Dark Net

Data Science 440- Capstone Project Description

Authors: Austin Gongora, Jack Haser

1. Problem Descriptions:

As a group we will focus on TCP SYN requests on the Telnet traffic, specifically ports 23 and 2323. The presentation suggested that we focus on the features: number of "unique srcIPs" per minute or "number of packets/bytes per minute". The issue at hand is to detect anomalies in the network traffic. My team will use unsupervised learning approaches to find these patterns in the dataset. A necessary assumption is that the majority of instances in the data set are normal. We also plan to use visualizations to provide analysis of the Mirai Botnet activity over time and a geo-location heat map to analyze areas around the world with high-density Mirai Botnet cases.

2. Expected Significance of Models:

Discovering a method to locate potential bad actors on a network is an increasing field of interest for many industries. This will serve to expose our team to real-world problems that need solving. If Jack and I are able to create an accurate and precise model, then there are actual areas in cybersecurity that this can be applied to. Considering data privacy and security have recently been in the headlines, we think most areas concerning network security are among some of the most important topics.

3. Potential Insights from Models: What are the potential insights regarding the problems above that you hope to gain from your project? How do you plan to extract these insights from the models you constructed?

The insight we are trying to gather from our model is the ability to predict if some ports are more associated with malicious activity and attacks than other ports. We plan to extract this insight from the model since it will be learning normal traffic, categorizing malicious traffic and allow us to put a numeric value to which ports are more prone to malicious activity. Mirai's activity in the network traffic could be abnormal seeing as it is such a large bot-net,

4. Planned Model Construction: What machine learning tools/modules you plan to use to create predictive models for your project? What specific types of machine learning models did you plan to use for each problem listed in Problem 1?

**K means

The vast number of features in our data leads to working with a problem with many dimensions. Dimensionality reduction is going to be an important tool to gain further insight. Principle Component Analysis or PCA was a suggested technique to use. The problem statement asked for visualization tools to be used, so we will also use t-distributed stochastic neighbor embedding to provide visualizations. Other visualizations will be Time Series to demonstrate network activity across different ports and a heat map to analyze the geo-location information of Mirai-infected IoT devices.

Our main task will be to use anomaly-based detections- thus, our assumption is that the majority of instances in the data set will be normal. Our high-level goal will be to model the normal network and system behavior and identify anomalies as deviations from normal behavior. The plan will be in the form of multiple stages. The first stage is to use a Self-Organizing MAP (SOM), an unsupervised clustering

approach. The architecture will be a feed-forward neural network with a single layer of neurons. SOM's have shown to work well with high dimensional data on a two-dimensional plane. The hypothesis is that we will be able to learn normal traffic patterns over time- like the TCP/IP port numbers.

The next stage will be to use the learned network behavior and feed this into a supervised multi-layer perceptron model to find patterns in the data that are anomalies. The number of nodes and layers will be determined by the first stage. The quality of insight gained from stage 1 will determine the quality of results given in stage.

5. Planned Model Evaluation/Comparison: How do you plan to evaluate your models?

We will measure Precision, Sensitivity, Specificity, 1-specificity or FAR & Negative Prediction Values. We will use a ROC curve to handle trade-offs between the different metrics. Using the SOM algorithm will require us to tweak it in different ways. The number of clusters must be specified and this will be a tricky part considering there isn't any prior knowledge to the data. We will combat this by running the model with a multitude of different clusters.

6. Model Refinement: How do you plan to refine your initial set of models after they have been constructed and evaluated?

The algorithms mentioned in section 4 will require a deep understanding of the model and the parameters that accompany it. Running the model with different layers, initiating different numbers of clusters to our SOM algorithm will generate variable results, so looking for a proper balance of our data will be important. We are planning on retesting our models with different types of weights and acceptable conditions.

Observing misclassified samples is a good way of determining how well our model is predicting. The issue with unsupervised learning is that isn't a possible approach. However, after our first stage, we will have learned the normal traffic patterns and can use this as "labeled data". After our second stage, we will be able to test if the model is classifying normal traffic as it should be.

7. Planned Extraction of Insights from Models: How do you plan to extract insights from the models you constructed?

We plan to extract normal network behavior using an unsupervised approach and later determining the anomalies that occur within the proposed normal behavior. By creating an approach to detect these anomalies we will gain a deeper understanding of the Mirai Botnet. This can lead to future work such as creating a method to determine networks most vulnerable stress points. The idea would be to prevent malicious attacks prior to them happening. Of course, this is the future scope and we as a team are limited with time and knowledge. But I believe that we can extract valuable insights that could potentially lead to even greater findings.

8. Milestones for Midterm Project Report: Your milestones for Midterm Project Report should include (1) the construction and evaluation of predictive models for problem 1, and (2) the design of model construction for at least one more problem described in Section 1.

*Find Open Source SOM model designed for Anomaly Detection
*Prepare data for Visualizations (Time Series & Heat Map)
*Begin Stage 1
*Implement the SOM model and run test with the data
*Generate normal network behavior
*Finish Visualizations
*Determine ideal Cluster number for SOM model (*Use t-sne or PCA??)
*Evaluate Key metrics of stage 1
*Begin Stage 2- find open source multilayer perceptron to run inferences on normal network behavior
*Stage 2 model predictions- determining anomalies within data
* Evaluate key metrics of stage 2
*Extract value from predictions

9. Project Plan: List weekly tasks to be completed. The end date of each task should be the due date of the corresponding labs.

*Our team will be using Wrike to keep track of tasks in a timely manner

Week 1	Task	Due Date
Austin	Model research	Feb 15th
Jack	Visualization research	Feb 15th
Both	Exploratory Analysis	Feb 15th

Week 2	Task	Due Date

Austin	Begin Creating Model SOM Model	Feb 25th
Jack	Begin Time Series Visualizations Creation or Heat Map	Feb 25th
Both	Data Cleaning/Preparation	Feb 20th

Week 3	Task	Due Date
Austin	Finished SOM Model	March 10th
Jack	Visualizations should be finished- could potentially work on t-sne or PCA	March 10th
Both	Begin Stage 2	March 15th

Week 4	Task	Due Date
Austin	Begin Multi-layer perceptron model	March 20th
Jack	T-sne visualizations	March 20th
Both	Finished Stage 2	March 25th

Week 5	Task	Due Date
Austin	Test data Pipeline. Stage 1 → Stage 2 → Predictions	April 5th
Jack	Extract Model Insights	April 5th

Both	Begin Testing Predictions	March 30th

Week 6	Task	Due Date
Austin		
Jack		
Both	If all goes according to plan, we can begin working on problem 2, this wasn't mentioned in our problem description but there is a second challenge	April 20th