Review on Clustering and Classification techniques in Intrusion Detection Systems

Dr. S. Sandosh[1], Akila Bala[2], Nithin Kodipyaka[3]

Assistant Professor Sr., School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India [1]
Student, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India [2]
Student, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India [3]

Abstract: *In the modern cyber world, the proportion of security threats is cumulating every day, and many researchers and security specialists are focusing on IDS (Intrusion Detection Systems) and the patterns in recognizing the alerts / events to detect and prevent them. The researchers and security specialists believe that the IDS is best way to protect the network and information assets. Hence this paper is focused on various threats and potential of IDS with its patterns in detecting the alerts / events. Basically, IDS has three different styles in detection the threats: Signature-based Detection (SD), Anomaly-based Detection (AD), and Stateful Protocol Analysis (SPA). The main area where the researchers and security specialists focusing is on techniques and algorithms used for clustering and classification. This paper mainly supports in understanding and analysing the various patterns on clustering and classifying the previous alerts / events which mainly supports in detection of threats with accuracy. This review will help to increase the detection accuracy of the IDS by enhancing the clustering and classification techniques which supports efficient execution of IDS over the network.*

Keywords: Networking, Clustering, Classification, Intrusion Detection System.

I. INTRODUCTION

A network is comprised of information-sharing devices, such as machines, servers, mainframe computers, network equipment, peripheral, and other electronic devices, which are all linked to one another. This collection of information-sharing devices is also known as a computer network. Because of the Internet, millions of individuals in every region of the globe may now communicate with one another.

II. NETWORK PROTOCOLS

The way data is transported from one device to another within the same network is predetermined by a protocol known as the network protocol. It makes it possible for linked devices to interact with one another even though their processes and the structure of it may be different.

Because of the standards that govern networks, we can readily connect with individuals located all over the globe. Computers can interact with one another under the control of a system. Protocols are sets of rules that govern how connection between the functional elements of communicative networks should take place. They are sometimes known as "communication standards".

To send and receive information successfully, devices which are on both sides of the communication exchange must accept and follow protocol conventions. The support for protocols can be built into software, hardware, or both in networking. The following are some concrete instances of network protocols and their applications:

- A common mechanism that is used to receive incoming emails is called Post Office Protocol 3, or POP3.
- Outgoing emails are sent and distributed via SMTP.
- File transfers between computers are done using FTP.
- Telnet is a set of guidelines used to establish remote login connections between systems. The distant computer accepts the connection request that the local computer makes.
- The IETF protocol suite was suggested in [1] as a means of supporting IoT devices and applications.

DOI: 10.48175/IJARSCT-12782

2.1 Functioning of network protocols:

At each stage of the network, protocols break down the bigger operations into a series of distinct roles and functions for each participant. According to the generally accepted paradigm, which is frequently referred to as the Open Systems Interconnection (OSI) model, the operations that take place at each layer of the communication exchange are regulated by one or more network protocols. The upper levels of the OSI model are concerned with software and applications, whereas the lower levels are concerned with the movement of data.

The collection of network protocols that work together is referred to as a protocol suite. The Transmission Control Protocol/Internet Protocol (TCP/IP) suite is what enables connectivity to the internet. This suite is typically utilised in client-server models and includes numerous protocols that are spread across layers [2]. These layers include the data layer, the network layer, the transport layer, and the application layer. The following are some examples of these:

At the level of the information packet, the TCP protocol uses a set of rules to facilitate the exchange of messages with other points on the internet.

- User Datagram Protocol (UDP), which functions as an alternate inter process communication to TCP and that is used to create links across applications and the internet that can tolerate loss and have minimal latency; this protocol was developed by the Internet Engineering Task Force.
- Internet Protocol, or IP, which employs a predetermined set of protocols to communicate messages on the layer of IP addresses
- Additional network protocols, such as the Hypertext Transfer Protocol (HTTP) and the File Transfer Protocol (FTP), all of which have a specified set of rules to facilitate the sharing of information and the presentation of that information.

TCP is an alternate communication protocol to UDP that is used to establish connections among applications and the internet that are both low-latency and loss-tolerant. UDP employs a list of norms to facilitate communication with other internet sites somewhere at information packet level.

Numerous different network protocols are useable, including such HTTP and FTP, most of which have established rules and standards to exchange as well as display information. IP is one of these protocols, which uses a list of norms to transmit and receive communications at the level of IP addresses. Other network protocols, such as HTTP and FTP, are also available.

2.2 Weakness in network protocols:

During the process of developing network protocols, security is not a top focus. Due to the lack of security measures taken by them, the system may occasionally be susceptible to malicious attacks such as eavesdropping and cache poisoning.

The publication of bogus routes is the type of attack that occurs most frequently against network protocols. This attack causes traffic to be rerouted through infiltrated sites instead of just the correct ones.

Firewalls, antivirus software, and antispyware software all play an important role in protecting computers from potentially malicious activities. Network protocol analysers are supplemental tools that help strengthen this protection.

III. NETWORK TRAFFIC

The quantity of data which is being moved over a computer network at any given instant is referred to as the "network traffic" of that network. Network traffic, which is also known as data traffic at times, is broken up into data packets and supplied over a network prior to getting patched back together by the computer or device that is receiving it. Network traffic can move in either a north-to-south or an east-to-west direction. Traffic influences the quality of a network since it can cause download speeds to become sluggish or can cause Voice over Internet Protocol (VoIP) connections to become unstable. Traffic and security are intertwined since a spike in traffic which is not consistent with previous patterns may indicate an attack.

The data must first be broken into smaller batches before it can be efficiently transferred via a network or internet. This is necessary for the transfer of larger files. The data is disassembled, arranged, and bundled into data packets by the network so that it may be reliably transmitted over the network and so that another user of the network can open and

DOI: 10.48175/IJARSCT-12782

read the data. In order to ensure that there is no imbalance in the distribution of network traffic, every packet takes the optimal route available to it.

- *North-south traffic:* The client-server traffic that occurs between the data centre and the rest of the network is known to as north-south traffic. This type of traffic originates from a point that is not within the data centre.[4]
- *East- west traffic:* Mobility within a data centre is referred to as "east-west traffic," which more specifically refers to the movement of data from one server to another within the data centre.
- *Real-time Traffic:* The traffic that is deemed essential or significant to the operations of the business must be provided in a timely manner and to the best of one's capacity. Real-time network traffic includes activities such as voice over internet protocol (VoIP), video conferencing, and web browsing.
- *Non- Real time traffic:* Real-time traffic is considered by network management to be more important than non-real-time traffic, which is also known as best-effort traffic at times. Non-real-time traffic Email applications and the File Transfer Protocol (often known as FTP) are two examples of online publishing tools.

3.1 Importance of the analysis and monitoring of network traffic:

Using a method known as network traffic analysis, managers of networks can investigate network behaviour, maintain network availability, and identify activity that is not typical (NTA). In addition, NTA provides administrators with the opportunity to determine whether there are any operational or security risks that are currently present or will occur in a futuristic sense. When problems of this nature are addressed as soon as they appear, the organization's resources are better utilised, and the possibility of an assault is mitigated. Therefore, NTA and increased safety are intertwined[5].

1. *Find the bottlenecks:* A rise in the number of people using a certain location is likely to result in the formation of bottlenecks in that location.
2. *Troubleshoot problems with bandwidth:* If a connection is slow, a network might not be built to manage an increase in the number of users or the amount of activity.
3. *Improve visibility of devices on your network:* Increase the visibility of the devices in the network so that administrators can plan for internet traffic and adjust as necessary with the help of greater endpoint awareness.

IV. NETWORK INTRUSION

The abbreviation "IDS" comes from the combination of the words "intrusion" and "detection system," which together form the full phrase "intrusion detection system." An intrusion has taken place, which puts at risk the availability, integrity, and confidentiality of any information that may have been stored in a computer or network system.

A network intrusion is any operation that is performed on a computer system without the administrator's permission. The level of the defenders' understanding of how the attackers carry out their missions is directly correlated to their capacity to spot an incursion.

This kind of undesired conduct will often use network resources that were intended to be used for other purposes, and it will nearly always put the network's and/or its data's security at risk. The intruders can be stopped if a network intrusion detection system is developed and put into place in the right way. [6] [7]. As a first line of defence, you can use the following brief list of frequent threat vectors that you should be aware of.

4.1 Asymmetric Routing

This tactic is utilised by the adversary to gain access to the network device that is the focus of the assault via multiple channels. The objective is to keep the attack from being discovered by making it such that a significant number of both the malicious packets steer clear of certain network segments and the network intrusion sensors located within those segments.

4.2 Buffer overflow attacks

This method seeks to overwrite specified sections of computer memory across a network by replacing the typical data found in certain memory locations with a set of commands that will afterwards be carried out as part of the assault. The commands will be carried out as part of the attack. Most of the time, the goal is to either build a channel via which the attacker can access the network remotely or begin a denial-of-service assault (DoS).

4.3 Protocol-specific attacks

When devices participate in network activities, they do so in accordance with appropriate of rules and protocols. A number of these protocols, such as ARP, IP, TCP, UDP, and ICMP, as well as various application protocols, may inadvertently leave openings through which network attacks can be launched, either through protocol impersonation (also known as "spoofing") or through improper protocol messaging. Since Address Resolution Protocol (ARP) will not provide authentication on messages, it is possible for attackers to carry out attacks such as "man-in-the-middle" attacks, for example.

4.4 Traffic Flooding

Targeting network intrusion detection systems by generating traffic loads that are too high for the system to properly filter is a clever way of committing network intrusion. Attackers may occasionally carry out an undetected attack and even start an undetected "fail-open" state in the resultant crowded and chaotic network environment.

4.5 Trojans

These programmes do not replicate themselves in the manner of a virus or worm and instead pose as benign software. They instead resort to launching denial-of-service attacks, wiping out data that has been saved, or creating backdoors that can be used by outside adversaries to take control of the system. Peer-to-peer file sharing is one of the most common ways that Trojans might infect a network. Other methods include unauthorised file storage and internet archives.

4.6 Worms

Worms are a common sort of autonomous virus that may infect computers. Worms are any computer programmes that are designed to reproduce themselves without altering the files that are permitted to be changed. Worms can typically spread using email attachments as well as the Internet Relay Chat (IRC) protocol. [Case in point:] [Case in point:] Unidentified worms will eventually consume up so very much bandwidth or processing time on the network that legitimate activity will be forced out of the network entirely. Some worms actively look for sensitive information, such as files that contain the word "financial" or "SSN," and then send that data to attackers who are waiting outside of the network.

Once they have a complete understanding of the various attack vectors, network security teams can begin looking for opportunities to apply technologies and approaches that will minimise the potential efficacy of each attack vector.

V. INTRUSION DETECTION SYSTEM (IDS)

There has been a lot of research done on intrusion detection systems (IDSs) in the area of computer science due to the rising network traffic and security concerns. In addition to arbitrary intrusion categories, current IDSs also require a lot of processing capacity. Even though IDS-related topics are extensively covered in literature, we aim to provide a more complete image for a thorough review. The taxonomy to delineate contemporary IDSs based on the extensive survey and sophisticated arrangement is given by experts.

Additionally, the tables and figures we provided in the material help readers quickly understand the big picture of IDSs. To begin with, it is important to understand the differences between intrusion, intrusion detection, intrusion detection system (IDS), and intrusion prevention system. NIST (Bace and Mell, 2001) defines an intrusion as an effort to breach the CIA or undermine the security of a computer system or network.[1] The technique of closely monitoring network or computer system activity and searching for signs of intruders is known as intrusion detection. Wireless networks are now widely used, and they are significantly more vulnerable to attack than any wired network.

Three main subcategories of intrusion detection approaches exist: Stateful Protocol Analysis (SPA), Anomaly-based Detection (AD), and Signature-based Detection (SD) (SPA).[8][9]

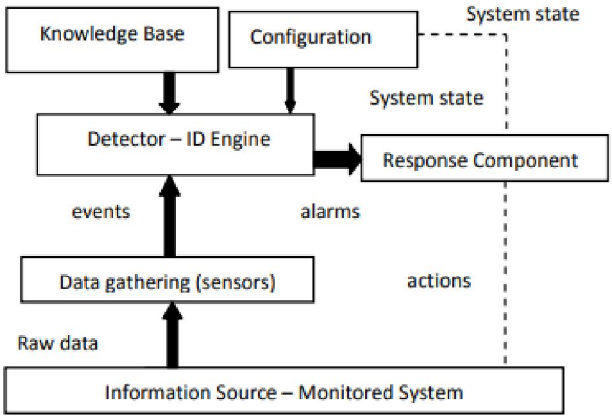

Figure 1: Basic architecture of Intrusion Detection System [3]

5.1 Advantages and Disadvantages of Intrusion Detection Methodologies

A. Signature-based Detection:

Advantages:

- Easiest and most reliable way to identify known attacks.
- Analyses the context in depth.

Disadvantages:

- Effective at identifying undiscovered attacks, evasion attacks, and variants of known attacks.
- States and protocols are not easily understood.
- It is challenging to maintain signatures and patterns.
- Maintaining the knowledge takes time

B. Anomaly-based Detection:

Advantages:

- Effective at finding new, unforeseen weaknesses.
- Less reliant on OS.
- Assist in identifying privilege abuse.

Disadvantages:

- Inaccurate profiles caused by the continual modification of observable events.
- Unavailable while behaviour profiles are being rebuilt.
- It's challenging timing the correct notifications to go off.

C. Stateful protocol Analysis:

Advantages:

- Recognize and abide by the protocol's instructions.
- Know how to spot strange command sequences.

Disadvantages:

- Examining and tracking protocol states require a lot of resources.
- Unable to examine attacks that mimic safe protocol behaviour.
- Perhaps incompatible with specialized OSs or APs

DOI: 10.48175/IJARSCT-12782

VI. CLUSTERING ALGORITHMS

In order to accurately assess the vast amount of data produced by contemporary applications, clustering algorithms have become a viable alternative meta-learning technique. Their major goal is to organise the data into clusters in such a way that things are grouped together when they have specific qualities or when they are like one another. There is a substantial body of research on clustering, and there have been efforts to classify and organize them for a wider range of uses [10]. However, there is a lack of consensus in the description of their qualities, and there is also a lack of formal categorization, which are two of the primary challenges that mislead practitioners when it comes to utilising clustering algorithms for enormous data sets. The recommended classification scheme was developed from the point of view of an algorithm designer, with the primary focus being placed on the operational characteristics of the clustering procedure. As a result, the following categories can be used to classify the various clustering algorithms' processes:

*Partitioning-based***:** Such algorithms quickly identify each cluster. A union is created by reallocating initial groups to it. These clusters ought to meet the following criteria:

a. At least one object must be present in each group
b. Each object must be a precise member of only one group.

There are a variety of techniques for partitioning data, including CLARANS, K-means, FCM, K-modes, CLARA, and PAM.

- *Hierarchical-based***:** Depending on the closest media, data are arranged hierarchically. The intermediate nodes can determine proximity. The datasets are shown as a dendrogram, where leaf nodes show individual data. Agglomerative (bottom-up) and divisive (top-down) hierarchical clustering are the two forms of hierarchical clustering. However, the hierarchical approach has a significant flaw in that once a step (such as a merge or split) has been made, it cannot be undone. Some of the well-known algorithms in this area include BIRCH, CURE, ROCK, and Chameleon.
- *Density-based:* In this section, data items are segmented based on the densities, connectedness, and boundaries of their respective areas. By analysing a point's total density, it is possible to determine the functions of datasets that have an influence on a particular data point. In order to discover clusters of any form and filter out noise (also known as outliers), several algorithms, such as OPTICS, DBSCAN, DENCLUE, DBCLASD adopt a method like this one.
- *Grid-based:* Grids are used to partition the data objects' space. The speed with which this approach processes data is the primary advantage it offers. It only must go through the information once for it to be able to determine the statistical values of the grids. To reach the appropriate level of clustering quality or to complete the work in a timely manner while dealing with severely unpredictable data distributions, the use of a single uniform grid may not be sufficient. Wave-Cluster and STING are two examples that often fall into this category.

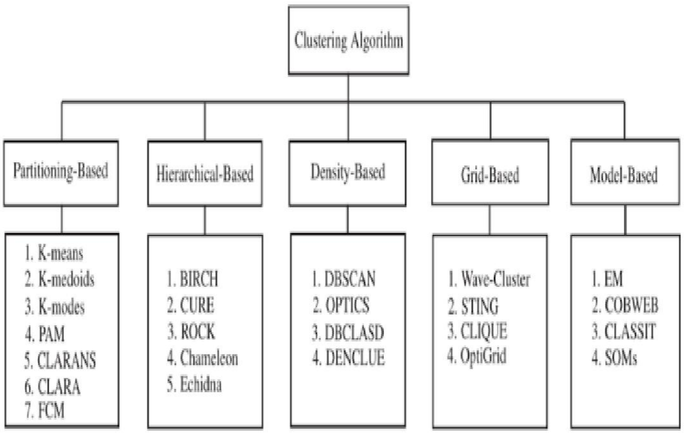

Figure 2: Types of Clustering Algorithms

- *Model-based:* Using such a strategy, the optimal fit may be achieved between the data that is presented and a mathematical model that has already been created. The model-based method is the foundation for the statistical and neural network approaches, which are the two main approaches. The most well-known model-based algorithm is arguably MCLUST, although there are also other practical algorithms, methods such as EM (which use a mixture density model), conceptual clustering (such as COBWEB), and neural network algorithms are all included such as self-organizing feature maps.

VII. CLASSIFICATION TECHNIQUES FOR IDS

- *Decision Tree Classifier:* Because they perform well and have certain benefits over other machine learning approaches, decision trees (DTs) are widely utilised in abuse detection systems [12]. This is since DTs are used frequently. Although the DT obtained a decent accuracy, it does not perform as well as other approaches on certain intrusion classes. It does not perform as well as other techniques on U2R and R2L assaults, which are both minor classes and include a considerable quantity of innovative attack types. Furthermore, they discovered that the performance of DTs and Random Forests (an ensemble of DTs) varied dramatically on different folds (subsets) of the training data due to how sensitive these models are to the data used for training.[6]
- *Support Vector Machine:* According to the description provided by Cortes and Vapnik, the Support Vector Machine (SVM) is a binary classification method that looks for the optimum hyperplane as a decision function in a high-dimensional space [11][13]. Speed and scalability are the two key benefits of utilizing SVM for intrusion detection. The DARPA 1998 dataset was used in the trials. By running training and testing on the dataset, SVM IDS was created. The testing set had an accuracy of 99.50% and a runtime of 1.63 seconds whereas the trained set had a runtime of 17.77 seconds. The effectiveness of SVM demonstrated that SVM IDS have a somewhat greater rate of correct detection.

VIII. CONCLUSION

In this paper, a complete survey on the analysis of the IDSs over the network environment is discussed. It is clearly renowned that though the network is monitored with all the available resources, the security specialists face lot more challenges and unresolved problems which triggers difficulties ahead. The wireless Intrusion Detection Systems (IDSs) raise difficulties with administration, communication, and security due to features such as mobility, lack of central points, the limited bandwidth of wireless networks, and limited resources. Most wireless Intrusion Detection Systems (IDSs) need to go through testing in a variety of topology and mobility scenarios before their protective capabilities can be validated. Heuristics — Some fuzzy, neural, and immune-based heuristic IDSs have been created and enhanced further to improve the detection accuracy of the IDS. Nevertheless, in order to limit the number of false alarms, security specialists should manage the sensitivity of warning harmful assaults.

In future, this review will be extended to implement and deploy the novelty in IDS through the extension of detection accuracy by new clustering and classification techniques. The main focus of this paper is to enhance the IDS detection of threats over the network with all possible support in the detection.

REFERENCES

[1] Erman, J., Arlitt, M., Mahanti, A., 2006. Traffic classification using clustering algorithms. In *Proceedings of the 2006 SIGCOMM workshop on Mining network data* (pp. 281-286).

[2] Z. G. Sheng, S. S. Yang, Y. F. Yu, A. V. Vasilakos, J. A. McCann,and K. K. Leung, “A survey on the ietf protocol suite for the internet of things: standards, challenges, and opportunities,” IEEE Wireless Communications Magazine, vol. 20, no. 6, pp. 91– 98, 2013.

[3] Heady, R., Luger, G., Maccabe, A., &Servilla, M. (1990). The architecture of a network level intrusion detection system (No. LA-SUB-93-219). Los Alamos National Lab.(LANL), Los Alamos, NM (United States); New Mexico Univ., Albuquerque, NM (United States). Dept. of Computer Science.

[4] Maswood, M.M.S., Develder, C., Madeira, E. Medhi, D., 2017. Energy-efficient dynamic virtual network traffic engineering for north-south traffic in multi-location data center networks. *Computer Networks*, *125*, pp.90-102.

[5] Lackman, R.A., Spragins, J.D., Tipper, D., 1992. Scheduling real-time., non-real-time traffic under nonstationary conditions. *Annals of Operations Research*, *36*(1), pp.193-224.

[6] Mukkamala S, Janoski G, Sung A. Intrusion detection using neural networks and support vector machines. Paper presented at: Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290). Honolulu, HI, USA: IEEE; vol. 2, 2002:1702-1707.

[7] Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, Vázquez E. Anomaly-based network intrusion detection: techniques systems and challenges. ComputSecur. 2009;28(1-2):18-28.

[8] Liao, H.J., Lin, C.H.R., Lin, Y.C., Tung, K.Y., 2013. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, *36*(1), pp.16-24.

[9] Sandosh, S., Govindasamy, V., Akila, G., Deepasangavy, K., FemidhaBegam, S., Sowmiya, B., 2019. A progressive intrusion detection system through event processing: challenges and motivation. In *2019 IEEE International Conference on System, Computation, Automation, and Networking (ICSCAN)* (pp. 1-7). IEEE.

[10] Zhang, S., Liu, Y. and Yang, D., 2022. A Novel IDS Securing Industrial Control System of Critical Infrastructure Using Deception Technology. *International Journal of Digital Crime and Forensics (IJDCF)*, *14*(2), pp.1-20.

[11] Scherer, P., Vicher, M., Drazdilova, P., Martinovic, J., Dvorsky, J., Snasel, V., 2011. Using SVM and clustering algorithms in IDS systems. In *Proc. Int Conf. Dateso 2011, 2011*.

[12] Amudha, P., Karthik, S., Sivakumari, S., 2013. Classification techniques for intrusion detection-an overview. *International Journal of Computer Applications*, *76*(16)

[13] Erman, J., Arlitt, M., Mahanti, A., 2006. Traffic classification using clustering algorithms. In Proceedings of the 2006 SIGCOMM workshop on Mining network data (pp. 281-286).