On Entropy in Network Traffic Anomaly Detection

Jayro Santiago-Paz, Deni Torres-Roman.
Cinvestav, Campus Guadalajara, Mexico

November 2015

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

Outline

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

Outline

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

Outline

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

Outline

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

Outline

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

Outline

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

Chandola et al. (2009) states that the term anomaly-based intrusion detection in networks refers to the problem of finding exceptional patterns in network traffic that do not conform to the expected normal behavior.
Given a traffic network and its set of the selected traffic features $X = \{X_1, X_2, \ldots, X_p\}$, and N time instances of X, the normal and abnormal behaviors of the instances can be studied. The space of all instances of X builds the feature space which can be mapped to another space by employing a function such as entropy. In the literature, Shannon and generalized Rényi and Tsallis entropy estimators, as well as probability estimators (Balanced, Balanced II), are used.

The A-NIDS usually consists of two stages: training and testing stage. In the training stage using a database of “normal” or free-anomaly network traffic, the feature extraction, windowing and entropy calculation modules, a “normal” profile is found. In the testing stage, using the feature extraction, windowing and entropy calculation modules, anomalies in the current network traffic are detected and classified.

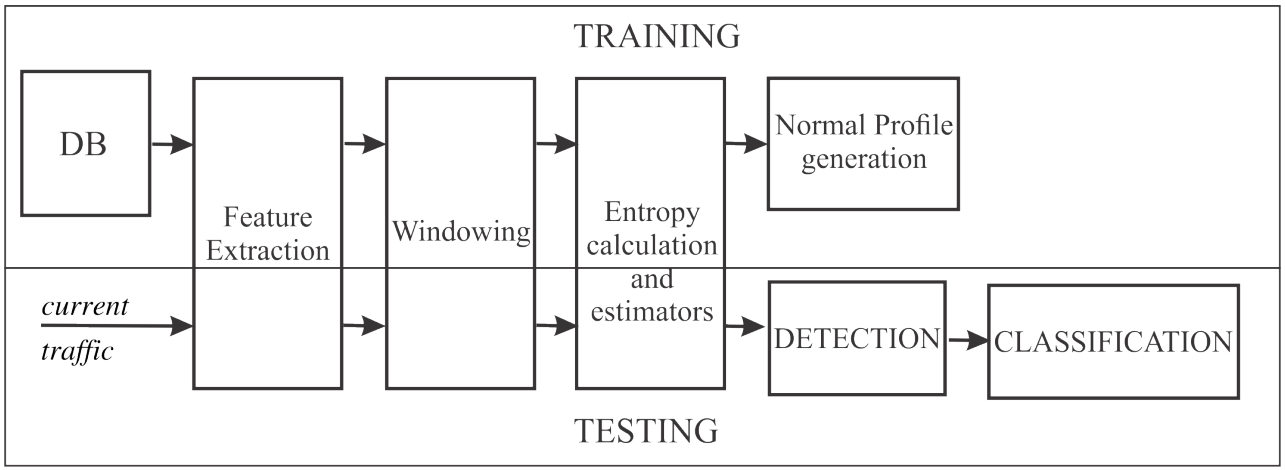


Figure 1: General architecture of entropy-based A-NIDS.

Synthetic

The synthetic databases are generated artificially, e.g., the MIT-DARPA 1998, 1999, 2000 databases[a], which include five major categories: Denial of Service Attacks (DoS), User to Root Attacks (U2R), Remote to User Attacks (R2U) and probes.

[a] http://www.ll.mit.edu/ideval/index.html

Real

Some real public databases are: CAIDA[a], which contains anonymized passive traffic traces from high-speed Internet backbone links, and the traffic data repository, maintained by the MAWI[b] Working Group of the WIDE Project. Other researchers have created their own databases in different universities, e.g., Carnegie Mellon University, Xi'an Jiaotong University, and Clemson University (GENI), or traffic collected from backbone in SWITCH, Abilene, and Géant.

[a] https://www.caida.org/data/passive/passive_2012_dataset.xml
[b] http://mawi.wide.ad.jp/mawi/

Motoda H. and Liu H. (2002)

Feature selection is a process that chooses a subset of M features from the original set of N features $M \leq N$ so that the feature space is optimally reduced according to a certain criterion.

Feature extraction is a process that extracts a set of new features from the original features through some functional mapping. Assuming that there are N features $Z_1, Z_2, \ldots, Z_N$ after feature extraction, another set of new features $X_1, X_2, \ldots, X_M (M < N)$ is obtained via the mapping functions F_i, i.e. $X_i = F_i(Z_1, Z_2, \ldots, Z_N)$.

Among the algorithms used to reduce the number of features in network traffic anomaly detection are: PCA, Mutual Information and linear correlation, decision tree, and maximum entropy.

In network traffic, the most commonly employed features are: source and destination IP addresses and source and destination port numbers. Other features extracted from headers are: protocol field, number of bytes, service, flag, and country code. Zhang et al. (2009) divided the size of packets into seven types and Gu et al. (2005) defined 587 packet classes based on the port number.

At flow[a] level the features selected were: flow duration, flow size distribution (FSD), and average packet size per flow. For KDD Cup 99, 41 features or a subset were employed. On the other hand, Tellenbach et al. (2011) used source port, country code and others, constructing the TES as input data.

[a] An IP flow corresponds to an IP port-to-port traffic exchanged between two IP addresses during a period of time T.

tronic
y and

15-30 November 2015

Windowing in Network Traffic

Window-based methods group consecutive packets or flows based on a sliding window. The ith window of size L packets is represented as $W_i(L,\tau) = \{pack_k, pack_{k+1}, \ldots, pack_{k+L}\}$, with $k = iL - i\tau$, where τ is the overlapping and $\tau \in \{0, 1, \ldots, L-1\}$. When the window size is given by time, L can be different in each window. Windowing is performed in two ways: overlapping ($\tau \neq 0$) and non overlapping ($\tau = 0$) windows.

The window sizes most commonly used are: 5 min, 30 min, 1 min, 100 sec, 5 sec and 0.5 sec. Some researchers use windows with a fixed length $L = 4096$, 1000, and 32 packets.

Let X be a random variable which takes values of the set $\{x_1, x_2, ..., x_M\}$, $p_i := P(X = x_i)$ the probability of occurrence of x_i, and M the cardinality of the finite set; hence, the Shannon entropy is:

$$H^S(X) = -\sum_{i=1}^{M} p_i \, log\,(p_i)\,. \tag{1}$$

The Rényi entropy is defined as:

$$H^R(X, q) = \frac{1}{1-q} log\left(\sum_{i=1}^{M} p_i^q\right) \tag{2}$$

and the Tsallis entropy is

$$H^T(X, q) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{M} p_i^q\right), \tag{3}$$

when $q \to 1$ the generalized entropies are reduced to Shannon entropy. In order to compare the changes of entropy at different times, the entropy is normalized, i.e.,

$$\bar{H}(X) = \frac{H(X)}{H_{max}(X)}. \tag{4}$$

Consider two complete discrete probability distributions $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$, with $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1, \quad 1 \geq p_i \geq 0, \; 1 \geq q_i \geq 0, \; i = 1, 2, \ldots, n$. The information divergence is a measure of the divergence between P and Q and is defined by Rényi (1961):

$$D_\rho(P||Q) = \frac{1}{\rho - 1} log \left(\sum_{i=1}^{n} p_i^\rho q_i^{1-\rho} \right), \quad \rho \geq 0, \tag{5}$$

where ρ is the order of the information divergence. Consequently, the smaller $D_\rho(P||Q)$ is, the closer the distributions P and Q are. $D_\rho(P||Q) = 0$ iff $P = Q$. When $\rho \to 1$ the Kullback-Leibler (KL) divergence is obtained

$$D_1(P||Q) = \sum_{i=1}^{n} \left(p_i \, log \left(\frac{p_i}{q_i} \right) \right), \; \rho \to 1. \tag{6}$$

Conditional Entropy

The conditional entropy of a variable Y given X, with alphabet $\mathfrak{X}$ and $\mathfrak{Y}$, respectively, is defined as:

$$H(Y|X) = -\sum_{x\in\mathfrak{X}} p(x) \sum_{y\in\mathfrak{Y}} p(y|x)\, log\left(p(y|x)\right) \tag{7}$$

$$= -\sum_{x\in\mathfrak{X}} \sum_{y\in\mathfrak{Y}} p(x,y)\, log\left(p(y|x)\right). \tag{8}$$

Joint Entropy

The joint entropy of X and Y, defined as

$$H(X;Y) = -\sum_{x\in\mathfrak{X}} \sum_{y\in\mathfrak{Y}} p(x,y)\, log\left(p(x,y)\right) \tag{9}$$

where $p(x,y)$ is the joint probability mass function.

The mutual information (MI) between two random variables X and Y is a measure of the amount of knowledge of Y supplied by X or vice versa. If X and Y are independent, then their mutual information is zero. The MI of two random variables X and Y is defined as:

$$I(X;Y) = H(X) - H(X|Y) \; = \; H(Y) - H(Y|X) \; = \; H(X) + H(Y) - H(X;Y) \quad (10)$$

where $H(\bullet)$ is entropy, $H(X|Y)$ and $H(Y|X)$ are conditional entropies, $H(X;Y)$ is the joint entropy.
The MI equation can be written as:

$$I(X;Y) = \sum_{x \in \mathfrak{X}} \sum_{y \in \mathfrak{Y}} p(x,y) \, log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (11)$$

where $p(x)$ and $p(y)$ are marginal probability mass functions of X and Y, respectively.
In order to estimate the MI between X, Y, it is necessary to estimate $p(x,y)$

Different probability estimators are used, e.g., relative frequency, Balanced, and Balanced II, and consequently, a "true" probability distribution is built. The entropy is calculated using these estimators; the more accurate the estimators, the better the entropy estimates.
Rahmani et al. (2009) noted that time series of IP-flow number and aggregate traffic size are strongly statistically dependent, and when an attack occurs, it causes a rupture in the time series of joint entropy values. In order to calculate the joint entropy $H(X;Y)$ they employed $p(x,y)$ of the time series X and Y using either the Gamma density probability function (when the number of connections was small) or the central limit theorem (when the number of connections was large enough). Liu et al. (2010) calculated the conditional entropy $H(Y|X)$ where Y and X are two of the most widely used traffic variables: source and destination Ip addresses.
Amiri et al. (2011) used an estimator of MI developed by Kraskov et al. (2004), which employs entropy estimates from k-nearest neighbors distances. Velarde-Alvarado et al. (2009) estimated entropy values using the balanced estimator II as a probability estimator.

An anomaly in network traffic is a data pattern that does not conform to those representing a normal traffic behavior.

Assuming that 1) $X \in \mathbb{R}^p$ is a p-dimensional real-valued random variable with a domain $S_X \subset \mathbb{R}^p$ representing traffic features, 2) x_i are instances of X, i.e. $x_i \in S_X$, and 3) data patterns of normal behavior are represented by the subspace $S_N \subset S_X$, anomaly detection determines whether an instance x_i belongs to S_N or not.
The space S_X can be partitioned or divided into classes with the help of decision functions, allowing further classification.

Specific Decision Functions

Zhang et al. (2009), Gu et al. (2005) used the KL divergence $D_1(P||Q)$, in addition, Zhang et al. (2009) classified the abnormal situations into different classes. Coluccia et al. (2013) employed both KL divergence and Maximum entropy. Yan et al. (2008) used $D_{0.5}(P||Q)$.
In Santiago-Paz et al. (2015), a decision function is based on the Mahalanobis distance $\mathbf{d}_M^2(\mathbf{x}_i)$, and a second decision function is given by $f(\mathbf{x}_i) = \sum_i^N \alpha_i k(\mathbf{x}_i, \mathbf{x}) - b$ for One Class-Support Vector Machine (OC-SVM), where $k(\mathbf{x}_i, \mathbf{x})$ is a kernel. Huang et al. (2006) computed the Rényi entropy ($q = 3$) of the Coiflet and Daubechies wavelets.
In Velarde-Alvarado et al. (2009), used the proportional uncertainty (PU) and the method of remaining elements (MRE) to detect anomalies. Tellenbach et al. (2011) used Kalman filter, PCA, and KLE as anomaly detection methods. Ma et al. (2014) established a function decision based on the entropy of the source IP address $\hat{H}_s$ and the entropy of the destination IP address $\hat{H}_d$. In Berezínski et al. (2015), Özçelik and Brooks (2015) a function decision based on entropy and a range of values was used to detect anomalies.

Table 1: Results of network traffic anomaly detection using entropy.

Author	Information metric	Database	Anomaly	TNR [%]
Gu et al. (2005)	KL divergence	–	Portscan	91.0
Zhang et al. (2009)	KL divergence	MIT-DARPA	DoS	87.10
			Probe	68.18
			R2L	79.49
			U2R	60.87
Liu et al.(2010)	Conditional entropy	CAIDA	DDoS	93.0
Ferreira et al. (2011)	Shannon, Rényi, Tsallis	KDD Cup 99	DoS	93.18
			Probe	79.20
			R2L	97.76
			U2R	95.05
Amiri et al. (2011)	Mutual Information	KDD Cup 99	DoS	90.02
			Probe	99.97
			R2L	99.98
			U2R	95.0
Santiago-Paz et al.(2015)	Shannon, Rényi, Tsallis	LAN, MIT-DARPA subset	Worms, DoS, Portscan	99.83

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

Gupta et al. (2014) state that given: 1) a training data set of the form $\{(x_i, \mathbf{y}_i)\}$, where $x_i \in S_X$ is a feature vector or data pattern and $\mathbf{y}_i \subset \{1, \ldots, G\}$ is the subset of the G class labels that are known to be correct labels for x_i, 2) a discriminant function $f(x; \beta_g)$ with class-specific parameters β_g for each class with $g = 1, \ldots, G$; then class discriminant functions are used to classify an instance x as the class label that solves $arg\, max_g f(x; \beta_g)$.

Lakhina et al. (2005) apply two clustering algorithms: k-means and hierarchical agglomeration, using a vector $\tilde{\mathbf{h}} = [\tilde{\mathbf{H}}(\text{srcIP}), \tilde{\mathbf{H}}(\text{dstIP}), \tilde{\mathbf{H}}(\text{srcPort}), \tilde{\mathbf{H}}(\text{dstPort})]$. Xu et al., (2005) define three "free" feature dimensions and introduce an *"Entropy-based Significant Cluster Extraction Algorithm"* for clustering.
Lima et al. (2011)use the WEKA[1] Simple K-Means algorithm. SVM is applied by Tellenbach et al. (2011) to classify the anomalies. Yao et al. (2012) use the Random Forests Test.
Santiago-Paz et al. (2014) present the *Entropy and Mahalanobis Distance (EMD) based Algorithm* to define elliptical regions in the feature space. In Santiago-Paz et al. (2015), OC-SVM and k-temporal nearest neighbors are used to improve accuracy in classification.

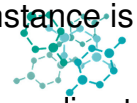
[1] http://www.cs.waikato.ac.nz/ml/weka/

The Classifier Metrics

Given a classifier and an instance, there are four possible outcomes:[2] TN, FP, FN, and TP. With these entries, the following statistics are computed: Accuracy (AC) is the proportion of the total number of predictions that were correct: $AC = \frac{TN+TP}{TN+FP+FN+TP}$; True Positive Rate (TPR) is the proportion of positive cases that were correctly identified: $TPR = \frac{TP}{FN+TP}$; True Negative Rate (TNR) is the proportion of negative cases that were classified correctly: $TNR = \frac{TN}{TN+FP}$; False Negative Rate (FNR) is the proportion of positive cases that were incorrectly classified as negative: $FNR = \frac{FN}{FN+TP}$; and F-measure is a measure of a test's accuracy: $F\text{-measure} = \frac{2*TPR*AC}{TPR+AC}$. In addition, Receiver Operating Characteristic[3] (ROC) graphs illustrate the performance of a classifier.

[2] TN is the number of correct predictions that an instance is negative, FP is the number of incorrect predictions that an instance is positive, FN is the number of incorrect predictions that an instance is negative, and TP is the number of correct predictions that an instance is positive.

[3] ROC graphs are two-dimensional graphs in which an (FP rate, TP rate) pair corresponding to a single point in Receiver Operating Characteristic space.



- Nowadays, there is no public database large enough to exhaustively test and compare different algorithms in order to extract significant conclusions about their performances and their capabilities of classification. Therefore, the construction of a common database with real “normal” and anomalous traffic for the evaluation of A-NIDS is needed.
- The value of the q parameter for generalized entropies is found experimentally; its correct choice for the best anomaly detection is an open research problem.
- For different networks, the larger the slot size, the more different the entropy behaviors. In the near future, this behavior including more and recent traces in order to determine whether the learned model from a certain network can be used in a different network should be addressed.
- Another open issue is related to the adequate window size for reducing the data volume, ensuring good entropy estimates and early detection of anomalies.
- The set of decision functions and classifiers with new closeness and farness entropy-based measures should be enhanced.

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015