Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 12211 - 12219
Received 25 April 2021; Accepted 08 May 2021.

Machine Learning Algorithms in Intrusion Detection and Classification

GunaSekhar Sajja[1], Malik Mustafa[2],Dr R Ponnusamy[3],Shokhjakhon Abdufattokhov[4],Murugesan G[5], Dr. P Prabhu[6]

[1] Research Scholar, University of the Cumberlands
[2] Center for foundation Studies, Gulf College
[3] Center for Artificial Intelligence & Research, Chennai Institute of Technology, Kundrathur
[4]Control and Computer Engineering, Turin Polytechnic University in Tashkent, Uzbekistan
[5]St. Joseph's College of Engineering, Chennai , Tamilnadu, India
[6]Department of Information Technology, Directorate of Distance Education
Alagappa University, Karaikudi

Abstract
The complexities that information systems face are quickly growing. Threats and attacks are framed and executed using new methods that exploit the information contained in networks. When going across subtle domains, knowledge is constantly changing due to the different categories of users, server managers, and those that need to access it. Information device protection is critical against threats such as denial of service attacks and intrusions. Intrusion is a big threat to unauthorized data or lawful network leveraging valid users' identities or any of the network's back doors and vulnerabilities. Intrusion Detection Systems are mechanisms designed to detect intrusions at different stages (IDS). The aim of this research is to improve the efficiency of intrusion detection systems (IDS) by using rule-based techniques and learning-based algorithms for intrusion detection and classification. Neural Networks (NN) , Random Forest and SVM algorithms.The output of rule-based techniques and machine learning algorithms is evaluated using regular datasets such as kddcup 99.
Keywords: Intrusion Detection, Machine Learning, Security, Privacy, KDD99, Attacks

1. Introduction

The Intrusion Detection System (IDS) is a critical component of network and data protection. Because of the rapid evolution of network technology, identification of attacks based on contextual knowledge processing can be unique to particular apps and networks. Such a challenge can be solved with the aid of a hybrid intrusion detection system (IDS). [1]
DoS attacks are typically focused on packet flooding with the aim of overburdening the victim's infrastructure. These attacks are now capable of disrupting networks of almost any scale. One of the major testing obstacles for developing high-performance hybrid IDS is dealing with huge volumes of records with a large number of features. A large number of features can make it difficult to identify malicious patterns, resulting in a long training and testing process, increased resource demand, and a low detection rate.[2]
Computer security is characterized as the defense of computing systems from threats in order to preserve resource confidentiality, integrity, and availability [3]. An intrusion is described as any series of acts that attempt to compromise network resources and the victim server [4]. The Intrusion Detection System (IDS) is primarily used for tracking incidents that occur in computer systems/networks, analyzing data, identifying, preventing, or reporting to the system administrator so that appropriate action can be taken. The increase in the number of attacks

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 12211 - 12219
Received 25 April 2021; Accepted 08 May 2021.

launched by attackers has increased users' skepticism about the Internet. Denial of Service is an effective security assault (DoS).

An intrusion detection system (IDS) is a monitoring system that tracks computer networks and network traffic and analyzes it for potential aggressive attacks from outside the organization as well as system abuse or attacks from inside the organization. In layman's words, an intrusion detection device is similar to a burglar detector. A car's lock system, for example, prevents it from burglary.

However, if anyone cracks the lock mechanism and attempts to rob the vehicle, the burglar detector senses the broken lock and alarms the owner by raising an alarm sound. Similarly, IDS will function as an alert in a system/network to detect incidents and notify if any malicious behavior occurs.

Attackers [5] continue to devise new ways to hack the host/network and conduct illegal operations. The Internet's scale and sophistication, as well as the operating systems on end hosts, make it more vulnerable to vulnerabilities. Because of these problems, existing Internet best practices depend on evidence of detecting attacking trends, monitoring security vulnerabilities, and closing them as soon as possible. Existing intrusion detection systems are seeing an increase in false alarms. Computational Intelligence (CI) components in IDS can be streamlined to minimize these. Many CI strategies were implemented by the researchers, and their accuracy was also measured using benchmark datasets.

The Intrusion Detection System (IDS) is a multicolored technique that inspects both inbound and outbound network traffic, detects unusual patterns, and discards them. IDS are made up of three major components: a data base, an analysis engine, and a response manager [6].

The primary component of any IDS, also known as an event driver, is the data base. Host-based monitors, network-based monitors, application-based monitors, and target-based monitors are the four types of data sources.

The research engine is the second part of an intrusion detection system. This component collects data from the data source and checks it for signs of attacks or other regulation breaches. Misuse/Signature-based detection and Anomaly/Statistical detection are the two most commonly used methods for IDS analysis.

The solution manager is the third aspect of an intrusion detection system. In general, the response manager can function only when potential inaccuracies in the mechanism are discovered, warning someone or something in the form of a response.

DoS is a form of attack in which unnecessary messages are created and directed at the victim resource, causing the server to be unable to provide service to legitimate users[7]. It results in unreliable and inaccessible networks, as well as disruptions in network traffic and connection interfaces.

2. Related Work

Machine learning (ML) [8] is a new branch of data mining that lets a computer program to become more accurate at predicting events without being explicitly programmed. These ML techniques are frequently divided into two types: supervised and unsupervised learning techniques. Supervised learning techniques use labeled training data for inference (classification, regression), whereas unsupervised learning techniques use unlabeled data to identify hidden existing patterns (clustering).

Classification is the process of translating an input collection of examples P into a unique collection of qualities Q, often known as target attributes or labels. In a number of applications,

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 12211 - 12219
Received 25 April 2021; Accepted 08 May 2021.

classification techniques such as decision tree classifiers, bayesian classifiers, and artificial neural networks, closest neighbor classifiers, random forest, and support vector machines are utilized [9]. We'll go through each one briefly. Each approach is built around the learning algorithm that it employs..

A decision tree is one of the most fundamental and straightforward classifiers used to address classification problems. A decision tree is a graph that categorizes events by sorting them based on their feature values. The decision tree is made up of nodes and branches, with each node indicating a classification instance and each branch indicating a possible value for the node. In decision, instance categorization begins at the root node and instance sorting is performed using feature values.

In certain cases, predicting the class label for a given set of input attributes might be difficult. Furthermore, class variables are non-deterministic even when the stated input attributes set values match some of the characteristics in the training data set. This is owing to the presence of certain noisy data and confusing aspects that are not taken into account during analysis. For example, predicting the likelihood of heart disease in a certain person based on that person's daily activities.

In this case, it is likely that the majority of individuals who eat healthy meals and exercise on a regular basis are at risk of developing heart disease due to other factors such as smoking, alcohol use, and maybe hereditary. In such cases, the classification model is built on well-known heart disease features, which cannot provide accurate information. In such cases, describing probabilistic connections between the attribute collection and the class label is required, and the Bayesian classifier is all about justifying such tasks [10].

An artificial neural network (ANN) is modeled after biological neural networks, which are used to construct animal brains. Because it is made up of connected nodes and directed interconnections, ANN is also known as a connectionist system. Each connected connection is given a weight and is in charge of conveying a signal from one node to the next. When a node receives a signal, it processes it before forwarding it to another node.

In most ANN implementations, the signal at the connection between artificial neurons is a real number, and the output of each neuron is governed by a non-linear function of the sum of all its inputs. As learning progresses, the signal intensity grows or falls according to the weights of artificial neurons and the connections between them [11].

In ML classification, there are two techniques to developing a learning model. One of them is that the model starts learning as soon as the training set is available; these models are referred to as eager learners. Another model observes all training examples but only achieves classification if the attributes of the test case precisely match any of the training cases. Such students are referred to as lazy students [12].

Each sample is treated as a data point in a d-dimensional space by the Nearest Neighbour (NN) classifier, where d is the number of features. The distance between the specified test example and all data points in the training set is calculated. The k-Nearest Neighbors of data point X are the k points closest to the X.

The data point is subsequently classified based on the class labels assigned to its neighbors. If a data point has more than one class labeled neighbor, the class label with the greatest number of class labels is given to the data point. It is necessary to determine the exact value of k's nearest neighbors. If k is set too low, it may misclassify due to noise in the training data. On the other hand, if the value of k is too large, there is a danger of misclassification since the collection of

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 12211 - 12219
Received 25 April 2021; Accepted 08 May 2021.

nearest neighbors may comprise data points located far away from the neighborhood of the test attribute.

To begin, random forest is a supervised machine learning methodology that consists of a forest of judgements generated by multiple decision trees generated using random vectors. This approach may be used to address both classification and regression issues. The outcome of the random forest is linked to the number of trees it combines in the forest in such a way that as the number of trees in the forest increases, so does the possibility of achieving greater accuracy. It is crucial to recognize that creating a forest is not the same as creating decision trees [12].

The primary distinction between decision trees and random forests is that finding the root node and dividing the feature nodes occurs at random in random forest classification. Because of its benefits, random forest classification is often used. One of these is that it may be used for classification as well as regression. Another advantage of this method is that if a significant number of trees are available, it avoids the problem of overfitting. In addition, a random forest classifier can manage missing data and can be modelled in the case of categorical data.

Random forest classifiers are utilized in many fields, including medicine, finance, e-commerce, and the stock market. Random classifiers are used in banking to distinguish between loyal and fraudulent customers. Random Forest is used in medicine to find the best combination of drugs and to detect sickness based on a patient's previous medical history. Random Forest classier is used in the stock market to monitor a stock's behavior and subsequently detect loss and profit. In the context of e-commerce, Random Forest may be used to estimate user product suggestions.

The supervised learning model utilized for categorisation is the Support Vector Machine (SVM). It has sparked a lot of curiosity in the categorizing industry. In the SVM model, a visible gap separates instances of the various categories in vector space. When a new sample comes, it is mapped into the given vector space and its label is assigned to a category based on which side of the gap it is located [13]. Using the kernel approach, an SVM can successfully do non-linear classification.

Researchers in [14-18] also did investigations over applications of machine learning algorithms in different real world applications and tried to identify different security threats in various real world applications.

3. Methodology

KDD 99 data set [19] is used as input in the framework. This KDD data set is preprocessed to remove noise from the data and make data consistent. After preprocessing, a clean and consistent input data is available. Now this input data is provided to machine learning algorithms like- SVM, Neural Network and Random Forest. These algorithms perform classification of input data. Then this classification data works as training data for the prediction task. When new intrusion related data is entered in to framework then on the basis of learning data available in the classes, framework predicts the new testing data to be normal or abnormal. Machine learning algorithms accuracy and error rate are shown below in figure and figure 2.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 12211 - 12219
Received 25 April 2021; Accepted 08 May 2021.

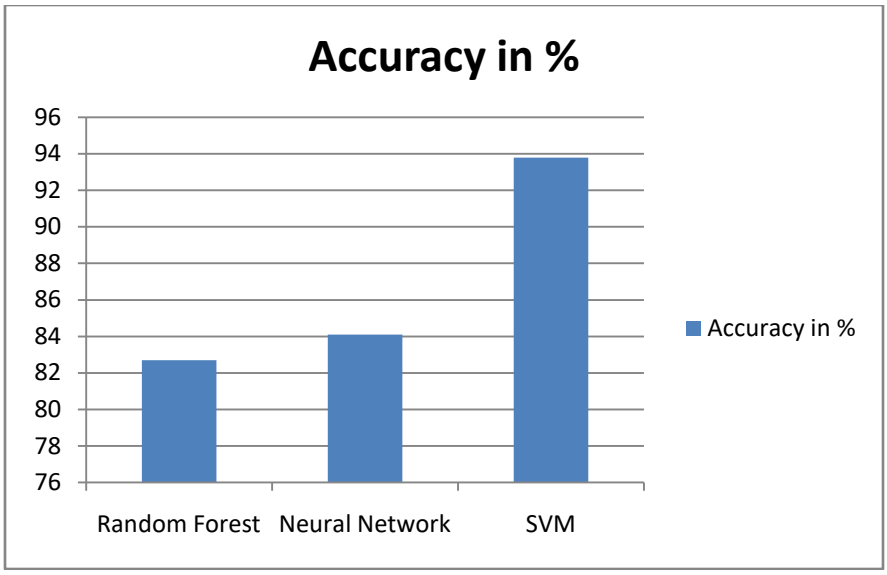

Fig.1 Accuracy Results of Classification Algorithms

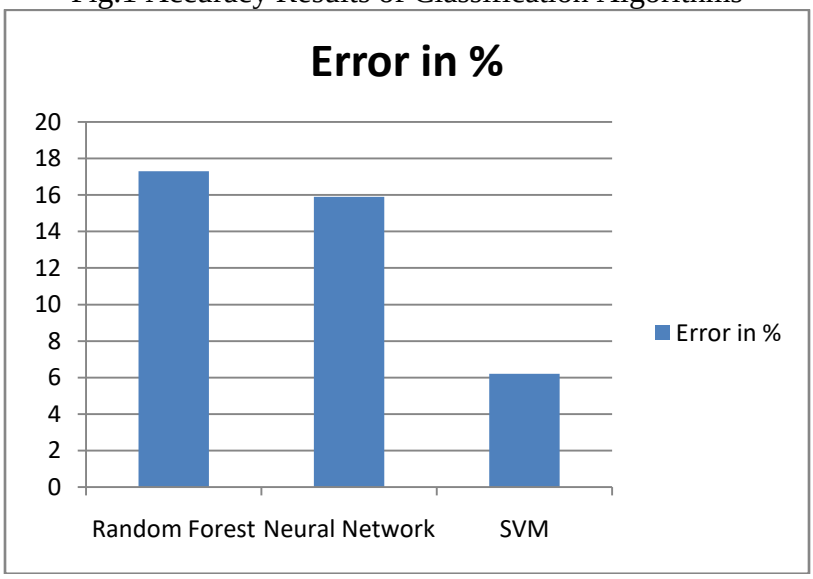

Fig.2 Error Rate Results of Classification Algorithms

4. Conclusion

The difficulties that information systems must deal with are rapidly increasing. Threats and assaults are constructed and carried out utilizing novel approaches that take use of the information stored in networks. When traversing subtle domains, knowledge is always changing owing to the many kinds of users, server admins, and those who require access to it. Protection of information devices is vital against threats such as denial of service attacks and invasions. Intrusion poses a significant risk to unauthorized data or a lawful network by utilizing legitimate users' identities or any of the network's back doors and weaknesses. Intrusion Detection Systems are techniques that detect invasions at various stages. KDD 99 data set is used as input. SVM algorithm has performed well. Accuracy of SVM is better than Random Forest and Neural Network.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 12211 - 12219
Received 25 April 2021; Accepted 08 May 2021.

References

[1] Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), Funchal, Portugal, 22–24 January 2018; pp. 108–116.

[2] Li, Z.; Rios, A.L.G.; Xu, G.; Trajkovi·c, L. "Machine Learning Techniques for Classifying Network Anomalies and Intrusions", IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 26–29 May 2019; pp. 1–5.

[3] Le, T.T.H.; Kim, Y.; Kim, H. Network Intrusion Detection Based on Novel Feature Selection Model and Various Recurrent Neural Networks. Appl. Sci. 2019, 9, 1392.

[4] Cordero, C.G.; Vasilomanolakis, E.; Wainakh, A.; Mhlhuser, M.; Nadjm-Tehrani, S. On generating network traffic datasets with synthetic attacks for intrusion detection.arXiv 2019, arXiv:1905.00304.

[5] Kabir, E.; Hu, J.; Wang, H.; Zhuo, G. A novel statistical technique for intrusion detection systems.Future Generation Computer. Syst. 2018, 79, 303–318.

[6] Hajisalem, V.; Babaie, S. A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection.Comput.Netw. 2018, 136, 37–50

[7] Hussain, J.; Lalmuanawma, S. Feature Analysis, Evaluation and Comparisons of Classification Algorithms Based on Noisy Intrusion Dataset. Procedia Computer. Sci. 2016, 92, 188–198.

[8]Garca, S.; Grill, M.; Stiborek, J.; Zunino, A.An empirical comparison of botnet detection methods.Computer Security. 2014, 45, 100–123.

[9] Divekar, A.; Parekh, M.; Savla, V.; Mishra, R.; Shirole, M.Benchmarking datasets for Anomaly-based Network IntrusionDetection: KDD CUP 99 alternatives. IEEE 3rd InternationalConference on Computing, Communication and Security(ICCCS), Kathmandu, Nepal, 25–27 October 2018; pp. 1–8.

[10] Belouch, M.; El Hadaj, S.; Idhammad, M. Performanceevaluation of intrusion detection based on machine learningusing Apache Spark. ProcediaComput. Sci. 2018, 127, 1–6.

[11] S. Abdufattokhov and B. Muhiddinov, "Probabilistic Approach for System Identification using Machine Learning," *2019 International Conference on Information Science and Communications Technologies (ICISCT)*, 2019, pp. 1-4, doi: 10.1109/ICISCT47635.2019.9012025

[12] S. Abdufattokhov and B. Muhiddinov, "Stochastic Approach for System Identification using Machine Learning," *2019 Dynamics of Systems, Mechanisms and Machines (Dynamics)*, 2019, pp. 1-4, doi: 10.1109/Dynamics47113.2019.8944452.

[13] Priyanka Sharma Dr. N.VasanthaGowri.,Dr. MohdNaved.,Bharat M N. (2021). An Internet of Things for prevention of security attack on cloud medical data using artificial intelligence. International Journal of Future Generation Communication and Networking, 14(1).

[14] Manne, Ravi, and Sneha C. Kantheti. 2021. "Application of Artificial Intelligence in Healthcare: Chances and Challenges". Current Journal of Applied Science and Technology 40 (6), 78-89. https://doi.org/10.9734/cjast/2021/v40i631320

[15] Malik Mustafa,&Omaima Ali Ahmed JalaAldein. (2020). Examining Perception of Malaysian autistic children social interaction for Virtual Reality (Version original).http://doi.org/10.5281/zenodo.4420802.

[16] Mustafa, M., 2021. The Technology of Mobile Banking and Its Impact on the Financial Growth during the Covid-19 Pandemic in the Gulf Region. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), pp.389-398.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 12211 - 12219
Received 25 April 2021; Accepted 08 May 2021.

[17] Mustafa, M., & Abbas, A. (2021). COMPARATIVE ANALYSIS OF GREEN ICT PRACTICES AMONG PALESTINIAN AND MALAYSIAN IN SME FOOD ENTERPRISES DURING COVID-19 PANDEMIC.PalArch's Journal of Archaeology of Egypt/Egyptology, 18(4), 254-264.

[18] Mustafa, M., &Alzubi, S. (2020). Factors Affecting the Success of Internet of Things for Enhancing Quality and Efficiency Implementation in Hospitals Sector in Jordan During the Crises of Covid-19. In Internet of Medical Things for Smart Healthcare (pp. 107-140).Springer, Singapore.

[19] KDD Cup (1999). University of California, Irvine (UCI). Available at: http://kdd.ics.uci.edu/databases/kddcup99/