First International Conference on Computer Engineering
International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
Volume 5 Issue 8, November-December-2020

Network Traffic Classification using Artificial Intelligence and Machine Learning

[1]Namita Parati , [2]Dr Salim Y. Amdani
[1,2]Department of CSE, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

ABSTRACT

Traffic Classification bunches comparable or related traffic information, which is one standard procedure of information combination in the field of organization the board and security. With the quick development of organization clients and the rise of new systems administration administrations, network traffic classification has drawn in expanding consideration. Many new traffic characterization methods have been created and broadly applied. Be that as it may, the current writing comes up short on exhaustive overview to sum up, think about and investigate the new advances of organization traffic characterization to convey a comprehensive point of view. This paper cautiously surveys existing organization traffic characterization strategies from a new and far reaching point of view by ordering them into five classifications in light of delegate grouping highlights, i.e., insights based arrangement, connection based grouping, conduct based order, payload-based arrangement, and port-based grouping. A progression of measures are proposed to assess the exhibition of existing traffic characterization techniques.

Keywords : Machine Learning, Clustering, Classification, Network, Analysis.

I. INTRODUCTION

The Internet has been one of the main creations of the 20th century. To adapt to the rising number of web clients, organizations are continually chipping away at upgrading their web speeds. The thriving of the web and its developing rates has permitted more traffic to stream all through the normal processing gadget. Notwithstanding, it additionally opens entryways for expected dangers and noxious assaults. Consequently, scientists have understood the requirement for proposing a few traffic grouping procedures that help oversee and control the progression of organization traffic to lighten the dangers implied with expected dangers. Traffic grouping is the relationship of organization traffic with the application or classification of uses that produced them (for instance, Skype, HTTP, SMTP, video web based, etc). Traffic characterization is significant for a long time [1], to be specific, guaranteeing the Quality of Service (QoS) and Service Level Agreement (SLA) and investigating unusual organization conduct during startling personal times by which network managers might actually utilize it to recognize weak spots inside the organization. Traffic arrangement is additionally utilized for traffic molding and data transfer capacity portion which directs the progression of organization bundles to guarantee consistence with a particular traffic profile. Ultimately, traffic characterization is utilized in network safety since it perceives pernicious classes of traffic that incorporate infections, trojans, spyware, and numerous others. When a particular progression of traffic has been marked as noxious, an Intrusion Detection System (IDS) can then shut out the malevolent classes before they arrive at the client. Traffic order procedures are partitioned into four principle instruments; port-based, Deep Packet Inspection (DPI) based, heuristic-based, and Machine Learning (ML) based strategies [2]. Port-based order procedures are to a great extent dependent on the port quantities of the vehicle layer of the Open System Interconnection (OSI). Be that as it may, as correspondence conventions develop, applications began fluctuating their port numbers powerfully to jumble any method for traffic grouping. DPI centers around intrusively checking the payload of the traffic searching for known marks that connect with explicit applications to order it. All things considered, it is additionally one of the most time and asset consuming procedures since design matching on application marks requires loads of processing power other than the time expected to contrast a mark with an information base of pre-saved marks for order purposes. Likewise, certain individuals are stressed over the protection of their conveyed information since they don't wish to be checked. Thusly, applications began conquering this component by scrambling the payloads of their bundles to safeguard their substance. In like manner, encryption renders DPI totally unfeasible. Besides, heuristic procedures will quite often consume lesser assets, produce the result in a more limited time to the detriment of forfeiting the characterization quality since it brings about exceptionally low correctness.

II. LITERATURE REVIEW

In the field of machine learning, related scientists apply profound learning calculations to the order of organization traffic, for example, applying limited Boltzmann machines to the characterization of DoS traffic [6], utilizing Artificial Neural Network (ANN) to distinguish the malignant traffic [7], the use of profound conviction network in network traffic grouping [8], etc. Since the organization traffic information itself additionally has possible fleeting and spatial highlights, the transient component is reflected in the current and past rush hour gridlock streams, and the spatial element is reflected in the topological relationship between's the traffic streams. Thusly, the spatial and fleeting highlights additionally impact the acknowledgment of typical and unusual traffic. Significant scientists have applied CNN to the spatial component extraction of organization traffic, and have accomplished specific accomplishments [9], [10]. Riyaz and Ganapathy [5] proposed a component determination strategy in view of restrictive arbitrary fields and straight connection coefficients to choose the most contributing elements, and afterward utilized the CNN model for additional element extraction to work on the exhibition of organization traffic acknowledgment. Xu et al. [11] proposed the LSTMs-AE model, which joins LSTM with the AutoEncoder (AE). The model uses LSTM's time series highlight extraction capacity and AE's component portrayal learning capacity to further develop execution. Azizjon et al. [12] utilized the 1D-CNN model for administered learning of organization traffic fleeting elements, and through trials to confirm that its presentation is superior to conventional AI models like irregular backwoods and SVM. Subsequent to preprocessing the first traffic information, Xu [13] utilized picture handling innovation to change over traffic information into grayscale pictures, and afterward utilized CNN to convolve the grayscale pictures of traffic to remove the relationship between's elements. Ling [14] handled the spatial highlights of the information by utilizing various CNNs with various scale convolution portions, and joined with LSTM to separate fleeting elements. Imrana et al. [15] proposed the bidirectional

LSTM (BidLSTM) model for the grouping of unusual traffic, and confirmed its exhibition to be preferable over LSTM and different models. Applying LSTM to the extraction of organization traffic highlights can really extricate the time series highlights between traffic streams. Albeit the utilization of CNN to the extraction of traffic spatial elements likewise has a specific presentation improvement, but CNN is more reasonable for handling Euclidean primary information like pictures. The type of organization traffic information is normally a one-layered structure, and the spatial connection between traffic streams is more like a geography structure. Diagram convolution model [16] has a decent element extraction capacity for topological design and has been generally applied in certain fields. Zhao et al. [17] proposed a mix of diagram convolution organization and Gated Recurrent Unit (GRU) to extricate the fleeting and spatial elements of traffic streets and make more exact expectations of street traffic stream. The outcomes show that its exhibition is superior to customary time series relapse models like ARIMA and SVR. Yao et al. [18] develop a solitary text diagram for the corpus in light of word co-event and archive word relationship, and afterward become familiar with the text chart convolution network for the corpus. Contrasted and different strategies, the presentation of this model is more noticeable. By investigating the application status and impediments of the above works, the diagram convolution model is as yet in the exploratory stage. In the field of organization security, the use of diagram convolution model in network traffic include extraction has significant examination importance.

III. Network Traffic Classification Model

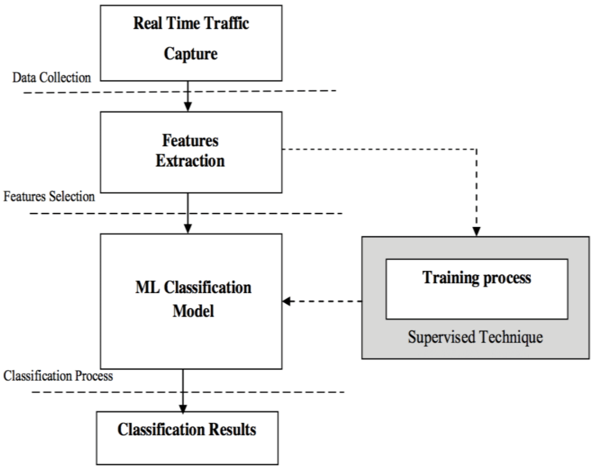


Figure 1. Network Traffic Classification Model

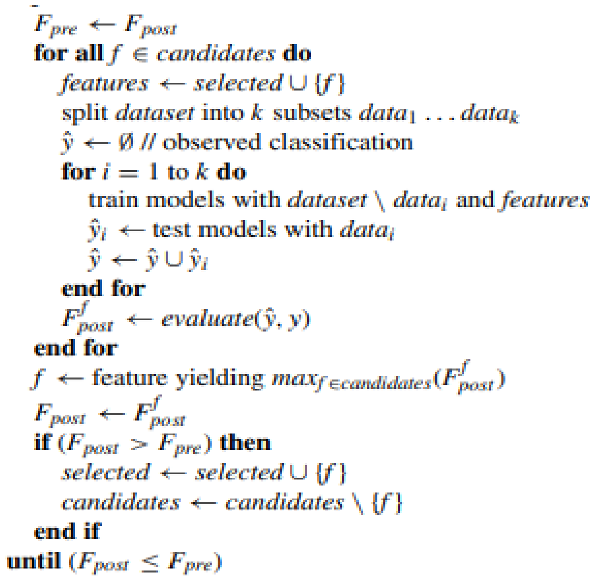
A. Data collection This is the first step, dealing with capturing real-time traffic from networks. Subsequently, the collected dataset is used for feature extraction, to train and test classifiers.

B. Features selection This step deals with extracting features from network flow. Inter-packet arrival times, packet length, and packet duration are examples of features that will be used to train the classifier. This is the core element to build a robust classifier. In this research will be based on a sub-flow instead of the full flow or the first few packets in the flow. This avoids having to wait for a full flow and minimises the required space for a buffer, which consequently increases the memory space. This aspect has not yet been explored by most of the published researches.

C. Training process The training phase involves data sampling and classifier training based on the (pre-labelled) samples. This step uses supervised algorithms to create classification rules and build the model. The information learnt during the training phase is used to classifY new unseen examples in the testing phase (classification process).

D. Classification process This is the final stage, where network traffic analysis accrues, by correlating network traffic patterns with the generating applications. In this stage our research project will examine different classification techniques, supervised, semi-supervised and unsupervised ML algorithms based on statistical analysis of sub-flows. By doing so we are aiming to explore the extent of their robustness and effectiveness in classifying real-world traffic correctly and to investigate their ability to detect new emerging applications, which could be malicious applications.

IV. Classification Algorithm



```
  $F_{pre} \leftarrow F_{post}$
  for all $f \in candidates$ do
    $features \leftarrow selected \cup \{f\}$
    split $dataset$ into $k$ subsets $data_1 \ldots data_k$
    $\hat{y} \leftarrow \emptyset$ // observed classification
    for $i = 1$ to $k$ do
      train models with $dataset \setminus data_i$ and $features$
      $\hat{y}_i \leftarrow$ test models with $data_i$
      $\hat{y} \leftarrow \hat{y} \cup \hat{y}_i$
    end for
    $F^f_{post} \leftarrow evaluate(\hat{y}, y)$
  end for
  $f \leftarrow$ feature yielding $max_{f \in candidates}(F^f_{post})$
  $F_{post} \leftarrow F^f_{post}$
  if ($F_{post} > F_{pre}$) then
    $selected \leftarrow selected \cup \{f\}$
    $candidates \leftarrow candidates \setminus \{f\}$
  end if
until ($F_{post} \leq F_{pre}$)
```

ML literature often utilizes two additional metrics known as Recall and Precision. These metrics are defined as follows:

• **Recall:** Percentage of members of class X correctly classified as belonging to class X.

• **Precision:** Percentage of those instances that truly have class X, among all those classified as class X

V. CONCLUSION

This paper gave a thorough review on the state of art of network traffic classification. We previously presented traffic highlights, summed up research datasets, and determined traffic order granularity. We then inspected include determination and characterization calculations that are generally utilized in rush hour gridlock arrangement. We further proposed a bunch of measures for assessing the exhibition of traffic grouping strategies. By utilizing the proposed standards, we completely studied and looked at the traffic arrangement strategies by grouping them into five classifications, in particular measurements based, relationship based, conduct based, payload-based and port-based. Toward the end, we demonstrate a few open issues and propose future examination bearings for working on the presentation of traffic arrangement. In no time, high proficiency, minimal expense, obscure application distinguishing proof, fine granularity with guaranteed exactness, encoded traffic characterization, grouping with little named information and progressed strength with protection concern are promising future exploration headings in the field of traffic order.

VI. REFERENCES

1) T. Karagiannis, A. Broido, M. Faloutsos, K. Claffy, Transport layer identification of P2P traffic, in: Proceedings of the Fourth ACM SIGCOMM Conference on Internet Measurement, 2004, pp. 121–134.
2) Y. Wang, Y. Xiang, S.Z. Yu, Automatic application signature construction from unknown traffic, in: Proceedings of IEEE International Conference on Advanced Information Networking and Applications, 2010, pp. 1115–1120.
3) T.T.T. Nguyen, G. Armitage, A survey of techniques for internet traffic classification using machine learning, IEEE Commun. Surv. Tutor. 10 (4) (2009) 56–76.
4) A. Callado, C. Kamienski, G. Szabo, B. Gero, J. Kelner, S. Fernandes, D. Sadok, A survey on

internet traffic identification, IEEE Commun. Surv. Tutor. 11 (3) (2009) 52.

5) Z. Cao, G. Xiong, Y. Zhao, Z. Li, L. Guo, A Survey on Encrypted Traffic Classification, Springer Berlin Heidelberg, 2014.

6) S. Seo, S. Park, and J. Kim, “Improvement of network intrusion detection accuracy by using restricted Boltzmann machine,” in Proc. 8th Int. Conf. Comput. Intell. Commun. Netw. (CICN), Tehri, India, Dec. 2016, pp. 413–417.

7) A. Shenfield, D. Day, and A. Ayesh, “Intelligent intrusion detection systems using artificial neural networks,” ICT Exp., vol. 4, no. 2, pp. 95–99, Jun. 2018.

8) I. Sohn, “Deep belief network based intrusion detection techniques: A survey,” Expert Syst. Appl., vol. 167, Apr. 2021, Art. no. 114170.

9) Y. Yang, Research on Convolutional Neural Network Intrusion Detection Model Based on Network Traffic Feature Map. Hangzhou China: Hangzhou Dianzi Univ., 2020.

10) S. Z. Lin, Y. Shi, and Z. Xue, “Character-level intrusion detection based on convolutional neural networks,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Rio de Janeiro, Brazil, Jul. 2018, pp. 1–8.

11) Y. Xu, Y. Tang, and Q. Yang, “Deep learning for IoT intrusion detection based on LSTMs-AE,” in Proc. 2nd Int. Conf. Artif. Intell. Adv. Manuf., Oct. 2020, pp. 64–68.

12) M. Azizjon, A. Jumabek, and W. Kim, “1D CNN based network intrusion detection with normalization on imbalanced data,” in Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC), Feb. 2020, pp. 218–224.

13) Y. Xu, “A research of intrusion detection based on image processing within the framework of deep learning,” M.S. thesis, Univ. Electron. Sci. Technol. China, Chengdu, China, 2020.

14) Y. Ling, Research on Intrusion Detection System Model Based on Deep Neural Network. Hangzhou, China: Hangzhou Dianzi Univ., 2020.

15) Y. Imrana, Y. Xiang, L. Ali, and Z. Abdul-Rauf, “A bidirectional LSTM deep learning approach for intrusion detection,” Expert Syst. Appl., vol. 185, Dec. 2021, Art. no. 115524.

16) T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016, arXiv:1609.02907.

17) L. Zhao, Y. Song, C. Zhang, Y. Liu, and H. Li, “T-GCN: A temporal graph convolutional network for traffic prediction,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 9, pp. 3848–3858, Sep. 2019.

18) L. Yao, C. Mao, and Y. Luo, “Graph convolutional networks for text classification,” in Proc. AAAI Conf. Artif. Intell., 2019, vol. 33, no. 1, pp. 7370–7377.

Cite this Article

Namita Parati, Dr Salim Y. Amdani, "Network Traffic Classification using Artificial Intelligence and Machine Learning", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 5 Issue 8, pp. 360-364, November-December 2020.

Journal URL : https://ijsrst.com/IJSRST205863