International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT206287

Intrusion Detection Using Hidden Markov Model and XGBoost Algorithm

Sanjana Gawali[1], Prerana Agale[1], Sandhya Ghorpade[1], Rutuja Gawade[1], Prof. Prabodh Nimat[2]

[1]Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, Maharashtra, India

[2]Professor, Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, Maharashtra, India

ABSTRACT

Security has been widely concerned and recognized as a critical issue in wireless communication networks recently, because the openness of the wireless medium allows unintended receivers i. e. intruders to potentially eavesdrop on the transmitted messages. Unauthorized access by an intruder can be monitored by Intrusion detection system. Machine learning algorithms such as Hidden Markov Model and Extreme gradient boost algorithm can be used for intrusion detection based on CICIDS dataset. Based on dataset, algorithms create classifiers of signatures of particular attack. These trained classifiers are tested against user data for intrusion detection. System reports attack in network. Here, XGBoost classifier gives higher accuracy compared to HMM classifier.

Keywords : Intrusion detection, HMM, XGBoost, CICIDS

I. INTRODUCTION

As network interconnections are expanding rapidly, media applications and data transfer rates are increasing. Various threats and attacks are occurring on network which violets security policies. It includes deleted, unexplained, additional, or modified files; additional accounts on the system; and unaccounted disk or memory usage. Intrusion detection system monitors possible attacks on system. It includes malicious instruction sequences, network traffic, unenumerated memory usage. Proposed system provides intrusion detection system to probe threats in sophisticated networks.

Intrusion Detection System:

Intrusion Detection System is needed to ensure the security of network which scans system against harmful activities. Intruder is a person trying to gain an unauthorised access to a system or a network Outside Intruder is the one who is not having authorised access to the system, whereas the inside intruder has some kind of authorised access to the system but it misuses the resources and privileges. The main goal of Intrusion Detection System is to monitor the data packets in the network.

Network-based intrusion detection systems(NIDS) detects malicious traffic on a network. NIDS monitors, captures and analyzes network traffic. NIDS monitors source and destination of the data packet, path of the data packet and content of data packet.

Host-based intrusion detection system (HIDS) is installed on individual host or device on a network. It monitors data packets only from the device and alerts the admin if any suspicious activity is detected. The

suspicious activity is detected by comparing the existing system and the previous system.

Network Security Issues:

In passive attack when the data is sent to the receiver, the attacker can only read the message. There is no modification in the message content. The main aspects of Passive attack are release of the message content and traffic analysis. In active attack the attacker not only reads the communication but can also write and modify the message. A DOS attack is an active attack and is performed by using only one system. It is done by flooding the website with packets and making impossible to the users to actually access the content of the website which is flooded. A DDOS (Distributed Denial Of Service) attack is more forceful as compared to DOS because the attacker tries to control many systems by using their IP and takes control over the systems.

Signature Based Method is used to detect the attacks which are based on specific patterns such as number of bytes in the network traffic. It detects attack on the basis of already known malicious instruction sequence that is used by the malware. The term signature in network security refers to the detected patterns in IDS. Signature-based IDS can easily detect the attack whose pattern (signature) already exists in the system.

II. PROPOSED SYSTEM

Proposed system uses network dataset and trains it with machine learning algorithms Hidden Markov Model and Extreme gradient boosting algorithm. Generated classifiers have signatures of particular intrusion attack. This classifiers are tested against user data for intrusion detection. Based on the trained dataset, classifier reports attack in system.

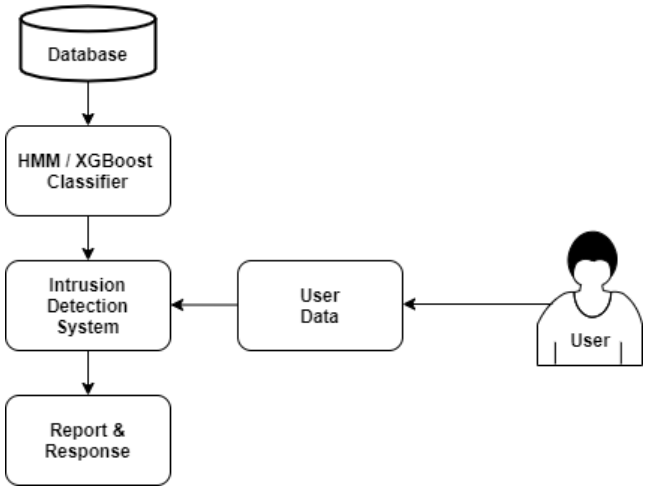

Fig 1. Flowchart of Intrusion Detection System

Hidden markov model (HMM):

Hidden markov model is a model with hidden states and observable emissions, which consists markov property i. e. , the future states depends only present state not any preceded state. Each state has probability distribution over possible output tokens. Hence, hidden markov model generates sequence of state to output token. HMM has Transition probability i. e. various state changes from initial or current state and emission probability i. e. output probability.

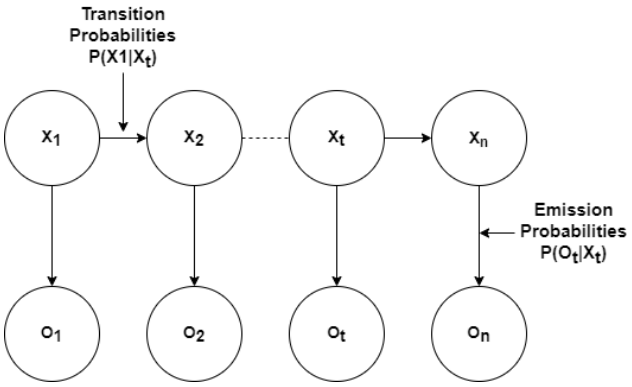

Fig 2. Hidden Markov Model

Consider a system with n states and at discrete time intervals (t), there is transition among states which gives output token. There are three sub-problems in HMM:

1. Likelihood Evaluation:

In this sub-problem, likelihood of the hidden state is determined with forward-backward reasoning of sequences with Forward and backward algorithm. Parameters or states are initialized randomly and then estimated over several training iterations.

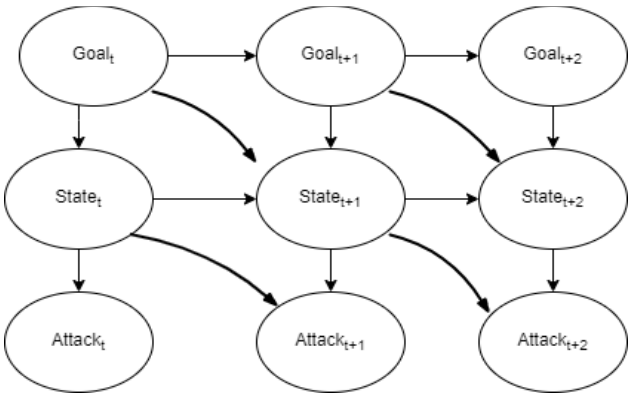

Fig 3. Forward-backward training

2. Decoding Problem:

In decoding problem, Optimal path to reach output state is created through forward and backward way calculated from likelihood evaluation. Viterbi algorithm is used to make inference based on trained model and some observed output. It has forward algorithm: Initialization, Recursion and Termination, but also the backtracking step to find the sequence of hidden states.

3. Inference problem:

This step gives observation sequence and set of hidden states in the HMM. Parameters of the model are maximized to get most specific model as output with greater accuracy with Baum-welch algorithm, case of expectation-maximization algorithm. It is iterative methods to find maximum probability and relations of estimated observation sequences.

HMM gives best observation sequence with hidden states, so it can be used to create signature sequence of specific attacks. These sequences can be used in sophisticated networks to detect intrusion.

Extreme Gradient Boosting (XGBoost):

Boosting is a mechanism used to solve complex real world problems. Boosting techniques combine the weak learners to for a strong learner in order to increase the accuracy of the model. In Gradient boosting base learners are generated sequentially in such a way that the present base learner in more efficient than the previous one. It tries to optimize the loss function of the previous learner or model.

XGBoost is the advanced version of gradient boosting method designed to focus on computational speed and model efficiency. XGBoost supports parallelization by creating decision trees parallel and uses cache optimization to use resources in optimal manner. It implements distributed computing methods to evaluate large and complex problems.

Implementation of XGBoost:

1) Each instance in the training dataset (D) is given weights respectively.
2) Gradients or Derivatives from the original dataset (D) is derived by random samling of the instances.
3) All instances are probable to selected in First derivative (D1) from which the model M1 gets generated.
4) When moving towards the next model updating the misclassified weights is mandatory. i. e. , the probability of misclassified instances increases as compared with others.

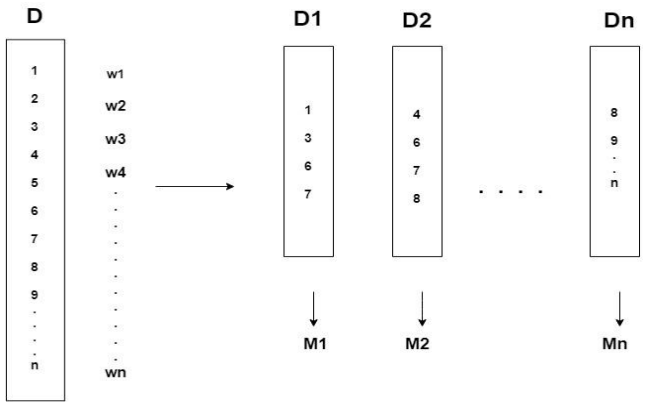

Fig: Gradient distribution and models forming

5) All the weak models (M1, M2,. . . . Mn) created are combined together to obtain a strong model.
6) This strong model is used as a classifier for implementation of XGBoost algorithm.

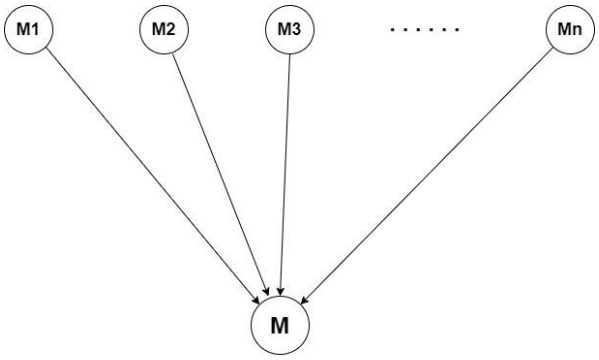

Fig: Final model

III. IMPLEMENTATION

A] Dataset Description:

Dataset CICIDS2017 is used for intrusion detection. The dataset consists of latest threats, features and most updated attack scenarios. The CICIDS2017 dataset contains total 83 features in which there are 15 class labels. In the available class labels, there are 14 attack label and 1 normal label. There are total 15 distinct classes.

B] Preprocessing:

The dataset is preprocessed to eliminate the missing data values or missing information which is of no use for prediction. In data preprocessing, the steps like data cleaning, feature reduction, selection are done which are independent methods. Elimination of redundant means columns which contains a single value that is 0 is dropped because it has no contribution in prediction. Features which are text type and nominal are selected. They are first factorized which convert them to nominal numeric feature. Further, one hot encoding is performed to get its binary feature form. The IP address in dataset are converted to an integer representation. The attack labels present in dataset are in text from which are also converted to numeric representation.

Table 1: Attack label in numeric form

Class label (Attack)	Numeric form
BENIGN	0
DDos	1
PortScan	2

The dataset is splited into training dataset and testing dataset. The train and test dataset have probability distribution which are deliberately different. The dataset is splits in 80:20 ratio in which the 80% is used to train the and remaining 20% is used for testing.

C] Results:

Results of Comparison between HMM and XGBoost :

Table 2: Classification report of algorithm classifier

Classifier	HMM	XGBoost
Precision	0. 61	1. 00
Recall	0. 88	0. 95
F1 Score	0. 75	0. 98
Support	35	62
Accuracy	62. 076 %	98. 3606 %

Fig. 1 shows graph of parameters of confusion matrix, i. e. Precision, recall, F1 score, support and overall accuracy of classifiers.

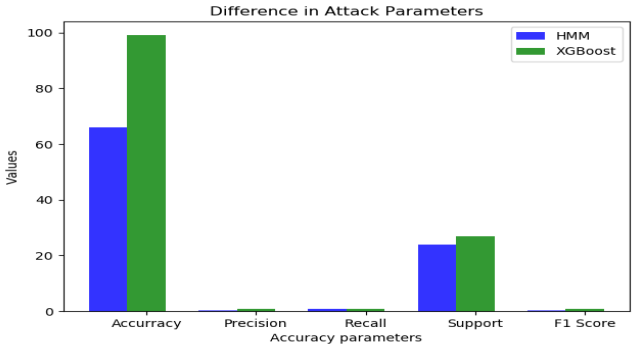

Fig. 1 Attack Parameters graph

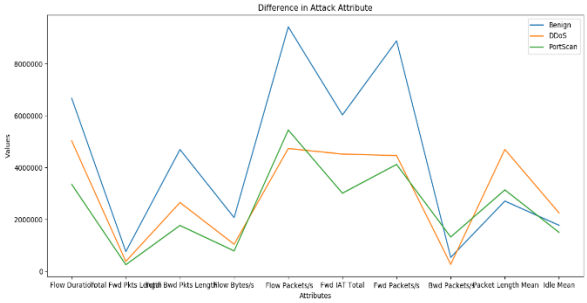

Fig. 2 Attack line graph based on attribute values

In Fig. 2, Graph of attributes of network data which are used to train classifiers with their average values is given. Graph shows the difference between behaviour of attacks such as benign, DDoS and Portscan.

As observed from above comparison, the XGBoost gives the higher accuracy in attack detection than the Hidden Markov Model.

IV. CONCLUSION

In this paper, Intrusion in network is detected by Hidden markov model and extreme gradient boosting algorithm with signature based detection method. Confusion matrix parameters of XGBoost classifier gives greater values than HMM classifier. Classification report shows that XGboost algorithm has more accuracy than HMM algorithm. Based on implemented results, we can conclude that extreme gradient boosting algorithm gives high accuracy and better performance than hidden markov model algorithm for intrusion detection.

V. REFERENCES

[1]. P. Mishra, V. Varadharajan, U. Tupakula and E. S. Pilli, “A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection, “in IEEE Communications Surveys & Tutorials, 2019.

[2]. K. Park, Y. Song and Y. Cheong, “Classification of Attack Types for Intrusion Detection Systems using a Machine Learning Algorithm,”IEEE Fourt International Conference on BigDataService ,Bamberg, 201.

[3]. Alhaidari, Sulaiman et al. “Detecting Distributed Denial of Service Attacks Using Hidden Markov Models. ” (2018).

[4]. Zheng, Ruijuan & Li, Guanfeng & Zhang, Juwei. (2011). Intrusion Intention Identification Methods Based on Dynamic Bayesian Networks. Procedia Engineering.

[5]. Megha Bandgar, Komal dhurve, Sneha Jadhav & Vicky Kayastha. (2013). Intrusion Detection System using Hidden Markov Model(HMM). IOSR-JCE. Vol 10, Issue 3(March-Apr. 2013).

[6]. J. Hu, X. Yu, D. Qiu and H. Chen, "A simple and efficient hidden Markov model scheme for host-based anomaly intrusion detection," in IEEE Network, January-February 2009.

[7]. P. Verma, S. Anwar, S. Khan and S. B. Mane, "Network Intrusion Detection Using Clustering and Gradient Boosting," 2018 (ICCCNT), Bangalore, 2018, pp. 1-7.

[8]. Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu and J. Peng, "XGBoost Classifier for DDoS Attack Detection and Analysis in SDN-Based Cloud," 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, 2018, pp. 251-256.

[9]. Ranjit Panigrahi, Samarjeet Borah, “A Detailed Analysis of CICIDS2017 Dataset for Designing Intrusion Detection Systems”,International Journal Of Engineering & Technology.

[10]. Sukhpreet Singh Dhaliwal, Abdullah-Al Nahid, Robert Abbas, “Effective Intrusion Detection System Using XGBoost”, School Of Engineering, Macquarie University.

Cite this article as :

Sanjana Gawali, "Intrusion Detection Using Hidden Markov Model and XGBoost Algorithm", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 2, pp. 466-470, March-April 2020. Available at doi : https://doi. org/10. 32628/CSEIT206287
Journal URL : http://ijsrcseit. com/CSEIT206287