**NETWORK INTRUSION DETECTION USING RANDOM**

**FORESTS, DECISION TREES AND KNN**

***ABSTRACT:***

Cyber physical systems (CPSs) are large scale, geographically dispersed, federated, heterogeneous, life-critical systems that comprise sensors, actuators and control and networking components. A device or software application that monitors a network or systems for malicious activity or policy violations is known as intrusion detection system. There are many types of intrusions and New attack techniques are coming out each month and the IDS technology must adapt to these rapid changes. In this we implemented random forests to detect the intrusion detection. Mainly they are categorised into two types i.e anomaly and normal. In this we calculate the accuracy of the classification and also accuracy, detection rate, false alarm for every individual method.

***KEYWORDS:***

Intrusion, KDDCUP 99, random forests, accuracy.

***INTRODUCTION:***

Intrusion detection (ID) is a type of security management system for computers and networks. An ID system gathers and analyzes information from various areas within a computer or a network to identify possible security breaches, which include both intrusions (attacks from outside the organization) and misuse (attacks from within the organization). The most common classifications are network intrusion detection systems (NIDS) and host-based intrusion detection systems (HIDS).

Network intrusion detection systems (NIDS) are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. It performs an analysis of passing traffic on the entire subnet, and matches the traffic that is passed on the subnets to the library of known attacks. Once an attack is identified, or abnormal behavior is sensed, the alert can be sent to the administrator. An example of an NIDS would be installing it on the subnet where firewalls are located in order to see if someone is trying to break into the firewall. Ideally one would scan all inbound and outbound traffic, however doing so might create a bottleneck that would impair the overall speed of the network.

We used KDDCUP data set in this classification. The original KDD DATA SET contains 4898431 instances with 41 features. Each instance will fall into either attack type or normal type. Every attack falls into four categories. They are: DOS, R2L, U2R and PROBE.

***LITERATURE SURVEY:***

Nowadays people are very much bothered about privacy,confidentiality,integrity and availability. Basically intrusion is something which causes interruption to above terms. Biswanath Mukherjee [1] described intrusion detection as follows:

*“ Intrusion detection is a new, retrofit approach for providing a sense of security in existing computers and data networks, while allowing them to operate in their current “open” mode. ”*

Previously system administrators sit in front of their system and used to detect intrusions by noticing. But when coming to our present huge internet system it is impossible to notice all the intrusion manually, so there is a need for intrusion detection systems which can effectively deploy in large networks. But it is not a easy task because the attacks which the systems encounter are increasing and our systems are unable to detect new attacks.

Many researchers and scholars developed many intrusion detection techniques and divided Intrusion Detection Systems(IDS) into three types. They are as 1.Host based IDS 2.Network based IDS 3.Application based IDS.

Susan C. Lee [2] uses simulated network and IDS. The IDS composed of a hierarchy of back propagation neural networks. They used vector method and neural networks to to summarize overall TCP status while distinguishing among the type of anomalies. Constantine [3] stated a statistical anomaly approach. They proposed the protocol of a Hierarchical Intrusion Detection System, Generalised Anomaly and Fault Threshold System which uses statistical processing and neural network classification to detect network faults and attacks.

Jiong [4] stated that In anomaly detection, novel intrusions are detected by the outlier detection mechanism of the random forests algorithm. After building the patterns of network services by the random forests algorithm, outliers related to the patterns are determined by the outlier detection algorithm. Chee-Wooi [5] described a new substation anomaly detection algorithm that can be used to systematically extract malicious “footprints” of intrusion-based steps across substation networks. An impact factor is used to evaluate how substation outages impact the entire system.

Robert Mitchell [6] proposed a behavior-rule specification-based IDS technique for intrusion detection of physical devices. Monowar [7] provides a structured and comprehensive overview of various facets of network anomaly detection so that a researcher can become quickly familiar with every aspect of network anomaly detection and discussed several evaluation criteria for testing the performance of a detection method or system.

Robert [8] proposed and analyze a behavior-rule specification-based technique for intrusion detection of medical devices embedded in a medical cyber physical system (MCPS). Anna L [9] presents the results of a literature survey of machine learning (ML) and data mining (DM) methods for cyber security applications.

Maciej [10] investigate the potential for a self-organizing, nonparametric distributed coordination framework inspired by those observed naturally in colonies of honey bees to provide dynamic individual detection thresholds for anomalous event pattern detection on networks. Jyothsna [11] elaborates the foundations of the main anomaly based network intrusion detection technologies along with their operational architectures and also presents a classification based on the type of processing that is related to the “behavioral” model for the target system.

Abhrajit [12] present Real-Time Flow Filter (RTFF) a system that adopts a middle ground between coarse-grained volume anomaly detection and deep packet inspection. RTFF was designed with the goal of scaling to high volume data feeds that are common in large Tier-1 ISP networks and providing rich, timely information on observed attack.

***RANDOM FORESTS:***

Breiman (2001) proposed random forests, which add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node.

The original KDD DATA SET contains 4898431 instances with 41 features. But training the whole data is difficult for our machines and so we took 2% of the data and train 70% of the data and remaining 30% of the data is used for the training.

In this we implement random forests which is made up of decision trees. In this features are selected arbitrarily after each split, this ensures a higher classification power and greater efficiency.

Random Forest Classification ( )

1.data ←read data set

2.data ←convert\_features\_text\_to\_integral (data)

3.(train\_features,train\_intrusion\_categories)←training\_function()

4.(test\_features,test\_intrusion\_categorie)←testing\_function()

5.model←randomforests\_train(train\_features,train\_intrusion\_categories)

6.classification\_result← randomforests\_classify (train\_features)

7.Accuracy ← (TP+TN) / (TP+TN+FP+FN)

8.Detection\_Rate ← TP / (TP+FP)

9.False\_Alarm ← FP / (FP+TN)

***KDD CUP DATA SET:***

This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment. The original KDD DATA SET contains 4898431 instances with 41 features. Each instance will fall into either attack type or normal type. Every attack falls into four categories. They are: DOS, R2L, U2R and PROBE.

*DENIAL OF SERVICE (DOS) :*

This occurs when an attacker takes action that prevents legitimate users from accessing targeted computer systems, devices or other network resources. The attacks which comes under this DOS are: back, land, neptune, pod, smurf, teardrop.

*USERS TO ROOT ATTACK(U2R) :*

These attacks are exploitations in which the hacker starts off on the system with a normal user account and attempts to abuse vulnerabilities in the system in order to gain superuser privileges. The following attacks come into this category: buffer\_overflow, loadmodule, perl, rootkit.

*REMOTE TO LOCAL ATTACK(R2L) :*

It is an attack in which a user sends packets to a machine over the internet, which s/he does not have access to in order to expose the machines vulnerabilities and exploit privileges which a local user would have on the computer. The following attacks comes under this: ftp\_write, guess\_passwd, imap, multihop, phf, spy, warezclient, warezmaster.

*PROBE:*

Probing is an attack in which the hacker scans a machine or a networking device in order to determine weaknesses or vulnerabilities that may later be exploited so as to compromise the system. It includes following attacks: ipsweep, nmap, portsweep, satan.

|  |  |  |
| --- | --- | --- |
| **OUTPUT** | **COUNT** | **PERCENTAGE** |
| NORMAL | 13484 | 13.763677 |
| DOS | 54351 | 55.47831 |
| R2L | 167 | 0.17046 |
| U2R | 6 | 0.00612 |
| PROBE | 569 | 0.58080 |

*TABLE: percentages of the output in the trained set.*

TRUE POSITIVE (TP) is something where the actual output and the predicted output is yes, TRUE NEGATIVE (TN) occurs when the actual output is wrong and the predicted output is also wrong, FALSE POSITIVE (FP) implies the predicted value is yes but it is actually not, FALSE NEGATIVE (FN) implies the prediction is no and the actual value is yes.

|  |  |  |
| --- | --- | --- |
|  | PREDICTION **YES** | PREDICTION **NO** |
| ACTUALLY **YES** | TRUE POSITIVE (TP) | FALSE NEGATIVE (FN) |
| ACTUALLY **NO** | FALSE POSITIVE (FP) | TRUE NEGATIVE (TN) |

The accuracy, detection rate and the false alarm rate can be calculated by using the following formulae:

Accuracy = (TP+TN) / (TP+TN+FP+FN)

Detection\_Rate = TP / (TP+FN)

False\_Alarm = FP / (FP+TN)

|  |  |
| --- | --- |
| Method | Accuracy |
| Random forests | 99.9455597142 |
| Decision trees | 99.8775093569 |
| knn | 81.3065668595 |

*TABLE:accuracy of each method.*

The accuracy for probe, dos, u2r, r2l in every method is found. In random forests for every attack the accuracy is more compared to decision trees and KNN.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Probe | dos | u2r | r2l |
| Pandom forests | 99.95008 | 99.99311 | 99.96519 | 99.93130 |
| Decision trees | 99.86642 | 99.98966 | 99.96509 | 99.79356 |
| knn | 62.82208 | 94.56629 | 78.65961 | 25.27379 |

*TABLE: accuracy for types of attacks*

***CONCLUSION:***

Intrusion Detection Systems (IDS) have become crucial components in computer and network security. There are many approaches to implement the intrusion detection and we implemented random forests algorithm on the KDDCUP 99 dataset, we used 41 features which are present in the KDD CUP dataset to classify the type of attack and by using random forests we got accuracy of 99.94555 %, by using decision trees we got 99.877% and by using KNN algorithm we got 81.306%. out of all random forests method gives the more accuracy.

***REFERENCES:***

[1]. Network Intrusion Detection by Biswanath Mukherjee, L. Todd Heberlein, and Karl N. Levitt.

[2]. Training a Neural-Network Based Intrusion Detector to Recognize Novel Attacks Susan C. Lee and David V. Heinbuch

[3]. Network Intrusion and Fault Detection: A Statistical Anomaly Approach Constantine Manikopoulos and Symeon Papavassiliou, New Jersey Institute of Technology

[4]. Random-Forests-Based Network Intrusion Detection Systems Jiong Zhang, Mohammad Zulkernine, and Anwar Haque

[5]. Anomaly Detection for Cybersecurity of the Substations Chee-Wooi Ten, Member, IEEE, Junho Hong, Student Member, IEEE, and Chen-Ching Liu, Fellow, IEEE

[6]. Behavior-Rule Based Intrusion Detection Systems for Safety Critical Smart Grid Applications Robert Mitchell and Ing-Ray Chen, Member, IEEE

[7]. Network Anomaly Detection: Methods, Systems and Tools Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita

[8]. Behavior Rule Specification-Based Intrusion Detection for Safety Critical Medical Cyber Physical Systems Robert Mitchell and Ing-Ray Chen, Member, IEEE

[9]. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection Anna L. Buczak, Member, IEEE, and Erhan Guven, Member, IEEE

[10]. Hive Oversight for Network Intrusion Early Warning Using DIAMoND: A Bee-Inspired Method for Fully Distributed Cyber Defense Maciej Korczyński, Ali Hamieh, Jun Ho Huh, Henrik Holm, S. Raj Rajagopalan, and Nina H. Fefferman

[11]. A Review of Anomaly based IntrusionDetection Systems V. Jyothsna V. V. Rama Prasad

[12]. Managing High Volume Data for Network Attack Detection Using Real-Time Flow Filtering Abhrajit Ghosh, Yitzchak M. Gottlieb, Aditya Naidu, Akshay Vashist, Alexander Poylisher, Ayumu Kubota, Yukiko Sawaya, Akira Yamada

[13]. Machine learning techniques for web intrusion detection – a comparison Truong Son Pham, Tuan Hao Hoang, Van Canh Vu.

[14]. Anomaly Detection in Network Traffic using K-mean clustering R. Kumari Sheetanshu M. K. Singh R. Jha N.K. Singh

[15]. Intelligent Feature Selection Method rooted in Binary Bat Algorithm for Intrusion Detection Adriana-Cristina Enache , Valentin Sgârciu and Alina Petrescu-Niţă

[16]. Reputation Prediction of Anomaly Detection Algorithms for Reliable System Guy Leshem, Esther David, Michal Chalamish

[17]. Machine-Learning-Based Feature Selection Techniques for Large Scale Network Intrusion Detection O. Y. Al-Jarrah, A. Siddiqui, M. Elsalamouny , P. D. Yoo , S. Muhaidat, K. Kim

[18]. Network Intrusion Detection System Based On Machine Learning Algorithms Vipin Das , Vijaya Pathak , Sattvik Sharma , Sreevathsan , MVVNS.Srikanth , Gireesh Kumar T

[19]. Random Forest , Support Vector Machine And Nearest Centroid Methods For Classifying Network Intrusion Sanjiban Sekhar Roy , Dishant Mittal1 , Marenglen Biba , Ajith Abraham

[20]. An Improved Method To Detect Intrusion Using Machine Learning A Lgorithms Urvashi Modi and Anurag Jain.

[21]. Machine Learning Techniques for Intrusion Detection on Public Dataset Udaya Sampath K. Perera Miriya Thanthrige, Jagath Samarabandu, Xianbin Wang.