ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

Using Data Mining Techniques for Intrusion Detection

Sathish.S.N

Project Manager, Infosys Limited, Mysore, India

ABSTRACT: The curve that plots the degree of advancement of the computer networks with the advent of internet and its amazing applications, security becomes the next factor that needs to be paid attention. With many intruders, hackers and computer users with malicious intent, any network user needs to secure his/her system in order to protect their own data.

Many techniques are proposed and available today to detect those intrusions and the misuse of the network. But these schemes have failed to make detection at the lower level (i.e.) the packet level. This paper aims at one such proposed scheme that works at lower level matching the incoming packet patterns with a set of intrusion patterns that are maintained in a database. This paper discusses Intrusion Detection using Data Mining techniques (IDDM) and to determine the feasibility and effectiveness of data mining techniques in real-time intrusion detection.

KEYWORDS*:* Data Mining; IDC; Intrusion Detection Systems; Detection Strategy, Host Intrusion Detection; Super plasticizers; w/c ratio

I. INTRODUCTION TO INTRUSION DETECTION SYSTEMS (IDS)

The well-known use of the Intrusion Detection Systems (IDS) has revealed the significance of their key role in network security as well as its evident performance shortfalls. Data Mining (DM) is a helpful practice to uncover new insights, associations and hidden patterns within large data set of logs and messages. Intrusion Detection Systems (IDS) have become a crucial element to secure the current and emerging networks, as well as the services and applications detecting, and responding to malicious activity. Organization can use IDS employing as **Data Source** the information from an individual host (Host Intrusion Detection HID) and/or network of computers (Network Intrusion Detection). On the other hand, it could be selected according to their **Detection Strategy** finding intrusions that matches with pre-defined patterns or signatures (Misuse Detection) and/or finding intrusions by the expected network behavior and its deviations (Anomaly Detection)

Misuse detection and anomaly detection: In misuse detection, each instance in a data set is labeled as 'normal' or 'intrusion' and a learning algorithm is trained over the labeled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labeled appropriately. Unlike signature-based intrusion detection systems, models of misuse are created automatically, and can be more sophisticated and precise than manually created signatures. A key advantage of misuse detection techniques is their high degree of accuracy in detecting known attacks and their variations. Their obvious drawback is the inability to detect attacks whose instances have not yet been observed. Anomaly detection, on the other hand, builds models of normal behavior, and automatically detects any deviation from it, flagging the latter as suspect. Anomaly detection techniques thus identify new types of intrusions as deviations from normal usage. While an extremely powerful and novel tool, a potential draw-back of these techniques is the rate of false alarms. This can happen primarily because previously unseen (yet legitimate) system behaviors may also be recognized as anomalies, and hence flagged as potential intrusions.

Currently intrusion detection can be broadly classified into three types

- Host based intrusion detection system

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

- Network based intrusion detection system and
- Network-Node based intrusion detection system.

- **Host-Based Intrusion Detection System (HIDS)**

In its narrowest sense, a HIDS is an IDS that monitors platform and application event logs from multiple sources for suspicious activity. Host computers may include user workstations (including specialized applications such as Web browsers), peripherals (such as printers), specialized servers such as Web servers, or network components (such as firewalls, routers, and switches). HIDS use software modules installed on each monitored host. HIDS can detect computer misuse from trusted insiders as well as from those who have infiltrated a corporate network. They look for unusual activity confined to the local host such as logins, improper file access, unapproved privilege escalation, or alterations on system privileges.

- **Network-Based Intrusion Detection System (NIDS)**

NIDS monitors all network traffic passing on the segment where the agent is installed, reacting to any anomaly or signature-based suspicious activity. NIDS come in the guise of turnkey appliances that just plug in to the network or software that installed on commercial off-the-shelf computers. A NIDS usually has two logical components: · A sensor and · A management station or console. The sensor sits on a network segment, analyzing every network packet for attack signatures. The console receives alarms from the sensor(s) and displays them to an administrator. The sensors are usually dedicated systems that exist only to monitor the network. They have a network interface in promiscuous mode, which means they receive all network traffic, not just, that destined for their IP address, and they capture passing network traffic for analysis.

- **Network-Node Based Intrusion Detection System (NNIDS)**

Network-Node Intrusion detection works by analyzing network traffic like Standard network-based intrusion detection does. Network-node pulls the packet-intercepting technology off of the wire and puts it on the host. NNID is positioned in such a way that it captures packets after they reach their final target, the destination host. The packet is then analyzed just as if it were traveling along the network through a conventional NIDS (A newer type termed as network-node intrusion detection is a hybrid of the two but is considered a subtype of network intrusion detection because it relies primarily upon network traffic analysis for detection

IDS often perform the following tasks:

- Monitoring and analyzing user and system activity
- Audit of system structure and fault
- Recognition activation model mapping known attacks and alert
- Statistical analysis of abnormal behavior model
- Evaluating the integrity of systems and data files

Evolution of IDS

The first IDS were host-based, and looked at operating system logs performing simple pattern matches against a small set of signatures. This approach quickly expanded to systems that looked at network traffic, initially also for simple patterns. As IDS gained a level of protocol-awareness, they were able to look for certain single packet traffic types known to be malicious, examining the source and destination IP addresses, along with source and destination ports. Further sophistication brought an awareness of network sessions and the ability to examine dialogs between systems for multi-packet activity. More recent IDS can examine and respond to entire conversations between hosts, using knowledge of protocols and network sessions to analyze traffic for malicious activity. Analysis of the network traffic is based on how that traffic would appear at the destination, a task often requiring specialized network drivers to operate at full wire-speed. The emerging classes of IDS take this one step further by combining log analysis, along with

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

information from other IDS and anti-virus software to correlate events in an effort to identify and respond to intrusions in real time.

II. DATA MINING AND INTRUSION DETECTION

ID using Data Mining (IDDM), use as basis the audited data from different sources (particularly records representing a network event, described with attributes as number of bytes transferred, access counts, etc), activity indexes (from normal and intrusion activity) and algorithms to search significant patterns; enabling the construction of misuse and anomaly detection models based on an intelligible set of rules. The raw data is archived and sampled in discrete records according to the attributes. Data mining programs are subsequently used over the traffic records to compute patterns. The connections and the patterns are then analyzed to construct additional features, getting an empirical and iterative approach. Data Mining can be applied to misuse detection and anomaly detection mode.

In **Data Mining-based Misuse Detection** each data record is classified and labeled as normal or anomalous activity. This process is the basis for a learning algorithm able to detect known attacks and new ones if they are cataloged appropriately under a statistical process. The basic known as discovery outliers, matches an abnormal behavior against an attack patterns knowledge base that capture behavioral patterns of intrusion and typical activity. To do this, it is needed to compute each measure with random variables implying more updating effort as more audit records are analyzed but more accuracy with more mined data. Although the activity needs to be analyzed individually, complementary visualization and data mining techniques can be used to improve performance and reduce the computational requirements. Some researches focused on this topic are JAM (Java Agents for Metal earning), MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection) and Automated Discovery and Concise On the other hand.

Data Mining-based anomaly detection goals are related with searching inherent but previously unidentified information from the collected data. A set of records is stored building a normal profile to be compared with the most recent activity (delimited in a time window) determining if it is far from the expected behavior and establishing the similarity degree with other historical profiles. The suspicious connection is classified as known, unknown or false alarm. The most popular anomaly detection system using data mining is ADAM (Audit Data Analysis and Meaning).

III. INTRUSION DETECTION USING DATA MINING TECHNIQUES

Technically the Knowledge Discovery in Databases (KDD) practice is associated with the extraction and discovery of useful information from large relational databases while Data Mining (DM) represents its core as decision support stage. Data Mining is a finding process of significant non-intuitive correlations and patterns from a variety of sources, making possible to get high-level knowledge information from low-level data.

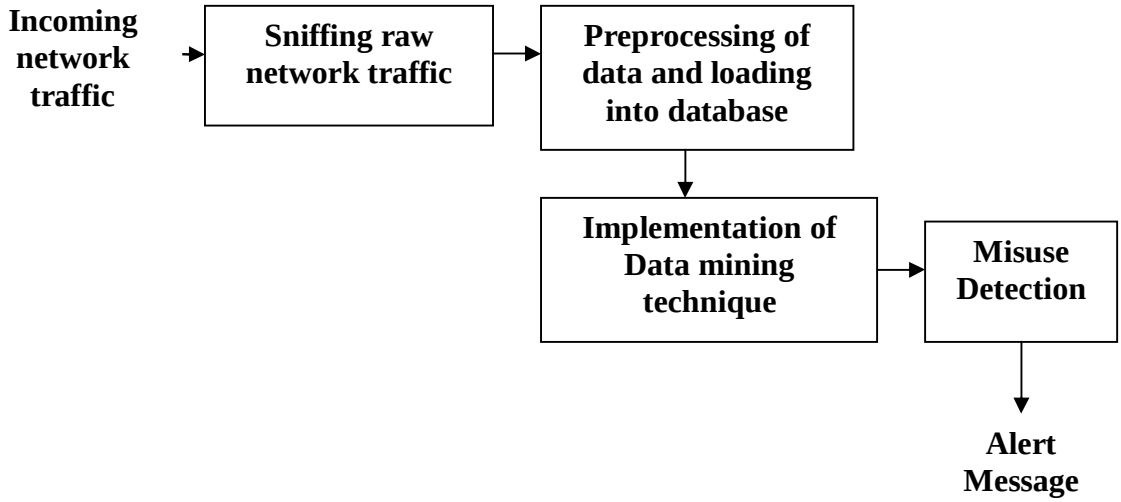
ID using Data Mining (IDDM), use as basis the audited data from different sources (particularly records representing a network event, described with attributes as number of bytes transferred, access counts, etc), activity indexes (from normal and intrusion activity) and algorithms to search significant patterns; enabling the construction of misuse and anomaly detection models based on an intelligible set of rules. The raw data is archived and sampled in discrete records according to the attributes. Data mining programs are subsequently used over the traffic records to compute patterns. The connections and the patterns are then analyzed to construct additional features, getting an empirical and iterative approach.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015



General Block Diagram of the IDS using Data Mining

One of the most critical and success determining selections is related with the data mining technique:

· **Classification** categorizes the data records (training data set) in a predetermined set of classes (Data Classes) used as attribute to label each record; distinguishing elements belonging to the normal or abnormal class (a specific kind of intrusion), using decision trees or rules. This technique has been popular to detect individual attacks but has to be applied with complementary fine-tuning techniques to reduce its demonstrated high false positives rate.

With support tools as RIPER (a classification rule learning program) and using a preliminary set of intrusion features, accurate rules and temporal statistical indexes can be generated to recognize anomalous activity. They have to be inspected, edited and included in the desired model (frequently misuse models).

· **Association Rules**: Associations of system features finding unseen and / or unexpected attribute correlations within data records of a data set, as a basis for behavior profiles.

· **Frequent Episode Rules** analyze relationships in the data stream to find recurrent and sequential patterns of simultaneous events, to compute them later. Its results have been useful for attacks with arbitrary patterns of noise or distributed attacks.

System Architecture

The basic intrusion detection system will have two parts

- Offline training process
- Online detection process

The operation of each part described as follows**:**

- **Data collection module**

  The application under investigation is executed under various normal usage scenarios. During each execution, the application spawns one or many processes are captured and recorded in a sequence. The output this module is a

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

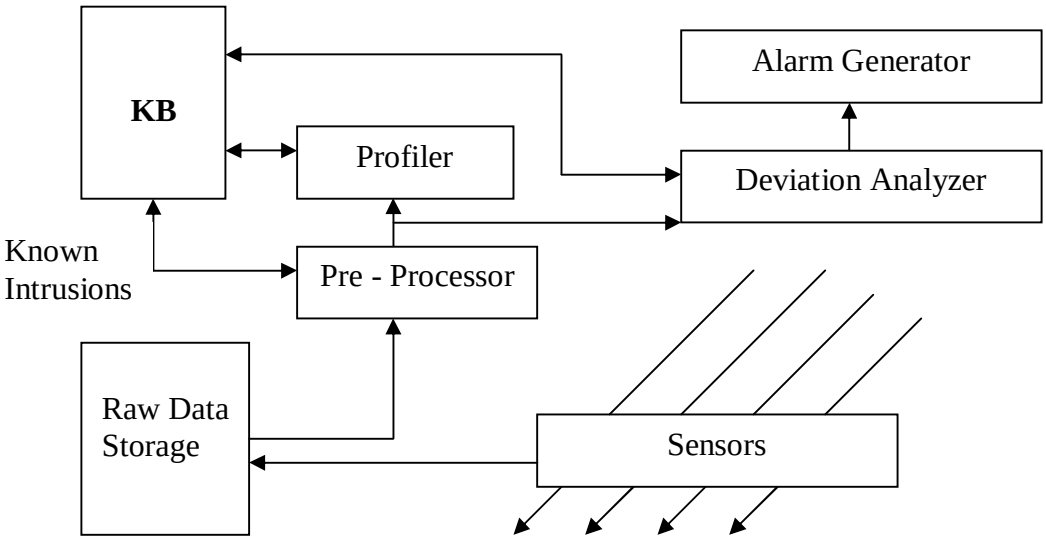
number of system calls if it is a host based detection system otherwise packet information if it is a network detection system.

- **Data preprocessing module**
  In the data preprocessing module, the unwanted information will be removed and information like packet length, packet type, source and destination address, header information, type of protocol, time stamp are collected.

- **Pattern extraction module**
  This module extracts maximal patterns from selected set of sequences generated by data preprocessing module.

- **Pattern overlap relationship identification module**
  Patterns are organized in adjacency list in which overlap relationship between patterns is located and maintained.

- **Pattern matching module**
In pattern matching module, the already stored patterns are matched with incoming packets. If there is any matching found then displaying that packet is intruded.



Intrusion Detection System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2015

The various components in the above figure are discussed below:

Component	Description
Sensors	• Collect raw network data
Raw Data Storage	• Archives the collected in a relational Database data and apply filters if needed ie. Transaction filters by source/destination IP/Port/protocol which hits in a specific time window • Support creation and tracking for security incidents
Pre-Processor	• Transforms data in useful formats for mining algorithms • Allows the use of Programmable and customized models • Filters and eliminates noise • Uses detection model to recognize known attack patters by storing them in the KB for more analysis and report
Knowledge Base	• Stores mining models, rules and associated information. The data can be manipulated for offline analysis, training and labeling. • Makes easy the data correlation from a variety of sources/collections from longer periods of time, enabling the detection of large scale attacks.
Profiler	• Get status of set of data in specific period of time for deviation analysis. • Uses historical information and the data set to generate new profiles, constructs features and redistributes the profiles.
Deviation Analyser	Finds differences to be analysed storing them in the knowledge base to obtain new profiles computed from the models and triggers the alarms
Alarm Generator	Notify the administrator abnormal network events using emails, alarms etc.

IV. CONCLUSION

The Intrusion Detection system developed using Data Mining Techniques provides a great flexibility for the network administrators and users to monitor the packets that are transferred over the network.

REFERENCES

1.Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Ning Tan Computer Science Department, 200 Union Street SE, 4-192, EE/CSC Building University of Minnesota, Minneapolis, MN 55455, USA {dokas, ertoz, kumar, aleks, srivasta, ptan@cs.umn.edu, Data Mining for Network Intrusion Detection .

2. Dary Alexandra Pena Maldonado, Data mining, A new intrusion detection technique IAC Security Essentials certification Practical Assignment Version No 1.4 Option 1 June 19th 2003

3. Vipin Kumar, Mahesh V. Joshi, Eui-Hong (Sam) Han, Pang-Ning Tan, Michael Steinbach , University of Minnesota, 4-192 EE/CSci Building, 200 Union Street SE, Minneapolis, MN 55455, USA, fkumar,mjoshi,han,ptan,steinbacg@cs.umn.edu, High Performance Data Mining