Review Article

A Review of Cyber Attack Detection System Using Machine Learning Technique

Arti Kushwaha[1], Vaibhav Patel[2] and Anurag Shrivastava[3]

[1]M.Tech. Scholar, Department of CSE, NIRT, Bhopal, M.P., INDIA
[2,3]Department of CSE, NIRT, Bhopal, M.P., INDIA

ABSTRACT

The aim of this research work is to design and development of an approach for improves cyber-attack detection mechanism. Growth of information system is increasing 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 in cyber-attack detection research area. Recently, many machine learning methods have also been useful to obtain high detection rate and accuracy. KDD Cup 99 dataset used for attack detection system. Shortcoming of all those techniques is low detection rate and high false alarm rate. The purpose of this paper is to review the various attack detection model based on machine learning technique and propose classification framework model based on decision tree at cloud platform. This model improves the classification performance. The Proposed work is tested on basis of Accuracy.

KEYWORDS

Cyber-attack, Classification, Machine learning technique, KDD Cup 99 Dataset.

1. 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 are 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.

2. RELATED WORK:

Security of Information is main important issue in modern Information system. Internet attacks are rising day by day and there have been different attack detection methods accordingly.

Intrusion detection systems have been using all along with the data mining and machine learning techniques to detect intrusions. In this survey discuss the data mining, cloud computing and machine learning technique which is used to develop the act of intrusion detection system.

Traditional Intrusion detection system using data mining and machine learning techniques are work on information system they are not working on cloud environment. Here give some literature about Intrusion detection system and using cloud for classification with machine learning techniques. Multiple choices of cloud computing models are available for different work load management, performance and computational requirements. The popular statistical tools and environments like Octave, R and Python are now embedded in the cloud as well [5].

Authors [3] worked on IDS for web proxy, taking inspiration from Intrusion Detection Systems that make use of machine learning capabilities to improve anomaly detection accuracy, this paper proposes that cloud-based machine learning can be used in order to detect and classify web proxy usage by capturing packet data and feeding it into a cloud based machine learning web service.

Authors [23] said about the cloud-based attack system. Authors add new valued feature to the cloud-based websites and at the same time introduces new threats for such services. DDoS attack is one such serious threat. Covariance matrix approach is used in this article to detect such attacks. The results were encouraging, according to confusion matrix and ROC descriptors.

In this research work authors [24] find that the results of k-means clustering showed that a higher efficiency rate is achieved when the correct number of clusters is applied, and increasing or decreasing the cluster beyond the number of data types only lessens the efficiency of the model.

In research paper [25], authors used novel feature reduction-based machine learning algorithms for detecting anomalous patterns in the recently provided dataset. High accuracy of 86.15 percent has been achieved. For large datasets, it is very critical to have a lesser number of features with the best accuracy results. Author's able to reduce the number of features from 49 to 37 using the novel Variance Threshold method. The results obtained are encouraging and in future distributed Machine Learning Algorithms can be applied to do the faster computation for large datasets.

In research [26] author's proposed feature selection methods using AR and compared it with three feature selectors CFS, IG, and GR. The experiment shows the detection rate of our method is higher than the detection rate of full data and is also as highly as detection rate of other methods. Also, false alarm rate is lower than full data and is as low as false alarm rate of other methods.

3. KDD CUP 99 DATA SET:

Since 1999, KDD'99 [11] has been the most wildly used data set for the evaluation of anomaly detection methods. This data set is organized by Stolfo et al. [11] and is related to the data captured in DARPA'98 IDS evaluation program. DARPA'98 is about 4 gigabytes of packed together raw tcpdump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each one with about 100 bytes. The two weeks of test data have approximately 2 million connection records. KDD training dataset have approximately 4,900,000 single connection vectors each of which have 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The virtual attacks fall into one of the following categories:

1) Probing Attack: is an attempt to gather information about a network of computers for the apparent reason of circumventing its security controls. The intruder attempts to gather information about potential target computers by scanning for vulnerabilities in software and configurations that can be exploited. This includes password cracking, port scanning.

2) User to Root Attack (U2R): is a class of exploit in which the attacker have right to use to a normal user account on the system (maybe gained by sniffing passwords, a dictionary attack, or social engineering) and is able to make use of some vulnerability to gain root access to the system. e.g. guessing password;

3) Remote to Local Attack (R2L): take place when an attacker who has the capability to send packets to a machine over a system but who does not have an account on that machine make the most of some vulnerability to gain local access as a user of that machine.

4) Denial of Service Attack (DoS): is an attack in which the invader makes some computing or memory resource too busy or too full to handle legal requests, or denies legitimate users access to a machine. The general purpose of DoS attacks is to disrupt some service on a host to prevent it from dealing with certain requests. This may be a step in a multi-stage attack, such as the Mitnick attack which is described below, or to be destructive

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.

4. 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 framework model that uses the machine learning technique for classification.

5. 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.

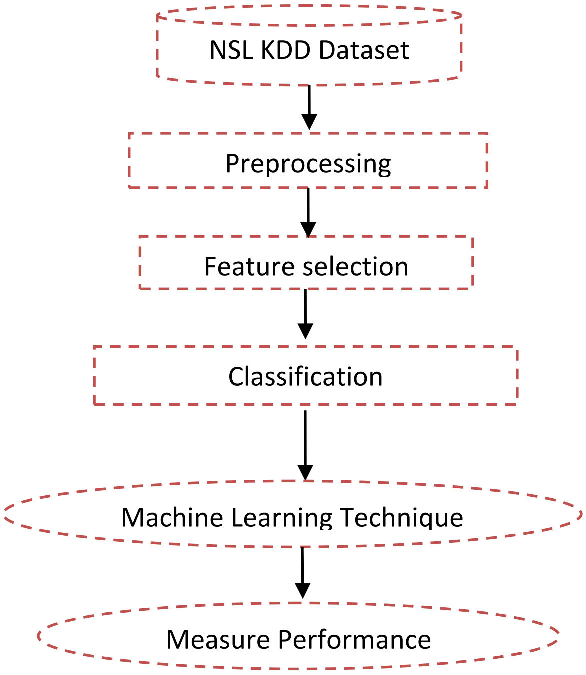

Figure 1. Architecture of the system

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.

6. 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.

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

7. CONCLUSION

In this paper, Machine Learning technique has been proposed in terms of accuracy, and accuracy for four categories of attack under different percentage of normal data. The purpose of this proposed method efficiently classifies 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 is 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 show 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] Pine II, B.J. and Gilmore, J.H. 1999. The Experience Economy. Boston: Harvard Business School Press.

[2] Ch.Ambedkar, V. Kishore Babu, "©ARC Page 25 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)

[3] Shane Miller, Kevin Curran, Tom Lunney " Cloud-based machine learning for the detection of anonymous web proxies" ISSC 2016.

[4] David Chappell, "introducing azure machine learning: a guide for technical professionals", Sponsored by Microsoft Corporation, 2015 Chappell & Associates.

[5] https://portal.azure.com

[6] Daniel Pop, "Machine Learning and Cloud Computing Survey of Distributed and SaaS Solutions", https://www.researchgate.net/publication/257068169.

[7] E.W.T. Ngai ,, Li Xiu, D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications 36 (2009) 2592–2602, Elsevier

[8] Suthaharan, S., "Big data classification: Problems and challenges in network intrusion prediction with machine learning" Performance Evaluation Review, 41(4), 70-73, ACM 2014.

[9] Maria Muntean, Honoriu Vălean, Liviu Miclea, Arpad Incze "A Novel Intrusion Detection Method Based on Support Vector Machines" IEEE 2010.

[10] Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest", R News, ISSN 1609-3631, Vol. 2/3, December 2002.

[11] https://www.MulticlassDecisionForest.html

[12] Apache Hadoop Website http://hadoop.apache.org/

[13] J. a. H. Friedman, Trevor and Tibshirani, Robert, The elements of statistical learning vol.1: Springer series in statistics Springer, Berlin, 2001.

[14] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, Ali A. Gorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", 2009 IEEE

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

[16] 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

[17] 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.

[18] Naeem Seliya , Taghi M. Khoshgoftaar "Active Learning with Neural Networks for Intrusion Detection" IEEE IRI 2010, August 4-6, 2010, Las Vegas, Nevada, USA 978-1-4244-8099-9/10/$26.00 ©2010 IEEE

[19] 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

[20] Shingo Mabu, Member, IEEE, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa, Member, IEEE "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming" IEEE, JANUARY 2011

[21] Liu Hui, CAO Yonghui "Research Intrusion Detection Techniques from the Perspective of Machine Learning" - 2010 Second International Conference on MultiMedia and Information Technology 978-0-7695-4008-5/10 $26.00 © 2010 IEEE

[22] 1sundus juma, 1zaiton muda, 1m.a. mohamed, 2warusia yassin "machine learning techniques for intrusion detection system: a review" Journal of Theoretical and Applied Information Technology 28th February 2015. Vol.72 No.3.

[23] Abdulaziz Aborujilah1 and Shahrulniza Musa2 "Cloud-Based DDoS HTTP Attack Detection Using Covariance Matrix Approach" Hindawi Journal of Computer Networks and CommunicationsVolume 2017, Article ID 7674594, 8 pages https://doi.org/10.1155/2017/7674594

[24] Solane Duquea*, Dr.Mohd. Nizam bin Omarb "Using Data Mining Algorithms for Developing Model for Intrusion Detection System (IDS)" Elsevier, Conference Organized by Missour University of Science and Technology 2015-San Jose, CA www.sciencedirect.com

[25] Anushka Srivastava, Avishka Agarwal, Gagandeep Kaur "Novel Machine Learning Technique for Intrusio Detection in Recent Network-based Attacks" 2019 4th International Conference on Information Systems and Computer Networks (ISCON) GLA University, Mathura, UP, India. Nov 21-22, 2019

[26] Sang-Hyun Choi and Hee-Su Chae "Feature Selection using Attribute Ratio in NSL-KDD data" International Conference Data Mining, Civil and Mechanical Engineering (ICDMCME'2014), Feb 4-5, 2014 Bali (Indonesia) http://dx.doi.org/10.15242/IIE.E0214081