Machine Learning and Deep Learning methods for Cybersecurity

Prof. Nanda M B[1], Parinitha B S[2],

[1]*Assistant Professor, Department of Computer Science and Engineering, Sapthagiri College of engineering, Bangalore, Karnataka, India*

[2]*UG Student, Department of Computer Science and Engineering, Sapthagiri College of engineering, Bangalore, Karnataka, India*

---***---

Abstract - *The detection and prevention of a network intrusion is a major concern. Machine Learning and Deep Learning methods detect network intrusions by predicting the risk with the help of training the data. Various machine learning and deep learning methods have been proposed over the years which are shown to be more accurate when compared to other network intrusion detecting systems. This survey paper gives a brief introduction about various machine learning and deep learning algorithms.*

***Key Words*: Cybersecurity, Machine learning, Deep Learning, Network Intrusion, Algorithm.**

1.INTRODUCTION

Cybersecurity is a set of technologies and processes designed to protect computers, networks, programs and data from attacks and unauthorized access, alteration, or destruction [1].

Machine Learning is a branch of artificial intelligence which predicts about future details by analysis of trained data. [1]

Deep Learning is a branch of machine learning, which analyses multi-layered representation of the input data and makes predictions and decisions.[2]

Manual detection of cybersecurity threats is not accurate as that of detection of cybersecurity threats by machine learning and deep learning. One of the simple and easy methods is to train the computer system, to do the work without human intervention. This is achieved by employing machine learning techniques along with the cybersecurity techniques. This is highly useful in detection, diagnosis of any medical abnormalities.

Earlier methodologies employed the use of only cybersecurity algorithms and techniques. This requires human effort to detect each and every cybersecurity threats. This includes even detection of already existing cybersecurity attacks. It requires as much effort to detect already existing type of attack so as to detect a new type of attack This is highly impossible task, since there are millions of cybersecurity attacks found among people all over the world. Hence, the detection of cybersecurity threats is of great concern. Therefore, the detection of these threats is very important. This can be done by using machine learning and deep learning algorithms in the field of cybersecurity.

The algorithms used for detection of cybersecurity threats is as follows:

1. Convolutional Neural Network (CNN)

2. Support Vector Machine (SVM)

3. K-Nearest Neighbor (KNN)

4. Decision Tree

5. Deep Belief Network (DBN)

6. Recurrent Neural Network (RNN)

2.Literature Survey

2.1. INTRUSION DETECTION USING CONVOLUTONAL NEURAL NETWORK

The intrusion detection model based on a CNN includes four steps:

STEP 1: Data pre-processing. This step adjusts the initial data format and normalizes the data values. In order to improve the performance of CNN model, it needs to convert normalized data into image data format.

STEP 2: Training. Training is performed on the datasets to improve the performance of CNN model performance.

STEP 3: Testing. The test data is used to check the accuracy rate of the CNN model. For example, if the accuracy rate could satisfy the training requirement, the training would cease; otherwise, the model would repeat the training step (STEP 2).

STEP 4: Evaluating. This step is used to evaluate the performance of CNN model following training. Generally, evaluation metrics include the accuracy rate, detection rate, and false alarm rate.[3]

2.2INTRUSION DETECTION USING SUPPORT VECTOR MACHINE

The intrusion detection model based on an SVM includes four steps:

STEP 1: Data Pre-Processing: It is the process of converting the available raw data in useful data by filtering and removal of useless data.

STEP 2: Conversion of datasets to LibSVM format: After preprocessing of data, datasets are converted to LibSVM format. In this process, categorical features from both training and testing datasets are converted to a numeric value and then target classes are determined for classification phase. The two target classes are, conversion and scaling. Linear scaling of datasets is done to improve the performance of SVM classification.

STEP 3: Optimization using SVM and PSO: The NSL-KDD dataset in LibSVM format is scaled in [0, 1]. The scaling is the method used to reduce the impact of bigger value on small value. It improves the performance of SVM. PSO is dynamic clustering algorithm which improves the quality of a candidate solution. It works better in integration of SVM.

STEP 4: Classification Using SVM: The SVM uses the data to train the system, finding several support vectors that represent the training data. A classification task involves the training and testing data, consisting of some data instances. Each instance in the training set contains one “target value" (class labels: Normal or Attack) and several “attributes" (features). The goal of SVM is to produce a model which predicts target value of data instance in the testing set which is given only attributes. To attain this goal there are four different kernels functions.in this experiment RBF kernel function is used.[4]

2.3. INTRUSION DETECTION USING K-NEAREST NEIGHBOR

K-nearest neighbor (KNN) classification algorithm is a data mining algorithm which is theoretically mature with low complexity. The basic idea is that, in a sample space, if most of its nearest neighbor samples belong to a category, then the sample belongs to the same category. The nearest neighbor refers to the single or multidimensional feature vector that is used to describe the sample on the closest, and the closest criteria can be the Euclidean distance of the feature vector.

Description of the process of this algorithm, we have the following definitions:

(1) the feature vector describing the node.

(2) the assemblage of all nodes in the network.

(3) the Euclidean distance of two different nodes.

(4) the -distance function of node is the value got by the summation of all the most adjacent nodes' Euclidean distance, divided by K. [5]

3. ALGORITHMS OF MACHINE LEARNING AND DEEP LEARNING IN CYBERSECURITY

3.1. SUPPORT VECTOR MACHINE:

Support Vector Machine (SVM) is a machine learning algorithm which is robust and accurate methods. It includes Support Vector Classification (SVC)and Support Vector Regression (SVR). The SVC is based on the concept of decision boundaries. A decision boundary separates a set of instances having different class values between two groups. The SVC supports both binary and multi-class classifications. The support vector is the closest point to the separation hyperplane, which determines the optimal separation hyperplane. In the classification process, the mapping input vectors located on the separation hyperplane side of the feature space fall into one class, and the positions fall into the other class on the other side of the plane. In the case of data points that are not linearly separable, the SVM uses appropriate kernel functions to map them into higher dimensional spaces so that they become separable in those spaces.[1]

3.2.K-NEAREST NEIGHBOR:

KNN algorithm is a machine-learning algorithm used for classification, regression. **KNN** is a **non-parametric, lazy** learning algorithm. It predicts the classification of a new sample point by using a dataset in which the data points are separated into several classes. KNN is used for classification,the output is a class membership (predicts a class—a discrete value). An object is classified by maximum presence of its neighbors, with the object being assigned to the class having maximum number of k nearest neighbors. KNN is used for regression, the output is the value for the object (predicts continuous values). The value predicted is the average of the values of its k nearest neighbors.[1]

The kNN classifier is based on a distance function that measures the difference or similarity between two instances. The standard Euclidean distance d (x, y) between two instances x and y is defined as:

$$d(x, y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

where, x_k is the kth featured element of the instance x,

y_k is the kth featured element of the instance y and

n is the number of features.[1]

3.3. DECISION TREE:

A decision tree is an efficient method used for classification and prediction. It is a tree structure in which each internal node represents a test on one property and each branch represents a test output, with each leaf node representing a category. In machine learning, the decision tree is a predictive model. It represents a mapping between object attributes and object values. Each node in the tree represents an object, each divergence path represents a possible attribute value, and each leaf node corresponds to the value of the object, which can be derived from the root node to leaf node. The decision tree only has a single output. To get complex output, an independent decision tree can be implemented to handle different outputs. Commonly used decision tree models are ID3, C4.5 and CART.[1]

3.4. CONVOLUTIONAL NEURAL NETWORK:

Convolutional Neural Networks is an efficient algorithm, which is widely used. It is a type of artificial neural network. CNN is the first successful learning algorithm for training multi-layer network structures. Convolutional Network is a multi-layered sensor specifically designed to recognize two-dimensional shapes that are highly invariant to translation, scaling, tilting, or other forms of deformation. Its weight-sharing network structure makes it more similar to a biological neural network, thus reducing the complexity of the network model and reducing the number of weights. This advantage is more obvious when the network input is a multi-dimensional image, and the image can be directly used as the input of the network to avoid the complex feature extraction and data reconstruction in the traditional recognition algorithm. There are three main means for CNN to reduce network-training parameters: local receptivity, weight sharing and pooling. The most powerful part of CNN is the learning feature hierarchies from large amounts of unlabeled data. Therefore, CNN are quite promising for application in the network intrusion detection field.[1]

3.5. RECURRENT NEURAL NETWORKS:

The recursive neural network (RNN) is deep learning algorithm, which is widely used. It is used to process sequence data. In the traditional neural network model, data from the input layer to the hidden layer to the output layer; The layers are fully connected and there is no connection between the nodes between each layer. Many problems exist that this conventional neural network cannot solve. The reason that RNN is a recurrent neural network is that the current output of a sequence is also related to the output before it. The concrete manifestation is that the network can remember the information of the previous moment and apply it to the calculation of the current output; that is, the nodes between the hidden layers become connected, and the input of the hidden layer includes both the output of the input layer and the last moment hidden layer output. Theoretically, any length of sequence data RNN can be processed. However, in practice, to reduce the complexity, it is often assumed that the current state is only related to the previous states.[1]

3.6. DEEP BELIEF NETWORK:

Deep Belief Network (DBN) is a deep learning model. It is a probabilistic generative model consisting of multiple layers of stochastic and hidden variables. The Restricted Boltzmann Machine (RBM) and DBN are interrelated because composing and stacking a number of RBMs enables many hidden layers to train data efficiently through activations of one RBM for further training stages. RBM is a special topological structure of a Boltzmann machine (BM). The principle of BM originated from statistical physics as a modelling method based on an energy function that can describe the high-order interactions between variables. BM is a symmetric coupled random feedback binary unit neural network composed of a visible layer and a plurality of hidden layers. The network node is divided into a visible unit and a hidden unit, and the visible unit and the hidden unit are used to express a random network and a random environment. The learning model expresses the correlation between units by weighting.[1]

4. COMPARISIONS OF RESULTS OF VARIOUS TECHNIQUES

Table -1: ML and DL methods and Dataset.

Methods	Dataset	Accuracy	Precision	Paper
SVM	KDD-CUP 99	82.31	74	**Pervez &Farid**
Mix-KNN	KDD-CUP 99	98.55	---	**E. G. Dada**
DBN	KDD-CUP 99	93.49	93.25	**N. Gao, et al**
RNN	KDD-CUP 99	77.55	84.6	**C.L. Yin, et al**
CNN	NetFlow	99.41	---	**W. Wang, et al**
DT	KDD-CUP 99	99.89	---	**Azad and Jha**

5. CONCLUSIONS

This paper presents a survey of ML and DL methods, that can be used for network security. Every method used for implementing an intrusion detection system has its own advantages and disadvantages. In this paper, a comparison is made among the various methods. Therefore, it is not possible to choose only one particular method to implement an intrusion detection system.

6. FUTURE ENHANCEMENTS

In this paper, different types of intrusion detection which can be mitigated by machine learning has been surveyed. The limitation in this area:

1.Scarcity of Dataset

2.Not able to cope up with latest cybersecurity mechanisms.

3.Cybersecurity field requires incremental, life-long and quick training model.

REFERENCES

[1] Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmilo Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, Chunhua Wang," Machine Learning and Deep Learning Methods for Cybersecurity", IEEE 2018.

[2] Giovanni Apruzzese, Luca Ferretti, Mirco Marchetti, Alessandro Guido, Michel Colajanni," On the Effectiveness of Machine and Deep Learning for Cyber Security", 2018, 10th International Conference on Cyber Conflict.

[3] Kehe Wu, Zuge Chen, Wei Li," A Novel Intrusion Detection Model for a Massive Network Using Convolutional Neural Networks",IEEE 2018.

[4] Vitthal Manekar, Kalyani Waghmare," Intrusion Detection System using Support Vector Machine (SVM) and Particle Swarm Optimization (PSO)", International Journal of Advanced Computer Research, September-2014.

[5] M. Govindarajan, Rlvl. Chandrasekaran, "Intrusion Detection Using k-Nearest Neighbor", IEEE 2009.

[6] M. S. Pervez and D. M. Farid, ''Feature selection and intrusion classification in NSL-KDD CUP 99 dataset employing SVMs,'' in Proc. 8th Int. Conf. Softw., Knowl., Inf. Manage.Appl. (SKIMA), 2014

[7] E. G. Dada, ''A hybridized SVM-kNN-pd APSO approach to intrusion detection system,'' in Proc. Fac. Seminar Ser., 2017.

[8] N.Gao, L.Gao, Q.Gao, H.Wang, ''An intrusion detection model based on deep belief networks,'' in Proc. 2nd Int. Conf. Adv. Cloud Big Data, 2014.

[9] C.L.Yin, Y.F.Zhu, J.L.Fei, X.Z.He, ''A deep learning approach for intrusion detection using recurrent neural networks,'' IEEE 2017.

[10] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, ''Malware traffic classification using convolutional neural network for representation learning,'' in Proc. Int. Conf. Inf. Netw. 2017.

[11] C. Azad and V. K. Jha, ''Genetic algorithm to solve the problem of small disjunct in the decision tree-based intrusion detection system,'' Int. J. Comput. Netw. Inf. Secur. 2015.