On Entropy in Network Traffic Anomaly Detection

Jayro Santiago-Paz * and Deni Torres-Roman

CINVESTAV, Campus Guadalajara, Av. del Bosque 1145, Col. El Bajio, Zapopan, Mexico.

* Author to whom correspondence should be addressed; E-Mail: jsantiago@gdl.cinvestav.mx; Tel.: +52(33)-3777-3600, +52(33)-3777-3609.

Published: 13 November 2015

Abstract: Different systems, e.g., the anomaly-based network intrusion detection system (A-NIDS), have been continuously developed in order to ensure integrity, availability, and confidentiality of networks. In this paper, we present a structured and comprehensive overview of the research into entropy-based A-NIDS with the intention of providing researchers a quick introduction to essential aspects of this topic. The main components of the general architecture of A-NIDS based on Entropy are discussed. The achieved high detection rates prove the effective use of entropy. Finally, some open issues in entropy-based network traffic anomaly detection are also highlighted.

Keywords: Network traffic anomaly detection; entropy; network security; A-NIDS, Mutual information, KL divergence

PACS classifications: one, two, three

1. Introduction

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 [1], Balanced II [2]), are used.

Chandola et al. (2009) [3] 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. This concept is accepted by the Internet Community.
A generic architecture of entropy-based A-NIDS is presented in figure 1, see [4]. It usually consists of two stages: the training and the testing stage. In the training stage, using a database of “normal” or free-anomaly network traffic, feature extraction, windowing, and entropy calculation modules, a “normal” profile is found. In the testing stage, using 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.

Table 1 presents the notation used in this paper.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_\rho(P</td>
<td></td>
</tr>
<tr>
<td>$D_N(P</td>
<td></td>
</tr>
<tr>
<td>$f(x; \beta_g)$</td>
<td>Discriminant function where $\beta_g$ are class-specific parameters</td>
</tr>
<tr>
<td>$H^R(\bullet, q)$</td>
<td>Rényi entropy</td>
</tr>
<tr>
<td>$H^S(\bullet)$</td>
<td>Shannon entropy</td>
</tr>
<tr>
<td>$H^T(\bullet, q)$</td>
<td>Tsallis entropy</td>
</tr>
<tr>
<td>$H(X;Y)$</td>
<td>Joint entropy</td>
</tr>
<tr>
<td>$H(Y</td>
<td>X)$</td>
</tr>
<tr>
<td>$I(X;Y)$</td>
<td>Mutual information</td>
</tr>
<tr>
<td>$p_i$</td>
<td>Probability of occurrence of $x_i$ element</td>
</tr>
<tr>
<td>$p(x,y)$</td>
<td>Joint probability</td>
</tr>
<tr>
<td>$p(x</td>
<td>y)$</td>
</tr>
<tr>
<td>$q$</td>
<td>Parameter of generalized entropies</td>
</tr>
<tr>
<td>$S_X$</td>
<td>$m$-dimensional space of all traffic features</td>
</tr>
<tr>
<td>$S_Y$</td>
<td>$p$-dimensional space of free-anomaly traffic features</td>
</tr>
<tr>
<td>$W_i(L, \tau)$</td>
<td>$i$-th sliding window with $L$ packets and $\tau$ as the overlapping parameter</td>
</tr>
<tr>
<td>$X, Y, Z \in \mathbb{R}^n, n = 1, 2, \ldots, k$</td>
<td>Random variables</td>
</tr>
<tr>
<td>$x, y, z$</td>
<td>Instances of random variables $X, Y, Z$.</td>
</tr>
</tbody>
</table>

Chandola et al. (2009) [3] present “a survey in anomaly detection, grouping existing techniques into different categories based on the underlying approach adopted by each technique”. Bhuyan et al. (2014)
[5] provides an excellent survey in network anomaly detection describing six distinct classes of methods and systems. However, these surveys are not focused directly on entropy-based A-NIDS.

This paper presents a structured and comprehensive overview of the research into entropy-based A-NIDS with the intention of providing researchers a quick introduction to essential aspects of this topic. Using a general architecture of entropy-based A-NIDS, the different techniques proposed in the state-of-the-art of the main modules are shown. The measures of information used by researchers and the most important metrics for testing the performance of the detection and classification are presented. We also highlight some open issues in entropy-based network traffic anomaly detection.

The next sections describe the main modules of the entropy-based A-NIDS. Section 2 presents the main databases used for researchers in the field. Section 3 shows the most commonly employed features in network traffic and the windowing approach. Section 4 introduces the mathematical background, including Shannon entropy, generalized entropies, Mutual information, Kullback-Leibler divergence, conditional entropy, and the techniques to estimate their values. Section 5 presents the decision functions defined by the researchers in the detection stage. Section 6 shows the approaches employed in the classification stage and the most widely used metrics for evaluating the A-NIDS. Section 7 contains the conclusions of the paper and some important open issues of this topic.

2. Databases

Different databases have been used to evaluate the A-NIDS, and these databases are divided into two groups: synthetic and real.

The synthetic databases are generated artificially, e.g., the MIT-DARPA 1998, 1999, 2000 databases\(^1\), which include five major categories: Denial of Service Attacks (DoS), User to Root Attacks (U2R), Remote to User Attacks (R2U) and probes.

Some real public databases are: CAIDA\(^2\), which contains anonymized passive traffic traces from high-speed Internet backbone links, and the traffic data repository, maintained by the MAWI\(^3\) 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\[^6\]) or traffic collected from backbone in SWITCH, Abilene, and Géant.

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.

3. Feature Extraction

Motoda H. and Liu H. (2002) \[^7\] state that 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 \[^8,9\]. Feature extraction is a process that extracts a set of new

---

\(^1\) http://www.ll.mit.edu/ideval/index.html
\(^2\) https://www.caida.org/data/passive/passive_2012_dataset.xml
\(^3\) http://mawi.wide.ad.jp/mawi/
features from the original features through some functional mapping [10]. 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 [11], Mutual Information and linear correlation [12], decision tree [13], and maximum entropy [14].

In network traffic, the most commonly employed features are [2,15–19]: 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) [20] divided the size of packets into seven types and Gu et al. (2005) [21] defined 587 packet classes based on the port number.

At flow 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 [12,13]. On the other hand, Tellenbach et al. (2011) [22] used source port, country code and others, constructing the TES as input data.

3.1. Windowing in Network Traffic

Window-based methods group consecutive packets or flows based on a sliding window. The \( i \)th window of size \( L \) packets is represented as \( W_i(L, \tau) = \{\text{pack}_{k}, \text{pack}_{k+1}, \ldots, \text{pack}_{k+L}\} \), with \( k = iL - i\tau \), where \( \tau \) 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 [15,16,23–25], 30 min, 1 min, 100 sec, 5 sec and 0.5 sec. Some researchers use windows with a fixed length \( L = 4096 \) [19], 1000 [26], and 32 [4] packets. Therefore, the main objective of windowing is to reduce the data volume.

4. Entropy Concepts Used in Network Traffic Anomaly Detection

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

\[
H^S(X) = - \sum_{i=1}^{M} p_i \log (p_i).
\] (1)

Based on the Shannon entropy [27], Rényi [28] and Tsallis [29] defined generalized entropies, which are related to the \( q \)-deformed algebra. The Rényi entropy is defined as:

\[
H^R(X,q) = \frac{1}{1-q} \log \left( \sum_{i=1}^{M} p_i^q \right)
\] (2)
and the Tsallis entropy is

---

4 An IP flow corresponds to an IP port-to-port traffic exchanged between two IP addresses during a period of time \( T \).
\[ H^T(X, q) = \frac{1}{q-1} \left( 1 - \sum_{i=1}^{M} p_i^q \right), \]  
(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.,
\[ \bar{H}(X) = \frac{H(X)}{H_{\max}(X)}. \]  
(4)

For generalized entropies, the values \( q = 0.9 \) and \( q = 1.1 \) to detect DoS and DDoS attacks have been used by [30,31], respectively. The sets \( q \in \{-3\} \cup \{-2, -1.75, ..., 1.75, 2\} \), \( q = \{1, 2, 3, ..., 15\} \) and \( q = \{1, 2, 3, ..., 10\} \) were used to detect DDoS and scanning attacks, and low-rate and high-rate DDoS by [22,32,33].

4.1. Kullback-Leibler divergence

Consider two complete discrete probability distributions \( P = (p_1, p_2, ..., p_n) \) and \( Q = (q_1, q_2, ..., q_n) \), with \( \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1, \ 1 \geq p_i \geq 0, \ 1 \geq q_i \geq 0, \ i = 1, 2, ..., n. \) The information divergence is a measure of the divergence between \( P \) and \( Q \) and is defined by [28]:
\[ D_\rho(P||Q) = \frac{1}{\rho - 1} \log \left( \sum_{i=1}^{n} p_i^\rho q_i^{1-\rho} \right), \ \rho \geq 0, \]  
(5)
where \( \rho \) 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 [34] is obtained
\[ D_1(P||Q) = \sum_{i=1}^{n} \left( p_i \log \left( \frac{p_i}{q_i} \right) \right), \ \rho \to 1. \]  
(6)

4.2. Mutual information

The conditional entropy of a variable \( Y \) given \( X \), with alphabet \( \mathcal{X} \) and \( \mathcal{Y} \), respectively, is defined as:
\[ H(Y|X) = - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log (p(y|x)) \]  
(7)
\[ = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log (p(y|x)). \]  
(8)

The mutual information (MI) [35] 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:
\[ I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X;Y) \]  
(9)
where $H(\bullet)$ is entropy, $H(X|Y)$ and $H(Y|X)$ are conditional entropies, $H(X;Y)$ is the joint entropy of $X$ and $Y$, defined as

$$H(X;Y) = -\sum_{x\in X} \sum_{y\in Y} p(x,y) \log (p(x,y))$$

(10)

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

The MI equation can be written as:

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

(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 $p(x,y)$.

### 4.3. Entropy calculation

As the full probability distribution is generally not known or not completely known, 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) [36] 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) [18] 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) [12] used an estimator of MI developed by Kraskov [37], which employs entropy estimates from $k$-nearest neighbors distances. Velarde-Alvarado et al. (2009) [2] estimated entropy values using the balanced estimator II as a probability estimator.

### 5. Anomaly detection

An anomaly in network traffic is a data pattern that does not conform to those representing a normal traffic behavior. Therefore, anomaly detection is a broad field, where numerous anomaly detection methods are used for different applications.

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.
5.1. Specific Decision Functions

The KL divergence $D_1(P||Q)$ is used as decision functions in [20,21]. In addition, [20] classified the abnormal situations into different classes. In [38], both KL divergence and Maximum entropy are used. Yan et al. (2008) [25] used $D_{0.5}(P||Q)$.

In Santiago-Paz et al. (2015) [4], a decision function is based on the Mahalanobis distance [39] $d_M^2(x_i)$, and a second decision function is given by $f(x_i) = \sum_i^N \alpha_i k(x_i, x) - b$ for One Class-Support Vector Machine (OC-SVM), where $k(x_i, x)$ is a kernel. Huang et al. (2006) [40] computed the Rényi entropy ($q = 3$) of the Coiflet and Daubechies wavelets.

In Velarde-Alvarado et al. (2009) [2], used the proportional uncertainty (PU) and the method of remaining elements (MRE) to detect anomalies. Tellenbach et al. (2011) [22] used Kalman filter, PCA, and KLE as anomaly detection methods. Ma et al. (2014) [41] established a function decision based on the entropy of the source IP address $H_s$ and the entropy of the destination IP address $H_d$. In [42,43], a function decision based on entropy and a range of values was used to detect anomalies.

The use of the entropy allows the A-NIDS to achieve high detection rates, see table 2. In addition, new measures based on entropy should be studied and used as a basis for other decision functions.

Table 2. Results of network traffic anomaly detection using entropy.

<table>
<thead>
<tr>
<th>Author</th>
<th>Information metric</th>
<th>Database</th>
<th>Anomaly</th>
<th>TNR [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gu et al. (2005)</td>
<td>KL divergence</td>
<td>–</td>
<td>Portscan</td>
<td>91.0</td>
</tr>
<tr>
<td>Zhang et al. (2009)</td>
<td>KL divergence</td>
<td>MIT-DARPA</td>
<td>DoS</td>
<td>87.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Probe</td>
<td>68.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R2L</td>
<td>79.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>U2R</td>
<td>60.87</td>
</tr>
<tr>
<td>Liu et al. (2010)</td>
<td>Conditional entropy</td>
<td>CAIDA</td>
<td>DDoS</td>
<td>93.0</td>
</tr>
<tr>
<td>Ferreira et al. (2011)</td>
<td>Shannon,</td>
<td>KDD Cup 99</td>
<td>DoS</td>
<td>93.18</td>
</tr>
<tr>
<td></td>
<td>Rényi, Tsallis</td>
<td></td>
<td>Probe</td>
<td>79.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R2L</td>
<td>97.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>U2R</td>
<td>95.05</td>
</tr>
<tr>
<td>Amiri et al. (2011)</td>
<td>Mutual</td>
<td>KDD Cup 99</td>
<td>DoS</td>
<td>90.02</td>
</tr>
<tr>
<td></td>
<td>Information</td>
<td></td>
<td>Probe</td>
<td>99.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R2L</td>
<td>99.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>U2R</td>
<td>95.0</td>
</tr>
</tbody>
</table>

6. Classification

Gupta et al. (2014) [44] state that given: 1) a training data set of the form $\{(x_i, y_i)\}$, where $x_i \in S_X$ is a feature vector or data pattern and $y_i \in \{1, \ldots, G\}$ 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 $\beta_g$ 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 $argmax_g f(x; \beta_g)$.
Lakhina et al. (2005) [16] apply two clustering algorithms: $k$-means and hierarchical agglomeration, using a vector $\tilde{h} = [\tilde{H}(\text{srcIP}), \tilde{H}(\text{dstIP}), \tilde{H}(\text{srcPort}), \tilde{H}(\text{dstPort})]$. Xu et al., (2005) [23] define three “free” feature dimensions and introduce an “Entropy-based Significant Cluster Extraction Algorithm” for clustering.


Santiago-Paz et al. (2014) [19] present the Entropy and Mahalanobis Distance (EMD) based Algorithm to define elliptical regions in the feature space. In [4], OC-SVM and $k$-temporal nearest neighbors are used to improve accuracy in classification.

6.1. The Classifier Metrics

Given a classifier and an instance, there are four possible outcomes:\textsuperscript{6} $TN$, $FP$, $FN$, and $TP$. With these entries, the following statistics are computed [46]: 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: $TPR = \frac{TP}{TN+TP}$; 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: $F$-measure $= \frac{2 \times TPR \times AC}{TPR + AC}$. In addition, Receiver Operating Characteristic\textsuperscript{7} (ROC) graphs illustrate the performance of a classifier.

Although, there are several types of distances, classifiers based on new closeness and farness measures of data patterns and pattern clusters should be studied.

7. Conclusions

This paper provides a structured and comprehensive overview of the state-of-the-art in entropy-based A-NIDS. Using a general architecture of Entropy-based A-NIDS, the different techniques proposed in the state-of-the-art of the main modules are shown. The measures of information used by researchers and the most commonly employed metrics for testing the performance of the detection and classification are presented. The achieved high detection rates prove the effective use of entropy.

Open Issues

\textsuperscript{5} http://www.cs.waikato.ac.nz/ml/weka/

\textsuperscript{6} $TN$ 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.

\textsuperscript{7} 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.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**


© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).