PORT SCANNING

Dr. Vijay Kumar M V[1], Rashmi P K[2]

[1]Professor, Head of the Department of ISE, Dr. A.I.T, Bengaluru, India
[2]M.Tech, Department of ISE, Dr. A.I.T, Bengaluru, India

ABSTRACT: *Developments in technologies have widely spread and it is having advanced changes. The convention of new technologies gives more advantages to individuals, companies, however, it also issues some glitches against them. e.g., the privacy which is really needed in dome information, security of platforms where we store the data, availability of facts etc. These complications making cyber terrorism has one of the major issue in world. Cyber terror, which is creating a lot of harms to people and industrials. Which intern threaten public and country security by various groups such as criminal organizations, professional persons and cyber activists. Thus, Intrusion Detection Systems (IDS) have been developed to avoid cyber-attacks. In this study, ANN, RF and support vector machine (SVM) algorithms were used to detect port scan attempts based on the new Kaggle dataset and 98.63%, 99.83% and 72.50 % accuracy rates were achieved respectively.*

Keywords: Cyber terrorism, SVM, RF, ANN

INTRODUCTION

Port scan is very dangerous technique where hacker can find the open door or dummy points to enter into any network. This also help cyber criminals to know open port is available and helps to check whether it is receiving or sending any kind of data. It will help to find whether the company is installed any security device to restrict the entry from outside world.

Though we are able to find the accuracy of the port scan attempt still we get place to enhance it. In the present training data set contains limited number of possibilities and it has used SVM algorithm for processing purpose. In SVM will not able to find he competency levels so many researches are finding the way for classification and regression. In this existing technology still need the many technologies to train, classification and layering approaches to produce the accurate results as expected.

The new system which helps to find and enable the modern network. Here we propose the deep learning methodologies which helps to learn the data fast. Here for solving this we introduce 2 new algorithm models like ANN and RF to compare with SVM.

LITERATURE SURVEY

R Christopher is the Professor at University of new Haven, who has proposed the paper on port scan and how can we defence against them in the year of 2001.

J A Hoagland and others are proposed to paper which helps to automated practically detect the port scan methods in the year 2002.

M Ouadanne and Ibrahimi are proposed on kdd99 data set Management of intrusion detection system based on KDD99 analysis with lda and pca in the year 2017.

Moustafa and J.Slay The significant features of the unsw-nb15 and kdd99 data sets for network intrusion detection system in the year 2015.

S M Almanson and others Addressing challenges for intrusion detection system using navie bayes at 2017.

M C Raja and M A Rabbni Combined analysis of SVM and principal component analysis for ids in the year 2016.

A A Resende and A C and Drummond engaged Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling in the year 2018.

N. Marir and others Distributed abnormal behavior detection approach based on deep belief network and ensemble svm using spark which has IEEE Access in the year 2018.

D. Aksu and others Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm in International Symposium in the year 2018

V. Vapnik and C. Cortes Support-vector networks in Machine learning 1995.

S. M. Almansob and others Addressing challenges for intrusion detection system using naive bayes and pca algorithm in the year 2017

1. METHODOLOGY

Python is the high level language which design in such way it will help to any kind of framework which in turn having attractive features like rich set of libraries, packages so it is also called as portable language.

They are many prerequisites are used here to build this project kaggle data set,SVM model,RF model and ANN model which uses the confusion matrix.

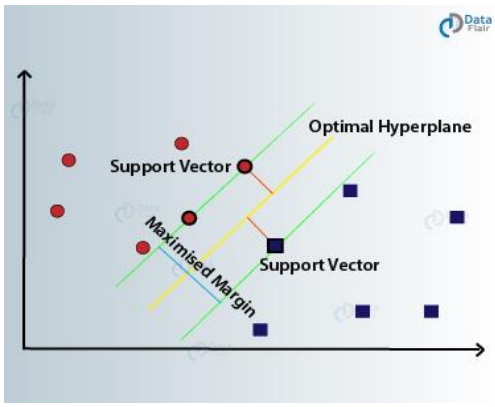
Kaggle dataset

Kaggle allows to find, collect and publish the data set and also helps to build our own data model to train and test data set it helps to solve the many data science and machine learning challenges

Support vector Algorithm

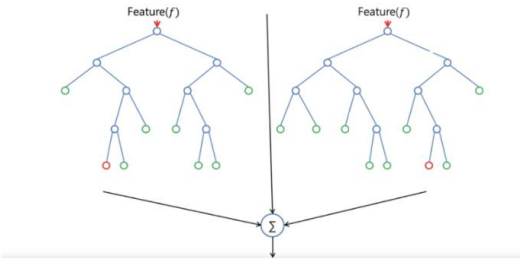
This algorithm works based on the supervised learning it details with both classification and regression.

In the plot each data element pointed as n dimensional space with the value of each feature value of particular coordinate.



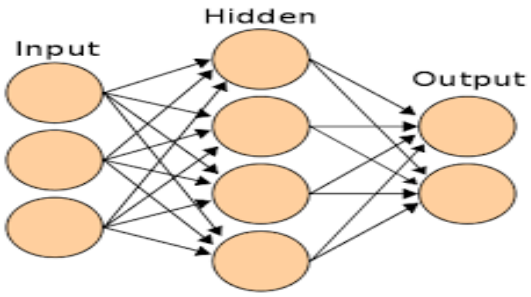
Random Forest Classifier:

Random forest is a flexible this is to very much used in the machine learning area and this process the data even though without the hyper parameter tuning and it always give high performance effect. It is most used algorithm because of its simplicity and it help us to solve the both classification and regression challenges.



Artificial Neural network

It is network of neutrons which helps stimulate the neuron that makes up human brain so that computer will be able the learn the things and provide the decision like human. ANNs are created by programming the computer regularly to behave act and take decision like humans.



2. IMPLEMENTATION

The below Fig 1 will show the working flow model and submodules in this project,

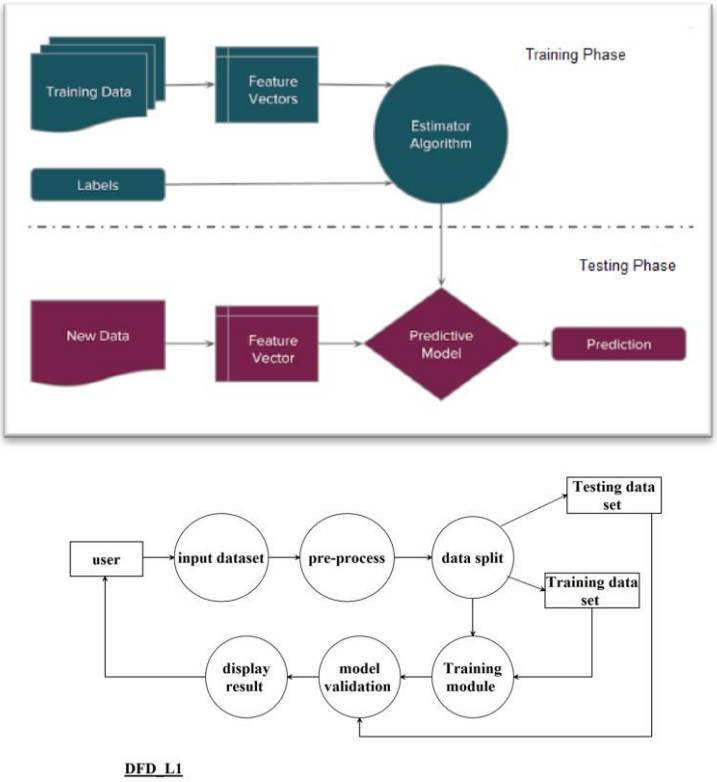


Fig 1: Flowchart for Proposed System

Project Execution steps:

- **Pre Processing of Data**

  Initially need collate and collaborate the data so that it can get process in to deep learning models. All the variables have to convert in to numeric values so that model can access it correctly.

- **Dividing the data set**

  After collecting the data need to divided into training and testing this is major part of operation in building the model.

- **Data Transformation**

  While building the model transformation of the data is more important to make computations are efficient.

- **Building CNN**

  For building computer neural network we need to use keras from there we can import sequential model to initialise the neural network.

- **Running prediction on the Test Set**

To start predicting the values of the results we should use the predict () function.

- **Creating the confusion matrix**

To analyse the correct and incorrect data we need to use the confusion matrix this is also called as error matrix. It has 4 classifiers True Positive(TP),False Positive(FP) ,True Negative(TN) & False Negative(FN)

- **Single Prediction**

Achieving the prediction in single model by using split data set can useful in many ways.

- **Model Accuracy**

After training the data multiple times we will get many variances so for avoiding and getting accurate results we should follow k fold cross validation

- **Over fitting**

Model take time to memorise the result in the training which as taken and it is unable to generalise the result.

- **Tuning**

After all the above steps we need to do the optimisation tuning to give accurate result. Here we consider the Adam optimiser it is kind of replacement optimisation algorithm.

- **Loss Function**

This function that maps an event values of one or more variable on to real number.

- **Matrix Function**

Matrices are used throughout the field of machine learning in the description of algorithm.

- **Optimizer Function**

Optimizers update the weight parameters to minimize the loss function. Optimization is process of searching parameters that minimum and maximum function.

- **Confusion Matrix**

Confusion matrix is combination of prediction results on classification problem. This matrix shows ways in which your classification model is confused when it makes prediction.

SVM Algorithm

Data : Dataset with p^* variables and binary outcome.

Output: Ranked list of variables according to their relevance.

Find the optimal values for the tuning parameters of the SVM model;

Train the SVM model;

$p \leftarrow p^*$;

while $p \geq 2$ **do**

  $SVM_p \leftarrow$ SVM with the optimized tuning parameters for the p variables and observations in **Data**;

  $w_p \leftarrow$ calculate weight vector of the SVM_p $(w_{p1}, \ldots, w_{pp})$;

  $rank.criteria \leftarrow (w_{p1}^2, \ldots, w_{pp}^2)$;

  $min.rank.criteria \leftarrow$ variable with lowest value in $rank.criteria$ vector;

  Remove $min.rank.criteria$ from **Data**;

  $Rank_p \leftarrow min.rank.criteria$;

  p ← p - 1 ;

end

$Rank_1 \leftarrow$ variable in **Data** $\notin (Rank_2, \ldots, Rank_{p^*})$;

return $(Rank_1, \ldots, Rank_{p^*})$

Random Forest Algorithm

For b = 1 to B:

(a) Draw a bootstrap sample Z* of size *N* from the training data.

(b) Grow a random-forest tree to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

i. Select ***m*** variables at random from the *p* variables.

ii. Pick the best variable/split-point among the *m*.

iii. Split the node into two daughter nodes.

Output the ensemble of trees.

To make a prediction at a new point *x* we do:

For regression: average the results

For classification: majority vote

Results Screen shot of ANN

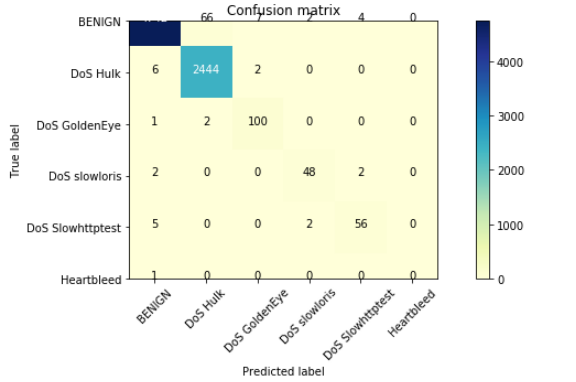
```
print(classification_report(y_true=y_test, y_pred=y_pred_rounded))
```

	precision	recall	f1-score	support
0	1.00	0.98	0.99	4821
1	0.97	1.00	0.98	2452
2	0.92	0.97	0.94	103
3	0.92	0.92	0.92	52
4	0.90	0.89	0.90	63
5	0.00	0.00	0.00	1
accuracy			0.99	7492
macro avg	0.79	0.79	0.79	7492
weighted avg	0.99	0.99	0.99	7492

```
acc_ann = accuracy_score(y_true=y_test, y_pred=y_pred_rounded)*100
print("Accuracy in percentage : %f"%acc_ann)
```

```
Accuracy in percentage : 98.638548
```

```
cm = confusion_matrix(y_pred=y_pred_rounded, y_true=y_test)
plot_confusion_matrix(cm,classes=class_names)
```

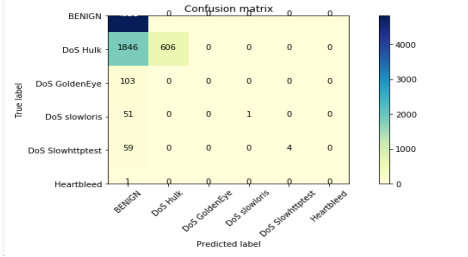


Results Screen shot of SVM

```
print(classification_report(y_test,predictions_svm))
```

	precision	recall	f1-score	support
0	0.70	1.00	0.82	4821
1	1.00	0.25	0.40	2452
2	0.00	0.00	0.00	103
3	1.00	0.02	0.04	52
4	1.00	0.06	0.12	63
5	0.00	0.00	0.00	1
accuracy			0.73	7492
macro avg	0.62	0.22	0.23	7492
weighted avg	0.79	0.73	0.66	7492

```
cm = confusion_matrix(y_pred=predictions_svm, y_true=y_test)
plot_confusion_matrix(cm,classes=class_names)
```



Results Screen shot of RF

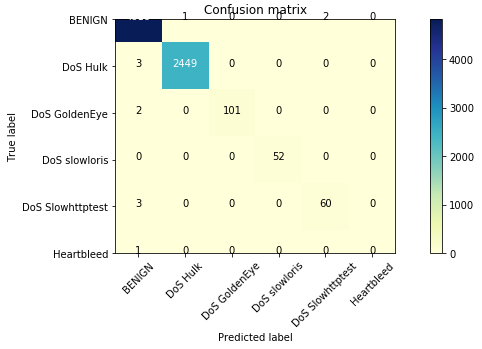
```
print(classification_report(y_test,predictions_rf))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4821
1	1.00	1.00	1.00	2452
2	1.00	0.98	0.99	103
3	1.00	1.00	1.00	52
4	0.97	0.95	0.96	63
5	0.00	0.00	0.00	1
accuracy			1.00	7492
macro avg	0.83	0.82	0.82	7492
weighted avg	1.00	1.00	1.00	7492

```
acc_rf = accuracy_score(y_true=y_test, y_pred= predictions_rf)
print("Overall accuracy of RF model using test-set is : %f" %(acc_rf*100))
```

```
Overall accuracy of RF model using test-set is : 99.839829
```

```
cm = confusion_matrix(y_pred=predictions_rf, y_true=y_test)
plot_confusion_matrix(cm,classes=class_names)
```



VI. CONCLUSION

We have built up the simple and easy to handle gadget framework, to deliver beneficial subordinate and provision for blind and We have tried our best to build up the simple and easy framework to handle port scan. This particular system or the project gone through the development life cycle modules lie design implementation testing and validation and verification steps. Intern it help us to achieve the desired promising results. Outcome of this project specify that it is well organized and exceptional in its competency and performance.

Here we concluding that as result of analysis of both random forest and ANN algorithm modules more accurate and consistent in the area of port scan detection compared to SVM of previous algorithm, hence we are concluding here as these two modules are met the maximum accuracy we can effectively use this algorithm for the application of using this.

REFERENCES

[1] Samson, Martin. in Internet Library of Law and Court Decision retrieved in the year 2021 for the paper Scott Moulton and Network Installation Computer Services.

[2] Poulsen and kevin worked on port scan legal judge says based on security focus in the year 2009.

Michell and Cuier in the university of Maryland they proposed Quantifying the computer security in the year 2009

[3] Shaun,Jamieson worked on The ethics and legality of port scanning in the year 2009

Maimon Uriel 1996 they tried for port scanning without the SYN flag.

[4] M Ouadanne and Ibrahimi are proposed on kdd99 data set Management of intrusion detection system based on KDD99 analysis with lda and pca in the year 2017.

[5] Moustafa and J.Slay The significant features of the unsw-nb15 and kdd99 data sets for network intrusion detection system in the year 2015.

[6] S M Almanson and others Addressing challenges for intrusion detection system using navie bayes at 2017.

[7] M C Raja and M A Rabbni Combined analysis of SVM and principal component analysis for ids in the year 2016.

[8]A A Resende and A C and Drummond engaged Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling in the year 2018.

[9] N.Marir and others Distributed abnormal behavior detection approach based on deep belief network and ensemble svm using spark which has IEEE Access in the year 2018.