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

Intrusion Detection Technique using Data Mining Approach: Survey

Meghana Solanki, Vidya Dhamdhere

PG Student, Department of Computer Engineering, G. H. Raisoni college of Engineering & Management, Wagholi, Pune, India.

Assistant Professor, G. H. Raisoni College of Engineering & Management, Wagholi, Pune, India.

ABSTRACT*:* Intrusion detection is an essential and important technique in research field. We propose many intrusion detection methods and systems in the literature. In this paper, we give a structured overview of various aspects of intrusion detection. Due to which a researcher can become quickly familiar with every aspect of intrusion detection. We give attacks normally identified by intrusion detection systems. We differentiate existing intrusion detection methods and systems based on the underlying computational methods used. We briefly describe and compare a large number of intrusion detection methods, techniques and systems. In addition, we also discuss tools which are used by network defenders and datasets.

KEYWORDS *:* Intrusion detection, attacks, dataset, classifier, tools.

I. INTRODUCTION

Due to development in Internet technologies and the increase in the number of network attacks, intrusion detection has become a important research issue. Intrusion detection is dynamic research area. Due to remarkable progress and a large amount of work, there are still many opportunities to advance the state-of-the-art in detecting and thwarting network-based attacks [1]. According to Anderson [2], an intrusion attempt or a threat is a unauthorized access to information, manipulate information, or render a system unreliable or unusable. For example, *Denial of Service (DoS)* attack attempts to deny a host of its resources, which are essential to work correctly during processing; *Worms and viruses* exploit other hosts through the internet and *Compromises* obtain privileged access to a host by taking advantages of known vulnerabilities. *anomaly-based intrusion detection* refers to the problem of finding exceptional patterns in network database that do not conform to the expected normal behavior. Intrusion detection has extensive applications in fraud detection for credit cards, intrusion detection for enemy activities, for cyber security, and military surveillance.
Our literature differs from the existing things in the following ways.

- We discuss origin, causes and aspects of intrusion, and also include brief information of sources of packet.
- We try to provide a classification of various intrusion detection methods, systems and tools.
- Our survey not only includes IP traffic classification and analysis but also a large number of up-to-date methods, systems and tools and analysis.

II. RELATED WORK

Intrusion is a set of actions attempt to compromise the security of system. It is compromised in terms of confidentiality, integrity and availability [8]. To gain unauthorized entry and control of the security mechanism, this can be made by an inside or outside agent. Intrusion detection functions include monitoring and analyzing user, system, and network activities. it configure systems for generation of reports of possible vulnerabilities, assessing system and file integrity. It recognizes patterns of typical attacks. It analyzes abnormal activity and it tracks user policy violations. An intrusion detection system (IDS) is a device or software application. it monitors network or system activities for malicious activities or policy violations. It produces reports to a management station. IDS come in a variety of form and approach. the goal is detecting suspicious traffic in different ways. There are two types of IDS system, network based

(NIDS) and host based (HIDS) intrusion detection systems. Some systems may attempt to stop an intrusion attempt. Intrusion detection and prevention systems (IDPS) give focus on identifying possible incidents, logging information about them, and reporting attempts. In addition, organizations use IDPSes for other purposes, such as identifying problems with security policies. It documents existing threats and determines individuals from violating security policies. IDPSes have become an important in addition to the security infrastructure of nearly every organization.

A. Different Classes of Attacks

There are different types of classes of attacks, they are following

1) Virus-

It is a self replicating program. It infects the system without any knowledge or permission from the user. If the system is accessed by another computer it increases the infection rate of a network file system.

2) Trojan-

It is a malicious program. It cannot replicate itself but can cause serious security problems in the computer system. It appears as a useful program but actually it has a secret code that can create a backdoor to the system, allowing it to do anything on the system easily, and can be called as the hacker gets control on the system without user permission.

3) Worm-

It is a self replicating program. It propagates through network services on computer systems without user intervention. It is highly harmful to network by consuming network bandwidth.

4) Denial of service (DoS)-

It is an attempt to block access to system or network resources. The loss of service is the inability of a particular network service, such as e-mail to function. It is implemented by forcing the targeted computer(s) to reset. It is also implemented by consuming resources. Intended users can no longer communicate adequately due to non-availability of service or because of obstructed Communication media.

5) User to Root-

It is able to exploit vulnerabilities to gain privileges of super user of the system while starting as a normal user on the system. There are various types of vulnerabilities such as sniffing passwords, dictionary attack, or social engineering.

6) Remote to Local-

It is an ability to send packets to a remote system over a network without having any account on that system. Performs attack against public services such as HTTP and FTP or during the connection of protected services (such as POP and IMAP).

7) Probe-

It scans the networks to identify valid IP addresses .it also scan network to collect information about host. It provides information to an attacker with the list of potential vulnerabilities that can later be used to launch an attack against selected systems and services.

B. Classification of Intrusion Detection and Intrusion Detection Systems

1) Host-based IDS (HIDS):

A HIDS monitors and analyzes the internals of a computing system [9]. it might detect internal activity such as which program accesses what resources and attempts illegitimate access.example of HIDS is a word processor that suddenly and inexplicably starts modifying the system password database.

2) Network-based IDS (NIDS):

It deals with detecting intrusions in network data. Intrusions typically occur as anomalous patterns though certain techniques model the data in a sequential fashion.it detects anomalous subsequences [9]. The NIDS monitor all incoming packets or flows, trying to find suspicious patterns.

C. Literature Survey

Network anomaly detection is a broad research area, which already boasts a number of surveys, review articles, as well as books. [3] This paper present an efficient technique for intrusion detection by making use of k-means clustering, fuzzy neural network and radial support vector machine. System uses different techniques for intrusion detection. [4] In this paper an intrusion detection system is developed using Bayesian probability. The system developed is a naive Bayesian classifier that is used to identify possible intrusions. [5] In this paper we propose a new, quantitative-based approach for the detection and the prevention of intrusions. Our model is able to probabilistically predict attacks before their completion by using a quantitative Markov model built from a corpus of network traffic collected on a *honey pot.* [6] Paper focus on improving intrusion system in wireless local area network by using Support Vector Machines (SVM). SVM performs intrusion detection based on recognized attack patterns. [7] In this paper we propose method Feature Vitality Based Reduction Method, to identify important reduced input features. We apply one of the efficient classifier naive bayes on reduced datasets for intrusion detection.

III. OVERVIEW OF INTRUSION DETECTION

A. ASPECTS OF INTRUSION DETECTION

1 input data type:

A main aspect of any intrusion detection technique is the input data nature used for analysis. Input is generally a collection of data instances also known as objects, records, points, vectors, patterns, events, cases, samples, observations, entities [10].

2) Appropriateness of proximity measures:

Proximity (similarity or dissimilarity) solve many pattern recognition problems in classification and clustering. Proximity measures are functions that take arguments as object pairs. Proximity measures return numerical values that become higher as the objects become more alike.

3) Data label:

The label related with a data instance denotes if that instance is normal or anomalous.

4) Classification of methods based on use of labeled data:

Based on availability of labels, anomaly detection techniques can operate in three modes: *supervised*, *semi-supervised* and *unsupervised.* In supervised mode, one assumes the availability of a training dataset which has labeled instances for the normal as well as the anomaly class. Semi-supervised techniques assume that the training data has labeled instances for only the normal class. Finally, unsupervised techniques do not require training data, and thus are potentially most widely applicable.

5) Relevant feature identification:

Feature selection plays an major role in detecting network anomalies. Feature selection methods are applicable in the intrusion detection domain for eliminating unimportant or irrelevant features.

IV.METHODS AND SYSTEMS FOR INTRUSION DETECTION

A. Statistical methods and systems

Normally, statistical methods use a statistical model for normal behavior to the given data and then apply a statistical inference test to determine if an unseen instance belongs to this model .low probability instances generated from the learnt model based on the applied test statistic are declared anomalies. Both parametric and nonparametric techniques have been applied in designing statistical models for anomaly detection. An example of statistical IDS is HIDE [11]. HIDE is an anomaly-based network intrusion detection system. It uses statistical models and neural network classifiers to detect intrusions.

B. Classification-based methods and systems

Classification techniques are based on establishing an explicit or implicit model. It enables categorization of network traffic patterns into several classes. An example of classification-based IDS is Automated Data Analysis and Mining (ADAM) [12]. It provides a testbed for detecting anomalous instances.

C. Clustering and Outlier-based methods and systems

Clustering is the task of assigning a set of objects into groups called *clusters*. The objects in the same cluster are more similar in some sense to each other than to those in other clusters. Clustering is used in data mining. Outliers are that point in a dataset that are highly unlikely to occur given a model of the data, For example, MINDS (Minnesota Intrusion Detection System) [13] is a data mining-based system for detecting network intrusions.

D. Soft computing methods and systems

Soft computing techniques are needed for network anomaly detection. Soft computing is generally thought of as encompassing methods such as Genetic Algorithms, Artificial Neural Networks, Fuzzy Sets, Rough Sets, Ant Colony Algorithms and Artificial Immune Systems.

1. Genetic algorithm approaches:

Genetic algorithms are population-based adaptive heuristic search techniques. It is based on evolutionary ideas.

2. Artificial Neural Network approaches:

Artificial Neural Networks (ANN) is motivated by the recognition that the human brain computes in an entirely different way from the conventional digital computer. An example of ANN-based IDS is RT-UNNID. This system is capable of intelligent real time intrusion detection using unsupervised neural networks (UNN).

3. Fuzzy set theoretic approaches:

Fuzzy network intrusion detection systems exploit fuzzy rules. it determine the likelihood of specific or general network attacks. A fuzzy input set can be defined for traffic in a specific network. NFIDS is a neuron-fuzzy anomaly-based network intrusion detection system.

4. Rough Set approaches:

A rough set is an approximation of a crisp set i.e., a regular set. It is in terms of a pair of sets that are in its lower and upper approximations. Rough sets have useful features such as enabling learning with small size training datasets and overall simplicity.

5. Ant Colony and Artificial Immune System approaches:

Ant colony optimization and related algorithms are probabilistic techniques. It is used for solving computational problems which can be reformulated to find optimal paths through graphs. Artificial Immune Systems (AIS) represent a computational method. This is inspired by the principles of the human immune system.

E. Knowledge-based methods and systems

In knowledge-based methods, network or host events are checked against predefined rules. It also checks patterns of attack. An example knowledge-based system is STAT (State Transition Analysis Tool)

1. Rule-based and Expert system approaches:

An expert system is a rule-based system, with or without an associated knowledge base. An expert system has a rule engine. Rule engine matches rules against the current state of the system.

2. Ontology and logic-based approaches:

It is possible to model attack signatures using expressive logic structure in real time by incorporating constraints and statistical properties.

F. Combination learner methods and systems

In this section, we present a few methods and systems which use combinations of multiple techniques, usually classifiers.

1. Ensemble-based methods and systems:

The idea behind the ensemble methodology is to weigh several individual classifiers. It combines them to obtain an overall classifier that outperforms every one of them. Octopus-IIDS is an example of ensemble IDS.

2. Fusion-based methods and system:

With an evolving need of automated decision making, it is important to improve classification accuracy. A suitable combination of these is the focus of the fusion approach. dLEARNIN is an ensemble of classifiers that combines information from multiple sources.

3. Hybrid methods and system:

Most current network intrusion detection systems employ misuse detection. They also employ anomaly detection. Misuse detection cannot detect unknown intrusions. Anomaly detection usually has high false positive rate. To overcome the limitations of the techniques, hybrid methods are developed by exploiting features from several network anomaly detection approaches .Hybridization of several methods improve performance of IDSs. For example, RT-MOVICAB-IDS, a hybrid intelligent IDS is introduced in.

V. EVALUATION CRITERIA

To evaluate performance, it is necessary that the system identifies the attack. System identifies normal data correctly. There are several datasets and evaluation measures. These are available for evaluating network anomaly detection methods and systems. The most commonly used datasets and evaluation measures are given below.

G. Datasets

Capturing and preprocessing high speed network traffic is very important prior to detection of network anomalies. Different tools are used for capture and analysis of network traffic data.

1. Synthetic datasets:

Synthetic datasets are generated to meet specific needs. It is used to meet conditions or tests that real data satisfy. This can be useful when designing any type of system which is used for theoretical analysis. The design can be refined. Synthetic data is used in testing. It is used in creating many different types of test scenarios.

2. Benchmark datasets:

We present some publicly available benchmark datasets. They are generated using simulated environments. It includes a number of networks. They execute different attack scenarios.

a. KDDcup99 dataset:

Since 1999, the KDDcup99 dataset is used dataset for the evaluation of network-based anomaly detection methods and systems.

b. NSL-KDD dataset:

Analysis of the KDD dataset showed that there were two important issues in the dataset. they highly affect the performance of evaluated systems which result in poor evaluation of anomaly detection methods. To solve these issues, a new dataset known as NSL-KDD [14], consisting of selected records of the complete KDD dataset was introduced. This dataset is publicly available for researchers. It has the advantages over the original KDD dataset such as it does not include redundant records in the training set. There are no duplicate records in the test set. The number of selected records from each difficulty level, it is inversely proportional to the percentage of records in the original KDD dataset. The number of records in the training and testing sets are reasonable.

c. DARPA 2000 dataset:

A DARPA6 evaluation project [15] targeted the detection of complex attacks. It contains multiple steps.

d. DEFCON dataset:

The DEFCON7 dataset is another commonly used dataset for evaluation of IDSs [16]. It contains network traffic captured when the hacker competition called Capture the Flag (CTF).

3. Real life datasets:

In this subsection, we show three real life datasets created by collecting network traffic on several days. It includes both normal as well as attack instances in appropriate proportions.

a. UNIBS dataset:

This dataset includes traffic captured or collected through many workstations. It stores traffic through 20 workstations running the GT client daemon.

b. ISCX-UNB dataset:

Real packet traces [17] were analyzed. it creates profiles for agents that generate real traffic for HTTP, SMTP, SSH, IMAP, POP3 and FTP protocols.

c. TUIDS dataset:

this data set has been prepared at the Network Security Lab at Tezpur University, India based on several attack scenarios.

B. Evaluation Measures

1) Accuracy:

Accuracy is a metric. It measures how correctly an IDS works. It measures the percentage of detection and failure as well as the number of false alarms that the system produces [18].If a system has 80% accuracy, it means that it correctly classifies 80 objects out of 100 to their actual classes.

2. Performance:

The evaluation of an IDS performance is a major task. It involves many issues that go beyond the IDS itself. These include the hardware platform, the operating system or even the deployment of the IDS.

3. Completeness:

The completeness criterion represents the space of the vulnerabilities. It shows an attack that can be covered by an ID.

4. Timeliness:

An IDS performs its analysis as quickly as possible. It enables the human analyst or the response engine to promptly react before much damage is done within a specific time period.

5. Data Quality:

Evaluating the quality of data is another essential task in NIDS evaluation. Quality of data is influenced by several factors, such as source of data which should be from reliable and appropriate sources, selection of sample which should be unbiased, sample size which is neither over nor under sampling, time of data which should be frequently updated real time data and complexity of data

C. TOOLS USED IN DIFFERENT STEPS IN NETWORK TRAFFIC ANOMALY DETECTION

1) Wireshark:

It is free and open-source packet analyzer. It can be used for network troubleshooting. It is used for analysis. It is applicable in software and communications protocol development. It is also used in education. Uses cross-platform GTK+ widget toolkit to implement its user interface. It uses pcap to capture packets. It has a graphical front-end.it also has some integrated sorting and filtering options. It functions in mirrored ports to capture network traffic to analyze for any tampering.

2) Gulp:

It allows much higher packet capture rate. It drops fewer packets. It is able to read disk. If the data rate increases, Gulp realigns its writes to even block boundaries. It optimizes writing efficiency. When it receives an interrupt, it stops filling its ring buffer. It does not exist until it has finished writing whatever remains in the ring buffer.

3) tcptrace:

it takes input files produced by several popular packet-capture programs. It includes tcpdump, snoop, etherpeek, HP Net Metrix, Wireshark, a WinDump. It produces several types of output containing information on each connection seen, such as elapsed time, bytes and segments sent and received, retransmissions, round trip times, window advertisements, and throughput. It can also produce a number of graphs with packet statistics for further analysis.

4) nfdump:

It collects and process net flow data on the command line. It is limited only by the disk space available for all the net flow data. It can be optimized in speed for efficient filtering. The filter rules look like the syntax of tcpdump.

5) nmap:

It is also known as Network Mapper. It is a free and open source utility. It is used for network exploration as well as security auditing. It uses raw IP packets in novel ways. It determines what hosts are available on the network. It determines what services (application name and version) those hosts offer. It determines what operating systems are running. It determines types of firewall or packet filter used, and many other characteristics. It is easy, flexible, powerful, well documented tool for discovering hosts in large network.

VI.CONCLUSION

In this paper, we have monitored the state-of-the-art in the modern network intrusion detection. Two well-known criteria can be used to classify and evaluate NIDSs: detection strategy and evaluation datasets. We have also showed many detection methods, systems and tools. In addition, we have seen several evaluation criteria for testing the performance of a detection method or system. A discussion of the different existing datasets and its taxonomy is also provided.

REFERENCES

[1] A. Sundaram, "An introduction to intrusion detection," *Crossroads*, vol. 2, no. 4, pp. 3–7, April 1996.

[2] J. P. Anderson, "Computer Security Threat Monitoring and Surveillance," James P Anderson Co, Fort Washington, Pennsylvania, Tech. Rep., April 1980.

[3] A.M. Chandrasekhar, "Intrusion Detection Technique By Using K-Means, Fuzzy Neural And Svm Classifier ", 2013 International Conference on Computer Communication and Informatics (ICCCI - 2013), Jan 04-06, 2013 Coimbatore, India.

[4] Hesham Altwaijry, "Bayesian Based Intrusion Detection System ", Journal of King Saud University – Computer and Information Sciences (2012) 24,1–6.

[5] Ammar Boulaiche, "A Quantitative Approach For Intrusions Detection And Prevention Based On Statistical N-Gram Models ", Procedia Computer Science 10 (2012) 450 – 457.

[6] Muamer N.Mohammed, "Intrusion Detection System Based On Svm For Wlan ",Procedia Technology 1 (2012) 313 – 317.

[7] Saurabh Mukherjee, "Intrusion Detection Using Naive Bayes Classifier With Feature Reduction", Procedia Technology 4 (2012) 119 – 128.

[8] R. Heady, G. Luger, A. Maccabe, and M. Servilla, "The Architecture of a Network Level Intrusion Detection System," Computer Science Department, University of New Mexico, Tech. Rep. TR-90, 1990.
[9] F. Wikimedia, "Intrusion detection system," http://en.wikipedia.org/wiki/Intrusion-detection system, Feb 2009.
[10] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005.
[11] Z. Zhang, J. Li, C. N. Manikopoulos, J. Jorgenson, and J. Ucles, "HIDE: a Hierarchical Network Intrusion Detection System Using Statistical Preprocessing and Neural Network Classification," in *Proc. IEEE Man Systems and Cybernetics Information Assurance Workshop*, 2001.
[12] B. Daniel, C. Julia, J. Sushil, and W. Ningning, "ADAM: a testbed for exploring the use of data mining in intrusion detection," *ACM SIGMOD Record*, vol. 30, no. 4, pp. 15–24, 2001.
[13] L. Ertoz, E. Eilertson, A. Lazarevic, P. Tan, V. Kumar, and J. Srivastava, *Data Mining - Next Generation Challenges and Future Directions*. MIT Press, 2004, ch. MINDS - Minnesota Intrusion Detection System.
[14] NSL-KDD, "NSL-KDD data set for network-based intrusion detection systems," http://iscx.cs.unb.ca/NSL-KDD/, March 2009.
[15] I. S. T. G. MIT Lincoln Lab, "DARPA Intrusion Detection Data Sets," http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/2000data.html, March 2000.
[16] Defcon, "The Shmoo Group," http://cctf.shmoo.com/, 2011.
[17] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Towards developing a systematic approach to generate benchmark datasets for intrusion detection," *Computers & Security*, vol. 31, no. 3, pp. 357– 374, 2012.
[18] S. Axelsson, "The base-rate fallacy and the difficulty of intrusion detection," *ACM Trans. Inf. System Security*, vol. 3, no. 3, pp. 186–205, August 2000.
[19] A.M. Chandrasekhar, "Intrusion Detection Technique By Using K-Means, Fuzzy Neural And Svm Classifier ", 2013 International Conference on Computer Communication and Informatics (ICCCI - 2013), Jan 04-06, 2013 Coimbatore, India.
[20] Hesham Altwaijry, "Bayesian Based Intrusion Detection System ", Journal of King Saud University – Computer and Information Sciences (2012) 24,1–6.
[21] Ammar Boulaiche, "A Quantitative Approach For Intrusions Detection And Prevention Based On Statistical N-Gram Models ", Procedia Computer Science 10 (2012) 450 – 457.
[22] Muamer N.Mohammed, "Intrusion Detection System Based On Svm For Wlan ",Procedia Technology 1 (2012) 313 – 317.