Intrusion Detection System

Abhishek Doshi
Computer Engineering
Universal College of Engineering
Mumbai, India,

Poonam Thakre
Computer Engineering
Universal College of Engineering
Mumbai, India.

Zil Shah
Computer Engineering
UniversalCollege of Engineering
Mumbai, India,

Abstract: ***Intrusion detection is the act of detecting unwanted traffic on a network or a device. An IDS can be a piece of installed software or a physical appliance that monitors network traffic in order to detect unwanted activity and events such as illegal and malicious traffic, traffic that violates security policy, and traffic that violates acceptable use policies. This article aims at providing a general presentation of the techniques and types of the intrusion detection and prevention systems and an in-depth description of the evaluation, comparison and classification features of the IDS and the IPS. Many IDS tools will also store a detected event in a log to be reviewed at a later date or will combine events with other data to make decisions regarding policies or damage control. An IPS is a type of IDS that can prevent or stop unwanted traffic. The IPS usually logs such events and related information.MachineLearning, it is a field of computer science that uses statistical techniques to give the ability to learn to the computer systems with data, Comparative study is based on Machine Learning, IDS and KDD dataset. As far as we know KDD is just a benchmark for IDS, so far, many people have researched where we came across many different algorithms ranging from decision to prediction. Almost all the parameters were used to identify which algorithm would be good for a particular metrics. We came across 42 parameters in KDD dataset. Using machine learning large amount of data to give statistical results and work can be done quickly. IDS was used to identify either the activity is malicious or non-malicious.***

Keywords—IDs intrusion detection system, KDD cup 1999 datasets, Idps intrusion detection and prevention system, Computer and network security.

Introduction:

IDS stands for intrusion detection system. It is a system that monitors network traffic for suspicious activity and issues alerts when such activity is discovered. While anomaly detection and reporting is the primary function, some intrusion detection systems are capable of taking actions when malicious activity or anomalous traffic is detected, including blocking traffic sent from suspicious IP addresses. Although intrusion detection systems monitor networks for potentially malicious activity, they are also prone to false alarms (false positives). Consequently, organizations need to fine-tune their IDS products when they first install them.That means properly configuring their intrusion detection systems to recognize what normal traffic on their network looks like compared to potentially malicious activity. An intrusion prevention system (IPS) also monitors network packets for potentially damaging network traffic. But where an intrusion detection system responds to potentially malicious traffic by logging the traffic and issuing warning notifications, intrusion prevention systems respond to such traffic by rejecting the potentially malicious packets.

Literature Survey:

Amri Danades, DeviePratama, Dian Anggraini, DinyAnggriani performed analysis on classification water quality in which they used K-Nearest Neighbour, SVM. Testing was done using 10-fold cross validation. The experimental result showed SVM having higher accuracy of 92.26 as compared to KNN which had accuracy of 71.28% in classifying quality of water.[2]

Preeti Aggarwal and Sudhir Kumar Sharma presented a comparative analysis of KDD data set with respect to four classes which are Basic, Content, Traffic and Host. The analysis is done with 2 important metrics Detection Rate and False Alarm Rate. Random Tree Algorithm was used to test dataset in Weka. This stud helped to understand to obtain higher accuracy DR should be high and FAR should be low.[6]

Nabila Farnaaz and M.A. Jabbar have built a model for IDS using Random Forest to identify and notify the activities as normal or anomaly .Experiment was conducted on NSL-KDD data set. RF was used to detect 4 types of attack like DOS,probe,U2R and R2L 10cross validations were applied. Experimental results proved RF had higher accuracy then J48.[4]

Mohammad Almseidin, MaenAlzubi, Szilvester Kovacs, MouhammdAlkasassbeh put forward an experiment by testing KDD dataset of IDS with various Machine Learning algorithms which were J48,Random Forest , Random Tree, Decision Table, MLP , Naïve Bayes and Bayes Network. Rate of attacks are approximately 79% of DOS,19% of Normal and 2% other. Experiment Demonstrated there is no single algorithm which can handle efficiently all the types of attack. Decision Table achieved lowest false negative rate of .0002, Random Forest registered highest accuracy rate of 93.77% ,Naïve Bayes execution was the best as compared to other algorithms.[8]

Anurag Jain, Bhupendra Verma and J. L. Rana Selecting right algorithm with accordance to IDS. With the help off KDD-99 Dataset which is found to be improved during research was used and feature elimination and feature selections were used to reduce and get more relevant features directly. After going through and understanding these many algorithms for classifier selection. Model evaluation and discussion is done to get best TP & Worst TP rates.[12]

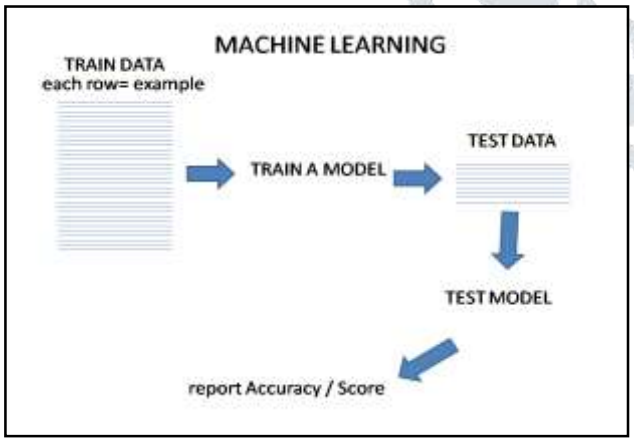
Dr. MalwanBahjatAbdulrazaq, Azarabidsalih used KDD dataset for comparative study of different algorithms. They used 60% for training and 26% for testing. Experimental results were based on following 3 algorithms J48,KNN,Naïve Bayes .The accuracy of

Classification gets best result when Pob class is used with DT while R2l and U2R used with KNN and DOS used with NB. The Naïve Bayes classifier underperform and gives less accuracy but it is faster as compared to other algorithms, KNN is slower as it takes more time to built and test data.[9]

Shikha Agrawal , Jitendra Agrawal made survey on anomaly detection using data mining techniques . In the present world huge amounts of data are stored and transferred from one location to another. The data when transferred or stored is primed exposed to attack. Although various techniques or applications are available to protect data, loopholes exist. Thus to analysedata and to determine various kind of attack data mining techniques have emerged to make it less vulnerable So,they did check on anomaly detection with this 3 ways –Classification,Clustering & Hybrid Approach and finally listed down their methodology and their pros and cons.[13]

Methodology:

Implementation



The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide

Data Set:

Cleaning and filtering of the data set is done to remove duplicate records, normalize the values, accounting for missing data and removing irrelevant data items.

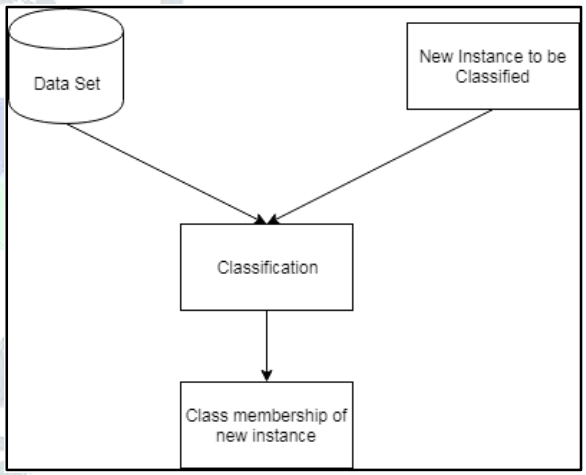
The training data set is provided to the classifier as inp ut. This classified data is also used for the purpose of testing. We used the CNN algorithm

The system will operate mainly in two stages:

- Training phase
- Testing phase

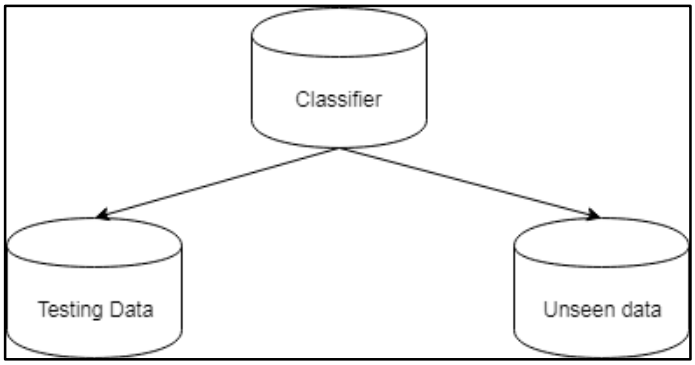
Training Phase

Classification assumes labeled data: we know ho w many classes there are, and for each class we have examples (labeled data).



- **Testing Phase**

Testing phase involves the prediction of unknown data sample



System Architecture:

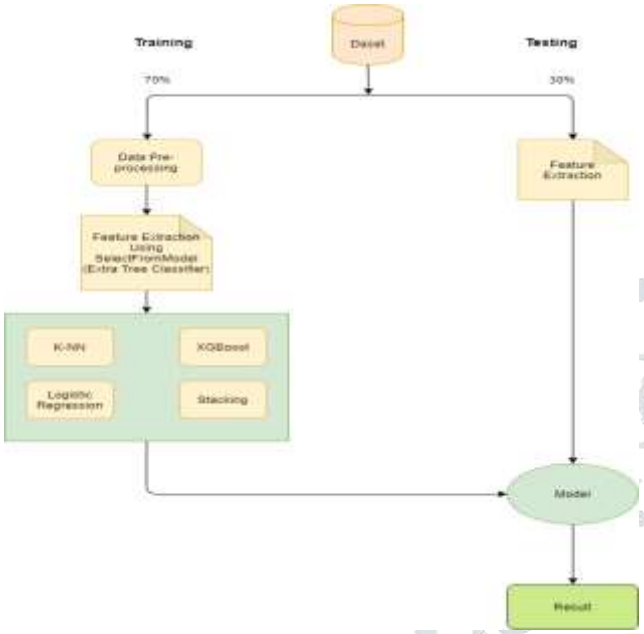
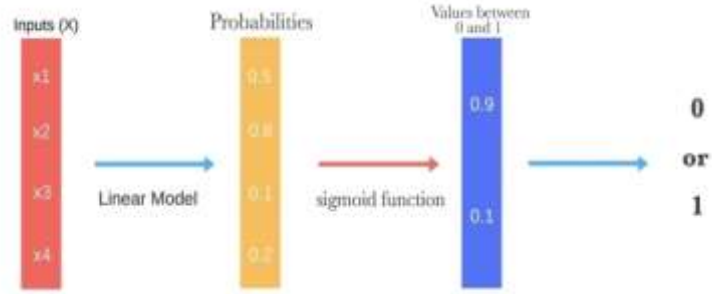


Fig displays the architecture of this thesis. The dataset used is divided into two parts. 70% of the dataset is used for training while 30% is used for testing. The training data is subjected to pre-processing. The pre-processing involves removal of garbage values and null values in the dataset. After pre-processing feature extraction is performed on the dataset. SelectFromModel is used along with Extra-tree classifier for extracting main features from the dataset. After feature extraction is performed the dataset is subjected to the algorithms for the training purpose and development of the model. The model developed will be used for testing purpose. Feature extraction is performed on testing data and then the dataset is fed to the model developed during training phase. Based on its training the model predicts the final results of whether there is a network intrusion performed or not.

Logistic Regression:

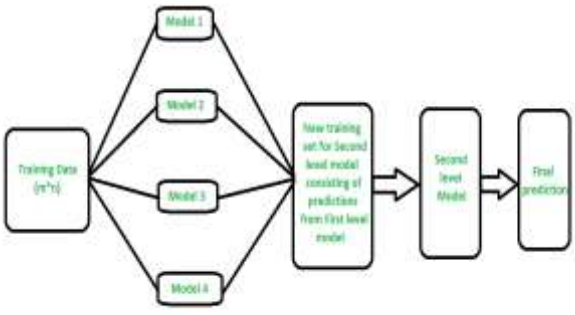
Logistic regression is used for making a binary classification. For exampleidentifying whether an email is spam or not, whether a tumor is malignant or not. That is, Logistic regression works well for classification between 0 and 1 kind of problems. Logistic regression makes use of the logistic function there by estimating the underlying probability for identifying the relation between a dependent and multiple independent variable. For making the final prediction the probability values are transformed in to binary measures. This is achieved with the help of a sigmoidal function also referred to as logistic function.

The sigmoidal function can take any value ranging between 0 and 1 as input. These values are transformed into binary values as 0 and 1 by sigmoidal function.



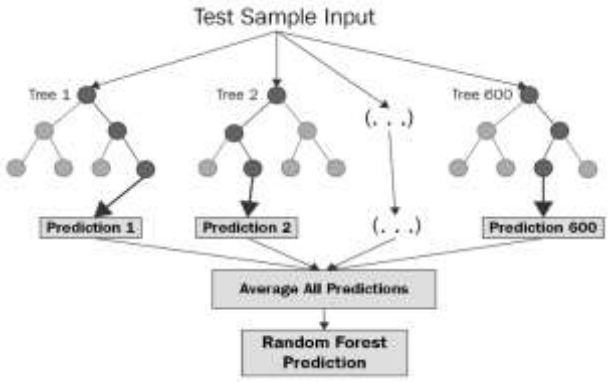
Stacking Algorithm:

Stacking is an ensemble-based machine learning algorithm. There are different kinds of ensemble/hybrid algorithms. The most famous form of ensemble are bagging and boosting. Bagging consists of multiple weak classifiers to operate in an isolation. The result of each classifier is combined in the end to obtain the final result

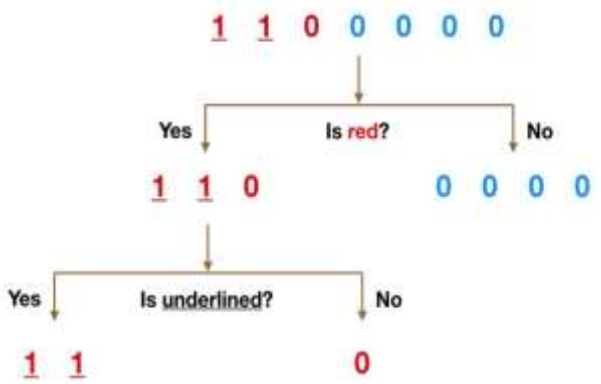


Random Forest:

Random Forest is a supervised form of machine learning algorithm. Random forest is generated by collaboration of multiple decision trees. Each decision operates independently the results obtained by each decision tree is finally averaged to generate a final result.



In case, of random forest algorithm each tree involved are trained individually and parallelly. These trees do not interact with each other during training phase. The results of all the trees are combined and a mean is obtained which serves as the final output for a random forest Regressor model.



The above figure shows an example of a decision tree. As we can see from the figure there are two leaf nodes present. That means two decision are made by the algorithm to reach a conclusion. At leaf node one the first decision is made, i.e whether the given data is red or blue. If the numbers are red then the second decision is made, whether the numbers are underlined or not.

Results:

Confusion Matrix**:**

Confusion Matrix is a matrix created on basis of the result obtained by machine learning classifier. Confusion matrix is also sometimes referred to as the error matrix. The matrix is used to describe the overall performance of the classifier in terms of classification. It provides an easy classification between different classes of the result section. The four boxes of confusion matrix represents the counts of correctly and incorrectly classified data. The confusion matrix represents the actual state of a classifier while making the prediction.

	Class 1 Predicted	*Class 2 Predicted*
Class 1 Actual	**TP**	**FN**
Class 2 Actual	**FP**	**TN**

TP (True Positive): The actual data point is positive and is predicted positive by our classifier.

TN (True Negative): The actual data point is negative and is predicted negative by our classifier.

FP (False Positive): The actual data point is negative and is predicted positive by our classifier.

FN (False Negative): The actual data point is positive and is predicted negative by our classifier.

Accuracy:

Accuracy is defined by the number of correct predictions done by our classifier divided by the total prediction done by the classifier.

Precision:

Precision is the measure of true positive obtained during the classification from all the positive results existing in the dataset.

Recall:

Recall is the measure of data points which are predicted as positive from all the prediction which are termed as positive by the classifier.

F1-score:

F1-score is value obtained by using both precision and the recall. F1-score is a harmonic mean of precision and recall.

Conclusion:

For this thesis we have used the dataset of size about 82,332. Out of it 57632 has been used for training the algorithms while 24700 is be used for testing purpose. Four machine learning models were buildXGBoost classifier, K-NN classifier, Logistic regression classifier, Stacking classifier. The Stacking classifier which is a hybrid algorithm produced best results as per as the prediction made by this thesis in the beginning as compared to other general classifiers. Hence, from the implementation of this thesis we can conclude that hybrid algorithm can produce a boost in the accuracy when compared with general machine learning algorithms.

Future Scope:

The Overall accuracy obtained is still not very great. Some different combination of algorithms can be used with different type of Hybrid approach like bagging or boosting to improve the overall accuracy. Model developed in this thesis can be transformed into a proper software that can run live with system to OS to continuously look for any ATP intrusion. A mobile version of the software can be developed for protection of android devices.

References:

1] Yihua Liao, V.RaoVemuri " Using K-Nearest Neighbor Classifier for Intrusion Detection", Oct 2002

[2] AmriDanades, DeviePratama, DianAnggraini, DinyAnggriani "Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status" in 2016 IEEE 6th Conference on System Engineering and Technology October 3-4,2016 Bandung-Indonesia

[3] GaganjotKaur, A Mit Chhabra "Improved J48 Classification Algorithm for Prediction of Diabetes" International Journal of Computer Application(0975-8887) Volume 98-No.22,July 2014

[4] Nabila Farnaaz and M.A.Jabbar "Random Forest Modeling for Network Intrusion Detection System" in Twelfth International Multi-Conference on Information Processing -2016(IMCIP-2016)

[5]G.V.Nadiammai , M.Hemalatha "Effective approach toward Intrusion Detection System using data mining techniques" 22 May 2013

[6]PreetiAggarwal,Sudhir Kumar Sharma "Analysis of KDD Dataset Attribute-Class wise for Intrusion Detection" in 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015)

[7]Okfalisa,Mustakim,IkbalGazalba,NurulGayatri Indah Reza "Comparative Analysis of K-Nearest Neighbor and Modified K-nearest Algorithm for Data Classification" in 2017 2nd International Conferences on Information Technology,Information Systems and Electrical Engineering(ICITISEE)

[8]MohammadAlmseidin,MaenAlzubi,Szilvester Kovacs ,MouhammdAlkasassbeh "Evaluation of Machine Learning Algorithms for Intrusion Detection Sysytem"

[9]Dr.MalwanBahjatAbdulrazaq,Azarabidsalih"Combination of Multi Classification Algorithm for Intrusion Detection System" in International Journal of Scientific & Engineering Research,Volume 6,Issue 1,January-2015

[10]RajeshWankhede,VikrantChole,Shruti Kolte "A Review on Intrusion Detection System using Classification Technique" in International Journal of Advanced Computational Engineering and Networking ,ISSN:2320-2106,Voulme-3,Issue-12,Dec-2015

[11]David Ahmad Effendy,KusriniKusrini,SudarmawanSudarmawan "Classification of Intrusion Detection System (IDS) Based on Computer Network" in 2017 2nd International Conferences on Information Technology

[12]AnuragJain,BhupendraVerma and J.L.Rana "Classifier Selection Models for Intrusion Detection System" in Informatics Engineering,an International Journal(IEIJ),Vol.4,No.1,March 2016

[13]Shikha Agrawal, Jitendra Agrawal "Survey on Anomaly Detection using Data Mining Techniques " In 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

[14]I.Dhanabal, Dr.S.P.Shantharajh "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithm" in International Journal of Advanced Research in Computer and Communication Engineering Vol.4,Issue 6,June 2015

[15]AbiramiSivaprasad, Neha Ghawalkar, Srushti Hodge, MaitriSanghavi, Vidhya Shinde "Machine Learning based Traffic Classification using Statistical Analysis".Volume 6 March 18 Volume 6 Issue 05-04-2018