IoT attacks detection

Yulia Zamyatins

April, 2021

Introduction

The **detection_of_IoT_botnet_attacks_N_BaIoT** data set[1] contains traffic data from 9 commercial IoT devices authentically infected by Mirai and BASHLITE (gafgyt).

The data set has 115 attributes:

1. It has 5 time-frames: L5, L3, L1, L0.1 and L0.01.
2. The statistics extracted from each stream for each time-frame:
   - *weight*: the weight of the stream (can be viewed as the number of items observed in recent history)
   - *mean*
   - *std (variance)*
   - *radius*: the root squared sum of the two streams' variances
   - *magnitude*: the root squared sum of the two streams' means
   - *covariance*: an approximated covariance between two streams
   - *pcc*: an approximated correlation coefficient between two streams
3. It has following stream aggregations:
   - *MI*: ("Source MAC-IP" in N-BaIoT paper) Stats summarizing the recent traffic from this packet's host (IP + MAC)
   - *H*: ("Source IP" in N-BaIoT paper) Stats summarizing the recent traffic from this packet's host (IP)
   - *HH*: ("Channel" in N-BaIoT paper) Stats summarizing the recent traffic going from this packet's host (IP) to the packet's destination host.
   - *HH_jit*: ("Channel jitter" in N-BaIoT paper) Stats summarizing the jitter of the traffic going from this packet's host (IP) to the packet's destination host.
   - *HpHp*: ("Socket" in N-BaIoT paper) Stats summarizing the recent traffic going from this packet's host+port (IP) to the packet's destination host+port. Example 192.168.4.2:1242 -> 192.168.4.12:80

Thus, the column *'MI_dir_L5_weight'* in the data set shows the weight of the recent traffic from the packet's host for L5 time-frame.

I've added extra *'botnet'* column, where I keep information about the attacks from the different botnets and benign traffic. I've used "ga_" prefix for gafgyt attacks, and "ma_" prefix for Mirai attacks.

The team that collected this data set used 2/3 of their benign traffic (their train set) to train their deep autoencoder. Then they used remaining 1/3 of benign traffic and all the malicious data (their test set) to detect anomalies with deep autoencoder. The detection of the cyberattacks launched from each of the above IoT devices concluded with 100% TPR.

[1] detection_of_IoT_botnet_attacks_N_BaIoT Data Set

On HarvardX PH125.9xData Science course we learned several algorithms that can be used for the classification, such as **KNN**, **rpart** or **randomForest**.

My aim in this project is to check how well all these algorithms can detect anomalies in the data set, how accurate they can perform classification.

The source data set consists of *.csv and *.rar files, each representing the benign traffic or the attack, that are divided into folders with devices names. Because R has no package that can unpack *.rar files for its own, so all *.rar archives have to be manually unpacked with command line or third-party applications, depending of OS. Therefor, I had to prepare the data set for this project that can be easily downloaded and unpacked.

But the whole data set size was more that 1TB. Nowadays it's not a problem to find a hosting to share this huge data set, but I thought that everyone who will check this project won't be happy to download it. Because the team, that collected this data set, trained and tested their autoencoder for each device separately, I decided that I can use the data only from one device for my project. I've used the data only from Danmini doorbell device, but the data for other devices can be downloaded from the source and prepared for the classification using download-data.R script.

The data set for Danmini doorbel has the size of 970MB in *.csv file and only 200MB in *.zip archive. The *.zip archive is downloading during the first time of using project scripts and saving as *.csv file in local project data folder. For the next time, saved *.csv files is used.

Difficulties

- rar
-

Dataset analisys

that explains the process and techniques used, including data cleaning, data exploration and visualization, insights gained,

Methods

your modeling approaches (you must use at least two different models or algorithms);

Result

that presents the modeling results and discusses the model performance

Conclusion

that gives a brief summary of the report, its potential impact, its limitations, and future work.