International Journal of Advanced Technology & Engineering Research (IJATER)
National Conference on Recent Trends in Science, Technology & Management (NCRTSTM-2018)

THE SCOPE OF MACHINE LEARNING TECHNIQUES IN AN EFFICIENT INTRUSION DETECTION SYSTEM

Ghanshyam Prasad Dubey[1], Dr. Rakesh K Bhujade[2]
PhD Research Scholar[1], PhD Supervisor[2]
Department of Computer Science and Engineering, Mandsaur University, Mandsaur, MP 458001, India

Abstract

Artificial Intelligence and Machine Learning help in effective decision making. Systems developed using AI or ML-based techniques possess the ability to learn and acquire new knowledge; yet operating with the same efficiency in the domain of their interest. Incorporating an ML-based technique for the development of an Intrusion Detection System will enhance the capabilities of the IDS and may increase the Sensitivity and reduce the Specificity. These IDS will work in an optimal manner on the known attacks and at the same time; will learn about new attacks and enhance their Performance. IDS are very important to avoid unauthorized access and overcome the Malware attacks, especially in Networks. This paper provides an overview of the various ML-based techniques that can be used to implement IDS.

Keywords: Intrusion Detection System, Machine Learning, KDD-99, NIDS, Neural Network.

1. Introduction

Intrusion Detection System can be considered as an interface for preventing a Network Host or System from unauthorized access, security breaches, and malware attacks. It continuously monitors the traffic coming in and going out of the Network. It also ensures that invalid or erroneous data must not be allowed to pass in either direction in any case. It will generate an alarm or can take a predefined action in case of an attack or abnormal event in the Network or System. IDS are really efficient to detect known attacks but its performance degrades in case of new attacks or unknown attacks; thereby leading to improper Specificity [1]. IDS can be considered as a solution that can identify, assess, and claim an abnormal network action.

An IDS is mainly based on two approaches, as Anomaly Detection and Signature Detection [2]. Anomaly Detection approach observes the behavior of systems and nodes, while Signature-based Detection methods already know the attack pattern. Through comparing observed signatures to signatures in the database, the signature-based approach works. The database is a repository that includes a list of documented attack signatures. Any signature pattern created by extracting a feature from packets in a controlled environment that matches signatures in the database is flagged as a security policy violation or an attack.

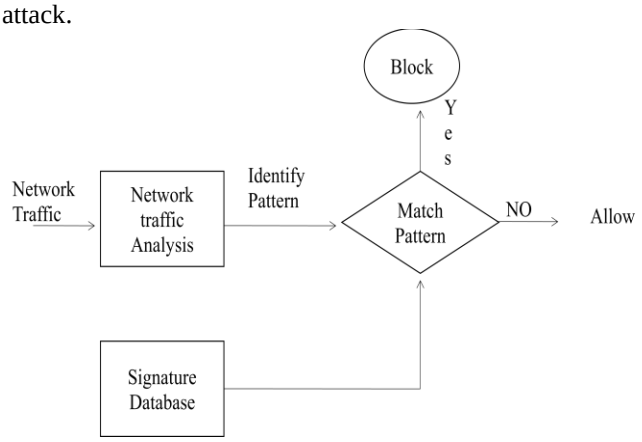

Figure 1: Working of signature-based IDS.

In Anomaly Detection, each node has a specific behavior and if there is a deviation in its behavior then the system raises an alarm and suspends the execution of this node. The system also raises an alarm, when certain part of the network starts misbehaving and takes other appropriate actions [1], [2]. In the Misuse Detection scheme, the characteristics of existing attacks form the basis of detection of an attack [3]. If malicious or suspicious activity is already known to the system, it will recognize the attack and raises an alarm; however, this scheme fails to detect new or unknown attack patterns [4].

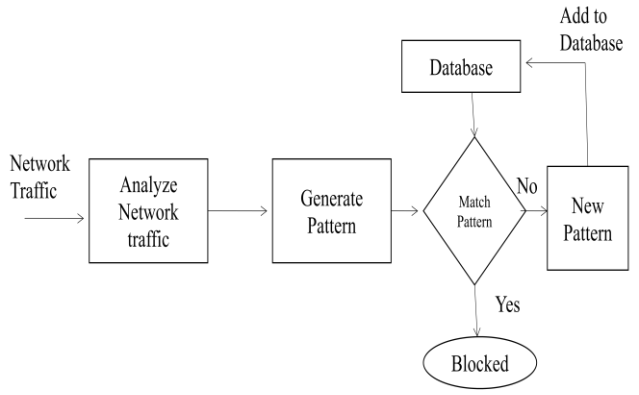

Figure 2: Working of Anomaly based IDS.

According to location where IDS can be deployed, it is classified into two categories namely host-based intrusion detection system abbreviated as HIDS and network-based intrusion detection system shortly known as NIDS [5], [6]. HIDS is deployed at a specific host to monitor its behavior; while NIDS is deployed at the gateway or server that observes traffic, behavior and monitors for identification of attacks. According to currents trends are concerned, IDS combines both host-based and network-based information to develop Hybrid systems for achieving better performance and security; thereby providing the benefits of both HIDS and NIDS. The client versions in such systems are used to monitor the behavior of hosts and send a report to the network manager, who validates the activity and decides whether it is malicious or not. If an attack or deviation in behavior is identified, the system takes an immediate action.

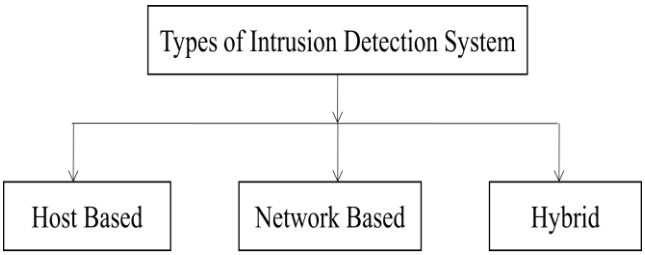

Figure 3: A simple classification of IDS.

HIDS can easily identify the attempts of legitimate system users for unprivileged access [7]. NIDS are mainly signature-based in nature and continuously monitors the network gateway for detecting packets containing malicious information. Few NIDS also use anomaly detection, as detection and notification of suspicious activities is their major objective. Anomaly detection has a bottleneck of high false positive rate; but has the benefit of predicting new or unknown attack style efficiently [8]. The performance of the Anomaly-based IDS can be improved significantly by using statistical-based, knowledge-based, and machine learning-based techniques. Statistical approach represents the conduct of the scheme using a number of viewpoints for analyzing the actual and malicious performance. The Knowledge-based approach compares the occurring activities with existing information and decides whether the activity is normal or malicious. Machine learning-based approach works well with both statistical as well as knowledge-based approaches; another benefit of using ML is that it not only detects malicious activities, but can predict the possible new or unknown attack patterns also [9].

2. Machine Learning Techniques

Machine Learning (ML) is a domain of artificial intelligence that imparts learning ability in the machines according to the situation and further applies this knowledge to solve the upcoming problems. To develop the ability of automatic learning of new attack patterns machine learning is used in anomaly-based IDS.

Anomaly-based IDS enriched with machine learning techniques can employ a system that will be able to learn from data (experience) and can take the decision for unseen data. Machine Learning approaches that can be used to implement an IDS with high effectiveness and efficiency are:

i. Neural Network
ii. Support Vector Machine (SVM)
iii. Genetic algorithms
iv. Fuzzy Logic
v. Bayesian Networks
vi. Decision Tree

a. Neural Network

Artificial neural networks or ANN is a collection of highly interconnected processing elements that processes a group of inputs to a certain group of desired outputs, inspired by biological nervous systems similar to the human brain. It is a computing technique that signifies the behavior and working of the human brain. The simplest, known and widely used neural network for developing IDS is Multilayer Perceptron (MLP). They are highly accurate in operation and can perform complex computations and processing by increasing the number of hidden layers. Hidden layers are integral part of neural networks, as they are majorly responsible for performing the computations. IDS developed using neural network approach will be able to classify an activity as normal or malicious and can also detect the type of attack, in case; it has information about

the distinguished variety of attacks. The only bottleneck that neural network exhibits is that their training time is long, as it incorporates the large number of training vectors for computations and learning. [10].

b. Support Vector Machine

SVM is another common and admired technique for machine learning [1]. The first step in SVM is to initialize the dataset and perform dimensionality reduction to remove redundant information and reduce the size of dataset. Support Vector Machine is highly efficient approach for implementing Classification with small datasets. SVM is highly applicable in various fields like Network Intrusion Detection, web page identification, and face identification. IDS based on SVM offer the benefits of increased training rate and relevant decision rate, regardless of the dimension of input data; continuous tuning of various parameters will lead to increase in accuracy of operation of the developed model. This makes SVM highly popular in the domain of Security.

c. Genetic Algorithm

Genetic algorithm (GA) is used to find fairly accurate results to an optimization task inspired by biological, evolution process, and natural genetics. GA based models mainly comprises of four operators, as initialization, selection, crossover, and mutation. GA based IDS can easily detect an intrusion activity, based on the past behavior [11]. GA based model will be trained on the normal behavior profile; after training, the model will classify any unknown pattern as malicious or normal. GA can also be used to specify the rules for NIDS. Attributes like service flags, login status and super user attempts can be easily stored as genes in a genetic chromosome using GA. These types of attacks are very common and can be accurately detected as compared to other unknown attacks. GA also plays an important role in the derivation of rules for implementing Classification and also in selection of optimal parameters for detection [12].

d. Fuzzy Logic

Fuzzy set theory forms the basis of Fuzzy Logic, which performs approximate reasoning for classification rather than precise and accurate logic-based classification. Fuzziness helps in handling probabilistic, uncertain, missing and inconsistent data and thus, Fuzzy techniques are highly suitable and useful in anomaly detection-based IDS. Fuzzy logic permits an object in a Fuzzy space to belong to multiple classes simultaneously. This helps in differentiating the object among several undefined classes. It is very useful and appropriate in the case of IDS; in IDS, the dissimilarities between normal and malicious classes are not properly defined [13], [14].

e. Bayesian Network

The Bayesian network is a model that encodes probabilistic relationships among the variables of interest. It can be applied to develop IDS models based on statistical schemes. The most striking feature of the Bayesian network is its ability and power to encode probabilistic relationships among the necessary features and to incorporate both Prior knowledge and data. The drawback of Bayesian approach is that it is difficult to handle continuous features and developing an efficient classifier is a tough and complex task, if existing facts and axioms are inconsistent [14].

f. Decision Tree

The Decision Tree algorithms are majorly used for classification of problems that deal with the data set is learned and modelled. Decision Tree algorithm can also be used for Intrusion Detection due to its efficient learning and training capability. It is highly accurate in classifying an activity as normal or malicious. The most important benefit of Decision Tree is that it is highly effective, even with huge datasets. It has the highest detection performance and can construct and interpret model easily; thereby making Decision Tree the most compatible technique for Real-time Intrusion Detection. High generalization accuracy of the Decision Tree is also another useful property that makes it highly applicable in IDS [15].

3. Literature Review

Here is the review of few machine learning algorithms that were proposed lately.

Gozde Karatas and Ozgur Koray Sahingoz [16] compared various network training function in a multi-layered artificial neural network. They preferred and compared neural network training functions like TRAINBR, TRAINC, TRAINCGP, TRAINLM, TRAINOSS, TRAINR and TRAINSCG. They used 2 hidden layers in their proposed artificial neural network for training and testing their selected functions using KDD-99 [17] data set. As per their experiment results, TRAINLM function is the best for the IDS application aresa which was implemented as a pattern recognition problem with five different patterns (DoS, U2R, R2L, Probe, and Normal).

Manjiri V. Kotpalliwar and Rakhi Wajgi [18] used Support Vector Machine (SVM) for classification of attack

in a large set of KDD-99 dataset [17]. SVM forms the basis for training the Network and validating the Data. As per experimental results, they were able to obtain the validation accuracy up to 89.85% along with classification accuracy of 99.9%. Kong, Lingjing, Guowei Huang, and Keke Wu [19] developed an abnormal traffic identification system (ATIS) using SVM to categorize several attack types and prevent local optimization to solve a 2-class classification problem. The ATIS can categorize and recognize numerous attack network flows which help better manage and defend the proposed network. According to the experimental results, scaling can much improve the speed of training time and accuracy.

Peiying Tao, Zhe Sun, and Zhixin Sun [20] suggested GA and SVM based intrusion detection systems, where GA and SVM were accustomed to choose the optimal feature subset and optimize the SVM parameters and feature weights to improve the network intrusion detection system. First, they select features based on GA and SVM followed by parameter optimization and data feature weighting using both techniques to enhance the true positive rate of IDS. A. Midzic, Z. Avdagic and S. Omanovic [21] recommended a hybrid model for intrusion detection using neural network and fuzzy logic. Self-Organizing Map (SOM) block was responsible for the reduction of training data through the process of clustering data in smaller subsets or clusters. These clusters are used in the Adaptive Network-Based Inference System (ANFIS) for training the system. Fuzzy logic is used in the ANFIS system. They used the KDD-99 dataset for the training and testing of the system.

Yunpeng Wang et al [22] proposed an advanced Naïve Bayesian classification-based machine learning model for the efficient intrusion detection system. This advanced Naïve Bayesian Classification (NBC-A) is a combination of traditional NBC and RELIEFF algorithm; both are used to train and test the network behavior using KDD-99 dataset. According to results obtained, NBC-A was suitable for large scale and complex dataset with a higher rate of a true positive.

Bhupendra Ingre, Anamika Yadav, and Atul Kumar Soni [23] recommended decision tree-based Classification and Regression Tree (CART) algorithm for effective attack categorization using NSL-KDD [24] dataset. The Correlation-based Feature Selection (CFS) subset evaluation algorithm was used to select optimal features; then the CART decision tree-based algorithm evaluates the performance of the dataset. The optimal feature selection method was used to enhance the accuracy; although significant enhancement has been found in NSL-KDD dataset after applying the decision tree algorithm and hence detection rate of all the attacks has improved.

R. A. R. Ashfaq et al (2016) proposed a Fuzzy based Semi-Supervised Learning Approach for Intrusion Detection. They proposed a Fast-Learning Mechanism for Single Hidden Layer Feed Forward Neural Network with Random Weights and Fuzziness for Intrusion Detection. Their approach is based on the principle of Divide and Conquer Algorithm Design technique. The fuzziness of each Sample is evaluated and classified into 3 categories as High Fuzziness Samples, Low Fuzziness Samples, and Mid Fuzziness Samples. Samples with High and Low Fuzziness are used to retrain the System. Neural Network with Random Weights has shown an Excellent Learning Performance and is Computationally Efficient. The proposed system has shown a High Accuracy Rate for Samples with High and Low Fuzziness but it has shown Low Accuracy Rate with Samples having Mid Fuzziness [25].

Javaid et al. [26] proposed an approach to trained Intrusion detection system using deep learning mechanisms. Self-taught Learning (STL) and deep learning-based technique used to improve the performance of the network intrusion detection system. Self-taught Learning (STL) is a kind of deep learning method consists of two classification stages. In the first step deal with the good classification of features from large unlabeled dataset (Unsupervised Feature Learning) and second step applied to leveled dataset for classification. They implemented sparse auto-encoder and soft-max regression-based NIDS using the NSL-KDD [24] dataset to calculate the accuracy of anomaly detection [26].

4. Comparative Study

IDS can be considered as Optimal if it exhibits 3 major characteristics, as High Sensitivity, Low Specificity and it must be able to learn and detect new or unknown attacks. IDS developed using ML-based techniques exhibits Learning ability and can fulfill the above 3 requirements for being an excellent IDS.

The table below shows the Comparative Analysis of the various ML-based techniques for implementing the IDS.

Table 1: Comparative Analysis of various ML-based Techniques for IDS.

Technique	Benefits	Drawbacks
Neural	~ High Accuracy	~ Training Time is

Network [16], [26]	~ High Sensitivity	more
Support Vector Machine [18], [19]	~ Better Precision ~ Low Training Time	~ Performance reduces with increase in Size of Data Set
Genetic Algorithm [20]	~ High Accuracy ~ Optimal Performance	~ Training Time is more
Fuzzy Logic [21], [25]	~ Easy to understand and implement ~ Handle Uncertainty	~ Low Accuracy ~ Can be used in Hybrid manner with GA or NN always
Bayesian Classifier [22]	~ Handle Uncertainty ~ Easy to understand	~ Low Accuracy
Decision Tree [23]	~ High Accuracy ~ Handle Large Data Sets efficiently	~ Accuracy lowers in case of Error or Missing Values in Data Set

5. Conclusion

This paper provides an introduction about the various Machine Learning based techniques that can be used for implementing the Intrusion Detection System. IDS developed using ML-based techniques are highly Accurate and Adaptive as compared to other state of the art techniques for implementing IDS. At the same time, ML-based IDS are self-learning in nature. Every Entity in this world has its advantages and disadvantages; the same applies to ML techniques also. Some techniques are highly accurate like NN and GA, but their Training Time is higher as compared to others. Fuzzy and Bayesian approaches are easy to implement, but they are having the bottlenecks of low Accuracy and Performance. SVM is accurate but only with small Data Sets. Decision Tree is highly accurate but it can't cater to missing data. For IDS, it is necessary that Accuracy can't be compromised. Hybrid models can be developed for implementing IDS using two or more ML techniques in such a way that the Training Time should be reduced and the Accuracy of the System must be enhanced as much as possible; at the same time, ensuring that the implementation must not become too complex.

References

[1] Rais, Helmi Md, and Tahir Mehmood. Dynamic Ant Colony System with Three Level Update Feature Selection for Intrusion Detection. International Journal of Network Security, 20(1), 184-192 (2018).

[2] Baig, Zubair A., Sadiq M. Sait, and AbdulRahman Shaheen. GMDH-based networks for intelligent intrusion detection, Elsevier Engineering Applications of Artificial Intelligence, 26 (7), 1731–1740 (2013)

[3] Wu, Han-Ching, and Shou-Hsuan Stephen Huang. Neural networks-based detection of stepping-stone intrusion. Expert Systems with Applications, 37 (2), 1431-1437 (2010)

[4] Lin, S.W., Ying, K.C., Lee, C.Y. and Lee, Z.J. An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection. Applied Soft Computing, 12 (10), 3285-3290 (2012)

[5] Javadzadeh, Ghazaleh, and Reza Azmi. IDuFG: Introducing an Intrusion Detection using Hybrid Fuzzy Genetic Approach. International Journal of Network Security, 17(6), 754-70 (2015)

[6] Sonawane, Sandip, Pardeshi, Shailendra and Prasad, Ganesh. A survey on intrusion detection techniques World Journal of Science and Technology, 2(3), 127-133 (2012)

[7] Morin, Benjamin, Ludovic Mé, Hervé Debar, and Mireille Ducassé. A logic-based model to support alert correlation in intrusion detection." Information Fusion, 10 (4), 285-299 (2009)

[8] Garcia-Teodoro, Pedro, J. Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. computers & security, 28(1-2), 18-28 (2009)

[9] Kumar, Vipin, Jaideep Srivastava, and Aleksandar Lazarevic, eds. Managing cyber threats: issues, approaches, and challenges, Vol. 5, Springer Science & Business Media, (2006).

[10] Tang, Hua, and Zhuolin Cao. Machine learning-based intrusion detection algorithms. Journal of Computational Information Systems, 5 (6), 1825-1831 (2009)

[11] Ojugo, A. A., A. O. Eboka, O. E. Okonta, R. E. Yoro, and F. O. Aghware. Genetic algorithm rule-based intrusion detection system (GAIDS). Journal of Emerging Trends in Computing and Information Sciences, 3 (8), 1182-1194 (2012)

[12] Zhao, Jiu-Ling, Jiu-Fen Zhao, and Jian-Jun Li. Intrusion detection based on clustering genetic algorithm. Proc. IEEE 2005 International Conference on Machine Learning and Cybernetics, 6, 3911-3914 (2005)

[13] Khan, M. Sadiq Ali, Rule based network intrusion detection using genetic algorithm. International Journal of Computer Applications, 18 (8), 26-29 (2011)

[14] Rajdeep Borgohain, FuGeIDS: Fuzzy Genetic paradigms in Intrusion Detection Systems,

International Journal of Advanced Networking and Applications, 3 (6), 1409-1415 (2012)

[15] Jalil, Kamarularifin Abd, Muhammad Hilmi Kamarudin, and Mohamad Noorman Masrek, Comparison of machine learning algorithms performance in detecting network intrusion, Proc. IEEE 2010 international conference on networking and information technology, 221-226 (2010)

[16] Karatas, Gozde, and Ozgur Koray Sahingoz. "Neural network based intrusion detection systems with different training functions." In IEEE 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1-6 (2018).

[17] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, accessed on march 2020.

[18] Kotpalliwar, Manjiri V., and Rakhi Wajgi. "Classification of Attacks Using Support Vector Machine (SVM) on KDDCUP'99 IDS Database." In IEEE 2015 Fifth International Conference on Communication Systems and Network Technologies, pp. 987-990 (2015).

[19] Kong, Lingjing, Guowei Huang, and Keke Wu. "Identification of abnormal network traffic using support vector machine." In IEEE 2017 18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), pp. 288-292 (2017)

[20] Tao, Peiying, Zhe Sun, and Zhixin Sun. "An improved intrusion detection algorithm based on GA and SVM." Ieee Access 6, pp. 13624-13631 (2018)

[21] Midzic, A., Z. Avdagic, and S. Omanovic. "Intrusion detection system modeling based on neural networks and fuzzy logic." In 2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES), pp. 189-194 (2016)

[22] Wang, Yunpeng, Yuzhou Li, Daxin Tian, Congyu Wang, Wenyang Wang, Rong Hui, Peng Guo, and Haijun Zhang. "A novel intrusion detection system based on advanced naive Bayesian classification." In International Conference on 5G for Future Wireless Networks, pp. 581-588. Springer, Cham, (2017)

[23] Ingre, Bhupendra, Anamika Yadav, and Atul Kumar Soni. "Decision tree based intrusion detection system for NSL-KDD dataset." In International Conference on Information and Communication Technology for Intelligent Systems, pp. 207-218. Springer, Cham, (2017)

[24] NSL-KDD dataset. https://www.unb.ca/cic/datasets/nsl.html. Accessed 15 August 2020.

[25] Ashfaq, Rana Aamir Raza, Xi-Zhao Wang, Joshua Zhexue Huang, Haider Abbas, and Yu-Lin He. Fuzziness based semi-supervised learning approach for intrusion detection system, Information Sciences, 378, pp. 484-497, Feb. 2017.

[26] Javaid, Ahmad, Quamar Niyaz, Weiqing Sun, and Mansoor Alam, A deep learning approach for network intrusion detection system. Proc. 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), pp. 21-26, May 2016.