Unsupervised Methods for Intrusion Detection Systems and Forensic Examination

Lisa Gopal, Samir Rana, Preeti Chaudhary, Vrince Vimal

Abstract: Crime is increasing with the widespread growth of digital world. The last decade has witnessed the elevation in the diversity and frequency of malicious usage of the network. Forensic investigators play a paramount role in the investigation based upon collection and analysis of facts from the crime scene. Intrusion Detection Systems, which are in use till date do not enlighten the trends in attack as they are built on various outmoded attack classes. IDSs that uses unsupervised techniques has been discussed in the literature. It is based on the requirement of labelled data as it is required in regular training or on the characteristics that elaborates each class without any knowledge in the prior. Despite of being widely popular among researchers and mammoth practical applications, fidelity of IDS Is yet debatable. This paper provides an exhaustive survey of the various unsupervised anomaly-based intrusion detection techniques and their potential usage in their respectivedomain.

Keywords: Forensic, IDS, Unsupervised Methods, Attacks.

I. INTRODUCTION

Software that is particularly designed to alert administrators when a malicious activity or any kind of security violation takes place is anIntrusion Detection System (IDS). Many IDS have been developing over the years that have been developed according to the requirement, reviewed accordingly and thenfinally is been evaluated. Internetworking made its debut in the year 1976. And that was the time when IDS came into existence. In the initial years, the intrusion detection was done manually by the administrator who was responsible for monitoring the traffic and other activities. But as the internetworking became popular, network size and complexity were increased and so was the probability of attacks and so the manual intrusion detection was a fail. So, at first data-mining approach came into existence that used supervised methods for intrusion detection but this approach was limited to the identify some limited number of attacks. Hence in the recent years, unsupervised methods and hybrid methods are gaining popularity in identifying anomaly intrusion detection.

Revised Manuscript Received on November 25, 2019.

Lisa Gopal, Assistant Professor, Graphic Era Hill University, Dehradun, India.
Samir Rana, Assistant Professor, Graphic Era Hill University, Dehradun, India.
Preeti Chaudhary, Assistant Professor, Graphic Era Hill University, Dehradun, India.
Vrince Vimal, Associate Professor, Graphic Era Hill University, Dehradun, India.

IDS that comes under four variants. IDS that checks for intrusion in the network is the Network Intrusion Detection System (NIDS), IDS that checks for intrusion internally in the host and also network is the Host-based Intrusion Detection System (HIDS), IDS that checks for any breach or intrusion in the Perimeter is the Perimeter Intrusion Detection System (PIDS) and virtual machine monitoring is done to check for intrusion is done by VM based Intrusion Detection System (VMIDS).NIDS identifies intrusion on the basis of network traffic and also by monitoring different hosts on the particular network. The main idea of NIDS is to monitor each packet to check for malicious activity. HIDS identifies malicious activity by analyzing the applications in the application logs, then analyzing the system calls, and also the system file modifications wherever required, and other activities. Here the sensor nodes usually consist of the software agent that checks for an intrusion.PIDS identifies malicious activities on the basis of disturbances on the perimeter of critical infrastructure. This can be done either by using electronics or more advanced fiber optic cable technology. The most recent type of IDS that is still under development is the Virtual Machine Intrusion Based System (VMIDS) that detects malicious activities using virtual machine monitoring so there will be no need to have a separate IDS.

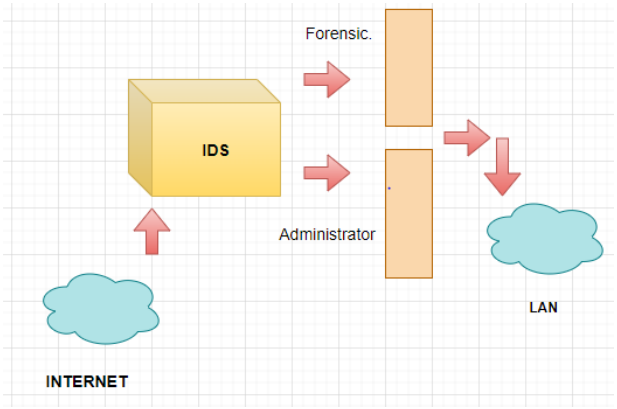

Fig. 1 Block Diagram of IDS

Since, time plays a crucial role in investigating the forensic data, and so combination of current forensic techniques with data-mining could be very effective solution and moreover such multistep and sophisticated attacks where complicated and progressive investigation is required that could lead to probably the reformation of occurred attacks, that can also be used to prevent from future attacks and might even lead to identifying the attacker, demands the collaboration of technologies such that of an IDS with the current forensic techniques.

Fig 1.1. shows a general representation of IDS where data before being transferred to the internal network is firstly monitored by the IDS and

then further sent to the forensic and then to the administrator if there is an alert. The traffic when considered to be safe is then passed to the local network. IDS can be collaboratively be combined with the forensic in order to provide more security in the network.Attacks that an IDS needs to detect are of four types. They are Denial of Service attack, Probe attack, User to Root attack and Remote to Local attack. DoS attack basically deals when the malicious nodes floods the attacking node with request packets till either all its resources are overwhelmed with packets or the attacking node crashes.The more prone risk in this is the DDoS attack where multiple connections are being established across the network. They generally target the infrastructure by making resources or service unavailable to the users. Recent biggest recorded DDoS attack was at GitHub with 1.3 Tbps of traffic that overwhelmedthe servers with 126.9 million packets of data each second.Such an attack lead GitHub system down for just 20 minutes then GitHub mitigation services identified the attack and was quick in taking the precautionary steps to prevent it. **Probe**is akind of attack for example port is been scannedandthat is used to range over the targeted network and then it collects the information related to hosts scanned by Probe, attacks as open ports, etc. Here the attacker deliberately wants to be identified so as to gather the required internal information. In **User to Root** attack, the malicious node has already access to the local targeted system. One of the common attacks is the buffer overwhelmingand this is done by the attacker so that it can execute the malicious code from backend. In **Remote to Local attack,** the malicious node tries to gain access tothe local machine because it does not have the access to targeted machine. Generally, this attack is combined with the User to Root attack. For example, SSH Brute Force. U2R and R2Lattacks are considered the most crucial of allattacks to detect as they are distinct and are generally misunderstood as normal traffic. This paper provides a preliminary survey on some of the best unsupervised methods used in detecting intrusion though IDS.

II. UNSUPERVISED IDS

The IDS architecture plays a crucial role in detecting the intrusion. Considering the IDS, architecture can be classified into three categories, **Centralized** where multiple sensor nodes sends data collectively to the central controlling unit from where the data is being analyzed and detected for intrusion. **Decentralized** is the hierarchical architecture. Here, there are multiple sensor nodes and multiple central controlling nodes. Before the data ends up at the main central unit, the senor nodes send the data to their nearest control unit from where data is pre-processed. Decentralized architecture is more efficient than the centralized architecture as the system performance is boosted during the pre-processing stage. The last is the **Distributed** architecture where there is no main control unit and the work is then distributed among all the nodes equally. Each node is responsible for gathering the data and processing it. Then using Peer to Peer network, the nodes communicate with each other. The agents act as sensor node as well as processing nodes at the same time. One of the major concerns for decentralized and distributed architecture is interacted between the agents for detecting the specific kinds of attacks as inability to interactamongst the nodes can lead to incapability of nodes to identify intrusion and also loss of analyzed data.

The main aim of using unsupervised techniques is to identify the malicious traffic from the normal one. The unsupervised method generally deals with cluster formation in order to identify the intrusion. The main idea behind this is to form cluster of dataset and then identifying the dissimilar behavioral dataset in order to identify intrusion. The unsupervised approaches that are proposed up till now are discussed here: Cases et al. have proposed their clustering technique into the IDS. They divided their approach into three steps. At first, after capturing the network, packets are converted into multiple flow packets and they are then being collected at flow-resolution levels prior to arriving to the change detection module. The main idea behind multi resolution flow is to capture and detect attacks of all level, be it small, large, single or distributed. If the changes are being detected then Sub-Space Clustering and Evidence Accumulation Algorithm (SSC-EA) clusters that particular data and then finally the threshold is defined to identify whether the cluster is malicious or not. The proposed methodology achieves high detection rate for all the four possible types of attack, i.e. Denial of Service, Probe attack, U2R and R2L attack. It is efficient and is good for real-time detection.

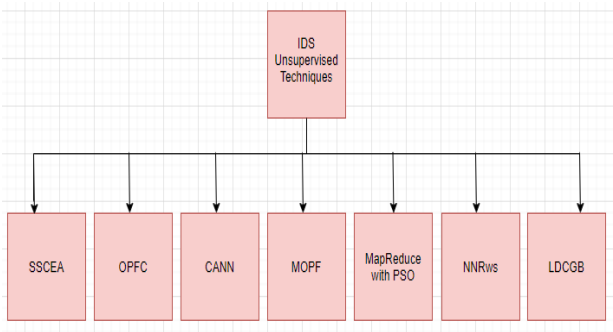

Fig. 2 IDS Unsupervised Techniques

Multistep Outlier-based detection approach is proposed by Bhuyan et al[6]. It consists of generalized entropy feature selection and mutual information(MIGE-FS). In order to improve efficiency and reduce the cost in terms of accuracy, it selects the most apropos subset of feature in each case. In order to identify whether the cluster is malicious or not, it identifies the cluster on the following basis. The TreeCLUS (TCLUS) algorithm produces a tree and each branch hererepresents a cluster. The nodes are further segregated on the basis of minimum feature-feature redundancy and maximum feature-class relevancy in a depth-first manner. At last, ROS outlier score is calculated, if it exceeds the predefined threshold value, then the cluster is considered to be malicious.

Costa et al. [7] developed Optimal Path Forest Clustering (OPFC). The OPF algorithm has been optimized via differentnature or inspired by Service Optimization Techniques. So, accurate

results are obtained by Bat Algorithm (BA) and Particle Swarm Optimization (PSO).

Optimum Path Forest modification algorithm (MOPF) is used by Bostani and Sheikhan [8], that consist of 3 modules, one for partitioning, next is pruning and finally detecting. Participating modules create training subsets using K-means later to be used for the detection. To modify the speed of the OPF, pruning is done to identify most informative subset. Lastly, the detecting module is totally based on the improved performance of OPF by 14.86%.

Linet al. [9] introduced Intrusion detection system based on Combining cluster centers and nearest neighbors (CANN).Thisapproach succeeded to identify the DoS and Probe attack only. This approach worked as follows: the cluster centers were first extracted using clustering techniques. Two distances are calculated, first between the cluster centers and dataset and the other one between its closest neighbor and each data point in the similar cluster. A new dimension feature is produced by adding both the distances. Finally, classification of data is done using K-NN by the newly constructed feature. Results shows that this approach is again useful in identifying DoS and Probe attack.

Hosseinpour et al. [10] combined the artificial immune system with Unsupervised learning. Data is divided into normal and malicious data and unsupervised learning is hereby worked as main to innate immunity. Based on the clustering results, detectors are hence generated which is done by supervised learning to represent the second adaptive immunity and then the results are distributed amongst the host once they are mature to receive the data. These detectors will be used by different hosts to interrupt the known attacks through the supervised methods.

MapReduce Technique with Particle Swarm Optimization (PSO) approach was used in Aljarah et al [11].In an approach the global optimal centroids, PSO is used for clustering. This was done after the data-processing. The main idea to introduce the MapReduce is that it gives the direction to IDS will adjust on huge networks and also to reduce the overall computational time. Proposed approach reached at 0.963 AUC with a 0.013 FPR. The capacities are specific in not determining the specific attack rather it is only capable in identifying the malevolent traffic from the non-malicious traffic.

A semi-supervised works on the concept of divide-and-conquer method was put forward by Ashfaq et al [12]. that basically categorize the unlabeled data by using the magnitude of the Fuzzy Logic. Authors have used as classifiers by adopting Neural Network with random weights (NNRw). The proposed model has an accuracy of 84%. Hence the model still required to be trained although it does not require labelled data in abundance.

Hui et al. proposed an approach based on local deviation coefficient that uses graph-based intrusion detection using outlier detection method (LDCGB). This method is able to detect the arbitrary shaped clusters, is able to distinguish the malicious cluster from the normal one. To decide the type of cluster, an outlier method is used based on the local deviation coefficient. It uses the graph-based (GB) algorithm.

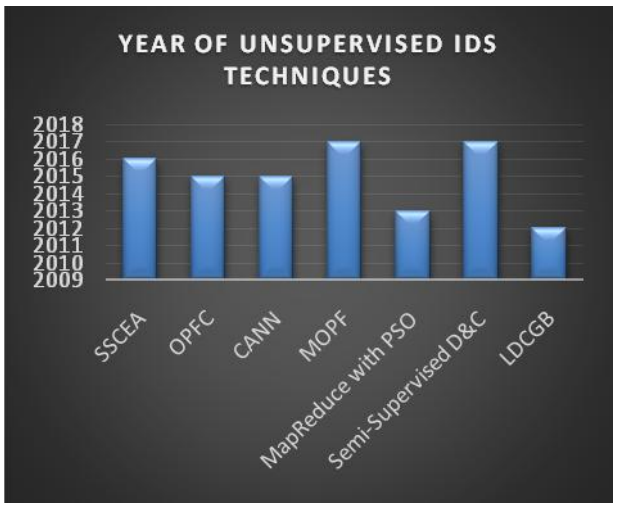

Fig.3 Various Unsupervised IDS Techniques

With the growing tendency of attack, IDS is becoming very important part of forensic in order to identify the type of malicious activity and response process of. The most important factor is the time in forensic for identifying the attack. By providing the output from the IDS to the forensic investigation team, thecorrelation of the evidence is done. It helps in identifying the scene of the crime and most importantly the time. It is important to identify the attack and combine all the investigation pieces so that the organization or the company can take the precautionary measures as early as possible. As automating the part of correlation has to be done manually and so it is a time taking process that could lead to late retrieval of time-sensitive data.

Awareness can be done amongst the employees so that no further loss of data takes place or more strong techniques can be implemented on the firewall or IDS in order to protect from malicious activities.

III. CONCLUSION

This paper surveys the various anomalies-based detection algorithms for intrusion based that uses unsupervised techniques for detecting the unknown attacks. The forensic investigation collects data from different source logs of different devices, network traffic, etc. Hence the unsupervised techniques are to be used in order to identify a malware. Various limitations of the IDS have also been discussed and also the possible extensions to the algorithms. The space between the forensic and the IDS needs to be completely removed by introducing the outputs of the IDS into investigation in the forensic in order to identify the attack and construct the timeline.

REFERENCES

1. A Nisioti, A Mylonas , P D. Yoo,, and V Katos," From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods", IEEE Communications Surveys & Tutorials, Vol. 20, No. 4, Fourth Quarter 2018.

2. E. Vasilomanolakis, S. Karuppayah, M. Mühlhäuser, and M. Fischer, "Taxonomy and survey of collaborative intrusion detection," ACM Comput. Surveys, vol. 47, no. 4, p. 55, 2015.
3. "State of the Internet, Q1 2017 report," Akamai, Cambridge, MA, USA, Rep., 2017. [Online].Available: https://www.akamai.com/fr/fr/multimedia/documents/state-of-the-internet/q1-2017-state-of-theinternet-connectivity-report.pdf.
4. M. A. Qadeer, A. Iqbal, M. Zahid, and M. R. Siddiqui, "Network traffic analysis and intrusion detection using packet sniffer," in Proc. IEEE 2nd Int. Conf. Commun. Softw. Netw. (ICCSN), Singapore, Feb. 2010, pp. 13–317.
5. A. Fahad, Z. Tari, I. Khalil, A. Almalawi, and A. Y. Zomaya, "An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion," Future Gener. Comput. Syst., vol. 36, pp. 156–169, Jul. 2014.
6. M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "A multi-step outlier-based anomaly detection approach to network-wide traffic," Inf. Sci., vol. 348, pp. 243–271, Jun. 2016.
7. K. A. P. Costa et al., "A nature-inspired approach to speed up optimum-path forest clustering and its application to intrusion detection in computer networks," Inf. Sci., vol. 294, pp. 95–108, Feb. 2015.
8. H. Bostani and M. Sheikhan, "Modification of supervised OPF-based intrusion detection systems using unsupervised learning and social network concept," Pattern Recognit., vol. 62, pp. 56–72, Feb. 2017.
9. F. Hosseinpour, P. V. Amoli, F. Farahnakian, J. Plosila, and T. Hämäläinen, "Artificial immune system-based intrusion detection: Innate immunity using an unsupervised learning approach," Int. J. Digit. Content Technol. Appl., vol. 8, no. 5, p. 1, 2014.
10. W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," Knowl. Based Syst., vol. 78, pp. 13–21, Apr. 2015.
11. I. Aljarah and S. A. Ludwig, "MapReduce intrusion detection system based on a particle swarm optimization clustering algorithm," in Proc. IEEE Congr. Evol. Comput., Jun. 2013, pp. 955–962.
12. R. A. R. Ashfaq, X.-Z. Wang, J. Z. Huang, H. Abbas, and Y.-L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system," Inf. Sci., vol. 378, pp. 484–497, Feb. 2017.
13. Z. Mingqiang, H. Hui, and W. Qian,"A Graph-based clustering algorithm for anomaly intrusion detection," in Proc. IEEE 7th Int. Conf. Comput. Sci. Educ. (ICCSE), Melbourne, VIC, Australia, Jul. 2012, pp. 1311–1314.
14. A. Bohara, U. Thakore, and W. H. Sanders, "Intrusion detection in enterprise systems by combining and clustering diverse monitor data," in Proc. ACM Symp. Bootcamp Sci. Security. Pittsburgh, PA, USA, Apr. 2016, pp. 7–16.
15. J. Song, H. Takakura, Y. Okabe, and K. Nakao, "Toward a more practical unsupervised anomaly detection system," Inf. Sci., vol. 231, pp. 4–14, May 2013.