PREDICTION OF NETWORK ATTACTS USING SUPERVISED MACHINE LEARNING ALGORITHM

Dr. K. Sridharan[1], Yogesh N[2], Ganajhaala Praveen G[3], Bennyhinn Joshua S[4], Sai Krishna CH[5]

[1]Associate Professor, Department of Information Technology, Panimalar Engineering College, Anna University, Chennai.

[2,3,4,5]Department of Information Technology, Panimalar Engineering College, Anna University, Chennai.

---***---

***Abstract* -** Human pose estimation is an important problem in the field of Computer Vision. Now-a-days, this world is more on automation, we make capturing all the activities in our surroundings using surveillances and cameras. It is difficult for computer to determine their poses for the analysis process. Pose Estimation is predicting the body part or joint positions of a person from an image or a video. This technology will have huge implications. Applications may include video surveillance, assisted living, advanced driver assistance systems and sports analysis. Human are flexible they can change their poses frequently. To analysis the human movement positions we use Generative Adversarial Network (GAN) which is an unsupervised machine learning algorithm. GAN can be trained to generate images fromrandom noises. GAN contains generator and discriminator in which generator generates the fake samples using noises and the discriminator tries to classify the fake and real images. GAN has the objective to produce a complex output from a simple input.

KeyWords - Machine learning-Classification method, python, Prediction of Accuracy result.False Negatives (FN), True Positives (TP), True Negatives (TN).

1. INTRODUCTION

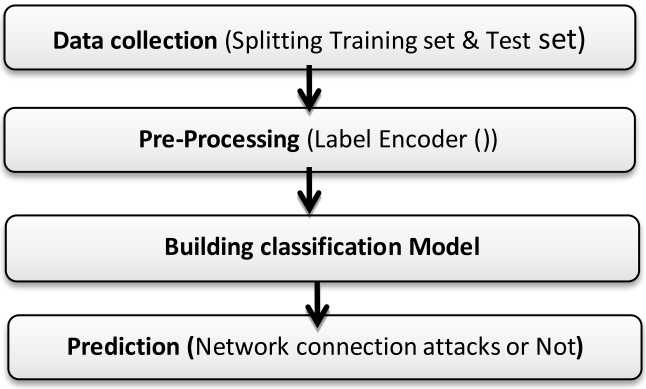
Due to the large volumes of data as well as the complex and dynamic properties of intrusion behaviors, data mining, based Intrusion Detection techniques have been applied to network-based traffic data. With recent advances in computer technology large amounts of data could be collected and stored. Machine Learning techniques can help the integration of computer-based systems in the network environment providing opportunities to facilitate and enhance the work of network security experts. It ultimately improves the efficiency and quality of data and information. Network Intrusion Detection aims at distinguishing the behavior of the network. This paper presents the implementation of four supervised learning algorithms, C4.5 Decision tree Classifier (J48), Instance Based Learning (IBK), Naive Bayes (NB) and Multi layer Perceptron (MLP) in WEKA environment, in an Offline environment. The classification models were trained using the data collected from Knowledge Discovery Databases (KDD) for Intrusion Detection. The trained models were then used for predicting the risk of the attacks in a web server environment or by any network administrator or any Security Experts. The Prediction Accuracy of the Classifiers was evaluated using K-fold Cross Validation and the results have been compared to obtain the accuracy. It have to find Accuracy of the training dataset, Accuracy of the testing dataset, Specification, False Positive rate, precision and recall by comparing algorithm using python code.

2. CLASSIFICATION OF ATTACKS

The data set in KDD Cup99 have normal and 22 attack type data with 41 features and all generated traffic patterns end with a label either as‘normal’ or any type of ‘attack’ for upcoming analysis. There are varieties of attacks 3 which are entering into the network over a period of time and the attacks are classified into the following four main classes.

- Denial of Service (DoS)
- User to Root (U2R)
- Remote to User (R2L)
- Probing

3. BUILDING THE DATA MODEL


Data collection (Splitting Training set & Test set)
Pre-Processing (Label Encoder ())
Building classification Model
Prediction (Network connection attacks or Not**)**

4. SYSTEM ARCHITECTURE

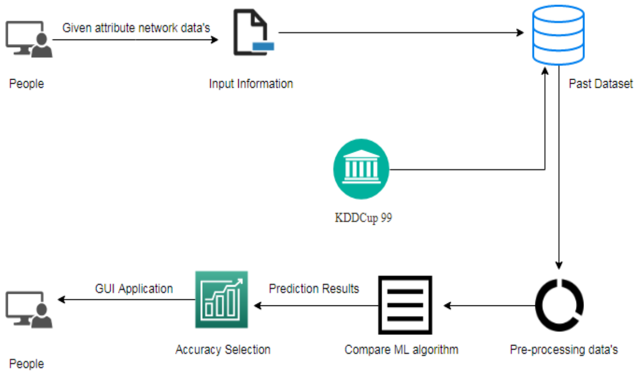
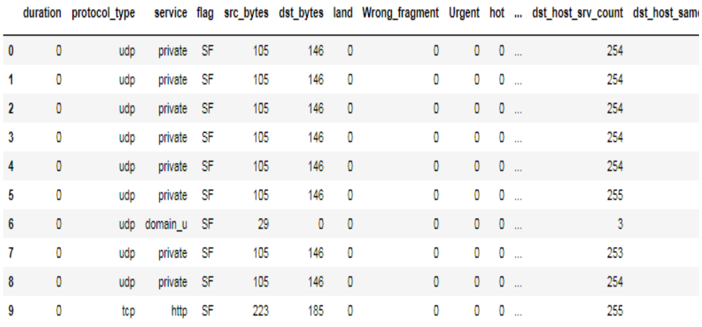


Fig 4.1 - Architecture Diagram

5. DATA VALIDATION

5.1 *Variable Identification Process :*

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques.The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.



	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	Wrong_fragment	Urgent	hot	...	dst_host_srv_count	dst_host_sam
0	0	udp	private	SF	105	146	0	0	0	0	...	254	
1	0	udp	private	SF	105	146	0	0	0	0	...	254	
2	0	udp	private	SF	105	146	0	0	0	0	...	254	
3	0	udp	private	SF	105	146	0	0	0	0	...	254	
4	0	udp	private	SF	105	146	0	0	0	0	...	254	
5	0	udp	private	SF	105	146	0	0	0	0	...	255	
6	0	udp	domain_u	SF	29	0	0	0	0	0	...	3	
7	0	udp	private	SF	105	146	0	0	0	0	...	253	
8	0	udp	private	SF	105	146	0	0	0	0	...	254	
9	0	tcp	http	SF	223	185	0	0	0	0	...	255	

Fig 5.1 - Sample of given Data Frame

The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. (For example to show the data type format of given dataset).

5.2 Cleaning / Preparing Process :

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

5.3 *Data Pre-Processing* :

Pre-Processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

6. PERFORMANCE MEASUREMENT OF EACH ATTACK

This topic focuses on measuring the performance of each algorithms on each attacks specifically. We apply all the algorithms to each attack, and measure the performance of these algorithms for each attack. So that we can determine the effective algorithm to find the type of the attack .

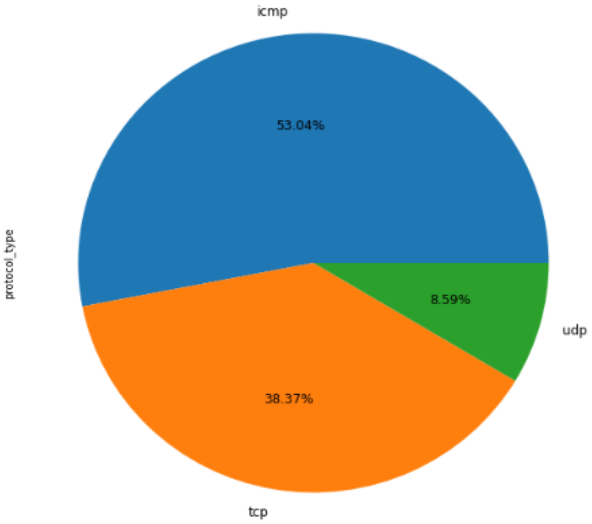


Fig 6.1 - Performance chart of Algorithms.

7. PERFORMANCE MEASUREMENT OF OVERALL ATTACKS

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

6.1 Comparing Algorithm with prediction in the form of best accuracy result:

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

8. PREDICTION OF RESULT BY ACCURACY

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct. (When the model predicts default: how often is correct?)

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

Recall = TP / (TP + FN)

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula: F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula: F1 Score = 2*(Recall * Precision) / (Recall + Precision) .

9. GUI BASED PREDICTION

Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface and Tkinter will come with Python as a standard package, it can be used for security purpose of each users or accountants.

10. ALGORITHMS USED

In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

✧ ***Logistic Regression***

✧ ***Decision Tree***

✧ ***K-Nearest Neighbor (KNN)***

✧ ***Random Forest***

✧ ***Naive Bayes algorithm***

✧ ***Support Vector Machines***

11. CONCLUSION AND FUTURE WORKS

11.1. Conclusions : The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracyonpublic testsetishigher accuracyscorewillbe findoutby comparing each algorithm with type of all network attacks for future prediction results by finding best connections. This brings some of the following insights about diagnose the network attack of each new connection. To presented a prediction model with the aid of artificial intelligence to improve over human accuracy and provide with the scope of early detection. It can be inferred from this model that, area analysis and use of machine learning technique is useful in developing prediction models that can helps to network sectors reduce the long process of diagnosis and eradicate any human errors.

11.2. Future Works :

➢ Network sector want to automate the detecting the attacks of packet transfers from eligibility process (real time) based on the connection detail.

➢ To automate this process by show the prediction result in web application or desktop application.

➢ To optimize the work to implement in Artificial Intelligence environment.

12. REFERENCES

1. P. Zheng and L.M. Ni, Smart phone & next generation mobile computing. San Francisco, CA: Morgan Kaufmann, 2006.

2. Noor AzahSamsudin, Shamsul Kamal Ahmad Khalid, MohdFikryAkmalMohdKohar, ZulkifliSenin, Mohd Nor Ihkassan, University Tun Hussein Onn Malaysia, UTHM Parit Raja, Malaysia, 2008.

3. Patel Krishna M., Patel Palak P., Raj Nirali R., Patel LatilA.Itm Universe, Jarod, Vadodara, 2015

4. AshutoshBhargave, NiranjanJadhav, Apurva Joshi, PrachiOke, Prof. Mr. S. R Lahane :Kigital. Ordering System For Restaurants Using Android, International Journal Of Scientific And Research, Publications, Volume 3,Issue 4, April 2013 I ISSN 2250-3153.

5. ReshamShinde, Dept. Of Computer Science & Engineering, Dr. BabasahebAmbedkar College OfEng& Research, Nagpur, Maharashtra – India.

6. Madhura M. Joshi, Department Of CSE, J.S.P.M's B.S.I.O.T.R(W), Pune, Pune University.