ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 5, Special Issue 12, April 2018

Overview on Data Mining in Intrusion Detection systems

C.Amali Pushpam[1] , J.Gnana Jayanthi[2]
Research Scholar, Rajah Serfoji College, Tamil Nadu, India
mailto:joemarycap@gmail.com [1] joemarycap@gmail.com
Professor, Dept.of Computer Science, Rajah Serfoji College, Tamil Nadu, India
[2]jgnanajayanthi@gmail.com

Abstract **- Communication is a part of our life. Nowadays tremendous wonders have been happening in communication. As a result enormous amounts of data are transferred through network. These data are having valuable, usable and novel information. This valuable information is being used in various applications like medicine, agriculture, education, market and business, etc. Along with valid information, sometimes fake information are also inserted and transferred and this fake information is becoming a very big threat to information assurance. Hence it is necessary to protect system / Network and information from attack. Intrusion detection system is an emerging trend in the path of security. In the process of detecting intrusion, data analysis is a basic and fundamental step. To support this activity, Intrusion detection system is integrated with data mining. This survey paper provides a detailed information about intrusion detection system based data mining and paves a new way to move on to achieve the best result.**

Keywords **- Intrusion, attack, data mining, intruders, security**

I. INTRODUCTION

Intrusion is an action which compromise system security in terms of confidentiality, availability and integrity. Intruder is a threat to system security [1][2][3]. They can be either insider or outsider. They are classified as Masquerader: The one who is an unauthorized to access computer, penetrates a system's access controls to exploit a legitimate user's account. He may be outsider

Misfeasor: The one who is an authorized to access system's resources, misuse his privileges. He may be insider. Clandestine user: The one who seizes supervisory control of the system and uses this control to avoid auditing or to suppress audit collection. He may be either an outsider or an insider.

Intrusion detection is a process which detects the intrusion in the system. An intrusion detection system is a combination of software and hardware designed to monitor the network activities and analyze them and identifies the malicious activity or policy violation and alerts the administrator.

IDS can be classified by location i.e where detection takes place and the detection method that is employed.

In terms of location:-

(i) Network intrusion detection systems (NIDS)
(ii) Host intrusion detection systems (HIDS)

In terms of detection method:-

(i) Signature based
(ii) Anomaly based

Network intrusion detection systems (NIDS)

Network intrusion detection system monitors activities of network and detect intrusion in network. As it monitors traffic to and from all devices on the network, it is kept at a strategic point or points within the network. It monitors and analysis passing events on the entire subnet and matches event with known attack already available in library. If there is match, abnormal behavior is sensed and alarm is given to the administrator.

Host intrusion detection system (HIDS)

A host intrusion detection system (HIDS) monitors the activities of system / device in network and detects intrusion. It runs on individual hosts or devices on the network and monitors the inbound and outbound packets from the device only. Monitoring important system files is an example of a HIDS. It takes a copy of existing system files and matches it to the previous snapshot. If the critical system files were modified or deleted, an alert is sent to the administrator to investigate.

Signature (knowledge) based detectionIt uses patterns of known attacks. A signature based IDS will monitor data that enter to and from the network and compare them against signatures or attributes of known malicious threats in database[1][2][3]. Here the signature means some evidence left by intruder while attacking system. By using data mining, these evidence / patterns are extracted from known attacks. Misuse detection is a supervised algorithm that tries to detect patterns of known attacks within the audit stream of a system, i.e. it identifies attacks directly. Signature-based IDS detects attacks by looking for specific patterns, such as byte sequences in network traffic, or known malicious instruction sequences used by malware. Signature-based IDS is designed to detect known attacks, but it is unable to detect new attacks, for which no pattern is available. This is similar to the way most antivirus software detects malware. The issue is that there will be an interval between arrival of a new threat being discovered and the uploading of signature in database used to detect that threat. During that interval, IDS would be unable to detect the new threat.

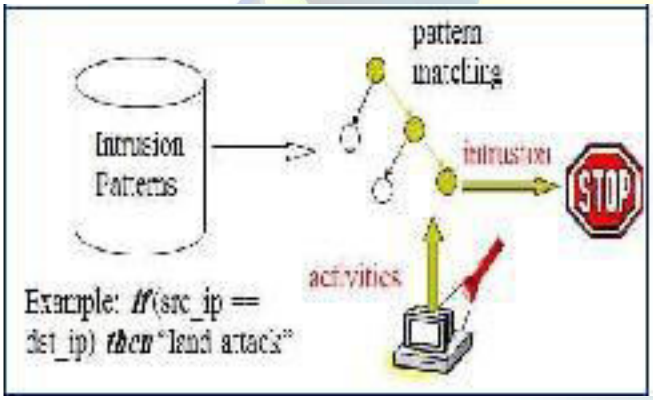


Fig : 1

The main disadvantage of this approach is that database of known attacks must be kept up-to-date and consistent. Manual coding of known intrusion patterns is another issue in signature based detection

Anomaly (Behavior) based detection

It identifies anomalies as deviation from "normal" behavior. An anomaly based IDS monitor network traffic and compares it against an established baseline. Established baseline is the normal behavior of that network, for example bandwidth size, type of protocols, type of ports / devicesused. If there is any deviation from this normal behavior, it is identified as anomalous and alerts the administrator or user[1][2][3].

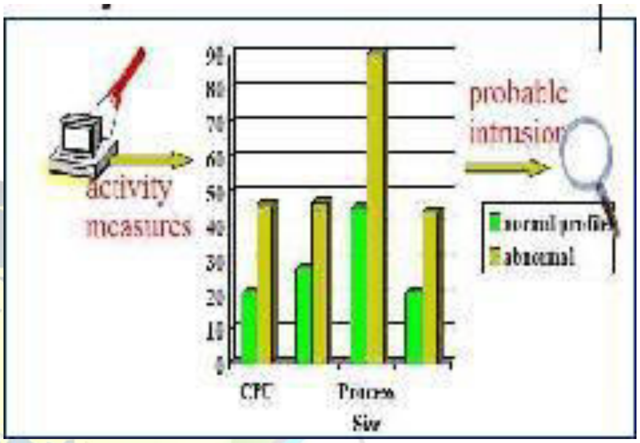


Fig : 2

Main problems:

- Selecting the right set of system features to be measured in based on experience.
- Unable to capture sequential interrelation between events.

II. INTRUSION PREVENTION SYSTEM (IPS)

Intrusion prevention systems are next level of intrusion detection systems. Like IDS, IPS monitors network traffics / system activities for threat [3]. The main difference is, it not only detect but also take necessary action to prevent system / network. It takes actions such as sending an alarm, dropping detected malicious packets, resetting a connection or blocking traffic from the offending IP address.

Passive IDS

A passive IDS simply detects and alerts. When it detects suspicious traffic, it generates an alert and sent to the administrator or user and leaves the responsibility to them to take necessary action.

Reactive IDS

A reactive IDS not only detect suspicious or malicious traffic but alert the administrator. It will take pre-defined proactive actions to respond to the threat. For example when it identifies traffic as malicious coming from particular address, it blocks any other traffic coming from that address / user.

Types of Attacks

Intruders affect the system security through various types of attacks. These attacks are classified into 4 categories named DOS, U2R, R2L, and Probe[1][2].

a) Denial of service (DOS) :

Dos attack tries to stop the authorized user for accessing or consuming the services

b) Remote to local (R2L):

Here the attacker sends a message to the host in a network over remote system and makes some vulnerability.

c) User to root (U2R):

In this intruder has the right to access local machine but tries to get the access of super user(administrator) .

d) Probe:

Probe attempts to steal the data of the target machine that would make some violation in the future.

III. STANDARD MEASURES FOR EVALUATING IDS

Performance of ids is measured [7][10][11][12] in terms of

1) True positive (TP):- At the time of IDS all numbers of normal data which have been found.

2) True negative (TN):- The total number of abnormal data which are detected in IDS.

3) False positive (FP):- It is also said as false alarm, the total set of normal data which are detected but that should be actual attack.

4) False negative (FN):- Total number of abnormal detected instance but that should be normal data.

Hence, to calculate the performance of the IDS in the form of detection rate, false alarm and accuracy.

So, Detection Rate (DR) = (TP/TP+FN)*100% false alarm rate (FAR) = (FP/number of attacks accuracy)

= (TP+TN/TP+TN+FP+FN)*100% [4]

IV. COMPARISON WITH FIREWALLS

Intrusion Detection System and firewall both relate to network security but differ in functionality. As Firewall stops intrusion from happening by looking outwardly, it limits access between networks and do not signal an attack from inside the network. An IDS signals an alarm once it suspects an intrusion. Also IDS perform another task that it also watches origin of attack from within a system.

V. ARCHITECTURE OF DATA MINING BASED IDS

It consists of sensors, detectors, a data warehouse, and a model generation component [7]. In this architecture, the following processes are being carried out like data gathering, sharing, analysis, data archiving, model generation and distribution. In the following section we describe the components depicted in Figure fig:3 in detail.

Sensors / Agents

Sensors or agents monitor the network activities and analyze it. In some places, the term agent is used particularly in host based IDS. Sensors monitor raw data on a monitored system and select features which will be used in model evaluation. Sensors protect the rest of the IDS from the specific low level properties of the target system being monitored. This is done by having all the sensors implement a Basic Auditing Module (BAM) framework.

Detectors

Detectors take formatted data from sensors and evaluate that formatted data using a detection model and determine if it is an attack. If it is an attack, that result be sent back to data warehouse for further analysis. There can be multiple layers of detectors monitoring the same system. Hence parallel mechanism is applied by distributing workload to different detectors to analyze events in parallel. There are “back-end” detectors and “front-end” detectors.

For correlation or trend analysis back-end detectors are used. Simple and quick intrusion detections are performed by front-end detectors. For more thorough and time consuming analysis front-end detectors pass data to back-end detectors.

Data Warehouse

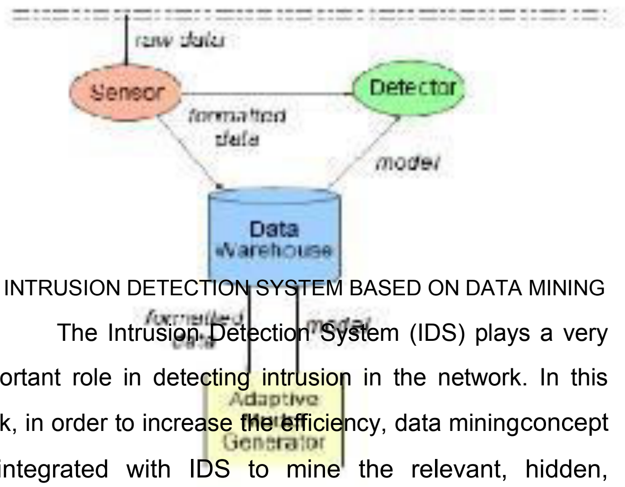
As the data warehouse is a centralized storage for data and models, different components like multiple sensors, detectors can manipulate the same data simultaneously with the existence of a database, such as off-line training and manually labeling and also data warehouse assist the integration of data from multiple sensors. By associating data/results from different IDSs the detection of complicated and large scale attacks becomes possible. As

Relational database support the feature of "stored procedure calls", complicated calculations can be carried out easily on

server. By using SQL query data are retrieved from server quickly.

Model Generator

The main purpose of the model is to generate new / updated intrusion detection models to facilitate the rapid development. In this architecture, when an attack is detected as an anomaly, its exemplary data is processed by the model generator. Model generator uses the archived (past) normal and intrusion data sets from the data warehouse and generates a model that can detect the new intrusion and distributes it to the detectors for detection. It is useful in unsupervised anomaly detection algorithms as it can operate on unlabeled data that are directly collected by the sensors from monitored system. Successful intrusion detection system implements data mining based IDS. In that a data mining engine, equipped with feature extraction programs and machine learning programs, serves as the model generator for several detectors. It receives audit data for anomalous events (encoded as a GIDO, the Generalized Intrusion Detection Objects) from a detector, computes patterns from the data, compares them with past normal patterns to identify the "unique" intrusion patterns, and constructs features accordingly. The Common Intrusion Specification Language (CISL) is used to represent intrusion detection models. Much of the design and implementation efforts are being carried in this area.



VI. INTRUSION DETECTION SYSTEM BASED ON DATA MINING

The Intrusion Detection System (IDS) plays a very important role in detecting intrusion in the network. In this work, in order to increase the efficiency, data miningconcept is integrated with IDS to mine the relevant, hidden, valuable data of interest for the user with less execution time. Intrusion detection system handle large amount of 5 V characteristics data i.e Volume (size of data), Velocity (speed of incoming data), Value (trustworthiness of data), Variety (heterogeneous nature) and Veracity (value of data changing dynamically). Hence IDS is integrated with data mining.

Why IDS is integrated with data mining

IDS are data analysis process[9]. Data mining analysis enormous amount of data with high speed in less amount of time. A number of tools and technologies are available and working on current and new platforms. Both normal and abnormal activities leave evidence on data. So it is quite easy to develop models through supervised and unsupervised learning

DATA MINING TECHNIQUES FOR INTRUSION DETECTION

CLASSIFICATION:

Classification categorized the datasets into predetermined sets / classes. It is effective for both misuse detection and anomaly detection, but more frequently used for misuse detection[1][3][4][5][6]. It is supervised machine learning technique and manages only labeled data. It cannot support to work on unlabeled data. Hence, here intrusion detection system performances will be degraded.

That's why it has low efficient in IDS in respect to clustering.

Different classification techniques

ANN, Decision tree, Naive bayes classifier, K-nearest neighbour classifier, Support vector machine etc. are used in IDS.

CLUSTERING:

Detect the intrusion over the network data. It is unsupervised machine learning technique. It detects the patterns from unlabeled data on the basis of various dimensions. Clustering technique is implemented in both anomaly as well as misuse detection[1][3][4][5][6].

Different classification techniques

K-Means, K-Medoids

ASSOCIATION:

You make a simple correlation between two or more items, often of the same type to identify patterns.

(i). Artificial Neural Network

A mathematical model inspired by biological neural networks. It consists of interconnected group of artificial neurons, and it processes information using this interconnection. It is an adaptive system that changes its structure during a learning phase. It models complex relationships between inputs and outputs in order to find patterns in data [2][3][4][7] [11].

Advantages

✓ Low error rate

✓ High accuracy

✓ Ease of maintenance

✓ Flexible with respect to, incomplete, missing and noisy data.

Limitations

✓ Long training times

✓ Neural networks are totally dependent on the quality and amount of data available.

ANN is much more stable and reliable than other models and algorithms in IDS

(ii). Support vector machines

A Support Vector Machine (SVM) is a binary classifier and is supervised machine learning algorithm. It is employed for both classification and regression purposes. SVMs are more commonly used in classification problems. SVMs use the concept of hyper plane not decision tree at all. It is working based on the idea of finding a hyper plane that best divides a dataset into two classes. In two-dimension, hyper plane is straight line. In multi dimension, hyper plane is called kernel and it is difficult to visualize and select that[1][6][7][11][12].

Advantages:

- Very accurate classifiers.
- Less over fitting, robust to noise.

Limitations:

- SVM is a binary classifier.
- In multi-class classification, kernel selection and interpretability are some weaknesses of SVM
- Computationally expensive, thus runs slow.

Execution time is less and produces high accuracy with smaller dataset in IDS

(iii). K-Nearest Neighbor

KNN is a classifier and a supervised learning algorithm. There are two types of learners i.e lazy and eager learner. C4.5 and SVM are eager learner. KNN is a lazy learner. Because during training, it just stores the training data rather than builds model. First KNN looks at k-nearest neighbors. Second, KNN classifies the data based on idea of the neighbors' classes[1][6].

Advantages:

- Ease of understanding and implementation.
- Depending on the distance metric, it can be quite accurate.

Limitations:

- Computationally expensive on a large dataset.
- kNN generally requires greater storage than eager classifiers.
- Selecting a good distance metric is crucial to accuracy.
- When value of K is large, takes large time for prediction and influence the accuracy by reduces the effect of noise.

(iv). Decision tree

Decision tree is a Classifier that constructs tree structure. In this, root and internal nodes are labeled and represents question and arc represents the answer to the associated question. Each leaf node indicates value of target variable. The result will be in the form of rules which are if-then-else expressions [1][4][6][7].

Advantages:

- Easy to understand• Easy to generate rules

Limitations:

- May suffer from over fitting
- Does not handle easily non numeric data
- If quite large – pruning is necessary

It has high detection rate in case of large dataset in IDS

(v). Naive Bayes Classifier

Naive Bayes classifier is probabilistic classifier and based on membership probability it predicts the class[3][8][13].

Advantages:

- Construction of Naive Bayes classifier is easy.
- Naive Bayes gives better performance in case of U2R and R2L attack.

Limitations:

The execution time of Naive Bayes is more as compared to other classifier.

(vi). k-means

It is a popular cluster analysis. When a set of objects is given as input, it creates *k* clusters on its own without any information about which cluster an observation belongs to. Hence it is called unsupervised learning [6][11][12].

Advantages:

- Faster and more efficient especially over large datasets.
- It can also be used to explore whether there are overlooked patterns or relationships in the dataset.

Limitations:

- It is sensitive to outliers and the initial choice of centroids.
- It is designed to operate on continuous data – extra tricks are needed to work on discrete data.

Execution time of KMeans clustering algorithm is less in case of small dataset, but when number of data point increases, K-Medoids performs better

VII. INTRUSION DETECTION SYSTEMS BASED DATA MINING - ISSUES AND CHALLENGES

In the above section, many data mining techniques for Intrusion detection systems were discussed. All these techniques are having their own limitations. In general, there are some common issues and challenges in this field [5][8][9][11]. It is given below.

- Data overload
- False positives / False Alarm rate
- False negatives
- Almost all techniques are designed to protect one or two web attacks only

VIII. RESEARCH RESULT

The work is done based on the literature research. The main purpose of this paper is to provide an overall view on Intrusion Detection System based on Data Mining. As it is given, the introduction part covers the fundamentals of IDS. Then the core concept of integrating data mining with IDS is explained. Data mining techniques are categorized into three concept i.e Classification, Clustering and Association Rule. Under this various data mining algorithms used in IDS were analyzed. Each algorithm has its own pros and cons. Performance of algorithms in IDS is evaluated in terms of detection rate, false alarm and accuracy. The degrees of these factors are varying based on applications. Sometimes hybrid methods yield good results. Hence, as per our research methodology, designing framework for intrusion detection systems based on data mining is current, critical and common field and also very complex task. Further researches are to be carried out in this area.

IX. CONCLUSION

Everyday huge amount of information are transferred from one network to another. The information may be exposed to attacks. The information and information system should be protected from unauthorized users. To provide and maintain the Confidentiality and Integrity of the information is a very tedious job. So Intrusion Detection plays a vital role. The Intrusion Detection System (IDS) plays a vital role in detecting anomalies and attacks in the network. In this work, data mining concept is integrated with IDS to identify the relevant, hidden data of interest for the user effectively and with less execution time. On the basis of detection rate, accuracy, execution time and false alarm rate the analysis has been done on different classification and clustering data mining techniques for intrusion detection.

After surveying many research papers, it is observed that most researches in intrusion detection use ANN. Because ANN is much more stable and reliable than other models and algorithms. Besides, the second most used model is SVM. Though a number of data mining techniques are used in Intrusion detection systems, some loopholes still exist. To rectify this, researchers try hybrid Methods that perform well than single methods. Combination of classifiers produce high detection rate and low false rate. Still, more research are to be carried out in this area to get better result and to meet the fast growing challenges in this field.

ACKNOWLEDGMENT

I sincerely thank my Guide Professor Dr. J.Gnana Jayanthi for her guidance and support given.

REFERENCES

[1]. Kavitha.N, Blessy Boaz, "A Survey on Intrusion Detection System Using Data Mining Techniques", in International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Special Issue 11, PP:460-465, Sep-2017

[2].J.Josemila Baby and J.R. Jeba, "Survey paper on various hybrid and anomaly based network intrusion detection system", in research journal of applied sciences 12 (3-4) : 304-310,2017

[3]. Liu Hua Yeo, Xiangdong Che, Shalini Lakkaraju, " Understanding Modern Intrusion Detection Systems: A Survey",

[4]. Hadi Barani Baravati et.al, " A new Data Mining-based Approach to Improving the Quality of Alerts in Intrusion Detection Systems", in JCSNS International Journal of Computer Science and Network Security, VOL.17 No.8, PP:194-198, Aug-2017

[5]. Shivangee Agrawal, Gaurav Jain, " A Review On Intrusion Detection System Based Data Mining Techniques", in International Research Journal Of Engineering And Technology (Irjet), Volume: 04 Issue: 09 | PP:402-407, Sep-2017

[6]. Rashmi Ravindra Chaudhari, Sonal Pramod Patil," Intrusion Detection System: Classification, Techniques And Datasets To Implement", In International Research Journal Of Engineering And Technology (Irjet), Volume: 04 Issue: 02, PP:1860-1866, Feb-2017

[7]. Tanmayee S. Sawant And Suhasini A. Itkar, " A Survey And Comparative Study Of Different Data Mining Techniques For Implementation Of Intrusion Detection System", In International Journal Of Current Engineering And Technology, Vol.4, No.3, PP:1288-1291,

June-2014

[8]. S. Latha, "A Survey on Network Attacks and Intrusion Detection Systems" in International Conference on Advanced Computing and Communication Systems (ICACCS -2017), 06 – 07, Jan-2017

[9]. Sattarova Feruza Yusufovna, " Integrating Intrusion Detection System and Data Mining" in International Symposium on Ubiquitous Multimedia Computing, 2008.[10]. Kapil Wankhade, "An Overview of Intrusion Detection Based on Data Mining Techniques", in International Conference on

Communication Systems and Network Technologies 2013.

[11]. Nadya EL MOUSSAID, "Overview of Intrusion Detection Using Data-Mining and the features selection",

[12]. Aasia Abdullah et.al, " Data Mining Approaches on Network Data: Intrusion Detection System", in International Journal of Advanced

Research in Computer Science, Volume 8, No. 1,PP:316-319, Jan-Feb 2017

[13]. Manoj and ,Jatinder Singh, "Applications of Data Mining for Intrusion Detection", in International Journal of Educational Planning & Administration. Volume 1, Number 1 (2011), pp. 37-42.