Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

**Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**
Preprint

# Unsupervised Hilbert Huang Transform with Pruned Exact Linear Time Algorithm for Anomaly Detection in Web Data

Emilio Gerardo Sotto Riveros[1]
Cristian Cappo[2]
Christian Schaerer[3]
Facultad Politécnica, UNA, San Lorenzo, Paraguay

With the massive use of digital technologies, many activities in society have transitioned to the web, from shopping and social interactions to business, industry, and, unfortunately, a crime. Recent reports reveal that criminals targeted more companies in 2020 than in 2019. The survey found that 28% of companies that suffered attacks were attacked more than five times in 2020. Studies on the changing cost of cybercrime show that between 2012 and 2019, in only two types of cybercrime, the kidnapping of data (Ransomware), and the identity theft of email account owners (BEC- Business Email Compromise) moved more than US$3.3 billion in the US and UK alone.

The traffic behavior in these web applications has characteristics such as non-stationarity, which presents random fluctuations, periodic patterns evolving with different frequencies, and trends over time. In this context, signal processing techniques arise as valid options to detect such attacks. For example, wavelet analysis permits detecting anomalies in web data [6] and network traffic [7]. Combining signal processing with other techniques has good results; for example, tracking anomalies in IP networks using statistical signal processing based on abrupt change detection effectively detects several network anomalies [8].

This work proposes an unsupervised anomaly detection algorithm based on Hilbert-Huang Transform(HHT) combined with a method to search change-points (Pruned Exact Linear Time - PELT) in data corresponding to the anomalies. The HHT is a method used in signal processing algorithms for extracting relevant information in data. It performs well in analyzing signals with fluctuations and patterns with many frequencies over time [4]. PELT is a method to find changes in statistical properties in a signal, minimizing a cost function [5].

Several web traffic characteristics are considered to find the abrupt variations of frequency and amplitude in signals representing data, such as the length of the request, the frequency of groups of characters, and characters' entropy [3]. Each of these web traffic characteristics is considered an independent signal. For each of them, the HHT is applied to decompose the signal, with the Empirical Mode Decomposition (EMD) process, into a set of Intrinsic Mode Functions (IMF). After decomposing the signal, we use only the first IMF in this approach. The first IMF generally carries the most oscillating (high frequency) components in terms of natural waves, providing enough information to detect attacks. The HHT is applied to the first IMF to obtain the instantaneous potency or energy density.We select the energy density to detect anomalies since this will change significantly with abrupt changes. Finally, the PELT method is used to find the change points where the energy changes abruptly concerning the data set's variance as a threshold for detecting the changes in its values.

[1]emiliosotto@gmail.com
[2]ccappo@pol.una.py - *corresponding author*
[3]cschaerer@pol.una.py

The resulting method achieves a linear computational cost to the number of data points. It reduces the computational cost (concerning state-of-the-art) without affecting the precision of the resulting segmentation. For the experiments were the public datasets CSIC 2010 [1] and CSIC TORPEDA 2012 [2], both contain HTTP requests made to an electronic commerce application, simulating different scenarios such as customer registration or product purchase.

The results for the Entropy feature show an F1-score (that relates true positives and false positives, with 1 being the best result) of 0.914, with windows of 256 and 512 data, in a dataset with 2% anomalies, presenting a False-Positive Rate (FPR) of 0.0003 in the worst case. Averaging an F1-score = 0.86 and FPR = 0.001 in all window sizes. With the Length feature analysis, the best results were in the window sizes of 256 and 128 data with 2% anomalies, presenting an F1-score = 0.876, with a FPR = 0.0006 in the worst case and averaging an F1-score equal to 0.832 and an FPR = 0.11 in all window sizes for 2% anomalies. These results are promising for an unsupervised (untrained) algorithm concerning other state-of-the-art algorithms.

# References

[1] CSIC. **CSIC Dataset 2010**. `https://www.isi.csic.es/dataset/`. Acceso: 17/10/2016. 2010.

[2] CSIC. **Torpeda CSIC Dataset 2012**. `https://www.tic.itefi.csic.es/torpeda/datasets.html`. Acceso: 17/10/2016. 2012.

[3] Ralf Funk, Nico Epp, et al. "Anomaly-based web application firewall using http-specific features and one-class svm". In: **Revista Eletrônica Argentina-Brasil de Tecnologias da Informação e da Comunicação** 2.1 (2018).

[4] N. E Huang et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis". In: **Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences** 454.1971 (1998), pp. 903–995.

[5] Rebecca Killick, Paul Fearnhead, and I.A. Eckley. "Optimal Detection of Changepoints With a Linear Computational Cost". In: **Journal of the American Statistical Association** 107 (Dec. 2012), pp. 1590–1598. DOI: `10.1080/01621459.2012.737745`.

[6] A. Kozakevicius et al. "URL query string anomaly sensor designed with the bidimensional Haar wavelet transform". In: **International Journal of Information Security** 14.6 (2015), pp. 561–581.

[7] K. Limthong, P. Watanapongse, and F. Kensuke. "A wavelet-based anomaly detection for outbound network traffic". In: **8th Asia-Pacific Symposium on Information and Telecommunication Technologies**. 2010, pp. 1–6.

[8] M. Thottan and Chuanyi Ji. "Anomaly detection in IP networks". In: **IEEE Transactions on Signal Processing** 51.8 (2003), pp. 2191–2204. DOI: `10.1109/TSP.2003.814797`.