Kuveyt Türk Katılım Bankası

Real Time Threat Hunting with Using HTTP Header Information

FACULTY OF ENGINEERING AND NATURAL SCIENCES

Sabancı Üniversitesi

Student(s)
Elif Pınar Ön
Ethem Tunal Hamzaoğlu
Yusuf Sar

Faculty Member(s)
Albert Levi

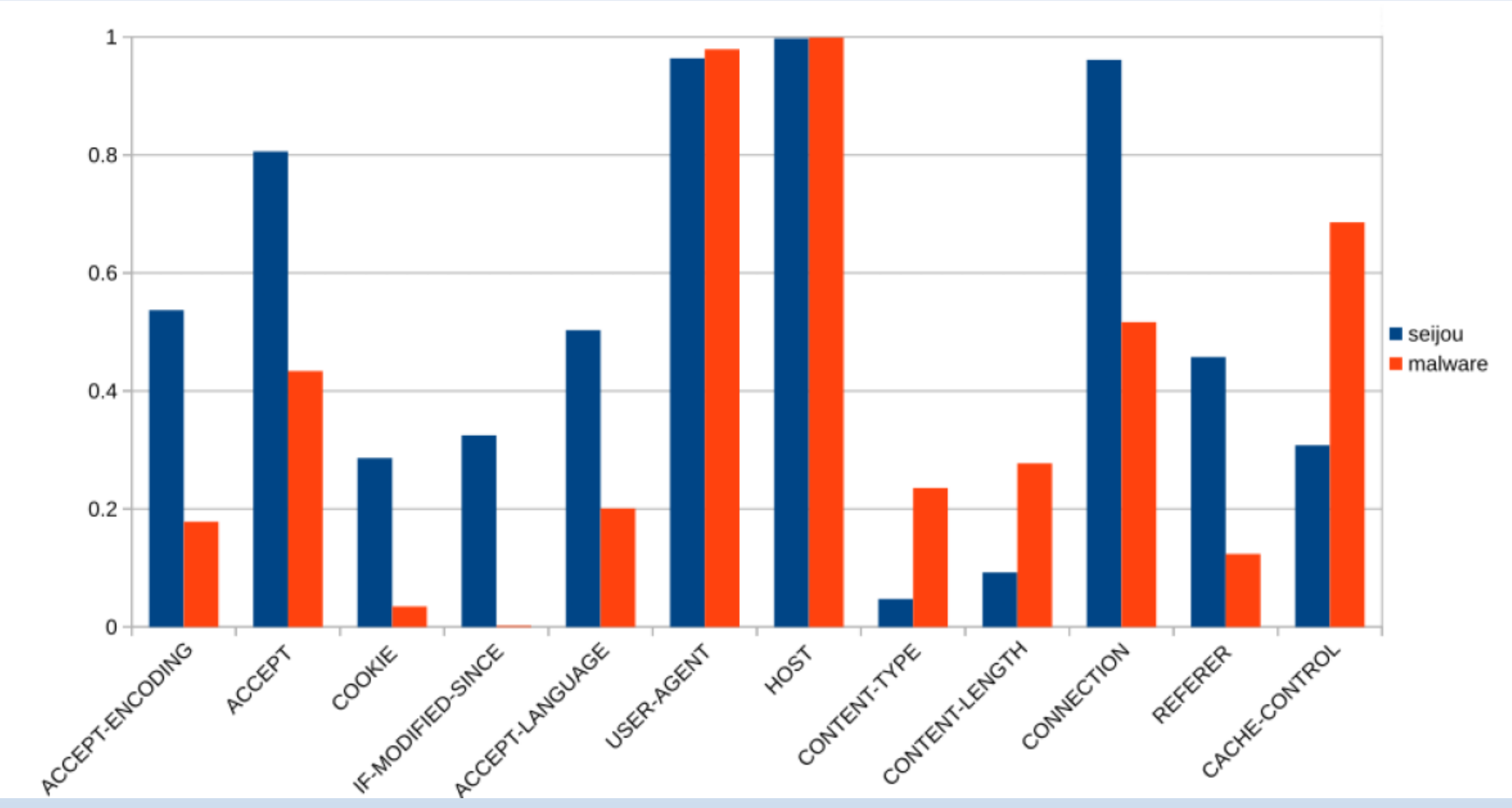
Company Advisor(s)
Ahmet Han
Ferhat Karakoç
Furkan Danış

MOTIVATION

This is an industry focused project sponsored and also the main topic revealed by Kuveyt Türk Katılım Bank. The main goal of this project is to detect malicious activities in real time and threat hunting by using HTTP header information with the machine learning algorithms in corporate networks. The main motivation of this project is to reduce the workforce of the Kuveyt Türk Katılım Bank workers by malicious activities detecting mechanism will be generated to automate the real time threat hunting. Because HTTP headers can be used for malicious purposes, every request sent or received has a HTTP header and it is not possible to check every single one of them by hand.

According to previous researches header information differences between malicious activities and normal activities and the most significant attributes that affects the manner of the HTTP activity are;

"CONNECTION", "ACCEPT", "ACCEPT- ENCODING", "ACCEPT-LANGUAGE", "COOKIE", "CONTENT TYPE", "CACHE-CONTROLS", "IF- MODIFIED-SINCE".



OBJECTIVES

- Is it possible to determine if a client is affected by a malware just looking at it's HTTP request header information?

Sub-Questions:

- Can models be trained to predict that the client is affected by a malware or not?
- Can models be trained to predict the type of the malware that client affected?

Data Collection & Preparation

What do we use?

- Normal-Crime PCAP files
- Bro and modified Bro module
- Python, Anaconda
- Pandas and Scikit-Learn libraries

Unprocessed data in pcap format which are shared public on the Internet are collected from Stratosphere IPS dataset. Four different botnet traffic has been used. These botnets are;

"Neris", "Virut", "Webcompanion" and **"Emotet"**.

Benign HTTP header information also used for classification.

The pcap files collected converted in to log files using Bro tool and modified Bro-Module. This module used for parsing the most significant HTTP Header attributes in to a single column in a log file. To manage collected dataset easily in the implementation part of machine learning algorithms, log files converted in to csv files and these csv files labeled for two different manners, **binary classification** and **multi classification**. Thus, two extra columns added which are indicating malicious or not and botnet type.

Machine Learning Algorithms

Count Vectorizer Method has been used to convert string values in dataset to numeric values. This method is collecting all variables from indicated column, then count and convert them in to integer vectors. This allows algorithms to predict according to the occurrence of the selected most significant HTTP header attributes in events. Only request header information is used when creating models to see if it is enough to predict the malware type and existence.

Three different machine learning algorithms has been used for comparison; **Multinomial Naive Bayes**, **Decision Tree Classifier** and **Random Forest Classifier**.

Two ways of classification has been performed; **binary** and **multi classification**.

For **multi classification** four type of botnet and normal activity data used and each botnet and normal traffic data include 1360 different activity header information. 80% of this data is used for training models and 20% of data is used for testing these models.

For **binary classification** 5440 different malicious and normal activity header information used. 80% of this data is used for training models and 20% of data is used for testing these models.

CONCLUSION

Binary Classification

	Multinomial NB	Decision Tree	Random Forest
True Negative	1100	1101	1101
False Positive	2	1	1
False Negative	0	0	0
True Positive	1074	1074	1074
Accuracy Score	0.99	0.99	0.99

Multi Classification

MultinomialNB

	0	1	2	3	4
0	260	0	0	0	0
1	0	310	0	0	0
2	0	0	263	0	0
3	0	2	0	254	29
4	0	0	0	78	164

Decision Tree

	0	1	2	3	4
0	260	0	0	0	0
1	0	309	1	0	0
2	1	0	262	0	0
3	0	0	0	259	26
4	0	0	1	71	170

Random Forest

	0	1	2	3	4
0	260	0	0	0	0
1	0	310	0	0	0
2	0	0	263	0	0
3	0	2	0	259	26
4	0	0	0	70	172

MultinomialNB

	Precision	Recall	F1-score	Support
Normal	1.00	1.00	1.00	260
Emotet	0.99	1.00	1.00	310
Webcompanion	1.00	1.00	1.00	263
Virut	0.77	0.89	0.82	285
Neris	0.85	0.68	0.75	242
avg/total	0.92	0.92	0.92	1360

Decision Tree

	Precision	Recall	F1-score	Support
Normal	1.00	1.00	1.00	260
Emotet	1.00	1.00	1.00	310
Webcompanion	0.99	1.00	0.99	263
Virut	0.78	0.91	0.84	285
Neris	0.87	0.70	0.78	242
avg/total	0.93	0.93	0.93	1360

Random Forest

	Precision	Recall	F1-score	Support
Normal	1.00	1.00	1.00	260
Emotet	1.00	1.00	1.00	310
Webcompanion	1.00	1.00	1.00	263
Virut	0.79	0.91	0.84	285
Neris	0.87	0.71	0.78	242
avg/total	0.93	0.93	0.93	1360

As result, all Binary Classification Models has resulted in high accuracy scores with only a few miss predictions As False Positive meaning that they predicted malicious when normal which creates less problem than false negative mistakes.

In multi classification models, all models has predicted normal accurately However, they made mistakes when predicting if the botnet type is Virut or Neris. It is probably because these botnets are creating similar http requests.

In general, results of the project has showed that machine learning algorithms can predict malware activities and types highly accurately using only HTTP request information.

The technical and environmeltal needs of proof of concept will be supported by Kuveyt Türk Katılım Bank and generated models will be tested with Kuveyt Türk Katılım Bank's network.

REFERENCES

- Zegers, University of Amsterdam, R. (2015). *HTTP Header Analysis*. Retrieved from https://www.os3.nl/_media/2014-2015/courses/rp2/p91_report.pdf
- Garcia S., Grill M., Stiborek J and Zunino A. "An Empirical Comparison of Botnet Detection Models", Computers and Security Journal, Elsevier. (2014) Vol 45, pp 100-123. Retrieved from http://dx.doi.org/10.1016/j.cose.2014.05.011