ITC INFOTECH

Business-friendly Solutions

Unsupervised Network Traffic Anomaly Detection Using Parameterized Entropy and LSTM AutoEncoders

Aaqib Waleed Bhat
Associate Data Scientist
Advanced Analytics, ITC Infotech
Bangalore, India

Anindya Neogi
Chief Data Scientist
Advanced Analytics, ITC Infotech
Bangalore, India

Keywords

Network Security Threat, LSTM Auto Encoders, Parameterized Entropy, Anomalous Traffic

Abstract

Detecting anomalous traffic provides one approach to network security threat detection. We propose a behavior-based anomaly detection method that detects anomalous traffic by applying a threshold to a reconstruction error given by the *LSTM AutoEncoder* model on the *Bro conn log data* collected as time series data. This approach relies on *network connection feature distributions*. Anomalous traffic is detected by inspecting abrupt changes in the statistical metrics of the *network connection feature distributions* given by parameterized entropy.

1. Introduction:

Computer networks are prone to various kind of threats and attacks, and hence reliable algorithmic mechanisms needed for security monitoring. Many researchers and cyber security professionals over the years have concentrated their efforts in securing computer networks from these threats and attacks. Detecting anomaly is related to classifying data in two different classes: *benign* and *anomalous.* With respect to the current context, the range of anomaly spans over a large spectrum *including Spam, Denial of Service (DoS), Scanning, Botnets* etc. Since there are multiple subclasses of anomaly each having its due importance and could impose serious and equal magnitude of threats to the network a multi-class classifier could be a potential solution to detect and discriminate between non-anomalous and different sub-classes of anomalous traffic behavior. One of the main drawback in using this approach is the necessity to have supervisory tags / labelled data corresponding to the observed non-anomalous and different subclasses of anomalous traffic behaviors. It also means that a system based on such a multi-classifier would be perhaps unable to detect a new anomaly which is usually the kind in the current networks, including *Internet.* Our proposed approach would, therefore, be based on unsupervised approach by focusing on data and attributes associated with *network traffic flows.* This approach relies on *aggregated traffic metrics,* thus show better scalability and therefore seem to be more promising. The proposed approach is based on two rational assumptions about the data.

Assumption 1 *The majority of the network connections are normal traffic. Only X% of traffic is malicious. (Portnoy, Eskin & Stolfo 2001)*

Assumption 2 The attack traffic is statistically different from normal traffic. (Javitz & Vadles 1993, Denning 1987)

A Bro conn record is composed of fields like *Source and Destination IP-address, Source and Destination port numbers, Traffic Volume in bytes, Number of Packets, Services, Duration and Protocol* for a session between the previous two endpoints.

Anomaly detection is an identification of events which do not conform to an expected behavior. Network traffic is *temporal* and the *IP Header fields are stochastic* in nature. Thus we implemented *Long Short-term Memory (LSTM) AutoEncoder* model to capture the long term dependencies in the Network traffic and *reconstruction error for parameterized entropy* statistic is used to account for any random change in the *IP header fields.*

2. Parameterized Entropy:

In Information theory, entropy is the average rate at which information is produced by a stochastic source of data. For a probability distribution P (X = xi) of a discrete random variable X, the Shannon entropy is defined as

$$H_s(xi) = \sum_{i=n}^{n} p(xi)\ log_{\propto} \frac{1}{p(xi)}$$

The more random the variable, the bigger the entropy, and in contrast, the greater the certainty of the variable, the smaller the entropy.

The Shannon entropy assumes a tradeoff between contributions from the main mass of the distribution and the tail. To control this tradeoff, two parameterized Shannon entropy generalizations were proposed, by Renyi (1970s) and Tsallis (late 1980s). If the parameter denoted as α has a positive value, it exposes the main mass *(the concentration of events that occur often),* and if the value is negative, with it refers to the tail *(the dispersion caused by seldom events).*

Renyi entropy is given by : $H_{R\propto}(\mathrm{X}) = \frac{1}{1-\propto} log_a \left(\sum_{i=1}^{n} \mathrm{p(xi)}^{\propto} \right)$

Tsallis entropy is given by : $H_{T\propto}(\mathrm{X}) = \frac{1}{1-\propto} \left(\sum_{i=1}^{n} \mathrm{p(xi)}^{\propto} - 1 \right)$

Both parameterized (Renyi and Tsallis) entropies:

- expose concentration for $\alpha > 1$ and dispersion for $\alpha < 1$,
- converge to the Shannon entropy for $\alpha \rightarrow 1$,
- correspond to cardinality of X for $\alpha = 0$.

3. LSTM AutoEncoders:

An *LSTM AutoEncoder* is an implementation of an AutoEncoder for sequence data using an Encoder-Decoder LSTM architecture. Autoencoders try to approximate representation of the original signal. During training, the network tries to minimizes the difference between the approximate representation to the original signal. While LSTM AutoEncoders are capable of dealing with sequence as input, regular AutoEncoders can't.

4. The Proposed multi Stage System:

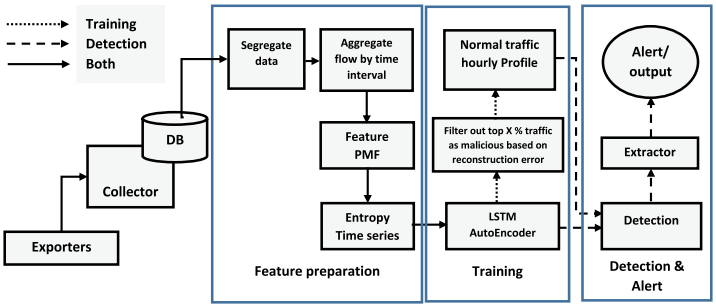

Figure 1: Network Anomaly Detector

a. Exporters

BRO enabled devices which permanently monitor network traffic, account statistics, and export connection-data to our system.

b. Collector

The function of this module is to collect connection-data exported from one or several exporters.

c. Feature Preparation module

Traffic distributions can be different on weekdays and weekends, so we segregated the flow data accordingly.

It is very unlikely that a single connection log can give much information about the activities happening in the network. Therefore, collected flows are analyzed for constant-length time intervals (say every 1 minute). A probability mass function (PMF) is created for every feature for a time interval. Next, depending on the version, Tsallis or Renyi entropy of positive and negative α values are calculated for traffic feature distributions and a time series is created out of the same.

d. Training

Over the years, several anomaly detection techniques have been proposed in the literature. The problem at hand was completely unsupervised and needed a sound approach for anomaly detection. We trained an AutoEncoder to learn an approximation of the original data and filtered out the top X% connection logs whose *reconstruction error* was high and thus capped the *reconstruction error* threshold. It is very likely that the network traffic at previous time intervals can show an effect on the traffic under study. In order to account for this characteristic, *Long Short-term Memory (LSTM) AutoEncoder* was implemented instead of a simple AutoEncoder.

Also, during the training phase, it builds a network profile of long term behavior over the data points which have less reconstruction error and are assumed to contain few or no attacks/anomalies. The network profile captures the hourly min and max concentration and dispersion Entropy Statistic for all the connection features to reflect the changes on weekdays and weekends.

e. Detection

In the detection phase, the observed data point for the time interval under study is made to pass through the LSTM AutoEncoder model. If the reconstruction error for short term behavior *(time interval under study)* is greater than the threshold, then the observed entropy is compared with the min and max values in the already created weekdays or weekends network profile according to the following rule:

$$r_{\propto} = \frac{H_{\propto}(Xi) - k * min_{\propto}}{k * (max_{\propto} - min_{\propto})} \qquad k \in (1...2)$$

Values $r_{\propto}(xi) < 0 \; or \; r_{\propto}(xi) > 1$ indicate abnormal concentration or dispersion respectively. This abnormal dispersion or concentration for different feature distributions is characteristic for different anomalies. For example, during a port scan, a high dispersion in port numbers and high concentration in *addresses* should be observed. *Detection* is based on the relative value of *entropy* with respect to the distance between min and max. Coefficient k in the formula determines a margin for min and max boundaries and may be used for tuning purposes. A high value of k, e.g. k = 2, limits the number of false positives, while a low value (k = 1) increases detection rate.

The final stage of the proposed system involves identifying the events that occurred, and gathering other related information as support,-from other related logs like *Bro dns logs, http logs, file logs, smtp logs* - to present status of the network to the administrator and create alarms with all the details as output.

5. Summary:

The proposed dynamic system provides early triggers in a completely unsupervised environment. The experimental results show that the use of *Bro connection logs* and extracting only features that significantly contribute to intrusion detection gives promising results.

There is a further scope to improve the accuracy and achieve a good tradeoff between true positives and false positives in the presence of a labelled dataset. This system can easily be extended, configured, and/or modified by replacing some features or adding new features for new types of attacks. The method we have used is a generic one and can be applied in various other domains, viz. Banking, Insurance, etc., to address fraudulent activities.

Reference:

1. Przemyslaw Bereziński, Marcin Szpyrka, Bartosz Jasiul, Michal Mazur: Network Anomaly Detection Using Parameterized Entropy.

2. Renyi, A.: Probability Theory. Dover Books on Mathematics Series. Dover Publ. Inc. (1973)

3. Tsallis, C., de Pesquisas Físicas, C.B.: Possible Generalization of Boltzmann-Gibbs Statistics. Notas de física. Centro Brasileiro de Pesquisas Físicas (1987)

4. Haakon Ringberg, Augustin Soule, Jennifer Rexford, Christophe Diot: Sensitivity of PCA for Traffic Anomaly Detection

5. Yousef Abuadlla1, Goran Kvascev2, Slavko Gajin3, and Zoran Jovanović3: Flow-Based Anomaly Intrusion Detection System Using Two Neural Network Stages.

6. Cynthia Wagner, Jérôme François, Radu State, Thomas Engel: Machine Learning Approach for IP-Flow Record Anomaly Detection.

7. Kingsly Leung Christopher Leckie: Unsupervised Anomaly Detection in Network Intrusion Detection Using Clusters