Brandon Rowe

Dr. Zhao

CSCI - 499

05-07-19

Business Intelligence: Machine Learning

Machine learning is the hottest topic in computer science today. Machine learning is a principle of Artificial Intelligence that suggests a software system can perform tasks as instructed but also, over time, learn and make decisions that simulate forms of consciousness within a machine. The idea of machine learning is that software will complete certain (pre-defined) tasks over and over while learning from the data it receives and results that it yields, creating the ability to predict or react to situations similar to that of a human but much faster. Many companies have already created excellent proofs of concept by designing systems like AlphaZero, formerly AlphaGo, that was able to learn the Go board game and defeat the best player in 4 hours. The significance of which lies in the fact that it takes human beings many years to learn and perfect a skill game such as Go or Chess but this can be done by a machine within several hours time. The potential to automate and make intelligent decisions within a fraction of a second has never been more readily available to the world. Most computers sold today have very powerful processing capabilities and are able to perform multiple tasks at once within milliseconds. The first commercial machine learning tools have been used for Business Intelligence or BI. Business intelligence software typically takes some kind of raw data, produced by a computer, analyzes the raw data to structure and identify things in a certain way, and then display that data in a meaningful way with visualization tools like graphs and charts. This software is typically used by tech affluent individuals to produce reports and metrics for executives or administrative staff who do not have the technical ability to understand or review the raw data. The business intelligence project conducted for this research project has consisted of three parts; gathering raw data, structuring and identifying anomalies in the data, and creating visual tools to represent that data.

The first important step in any Business Intelligence solution is to identify and collect the raw data that will be used for processing by the software. For the purpose of this research, we have used [Bro](https://www.zeek.org/) network monitoring software to gather data from a given network. “What is Bro? Bro is an open source software framework for analyzing network traffic that is most commonly used to detect behavioral anomalies on a network for cybersecurity purposes” (Bricata). Bro essentially provides similar functions to network intrusion detection systems but can also allow security operations to perform incident response, hashing, forensics, and extraction among other things. Bro is used to, “1. Convert data about network traffic into higher level events. 2. Provide a script interpreter - a robust programming language – which is used to interact with events and understand what those events mean in terms of network security” (Bricata). We used bro to monitor a local network with five devices connected and three actively sending and receiving packets on the network. Bro captures metadata of the activity on the network but also performs some analysis to determine what activity poses malicious or suspicious threats. On the local network, two of the connected devices are not active but one of the three active devices is intermittently sending malicious packets through Low Orbit Ion Cannon. Three separate tests are conducted; the first test is with minimal traffic other than the malicious device, the second test is with moderate traffic and the third test is with heavy traffic, including the two inactive devices. Bro is capable of monitoring HTTP, DNS and connections logs but we will be focusing on the HTTP log. To change the amount of traffic we simply increase pings to websites from each host device on the local network while running LOIC on at least 1 device. This gives us an idea of how well Bro works given different traffic inputs to see if it performs as well over larger data sets. After all three tests of difficulty are performed, Bro appears to hold up performance within a fair margin across all scales of difficulty (low, moderate, high). Bro is able to distinguish between random pings on a network, which simulate requests/response from general internet use, and the malicious packets that are sent from LOIC. Bro uses a command line interface (CLI) that can be started using BroControl and actively begins monitoring the preset network interface for traffic and identifying anomalies. Examples and screenshots of the tests performed in Bro are available on [Github](https://github.com/brandon-rowe/Bro-Bat-Dash) and tables are available at the end of this research paper, section titled Table 1-3.

The second important step in any Business Intelligence solution is to analyze and identify certain data in order to structure that data into meaningful tables to be read and reviewed by others. While Bro already uses some machine learning tactics to monitor the network and organize the data for professional IT personnel, it is still not formatted in such a way that it can be easily converted to readable data for visualization in charts and graphs. For this purpose, we have used the [Bat](https://github.com/SuperCowPowers/bat) software package to analyze and organize the data from Bro network monitoring in order to create readable tables that can then become actionable for visualization. Bat is a compilation of processing and analysis tools for Bro network data such as Pandas, scikit-learn, and Spark. Bat creates an easy interface to use so that the user does not need to download the aforementioned software packages and implement them individually. So if Bro already has a powerful scripting language of its own, why use BAT? Two reasons, “Offloading: Running complex tasks like statistics, state machines, machine learning, etc.. should be offloaded from Bro so that Bro can focus on the efficient processing of high volume network traffic. Data Analysis: A large set of support classes that help bridge from raw Bro data to packages like Pandas, scikit-learn, and Spark. Also, example notebooks that show step-by-step how to get from here to there” (SuperCowPowers, Bat). Bat can use Bro logs as Python Dictionaries, Pandas DataFrames, VirusTotal Queries, uncommon User Agents, extracted files, certificate verification and anomaly detection. We have used Bat to monitor the http.log from Bro for uncommon user agents and anomaly detection. After all three tests of difficulty are performed, the http.log file is analyzed and yields excellent tables for anomalies and identifying threats. We use Bat to separate the dataset from potential outliers. Outliers were identified as packets from the attacking host device with LOIC and were identified by the length of data being sent and recognizing that several packets were duplicated packets sent from the same host. Some of the anomalies identified by the model might be expected at times, given a server with large amounts of traffic in general, but for the purpose of this research, the software successfully analyzed and identified packets sent from the attacking host computer (outliers) and regular working devices sending and receiving normal traffic (dataset). Examples and screenshots of the tests performed in Bat are available on [Github](https://github.com/brandon-rowe/Bro-Bat-Dash) and tables are available at the end of this research paper, section titled Table 1-3.

The last important step in any Business Intelligence solution is to take the structured data and implement it into some type of meaningful representation like charts or graphs. This step is critical to any business intelligence software and proved to be the most challenging of the entire project. While Bro and Bat were synchronous tools that are made to use together, developing a web-based visualization of the data would require many different web development languages. We decided to use the Dash framework that is used for web application analytics. Dash is a python based framework which can be used with Bat since Bat can create python dictionaries that contain the dataset and representation. Dash framework is, according to plot.ly, is “Plotly's Python graphing library that makes interactive, publication-quality graphs online ” (plotly). This framework creates line, bar and circle charts and graphs for the data representation. We were using Plotly to create cluster charts of the data in order to display the amount of data being sent across the network, the length of the data being sent in packets, as well as the source and destination frequency information. These representations would make a meaningful display of anomalies on the network, highlight the threat level and amount of traffic that could be marked nefarious, and also show the inconsistencies between normal traffic and anomalies. Unfortunately, we were unable to get a working model of the dash framework onto the live server at cyntax.org at this time. Examples and screenshots of the tests performed using Dash are available on [Github](https://github.com/brandon-rowe/Bro-Bat-Dash).

We successfully demonstrated the ability to gather raw network data with Bro network security tool, analyze data for anomaly detection within the http.log file using Bat, and connect those tools to an outside web analytics application framework for visualization. Though the final finished project is not complete and live online, we have established that it is plausible to develop machine learning tools with existing packages. The most important aspect of the business intelligence solution we have demonstrated is the machine learning aspect that is able to identify, structure and take action on a given dataset or function. The visualization aspect is important for representing the data in a meaningful way but does not comprise a business intelligence solution without some part of the program thinking and taking action within its predefined set of instructions. Further experimentation and exploration in the field of machine learning using these tools and packages can greatly influence our ability to automate tasks and processes so that functions can be carried out at lightning speed by a strong computer as opposed to the slow reaction of a person. In closing, creating a successful business intelligence solution is possible with free and opensource software/tools that are readily available.

Table 1

Test 1: Low Traffic

| Anomalies (Outliers) | Duplicates | Body Length(Avg) | Dataset |
| --- | --- | --- | --- |
| 2 clusters | 4237 | 72993 | 5 clusters |
| 1 cluster | 1374 | 84993 | 8 clusters |
| 5 clusters | 11459 | 69993 | 10 clusters |

Table 2

Test 2: Moderate Traffic

| Anomalies (Outliers) | Duplicates | Body Length(Avg) | Dataset |
| --- | --- | --- | --- |
| 17 clusters | 22958 | 72453 | 30 clusters |
| 21 clusters | 31487 | 79993 | 45 clusters |
| 23 clusters | 36020 | 62933 | 60 clusters |

Table 3

Test 3: Heavy Traffic

| Anomalies (Outliers) | Duplicates | Body Length(Avg) | Dataset |
| --- | --- | --- | --- |
| 22 clusters | 101234 | 65933 | 100 clusters |
| 37 clusters | 89235 | 84993 | 125 clusters |
| 33 clusters | 104239 | 69993 | 150 clusters |

Works Cited

|  |  |
| --- | --- |
| **Author** | 22 Resources |
| **Article title:** | What is Bro? [Overview and Resources] | Bricata |
| **Website title:** | Bricata |
| **URL:** | https://bricata.com/blog/what-is-bro-ids/ |
|  |  |
| **Article title:** | SuperCowPowers/bat |
| **Website title:** | GitHub |
| **URL:** | https://github.com/SuperCowPowers/bat/ |
| **Article title:** | The Zeek Network Security Monitor |
| **Website title:** | Zeek.org |
| **URL:** | https://www.zeek.org |
| **Article title:** | plotly |
| **Website title:** | Plot.ly |
| **URL:** | https://plot.ly/python |