Expert Information Prediction Modeling In Pcap Files

S. Leelalakshmi[1], K. Rameshkumar[2]

[1]research Scholar,Bharathiar Universitysleein@Rediffmail.Com
[2]research Guide,Bharathiar University Rameshkumark.Dr@Gmail.Com

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: The cyber security field has more challenges and analysing pcap files for identifying the network traces provides important details. The web traffic consists of different types of transfer. Some data can be malicious and it can fall ***into different*** categories. Analysing and extracting features and applyying machine learning algorithms can prove to be more useful for indentifying the network information. Todays' big data environment provides more way to security threats.

Keywords: PCAP, Security, visualisation, ,Network Component,Prediction

1. Introduction

The ever increasing networks and complexity of applications requires more secured data. Presently security data visualisation[1] plays a very important role in identifying the problems.

ML-based methods can achieve higher prediction accuracy by extracting patterns from web data.

The patterns can help identifying the malicious data

There are several ways Network traffic techniques can be classified[2]

1.NetworkTraffic Techniques Classification

The technique is to classify the collected Network Traffic according to the network protocol and the IP address it belongs.

2.Port Based Technique

This technique is the traditionally used technique. A classification of networks done using well-known port numbers.

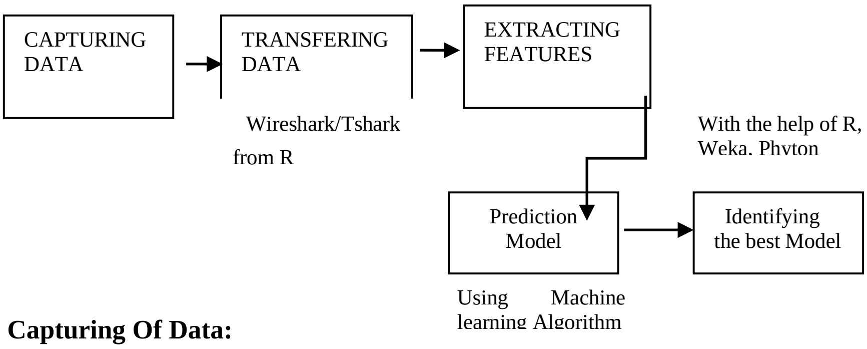
3. Machine Learning Technique Classification

Machine Learning technique which are based on the data set. This technique uses classifiers trained as an input. Then using this trained sample, the unknown classes are classified. Unsupervised and supervised machine learning techniques are the two types of machine learning techniques.

Supervised and Unsupervised machine learning can be used for prediction.

2. Methodology for Analysing the PCAP files

Diagram for the Process


CAPTURING DATA
TRANSFERING DATA
EXTRACTING FEATURES
Wireshark/Tshark
from R
With the help of R, Weka, Phyton
Prediction Model
Identifying the best Model
Using Machine learning Algorithm

Capturing Of Data:

PCAP files are generally difficult to read. They can be made in a readable format with either by converting it to a excel file or by functions which can read directly a PCAP file in R or by Python

Expert information of Wireshark

For the research purpouse normally data will be derived from the previous datasets available.Here the data used was captured live in a normal household computer. As the

Web traffic is more this traffic issues in the home environment itself is a challenge. This is a period where the home environment itself has become like a virtual office. As a novice user of networks if people want to know regarding the malicious software in the system , these paper can provide some information.

The capturing of data will be normally done in tshark or wireshark . In this paper discussed Wireshark will be used for capturing of data.

The packet capture normally results in datas like frame number, timestamp, number of bytes, protocol,source ip, destination ip, source port, destination port,error data etc and so much of information which provides us huge data for analysing. A single data capture can provide so much of data as the analysing itself becomes a big task.

One of Wireshark's strongest capabilities is to identify the network traffic and suggest a reason for it.There is a information called as Expert Information which gives deeper analysis of network, events and problems.

An expert information system can provide information l[3]

Expert Information Entries

Expert data are categorized by level of severity (described below) and contain:

Packet #	**Summary**	**Group**	**Protocol**
443	TCP: 80 → 59322 [RST] Seq=12761 Win=0 Len=0	Sequence	TCP
1202	DNS: Standard query response …	Protocol	DNS
592	TCP: [TCP Out-Of-Order] …	Malformed	TCP

Each data topic has a degree of severity. The levels from the lowest to the highest are used. Wireshark uses various colors seen in paranthesis to mark them.

Chat (blue)
Normal workflow detail, e.g. SYN flagset TCP packet.
Note (cyan)
Notable events e.g. a typical error code such as HTTP 404 has been returned.
Warn (yellow)
Warnings, like a link problem, for example an app returned an odd error code.
Error (red)
Severe concerns, including packets that are malformed.
Apart from these there can be blank spaces also

Expert information objects are classified by category in addition to severity ratings. Information discussed in the paper is only regarding the severity levels and visualisation.

2.1 Visualisation to Understand PCAP files

Here expert information and Protocol with Destination address will be taken for discussion.

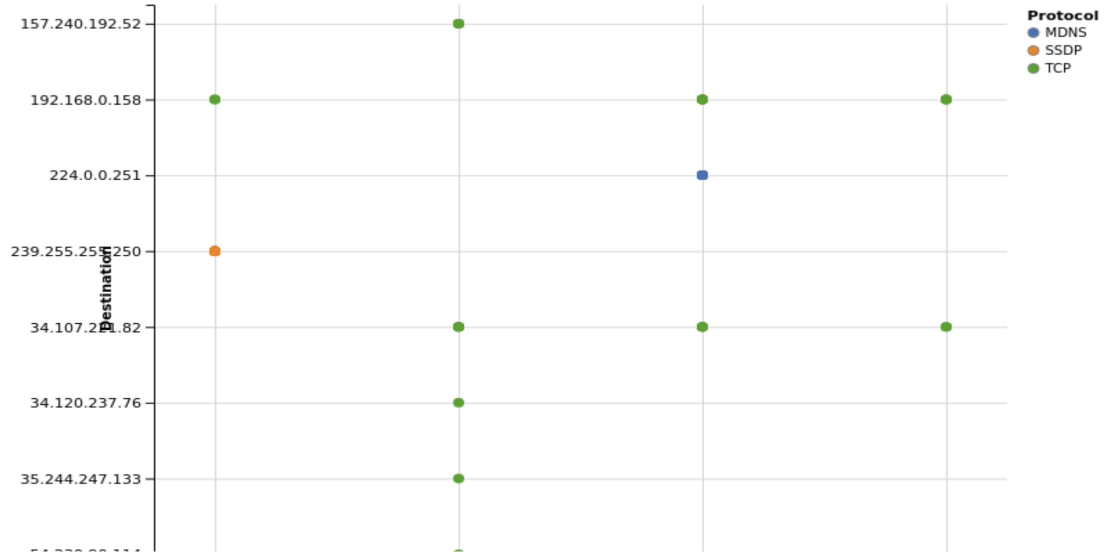
The data set described is a normal dataset which can range from any number of values till we stop the capturing. Here a dataset of 100 values is taken into consideration.

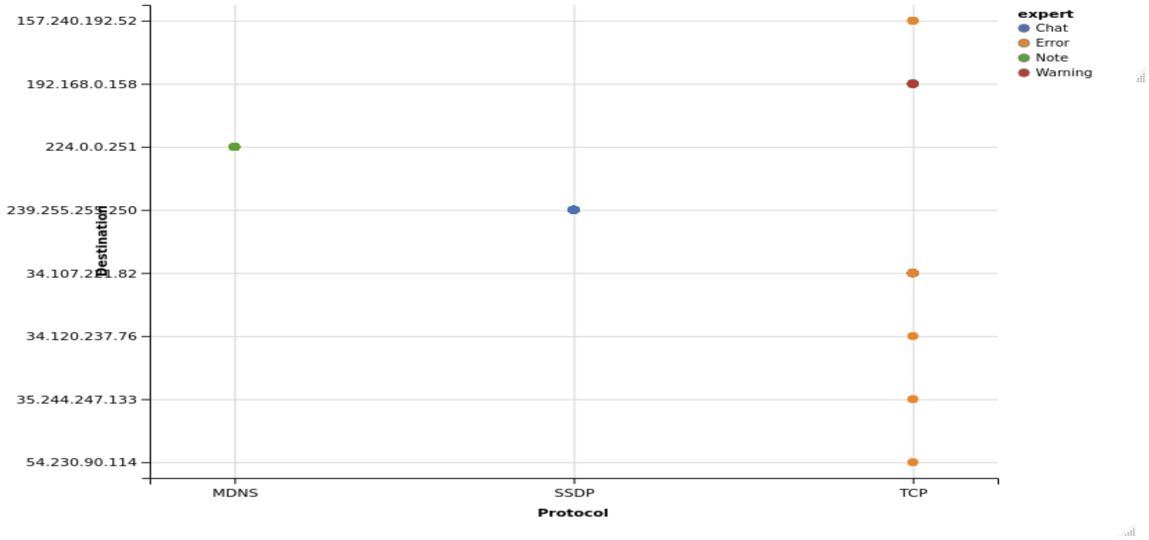
The concept of visualizing information is first an outline, zooming and filtering and the specifics of demand by Ben Schneider man. [4]

Just for understanding a Protocol and Destination plotted with expert information. This particular visualisation helps to view which particular protocol takes the expert information accordingly with destination machines.

Any number of combinations can be checked in the form of graph so that visualisations gives a lot of understanding in the binary formats of pcap.[5]

And also identifying the protocols which falls under the destination machines as well as expert information





2.2 Using classifiers for analysing the PCAP data using WEKA and R

Here is an example of using a Random forest for a classifier tree in weka. Instances here are 626
Scheme:weka.classifiers.trees.RandomForest -I 100 -K 0 -S 1

```
Relation:     vedium1
Instances:    626
Attributes:   7
No.
Time
Destination
Protocol
Length
source
expert
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===


Random forest of 100 trees, each constructed
while considering 3 random features.
```

```
Out of bag error: 0.1087

Time taken to build model: 0.17 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          381                              90.0709 %
Incorrectly Classified Instances        42                               9.9291 %



Kappa statistic                    0.853
Mean absolute error                        0.2205
Root mean squared error                    0.2792
Relative absolute error                 64.4776 %
Root relative squared error           67.5522 %

Total Number of Instances        423
Ignored Class Unknown Instances           203

=== Detailed Accuracy By Class ===
```

Class	Precision	F-Measure	Recall	ROC Area	TP RATE	FP RATE
Note	0.851	0.887	0.927	0.968	0.927	0.067
Error	0.897	0.924	0.953	0.988	0.953	0.047
Warning	0.000	0.000	0.000	0.683	0.000	0.000
Chat	0.948	0.932	0.918	0.96	0.918	0.030
	0.873	0.901	0.873	0.886	0.901	0.045

```
=== Confusion Matrix ===
```

a	b	c	d	<- Classified as
114	7	0	2	a= Note
4	122	0	2	b=Error
5	5	0	4	c= Warning
11	2	0	145	d=Chat

In the WEKA tool before preprocessing the classifier Random forest is applied and provides the following result with 90% correctly classified instances.

3. Feature extraction and reduction

There are several attributes in the PCAP file .But here there is no need of many such attributes,Since this is PCAP analysis to identify what is the protocols and expert notes relationship. Here taking all protocols into account and using a random forest provides the following result done with R

Class.error

ADwin Config	1.0000000
ARP	0.0000000
DNS	1.0000000
HTTP	1.0000000
ICMP	1.0000000
IGMPv2	1.0000000
IPv6	1.0000000
LLMNR	1.0000000
MDNS	0.1645570
NBNS	1.0000000
OCSP	1.0000000
SSDP	0.0000000
TCP	0.5544554
TLSv1.2	1.0000000

So for deriving at the things the attribute that is going to be associated here are the protocols of MDNS,TCP and SSDB. And since more protocols are reduced we can visualize it in a better manner.

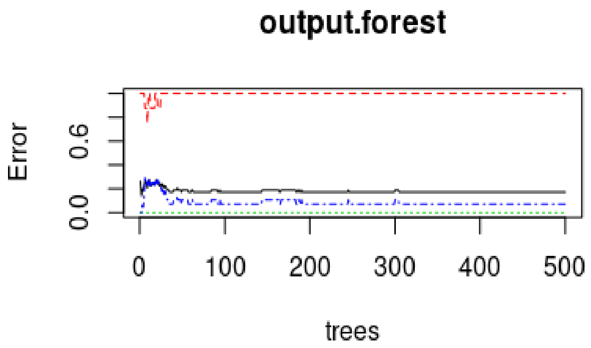
Here the instance number is reduced to 58.

3.1 Applying random forest classifiers for the reduced data set

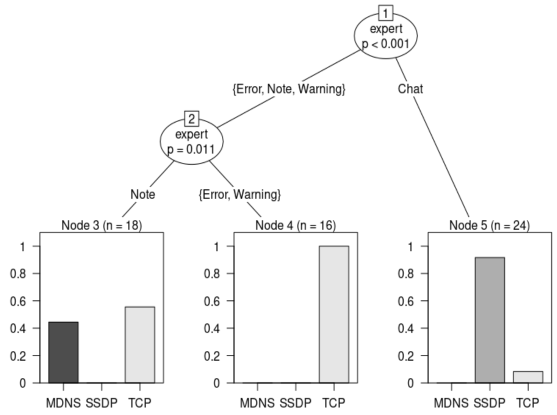
```
randomForest(formula = Protocol ~ expert, data = f)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 1

OOB estimate of  error rate: 17.24%
Confusion matrix:
```

	MDNS	SSDP	TCP	Class.error
MDNS	0	0	8	1.0000000
SSDP	0	22	0	0
TCP	0	2	26	0.0714286



The output forest for error and trees is depicted here.



Applying decision tree classifier for PCAP data.The same can be derived for the decision trees as in the figure above which gives the visualisation aspect in a better manner.

2. 4. Designing and Evaluating the model

In order to train the pcap data for prediction the two random models on random forest and boosted_logistic is selected

Prediction model using random forest

mtry	Kappa	Accuracy
2	0.6114058	0.734357
8	0.6086941	0.732077
14	0.6003884	0.72637

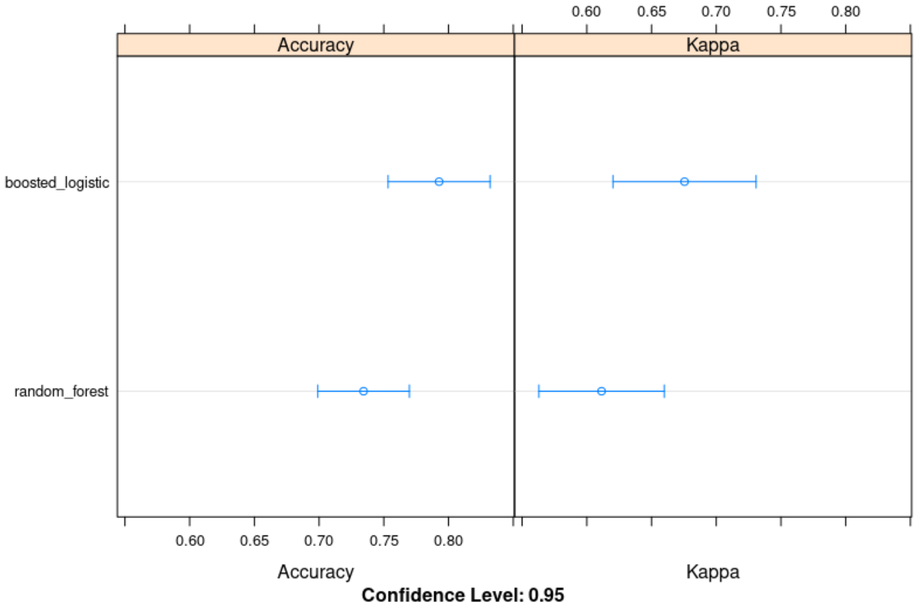
To choose the best model using the highest value, accuracy was used.
Resampling performance for parameters of tuning.
Prediction model using Boosted _logistic

nIter	Accuracy	Kappa
11	0.769007	0.6484441
21	0.767714	0.6449904
31	0.766987	0.6440484

Final value utilized for model was nIter = 11.
Here accurate model comparatively is boosted_logistic

Comparison of the model[5]

In addition to this sample analysis it is also possible to compare the scale and mean accuracy of particular models by drawing a plot of the models' assessment results.



We can see most accurate model is boosted_logistic regression

The variable importance can also be determined

Overall	
Length	5.4721
Destination239.255.255.250	5.2662
source192.168.0.158	3.2883
Destination224.0.0.251	2.0178
source34.107.221.82	1.2513
Destination192.168.0.158	1.1835
source192.168.0.169	1.024
Destination34.107.221.82	0.9665
source192.168.0.119	0.3073
source54.230.90.114	0.2456
source34.120.237.76	0.2396
Destination34.120.237.76	0.2374
Destination54.230.90.114	0.1769
Destination35.244.247.133	0.1422

Conclusion

Modeling the PCAP files for prediction is now happening with Python, R ,Weka and some more applications. By using this models if visualisation and Instrustion detection prediction are done it can be helpful. In this paper some of the attributes in the type of Protocol is selected as well as some of the headers of PCAP[6] files are selected. Careful selection of attributes on any criteria can be helpful in predicting the model. In this paper expert information provides some meaningful result. Searching and predicting a PCAP data is a challenging one. In future developing this can be helpful in predicting the message type a particular node is going to produce.

References

1. Raffel Marty , Applied Security Data Visualisation,1st Edition
2. M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn And F. Abdessamia, "Network Traffic Classification Techniques And Comparative Analysis Using Machine Learning Algorithms," *2016 2nd IEEE International Conference On Computer And Communications (ICCC)*, Chengdu, 2016, Pp. 2451-2455, Doi: 10.1109/Compcomm.2016.7925139.
3. Https://Www.Wireshark.Org/Docs/Wsug_Html_Chunked/Chadvexpert.Htmlshneiderman, Ben. "The Eyes Have It: A Task By Data Type Taxonomy For Information Visualisations." The Craft Of Information Visualisation. Morgan Kaufmann, 2003. 364-371.
4. Https://Medium.Com/@Wanjirumaggie45/Data-Science-For-Good-Machine-Learning- For-Heart-Disease-Prediction-289234651fed
5. Conti, Greg. Security Data Visualisation: Graphical Techniques For Network Analysis. No Starch Press, 2007