ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

A Survey of Intrusion Detection System Using Different Data Mining Techniques

Trupti Phutane, Apashabi Pathan

Dept. of Computer Engineering, G.H.Raisoni College of Engineering & Management, Wagholi, India

ABSTRACT: Now- a- day authentication is of prime concern. We have to look every aspects of security in every transaction to avoid the treats & intruders. In this paper, we will discuss the existing intrusion detection systems with data mining approach such as intrusion detection system using association rule mining [6] and intrusion detection system using event correlation data mining [5]. We also discuss the proposed system [3], [4], [10] in which the intrusion detection is a data analysis process, rather than previous approaches like knowledge engineering processes. In association rule mining, we firstly capture the network data using sniffers. The captured data will be filtered so that non relevant data is removed from analysis and finally we will extract the features which will be associated with given datasets where as in event correlation data mining method we will maintain the logs of every network system and on the basis of "event data" stored in logs we will try to maintain the link among them, if we found any suspected activity we will deny the access for data. In both the techniques we are only concerned with the allowing & denial of access but in proposed system we are maintaining the decision tree by analysis of data and its attributes rather than just guessing or finding any relations with previous data.

KEYWORDS: Threats, Intruders, association rule mining, event correlation data mining, decision tree

I. INTRODUCTION TO EXISTING INTRUSION DETECTION SYSTEMS

Data mining is the process of finding the unknown pattern from given set of patterns [6], [7]. In case of intrusion detection system, we use the concept of data mining we will find out the pattern which will track all users activity to find out the intruders. In existing system we are focusing on knowledge engineering processes in which the decisions are taken on the basis of some fixed rule. Mainly intrusion detection system is divided in two broad categories i.e. intrusion detection system using association rule mining and intrusion detection system using event correlation data mining.

A. Intrusion Detection System Using Association Rule Mining

In this method we are keeping track of every incoming packet of networkand will associate the collected log with different data sets like KDD cup, DARPA etc [1], [2]. We have to make analysis of network before intrusion detection on the basis of some assumptions like

i. FTP connections with some other designated ports are treated as attack
ii. Connections to same IP repeatedly is treated as attack
iii. Any remote login connection is treated as attack
iv. Any connection flow from any source port to destination port 445 (Worm) is treated as attack.

On the basis of such assumptions we filter the data as normal data or attacker data.This notion is called as pattern creation we will create this pattern on database if any user tries to access the system and if it is not the one who should be there then an alert is generated from it[2].
After pattern creation following are the phases of IDS system

i. **Data Collection phase**: **-** Capturing is performed using a special sniffer. The sniffer is used to capture all packets and store its header in database. The captured features for every packet includes source & destination IP, source/destination port, protocol, number of bytes, service type & flags

ii. **Data filtering phase**: **-** Captured data are filtered in order to remove traffic non relevant for analysis, this study concentrates on different network protocol like TCP, UDP, ICMP because majority of network connection falls under these protocols.

iii. **Feature Extraction Phase**: **-** In this phase, new features are extracted for association rule mining. Mostly extracted features are time and connection oriented. Then network connections are captured first & then processed offline within the system for granting the access of data or not

After feature Extraction association rule mining will be applied using following models
Known attack detection model: **-** It classifies network data as normal or intrusion based on the basis of labeled training data. For classification purpose decision tree classification algorithm is used to classify the data internally into multiple sets and uses combination of classifier to increase classification accuracy.

Anomaly Detection Model: **-** clustering technique is used to discover new attacks which were not detected in previous phase. It uses two step clustering algorithms.

Attack Summarization using association pattern: **-** discovered attack & suspicious connections are summarized using apriori algorithm which describe the features detected outlier to assist the analyst in generating new signature. Following diagram shows the proposed intrusion detection system using association rule mining.

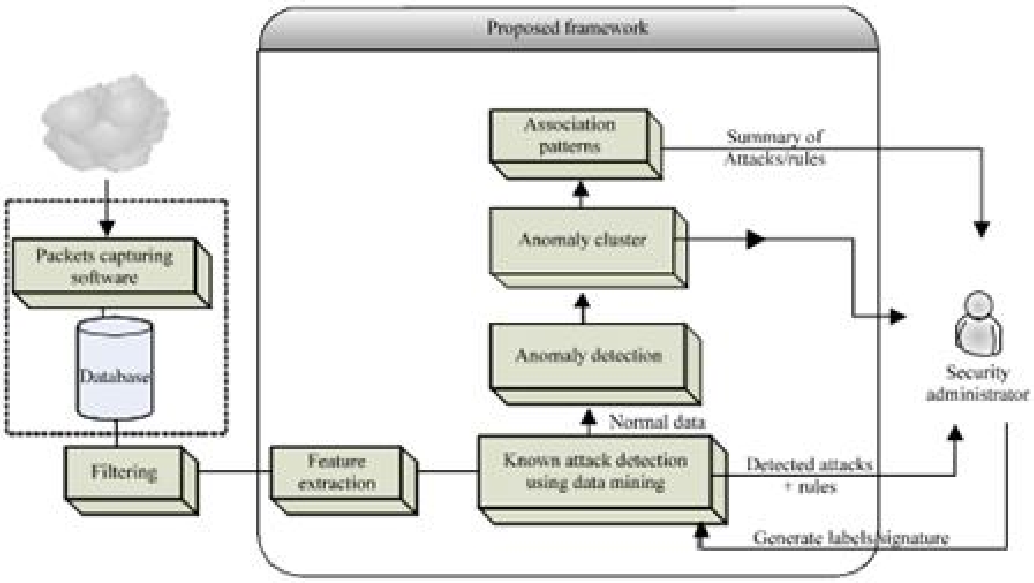

Figure1. Proposed IDS framework using association rule mining

B. Intrusion Detection System Using Event correlation data mining

In previous system, guessing the new intruders is difficult because of signature based method and another problem is high false alarm rate. These problems effects on efficiency of IDS so in this system we are trying to focus on these problems .To minimise false alarm , we can use Data mining using Event Correlation Technique (ECT) for Network Intrusion Detection in an adaptive manner, so that NIDS can identify and add new pattern on fly into the signature database. Event correlation analysis is useful for reducing false Positive alarms. In this proposed system we are trying to develop a system that will reduce the false alarm rate and add new signature on fly. In this system we will maintain the log of every activity of every user of every system within the network. To guess the intruder or threat we will compare those logs together and find out the relation among them. If any suspicious activity found, we will report it to be an attack. In this method by correlating the logs we can also guess the future attacks which were least possible in association rule mining.

Following figure shows the proposed system of Intrusion detection using event correlation technique

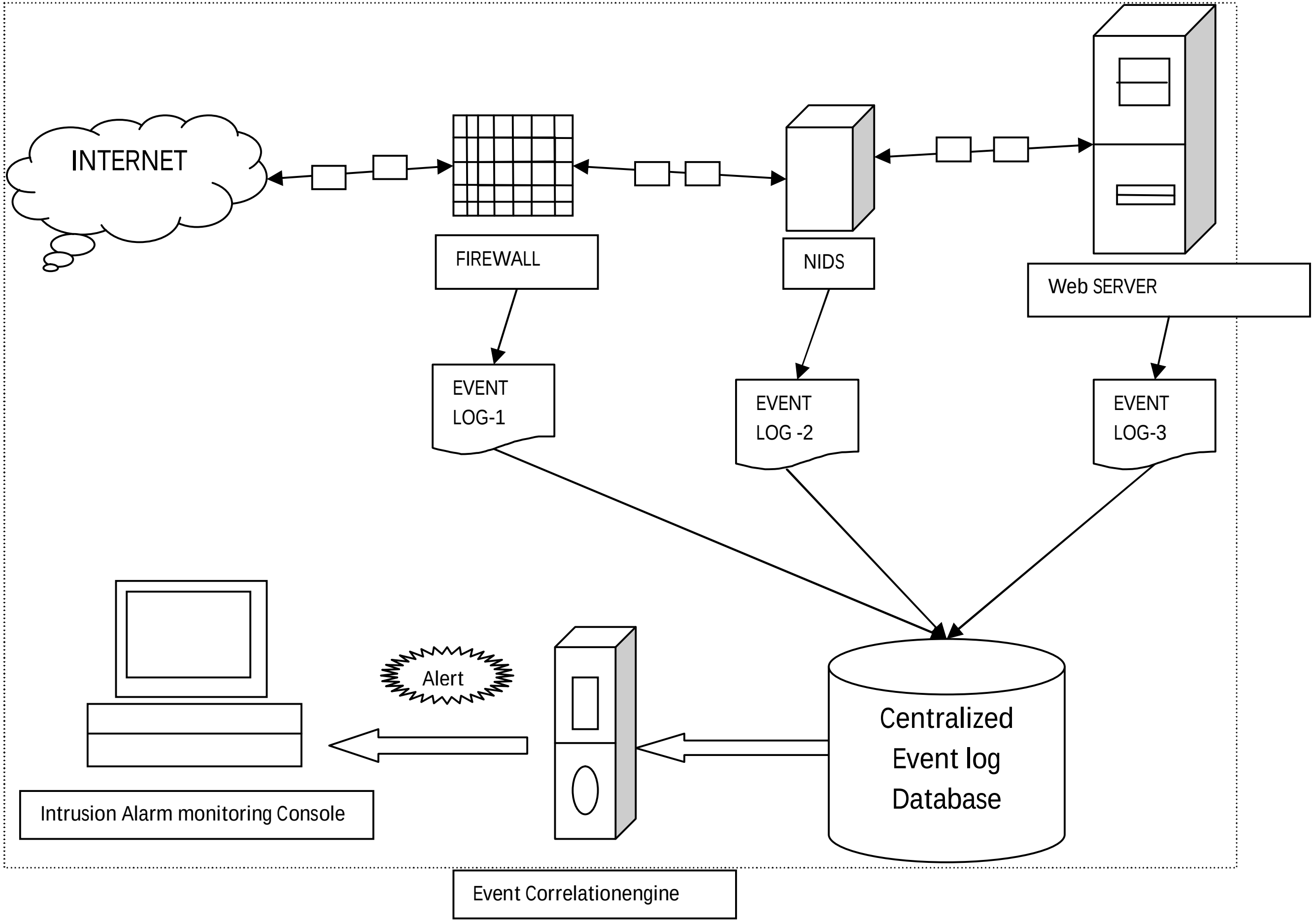

Figure2. Proposed IDS framework using event correlation technique

As shown in figure, to detect the intruders we collect the event logs from different workstations like firewall, NIDS and web server etc[6],[7]. those logs will be correlated with each other by event correlation engine. Event correlation engine will find the relations among the logs & if any suspicious activity find out it will report an attack and also predicts the future threats & attacks

Algorithm:
Input: simple alerts generated by the IDS (Event Logs)
Output: Correlated Attack Graph
Methods:

```
1- Let CAG (N, E) = null
2- Map elementary alerts to hyper alerts {ho, h1, h2,....}
3- Let h0 be the isolated hyper-alert


4- For k=1 to i
If
a.      at least one Prereq(hi+1) Є Conseq(hi)U Prereq(hi+1) Є ExConseq(hi)
   b.   Vul (hi+1) exists.
   c.   Conseq (hi).End_time>= Prereq (hi+1).Start_time U
        ExConseq(hi).End_time>=Prereq(hi+1).Start_time
Then
Add CAG (nhi, nhi+1)
5- Return result.
```

II. LITERATURE SURVEY

In case of intrusion detection we are testing the system on standard data set such as KDD cup and DARPA data set. We will see each of them in details

A.KDD cup-99 Dataset

This data set was used for The Third International Knowledge Discovery and Data Mining Tools Competition, The competition task was to build network intrusion detection system to distinguish the bad & good connections in other word we can say that with the help of this data set we can identify the intruders or attackers. This database contains a standard dataset to be audited, which includes a wide variety of intrusions simulated in military Applications.

B. KDD cup-2010 Dataset

This dataset represent a sampled snapshot of users' preferences for various items. The suggestion of items to users and the history of users' 'following' history. It is of a larger scale compared to other publicly available datasets ever released. Also it provides wealthier information in multiple domains such as user profiles, social graph, item category, which may hopefully evoke deeply attentive ideas and methodology. The users in the dataset, numbered in millions, are provided with information such as profile keywords; follow history, etc. for generating a good prediction model. To protect the privacy of the users, the IDs of both the users and the recommended items are anonymized as random numbers such that no classification is revealed. Furthermore, their information, when in Chinese, will be encoded as random strings or numbers, thus no contestant who understands Chinese would get advantages. Timestamps are provide if recommendation is required.

C. DARPA dataset

The Cyber Systems and Technology cluster (formerly the DARPA Intrusion Detection Evaluation Group) , under Defense Advanced Research Projects Agency (DARPA ITO) and Air Force Research Laboratory (AFRL/SNHS) sponsorship, has collected and distributed the first standard data set for evaluation of computer network intrusion detection systems. We have also coordinated, with the Air Force Research Laboratory, the first formal, repeatable, and statistically important evaluations of intrusion detection systems which are carried out in 1998 & 1999. These evaluations measured chance of detection and probability of false alarm for each system under test. These evaluations contributed significantly to the intrusion detection research field by providing direction for research efforts and an objective calibration of the technical state of the art. They are of interest to all researchers working on the general problem of workstation and network intrusion detection. The estimate was designed to be easy, to focus on technology issues, and to encourage the widest possible participation by untwining security and privacy concerns, and by providing data types that were used commonly by the majority of intrusion detection systems.

C. Intrusion detection Technique

As we know that, intrusion detection is the process of inspecting the computers or networks for unrecognized threats, activity, or file modification. IDS can also be used to inspect network traffic, thereby searching if a system is being targeted by a network attack such as a DOS attack. There are two basic types of intrusion detection: host-based and network-based. Each has a distinct approach for securing data, and each has distinct reward and disadvantages. In short, host-based IDSs examine data held on individual computers which serve as hosts, while network-based IDSs examine data exchanged between computers.

Host-Based IDS (HIDS)

These systems collect and examine data that created in a computer that hosts a service, such as a Web server. Once this data is aggregated for a given computer, it can either be examined locally or sent to a separate/central analysis machine. One example of a host-based system is programs that operate on a system and receive application or operating system audit logs. These programs are highly effective for detecting insider abuses. Residing on the trusted network systems themselves, they are close to the network's genuine users. If one of these users attempts illegal activity, host-based systems usually detect and collect the most pertinent information in the quickest possible manner. In addition to detecting unauthorized insider activity, host-based systems are also effective at detecting unauthorized file modification.

Network-Based IDS (NIDS)

As opposed to monitoring the activities that take place on a particular network, Network-based intrusion detection examines data packets which travel over the actual network. These packets are examined and sometimes compared with empirical data to verify their nature: malicious or benign. Because they are responsible for examining the network, Network-based intrusion detection systems (NIDS) tend to be more distributed than host-based IDS. Software, or appliance hardware in some cases, resides in one or more systems connected to a network, and are used to examine data such as network packets. Instead of analyzing information that originates and resides on a computer, network-based IDS uses techniques like "packet-sniffing" to pull data from TCP/IP or other protocol packets traveling along the network. This surveillance of the connections between computers makes network-based IDS great at detecting access attempts from outside the trusted network. In general, network-based systems are best at detecting the following activities:

- **Unauthorized outsider access:** When an unauthorized user logs in successfully, or attempts to log in, they are best tracked with host-based IDS. However, detecting the unauthorized user before their log on attempt is best accomplished with network-based IDS.
- **Bandwidth theft/denial of service:** These attacks from outside the network single out network resources for abuse or surplus. The packets that start/carry these attacks can best be noticed with use of network-based IDS.

IDS Techniques

Now that we have examined the two basic types of IDS and why they should be used together, we can investigate how they go about doing their job. In every case, there are four basic techniques used to detect the intruders namely anomaly detection, misuse detection (signature detection), target monitoring, and stealth probes.

III. PRPOPOSED SYSTEM

The proposed intrusion detection system is based on two important parts viz. Data Preprocessing and Analysis of pre-processing.

A.Data Preprocessing

Our proposed IDS System is represented in figure 3. Data preprocessing is used to translate the non-numeric value to numeric value. The information obtained by KDD Cup"99 is mixture of many system calls. A system call is a text based record to call the query. [1]

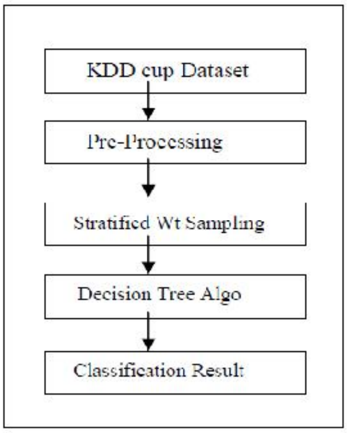

Figure3: Flowchart of Proposed Decision tree Approach for Intrusion detection

B.**Analysis of Pre-processing Step**

The proposed IDS system requires proper pre-processing activities as suggested [1]. Most of existing researches are conceptually deficient and also independent. Some researches provide adequate data formats, and some others are obvious on analysis. So we can say that each intrusion design has its different speciality so rather than considering a new analysis framework we can make the hybrid framework.

The obvious objective of our system analysis is accurate detection rates along with reduction of false filter rates. So, some issues are derived from these purposes:
(1) How to provide a best and well-organized computing data for IDS.
(2) How to filter false rates and improve detection rates.
(3) How to find out attack patterns and display suitable data types for administrators to make policies.

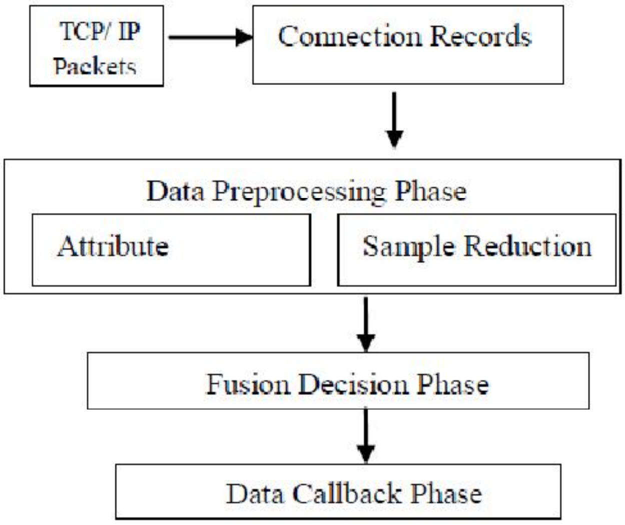

Figure 4: Framework of Integrated Decision System

So, our proposed system will enhance the intrusion detection by reducing the false rates & improves decision rates by taking advantages of association rule mining & event correlation data mining i.e. we are considering the benefits of independent as well as dependent frameworks in intrusion detection.

IV. CONCLUSION

The proposed system will develop an intrusion detection system for detecting the intrusion behavior normal or Attack using Decision tree and Stratified weighted Sampling as compared to existing technologies such as association rule mining and event correlation technique of data mining. A decision Tree generates to build the system more accurate for attack detection because it is taking advantage of both the frameworks explained in existing system. In this project, we need to use preprocessing step to KDD cup dataset which is classified in to three phase, data preprocessing phase, fusion decision phase and data call back phase. These techniques ensure the availability of our performance in terms of Accuracy Rate and Error rate. Stratified weighted sampling techniques to generate the samples from the original datasets and then apply the decision tree algorithm which overcomes the limitations of the ID3 algorithm. Hence the proposed method can be implemented for various datasets where size of data is large and result are very truthful with less Error rate than existing algorithm. Hence the CPU and memory utilization is decreased. Thus, proposed Approach is very apt and consistent for intrusion detection

REFERENCES

[1] Dr. R.C. Jain M.Tech, SATI Vidisha Director Devendra kailashiya, ‘Improve Intrusion Detection Using Decision Tree with Sampling’ SATI Vidisha.May 2012

[2] Sandhya Peddabachigari, Ajith Abraham, Johnson Thomas, ‘Intrusion Detection Systems Using Decision Trees and Support Vector Machines’ Department of Computer Science, Oklahoma State University, USA. June 2010

[3] Mrs. Snehal A. Mulay Department of Information Technology,Bharati Vidyapith's COE, Pune, India snehalmulay@gmaiLcom Prof. P. R. Devale HOD, Department of Information Technology Bharati Vidyapith's COE, Pune, India Prof. G.V. Garje HOD, Department of Computer and IT PVG's COET, Pune, ‘Decision Tree based Support Vector Machine for Intrusion Detection’ , India.2010

[4] Jashan Koshal, Monark Bag Indian Institute of Information Technology Allahabad, Uttar Pradesh-211012, India ‘Cascading of C4.5 Decision Tree and Support Vector Machine for Rule Based Intrusion Detection System’,Aug 2012

[5] 1K.P.Kaliyamurthie, 2D,Parameswari , 3DR. R.M. Suresh 1 Assistant Professor, Dept of IT, Bharath University. Chennai, Tamil Nadu-600073’Intrusion Detection System using Mimetic Algorithm Supporting with Genetic and Decision Tree’. Mar 2012

[6] K.V.R. Swamy, K.S. Vijaya Lakshmi Department Of Computer Science and Engineering V.R.Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India, ‘Network Intrusion Detection Using Improved Decision Tree algorithm’ Sept 2011

[7] Dewan Md. Farid1, Nouria Harbi1, and Mohammad Zahidur Rahman2 1ERIC Laboratory, University Lumière Lyon 2 – France 2Department of Computer Science and Engineering, Jahangirnagar University, Bangladesh ‘COMBINING NAIVE BAYES AND DECISION TREE FOR ADAPTIVE INTRUSION DETECTION’.Apr 2010.

[8] Yogendra Kumar Jain and Upendra , ‘An Efficient Intrusion Detection Based on Decision Tree Classifier Using Feature Reduction ‘Jan 2012

[9] Zeinab Kermansaravi 1, Hamid Jazayeriy2, Soheil Fateri3 ‘Intrusion Detection System in Computer Networks Using Decision Tree and SVM Algorithms’ June 2013.

[10] Snehal A. Mulay Bharati Vidyapeeth University, Pune, Maharashtra India P.R. Devale Bharati Vidyapeeth University Maharashtra India G.V. Garje Pune University Maharashtra India ‘Intrusion Detection System using Support Vector Machine and Decision Tree’.June 2010.