JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

K-MEANS ALGORITHM BASED DISTRIBUTED INTRUSION DETECTION SYSTEM FOR CLOUD COMPUTING ENVIRONMENT

Anthony Raj.A[1]K.Logesh[2] Dr. D. Sumathi[3]

[1]Research Scholar,Department of Computer Science,Kuppam Engineering College,Kuppam, Andhra Pradesh, India. anthonyraj171@gmail.com
[2]Associate Professor, HOD Dept. of CSE.Kuppam Engineering College, Kuppam, Andhra Pradesh, India. hod_cse@kec.ac.in
[3]Associate Professor, HOD Dept. of CSE, Kuppam Engineering College, Kuppam, Andhra Pradesh, India. d.sumathisenthil@gmail.com

Abstract
An intrusion spying or Detection system (IDS) is a software application program that managesnetwork or system activities for vicious activities or policy violations anddevelops reports to a management broadcast station.Absolute majority of research is going on neural network and machine learning technique for discovering intrusions. In this paper, we present aunsupervisedmodel integrated with anomaly detection to construct an effective intrusion detection system. Here we make use of K-means clustering algorithm along with Cloud based applicationnatural world. Thesuggested system is trained to discover abnormal behavior of the intrusions. The solutions show the system has high attack detection accuracy
Keywords—IDS, CIDD-001 data set, Detection Rate, K-means Algorithm.

1. Introduction

Intrusion detection system (IDS) is a tool that is being used to protect organization from attacks from different sources. It is expected that IDS can deal large amount of data without affecting performance and without dropping data and can detect attacks reliably without giving false alarms. An IDS is broadly classified as:

A. **Misuse Based System:** In misuse-based IDS, detection is performed by looking for the exploitation of known attacks in the system, which can be described by a specific pattern or sequence of events or data. That means these systems can detect only known pattern attacks for which they have a defined signature.

B. **Anomaly based system:** In anomaly based IDS, detection is executed by noticing changes in the patterns of utilization or behavior of the system. If it deviations from the normal profiles of the system is found then they are detected as intrusions and alter is raised in the system. The main gain of anomaly detection systems that they can find antecedently unknown attacks.

2. Literature Review:

Most of the IDS developers proposed a traditional system based on Snort to detect anomalies in the network traffic. The other methodologies includes in prior approach was implementing both supervised and unsupervised machine learning techniques which included applied Machine learning Hybrid classifiers like K-mean clustering and Naïve Bayes, Decision tree, Random forest, back propagation Neural Networks, fuzzy classification methods, synergetic neural network approach and others data mining techniques like signature apriori Algorithm were extensively used to detect network intrusions.

In recent authors have applied Neural Network to detect intrusions in the cloud and most of the developed systems evaluated using KDD data set yielding satisfactory attack detection results. In all these approaches some are good at detecting both kinds of signature and anomaly intrusions and some are very bad at detecting specific attacks but considered all attacks as anomalies.

However, most of the IDS systems considered to be following the traditional approach and they are limited for standard information system infrastructure or private cloud and need to be face with new distributed and sophisticated attacks. Therefore, it is recommended to consider for further scalable and apply distributed techniques to adapt for the public cloud IDS.

Anomaly detection has appealed the attention of many investigators to get the most effectivethe weakness of signature-based IDSs in detecting novel attacks, and KDDCUP'99 is the mostly widely applied data set for the evaluation of these systems. After performing on a statistical investigation on this data set, we found out two significant issues which highly affect the performance of evaluated systems, and results in a very poor evaluation of anomaly detection approaches. To compute these consequences, we have suggested a new data set,

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

CIDDS-001(Coburg Intrusion Detection Data Set), which consists of selected records of the complete KDD data set and does not tolerate from any of mentioned defects.[3]

Methodology

K-means Algorithm one of the simplest method of unsupervised learning techniques that is used to discretize the features and analyze the patterns that are normal and intrusion clusters in a given data set.

A. The proposed Intrusion Detection System Method

The suggested intrusion detection system is compiled of the following components:

a). **Cloud network traffic data collector module**: This allows to capture incoming network traffic data to the Cloud services hosted on the same infrastructure. The module executes the conquer in a 5 minutes time window basis. For each time window the captured network traffic data is stored in a storage server.

b). **Network traffic data preprocessing module:** preprocesses the captured network traffic data during each time window. This module is planned to be inserted near to each Cloud network traffic data collector module. The preprocessing tasks are implemented in MapReduce and deployed on the Hadoop cluster.

c). **Anomaly detection module:** classifies the network traffic data of each time window into normal or abnormal traffic. K-Mean clustering model constructed from the CIDDS-001 dataset is used to observe malicious traffic. This module is distributed side by side to each Cloud network architecture. Data preprocessing module which allows speeding up the anomaly detection with minimum resources.

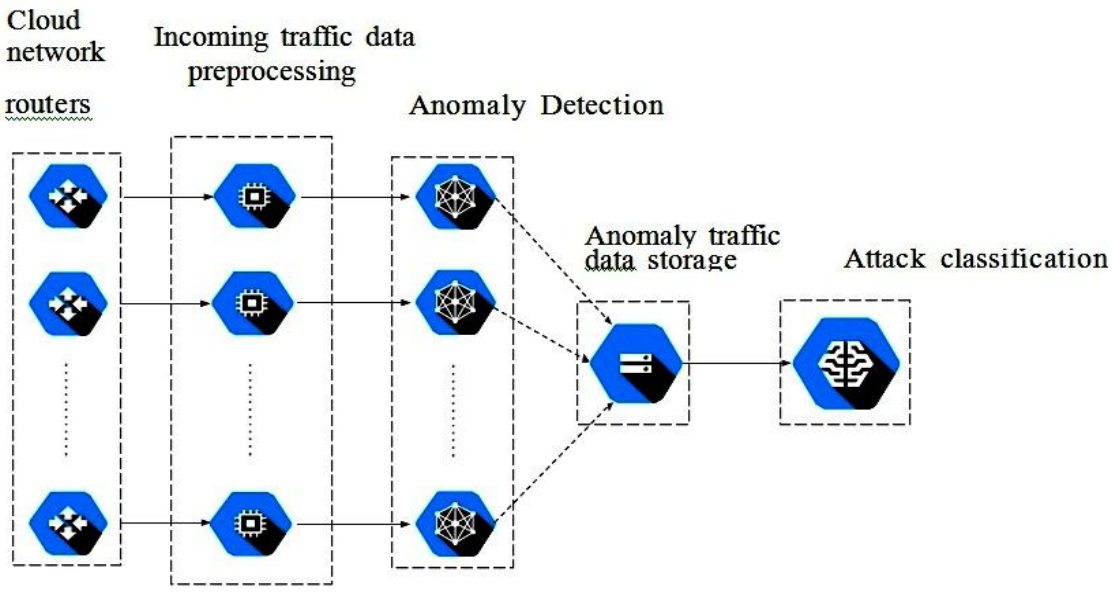
d). Network traffic synchronization module: synchronizes the malicious network traffic of each Cloud router side to a centralized server. The synchronization is Carry out an action for each time window separately.

e). **Attacks traffic categorization module:**separates the malicious network traffic data synchronized to the central storage server. This allows detecting the type of each attack and early performing convenient mitigation actions for it.

3. Architecture design of the system

This section brings out the process accompanied by the proposed IDS to discover attacks. For each time window the entering network traffic data to the Cloud is caught at each one of the edge routers. The captured data are then preprocessed and passed to a first anomaly detection. This allows detecting and blocking suspected traffic while allowing access to the normal traffic. The suspected traffic at each network router side are synchronized to a central server.

The intrusions detection process and the flowchart of the entire proposed IDS are illustrated in figure 1.



4. Experiment set up design and implementation Details

To begin with in Cloud network routers the incoming IP packets are captured by the IPpacket sniffers like snort are then subjected to decoding &analysis of features that features likesource ip, hop limit, destination ip, length of the packet, SYN flag, ACK flag, RST flag, port number. Etc. Can be used for classification after they are extracted and discretized using K-means clustering algorithm.

To facilitate the experiments, we used eclipse java and weka tool to implement the algorithms on a PC with 64-bit window 7 operating system, 4GB RAM and a CPU of Intel core i3-4010U CPU with 1.70GHz. Data come

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

from CIDD-001 data set. The table lists the number of instances available in the whole dataset, 10% of CIDD-001 dataset. The analysis is performed by using K-means algorithms. We adopt K-means algorithm to bring forth heterogeneous dataset to nearly homogeneous dataset..

5. CIDD-001 (Coburg Intrusion Detetion Data Set

In this research project we have used CIDD-001 (Coburg Intrusion Detetion Data Set) dataset which is an up-to-date labeled flow-based dataset created by M.Ring et.al[19] in a Cloud environment based on Open Stack Platform. This surroundings includes various clients, emulated using a set of python scripts, and typical servers including E-Mail server, Web server, etc.

The dataset comprises realistic normal and attack traffic admitting important bench mark of network intrusion discovery systems in a cloud environment. The dataset is split into four parts each is created during a week.

Features of the CIDDs-001 dataset

Feature	Description
1. Attack Description	Provides additional information about the set attack parameters (e.g. the number of tried password guesses for SSH-Brute-Force attempts)
2. Attack ID	Unique attack id. All flows which go to the same attack carry the same attack id.
3. Attack Type	Type of Attack (PortScan, dos, bruteForce,)
4. Bytes	Number of flow
5. Date first checked	Start time flow first checked
6. Class	Class label (normal, attacker, victim, suspicious or unknown)
7. Dest IP	Destination IP Address
8. Dest Port	Destination Port
9. Duration	Duration of the flow
10. Flags	OR concatenation of all TCP Flags
11. Packets	Number of Transmitted packets
12. Proto	Transport Protocol (e.gICMP,TCP, or UDP)
13. Src IP	Source IP Address
14. Src Port	Source Port

6. Unsupervised Learning Techniques

Clustering Techniques: Clustering partitions the data set into several clusters or equivalence classes or clusters. The property of a cluster or equivalence of a class is similarity among given members of class in a given cluster. There are several measures of measures of similarity and to compute the similarity. The given K-Means Clustering algorithm proposed in our research work.

Iterative Partitioned clustering algorithm:

Given n elements x1,x2…xn and k clusters, each with a center.[centroid]

1. Assign each element to its nearest cluster center
2. After all designations have been built, calculate the cluster centroids for each of the cluster
3. Repeat the above two steps with new centroid until the algorithm converges

7. Results of the Model

The experiment aims to research the performance of the intrusion detection system using K-Mean clustering techniques. By selecting Different threshold and cluster radius, the effect on results could be observed. Data clustering results greatly affected by changing the cluster radius. The More network behavior pattern classes will be generated as the cluster radius reduces. The fewer network behavior pattern classes will be generated as the cluster radius increases.

Data clustering table

Cluster Radius	Network normal behavior pattern class	Network abnormal behavior pattern class
1	0	145
2	1	136
3	3	128
4	6	112
5	9	91

The Results show that the Model detects the given IP packets whether the cluster is Malicious or Normal. And also detects the normal and abnormal behavior pattern of the given intrusion data set.

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

8. Contributions in Information security and identification of system vulnerabilities.

C1. Identification of novel attacks:

The qualitative analysis in this thesis showed that 61% of the vulnerabilities were of web application. Attackers more often found to be exploiting the trusted entities and intended to break the integrity of the system. The discovered knowledge found to be very useful in monitoring the trusted entities on regular basis and provides much information on preventive measures and discussed more in detail in relevant chapters.

C2.Challenges of managing of an Intrusion Detection System in the Enterprise:

The qualitative analysis of the Explored Challenges of Managing an Intrusion Detection System (IDS) in the Enterprise showed that control measures have to be taken in monitoring the internal network and in managing the intrusion flood of alerts, Following up on the intrusion alerts, Creating Actionable Reports for Follow-up and finally program improvements which are recommended in the thesis are very useful for researchers for design and implementing the real time IDS and discussed more in detail in relevant chapters.

9. Limitation of the Study

Even though IDS' tools considered as integral part of thorough and complete security system. IDS's don't fully guarantee security of the system. By just implementing security policy, data encryption, vulnerability assessments, access control, user authentication, and firewalls we can only enhance network safety of the system but we cannot fully trust on IDS. The design of IDS techniques hold changes as the course of data network advanced attack methods gets updated day by day. Hence there is no single complete solution is found for noticing the intrusion in data network. In general IDS systems are building complex and it is an ongoing cognitive operation.[4]

10. Conclusions & Recommendations

Intrusion detection systems try to identify attacks or intrusions by analyzing network data (cloud network-based systems) or operating system and application logs(host-based systems), possibly in real-time. These systems either look for normal of well-known attacks in the data (misuse detection) or try to find irregularities in the data by first building the normal profile of the system under observation and then discovering deviations from this profile(anomaly detection). Anomaly detection is very significant due to the unfitness of misuse detection techniques in detecting unidentified attacks. In our approach we used only anomaly based detection system

Further study ofneural network methods arehighly recommended which are promising techniques and hybrid model for both kinds of signature and anomaly based intrusions and used for multi classification problems.

Reference

[1]. Mohammed Sammany, MarwaSharawi, Mohammed El-Beltagy and ImaneSaroit, "Artificial Neural Networks Architecture For Intrusion Detection Systems and Classification of Attacks", Cairo University, Egypt 2019

[2]. PrzemysławKukiełka, ZbigniewKotulski, "Adaptation of the neural network-based IDS to new attacks detection",Research and Development Department, Polish TelecomInstitute of Telecommunications, Warsaw University of Technology.

[3].KhaledAlsabti, Sanjay Ranka, Vineet Singh, "An Efficient K-Means Clustering Algorithm", 2018

[4].Moustafa, N., & Slay, J. (2015, November). The significant features of the UNSW-NB15 and the KDD99 data sets for Network Intrusion Detection Systems. In Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2015 4th International Workshop on (pp. 25-31). IEEE. [17] KDD Cup 1999. (2014, Nov.) [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/.

[5]. D. A. Fernandes, L. F. Soares, J. V. Gomes, M. M. Freire, P. R. Ina´cio, Security issues in cloud environments: a survey, International Journal of Information Security 13 (2) (2018) 113–170.

[6]. P. Mell, T. Grance, The nist definition of cloud computing.S. Iqbal, M. L. M. Kiah, B. Dhaghighi, M. Hussain, S. Khan, M. K. Khan, K.-K. R. Choo, On cloud security attacks: A taxonomy and intrusion detection and prevention as a service, Journal of Network and Computer Applications 74 (2017) 98–120.

[7]. Wikipedia, 2016 dyncyberattack[Online; accessed 10-November-2018)].theguardian, Ddos attack that disrupted internet was largest of its kind in history, experts say[Online; accessed 10-April-2018)].

[8]. D. Miner, A. Shook, MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems, " O'Reilly Media, Inc.", USA, 2019.