ISSN 2395-1621

A Machine learning Powered Cyber Security System

Ganesh Sunil Bayas, Charudatta Kapade, Diksha Kadam, Raviraj Nagul

ganeshb979@gmail.com
dikshakadam126@gmail.com
nagulraviraj@gmail.com

Jayashree C. Pasalkar

Assistant Professor Information Technology

AISSMS Institute of Information Technology, Pune, India
jayashree.pasalkar@aissmsioit.org

ABSTRACT

Security of a system is a major issue in any enterprise. It contains important data related to the company or any other system. SIEM (Security Information and Event Management) System is an approach to security management that combines SIM (security information management) and SEM (security event management) functions into one security management system. Thus the network has to be monitored continuously so as to detect the suspicious traffic. In our system we use SVM and Random Forest algorithms are used as compared to other algorithms as its generalization ability and their efficient validation results.

Keywords: machine learning,cyber Security, Deeplearning, SIEM.

ARTICLE INFO

Article History

Received: 8th March 2020

Received in revised form :

8th March 2020

Accepted: 10th March 2020

Published online :

11th March 2020

I. INTRODUCTION

The The purpose of this literature survey is to compare the various threats or intrusion detection problems and generate best algorithm for complex problem solving. It also consists of different methods used in Data Mining and machine learning alorithms. Cyber Security is the main component which is essential to protect the programs, networks, unauthorized access and so on. There are three main Cyber Analytics which fulfills IDS i-e misue based , anomaly based & hybrid. A network contains various and enormous amount of data which automatically generates traffic which has malicious data in it. Hence it is important that the network goes under continous check so that the possible intrusions are detected and the voilation of the policies of organization is avoided[2]. One of the most important system for detecting malicious activities, the Security Information and Event Management SIEM.

When huge amount of data is collected daily by SIEM they are been sorted within different catagories. This system correlates with the end points from IDS/IPS, DNS, VPN logs etc. The classification of different algorithms is carried out by performing confusion matrix[2]. The classification carried out by the number of instances which are predicted accurately or inaccurately in tabulated form[2].

As the SIEM generates too many alerts. And also with the high positive rates. The number of alert increases day by day which may also extend over thousand. The SOC Security Operation Centre investigate the alerts to decide the alert is truly malicious or not. The capacity of SOC is less then SIEM to handle the alerts because of too many alerts SOC will investigate only those alerts which has high severty[1].

II. LITERATURE SURVEY

SR.NO	TITLE	AUTHORS	YEAR	METHODOLOGY
1.	A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection.	Anna L. Buczak, Member, IEEE, and Erhan Guven, Member, IEEE	2016	There is different types of datasets which are trying to analyse patterns of network which are going to collect record of all system and notify to the administrator that they can disable the particular network.
2.	Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection.	Sumouli Choudhury and Anirban Bhowal	2015	Continues monitoring is necessary so as to detect policy violation hence machine learning algorithms are applied using classifiers in WEKA tool and also including Random Forest and Bayes Net.
3.	A Comparative Analysis of SVM and its Stacking with other Classification Algorithm for Intrusion Detection.	Nanak Chand, Preeti Mishra, C. Rama Krishna, Emmanuel Shubhakar Pilli	2016	With view of network security, we intend to apply machine learning specially using SVM as base classifier for detecting intrusions in network. We have studied performance of SVM and its stacking with 9 other machine learning algorithms. We have used NSL-KDD99 data set to analyse performance of all the classifiers.
4.	A Survey of Intrusion Detection System and Internal Intrusion Detection and Protection System.	Amol Borkar,Akshay Donode,Anjali Kumari	2017	This model introduce information of different types of attacks and Techniques like decision free,Random forest.
5.	A Survey of Attack Projection and Forecasting in Cyber Security	Martin Husak,Jana Komarkova,Elias Bob-Harb,and Pavel Celeda	2018	This model presented prediction methods.The problem was set in a context of research on intrusion detection and cyber situational awareness.
6.	A Detailed Investigation and Analysis of using Machine Learning Techniques.	Preeti Mishra,Vijay Varadharajan,Uday Tupakula,Emmanuel S,Pilli,Senior	2018	In this model different comparison of machine algorithms are made and TP and NP graphs are analysed.
7.	Network Anomaly Detection: Methods, Systems and Tools	Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita	2014	The state-of-the-art in the modern anomaly-based network intrusion detection have examined.two well-known criteria to classify andevaluateNIDSs:detectionstrategyandevaluationdatasets
8.	Network Digest analysis by means of association rules	Daniele Apiletti, Elena Baralis, Tania Cerquitelli, Vincenzo D'Elia	2015	This paper presents a Network Digest framework, which performs network traffic analysis by means of data mining techniques to characterize traffic data and detect anomalies

III. RELATED WORK

A. Algorithms Used In Proposed System:-

- Random forest

random forest combines the decision trees and esemble learning. The trees are picked Randomly from the forest as their input . The prediction i-e the resulting prediction is done by majority voting or weighted voting. Major Advantage of Random Forest is that the variance of the model increases as the Forest trees increases.The Study collected the data set into attacks such as Dos and Probe and minority attacks such as U2R and R2L .The accuracies reported by Random Forest are 97%, 76%, 5% and 35% for DoS, Probe , R2L, and U2R respectively.[3]This classification algorithm is used to obtain better predictive Performance. Logistic regression
This type of regression is used to predict the outcome which can have only two values. It is also know as predictive Analysis Regression . It produces a logistic curve that denotes the value between 0 and 1. It is used to describe the data and Explain Relationship between one dependent variable and one or more independent variables.It is not Regression model it is a Classification model.

- Multi-layer neural network (MNN)

It is used in solving Non linear sets which has hidden layers. Its neurons are not directly connected to the output. The training of the sets is in Supervised style. Its main objective is to present the input vector to the network and calculate output of forward direction and generate final output for network .

- Support Vector Machine (SVM)

This type of technique in ML is used for both classification as well as for regression.It mainly has two variant which are support linear and non linear problems. They are mainly used for text classification , bioinformatics . Etc. SVM are well known for its generalization ability. SVM classifier is based on finding a separating hyperplane between two classes. Such that the distance between the closest data points and hyperplane of each class is maximized.

IV. PROPOSED METHODOLOGY

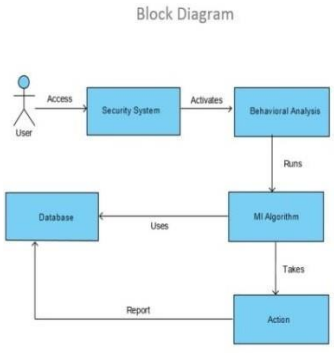

Fig 1. Architecture of Proposed Model

In fig.1.The system working is explained in which the principal parts of function are represented by blocks connected by lines that show the relationship of blocks and working of system.System contains five blocks name Security System, Behavioural Analysis,Database,Ml Algorithm,Action.In First phase user Access the Security System then he select the option Activates Behavioural analysius and Runs the Ml Algorithm if any malware of Network intrusion or any unusual activities is found the System Takes action and alert the database and runs ml algorithm.This is working of system.

V. PERFORMANCE MEASURES

TABLE I. Performance measures play an important roles in machine learning. They are not only used as the criteria to evaluate learning algorithms, but also used as the problem solving technique to construct learning models. It is comparable with the measurable terms like accuracy, specificity, sensitivity, training time etc. On a basis of confusion matrix different aspects can be calculated. Occurrences which may be predicted accurately or inaccurately by a model which is a categorised tabulated form of confusion matrix. It can be shown by four values which are TP, FN, FP and TN.

True positive:It refers to the occurrences which are predicted normally.
False negative:It refers to the incorrect prediction,
i.e. detects occurrences which are attacks in actuality, as normal.
False positive:It gives an indication of the number of detected attacks which are normal in actuality.
True negative:It refers to occurrences which are correctly detected as an attack.

		Predicted	
		Normal	Anomaly
Actual	Normal	TP	FN
	Anomaly	FP	TN

ROC (Receiver Operating Characteristics):It is required to draw the curve between true positive rate (TPR) and false positive rate (FPR). The Area under the curve (AUC) is directly proportional to ROC. More the value of AUC, more will be the value of ROC.

Sensitivity:Also called as true positive rate and reference of the actual positives which are correctly recognized. The algorithm can then predict positive occurrences correctly.
Sensitivity: TP / (TP+FN)
Specificity: Also called as true negative rate and determines the actual negatives which are identified correctly.
Specificity: TN / (FP+TN)

Precision: Calculation of the chances of a positive prediction which are being correct.
Precision: TP / (TP+FP) .
Accuracy: (TP+TN) / (TP+TN+FP+FN)
Kappa: Used to check the dependency of the classification of algorithm on the dataset. Shown by discrete values 0 and 1. 0 refers to total unacceptance and 1 refers to full acceptance.
Mean absolute error (MAE): Error should be least for an algorithm to be the best in performance.
F1 score:It is called as the symmetrical mean of sensitivity and precision. The calculation is below shown,
F = 2*TP / (2TP+FP+FN)
False positive rate:It indicates the chances of an algorithm to predict occurrences as attacks which are actually normal.
FPR = FP / (FP+TP) = 1 – PPV
Negative predictive rate: It refers to the predictability of an algorithm to correctly detect occurrences as attack.
NPV = TN / (TN+FN)
Training time:Here the value of the parameter is directly proportional to the classifier. More the value of the parameter, more will be the classifier.

VI. DESCRIPTION ON DATASETS

A. Defense AdvanceResearch Project Agency(DARPA)-

The system prior used DARPA data sets 1999 its major drawback was that it was time consuming and was difficult to obtain a representative data set. These data sets was so huge that the researchers too worked of other subsets.

B. Netflow-

Later DARPA the netflow dataset came into existence which were used in prior System. Netflow does not have the important feature such as tcp dump. It was introduced as a router feature by the CISCO. Usually the netflow contains three main components which are

- A netflow exporter
- Netflow Collector
- Analysis console.

Its feature were limited only to flow the information which was generated by higher end routers.

C. NSL-KDD –

The survey that many studies used in KDD and DARPA datasets and also applied other types of ML Methods.The datasets used in other papers were NSL-KDD dataset. It classifies the network traffic into two types such as anamaly and normal.The NSL-KDD consists some selected records of the complete KDD data sets.

VII.CONCLUSION

This paper describes the literature review of machine learning methods used for cyber security by using these methods we shall present a user – centric machine which manages the big data of various security logs. And also initialize to focus on the identification of risky user. The purpose system framework provides a solution on risky user detection. For the enterprise security operation center. In our paper we put forth improved and efficient algorithms necessary for proper detection of risky user in our network.

REFERENCES

[1] Charles Feng ZhongDu Technologies, Inc. “A User - Centric Machine learning framework for cyber security operations center.

[2] Sumouli Choudhary and Anirban Bhowal. “Ccomparative Analysis of Machine Algorithms alon with Classifiers for Network Intrusion Detection”.

[3] A. L. Buczak and E. Guven. “A Survey of data mining and machine learning methods for cyber security intrusion detection”, IEEE Comunications Surveys & Tutorials.

[4] N. Chand et al. “A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection”.

[5] K. Goeschel. “Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis”.

[6] M. Bhuyan, D. Bhattacharyya, and J. Kalita, “Network anomaly detection: Methods, systems and tools,” IEEE Commun. Surv. Tuts.

