proceedings

MDPI

Extended Abstract

Network Data Unsupervised Clustering to Anomaly Detection †

Manuel López-Vizcaíno *, Carlos Dafonte, Francisco J. Nóvoa, Daniel Garabato and M. A. Álvarez

CITIC, UDC, Campus de Elviña s/n, 15071 A Coruña, Spain; carlos.dafonte@udc.es (C.D.); francisco.javier.novoa@udc.es (F.J.N.); daniel.garabato@udc.es (D.G.); marco.antonio.agonzalez@udc.es (M.A.Á.)

* Correspondence: manuel.fernandezl@udc.es

† Presented at the XoveTIC Congress, A Coruña, Spain, 27–28 September 2018.

Published: 17 September 2018

Abstract: In these days, organizations rely on the availability and security of their communication networks to perform daily operations. As a result, network data must be analyzed in order to provide an adequate level of security and to detect anomalies or malfunctions in the systems. Due to the increase of devices connected to these networks, the complexity to analyze data related to its communications also grows. We propose a method, based on Self-Organized Maps, which combine numerical and categorical features, to ease communication network data analysis. Also, we have explored the possibility of using different sources of data.

Keywords: Self-Organizing Maps; IDS; network security; categorical SOM; visualization; unsupervised clustering

1. Introduction

These days network data analysis has become essential to provide adequate levels of security in mid and big sized networks. The number of connected devices has increased to 20 thousand million of devices in 2017 and will exceed 30 thousand million devices in 2020, as it is reflected in the forecast from Statista [1]. As a result of the exponential increase of the traffic generated, classical analysis techniques based on payload packet inspection become unfeasible [2].

One possible approximation to this problem is unsupervised clustering. These techniques allow to cluster elements with similar characteristics without prior knowledge, easing the analysis of the data as it was shown in previous research [3]. In particular for the scope of this work we have chosen Self-Organized Maps (SOM) technique [4] as it allows to perform clustering as well as dimensionality reduction. Also, this technique has been successfully applied to Intrusion Detection Systems [5,6].

As it is said in [7], a habitual traffic profile, called baseline, is present in communication networks. Different kind of attacks present deviations from this baseline and these features could be used to detect certain anomalies in traffic behavior (DoS [8], DDoS [9], brute force attacks [10]).

The objective of this work is to present a method to ease the analysis of communication network data. Providing a system to allow the study and detection of anomalies out of data gathered from different sources.

2. Methods

For the scope of this work, to generate the clusters, we have modified SOM technique to accept numerical and categorical features, as explained in [11]. Besides we have only used information present on IP packet headers or values derived from them. From the data available on the IP header we have selected a number of features such as source, destination, source port, destination port, protocol, duration and bytes transmitted.

Two different datasets have been used to perform the experiments. One the one hand, the UNB ISCX which is a synthetic and labeled flow dataset generated by the Cybersecurity Institute of Canada intended for Intrusion Detection research [12]. On the other hand, we have used a log dataset gathered in the firewall of the Computer Science faculty of the University of A Coruña. We have divided both datasets to use the 80% of them to train and the 20% were left to test.

Before the clustering technique could be applied, a preprocessing step must be preformed in order to categorize certain variables and to normalize numeric features. We performed a shallow approach to the analysis of the map configuration with three different map sizes: 10 × 10 (100 neurons), 20 × 20 (400 neurons) and 30 × 30 (900 neurons). Increasing the number of neurons could help to get better detection rates but it also rises the complexity of map analysis.

Finally, to evaluate the clustering we have used some tags referred to the nature of the connection. In the case of the flow dataset we have used the synthetic labels showing if it is part of an attack or normal traffic. On the other dataset we have taken the actions of the firewall as an approach. In the last case it also allows the revision of the firewall rules by studying the misclassification.

3. Results

The aim of these experiments was to determine if a mixed numerical-categorical version of SOM technique was suitable for network data classification, by using only IP header information gathered from different sources. Also, other objective was to study how the information obtained could be used in relation with the source of the data. For example, network flows could help to detect attacks and firewall logs analysis could help to detect misconfiguration.

As it can bee seen in Table 1 where the results are shown both for the flow labeled dataset (ISCX) and the firewall log (FIC), there are similarities between their results. Bigger map sizes tend to increase the performance of the technique for both datasets but with the drawback of a more complex map analysis. Also, it should be noticed that the better overall results are achieved with logs rather than with flows.

Table 1. Experiment results.

	Flows			**Logs**		
	10 × 10	**20 × 20**	**30 × 30**	**10 × 10**	**20 × 20**	**30 × 30**
Sensitivity	90.33%	94.09%	94.28%	87.78%	90.20%	94.66%
Specificity	98.36%	99.00%	99.26%	96.37%	99.24%	99.12%
Precision	67.06%	77.80%	82.44%	86.56%	96.95%	96.62%
Accuracy	98.07%	98.83%	99.08%	94.56%	97.34%	98.18%

4. Discussion

As it can be seen in the results, despite the differences between both datasets, we can conclude that the technique could be applied to different sources of network data. This difference should be studied in order to determine if it is related to the nature of the dataset, the differences in the features or other reasons. Also, additional research using other sources of data and different configurations over the proposed technique should be performed.

Funding: This work has received financial support from the Xunta de Galicia (Centro singular de investigación de Galicia accreditation 2016-2019) and the European Union (European Regional Development Fund - ERDF).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Statista. IHS. Internet of Things (Iot) Connected Devices Installed Base Worldwide from 2015 to 2025 (in Billions) 2018. Available online: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/ (accessed on 17 September 2018).
2. Umer, M.F.; Sher, M.; Bi, Y. Flow-based intrusion detection: Techniques and challenges. *Comput. Secur.* **2017**, *70*, 238–254, doi:10.1016/j.cose.2017.05.009.
3. Buczak, A.; Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1153–1176, doi:10.1109/COMST.2015.2494502.
4. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59–69, doi:10.1007/BF00337288.
5. Ibrahim, L.M.; Basheer, D.T.; Mahmod, M.S. A comparison study for intrusion database (KDD99, NSL-KDD) based on self organization map (SOM) artificial neural network. *J. Eng. Sci. Technol.* **2013**, *8*, 107–119.
6. Ramadas, M.; Ostermann, S.; Tjaden, B. Detecting Anomalous Network Traffic with Self-organizing Maps. In *Recent Advances in Intrusion Detection*; Vigna, G., Kruegel, C., Jonsson, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 36–54.
7. Xu, K.; Zhang, Z.L.; Bhattacharyya, S. Internet Traffic Behavior Profiling for Network Security Monitoring. *IEEE ACM Trans. Netw.* **2008**, *16*, 1241–1252, doi:10.1109/TNET.2007.911438.
8. Fadlullah, Z.M.; Taleb, T.; Vasilakos, A.V.; Guizani, M.; Kato, N. DTRAB: Combating Against Attacks on Encrypted Protocols Through Traffic-Feature Analysis. *IEEE ACM Trans. Netw.* **2010**, *18*, 1234–1247, doi:10.1109/TNET.2009.2039492.
9. Lee, K.; Kim, J.; Kwon, K.H.; Han, Y.; Kim, S. DDoS attack detection method using cluster analysis. *Expert Syst. Appl.* **2008**, *34*, 1659–1665, doi:10.1016/j.eswa.2007.01.040.
10. Hofstede, R.; Jonker, M.; Sperotto, A.; Pras, A. Flow-Based Web Application Brute-Force Attack and Compromise Detection. *J. Netw. Syst. Manag.* **2017**, *25*, 735–758, doi:10.1007/s10922-017-9421-4.
11. Del Coso, C.; Fustes, D.; Dafonte, C.; Nóvoa, F.J.; Rodríguez-Pedreira, J.M.; Arcay, B. Mixing numerical and categorical data in a Self-Organizing Map by means of frequency neurons. *Appl. Soft Comput.* **2015**, *36*, 246–254, doi:10.1016/J.ASOC.2015.06.058.
12. Shiravi, A.; Shiravi, H.; Tavallaee, M.; Ghorbani, A.A. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Comput. Secur.* **2012**, *31*, 357–374, doi:10.1016/j.cose.2011.12.012.