal, R2L and U2R attacks. The features of SVM dataset are the decline for each type of attacks using correlation-based selection feature method. Then with the reduced feature set, discriminant analysis has done for the classification of different records. Comparison with other techniques shows that modified approach provides good classification rate for Normal, Dos, R2L (Remote-to-Local) and U2R (User-to-Root) attacks. A NIDS can be a software or piece of hardware. Many NIDS tools will store event or log of the event at a later date or will combine events with other data to make decisions about damage control or regarding policies. This paper shows the comparison of the different types of attacks that can be detected in a simulated core network environment. The different types of attacks are normal, DoS, Probe attacks, R2L and U2R attacks. The proposed method is implemented by the Python (Anaconda Navigator) and R programming software and tested on NSL-KDD dataset.

***Keywords**: intrusion detection system, support vector machine, decision tree, Decision Tree.*

*****

I. INTRODUCTION

The importance of security problem for the data has been increasing day by day along with the rapid development of the computer network. Security means the degree of protection given to the network or system. The main achievement of data security is integrity, confidentiality and availability of data [1].Different Attacks on the host can be referred as Intrusion. Intrusion means any set of malicious or fake activities that attempt to compromise the security standards of the information. Intrusion detection is one of the enormous information security problems. NIDS (Network Intrusion Detection System) assist the host in resisting internal and external attacks [2] [3] [4]. In early days, only traditional approaches were used for core network such as cryptography, access control list, firewalls, virtual private network VPN etc. but they were not enough to highly secure network completely. It is difficult to depend completely on fix define techniques. This increases the need for dynamic or unique technique, which can be observed and monitors system and identifies malicious activities. Thus to enhance the core network security dynamic or changing approach has been introduced and known as Intrusion Detection System NIDS. Intrusion detection system collects online data information from the large network after that monitor (observe) and analyzes this data information and divide it into normal abnormal activities, provide the result to network administrator [5]. NIDS is the key area, where Data mining or cryptography [6] is used wide extensively, this is due to limited scalability, adaptability and validity. In NIDS data is collected from various sources like firewall log data, host or server data etc. Since the network traffic is very large, the analysis of TCP data is too difficult. This gives rise to the need of using NIDS along with different Data mining KNN techniques for intrusion detection.

Support vector machines (SVM) are the classifiers which were originally designed for binary classification [7] [8] [9], can be used to classify the attacks. If we combine binary SVMs with decision trees, we can have multi-class SVM, which can classify the four types of attacks, Probing, DoS, U2R, R2L attacks and Normal data, and can prepare five classes for anomaly detection.

This paper is organized as follows. Section I gives Introduction. Section II discusses the literature survey. Section III overviews of different type of network intrusion detection system attack and its classification. Section IV gives various data mining techniques for NIDS. Section V discusses the various datasets that are used to build a NIDS and the next section VI is in conclusion.

II. BACKGROUND

Various algorithms or data flows have been used in data-mining or security area and machine-based learning methods [10] [11] [12]. In this paper, compare very famous three mining algorithm like SVM, decision tree and KNN suing KDD’90 dataset. Thus, here is a short overview of these three the algorithms to the study of intrusion detection systems.

Support Vector Machine (SVM) this is a supervised learning method used in machine learning or in mining for identifying objects. Linear classifiers find if an object is that object or is not that object by finding a hyperplane or demarcating line, that clearly segments the TCP objects from each other. By doing so, the SVM algorithm can determine the classification of the object by determining on which side of the hyperplane the object falls. By modifying the kernel function, it is possible to find a hyperplane to determine non-linear classifications by creating hyperplane lines that appear to weave through the data set. This determination is based on Gaussian radial basis and tangents. However, this study uses linear SVMs.

Decision trees are a widely used top-down, divide-and-conquer approach. A decision tree is a tree comprised of several different nodes. These nodes may be leaf nodes or decision nodes connected via edges. A leaf node is a node at which a test is executed on a feature to determine which edge to travel on to the next decision node. An edge connects leaf nodes on the tree. An instance starts at the root of the tree. A different objective and function are tested at each leaf node. Based upon the results at each leaf node a path down the tree for that instance is created. At each subsequent leaf node, another decision is made until the item reaches a leaf node. If a leaf node is reached a successful classification of that instance is determined. This is the best learning technique and requires a tree to be built from test data. Trees can be huge and best routing instances through the tree can be both computationally and time expensive. Instances without a routing path down the tree are not correctly classified causing the need to regenerate the tree.

Genetic algorithm generates a large rule set after the verified clustered set from the k-mean clustering must have been taken as input into it; every row in the GA is a rule. One of the rules specifies that if a certain procedure is being seen then it is regarded as an intrusion and if it’s the opposite then it’s not an intrusion. When an activity is being investigated, the K-Nearest Neighbor module extracts the characteristics of that activity and compare it with the characteristics described in the rules to see how close the characteristics of the observed activity is to the characteristics in the rule set, if the characteristics are so near (that is similar) then we regard it as intrusion but if its far away then its not an intrusion, KNN judges by NEARNESS. When characteristics are extracted from an observed activity, it compares it with every line of rules in the rule set, so assuming there are 5 million lines of rules, the KNN has to do the comparison 5 million times which consumes classification time and affects prediction accuracy.

III. THE DIFFERENT TYPE OF ATTACKS THAT CAN OCCUR IN A NIDS SYSTEM ARES

- Probe attacks The probe attacks are aimed at acquiring information about the target network from a source that is often external to the network. Hence, basic

  connection level features such as the "duration of connection" and âA˘ IJsource bytes" are significant while features like "number of files creations" and "number of files accessed" are not expected to provide information for detecting probes.

- DoS Attacks The DoS attacks are meant to force the target to slow any service(s) that is (are) provided by flooding or sending garbage packets towards servers with probes fack requests. Hence, for the DoS attack to be detected, traffic features such as the "percentage of connections having same service and same destination host" and packet level features such as the "source bytes" and "percentage of packets with errors" are significant. To detect DoS attacks, it may not be important to know whether a user is "logged in or not."

- R2L Attacks The R2L attacks are one of the most tuff to detect as they involve the host and network level features. We, therefore, select both the network level features such as the "duration of connection" and "service requested" and the host level features such as the "number of failed login attempts" among others for detecting R2L attacks.

- U2R Attacks The U2R attacks involve the semantic details that are very tuff to capture at an early phase. Such attacks are often content based, packet base and target an application. Hence, for U2R attacks, features such as "number of file creations" and "number of shells prompts invoked," are selected while features such as "protocol" and "source bytes are ignored.

IV. KDD CUP’90 DATASET

KDD Cup ’90 intrusion detection datasets [10] which are based on DARPA ’98 dataset provides main data for the researcher working in the field of intrusion detection and is the only fix dataset publicly available. The details of KDD dataset are given in the next section. The KDD dataset has generated using a simulation of a military network consisting of three target machines running various operating systems and traffic. Finally, there is a watchdog that records all network traffic using the tcpdump format. The total simulated period is few weeks. Normal TCP connections are created to profile than expected in a military network and attacks fall into one of the four categories.

- Denial of Service (Dos): Dos Attacker tries to slow service server and send continuous garbage packets.
- Remote to Local (r2l): Attacker tries to gain access to remote machine because they do not have rights to access or does not have control of same.
- User to Root (u2r): Attacker does not have super or root privilege on the machine, it has the local machine but does not has full rights.
- Probe: Attacker tries to get information of remote host without knowing actual users.

There are 41 features for each connection, which are detailed in Table I. Specifically, "a connection is a sequence of TCP packets starting and ending at some well-defined times, between which data flows from a source IP address to a target IP address under some well-defined protocol". Features are grouped into four categories:

- Host-based Traffic Features: Utilize an estimate window over the number of connections instead of time in Sec. Host-based features are designed to access attacks, which span intervals longer than min 2 seconds

- Basic Features:: Basic features can be derived from TCP packet headers without inspecting the data payload.
- Content Features: Domain knowledge is used to access the data payload of the main TCP packets. This includes features such as a number of various failed login attempts or successful events.
- Time-based Traffic Features: These main features are designed to capture properties that mature over a min 2 second TCP window. One example of such a main feature is the number of different connections to the same node over the 2 second interval time.

V. EXPERIMENTAL RESULTS

In the experiments, instead of using standard KDD99, the NSL-KDD dataset is used. This dataset has several benefits in comparison with KDD'99 [12]:

- The redundant record is removed from the train set to eliminate the bias to the most frequent records.
- Duplicate records in test sets are removed.
- The number of records in the test and train datasets is reasonable. In the experiment, subsets of training and test datasets are utilized.

In the experiments, instead of using standard KDD99, the NSL-KDD dataset is used. This dataset has several benefits in comparison with KDD'99 In [13] the NSL-KDD dataset is analyzed using Simple K-Means clustering algorithm. The dataset is clustered into normal and four of the major attack categories,i.e. DoS, Probe, R2L, U2R. It is shown that NSL-KDD dataset has reasonable accuracy in comparing with KDD99. The proposed method is implemented by the Python (Anaconda Navigator) and R programming software and tested by NSL-KDD dataset. The number of training and testing datasets which are used for the experiments are shown in Tables and Graph[12]:

TABLE I: Summary of Network Intrusion Detection Techniques

Attack	Attack Class	Frequency	Attack Class	Frequency Percent
DoS	312251	79.01	79207	80.17
Probe	3796	0.96	311	0.31
R2L	1125	0.28	1	0.00
U2R	35	0.01	17	0.02
normal	78010	19.74	19268	19.50

Table I The number of training and test datasets which is used for the experiments.

Two scenarios are used to investigate the performance of the proposed method is compared with the KNN method: Scenario 1: in this experiment, just training datasets are used for the algorithm. Thus the training and test datasets are completely separated from each other.
Scenario 2: in this experiment, in training not only train dataset, is used, but also a subset of test dataset is used. Thus the training and test datasets are not completely separated from each other.

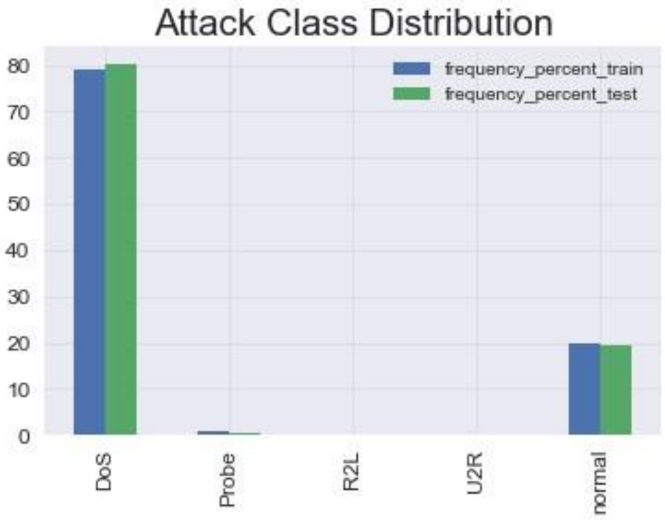

Fig. 1: Attack class bar plot

The simulated analysis of the IDS method KNN classifier with Ant Colony Optimization (ACO) and Decision tree [14] and Support Vector Machine (SVM) is done using well define performance measuring parameters which are accuracy and Cross-Validation Mean Score. Here, table II shows accuracy result of the Evaluates models and Test models using SVM, Decision tree and KNN algorithms. After analysis, it is found that the overall accuracy rate for Evaluates method of SVM is about 99.82% whereas the Test models are 99.94%. Decision tree accuracy is 100% during evaluates models and 99.83% during Test models. In KNN algorithm accuracy become 99.99% whereas in test models 99.98%. So it is concluded that Evaluates models generate more accurate result for intrusion detection as compared to Test method.

Table II Comparison for the accuracy rate of Evaluates models and test models with DOS attacks of SVM, Decision Tree and KNN model.

Models	**Evaluates Models**	**Test Models**
SVM	0.998291438618	0.999408938857
Decision Tree	1.0	0.998327422722
KNN Model	0.999971177034	0.999823939234

TABLE II: Accuracy

Table III Comparison for Cross-Validation Mean Score of Evaluates models and test models with DOS attacks of SVM, Decision Tree and KNN model.

Models	**Evaluates Models**
SVM	0.993037729178
Decision Tree	0.997446033224
KNN Model	0.997932749735

TABLE III: Cross-Validation Mean Score

Here, table III also shows the cross-validation mean score of evaluates a model for SVM, KNN and Decision tree algorithm which shows that KNN gives good results compare to Decision tree and SVM.

VI. Conclusion

The Network Intrusions detection system (NIDS) is the tool with the help of observing and analyzing the different traffic in order that the data packets that may be infected with virus or harm to the network can be detected and discarded. The presence of missing values in a KDD cup 90 datasets can influence the performance of a classifier developed using that dataset as a training sample. In this paper compare an SVM, KNN and decision tree algorithm to improve the network intrusion detection system (NIDS) and after observing conclude that the performance of the KNN algorithm has significantly improved the classification accuracy and thus it reveals the importance of preprocessing in NIDS. As compared to the existing methods, Evaluates model fairly improves the KNN classification accuracy of Dos attacks. Hence conclude that the KNN analysis proves to be an efficient classifier for DoS attacks.

Disclaimer :

The authors declare that there is no conflict of interest regarding the publication of this paper.

VII. References

[1]. S. Axelsson, “Intrusion detection systems: A taxonomy and survey,” Technical Report No 99-15, Dept. of Computer Engineering, Chalmers University of Technology, Sweden, March 2016.

[2]. D. R. R. S. Selvakani Kandeeban, “A genetic algorithm based elucidation for improving intrusion detection through the condensed feature set by kdd99 dataset,” information and knowledge management ISSN 2224-5758, ISSN 2224-896X Vol. 1, No.1, 2011, www.iiste.org.

[3]. M. T. Mouaad KEZIH, “evaluation effectiveness of intrusion detection system with reduced dimension using data mining classification tools,” 2nd International Conference on Systems and Computer Science (ICSCS) Villeneuve d’Ascq, France, August 26-27, 2013; 978-1-4799-2022.

[4]. M. S. Mrs. D. Shona, “An ensemble data preprocessing approach for intrusion detection system using variant firefly and bk-nn techniques,” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 6 (2016) pp 4161- 4166.

[5]. T. F. Lunt, “Detecting intruders in computer systems,” in the proceeding of 1993 Conference on Auditing and Computer Technology, 2001.

[6]. T. M. B. Ayman I. Madbouly, Amr M. Gody, “Relevant feature selection model using data mining for intrusion detection system,” International Journal of Engineering Trends and Technology (IJETT) - Volume 9 Number 10 -, Mar 2014.

[7]. D. E. Denning, “An intrusion detection model,” In IEEE Transactions on Software Engineering, Vol.SE 13, Number 2, page 222-232, February 2017.

[8]. M. S. Divyatmika, “A two-tier network-based intrusion detection system architecture using machine learning approach,” International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) - 2016 in the proceeding of IEEExplore., 2016.

[9]. Y. L. JianfengPu, Lizhi Xiao and X. Dong, “A detection method of network intrusion based on SVM and ant colony algorithm,” National Conference on Information Technology and Computer Science (CITCS 2012) Published by Atlantis Press., 2012.

[10]. W. Lee and S. J. Stolfo, “Data mining approaches for intrusion detection,” In Proceedings of the7th USENIX Security Symposium - Volume 7, SSYMâA˘ Z´98, pages 6-6, Berkeley, CA, USA, 2012.

[11]. M. Panda and M. R. Patra, “Comparative study of data mining algorithms for network intrusion detection,” First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008.

[12]. M. H. G.V. Nadiammai, “Effective approach toward intrusion detection system using data mining techniques,” Egyptian Informatics Journal 2015, in 0 50 100 DOS PROB U2R R2L Accuracy Accuracy Comparison 0 BKP 0 0 10 20 30 40 DOS PROB U2R R2L FAR FAR Comparison FAR SVM BKP FAR Proposed International Journal of Computer Applications (0975-8887) Volume 171-No. 10, August 2017 23 proceeding Elsevier Pp 37-50.

[13]. P. K. VivekNandanTiwari, Prof. SatyendraRathore, “Enhanced method for intrusion detection over kdd cup 99 dataset,” International Journal of Current Trends in Engineering Technology, Volume: 02, Issue: 02 (MAR-APR, 2016), ISSN: 2395-3152., 2016.

[14]. M. R. Norouzian and S. Merati, “Classifying attacks in a network intrusion detection system based on artificial neural networks,” in Advanced Communication Technology (ICACT), 2011 13th International Conference on, pp. 868-873, IEEE, 2014.