J. Basic. Appl. Sci. Res., 4(8)134-139, 2014

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Machine Learning: A Solution for Intrusion Detection

Mehr Yahya Durrani[1], M. Taimoor Khan[1], Armughan Ali[1], Ali Mustafa[1], Shehzad Khalid[2]

[1]COMSATS Institute of Information Technology, Attock, Pakistan
[2]Bahria University, Islamabad, Pakistan

Received: May 3, 2014
Accepted: July 6, 2014

ABSTRACT

Millions of users share resources and send and receive data daily through Internet. However, they are certainly at risk of data theft and other attacks due to this connectivity. Researchers are showing increasing trends in security related attacks. Network security has thus become one of the most active research fields. Intrusion Detection Systems (IDS) are commonly used for detection of attacks in a Network due to its ability to detect unknown attacks. Many techniques, ranging from statistical approaches to Artificial Intelligence (AI) based approaches have been presented in literature. AI based techniques have gained a lot of popularity in research community due to its various benefits. In this paper, we present a survey of Intrusion Detection Systems based on machine learning techniques.
KEYWORDS: ANN, Markov Model, Bayesian Network, Intrusion Detection System

I. INTRODUCTION

With the emergence of internet, resource sharing has become very easy and has created many interesting possibilities. However, this ease of use has its price; security related attacks are increasing day by day. Network security has, therefore, taken center stage in the research community to protect the system from intrusion activities. IDS have undergone significant growth recently due to its importance in network security. Unwanted network activities over the network have increased significantly. This has resulted in the spur of research activities aimed at the development of intelligent intrusion detection system to protect the network. Network security techniques range from Cryptography, firewall to Intrusion Detection Systems (IDS). Previously, network security systems were rule-based which can detect only specific events. However, the nature of attacks keep on changing which requires an intelligent and adaptive systems to detect variety of attacks that deviate from normal usage of networks. Recently, IDS have gained popularity due to its flexibility and detection of unknown attacks. IDS is an area which incorporates techniques from various disciplines.

Artificial Intelligence (AI) is a field of study which tries to make intelligent machines. AI is widely used in many fields including intrusion detection. AI based techniques include machine learning approaches, genetic algorithm based techniques, fuzzy logic based approaches and data mining based algorithms. Machine learning based techniques have shown promise and are very successful in accomplishing tasks related to intrusion detection. The incorporation of machine learning concepts in intrusion detection systems is appealing as existing techniques are not able to cater for the unknown and increasingly complex nature of security requirements in networks.

The beauty of machine learning is that it can deduce new information from the previously seen data. The objective of this research paper is to provide an in depth study of various techniques of intrusion detection which are based on machine learning.

The remainder of this paper is organized as follows; Section 2 of this paper provides a brief introduction of IDS. Section 3 introduces ANN and approaches of ANN for IDS. Discussion of Bayesian Networks and its techniques for IDS is covered in Section 4. Markov Model is described in Section 5. Section 6 discusses results of different schemes. The last section concludes the discussion.

***Corresponding Author:** Mehr Yahya Durrani, COMSATS Institute of Information Technology, Attock, Pakistan

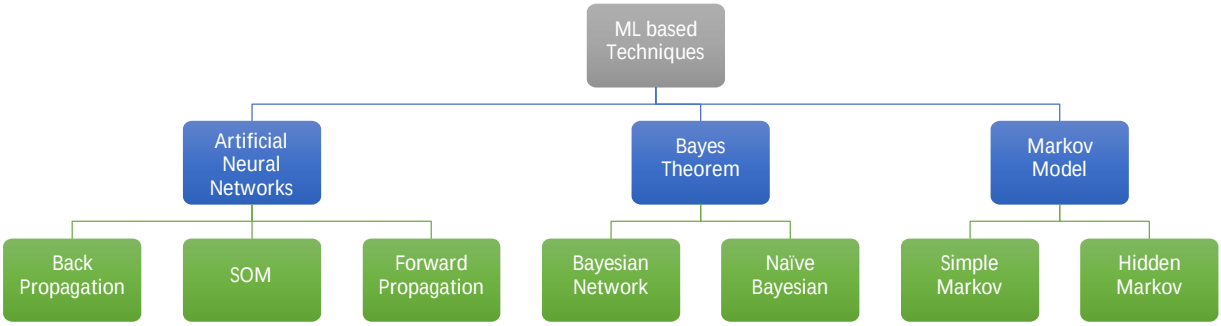

Figure 1: Taxonomy of Machine Learning based IDS

II. INTRUSION DETECTION SYSTEM

Intrusion Detection System is considered as a second line of defense in network Security. It is a class of application which detects, prevent and can take action against certain attack. In a generalized architecture [1],the IDS have a data storage unit and data analysis unit which take the decision of intrusion. The data to these units is fed by a monitoring unit which captures the data and forwards it to these units. Based on the result from data analysis unit, IDS can alert the administrator about the intrusion or can take action on its own. IDS can be categorized in to various categories.

On the basis of data storage, the IDS can be categorized as host based IDS which monitors a single host or it can be Network based IDS which monitors a certain network. On the basis of data processing, the IDS can be categorized as Signature based IDS which matches the attack signature with that of stored pattern or anomaly based IDS which makes a certain profile of a system and on the basis of deviation from a threshold can identify normal or abnormal traffic. Taxonomy of IDS can be Active IDS if it takes action on its own against anomalous entity or a *Passive IDS* which will only alert the Administrator about the anomaly.

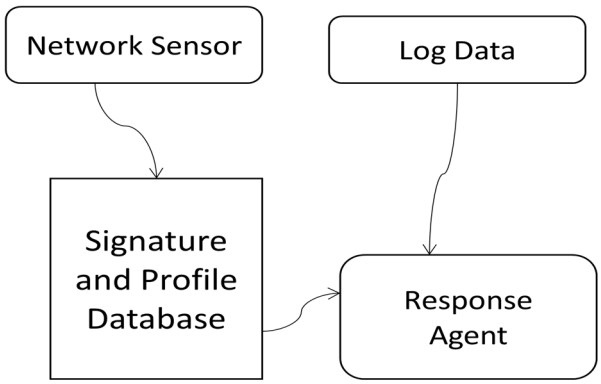

Figure 2: Architecture of IDS

III. ARTIFICIAL NEURAL NETWORK (ANN)

ANN is branch of AI which mimics human nervous system. The major component of ANN is neuron which applies certain function on input values and gives an output. The beauty of ANN is that it can tolerate imprecise data and can learn even if there are irregularities in the data.

In IDS, ANN is used for predicting a user's behavior. For this task, we feed data related to attacks and normal traffic using which ANN learns the characteristics of normal and abnormal traffic in training phase. The network adjusts its parameters to generate representation of normal and abnormal behaviors. Variety of neural networks based approaches have been proposed in literature [2-4]. These approaches can be further classified as supervised and unsupervised techniques. Cunningham and Lippmann [2] used key words related to attacks for intrusion identification. They searched these key words in the network audit data through which the decision regarding anomalous data was taken. Kayacik et.al [3] used Self Organizing Maps (SOM) for identification of intrusive data.

They employed a three layer SOM model for intrusion detection. The first layer employs some basic TCP protocol features. The second layer integrates the results of first layer in to a single view and third layer provides the output which classifies the event as normal or intrusive. Sarasamma et.al [4] proposed another SOM-based intrusion detection approach based on some selected attributes of dataset such as Protocol, Service, srcBytes, dstBytes, etc. Min Li et. Al [5] presented SOM based intrusion detection by creating special SOM for each feature. Each special SOM identifies a specific intrusion. Mortezaet.al [6] presented an intrusion detection system based on Adaptive Resonance Theory (ART)which is an unsupervised approach. They also compared their results with SOM based IDS and showed that their scheme performs better than the SOM based schemes. Yatimet.al [7] advocated the use of back propagation ANN for intrusion detection. On the contrary, Nabil et. al [8] proposed a feed forward neural network to identify intrusions. Ramakiet.al [24] used an ANN based IDS which incorporate correlation of alerts generated by IDS. Their scheme provides excellent results in distributed environments where one needs to provide a true picture to system administrator regarding the intrusion. Jahanbani et. al [25] proposed a feed forward based ANN based on a self organizing feature map. The focus of their scheme is training the ANN with abnormal traffic, unlike other schemes which use normal traffic, for this purpose. This scheme provides excellent results in comparison to other ANN based schemes.Alrajehet.al [28] proposed an IDS for wireless sensor networks based on ANN. The focus of the paper is the detection of energy exhaustion attack. Their scheme provides up to 95% detection rate for the said attack. A genetic algorithm based intrusion detection approach is proposed in [33]. The proposed system operates by capturing firewall entries as the features to monitor network activities. The proposed system achieves an accuracy of more than 95% in the detection of malicious connections. However, heavy resource requirements make this approach unsuitable for Wireless Sensor Networks (WSN).

Among the various approaches based ANN for intrusion detection, SOM based approach is very popular due to its high accuracy and short learning time. However, emphasis is also given on back propagation based techniques recently which improve the accuracy of the IDS.

IV. BAYESIAN NETWORK

Bayesian Network (BN), also referred to as Probabilistic Graph Model, is used to represent random variables and their relationships. More specifically, each random variable in a BN is represented by a node whereas their dependencies are represented by edges between them [23]. BN are widely used in Machine Learning, Speech Recognition and Bio informatics etc. BN are used in combination with some statistical techniques for intrusion identification.

Pearl & Russell [9] described Bayesian Network as graphical model for multivariable analysis. Various researchers have proposed IDS based on Bayesian Network. Krugelet.al [10] proposed an IDS which used multi-sensory approach from where data is aggregated and a single alarm is generated for intrusion identification. They created an event classification schemed which is based on BN. Johanssen and Lee [11] presented a simple model based on Bayesian Networks which learn from normal and intrusion data in the form of matrices. They used this model of normal and intrusive data for intrusion detection. Puttiniet.al [12] used Bayesian networks approach for intrusion detection that require human expert to check intrusion and normal behavior. XU and Shelton [13] used Continuous Time Bayesian Network for detecting intrusion. Their model incorporates system calls from Sun Solaris Security Module. They have identified three important attributes header, subject and return of a system call and used the same for intrusion detection. Liang et.al [14] proposed a technique which combines Naïve Bayes with decision trees. They claim to achieve one of the best results in IDS using Bayesian schemes. Mukherjee et. al [25] proposed an architecture for IDS which is based on Naïve Bayes Classifier. Their approach implies feature vitality based reduction and use 24 out of 41 features of KDD CUP dataset. Their results are comparable to the results of other schemes. Mukherjee et al. [26] focused on Remote to User (R2L) attack which is one of the most sophisticated attacks on computer network. They perform feature reduction and employ Bayesian classifier on compress feature space representation. The authors used 3 features (count, serror_rate, Srv_diff_host_rate) out of 41 features of KDD CUP dataset for intrusion identification. Their scheme provides one of the best results in detecting R2L attacks.

The major problem with Bayesian based approach is that it is a computationally expensive process which requires extra power and tends to be slow because of a large No. of features and their relationships in the intrusion detection datasets.

V. MARKOV MODEL

Markov Chaining is a stochastic process in which there is a chain of states and each state is dependent on its previous state. The change of state is governed by various probabilities associated with it. Hidden Markov Model

(HMM) is a double stochastic process where states are unobservable. The change in transition creates a chain known as Markov chain. For learning normal behavior an expectation maximization algorithm is used in HMM.

In Markov Chaining for intrusion detection, a set of variables with some probabilities and state transitions are estimated from sample training data. In detection stage, the data is checked against these probabilities and on the basis of a certain threshold, decision regarding normal and abnormal traffic is taken. Jalilzadehet.al [30] defines that Markov Chaining is a process which runs at discrete time intervals and exhibits markovian chain property. Due to its flexibility and simplicity it has gained popularity in many domains. Markov Chaining is widely used for intrusion detection. Most schemes based on this model inspect packet headers in order to identify intrusion. PHAD (Packet Header Anomaly Detection) [15] and ALAD (Application Layer Anomaly Detection) [16] are based on this approach. PHAD checks 33 fields in protocols of all the layers and assign anomaly score to each field. These scores are then combined to yield net anomaly score. However, the problem with PHAD is that it assumes independence of features which is normally not the case. Similarly, ALAD learns normal behavior of protocols on application layer of the network. The packet which deviates from normal behavior with a certain threshold is identified as suspicious. Yeunget.al [17] proposed a model based on sequence of system calls especially shell based calls in Unix environment. Srivastavaet.al [18] proposed a HMM based scheme which is used for fraud detection and provide interesting insight in HMM based anomaly detection.

Yi Xie and Shun Zeng [19] proposed a scheme for observing a user's browsing behavior based on Semi Markov model. Their approach identifies Distributed Denial of Service attacks (DDoS) attack on a target website. Ariu et al. [20] proposed HMM PAYL which is an IDS based on *n*-gram based approach which uses frequency distribution in protocol's payload for intrusion identification. Rafi et. al [23] combined HMM with Clustering for intrusion identification which gives promising results. Wang et. Al [21] provided a good survey about the HMM techniques applied on the domain. They have advocated incorporation of statistical techniques such as Rough Sets and ANN. Markov models are best suited for intrusion detection if the mechanism for checking system calls is implied. Bao et. al [29] proposed an IDS for Mobile Ad Hoc Network (MANET) based on HMM. Their scheme is a cluster based scheme which gives concept of Trust based intrusion detection. The authors claim that trust based IDS is better than other IDS schemes. Anari et. al [31] focuses on intelligently predicting user behavior using Markov Chaining process. The authors claim that their results are better than ANN for student behavior prediction.

VI. COMPARISON OF DIFFERENT APPROACHES

Intrusion Detection is a non-trivial task. Many researchers have worked and proposed various IDS based schemes. These schemes range from Networks to Machine learning approached and fuzzy rules etc. Although many sophisticated approaches have been proposed, no system can reliably detect all intrusive activities. Table 1 presents a comparative analysis of various machine learning based IDS schemes.

ANN based approaches provide a very high success rate for intrusion identification. The beauty of these approaches is that it require less training time and their resource requirement are not too high which makes it feasible to even apply on those devices which have low computation power. Bayesian Network based approaches perform well but their accuracy is less than the approaches of ANN. Further, they require very high computing resources but their training time is very low as compare to ANN based techniques. Another reason for low accuracies of Baeysian techniques is their assumption of feature independence which is normally not the case. Techniques based on HMM provide accurate results which are comparable to ANN. However, HMM based approaches are computationally expensive especially in terms of training time. Most of the HMM based IDS techniques work in offline mode in the training stage. However, some recent work has suggested optimizations in HMM techniques through which training time can be reduced and online training is possible. HMM based techniques are extremely popular in host based environments where system calls are monitored. On the basis of CPU utilization and other resource consumption, HMM techniques can effectively identify intrusions. Kumari et al. [34] has presented comparison of variety of classifiers including Decision Tree, Naïve Bayes, *k*-Nearest Neighbor and Support Vector Machines (SVM). They found the performance of decision trees and k-NN better than other classifiers on NSL-KDD dataset.

Latest trend in research shows that more researchers are now focusing on hybrid approaches that combine variety of approaches to identify intrusions based on outputs from multiple IDS. Further classical data mining based approaches such as classification, clustering and association rules mining are also incorporated thus effectively creating a hybrid IDS which can be effective as well as adaptive.

Type	Name	Dataset	Accuracy	Training Time	Resource Requirements
Artificial Neural Network	Back Propagation (Yatimet. al)	RLD 2009	98.6%	Very High	High
	SOM (Li Min et. al)	Sendmail UNM	95%	Low	High
	ARTNet (Aminiet. al)	Sendmail	97.19%	Very Low	Low
Bayesian Network	Simple Bayesian Network (Krugelet. al]	MIT Lincoln Lab	89%	Low	Very High
	Naïve Bayes	KMUTT 2009	79%	Low	Very High
	Naïve Bayes + DT (Shengliet. al)	--------	97%	High	High
Markov Model	HMM (Mahoney)	DARPA 1999	71%	High	Very High
	MultiLayer HMM (Ariuet. al)	KDD 99	98.3%	High	Very High
	HMM with Clustering (Rafi et. al)	KDD 99	98.5%	High	Very High

Table 1: Comparison of various Machine Learning based IDS schemes

VII. CONCLUSION AND FUTURE WORK

Machine learning is a domain of AI which is used vastly in many areas of computer science. Machine learning based intrusion detection techniques exhibit very good performance and their accuracy is very high. At the same time, they perform extremely wellin the presence of novel attacks and security issues introduced by the attackers to disrupt the network operations [32].This paper provides an in depth review of Machine Learning approaches used in IDS. It provides a solid understanding of the major techniques and approaches thus providing a basis for further research in this area. We are planning to further investigate intrusion detection approaches which incorporate ANN, BN and HMM based techniques together. Similarly, we plan to look into other AI based techniques including data mining approaches that can provide better results in the field of intrusion identification in the presence of novel attacks.

VIII. REFERENCES

[1] Axelsson, Stefan, 2000. Intrusion detection systems: A survey and taxonomy. Vol. 99. Technical Report

[2] Cunningham R, Lippmann R, 2000. Detecting computer attackers: recognizing patterns of malicious stealthy behavior. MIT Lincoln Laboratory—presentation to CERIAS

[3] Kayacik G, Zincir-Heywood N, Heywood M., 2003. On the capability of an SOM based intrusion detection system. In the Proceedings of the 2003 IEEE IJCNN Conference, pp: 1808-1813.

[4] Sarasamma, Suseela T., Qiuming A. Zhu and Julie Huff, 2005. Hierarchical Kohonenen net for anomaly detection in Network Security: Systems, Man, and Cybernetics, Part B: IEEE Transactions on Cybernetics.,35 (2): 302-312.

[5] Min, L. I., and Wang Dongliang., 2009. Anomaly Intrusion Detection Based on SOM.In the Proceedings of the 2009 WASE International Conference, pp: 40-43.

[6] Amini, Morteza, Jalili, and Shahriari, 2006. RT-UNNID: A practical solution to real-time network-based intrusion detection using unsupervised neural networks. Computers & Security., 25 (6): 459-468.

[7] Yatim, Utomo, 2006. Optimization of Variable Speed Induction Motor Drive using Online Backpropagation.In the Proceedings of the International Conference on Power and Energy, pp: 441-446

[8] Hachem, Nabil, Alfaro and Debar, 2013. An Adaptive Mitigation Framework for Handling Suspicious Network Flows via MPLS Policies. Secure IT Systems Springer Berlin., (2013): 297-312.

[9] Pearl, Judea, Russell, 2000.Bayesian networks. UCLA Cognitive Systems Laboratory. Technical Report.

[10] Kruegel, Christopher, Mutz, Robertson, Valeur, 2003. Bayesian event classification for intrusion detection. In the Proceedings of 2003 IEEE Conference of Security Applications, pp: 14-23.

[11] Johansen, Lee, 2003. CS424 network security: Bayesian Network Intrusion Detection (BINDS).

[12] Puttini, Ricardo, Marrakchi, Ludovic Mé, 2003. A Bayesian classification model for real-time intrusion detection. In the Proceedings of 2003 AIP Conference, pp: 150-162.

[13] Shelton, Fan, Lam, Lee, Xu, 2010. Continuous time Bayesian network reasoning and learning engine. The Journal of Machine Learning Research.,11 (2010): 1137-1140.

[14] Sun, Haitao, Liu, Chen and Zhang, 2011. HTTP tunnel Trojan Detection Based on Network Behavior. Energy Procedia, 13 (2011): 1272-1281.
[15] Mahoney, Matthew, Chan, 2001. PHAD: Packet header anomaly detection for identifying hostile network traffic. Florida Institute of Technology Technical Report CS-2001-04
[16] Mahoney, Matthew, Chan, 2002. Learning Nonstationary Models of Normal Network Traffic for Detecting Novel Attacks.In the Proceedings of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, pp: 376-385.
[17] Yeung, Dit-Yan, and Yuxin Ding, 2003. Host-based Intrusion Detection Using Dynamic and Static Behavioral Models.Pattern Recognition., 36 (1): 229-243.
[18] Srivastava, Kundu, Sural and Majumdar, 2008. Credit Card Fraud Detection using Hidden Markov Model. IEEE Transactions on Dependable and Secure Computing.,5(1): 37-48.
[19] Xie, Yi, and Yu, 2009. A large-scale Hidden Semi-Markov Model for Anomaly Detection on User Browsing Behaviors. IEEE/ACM Transactions on Networking.,17 (1): 54-65.
[20] Ariu, Davide, Tronci and Giacinto, 2011. HMMPayl: An Intrusion Detection System Based on Hidden Markov Models. Computers &Security., 30 (4): 221-241.
[21] Wang, Panhong, Shi, Wang, Wu, Liu, 2010. Survey on HMM Based Anomaly Intrusion Detection Using System Calls. In the Proceedings of 2010 IEEE International Conference of Computer Science and Education, pp: 102-105.
[22] Hassan, Rafiul, Nath, Kirley, 2006. A Data Clustering Algorithm Based on Single Hidden Markov Model.In the Proceedings of the 2006 International Multi conference on ISSN. pp: 57-66.
[23] Ben-Gal, Irad, 2007. Bayesian Networks. Encyclopedia of Statistics in Quality and Reliability.
[24] Ramaki, Atani, Abadi and Tavaghoe, 2013. Enhancement Intrusion Detection using Alert Correlation in Co-operative Intrusion Detection Systems. Journal of Basic and Applied Scientific Research., 3 (6): 272-279.
[25] Jahanbani, Keshtgari and Monadjemi, 2012. Intrusion Detection System Using New Synthetic Neural Networks. Journal of Basic and Applied Scientific Research., 2 (5): 4667-4671.
[26] Mukherjee, Saurabh and Sharma, 2012. Intrusion Detection using Naïve Bayes Classifier with Feature Reduction. Procedia Technology., 4 (2012): 119-128.
[27] Altwaijry and Hesham, 2013. Bayesian based Intrusion Detection System. IAENG Transactions on Engineering Technologies., (2013): 29-44.
[28] Alrajeh, Khan, Lloret, and Loo, 2013. Artificial Neural Network based Detection of Energy Exhaustion Attacks in Wireless Sensor Networks Capable of Energy Harvesting. Journal of Ad Hoc & Sensor Wireless Networks., (2013): 1-25.
[29] Bao, Chen, Chang, and Cho, 2012. Hierarchical Trust Management for Wireless Sensor Networks and its Applications to Trust-based Routing and Intrusion Detection. IEEE Transactions on Network and Service Management.,9 (2): 169-183.
[30] Jalilzade and Jamali, 2013. Modeling Based on Hidden Markovian Chain in Mobile Ad Hoc Networks. Journal of Basic and Applied Scientific Research., 3 (1): 40-44
[31] Anari and Sabri, 2012. Intelligent E-Learning Systems Using Student Behavior Prediction. Journal of Basic and Applied Scientific Research., 2 (12): 12017-12023
[32] Jeyanthi, Iyengar, Kumar and Kannammal, 2013. An Enhanced Entropy Approach to Detect and Prevent DDoS in Cloud Environment. International Journal of Communication Networks and Information Security., 5 (2): 110-119.
[33] Dhak, Bharat and Lade, 2012. An Evolutionary Approach to Intrusion Detection System Using Genetic Algorithm. International Journal of Emerging Technology and Advanced Engineering., 2 (2): 632-637.
[34] Kumari and Ranjitha, 2013. Intrusion Detection- A Comparative Analysis Using Classification Algorithms. Networking and Communication Engineering, 5 (2): 85-89.