Machine Learning Applications for the Cyber Security Threats

HGCI Summit 2017

Muhammad Najmi Ahmad Zabidi

28th Nov 2017

About

Work as a Linux Consultant (Contractor)

- Work at home, remotely for End Point Corporation
- This research was an academic work, now I still keep and enchance it whenever possible so that it will not go obsolete
- Earned a Master degree in Computer Science from USM Penang (2007), a Bachelor degree from UIAM Gombak (2004)

State of the cyber security problems

- From generic threat (mass scan, intrusion, malware) to organized, sponsored and customized threat (APT- Advanced Persistent Threat)
- We could reduce the analysts work as much as possible to *reduce* threat and automate them
- Many attack vectors that we have to consider

Machine Learning at Large

- machine learning being used in almost every aspect in life
  - hand writing recognition
  - e-commerce sites - product suggestion
  - camera - face recognition
  - camera - speed trap!

Academic Research in Cyber Security

- notably Intrusion Detection System (IDS) topics before 2010
- many on Windows malware around 2000-2015
- later research moved to Android malware from 2005-now
- recent trends attempt to focus on cloud security as well as blockchain

Datasets

- IDS (focuses on the network stream)
  - IDS researches usually use MIT Lincoln 1999 dataset
  - recently researchers attempt to introducte a more recent, better dataset to replace it with the inclusion of new protocol, for e.g SIP
- Malware (focuses on the host level)
  - No standard dataset exists to date
  - Datasets exists, but not for long (due to many reasons)
  - Malware being used for analysis ranged from 500 samples to millions
  - We probably could use dataseet from Kaggle
    `https://www.kaggle.com/c/malware-classification`

Malware's taxonomy

- Malware is a software so it has certain traits which make it different compared to the benign software (detection)
- We can also use these traits to differentiate it from its variance (classification)
    - For e.g : Conficker A, Conficker B and Conficker C

Network stream based threats vs host

- Attack like DDoS, mass scan - network based attack
- Malware (APT, virus, ransomware) - host based attack

I will focus on the malware

- Datasets could be obtained from the wild (honeypot sensors, filesharing service, Vxhaven)
- There are publicly accessible malware dataset, but since it’s hosted by researchers from a department - the dataset is gone when the dept got reorganized

Analyzing data

- Preprocessing (data filter, data massage)
- Feature Selection
- Feature Reduction
- Learning (Classification, Clustering)

Malware

- Malware is a software
- We need to extract data from the software binary
- We could use *static* analysis or *dynamic* analysis
- *Static* involves reverse engineering while *dynamic* analysis involves the malware behavior monitoring (at least)

Quality of the data

- Static analysis
  - Slow, but many features could be obtained
- Dynamic analysis
  - Fast and could be automated with many ways
  - There is probability to miss some features, the data capture need to be planned carefully

Where Machine Learning Could Play its Role?

- Data could be labelled non no labelled
- Use classification for labelled data, clustering for non labelled data

Platform

- Initially worked on a Ubuntu machine (laptop)
- Later, I worked on my Ubuntu desktop - Intel i3, 16GB RAM
- Also, I applied and received Amazon research grant - twice (in usage credit) from Amazon which I used for AWS' EC2

Related Tools for Machine Learning

- Weka
- Python Scikit-learning
- Built in libraries for any programming language or write our own

Dataset which I used

- I used two datasets from Malaysia based honeypot sensors (here I call it Dataset A, 3000-ish malware) and publicly available dataset from a foreign university (Dataset B, which is around 500-ish malware)
- Both have different nature of features

Classification

- Malware detection is a binary classification (two groups, malware and non malware (benign)
- Malware *family* classification is the next step, categorizing the families of the verified malware

Clustering

- This is NOT a ”computer cluster” as the system admin knows
- Clustering in machine learning involve grouping of unknown items into its own group
- Useful for zero-day attacks/malware

Machine learning algorithms

- K-means
- Decision trees
- Random Forest
- SVM
- etc

Sequence based algorithms that I used

- N-gram
- LCS

Sequence aligning, a dynamic programming method

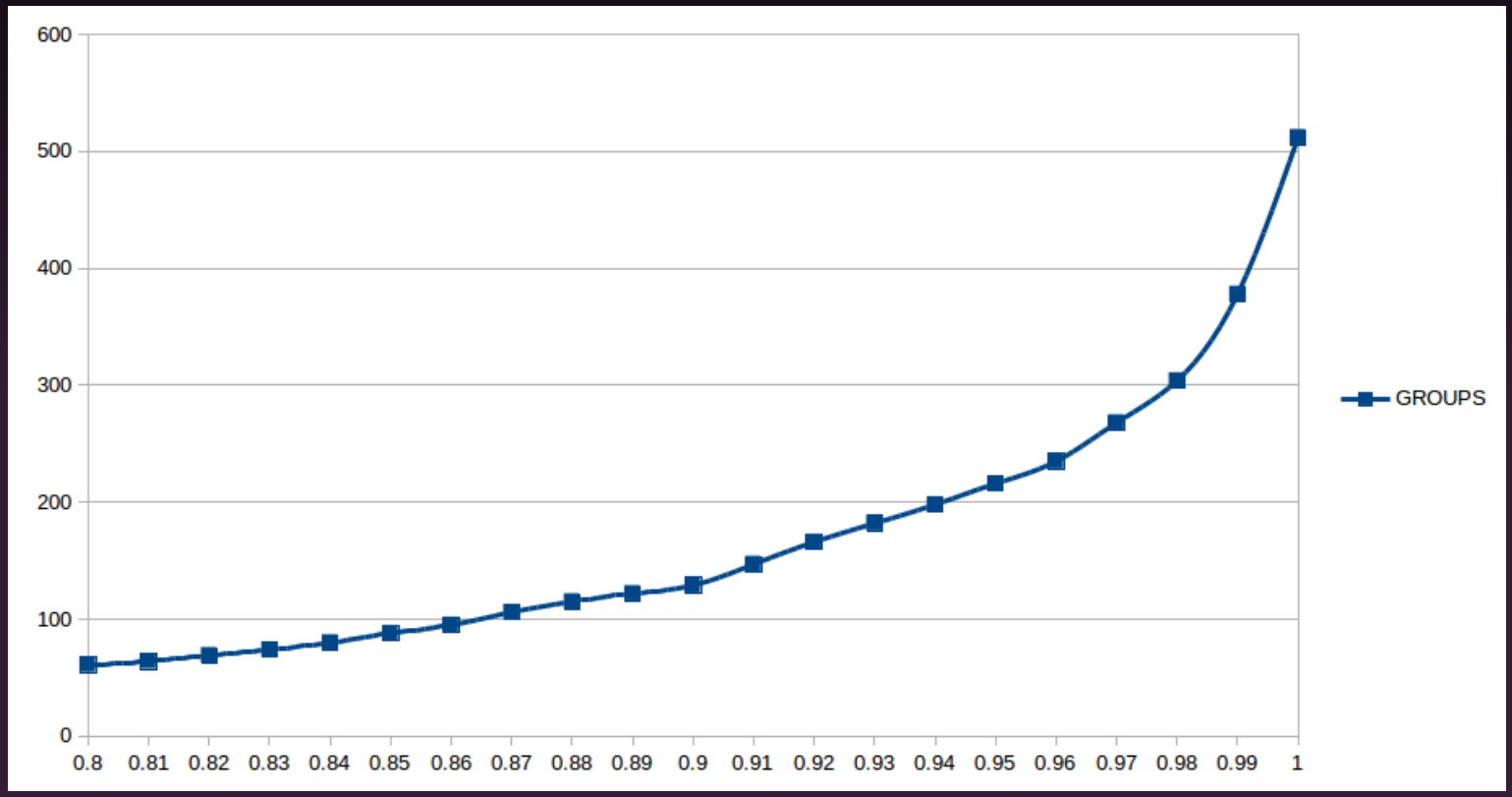
- We tested out Longest Common Subsequence (LCS) algorithm
- LCS is defined as an algorithm which is part of the Reinforcement Learning, which is also one of the branches in Machine Learning
- I then focused on the dynamic programming method, with sequence based algorithm
- The reason is I need to understand if the sequence is important to classify the malware and non malware
- For dataset A, I got a good result for classifying malware and non malware
- I replicated the process of the getting the dataset stream as in Dataset A with the researcher's tool (API hook at Ring 3 in the virtual machine)
- I also replicated the Dataset B's method to increase the number of benign software, to test out the algorithm

Sample input with of one of the datasets I

APITrace started for 124ef237c006cb419ad60e3bb509d7f4.exe
ProcessId: 1340
Timestamp, ThreadId, DLL, API
81367290074,1476,ADVAPI32.dll,GetSecurityDescriptorControl

81492369080,1476,kernel32.dll,GetProcAddress
81499913709,1476,kernel32.dll,GetProcAddress
81504322431,1476,kernel32.dll,GetProcAddress
81508904906,1476,kernel32.dll,GetProcAddress
81512476374,1476,kernel32.dll,GetProcAddress
81517542883,1476,kernel32.dll,GetProcAddress

Draft of the expriment's result



Example of running LCS analysis for Dataset A, at ratio 0.55

```
Group 0 -------------------------------------------------
Worm.Win32.Allaple.A-096d7f4d3ff56c90060d2effad34fdbe.MS.exe.csv
Worm.Win32.Allaple.A-6aaf5af66183ea26be412605b8c559a.MS.exe.csv
Worm.Win32.Allaple.A-45f071cd11b0bb9bfd3fe7a21cbb9f4d.MS.exe.csv
Worm.Win32.Allaple.A-014678e722fec9f5eb6a06776f3e64f2.MS.exe.csv
Worm.Win32.Allaple.A-06aeb0d07707f3786e5cff438b6421dc.MS.exe.csv
Worm.Win32.Allaple.A-179e1f455c511340fe9e0e73da35a39a.MS.exe.csv
Worm.Win32.Allaple.A-260cc0a2e6cc941ebba0e1f6c7cf6331.MS.exe.csv
Worm.Win32.Allaple.A-18e2b135d5a1e56a652b753d4ffd8166.MS.exe.csv

Group 2 -------------------------------------------------
Worm.Win32.Korgo.V-17f02ee76ed5de5f37c8f2c761bde2cc.MS.exe.csv
Worm.Win32.Korgo.V-1e5df7ba7491bf8c40d866736e5a96be.MS.exe.csv
Worm.Win32.Korgo.V-3c19c5ed7755aa8e5fec14def0b5ceae.MS.exe.csv
Worm.Win32.Korgo.V-33ffb2cb88d90c9e32317348db320d69.MS.exe.csv
Worm.Win32.Korgo.P-09375c2f4c66539a730e25208d5a1c60.MS.exe.csv
Worm.Win32.Korgo.AB-2b52862acfb0447cbe86ed8543fc01b3.MS.exe.csv
Worm.Win32.Korgo.S-55fe9d9adeb3e79831c7048906fb2201.MS.exe.csv
```

LCS analysis, on the benign software vs malware

```
Group 1 ------------------------------------------------
tsort.00000000000000000000000000000000.exe.csv
sha1sum.00000000000000000000000000000000.exe.csv
comm.00000000000000000000000000000000.exe.csv
indxbib.00000000000000000000000000000000.exe.csv
dos2unix.00000000000000000000000000000000.exe.csv
which.00000000000000000000000000000000.exe.csv
getfacl.00000000000000000000000000000000.exe.csv
truncate.00000000000000000000000000000000.exe.csv
freshclam.00000000000000000000000000000000.exe.csv
zipnote.00000000000000000000000000000000.exe.csv
unexpand.00000000000000000000000000000000.exe.csv
troff.00000000000000000000000000000000.exe.csv
```

Example of running LCS analysis for Dataset B, at ratio 0.9

```
Group 38 -------------------------------------------------
XTrojan.Win32.Cacogen.apm.txt
XVirus.Win32.HLLP.Alcaul.b.apm.txt
XVirus.Win32.HLLP.Alcaul.c.apm.txt

Group 52 -------------------------------------------------
XVirus.Win32.HLLP.Semisoft.g.apm.txt
XVirus.Win32.HLLP.Semisoft.i.apm.txt
XVirus.Win32.HLLP.Semisoft.k.apm.txt
XVirus.Win32.HLLP.Semisoft.l.apm.txt
XVirus.Win32.HLLP.Semisoft.m.apm.txt
XVirus.Win32.HLLP.Semisoft.n.apm.txt
```

Possible future plans

- Check for ransomware and APT
- Check if the same method is applicable for Android based malware

Conclusion

- Machine learning is possible to be used to assist analysts to detect and classify malware
- It should not be relied 10 percent, human expertise is still needed
- Helps out to segregate common attacks with the new and unknown attacks

Thanks!

najmi@endpoint.com

created with LaTeX