Network Flow based botnet detection using supervised learning

Praveen Keshavamurthy

Problem Statement

- Botnets deploy C&C channel using a variety of communication protocols, such as: IRC, HTTP / HTTPS, P2P
- A large set of netflow data is available from the IP backbone traffic
- The metadata information in netflow data is not ideal
- Huge availability of netflow data is a motivation to find features which might help us to build supervised model to detect botnet traffic

Dataset

- CTU-13 DATASET consists of labelled data from 13 different botnet scenarios simulated over a period of 9 days
- The netflow data contains bidirectional and unidirectional flows ~20M flows
  - Background flows ~19M
  - Normal flows ~ 350k
  - Botnet flows ~ 400k
- Dataset after Feature extraction per destination IP ~53k unique destination IPs
  - Botnet data - 48479
  - Normal data - 4101

Approach

- Goal: To build a classifier to distinguish malicious from legitimate destination IPs
- Steps:
    - Extract features from the CTU-13 Netflow dataset for each of destination IPs
        - 48k botnet labelled IPs
        - 4k normal labelled IPs
    - Compare 3 classifiers - Logistic Regression, Neural Network and Random Forest
    - For each classifier, run 10 iterations of cross-validation and in each iteration:
        - Sample 4K malicious IPs from the malicious set at random
        - Split the 4k malicious + 4K benign IPs such that the test_size is 40% & train_size is 60%
        - Run 10-fold cross validation on the training set
        - Test the model for accuracy, precision and recall on test set

Feature Extraction

- Feature Set 1: Generic features
  - F1 - Total Source IPs per destination IP
  - F2 - Total Protocols used for communication per destination IP
  - F3 - Total Bidirectional flows per destination IP
  - F4 - Total Client flows per destination IP
  - F5 - Total Server flows per destination IP
  - F6 - Protocols used for communication represented as bit string
- Feature Set 2: Aggregate features
  - F7 : F12 - Total, Max, Min, Mean, Variance, Std of Flows per destination IP
  - F13 : F18 - Total, Max, Min, Mean, Variance, Std of Packets per destination IP
  - F19 : F24 - Total, Max, Min, Mean, Variance, Std of Bytes per destination IP
  - F25 : F30 - Total, Max, Min, Mean, Variance, Std of SourceBytes per destination IP

Features (contd..)

- Feature Set 3: Subnet Features
  - F31 - No. of distinct IPs in dstIP/24 subnet
  - F32 - Total Flows in dstIP/24 subnet
  - F33 - Total Packets in dstIP/24 subnet

- Feature Set 4: Periodic Communication Features based on IAT
  - F34 - Total periodic communications involved per destination IP
  - F35 - Ratio of total source IPs involved in periodic communication over total source IPs involved per dst IP

Result

Classifiers	*Accuracy(%)*	*Precision(%)*	*Recall(%)*
Logistic Regression	83.7	78.3	90
Neural Network (1 Layer)	91.1	86.7	97.1
Random Forest	99.8	99.9	99.8

Feature Analysis

Features	Logistic Regression			Neural Network			Random Forest		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
F1	77.6	71.2	92.1	84.3	76.4	99.4	99.7	99.8	99.6
F2	50.8	21.4	0.4	54.8	62.3	60	99.6	99.9	99.3
F3	68.5	59.7	88.5	72.6	66.3	92.6	94.7	94.6	94.6
F4	50.6	0	0	50.9	45.5	90	50.6	11.2	20

Feature Analysis (contd..)

Features	Logistic Regression			Neural Network			Random Forest		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
F1,F2	78.2	71.8	92.1	86.9	80	98.5	99.7	99.9	99.3
F1,F3	82.8	77.2	89.3	89.9	85.5	96.2	99.5	99.9	99.1
F1,F4	78.7	72.4	92.2	85.2	77.7	98.7	99.7	99.8	99.6
F2,F3	68.9	60	88.5	71.3	70.3	83.4	99.5	99.9	99.2
F2,F4	61.8	56.6	98.4	53.8	64.6	49.9	99.6	99.9	99.3
F3,F4	68.7	59.8	88.6	73.8	67.5	92.1	94.7	94.5	95

Feature Analysis (contd..)

Features	Logistic Regression			Neural Network			Random Forest		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
F2,F3,F4	69	60.1	88.6	75.0	68.3	93.7	99.6	99.9	99.2
F1,F3,F4	83.4	78	89.9	90.4	85.8	97	99.5	99.9	99
F1,F2,F4	79.1	72.9	92.2	87.2	80.2	99	99.7	99.9	99.6
F1,F2,F3	83.2	77.9	89.6	90.8	86.7	96.5	99.7	99.9	99.4

Software Details

- https://github.com/praveenkmurthy/BotnetDetection.git
  - Feature Extraction - Hadoop
  - Logistic Regression and Random Forest - sklearn package
  - Neural Network - tensor flow package