DOI: https://doi.org/10.24113/ijoscience.v4i8.153

A Review on Intrusion Detection System using Artificial Intelligence Approach

Apoorva Deshpande
P.G. Student
Department of Computer Science and Engineering
MPCT
Gwalior, India
deshpande.apoorva@rocketmail.com

Ramnaresh Sharma
Associate Professor
Department of Computer Science and Engineering
MPCT
Gwalior, India

***Abstract*—Today, intrusion detection system using the neural network is an interested and considerable area for the research community. The computational intelligence systems are defined on the basis of the following parameters: fault tolerance and adaptation; adaptable the requirements of make a better intrusion detection model. In this paper, provide an overview of the research progress using computational intelligence to the problem of intrusion detection. The goal of this paper summarized and compared research contributions of Intrusion detection system using computational intelligence and neural network, define existing research challenges and anticipated solution of machine learning. Research showed that application of machine learning techniques in intrusion detection could achieve high detection rate. Machine learning and classification algorithms help to design "Intrusion Detection Models" which can classify the network traffic into intrusive or normal traffic. This paper discusses some commonly used machine learning techniques in Intrusion Detection System and also reviews some of the existing machine learning IDS proposed by researchers at different times.**

***Keywords*— Intrusion Detection System; Anomaly Detection; Supervised learning; Unsupervised; Detection Rate;**

I. INTRODUCTION

Nowadays, protection of network and system via malware and attacks is a very challenging task. The requirement of intrusion detection system depends upon the intrusion recognition, data collection and preprocessing. Intrusion recognition is most vital, among these. Observe data and detection model are compared and identify the patterns of intrusion behavior either its successful or unsuccessful. At the earlier age of intrusion detection system e focused only ho to implement the intrusion detection system, according to Denning [1] observed this major research identification in 1987. During 1980s to1990s, a hybrid expert system and statistical approaches were very much famous. Detection models were generated from the domain knowledge of security experts. Set of training data discovery using artificial intelligence and machine learning techniques. Generally, we adapt the methods for a set of training data is data classification, rule-based induction, and data-clustering. In intrusion detection problems, data are not trivial when the process of automatically constructing models [2,3].

There are several challenges during intrusion detection system such as distinguishes criteria for normal and abnormal behavior, dynamically update the process of learning, huge network traffic. Thus, at that criteria machine learning and artificial intelligence unable to handle the solution so computational intelligence play measure role into this significance.

Intrusion Detection System (IDS)s are security tools that detect intrusions to a network or a host computer. An IDS is either host based or network based. A host based IDS detects attacks on a host computer, whereas, a network based IDS, also called Network Intrusion Detection System (NIDS), detects intrusions into a network by analyzing network traffic and are generally installed in network gateway or server. Host based intrusion detection systems can be divided into four types, namely (a) File System Monitors, (b) Log file analyzers, (c) Connection analyzers, (d) Kernel-based IDS [4-10].

Furthermore, based on the data analyzing technique there are principally 2 classes of IDSs, signature-based and anomaly primarily based. A signature-based system detects attacks by analyzing network data for attack signatures hold on in its database. this kind of IDS detects previously best-known attacks, whose signatures are stored in its database. On the other hand, an anomaly-based IDS appearance for deviations from traditional behavior of the subject. Anomaly-based systems are capable of detecting novel attacks [11,12].

Machine learning techniques can be effective for detecting intrusions. Many Intrusion Detection Systems are modeled based on machine learning techniques. Learning algorithms are designed either on offline dataset or real data collected from university or organizational networks. Usually machine learning techniques is classified into two classes i.e. supervised Learning and unsupervised Learning [14-18].

In supervised learning the training dataset is instantly accessible together with its target vector. The learner learns from available data taking guidance of the output vector. In contrast to supervised learning, unsupervised learning systems learn from its atmosphere. Systems learn from coaching knowledge; however there's no target vector accessible. Some usually used machine learning techniques within the field of intrusion detection are like Artificial Neural Network (ANN), decision Tree, Support Vector Machine, Bayesian Classification, Self-organizing Map, etc [12-18].

II. Related Work

Meng et al. [8] compared ANN, SVM and DT schemes for anomaly detection in a uniform environment and concluded that J48 algorithm of DT gives better performance than the other two schemes. The detection rate of low frequent attack types (U2R, R2L) was also high. Feng et al. [9] introduced a new classification technique and utilized the advantages of SVM and Clustering based on Self-Organized Ant Colony Network.

Sumaiya Thaseen Ikram et al. [9] proposed an intrusion detection model using chi-square feature selection and multi class support vector machine (SVM). A parameter tuning technique is adopted for optimization of Radial Basis Function kernel parameter namely gamma represented by '!' and over fitting constant 'C'. These are the two important parameters required for the SVM model. The main idea behind this model is to construct a multi class SVM which has not been adopted for IDS so far to decrease the training and testing time and increase the individual classification accuracy of the network attacks.

Manjula et al. [10] proposed a classification and predictive models for intrusion detection which is built by using machine learning classification algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest. An experimental result shows that Random Forest Classifier out performs the other methods in identifying whether the data traffic is normal or an attack.

Saad Mohamed et al. [11] presented a hybrid approach to anomaly detection using of K-means clustering and Sequential Minimal Optimization (SMO) classification.

To reduce the false alarm rate of anomaly-based IDS, many machine learning techniques, including support vector machine (SVM) Feng et al. [13] applied extreme learning machine (ELM) along with models combining several techniques. Each model offers specific strengths and weaknesses, with overall generic detection rates steadily increasing. SVMs exhibit good detection performance with IDSs in terms of classifying the flow of a network into normal or abnormal behaviors.

Kuang et al. [15] proposed an IDS based on a combination of the SVM model with kernel principal component analysis (KPCA) and genetic algorithm (GA). KPCA was used to reduce the dimensions of feature vectors, whereas GA was employed to optimize the SVM parameters. The average detection rate was 95.26%, whereas the average false alarm rate was 1.03%. ELMs exhibit performance comparable with that of SVMs in terms of classifying instances of IDS.

Gogoi, Bhattacharyya et al. [16] proposed a multi-level hybrid IDS using a combination of supervised, unsupervised, and outlier methods. This system was evaluated with three datasets, namely, real-time flow dataset, DDoS dataset, and the KDD Cup 1999 with NSL-KDD datasets. The system performance was good with a false alarm rate of 3.4% with the corrected KDD Cup 1999 dataset.

Wathiq Laftah Al-Yaseen et al. [17] proposes a multi-level hybrid intrusion detection model that uses support vector machine and extreme learning machine to improve the efficiency of detecting known and unknown attacks. A modified K-means algorithm is also proposed to build a high-quality training dataset that contributes significantly to improving the performance of classifiers. The modified K-means is used to build new small training datasets representing the entire original training dataset, significantly reduce the training time of classifiers, and improve the performance of intrusion detection system. The popular KDD Cup 1999 dataset is used to evaluate the proposed model. Compared with other methods based on the same dataset, the proposed model shows high efficiency in attack detection, and its accuracy (95.75%) is the best performance thus far.

III. Existing System & Problem Identification

Traditional IDS/IPS techniques such as signature based detection, anomaly detection, artificial intelligence (AI) based detection etc.

Signature based intrusion detection attempts to define a set of rules or signatures or predefined knowledge base that can be used to decide that a given pattern is that of an intruder. As a result, signature based systems are capable of attaining high levels of accuracy and minimal number of false positives in identifying even very subtle intrusions. Little variation in known attacks may also affect the analysis if a detection system is not properly configured. Therefore, signature based detection is an efficient solution for detecting known attacks but fails to detect unknown attacks or variation of known attacks. One of the motivating reasons to use signature based detection is ease in maintaining and updating preconfigured rules. These signatures are composed by several elements that identify the traffic. For example, in SNORT the parts of a signature are the header

(e.g. source address, destination address, ports) and its options (e.g. payload, metadata), which are used to determine whether or not the network traffic corresponds to a known signature. Signature based intrusion detection technique can be used to detect known attack.

Anomaly (or behavioral) detection is concerned with identifying events that appear to be anomalous with respect to normal system behavior. A wide variety of techniques including data mining, statistical modeling and hidden markov models have been explored as different ways to approach the anomaly detection problem. Anomaly based approach involves the collection of data relating to the behavior of legitimate users over a period of time, and then apply statistical tests to the observed behaviour, which determines whether that behaviour is legitimate or not. It has the advantage of detecting attacks which have not been found previously. The key element for using this approach efficiently is to generate rules in such a way that it can lower the false alarm rate for unknown as well as known attacks. Anomaly detection techniques can be used to detect unknown attacks at different levels. The ability of soft computing techniques to deal with uncertain and partially true data makes them attractive to be applied in intrusion detection. There are many soft computing techniques such as Artificial Neural Network (ANN), Fuzzy logic, Association rule mining, Support Vector Machine (SVM), Genetic Algorithm (GA) etc. used to improve detection accuracy and efficiency of signature based IDS or anomaly detection based IDS.

An Intrusion Detection System is designed to detect an intrusion while it is in progress, or after it has occurred. That means properly configuring their intrusion detection systems to recognize what normal traffic on their network looks like compared to potentially malicious activity. An intrusion detection system (IDS) is a device or software application that monitors a network or systems for malicious activity or policy violations. Any malicious activity or violation is typically reported either to an administrator or collected centrally using a security information and event management (SIEM) system. A SIEM system combines outputs from multiple sources, and uses alarm filtering techniques to distinguish malicious activity from false alarms.

So far, IDs have poor perfection. There are so many shortcomings associated with IDS. As these SDIs evolve, many gaps are continually corrected by improving and refining existing techniques, but some of them are inherent in the way in which SDIs are built.

The following are the most common shortcomings[19]:

Lack of Efficiency: IDS are necessary to evaluate activities in real time. It is very difficult to manage a large number of events, as is the case with today's networks. As a result, HIDS generally slows down a system in which NIDS release network packets for which processing time is insufficient.

High Number of False Positives: Most IDS detects attacks on the enterprise by analyzing information from a single host, application, or network interface at many points in the network. False alarms are high and attack detection is not perfect. Lowering the thresholds to reduce false alarms increases the number of attacks not detected as false negatives. Improving the ability to accurately identify attacks is the main problem faced by IDS manufacturers today.

Burdensome Maintenance: There is a specific knowledge of the requirements and a considerable effort to configure and maintain IDS. Here are some examples, such as detection of abuses, which typically use expert system shells for their implementation that encode and map signatures using rule sets. The updating of these rule sets includes details specific to the expert system and its language for the expression of rule sets and can only allow an indirect specification of sequential relations between events. These considerations also apply to the addition of a statistical measure generally used to find unusual variations in nature.

Limited Flexibility: When an intrusion detection system is designed for a typical specific environment, it can be difficult to use in other environments, whether it is the same concerns or policies themselves. The mechanism used for detection can also be difficult to adapt to different usage patterns. The process of identifying personalization, particularly for the system with some yews and chains replacing them with time-enhanced detection techniques, is also problematic in many IDS implementations. Often, the IDS must be restarted completely for the changes and additions to take effect.

High Speed Communications: the increase in communication speed also changes the processing speed because the communication speed is directly proportional to the processing speed. It is therefore necessary to analyze the contents of the communication packages. This is where packet loss can occur. With NIDS, to observe different communication flows at the same time, the difficulty increases when communication is changed, i.e. Conventional communication is used instead of broadcast communication. So far, we have discussed some of the existing approaches which are incorporating IDS. However, there is no universal efficient solution found yet. Each has

some limitations. Here we summarize some common pros and cons of existing techniques of IDS:

Requires more training time and samples for detection accuracy.

Cannot be used for all types of attacks.

Computational overhead high.

IV. Machine Learning Approach

Artificial Neural Network (ANN)s are the computational models of neural structure of human brain. Neurons are the basic building blocks of human brain. An ANN is a layered network of artificial neurons. An ANN may consist of an input layer, one or more hidden layer(s) and an output layer. The artificial neurons of one layer are fully or partially connected to the artificial neurons of the next layer. Each of these connections is associated with a weight, and feedback connections to the previous layers are also possible [2].

Decision tree is one of the simplest machine learning techniques. A decision tree can easily be represented as a set of if-then rules. The classification starts from root node, traversing down the tree till the suitable leaf node. Each node of the tree represents a solution. Each node tests on an attribute of the instance and each descending branch of that node corresponds to one of the values of that attribute. Starting from the root node, each node tests the attribute specified by that node and moves down the tree through the branch matching the value, till it reaches a leaf node [10].

Support Vector Machine (SVM) maps the input vector into a higher dimensional feature space. It is a binary classification technique that classifies input instances into two classes. Only the Support Vectors determine the optimal separating hyper-plane to classify input instance into one of the two classes. Support Vectors are the points closest to the separating hyper-plane. During classification, mapped input vectors placed on one side of the separating hyper-plane in the feature space falls into one class and placed on the other side of the plane falls into the other class. In case the data points are not linearly separable, SVM uses suitable kernel function to map them into higher dimensional space, so that, in that higher dimensional space they become separable [9].

Bayesian learning is a statistical learning method based on probabilities of hypotheses. A prior probability is assigned to each candidate hypothesis based on prior knowledge. Training examples may increase or decrease the probability of a hypothesis to be correct. This probability can be calculated using Bayes' theorem. Classification is done by combining the predictions of multiple hypotheses, weighted by their probabilities. These probabilities in Bayesian method could be calculated using Bayes' theorem. Requirement of initial knowledge of many probabilities make practical application of Bayesian methods difficult [10].

V. Proposed Model

The biggest challenge for today is to protect the users from Intrusion due to wide use of internet. Intrusion Detection Systems (IDS) are one of the security tools available to detect possible intrusions in a Network or in a Host. Research showed that application of machine learning techniques in intrusion detection could achieve high accuracy rate as well as low false alarm rate. Accurate predictive models can be built for large data sets using supervised machine learning techniques, that is not possible by traditional methods. IDS learns the patterns by the training data, so it can detect only the known attack, new attacks cannot be identified. This research work is based on designing an optimized feature based classifier and performing analysis on three different datasets. This section describes the proposed hybrid model for intrusion detection. The KDD-99 dataset is used as a benchmark to evaluate the performance of the proposed model.

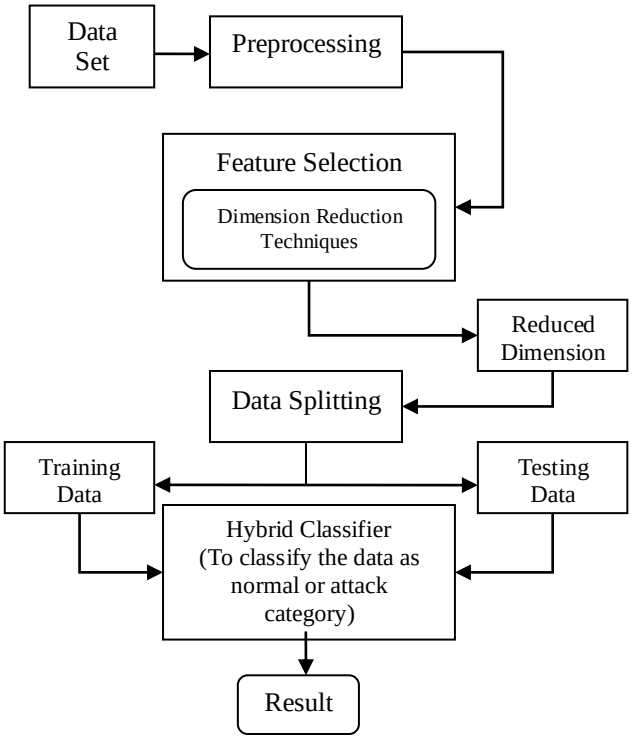

Figure 1: Proposed Flow Diagram of Intrusion Detection System

The algorithm flow of the proposed method is described as follows:

Following steps will be used to build the proposed model for intrusion detection:

Step 1: Convert the symbolic attributes protocol, service, and flag to numerical.

Step 2: Normalize data to [0,1].

Step 3: Separate the instances of dataset into two categories: Normal, DOS, R2L, U2R and Probe.

Step 4: Feature Reduction and Extraction.

Step 5: The data set is divided as training data and testing data.

Step 6: Train classifier with these new training datasets.
Step 7: Test model with dataset.
Step 8: Finally computing and comparing Accuracy and FAR for different classifiers.
The proposed algorithm flow diagram of intrusion detection model is illustrated in figure 1.

VI. IDS TERMINOLOGIES

To evaluate the proposed algorithm, it is concentrated on three indications of performance: detection rate, accuracy and False Alarm Rate (FAR).
If one sample is an anomaly and the predicted label also stands anomaly, then it is called as true positive (TP).
If one sample is an anomaly, but the predicted label stands normal, then it is called as false negative (FN).
If one sample is a normal and the predicted label also stands normal, then it is true negative (TN).
If one sample is normal, but the predicted label stands anomaly, then it is termed as false positive (FP).
TP stands the number of true positive samples, FN stands the number of false negative samples, FP stands the number of false positive samples, and TN stands the number of true negatives.
To evaluate the proposed algorithm, it is concentrated on three indications of performance: detection rate, accuracy and False Alarm Rate (FAR).
If one sample is an anomaly and the predicted label also stands anomaly, then it is called as true positive (TP).
If one sample is an anomaly, but the predicted label stands normal, then it is called as false negative (FN).
If one sample is a normal and the predicted label also stands normal, then it is true negative (TN).
If one sample is normal, but the predicted label stands anomaly, then it is termed as false positive (FP).
TP stands the number of true positive samples, FN stands the number of false negative samples, FP stands the number of false positive samples, and TN stands the number of true negatives.
From equation (5.1) and (5.5), the accuracy, detection rate, False Positive rate (FPR), False Negative Rate (FNR) and False Alarm rate (FAR) is calculated.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)*100 \quad (1)$$

$$\text{Detection Rate} = TP/(TP+FN)*100 \quad (2)$$

$$\text{False Negative Rate (FNR)} = FN/(FN+TP)*100 \quad (3)$$

$$\text{False Positive Rate (FPR)} = FP/(FP+TN)*100 \quad (4)$$

$$\text{False Alarm Rate (FAR)} = (FPR+FNR)/2 \quad (5)$$

VII. CONCLUSION

In modern society, the security of computer networks becomes an increasingly vital issue to be solved. Traditional intrusion detection techniques lack extensibility in face of changing network as well as adaptability in face of unknown attack type. Machine learning techniques are proved to be efficient for intrusion detection. High accuracy in intrusion detection can be achieved using machine learning techniques even though the detection accuracy depends on some other factors too. Some of them are selection of correct feature set, selection of appropriate training and testing data, etc. With the selection of the appropriate attributes for these factors, a higher performance could be achieved.

REFERENCES

[1] Garcia-Teodoro, P., "Anomaly-based network intrusion detection: techniques", systems and challenges. Comput. Security Vol. 28.Issue, pp. 18–28, 2009.
[2] Sufyan T Faraj Al-Janabi, Hadeel Amjed Saeed, "A neural network based anomaly intrusion detection system", IEEE, 2011.
[3] J. Ryan, M. Lin, and R. Miikkulainen, "Intrusion Detection with Neural Networks," Conference in Neural Information Processing Systems, 943–949.
[4] A. K. Ghosh and A. Schwartzbard, "A Study in Using Neural Networks for Anomaly and Misuse Detection," Conference on USENIX Security Symposium, Volume 8, pp. 12–12, 1999.
[5] P. L. Nur, A. N. Zincir-heywood, and M. I. Heywood, "Host-Based Intrusion Detection Using Self-Organizing Maps," in Proceedings of the IEEE International Joint Conference on Neural Networks, pp. 1714–1719, 2002.
[6] K. Labib and R. Vemuri, "NSOM: A Real-Time Network-Based Intrusion Detection System Using Self-Organizing Maps," 2000.
[7] Sharma, R.K., Kalita, H.K., Issac, B., "Different firewall techniques: a survey", International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 2014.
[8] Meng, Y.-X., "The practice on using machine learning for network anomaly intrusion detection", International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 2, IEEE, 2011.
[9] Sumaiya Thaseen Ikram, Aswani Kumar Cherukuri, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM", Journal of King Saud University –Computer and Information Sciences, 2016.
[10] Manjula C. Belavagi and Balachandra Muniyal, "Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection, Procedia Computer Science", Elsevier, 2016.
[11] Saad Mohamed Ali Mohamed Gadal and Rania A. Mokhtar, "Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique", International Conference on Communication, Control, Computing and Electronics Engineering, IEEE, 2017.
[12] Ibrahim, H. E., Badr, S. M., & Shaheen, M. A., "Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems", International Journal of Computer Applications, Vol. 56, Issue 7, pp. 10–16, 2012.
[13] Wen Feng, Qinglei Zhang, Gongzhu Hu, Jimmy Xiang Huang, "Mining network data for intrusion detection through combining SVMs with ant colony networks", Elsevier, Vol 37, pp 127-140, 2014.
[14] Shi-JinnHorng, Ming-Yang Su, Yuan-Hsin Chen, Tzong-Wann Kao, Rong-Jian Chen, Jui-Lin Lai, Citra Dwi Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines" Expert Systems with Applications, Elsevier, Vol. 38, pp. 306–313, 2011.
[15] Kuang, F., Xu, W., & Zhang, S., "A novel hybrid KPCA and SVM with GA model for intrusion detection", Applied Soft Computing Journal, Vol. 18, pp. 178–184, 2014.

[16] Prasanta Gogoi, D.K. Bhattacharyya, B. Borah1 and Juga, K. Kalita, "MLH-IDS: A Multi-Level Hybrid Intrusion Detection Method", The Computer Journal, Vol. 57 Issue 4, pp. 602-623, 2014.

[17] Wathiq Laftah Al-Yaseen , Zulaiha Ali Othman ,Mohd Zakree Ahmad Nazri, "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", International Journal in Expert Systems With Applications, Elsevier, 2017.

[18] He, L., "An improved intrusion detection based on neural network and fuzzy algorithm. Journal of Networks, Vol. 9, Issue 5, pp. 1274–1280, 2014.

[19] Hoque, M. S., Mukit, M. A. , & Bikas, M. A. N., "An implementation of intrusion detection system using genetic algorithm", International Journal of Network Security & Its Applications, Vol 4, Issue 2, pp. 109–120, 2012.