IRJEAS
NURTURING RESEARCH

INTERNATIONAL RESEARCH JOURNAL
OF
ENGINEERING & APPLIED SCIENCES

ISSN: 2322-0821(O) VOLUME 5 ISSUE 2
ISSN: 2394-9910(P) Apr 2017 - Jun 2017

www.irjeas.org

CYBER ATTACK DETECTION AND CLASSIFICATION BASED ON MACHINE LEARNING TECHNIQUE USING NSL KDD DATASET

Anurag Shrivastava[1], Jyoti Sondhi[2], Sunita Ahirwar[3]

[1]*Asso. Professor &Head, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*
[2]*Assistant Professor, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*
[3] *M.Tech Scholar, Department of Computer Science& Engineering, NRI Institute of research &Technology, Bhopal*

***Abstract*: This research work aims in design and development of an approach for boost cyber attack detection. Growth of information system is increase the data size and attention of intruders now days. Intrusion Detection System (IDS) as the security technique and is widely used against intrusion. Researchers use Data Mining and Machine learning techniques intrusion detection research area. Recently, many machine learning methods have also been useful to obtain high detection rate and low false alarm rate. NSL KDD dataset used for intrusion detection system. Shortcoming of all those techniques is low detection rate and high false alarm rate. The purpose of this paper is to propose IDS framework model based on machine learning technique. This model improves the classification performance. The Proposed work is tested on basis of Accuracy, Error rate, Detection rate and False Alarm rate.**

***Keywords*—Cyber attack, Classification, Machine learning technique, NSL KDD Dataset.**

I. INTRODUCTION

The information security research that has been the subject of much attention in recent years is that of cyber attack detection systems. As the cost of information processing and internet accessibility falls, organizations are becoming increasingly vulnerable to potential cyber threats such as network cyber attacks. So, there exists a need to provide secure and safe transactions through the use of firewalls, Cyber Attack Detection Systems (CADSs), encryption, authentication, and other hardware and software solutions. However, completely preventing breaches of security appear, at present, unrealistic. Efforts can be made to detect these attacks attempts, so that action may be taken to repair the damage later. This field of research is called Cyber Attack Detection. System vulnerabilities and valuable information magnetize most attackers' attention. Traditional intrusion detection approaches such as firewalls or encryption are not sufficient to prevent system from all attack types. The number of attacks through network and other medium has increased dramatically in recent years. Efficient intrusion detection is needed as a security layer against these malicious or suspicious and abnormal activities. Thus, intrusion detection system (cyber attack) has been introduced as a security technique to detect various attacks. IDS can be identified by two techniques, namely misuse detection and anomaly detection. Misuse detection techniques can detect known attacks by examining attack patterns, much like virus detection by an antivirus application. However they cannot detect unknown attacks and need to update their attack pattern signature whenever there is new attacks .On the other hand, anomaly detection identifies any unusual activity pattern which deviates from the normal usage as intrusion. Although anomaly detection has the capability to detect unknown attacks which cannot be addressed by misuse detection, it suffers from high false alarm rate .In recent years, and interest was given into machine learning techniques to overcome the constraint of traditional intrusion techniques by increasing accuracy and detection rates. New machine learning based IDS is used in our detection approach. Boost the performance of IDS and the low false alarm rate.

A. Data Mining

Data Mining is defined as the technique of extracting information or knowledge from huge amount of data. In other words, we can say that data mining is mining knowledge from large data.

B. Machine Learning Technique :

When a computer needs to perform a certain task, a programmer's solution is to write a computer program that performs the task. A computer program is a piece of code that instructs the computer which actions to take in order to perform the task. The field of machine learning is concerned with the higher-level question of how to construct computer programs that automatically learn with experience. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E Thus, machine learning algorithms automatically extract knowledge from machine readable information. In machine learning, computer algorithms (learners) attempt to automatically distill knowledge from example data. This knowledge can be used to make predictions about novel data in the future and to provide insight into the nature of the target concepts applied to the research at hand, this means that a computer would learn to classify alerts into incidents and non-incidents (task T). A possible performance measure (P) for this task would be the Accuracy with which the machine learning program classifies the instances correctly. The training experiences (E) could be labeled instances.

II. Related work:

Author [1] says common problem of IDS that are high false positives and low detection rate. An unsupervised machine learning using k-means was used to propose a model for Intrusion Detection System (IDS) with higher efficiency rate and low false positives and false negatives. The NSL-KD data set was used which consisted of 25,192 entries with 22 different types of data.

In [2], the author said performance of a Machine Learning algorithm called Decision Tree is evaluated and compared with two other Machine Learning algorithms namely Neural Network and Support Vector Machines which has been conducted by A.. From the experiments conducted, it was found that the Decision tree algorithm outperformed the other two algorithms. Compare the efficiency of Neural Networks, Support Vector Machines and Decision Tree algorithms against KDD-cup dataset.

Author [3 Presented a comprehensive analysis on Probe attacks, by applying various popular machine learning techniques such as Naïve Bayes, SVM, Decision Trees etc. Author used KDDcup99 data set to build the model. Author proposed three layers architecture for detection of probe attacks. Principal Component Analysis is used for dimensionality reduction. Author removed duplicate samples from the training data set. Here author compared the performance of each classifier with the help of a line chart.

In [4] authors said The Intrusion detection system deals with large amount of data which contains various irrelevant and redundant features resulting in increased processing time and low detection rate. Therefore feature selection plays an important role in intrusion detection. There is various feature selection methods used. Author's comparer the different feature selection methods are presented on KDDCUP'99 dataset and their performance are evaluated in terms of detection rate. Feature selection can reduce the computation time and model complexity.

The authors [5] have proposed to use data mining technique including classification tree and support vector machines for intrusion detection. Utilize data mining for solving the problem of intrusion because of following reasons: It can process large amount of data. User's subjective evolution is not necessary, and it is more suitable to discover the ignored and unknown information. Machine learning based ID3 and C4.5 two common classification tree algorithms used in data mining. Author said C4.5 algorithm is better than SVM in detecting network intrusions and false alarm rate in KDD CUP 99 dataset.

III. NSL KDD data set:

In Earlier days the researcher focused on DARPA dataset for analyzing intrusion detection [13]. It consist of seven weeks of training and also two weeks of testing raw tcpdump data. The main drawback is its packet loss. The refined version of DARPA dataset which contains only network data (i.e. Tcpdump data) is termed as KDD dataset [11]. Which consist on 5 million single connection for training records and 2 million connection for testing. Due to its huge size the researcher used on 10%% of dataset to analysis intrusion accuracy which affects the performance of the system, and results in a very poor estimation of anomaly detection approaches. To solve these issues, a new data set as, NSL-KDD [9] is proposed, which consists of selected records of the complete KDD data set. The advantage of NSL KDD dataset is 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.

A. Feature selection

Due to the large amount of data flowing over the network real time intrusion detection is almost impossible. Feature selection can reduce the computation time and model complexity. Research on feature selection started in early 60s [11]. Basically feature selection is a technique of selecting a subset of relevant/important features by removing most irrelevant and redundant features from the data for building an effective and efficient learning model [11].A number of feature selection algorithms are proposed by various authors. Attribute evaluator is basically used for ranking all the features according to some metric.

IV. Proposed Work

Some research in machine learning community has addressed the strategy for improve the performance of cyber attack detection system. Intrusion Detection Systems (IDSs) are designed to defend computer systems from various cyber attacks and computer viruses. IDSs build effective classification models or patterns to distinguish normal behaviors from abnormal behaviors that are represented by network data. To classify network activities as normal or abnormal while minimizing misclassification .To defend computer systems from various cyber attacks and computer viruses. In this approach we proposed a classification farmwork model that uses the machine learning technique for classification.

V. ARCHITECTURE OF THE PROPOSED CLASSIFICATION MODEL

In Architecture of the proposed model shows that in NSL Dataset. Firstly we are applying preprocessing technique and get preprocessed dataset now we are using feature selection technique in preprocessed data.

Now going to classification part and determine the training and testing data in very short period after that applying classification technique in trained data and evaluate the result. Same procedure is applying in different machine learning classifier and measure result. Parameter of the performance measures in the terms of high detection rate, low false alarm rate, less training and testing time, and high accuracy.

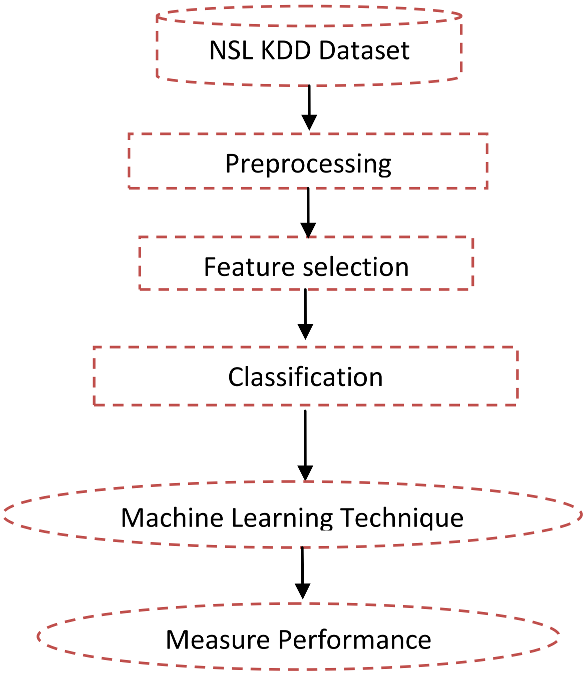

Figure 1. Architecture of the system

VI. RESULT ANALYSIS

Following fundamental definition and formulas are used to estimate the performance of the classifier: accuracy rate (AR) and Error Rate (ER).

True Positive: When, the number of found instances for attacks is actually attacks.

False Positive: When, the number of found instances for attacks is normal.

True Negative: When, the number of found instances is normal data and it is actually normal.

False Negative: When, the number of found instances is detected as normal data but it is actually attack.

The accuracy of IDS classifier is measured generally on basis of following parameters:

Detection Rate: Detection rate refers to the percentage of detected Attack among all attack data, and is defined as follows:

$$\text{Detection rate} = \frac{\text{TP}}{\text{TP} + \text{TN}} * 100$$

With this formula detection rate for different types of Attacks can be calculated.

False Alarm rate: False alarm rate refers to the percentage of normal data which is wrongly recognized as attack. , and is defined as follows:

$$\text{False Alarm rate} = \frac{\text{FP}}{\text{FP} + \text{TN}} * 100$$

VII. CONCLUSION

In this paper, Machine Learning technique have been proposed in terms of accuracy, detection rate, false alarm rate and accuracy for four categories of attack under different percentage of normal data. The purpose of this proposed method efficiently classify abnormal and normal data by using very large data set and detect intrusions even in large datasets with short training and testing times. Most importantly when using this method redundant information, complexity with abnormal behaviors are reduced. With proposed method we get high accuracy for many categories of attacks and detection rate with low false alarm. The proposed method results compare with other machine learning technique using intrusion detection to improve the performance of intrusion detection system. Experimental results and analysis shows that the proposed system gives better performance in terms of high detection rate, low false alarm rate, less training and testing time, and high accuracy.

References

[1] Solane Duquea*, Dr.Mohd. Nizam bin Omarb Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS) www.sciencedirect.com, Elsevier 2015.

[2] Chi Cheng, Wee Peng Tay and Guang-Bin Huang "Extreme Learning Machines for Intrusion

Detection" - WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[3] Ch.Ambedkar, V. Kishore Babu Detection of Probe Attacks Using Machine Learning Techniques International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) Volume 2, Issue 3, March 2015, PP 25-29 ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online) www.arcjournals.org

[4] Kamarularifin Abd Jalill, Mohamad Noorman Masrek "Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion" 201O International Conference on Networking and Information Technology 978-1-4244-7578-0/$26.00 © 2010 IEEE

[5] YU-XIN MENG "The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" 2011 IEEE

[6] 1Premansu sekhara rath, 2Manisha mohanty, 3Silva acharya, 4Monica aich Optimization of ids algorithms using data mining technique International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Volume-4, Issue-3, Mar.-2016

[7] S. Revathi 1 Dr. A. Malathi2 Network Intrusion Detection Based On Fuzzy Logic International Journal of Computer Application Issue 4, Volume 1 (February 2014) Available online on http://www.rspublication.com/ijca/ijca_index.htm ISSN: 2250-1797

[8] M.Saraswathi1, N.Kowsalya2 Anomaly Detection via Online Oversampling Principal Component Analysis International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 8, August 2015.

[9] Mr. Kamlesh Patel1, Mr. Prabhakar Sharma2 An Implementation of Intrusion Detection System Based on Genetic Algorithm International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 11, November 2016.

[10] S. Revathi 1 Dr. A. Malathi2 Detecting User-To-Root (U2R) Attacks Based on Various Machine Learning Techniques International Journal of Advanced Research in Computer and Communication Engineering.

[11] Megha Aggarwal, Amrita " Performance Analysis Of Different Feature Selection Methods In Intrusion Detection" International journal of scientific & technology research volume 2, issue 6, june 2013 issn 2277-8616

[12] Purushottam R. Patil, Yogesh Sharma, Manali Kshirasagar," Performance Analysis of Intrusion Detection Systems Implemented using Hybrid Machine Learning Techniques" International Journal of Computer Applications (0975 – 8887) Volume 133 – No.8, January 2016

[13] Rajesh Wankhede, Vikrant Chole " Intrusion Detection System using Classification Technique" International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[14] S. Revathi, Dr. A. Malathi " A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December – 2013

[15] Koushal Kumar, Jaspreet Singh Batth Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms International Journal of Computer Applications (0975 – 8887) Volume 150 – No.12, September 2016