Piradon (Tien) Liengtiraphan
Project: Milestone
Instructor: Dr. Pablo Rivas

CMPT 404L-111: Artificial Intelligence
Due: 2016-11-07

Analysis of Failed Access

Exercise 1a. *The milestone report should describe what you've accomplished so far, and very briefly state what else you plan to do.*

Proof.

Abstract

To see whether there is a relationship between attacks to protected resources on a college campus and methods of attack chosen, that can be identified by Unsupervised Algorithms.

Introduction

The original idea behind the pursuit of this idea comes from my summer working in the IBM-Marist Joint Study. Each day the college receives an incredibly large number of attempted access, log-in attempts, extending all the way to DDoS attacks. The system is currently being protected by a Security System from BlackRidge technology, which keeps a log of all the ”Discard” and ”Forward” actions. These logs allow us to see where and how each rouge IP attempted to access the protected resource. Analyzing this data with a Machine Learning Algorithm might provide cyber-security professionals useful insight to the inner working of an attempted access and/or possible identify correlations that might have gone missing under the human eye. The Unsupervised Algorithm of K-Means was chosen to help identify possible correlations. This choice was made given the fact that we are currently uncertain whether a relationship exists or not, while over-fitting might become an issue, I believe it is first important to establish a base-case from which work can continue.

Background and/or Related Work

I am uncertain if similar research has been done in this particular area, given that the data being used was collected and analyzed using tools that were written by students and the IBM-Marist Joint Study. However identifying pattern in and between cyber-security attacks is not a new area of studying. The usage of log files to dynamically update firewall setting is an avenue to researching being pursued by many in the industry. What I aim to achieve with this Machine Learning Project is to possible identify patterns in attacks as they coming in, allowing slightly delayed, if not real time identification of different attacks and/or attackers. As for how this project could be improved, more data from other systems should be collected, once that data is acquired we would run the Unsupervised Algorithm to see if the correlations found in this experiment hold true for multiple cases.

Methodology

The approach taken in this experiment is slightly different from what one might find if one had chosen a project from Kaggle. First and foremost this experiment focused on data collection, the analysis stage coming afterwards. This project has taken slightly longer than initially expected, due to the need for data collection and the code that has to be written to parse the data into a form that would be considered useful for analysis later on. Once the data was collected an appropriate Machine Learning Algorithm must be selected to analyze the data. As of last week we learned about Unsupervised Algorithms, specifically K-Means, given the nature of this experiment it seemed ideal to use K-Means due to its nature of finding possible correlations and relationship between data points that are unspecified. Currently the data collected from the BlackRidge Boxes are formatted similarly to the 'features.csv' file provided in class, with 3 separate data point located in the first cell using a delimiter of a single white-space. Initial attempts to run the Nearest_Neighbors algorithm of the data set resulted in the following error

```
"Traceback (most recent call last):
File "C:/Users/Tien/Documents/GitHub/CMPT404-ARTIFICAL-INTELLIGENCE/Liengtiraphan-Project/Project
Resources/Nearest Neighbors.py", line 18, in ¡module¿
X, y, ytrue = genfromtxt('Dataset/features.csv', delimiter=' ') ValueError: too many values to unpack (expected 3)"
```

The issue lies with the formatting of the data, with help from Professor Rivas, I hope to properly format the data so that the experiment can proceed further. As discussed this morning Nearest_Neighbors was decided not the be the proper algorithm to use, instead we will shift to using K-Means while using Silhouette Coefficient to attain the K-Value.

Experiments

This data was obtained from the syslog of the 1G BlackRidge Authentication pair, which sits in the ECRL (Enterprise Computing Research Lab) in Hancock's Basement. The data was then parsed out using a script that I wrote to read the data and organize it in a way that was useful and usable for this particular experiment. The script leverages Python's re (Regular Expressions) library to identify and retrieve the data. This data included: src_ip, src_port, dest_ip, dest_port, city, subdivision, lat, country, postal, long, host, host_name, isp_ip, asn. For usage in this experiment, given that the algorithms we learned cannot be used against data sets that are not numbers, each was assigned a unique value in decimal form to be used in this experiment. The experiment that will be performed on this particular data set will be different combinations of the data collected to see if a relationship and/or correlation and be found between the data points by the Near_Neighbors Unsupervised Algorithm. Currently attempted analysis for a relationship between IP address and Source Port has begun.

Discussion and/or Analysis

Unfortunately given that the desired Machine Learning Algorithm selected to analyze the data has to be tweaked further to ensure compatibility with the data set provided. With the help of Professor Rivas, I hope to rectify this situation as soon as possible so that continued experiments with different data combinations can commence. As soon as this situation is rectified and conclusions can be drawn, this section will be updated periodically to include each new instance of analysis on data combinations.

Conclusion

Much like the previous section, not much can be written base on solid concrete data acquired in this section, due to the roadblock that I am currently facing. However, what can be outlined is what I hope to achieve with this experiment. Ideally a correlation between the different data combinations can be found. This data will the be used to help identify specific attacks and/or sources of attacks to our system using pattern recognition from the data provided in this experiment. This pattern will hopefully be distinct enough to identify; however, not so unique to this particular use case that it is inapplicable to other similar environments. As soon as my current roadblock is circumvented more updates will be made to this section as well as the one above, when concrete data is generated to draw conclusions and analysis from. □