Effectiveness of Hard Clustering Algorithms for Securing Cyber Space

Sakib Mahtab Khandaker[1,2](✉), Afzal Hussain[1,2], and Mohiuddin Ahmed[1,2]

[1] Islamic University of Technology, Gazipur City, Bangladesh
{sakibmahtab,afzalhussain}@iut-dhaka.edu,
m.ahmed.au@ieee.org
[2] Canberra Institute of Technology, Reid, Australia

Abstract. In the era of big data, it is more challenging than before to accurately identify cyber attacks. The characteristics of big data create constraints for the existing network anomaly detection techniques. Among these techniques, unsupervised algorithms are superior than the supervised algorithms for not requiring training data. Among the unsupervised techniques, hard clustering is widely accepted for deployment. Therefore, in this paper, we investigated the effectiveness of different hard clustering techniques for identification of a range of state-of-the-art cyber attacks such as *backdoor, fuzzers, worms, reconnaissance* etc. from the popular UNSW-NB15 dataset. The existing literature only provides the accuracy of identification of the all types of attacks in generic fashion, however, our investigation ensures the effectiveness of hard clustering for individual attacks. The experimental results reveal the performance of a number of hard clustering techniques. The insights from this paper will help both the cyber security and data science community to design robust techniques for securing cyber space.

Keywords: Network traffic analysis · Cyber attacks
Unsupervised clustering · Big data

1 Introduction

Technology have been increasing at breakneck speed and with it the amount of data generated thus the word big data has become ubiquitous in both academic and industrial domains. Although it is a relatively new term which was only coined in 2008 [2], it became a buzzword after the Mckinsley Global Institute report [3] but there still remains confusion as to the amount of data denoted by it. Big data is the driving force behind many digital transformation waves like internet of things, data science and artificial intelligence. The term can properly be defined using the "5V" – volume (referring to the large amount of data requiring no traditional processing methods), velocity (the high speed at which data is produced), variety (denoting the structured, semi structured and unstructured form of generated data), veracity (the quality of the data)

A.-S. K. Pathan et al. (Eds.): SGIoT 2018, LNICST 256, pp. 113–120, 2019.
https://doi.org/10.1007/978-3-030-05928-6_11

and value (referring to the added value big data brings) [1]. By reflecting on all existing definitions, Big data has been [4] defined as *"Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value."* Use of big data implies dealing with large amount of structured, unstructured and semi structured data ranging from petabytes, exabytes to even yotabytes [5]. Big data comes with a cost, it creates information security and privacy breach but big data analytics on the other hand promises detection and prevention of such cyber attacks. Real time analysis of patterns, outlier detection and attack recognition through correlation and machine learning is where big data analysis can play a big role in cyber security. Big data holds the key to solve information security threat by providing real time actionable insights [6]. Extension of traditional security and real time large scale analysis, compassion, anomaly detection of heterogeneous data sets at greater speed have been made possible with the newer big data technologies like Hadoop ecosystem, stream mining, complex-event processing and NoSQL databases [7].

The contribution of this paper lies in the fact that 3 different unsupervised algorithms were used here for anomaly detection namely - kmeans, kmedoids and kmodes. We used them to identify different types of cyber attacks like - analysis, exploits, reconnaissance, worms, backdoor, fuzzers, shellcode, DoS and generic attacks. Among the three, kmodes showed the highest accuracy with an overall accuracy of 69%. It was able to segregate different types of cyber attacks from the UNSW-NB15 dataset which is used as a standard for cyber security research. This paper sheds light on the potential of unsupervised algorithms in cyber security.

1.1 Roadmap

The Rest of the paper is organized as follows: Sect. 2 discusses about different network anomaly techniques then Sect. 3 focuses on hard clustering algorithms. In Sect. 4 different types of cyber attacks are described. Finally experimental results are highlighted in Sect. 5 and Sect. 6 is where we draw the conclusion.

2 Network Anomaly Detection Techniques

Anomaly detection can be referred to the detection of nonconforming patterns in regular data. These abnormal data are sometimes termed as outliers, exceptions, peculiarities or contaminants in different paradigm. The most common terms used terms are anomaly and outlier. Actionable intelligence anomaly detection provides where its most importance lies. Outlier or anomaly detection can be dated back to the 19th century [8]. Some of the challenges which come up when detecting outliers keeping up with constantly evolving normal behavior, availability of labeled data for training or validation and often data contains noise which gives rise to false positives [9]. Compared to statistical approach which focuses on understanding the process of data generation, machine learning focuses on providing the necessary answer based on previous available data thus creating a dynamic system with adaptive capability. Now machine learning based anomaly detection can be divided into three categories - supervised, semi supervised and unsupervised.

Supervised anomaly detection presumes that training data set is available with labeling denoting normal and anomalous instances. The usual approach for such a method is build a predictive model of normal vs anomaly. New data entries are compared with preexisting ones to determine its position among the two cases. Sadly this method has drawbacks. Firstly the training data set contains fewer anomalous incidents compared to normal ones. Secondly obtaining accurate and suitable labels especially in case of anomaly is difficult.

Semi supervised techniques work with the presumption that training data has been labeled for normal instances. As anomalous instances are not labeled here its add flexibility to their application. There are a few number of semi supervised techniques which assumes availability of anomalous cases [10,11]. But these techniques are less used because of the fact that obtaining training data set for all anomalous cases is not possible.

Lastly we have unsupervised anomaly detection technique where there id no need of a training data set. Such techniques make two implicit assumptions. First, the major part of the data set is constituted of normal cases and only a small percentage are anomalies and the other assumptions is that anomalous cases are statistically different from normal cases.

Application of unsupervised algorithm for cyber security has gained momentum in academic paradigm. The accuracy and effectiveness of fixed width clustering, optimized k nearest neighbor support vector machine have been shown in Eskin et al. [12]. Old - meadow et al. showed improvement of cluster accuracy when the said clusters are adaptive [13]. A two tier novel intrusion detection system was proposed in by Zanero et al. [14]. Phad et al. examines IP headers connections to various ports as well as packet headers of Ethernet, IP and transport layers packet headers [15]. Alad et al. detects anomalies in inbound TCP connections to well known ports on the server [16]. Lerad et al. detects TCP stream anomalies like alad but uses a learning algorithm to pick good rules from training set rather than using a fixed set of rules [17]. K-mean was used by Dragon Research and Nairac for novelty detection [18,19]. Gaddam et al. showed a method to detect anomalous activities using k means clustering [20].

3 Hard Clustering

When it comes to the taxonomy of clustering algorithms we can categorize them into two broad groups as shown in Fig. 1. Hierarchical clustering algorithms can either be successive splitting (divisive) or merging (agglomerative) of groups to form a hierarchy based on similarity or a specific measure of distance. Hierarchical algorithms can also be sub classified according to the process the distances or similarities between objects are updated after splitting or merging groups. On the other hand partitional clustering algorithms focus on portioning data based

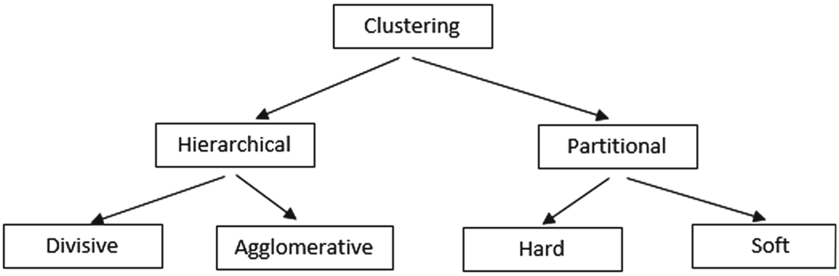

Fig. 1. Taxonomy of clustering algorithm

on distance between objects. On further classification we have hard and soft clustering. A hard clustering algorithm allocates each pattern to a single cluster and no data points is assigned to two clusters while soft clustering assigns degrees of member in several cluster to each input pattern.

The field of cyber security requires a clear distinction between normal cases and anomalous cases. Hard clustering was more suitable for such application. We have evaluated the performance of three hard clustering algorithms- k-means, k-modes and k-medoid for identification of different types of cyber attacks.

4 Attack Description

We have used the UNSW-NB 15 data set for our experiment [21]. A short description of the attacks found in this data set has been described [22].

1. Fuzzer: is a type of attack when the attacker attempts to exploit weakness of a security system by forcing it to crash by providing larger quantities of random data.
2. Analysis: combination of multifarious intrusion techniques including penetration of ports, email address and even web scripts.
3. Backdoor: Such type of attack attempts to bypass the authentication thus allowing unauthorized remote access.
4. DoS: This type of attack overloads memory to a state of unresponsiveness
5. Exploit: A series of instructions designed to work around vulnerabilities of the system from bugs to glitches.
6. Generic: Uses functions to cause collision of block ciphers.
7. Reconnaissance: a probe that attempts to gather information necessary to exploit security weakness.
8. Shellcode: An attack where a simple injection of command into the running system leads to full control of the system.
9. Worm: A replication based attack where the attacker replicates itself into multiple hosts.

5 Experimental Analysis

In order to evaluate the performance of the machine learning algorithms we have used confusion matrix [23]. The confusion matrix has four cells as shown in Table 1, True Positive (TP) which shows the number of attacks detected, True Negative (TN) which shows the number of normal instances detected, False Positive (FP) showing false alarms and False Negative (FN) where the algorithm mistakes an attack for normal case.

To find the accuracy we have used the formula:

$$\frac{TN + TF}{TN + TF + FP + FN} \tag{1}$$

We have evaluated the accuracy of all three algorithms for each type of attacks as shown in Table 2.

The dataset used here is the UNSW-15 which was created using an IXIA Perfectstorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) to generate a set of realistic normal cyber activities and synthetic contemporary attack behaviors. Through the use of a tcpdump tool 100 GB of raw network traffic with a total of 2,540,044 records.

We applied the algorithms on each type of attacks for different split to find accuracy as shown in Fig. 2. It has been found that among the three algorithms kmodes showed the highest accuracy for all types of attacks with an overall average of 70% followed by kmedoid which had an average of 63% for cyber attack detection. The performance of all the algorithms for different split ratios have also been shown in Fig. 2 for better understanding of their performance.

Table 1. Confusion matrix

	Negative	Positive
Actual negative	TN	FP
Actual positive	FN	TP

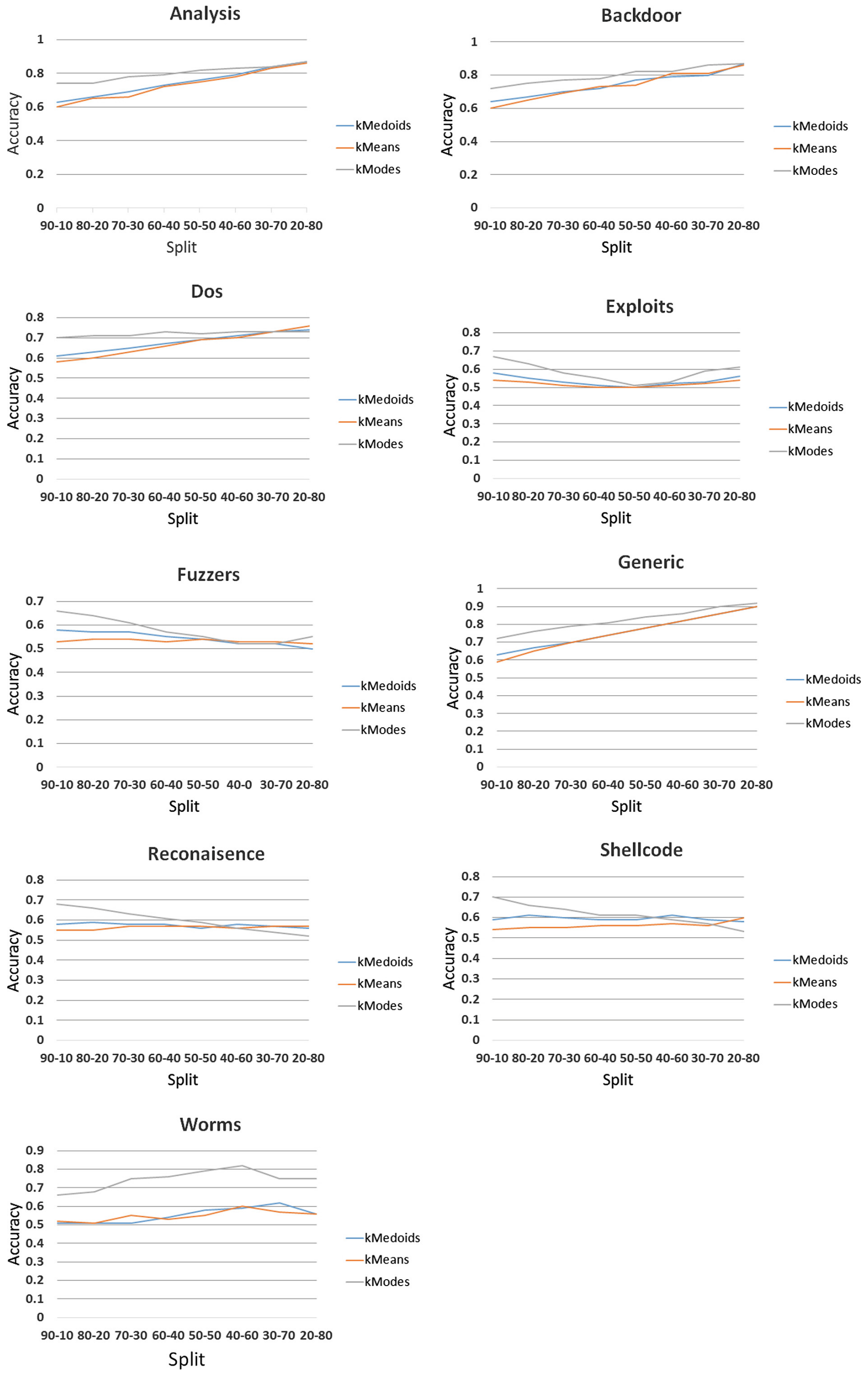

Fig. 2. Accuracy of algorithms

Table 2. Accuracy of different algorithms

	kmedoid	kmean	kmodes
Analysis	0.74625	0.73125	0.80125
Exploits	0.535	0.51875	0.58375
Reconaisence	0.575	0.56375	0.59875
Worms	0.5525	0.54875	0.745
Backdoor	0.745	0.73625	0.79875
Fuzzers	0.54375	0.5325	0.5775
Shellcode	0.595	0.56125	0.61375
DoS	0.67875	0.66875	0.72
Generic	0.7625	0.755	0.825
Overall avg.	0.637083	0.624028	0.695972

6 Conclusions and Future Research

In this paper, we investigated the performance of unsupervised approaches to identify different types of network attacks. From our experimental analysis, we come to a conclusion that the k-modes clustering algorithm outperforms the k-means and k-medoid algorithms in terms of accurately identifying nine different types of cyber attacks from the state-of-the-art datasets. We have used different combination to avoid any bias in our investigation and it turns out that k-modes algorithm consistently performs better than the rest. In future we are going to address the issue of automatic centroid calculation and mixed type of attributes in the datasets.

References

1. Baaziz, A., Quoniam, L.: How to use Big Data technologies to optimize operations in Upstream Petroleum Industry. Int. J. Innov. **1**(1), 19–29 (2013)
2. Editorial: community cleverness required. Nature **455**(7209), 1 (2008). http://www.nature.com/news/specials/bigdata/index.html
3. Manyika, J., et al.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, New York (2011)
4. De Mauro, A., Greco, M., Grimaldi, M.: What is big data? A consensual definition and a review of key research topics. In: AIP Conference Proceedings, vol. 1644, pp. 97–104. AIP (2015). http://aip.scitation.org/doi/abs/10.1063/1.4907823
5. Akerkar, R.: Big Data Computing, International Standard Book Number 13: 978-1-4665-7838-8
6. Mahmood, T., Afzal, U.: Security analytics: big data analytics for cybersecurity: a review of trends, techniques and tools. In: 2013 2nd National Conference on Information Assurance (NCIA), Rawalpindi, pp. 129–134 (2013)
7. Alguliyev, R., Imamverdiyev, Y.: Big data: big promises for information security. In: Proceedings of the 2014 8th IEEE International Conference on Application of Information and Communication Technology AICT, pp. 1–4, October 2014

8. Edgeworth, F.Y.: On discordant observations. Philosoph. Mag. **23**(5), 364–375 (1887)
9. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection. ACM Comput. Sur. **41**(3), 1–58 (2009). https://doi.org/10.1145/1541880.1541882
10. Dasgupta, D., Andmajumdar, N.: Anomaly detection in multidimensional data using negative selection algorithm. In: Proceedings of the IEEE Conference on Evolutionary Computation, pp. 1039–1044 (2002)
11. Dasgupta, D., Andnino, F.: A comparison of negative and positive selection algorithms in novel pattern detection. Proc. IEEE Int. Conf. Syst. Man Cybernet. **1**, 125–130 (2000)
12. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S.: A geometric framework for un-supervised anomaly detection: detecting intrusions in unlabeled data. In: Barbará, D., Jajodia, S. (eds.) Applications of Data Mining in Computer Security, vol. 6. Springer, Boston (2002). https://doi.org/10.1007/978-1-4615-0953-0_4
13. Oldmeadow, J., Ravinutala, S., Leckie, C.: Adaptive clustering for network intrusion detection. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS, vol. 3056, pp. 255–259. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24775-3_33
14. Zanero, S., Savaresi, S.: Unsupervised learning techniques for an intrusion detection system. In: Proceedings of the ACM Symposium on Applied Computing, SAC 2004. ACM (2004)
15. Mahoney, M.V., Chan, P.K.: PHAD: Packet Header Anomaly Detection for Identifying Hostile Network Traffic Department of Computer Sciences, Florida Institute of Technology, Melbourne, FL, USA, Technical report CS- 2001-4, April 2001
16. Mahoney, M.V., Chan, P.K.: Learning nonstationary models of normal network traffic for detecting novel attacks. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, pp. 376–385 (2002)
17. Mahoney, M.V., Chan, P.K.: Learning Models of Network Traffic for Detecting Novel Attacks Computer Science Department, Florida Institute of Technology CS-2002-8, August 2002
18. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study: final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop (1998)
19. Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P., Tarassenko, L.: A system for the analysis of jet system vibration data. Integr. Comput. Aided Eng. **6**(1), 53–65 (1999)
20. Gaddam, S.R., Phoha, V.V., Balagani, K.S.: K-Means+ID3: a novel method for supervised anomaly detection by cascading K-means clustering and ID3 decision tree learning methods. IEEE Trans. Knowl. Data Eng. **19**(3), 345–354 (2007)
21. Moustafa, N., Slay, J.: UNSW-NB15 DataSet for Network Intrusion Detection Systems, May 2014. http://www.cybersecurity.unsw.adfa.edu.au/ADFA20NB15
22. Moustafa, N., Slay, J.: The evaluation of Network Anomaly Detection Systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. Inf. Secur. J. Glob. Perspect. **25**(1–3), 18–31 (2016)
23. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**, 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010