MACHINE LEARNING TECHNIQUES FOR INTRUSION DETECTION

Ajeesha M I[1], Dr. D Francis Xavier Christopher[2]

[1]Research scholar, School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Coimbatore, Tamilnadu, India.

[2]Director, School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Coimbatore, Tamilnadu, India.

ABSTRACT

In cyber security, intrusion detection is the act of detecting malicious attacks. Unauthorized users would never gain access to the system. The computer security is to limit the access to a computer system. An Intrusion detection system (IDS) is a software that monitors a single or a network of computers from malicious activities that steals the system information. IDS can distinguish between legitimate and illegitimate traffic and can able to signals attacks in real time, before malicious attacks occur. In this paper we describe a framework for building intrusion detection (ID) models. Machine learning algorithms are used for detecting attacks and helps the users to develop secure information systems.

KEYWORDS: Machine learning, Intrusion detection system, network security.

INTRODUCTION

As computers have become more pervasive and liaison, their security has become a major concern. The internet has developed tremendously in recent decades. The correlation of computers and network has made the cyberspace more complex. Conventional security software requires a lot of human effort to analyse threats and to obtain the characteristics of threats. Manual analysis to extract relevant threat information is also a time consuming and error-prone process. This process can be more efficient by applying machine learning algorithms. Data mining techniques and machine learning are a new approach for intrusion detection.

Intrusion detection system(IDS) protects the system from malicious attacks. There is a strong demand for effective intrusion detection system. IDS is designed to construe intrusion attempts of incoming traffic accurately, intelligently and effectively. One of the main challenges in the security management of the high speed networks is the detection of suspicious anomalies. A computer security is expressed as confidentiality, availability and integrity of information to authorised users.

Data Confidentiality: Data transferred through a network should only be accessible to authorized users.

Data Integrity: Data should maintain integrity in the network. No corruption or data loss is accepted either from random events or malicious activity.

Data Availability: The network should recover from Denial of Service attacks (DOS).

The KDDCup 99 dataset is the most commonly used intrusion detection dataset. KDDCup 99 dataset was built by processing tcpdump data of 1998 DARPA intrusion detection challenge dataset. The KDDCup 1998 dataset was created by MIT Lincon laboratory using 1000's of UNIX machines and 100's users accessing those machines. The network traffic data was stored in tcpdump format for 10 weeks[2]. Later this can be processed and converted into connection records. A connection is a sequence of TCP packets starting and ending at well-defined times with well-defined protocols. The data of first seven weeks are used as training dataset and rest as testing dataset. The KDDCup 99 are of two forms : full dataset and 10% dataset. The dataset contains 41 features and 5 classes (Normal, DOS, Probe, R2L, U2R)

TABLE-1 TYPE OF ATTACKS IN KDDCup 99 DATASET[14]

ATTACK TYPE	PATTERN OF ATTACK
Denial of attack	Back, Land, Neptune, Pod, Smurf, Teardrop
Probe	Ipsweep, nmap, portsweep, satan
Root to Local	Fp_write, guess_password, imap, multihop, phf, spy, warezclient, warezmaster
User to Remote	Buffer_overflow, loadmodule, perl, rootkit

The factors for the measurement of Intrusion detection system are:
False Positive (FP): represents the number of instances classified by intrusion detection system as being anomalous when infact they are legitimate.
False Negative (FN): represents the number of instances classified by intrusion dtection system being legitimate infact they are anomalous.
True Positive (TP): represents the number of instances that are classified by the intrusion detection systems as being anomalous and they are really anomalous.
True Negative (TN): represents the number of instances that are classified but the intrusion detection systems as being legitimate and they are really legitimate[13].

AN INTRUSION DETECTION SYSYTEM

An Intrusion detection System is a software to detect malicious activities in a network. For a secure network various preventive mechanisms were developed. Primarily Intrusion detection system can be classified into two:

- Host based intrusion detection system
- Network based intrusion detection system

1) Host based intrusion detection system (HIDS)

It is capable of monitoring and analyzing attacks in a network. This was the first type of intrusion detection software designed. A HIDS is capable of verifying all parts of dynamic behavior and the state of a computer system based on the configuration. HIDS are installed in a host and they monitor traffics that are originating and coming to that particular hosts only. If the attack is on any other part of network they will not be detected by HIDS. It monitors all the user activities. They do not require any extra hardware since they can be installed in the existing host servers. In small scale network HIDS is preffered.

2) Network based intrusion detction system (NIDS)

It is used to monitor and analyze network traffic to protect a system from network based threats. To capture all the data passing through the network, you need to position your IDS at the entry and exit point of data from your network to the outside world. Depending on the function NIDS are divided into two :
Statistical anomaly IDS and Pattern matching IDS.

In statistical based IDS model, the IDS try to find out users' or system's behavior that seem abnormal.The main advantages of this type of IDS is that they can detect the type of intrusion that has no records of its previous occurrence. In that sense, statistical anomaly can detect new type of attack patterns. A large number of false alarms are the main problem with this system.

In pattern based system, the IDS maintain a database of known exploits and their attack pattern. During the analysis of network packets if it finds any pattern match to one of those known attack patterns then it triggers alarm.

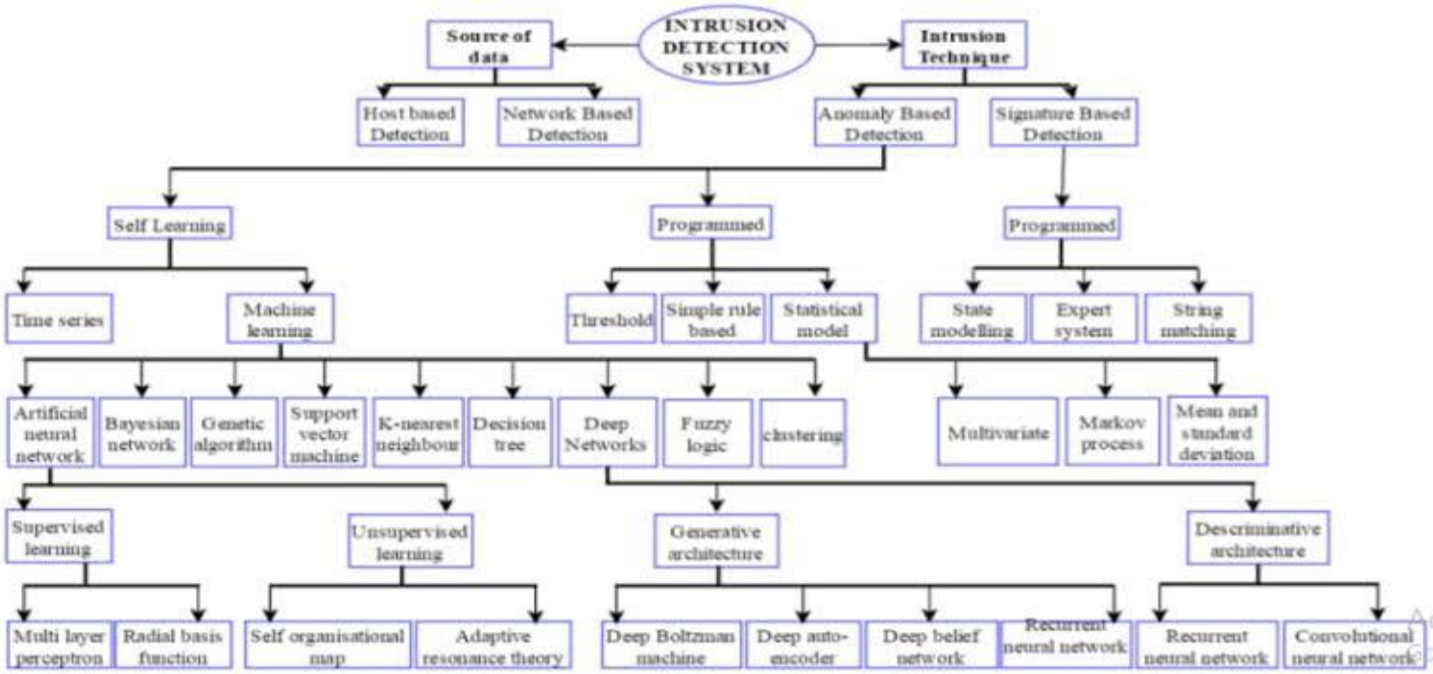

Fig 1 : Intrusion Detection System

MACHINE LEARNING METHODS FOR INTRUSION DETECTION

K-MEANS CLUSTERING ALGORITHM

K-Means algorithm is an unsupervised learning algorithm. It categorize a given dataset of items into groups. The algorithm will categorize the items into k groups of similarity. The Euclidean distance measurement is used to calculate the similarity. The algorithm works as follows:
1. Initialize number of clusters K
2. Categorize reach item to its closest mean and update the mean which are the average of the selected items.
3. Continue the process till there is no change to the centroids.
The applications of K-Means algorithm is in Pricing Segmentation, Document clustering, Image segmentation, Image compression etc.

SUPPORT VECTOR MACHINE ALGORITHM

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It constructs a hyperplane or a set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression and other outlier detection.As there are many such linear hyperplanes, SVM algorithm tries to maximize the distance between the various classes that are involved and this is referred as margin maximization. SVM is a global classification model that generates non overlapping partitions and usually employs all attributes.
SVM's are classified into two categories:
Linear SVM: The classifiers are seperated by hyperplane.
Non-linear SVM: It is not possible to seperate the classifier using a hyperplane. In such conditions the classifier is too complex and is impossible to find a representation to every feature vector. The applications of Support vector machines are in Permutation tests, Classification of images, Text and hyper text categorization etc.

NAIVE BAYES CLASSIFIER ALGORITHM

The Naive Bayes classifier is a probabilistic machine learning model used for classification. Naive bayes classifiers are a collection of algorithms based on Baye's Theorm. These algorithms are independent. It is a straightforward and powerful algorithm for classification. The statistical classification technique based on Bayes Theorm. It is fast, accurate and reliable algorithmwith high accuracy and speed on large datasets. The applications of naive bayes classifiers is in Real time prediction, Spam filtering, Text classification, Multi class prediction etc.

APRIORI ALGORITHM

Apriori algorithm is an unsupervised machine learning algorithm, that generates association rules from a given dataset. It is a categorization algorithm.
Apriori algorithm is for finding frequent itemsets in a dataset for boolean association rules. It uses prior knowledge of frequent itemset. To improve the efficiency of level-wise generation of frequent itemsets, Apriori property is used which helps by reducing the search space. The Apriori property is defined as : All subsets of a frequent itemset must be frequent. If an itemset is infrequent, all its supersets will be infrequent. Apriori algorithm has many applications in the field of software bugs, data analysis, market basket analysis, cross marketing etc.

LINEAR REGRESSION ALGORITHM

Linear Regression is a machine learning algorithm based on supervised learning. This algorithm is referred as the combination of input variable(x) in order to predict the output variable (y). If there is a single input value (x), then the linear model is represented as Simple Linear Regression. If there are multiple input values (x), the model is represented as Multiple Linear Regression. Different regression models differ based on the kind of relationship between dependent and independent variables being used. The most popular applications of linear regression algorithm are in financial portfolio prediction, salary forecasting, real estate prediction etc.

LOGISTIC REGRESSION ALGORITHM

Logistic regression is the most famous machine learning algorithm after linear regression. It is a supervised classification algorithm. Linear regression algorithms are used to predict values. This algorithm measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic/sigmoid function. The output is in binary or in the form of 0/1,-1/1. The major difference between logistic regression and linear regression is that: Logistic regression is used when the dependent variable is binary in nature. The Linear regression is used when the dependent variables are continuous and the nature of regression line is linear. Logistic regression is used in various fields, it can be used for cancer detection problems, it computes the probability of an event occurrence.

DECISION TREE ALGORITHM

Decision Tree algorithm belong to the family of supervised learning. It is used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. For predicting a class label, compare the values of root attributes with record's attribute. On the basis of comparison follow the branch corresponding to that value and jump to the next node. The comparison continues until reaches the leaf node with predicted class values. In decision tree the major challenge is the identification of the attribute for the root node in each level. This process is called as attribute selection. The popular attribute selection measures are : Information Gain and Gini index. The decision tree is effectively used to determine the species of animal. The decision tree method is used in classification, prediction, interpretation and data manipulation.

ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN) are computational algorithms. This computing systems inspired from biological neural network. Neurons are the atomic unit of a biological neural network. It is capable of machine learning as well as pattern recognition. ANN is an oriented graph. It consists of nodes which are connected by arcs. Each arc associated with weight at each node. The input to the node define Activation function along the incoming arcs. A neural network sends and process signals in the form of electrical and chemical signals. They are connected with synapses, which allow neurons to pass signals. An ANN is an information processing technique. ANN include a large number of connected processing unit that works together to process information. It consists of input layer, hidden layer and output layer. The input layer represents the raw information to the network. The hidden layer determine the activity of each hidden unit. The output layer represents the behaviour of output unit depends on the activity of hidden unit. Neural network is applied both for classification and regression of continuous target attributes. ANN have been used on variety of tasks including computer vision, speech recognition, medical diagnosis, video games etc.

K-NEAREST NEIGHBOR ALGORITHM

K-Nearest Neighbor (KNN) is supervised machine learning algorithm and is used for both classification and regression predictive problems. The two properties of KNN algorithm is:
Lazy learning algorithm: KNN is a lazy learning algorithm. It does not have a specialized training phase and uses all the data for training while classification.
Non-parametric algorithm: KNN is also non parametric learning algorithm. It does not assume anything about the underlying data.
KNN uses feature similarity to predict the values of new data points. The new data point will be assigned a value based on how closely it matches the point in the training set. The input consist of k closest training examples in feature space, the output depends on whether KNN used for classification or regression. A peculiarity of KNN algorithm is that it is sensitive to the local structure of the data. The applications of KNN algorithm includes in pattern recognition, gene expression, to measure document similarity etc.

ASSOCIATION RULE LEARNING ALGORITHM

Association rule learning is a rule-based machine learning method. It detects interesting relations between variables in a large databases. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Basket Analysis. The association rule mining is defined as:
Let I be the set of items of 'n' binary attributes.
Let D be the set of transactions called the database.
Each transaction in D has an unique ID and contains a subset in I items. The rule is represented as:
X => Y, where X, Y $\underline{C}$ I
The Rule evaluation metrics:
Support(s): Indicates the number of transactions in the dataset. It is interpreted as the fraction of transaction of two sets.
Confidence(c): Ratio of number of transactions that include all items in {B} as well as in {A} to the number of transactions that includes all items in {A}. It measures how often each item in Y appears in transactions that contains items in X also.
Conf(X=>Y) = Supp(X *U* Y) / Supp(X)
Association rules satisfies a user-specified minimum support and a user specified minimum confidence simultaneously. A minimum support is used to find all frequent itemsets in a database. A minimum confidence is applied to these frequent itemsets to form rules.
Applications of association rule mining are stock analysis, web log mining, medical diagnosis etc.

DEEP LEARNING ALGORITHM

Deep learning is a machine learning technique based on artificial neural networks. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled. Deep learning utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. Deep learning is an artificial intelligence function that act like a human brain in processing data in decision making. It is able to learn from data that is both unstructured and unlabeled. In deep learning, a computer model learns to perform classification tasks directly from images, text or sound[15]. Deep learning models are trained by using a large set of labeled data and neural network architectures that contain many layers. Deep learning achieves recognition accuracy at higher levels. The applications are used in industries from automated driving to medical devices.

DIMENSIONALITY REDUCTION ALGORITHM

Dimensionality reduction is the process of reducing the random variables under consideration, by obtaining a set of principal variables. The two components of dimensionality reduction algorithm is feature selection and feature extraction.

Feature selection: To find a subset of original set of variables or features to get a smaller subset which is used to model the problem. It involves Filter, Wrapper and Embedded methods.
Feature extraction: This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser number of dimensions.
Dimensionality reduction can be both linear or non-linear, depending upon the method used. It helps in data reduction and hence reduce the storage space. It reduces the computation time. It also helps to remove if any redundant features exist. The methods for dimensionality reduction algorithm is Principal Component Analysis(PCA), Linear Discriminant Analysis(LDA) and Generalized Discriminant Analysis (GDA). Dimensionality Reduction is applied in the fields of text mining, intrusion detection, face recognition etc.

RANDOM FOREST CLASSIFIER

Random forest classifier is ensemble algorithm. An ensemble algorithm combines more than one algorithms of same or different kind for classifying objects. Random forest is a supervised classification tree algorithm. The goal of this classifier is to enhance trees classifiers based on the concept of forest. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. Basic parameters to Random Forest Classifier can be total number of trees to be generated and decision tree related parameters like minimum split, split criteria etc[7]. Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. Random forest is good in accuracy, robustness and ease of use. The applications include in banking, Medicine, Stock market, e-commerce etc.

CONCLUSION

The objective of this paper is to provide an overview of the machine learning algorithms for intrusion detection. This paper gives a complete study of important machine learning methods to detect an intrusion. The large volume of database is increasing rapidly resulting in gradual rise in the security attacks. IDS are essential for day today security in corporate world and for network users. The KDDCup 99 dataset is commonly used for intrusion detection. The testing phase is implemented based on random instances of records. Several performance metrics are computed such as accuracy rate, precision, false negative, false positive, true negative, true positive. There is no single machine learning algorithm which can handle efficiently all the types of attacks.

REFERENCES

[1] Chie-Hong Lee, Yann-Yean Su,Yu-Chun Lin and Shie-Jue Lee, "Machine Learning Based Network Intrusion Detection,"2017 2nd IEEE International Conference on Computational Intelligence and Applications.

[2] Dhikhi T and M.S. Saravanan, "An Enhanced Intelligent Intrusion Detection System using Machine Learning," International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278- 3075, Volume-8 Issue-9, July 2019

[3] Dr. S.Vijayarani1 and Ms. Maria Sylviaa.S, " INTRUSION DETECTION SYSTEM – A STUDY," International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1, February 2015

[4] https://www.geeksforgeeks.org/machine-learning/#int

[5] Hassan Azwar,Muhammad Murtaz,Mehwish Siddique,Saad Rehman, "Intrusion Detection in secure network for Cybersecurity systems using Machine Learning and Data Mining,"2018 IEEE 5th International Conference on Engineering Technologies & Applied Sciences, 22- 23 Nov 2018, Bangkok Thailand.

[6] http://dni-institute.in/blogs/k-means-clustering-examples-and-practical- applications/

[7] https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine- learing/

[8] https://securitywing.com/host-based-ids-vs-network-based-ids/

[9] Kajaree Das, Rabi Narayan Behera, " A Survey on Machine Learning: Concept, Algorithms and Applications," DOI: 10.15680/IJIRCCE.2017. 0502001

[10] Md Rakibul Islam and Aayush Nagpal, "Intrusion Detection System: A Robust Machine Learning Approach,"

[11] Md Zahangir Alom and Tarek M. Taha, "Network Intrusion Detection for Cyber Security using Unsupervised Deep Learning Approaches,"978-1-5386-3200-0/17/$31.00 ©2017 IEEE

[12] Mohammad Almseidin, Maen Alzubi , Szilveszter Kovacs and Mouhammd Alkasassbeh, "Evaluation of Machine Learning Algorithms for Intrusion Detection System,"SISY 2017 • IEEE 15th International Symposium on Intelligent Systems and Informatics • September 14-16, 2017 • Subotica, Serbia.

[13] Norbert ÁDÁM, Branislav MADOŠ, Marek ČAJKOVSKÝ, Ján HURTUK, Tomáš TOMČÁK, "METHODS OF THE DATA MINING AND MACHINE LEARNING IN COMPUTER SECURITY," Acta Electrotechnica et Informatica, Vol. 14, No. 2, 2014, 46–50, DOI: 10.15546/aeei-2014-0017

[14] Philip K. Chan, Richard P.Lippman, " Machine Learning for Computer Security," journal of Machine Learning Research 7 (2006) 2669-2672 Submitted 12/06; Published 12/06

[15] Rahul Vigneswaran K , Vinayakumar R , Soman KP and Prabaharan Poornachandran, "Evaluating Shallow and Deep Neural Networks for Network Intrusion Detection Systems in Cyber Security,"

[16] Rohit Kumar Singh and Gautam Er.Amit Doegar, "An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms,"978-1-5386-1719-9/18/$31.00 #2018 IEEE.

[17] R. VINAYAKUMAR, MAMOUN ALAZAB , (Senior Member, IEEE), K. P. SOMAN,PRABAHARAN POORNACHANDRAN , AMEER AL- NEMRAT ,AND SITALAKSHMI VENKATRAMAN, "Deep Learning Approach for Intelligent Intrusion Detection System,"Digital Object Identifier 10.1109/ACCESS.2019.2895334

[18] Rung-Ching Chen,Kai-Fan Cheng,Ying-Hao Chen and Chia-Fen Hsieh, "Using Rough Set and Support Vector Machine for Network Intrusion Detection System,"2009 First Asian Conference on Intelligent Information and Database Systems.

[19] Syam Akhil Repalle1, Venkata Ratnam Kolluru, " Intrusion Detection System using AI and Machine Learning Algorithm," Volume: 04 Issue: 12 Dec 2017 p-ISSN: 2395-0072

[20] SYDNEY MAMBWE KASONGO AND YANXIA SUN, "A Deep Learning Method With Filter Based Feature Engineering for Wireless Intrusion Detection System,"Digital Object Identifier 10.1109/ACCESS.2019.2905633.

AUTHORS PROFILE

Ms. Ajeesha M I currently pursuing Phd in the area of Data mining/Machine Learning from RathnavelSubramaniam College of Arts & Science affiliated to
Bharathiar University Coimbatore. She obtained M.Phil
in the area of Data Mining from RathnavelSubramaniam College of Arts & Science affiliated to Bharathiar University Coimbatore in 2018. Her research interests are Data Mining and Machine Learning.

Dr. D. Francis Xavier Christopher received his Ph.D., in the area of Networking from Bharathiar University, Coimbatore in 2014 from Bharathiar University, Coimbatore. He obtained his M.Phil, in the area of Networking from Bharathiar University, Coimbatore in 2002. At present he is working as a Director, School of Computer Studies in Rathnavel Subramaniam College of Arts and Science, Coimbatore. His research interest lies in the area of Networking and Software Engineering. He has published 27 research papers in various reputed journals ranking with international standard. He served as a key note speaker for various research conferences country wide.