PHD ASSISTANCE Labs

Building an Effective Machine Learning Algorithms for intrusion Detection systems (IDS)

HOW TO BEST SELECT THE DATASET AND SIZES FOR BUILDING IDS MODELS

PHD ASSISTANCE
YOUR TRUSTED MENTOR SINCE 2001

EXECUTIVE SUMMARY

- The size of the IDS dataset is a major factor in the performance of machine learning (ML) applications.

- The amount of IDS data that is required is completely dependent on the complexity of the selected technique and the requirement.

- A right amount of IDS data size must be identified for obtaining better performance of the classifier and better

INTRODUCTION:

Network attacks are unauthorised activities and transfers in a network also known as intrusions. These intrusions may take place in a small local computer network or a large network with hundreds of servers. These intrusions have different consequences ranging from data theft to damage to the equipment. Intrusion Detection Systems (IDS) are models that identify activities and transfers in a network that shouldn't have taken place[1].

CHARACTERISTICS OF THE DATASET

Each type of attack is different from the other, and hence there is a necessity to understand the characteristics of the different attacks for better identification. Real-time learning for machines is not feasible since the machines will not know the difference between intrusions and normal transfers.

The characteristics of the packets must be accompanied by labels that distinguish between the attacks and normal packets. Hence, these characteristics are compiled into a single document known as a dataset. The size of the dataset is a major factor in the performance of machine learning applications. The amount of data that is required is completely dependent on the complexity of the selected technique and the requirement. This selection is necessary since having very little data might give less accuracy whereas having lots of data may lead to unnecessary time wastage. Therefore, the right amount of data size must be identified for obtaining better performance of the classifier and better

Data size

Some of the techniques with high complexity are neural networks, where a comparatively large dataset is required. It is usually thought that neural networks especially deep neural networks require large amounts of data in the range of Big data. However, a large amount of computational resources are not required for running neural network. A neural network can run with just one GPU. Many GPUs are necessary only for large corporations who work with millions of data and images on a daily basis. On the other hand, a researcher does not need to work with such large data for intrusion detection. Anywhere between 100,000 and 200,000 data size is sufficient for optimum operation. Usually, around 80% of the data is used for training, and 20% is set aside for the actual classification. Other lightweight algorithms which do not use neural network do not require even 100,000 instances. Depending on the machine learning algorithm between 10,000 and50,000 data rows are sufficient.

Data scaling

This is a very important process during the pre-processing stage. Since most of the datasets will have large number of data, unnecessary rows can be removed in order to eliminate overfitting. Since some attacks will have more data when compared to other attacks, the overrepresented attacks must be reduced. This will clear the overfitting problem and also reduce the size of the dataset

IDS DATA TYPE

In an intrusion dataset, the most common features are date & time, source & destination IP, duration, transport protocol, source & destination port, size of the data and TCP flags. This is accompanied by labels which classify it as a normal and intrusion. For single class execution, the type of attack is not required. It will only have two types of labels which are normal and attack. It does not matter what kindof attack is present. One example of such dataset is UNSW NB15[2]. On the other hand, the multi-class dataset has different attacks labelled as such. Their attacks might have different percentages of data where some attacks have higher prevalence than others. Multiclass data are more preferred while conducting research. E.g. UGR16[3] The most common attacks are given in Figure 1.

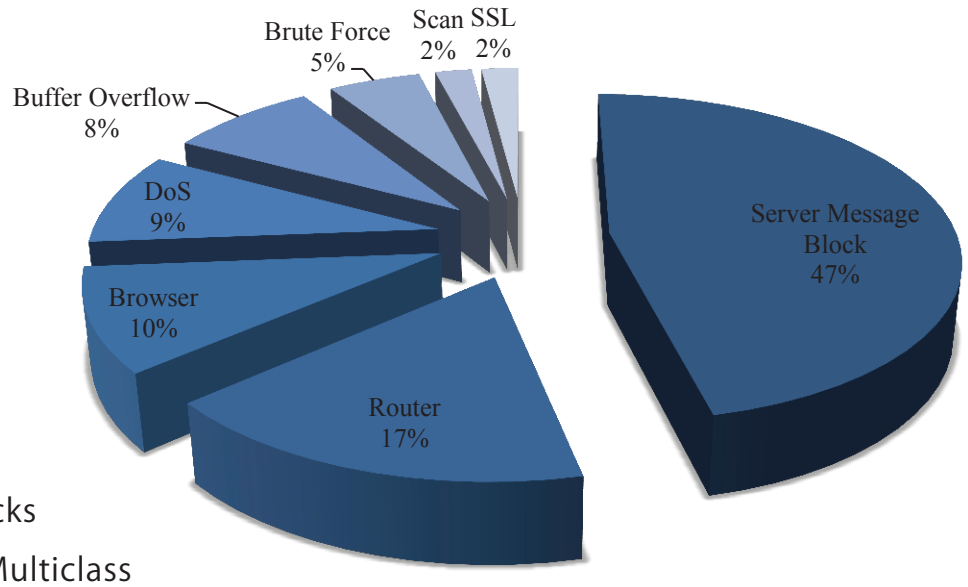

Figure 1[4]

SUMMARY

There are not many datasets which contain intrusion data in the public domain. of research, and many researchers use the same datasets multiple times. KDD CUP dataset[5] is one of the most widely used datasets in the last decade use by lots of researchers[6]. This lack of data is a major challenge for IDS[7]. Most of the public datasets have been researched extensively, and their characteristics are published online. Since the nature of attacks is changing over time, it is necessary to identify more comprehensive datasets. Depending upon the requirements and the machine learning algorithm, an appropriate data size may be selected or scaled down.

References

[1] Subramanian N, Jeyaraj A. Recent security challenges in cloud computing. Comput Electr Eng 2018; 71: 28–42.

[2] Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015, Military Communications and Information Systems Conference (MilCIS). IEEE, pp. 1–6.

[3] Maciá-Fernández G, Camacho J, Magán-Carrión R, et al. UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs. Comput Secur 2018; 73: 411–424.

[4] Raj Samani, Beek C. McAfee Labs Threats Report, https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-dec-2018.pdf (2018).

[5] KDD U. KDD Cup 1999 Data. UCI KDD, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (1999, accessed 29 August 2019).

[6] Özgür A, Erdem H. A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015. PeerJ Prepr. Epub ahead of print 2016. DOI: 10.7287/peerj.preprints.1954v1.

[7] Sharafaldin I, Habibi Lashkari A, Ghorbani A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy. SCITEPRESS - Science and Technology Publications, pp. 108–116.

About PHD ASSISTANCE

PhD Assistance, is world's reputed academic guidance provider for the past 15 years have guided more than 4,500 Ph.D. scholars and 10,500 Masters Students across the globe. We support students, research scholars, entrepreneurs, and professionals from various organizations in providing consistently high-quality writing and data analytical services every time. We value every client and make sure their requirements are identified and understood by our specialized professionals and analysts, enriched in experience to deliver technically sound output within the requested timeframe. Writers at PhD Assistance are best referred as 'Researchers' since every topic they handle unique and challenging. We specialize in handling text and data, i.e., content development and Statistical analysis where the latest statistical applications are exhausted by our expert analysts for determining the outcome of the data analysed. Qualified and experienced researchers including Ph.D. holders, statisticians, and research analysts offer cutting edge research consulting and writing services to meet your business information or academic project requirement. Our expertise has passion towards research and personal assistance as we work closely with you for a very professional and quality output within your stipulated time frame. Our services cover vast areas, and we also support either part or entire research paper/service as per your requirement at competitive prices.

UK: 10 Park Place,
Manchester M4 4EY,
UK: +44-1143520021
Email:info@phdassistance.com,
Web:www.phdassistance.com.