INTRUSION DETECTION SYSTEM BASED ON DATA MINIG AND ARTIFICIAL INTELLIGENCE

[1]Prof.Kirti Randhe, [2]Prof.Vinay Thamke, [3]Prof.Kavita Ugale
[1]Assistant Professor, [2]Assistant Professor , [3]Assistant Professor
[1]Computer Engineering Department,
[1]International School of Buisiness and Media,(School of Tecnology), Pune, India

Abstract : **Intrusion Detection System (IDS) is meant to be a software application which monitors the network orsystem activities and finds if any malicious operations occur. Tremendous growth and usage of internet raises concerns about how to protect and communicate the digital information in a safe manner. We can build Intrusion detection system model to find attacks on system and can improve the system using captured data. By applying feature selection approach in machine learning, the NSLKDD data set obtained can be reduced and also can improve the intrusion detection using the captured data. By machine learning techniques, we can increase number of new unseen attacks the system of intrusion detection can be developed .Though efficient adaptive methods like various techniques of machine learning can result in higher detection rates, lower false alarm rates and reasonable computation and communication cost. With the use of data mining can result in frequent pattern mining, classification, clustering and mini data stream. This survey paper describes a focused literature survey of machine learning and data mining methods for cyber in support of intrusion detection.**

IndexTerms **- Local Area Network, Wide Area Network, Metropolitan Area Networks, Close Circuit Television, Security through Obscurity GPS, Global Positioning System, Point Of Access, Network Intrusion Detection System**

I. INTRODUCTION

Data Mining methods are described, as well as several applications of each method to cyber intrusion detection problems. The complexity of different machine learning and data mining algorithms is discussed, and the paper provides a set of comparison criteria for machine learning and data mining methods and a set of recommendations on the best methods to use depending on the characteristics of the cyber security.
Problem to solve Cyber security is the set of technologies and processes designed to protect computers, networks, programs, and data from attack, unauthorized access, change, or destruction. Cyber security systems are consist of network security systems and computer security systems. Each of these has, at a minimum, a firewall, antivirus software, and an intrusion detection system. System help to discover, determine, and identify unauthorized use, duplication, alteration, and destruction of information systems. The security breaches include external intrusions attacks from outside the organization and internal intrusions. There are three main types of cyber analytics in support of intrusion detection systems: misuse-based, anomaly-based, and hybrid. Misuse-based techniques are designed to detect known attacks by using signatures of those attacks. They are effective for detecting known type of attacks without generating an overwhelming number of false alarms. They require frequent manual updates of the database with rules and signatures. Misuse-based techniques cannot detect novel attacks. Anomaly-based techniques model the normal network and system behaviour, and identify anomalies as deviations from normal behaviour.The Machine learning and data mining methods covered in this paper are fully applicable to the intrusion and misuse detection problems in both wired and wireless networks.

II. INTRUSION DETECTION IN WIRED NETWORKS

An IDS should monitor traffic and detect malicious activities.IDSs can be categorized based on different modules. Figure shows IDS classes upon three modules [4]: data source, data analysis, and response. A data source can be gathered from either an individual computer (host-based IDS, HIDS) or network traffic (network-based IDS, NIDS).

IDS classes

There are two methods for analyzing the collected data: anomaly-based detection and misuse-based detection, which will be explained in the following sections. Anomaly- and misuse-based detection has general meaning for all environments. The third module specifies a suitable response for the suspicious data. This response can be passive or active in terms of behavior. As opposed to passive methods, active IDSs detect and respond to attacks [1, 2].

II. ANOMALY BASED DETECTION:

The pattern of normal behavior is used in anomaly-based detection, which can be either self-learned or programmed [4]. In the self-learned anomaly detection, the normal behavior of a system is built automatically. On the other hand, in the programmed detection, a system developer provides the model of normal behavior. Although an anomaly-based IDS is able to detect unknown attacks, it has a high false alarm rate and cannot distinguish between different types of attacks.

- **Statistical techniques:** This technique uses a statistical model for defining normal behavior of the components of the system. This technique shows that the probability of normal data instances is higher in a stochastic model in comparison with the probability of an anomaly occurrence [6].

• **Clustering-based methods:** In this method, normal data belong to a cluster, and data not included in any cluster is detected as anomalies [6].
• **Information theoretic**: This method assumes that anomalies cause irregularities in information content of the data set. Different information theoretic measures are used to analyze the information content, for example, entropy and Kolmogorov Complexity [6].
• **Bayesian networks:** The probabilistic relationships among variables are encoded in the Bayesian method. The combination of this method with a statistical scheme offers better detection capability [8].
• **Data mining methods:** Data mining is the application of machine learning in large databases to provide simple models [9].

• **Neural networks:** This method is inspired by the human brain.Neural networks are flexible and adaptable to environmental changes. They can be deployed to create a user profile, to detect intrusion, and to predict the future behavior of the traffic [8].
• **Support vector machines (SVMs):** The SVM is based on statistical learning theory. One class SVMs and multi-class SVM are well-known methods in classification and regression .
•**Nearest neighbor-based techniques:** In this method, the distance or similarity between two data instances is measured. Although normal data instances occur in dense neighborhoods, anomalies occur far from their closest neighbors [6].
• **Pattern matching**: Online learning is used in the pattern matching method to generate a traffic profile for each network. The profiles are used for anomaly detection. However, pattern matching may need to build traffic profiles for new networks,which results in a time-consuming process [7].

III.MISUSE BASED DETECTION SYSTEM

A misuse-based IDS, which is a programmed method, compares the user's activities with predefined signatures to find malicious traffic. Although this method is very accurate in detecting known attacks, it is not able to detect unknown attacks.

Misuse detection methods

• **Expert system:** There are a set of rules in an expert system in which rules describe attack behavior. The expert system can be used to consider the security state of the system.
• **String matching:** Numbers of misuse or signature-based IDSs use this technique, which is a substring matching of characters in texts. If there is a change in an attack signature, this method is unable to detect the attack. NSM is a model that is proposed .
• **Simple rule-based:** Expert knowledge about attacks can be modeled by the rule-based method. NADIR and ASAX are two methods that use a rule-based method.

IV.INTRUSION DETECTION IN HIGH-SPEED NETWORKS

- **Using Specialized Hardware**
  Hardware-based intrusion detection is a scalable method as it is able to inspect packets in high-speed networks. Most of the hardware-based NIDSs have been proposed to improve deep packet inspection (DPI), using some specialized hardware, such as a field-programming gate array (FPGA), an application-specific integrated circuit (ASIC), and ternary content addressable memories (TCAM) [5–6].
- **Using Flow-Based Traffic for Intrusion Detection**

A flow record is defined as a group of packets with a number of common properties, which pass a monitoring point in a certain time interval. Flow-based traffic contains only packet headers, and hence it reduces data. Flow-based IDS (FIDS) cannot detect attacks related to packet payload.

Therefore, it is not a replacement for packet-based IDS. The FIDS can detect attacks such as denial of service (DoS), scans, worms, and botnets. DoS attacks caused by payload contents cannot be detected by FIDS [5].

- **Using Sampling Techniques**

Sampling data is a method used for anomaly detection [2–3] and change detection, for example, DoS attack detection. Cisco NetFlow [3] is a sampling technique to decrease the heavy load on router CPU in high-speed networks. However, sampling has negative impacts on the statistical characteristics of traffic and hence on the performance of intrusion detection. There are different typesof sampling [5, 3, 3]:

1.Packet sampling: There are two types of packet sampling methods: systematic and random. In systematic packet sampling, a time interval, or a sequence of packet arrival, is chosen to select a packet. In random packet sampling, the probability distribution function is used as a basis of sampling.

2.**Flow sampling:** This method is more accurate than packet sampling. Random probability is used in random flow sampling to select flows**.**

3. **Smart sampling:** This method is proposed to control the size of sampling data. Both the smart and the sample-and-hold sampling are flow-sampling methods proposed to reduce required memory.

4. **Sample-and-hold:** The smart and sample-and-hold sampling methods try to provide precise traffic estimation for larger flows.

5. **Adaptive packet sampling:** In order to have an accurate traffic statistic, this method identifies the current traffic load to adjust the sampling rate**.**

6. **Selective flow sampling:** Although sampling techniques address the scalability problems, they affect anomaly detection efficiency. Selective flow sampling provides an appropriate balance between the performance and the amount of sampled information.

V.Learning Methods

1.Unsupervised Learning

There is no supervisor in unsupervised learning, and it is trained using unlabeled data only. Unsupervised learning is similar to a statistical clustering, in which they identify various groups of inputs using their similarity . The self-organizing maps (SOM) and the adaptive resonance theory (ART) are two examples of unsupervised learning. The SOM is an important neural network method used for the anomaly and misuse detection [1].

2.Semi-Supervised Learning

The semi-supervised learning method combines the supervised and the unsupervised learning capabilities. In this method, often some unlabeled data is provided in a data set besides labeled data. When labeling is expensive, this method can be useful even with less labeled data [4].The semi-supervised method can act as a supervised or an unsupervised learning according to the availability of labeled data. This method is employed for intrusion detection in several studies [4, 4, 5]. A semi-supervised learning-based method can be trained by both labeled and unlabeled data and has more accurate prediction. Low density separation (LDS) [5] and transductive SVM (TSVM) [5] are examples.

3.Reinforcement Learning

Reinforcement learning (RL) is a combination of the supervised and unsupervised learning. In reinforcement learning, there is an agent acting upon the environment. The state of the environment changes with the agent's actions, and the environment, in return, gives feedback for those

actions. The feedback is either a reward or a punishment. Thus, due to the existence of the feedback from the environment, RL is a form of supervised learning, but it is known as weak supervised learning because RL never presents the correct input/output pairs.

V.Methods used in AI

The improvement of theoretical and methodological ways to deal with transfer of big data from illustrative and parallel research and applications to ones that investigates easygoing and illustrative connections.

In [9] Yuehu Liu, Bin Chen et al. have proposed another technique for overseeing gigantic remote sensing image data by utilizing HBase and MapReduce framework.At first they have dividednormally exceptionally troublesome in general ways.Finally they notice that the speeds of data commerce and processing increase because the cluster of HBase grows.The outcomes demonstrate that HBase is extremely reasonable for large image information stockpiling and handling.The recovery technique is actualized dependent on Map Reduce. The Argo data is utilized to exhibit the proposed technique. The execution is looked at under a disseminated domain taking into account PCs by utilizing distinctive data scale and diverse task numbers. The examinations result demonstrates that the parallel strategy can be utilized to store and retrieve the vast scale NetCDF productively. Big data has turned into a noteworthy center of worldwide interest that is progressively pulling in the acknowledgment of the educated community, industry, government and other association. The incremental development in volume and changing.

- **Data Fusion**

Conventional data processing sometimes consider data from one domain. In this big data era, everyone has to make wide selection of datasets from totally different sources in several domains. Each of these datasets comprise of various strategies such as alternate representation, measurements, scale, dissemination, and consistency. Removing the force of information from numerous diverse (however conceivably associated) data sets is an extraordinary arrangement in big data research, which incorporates basically isolating big data from customary data mining undertakings. Which itself prompts propelled procedures that can comb data fusion and conventional data fusion contemplated in the database group [6].

Crowdsourcing

The term crowd sourcing means to data acquirement by vast and various gatherings of individuals, who much of the time are not prepared measurer and who don't have exceptional PC learning, utilizing web innovation. An assortment of data mining techniques can be applied to find associations and regularities in data, extract knowledge in the forms of rules and predict the value of the dependent variables. Common data mining techniques which are used in almost all the sectors are listed as: Naive Bayes, Decision Tree, Artificial neural network (ANN), Bagging algorithm, K- nearest neighborhood (KNN), Support vector machine (SVM) etc. Data mining is an important step of knowledge discovery in databases (KDD) which is an iterative process of data cleaning, integration of data, data selection, pattern recognition and data mining knowledge recognition. KDD and data mining are also used interchangeably. Data mining encompasses association, classification, clustering, statistical analysis and prediction. Data mining has been widely used in areas of communication, credit assessment, stock market prediction, marketing, banking, education, health and medicine, hazard forecasting, knowledge acquisition, scientific discovery, fraud detection, etc. Data mining applications include analysis of data for better policy making in health, prevention of various errors in hospitals, detection of fraudulent insurance claims early detection and prevention ofvarious diseases, value for more money, saving costs and saving more lives by reducing death rates.

CONCLUSION

Common data mining techniques which are used in almost all the sectors are listed as: Naive Bayes, Decision Tree, Artificial neural network (ANN), Bagging algorithm, K- nearest neighborhood (KNN), Support vector machine (SVM) etc. Data mining is an important step of knowledge discovery in databases (KDD) which is an iterative process of data cleaning, integration of data, data selection, pattern recognition and data mining knowledge recognition. KDD and data mining are also used interchangeably. Data mining encompasses association, classification, clustering, statistical analysis and prediction. A steeper Subthreshold Slope (SS) is obtained compared to conventional CMOS, because of the better electrostatic control and absence of doping.High-speed IDSs are required to handle this huge amount of traffic. Different high-speed intrusion detection techniques were described in this chapter. Use of artificial intelligence techniques has numbers of advantages due to their learning ability and adaptability. An artificial intelligence–based IDS is adaptable to environmental changes and is trained to detect even unknown attacks. The intelligent IDS may also be able to work in high-speed networks.

REFERENCES

[1.] Yin, C., Huang, S., Su, P., and Gao, C. Secure routing for large-scale wireless sensor networks. In Proceedings of IEEE ICCT 2003, 2 (April 9–11, 2003)

[2.] Hass, Z. J. Design methodologies for adaptive and multimedia networks. IEEE Communications Magazine, 39(11), (November 2001)

[3.] Heinzelman, W. B., Chandrakasan, A. P., and Balakrishnan, H. An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wire. Commun., 1(4) (2002)

[4.] P. P. Bonissone. Soft computing: the convergence of emerging reasoning technologies. Soft Computing— A Fusion of Foundations, Methodologies and Applications, 1(1):6–18, 1997

[5] J. Cannady. Artificial neural networks for misuse detection. In Proceedings of the 1998 National Information [10.] J. Frank. Artificial intelligence and intrusion detection: Current and future directions. In Proceedings of the 17th National Computer Security Conference, Baltimore, MD, 1994

[6.] J. Frank. Artificial intelligence and intrusion detection: Current and future directions. In Proceedings of the 17th National Computer Security Conference, Baltimore, MD, 1994.

[7.] A. H. M. Lichodzijewski, P.; Nur Zincir-Heywood. Host-based intrusion detection using self-organizing maps. In Proceedings of the 2002 International Joint Conference on Neural Networks, 2002.

[8.] M. Moradi and M. Zulkernine. A neural network based system for intrusion detection and classification of attacks. In 2004 IEEE International Conference on Advances in Intelligent Systems.

[9.] A. Mounji. Rule-Based Distributed Intrusion Detection. PhD thesis, University of Namur, 1997.