International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Analysis of Various Network Traffic Classification Techniques for Cyber Security

[1]Namita Parati , [2]Dr Salim Y. Amdani

[1,2]Department of CSE, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

Article Info
Volume 8, Issue 5
Page Number: 115-120

Publication Issue :
September-October-2021

Article History
Accepted : 02 Oct 2021
Published : 26 Oct 2021

ABSTRACT

The quantity of supposed violations in PC networks had not expanded until a couple of years prior. Constant examination has become fundamental to identify any dubious exercises. Network classification is the initial step of organization traffic examination, and it is the center component of organization interruption recognition frameworks (IDS). Albeit the procedures of arrangement have improved and their precision has been upgraded, the developing pattern of encryption and the demand of use engineers to make better approaches to stay away from applications being separated and recognized are among the reasons that this field stays open for additional examination. This paper examines how specialists apply Machine Learning (ML) calculations in a few arrangement procedures, using the factual properties of the organization traffic stream. It additionally frames the following phase of our exploration, which includes examining different characterization procedures (managed, semi-administered, and unaided) that utilization ML calculations to adapt to true organize traffic.

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

I. INTRODUCTION

Classifying network traffic with a generated application, and is a vital first step for network analysis. Significant data can be assembled from traffic investigation, particularly for security purposes, for example, separating traffic and distinguishing and identifying pernicious movement. By realizing what sort of utilization is streaming over their organizations, network administrators can respond rapidly to potential occurrences in view of their episode reaction plans. A few organization traffic order methods have been created in the course of the most recent twenty years to adapt to the difficulties that classifiers face. Authentic advancements have uncovered huge error and trickiness of the customary procedures (port-based arrangements) [I, 2], which rely upon port numbers to classifications network traffic. This is on the grounds that the quantity of uses that stream over networks utilizing arbitrary or non-standard ports has expanded dramatically. To defeat this issue, payload-based arrangement arose, and examines the headers of the bundles as well as their substance [2, 3, 4, 5, 6]. This grouping is viewed as a solid strategy with exact outcomes, however the viability of Deep Packet Classifying network traffic

joins network traffic with a produced application, and is an indispensable initial step for network investigation. Important data can be assembled from traffic investigation, particularly for security purposes, for example, sifting traffic and recognizing and distinguishing pernicious movement. By realizing what sort of utilization is streaming over their organizations, network administrators can respond rapidly to potential occurrences in light of their episode reaction plans. A few organization traffic arrangement strategies have been created in the course of the most recent twenty years to adapt to the difficulties that classifiers face. Authentic advancements have uncovered huge error and lack of quality of the customary strategies (port-based groupings) [I, 2], which rely upon port numbers to order network traffic. This is on the grounds that the quantity of utilizations that stream over networks utilizing arbitrary or non-standard ports has expanded dramatically. To defeat this issue, payload-based order arose, and reviews the headers of the parcels as well as their substance [2, 3, 4, 5, 6]. This characterization is viewed as a solid strategy with precise outcomes, yet the viability of Deep Packet.

II. RELATED WORK

A few examinations have shown a near examination of traffic characterization and classification in light of AI using different datasets, for example, the spine organization, while others have utilized AI for traffic order or researched QoS support for savvy city applications across various layers, for example, the information interface layer and transport layer. For example, Aureli et al. [7] proposed a powerful arrangement technique called learning-based Differentiated Services to find traffic qualities and progressively allot administration classes to IP bundles. They applied AI techniques (e.g., direct discriminant examination, k-implies bunching) considering parcel qualities like the uneven traffic conveyance between classes. Their proposed strategy changed the order results powerfully. In spite of the fact that our methodology and that of the creators' portion a similar goal, which is to arrange traffic, the creators applied semisupervised procedures to produce an alternate number of subclasses from the Differentiated Services marks. Nonetheless, in our methodology, we apply four directed AI calculations to order network traffic, utilizing 11 classes. Zhongsheng et al. [8] proposed a SVM to order network traffic in grounds spine organizations. They applied the SVM to traffic characterization through information assortment and component age. The SVM accomplished solid and exact outcomes, arriving at 99.31% and 96.12% precision utilizing one-sided and impartial test tests, separately. Be that as it may, they just dissected the SVM, dismissing other AI calculations since calculation exactness isn't the most required objective 100% of the time. As a matter of fact, continuous applications are more touchy to defer than to precision. Thusly, the execution season of various AI calculations must be thought of. Al-Turjman [9] dealt with the remote medium access issue under quick versatility in savvy urban communities. The subsequent structure utilizes LTE (Long Term Evolution), while working on the QoS of versatile applications. Moreover, it limits the postponement and blunder progressively shrewd transportation. The structure incorporates a Markovian interaction into the IEEE 802.16 norm to explore different QoS measures, for example, the normal parcel delay. Also, a plan for portable vehicular cloud is proposed thinking about different circumstances, like traffic and climate. The plan utilizes the cell foundation to transfer information and video however doesn't consider AI strategies that could give better and more-productive choices. Yao et al. [10] proposed a traffic order technique basically planned for shrewd city organizations. Their technique depends on profound learning (DL), utilizing a container network model for productive characterization. The proposed technique plans to eliminate the manual determination of Sensors 2020,

21, 4677 5 of 17 organization traffic highlights. While this strategy utilizes just an improved convolutional brain network model to upgrade the element choice, we depend on four administered AI calculations and analyze their outcomes for traffic order, planning to work on the QoS in shrewd city networks by arranging the organization traffic. Miao et al. [11] looked at six AI calculations for traffic characterization: Naive Bayes, RF, SVM, H2O, KNN, and DT. They involved head part investigation for include extraction and broke down its effect on the grouping results. Trial results showed that RF and KNN were the top performing calculations in general. Without head part investigation, the precision was 92.92% and 84.56% for RF and KNN, individually. Conversely, our traffic grouping calculations accomplished higher exactness, coming to 99.08% and 97.16% for RF and KNN, individually. In spite of the fact that our datasets contain grounds information traffic and their datasets contain ISP information traffic, they are both viewed as spine network traffic types. Accordingly, they share comparable information traffic. Perera et al. [12] analyzed six administered learning calculations for traffic arrangement: Naive Bayes, Bayesian organization, RF, DT, Naive Bayes tree, and multi-facet perceptron. Tests were directed utilizing two element determination strategies and five traffic classes. The outcomes showed that the RF and DT calculations gave the most noteworthy arrangement precision, with 96% and 95% normal exactness, separately. Be that as it may, our traffic characterization calculations accomplished prevalent execution, with 99.08% and 99.18% normal precision for RF and DT, individually. Rahman et al. [13] proposed a cloud advanced mechanics system that is appropriate for brilliant city applications. In the structure, an automated specialist use cloud administrations through task offloading to work on the QoS and framework execution. An improvement issue is figured out for a coordinated non-cyclic diagram, and a hereditary calculation decides the ideal offloading choices and tackles the advancement issue. Not at all like this turn of events, we work on the QoS in savvy city networks by taking on traffic grouping in light of AI. To sum up, AI calculations have been utilized to think about classifier execution thinking about regulated calculations. Also, profound learning strategies have been considered and different techniques have been proposed to further develop QoS in savvy city organizations. Dissimilar to existing investigations, we give an extensive report and assess the exhibition of managed grouping calculations in particular, SVM, RF, KNN, and DT-to work on the QoS in shrewd city organizations and characterize network traffic as indicated by measurable highlights. Besides, we plan and carry out a port-based traffic characterization strategy for correlation with the AI calculations.

III. METHODOLOGY

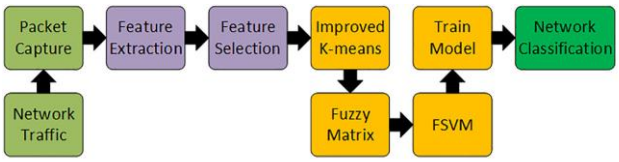

Figure 1. Network Traffic Classification

To address the challenges of obtaining high quality ground-truth data incorporating flow class segregation and identification in each of the examined applications, our proposed classification technique utilizes unsupervised cluster analysis and supervised classifier training in tandem. A high level overview of the traffic classification scheme is shown in Figure 1 with a description of principal steps as follows.

(i) ***Preprocessing.*** Internet traffic is collected from end-user machines and marked with application labels accordingly (e.g., Skype and YouTube) using a localized operational packet-level classifier. Application labeled traffic is afterwards exported as flows using a flow exporting utility for unsupervised cluster analysis.

(ii) ***Cluster Analysis.*** Using unsupervised -means, flows belonging to individual applications are separately cluster analyzed to extract unique subclasses per application, offering a finer granularity of the classification (e.g., YouTube and Netflix flows would be classed as streaming and browsing).

(iii) ***Classifier Training.*** Flows marked with their - means clusters, indicating the subclass they belong to, are afterwards fed to a C5.0 classifier for supervised training, leading to a decision tree.

(iv) ***Evaluation.*** A separate data set is used for testing the accuracy of the algorithm. For each NetFlow record the trained C5.0 classifies the application and the subclass of the flow based on their respective attributes, ingrained during decision tree creation.

IV. Classification Approaches

Next, labeled data was used to train classification models. There are multiple classification models available and each and every model classify data with different mathematical models. Therefore, results of each model could be different from each other. Some models could perform better and some models perform poorly. In other word, it is better to train and test multiple classification models to find out which model fit better for the project. The tested models are briefly described below.

a. **Support Vector Machine (SVM)** algorithm is a supervised learning algorithm that uses labeled data to train the model. SVM model will calculate decision boundaries between labeled data also known as hyper planes. And points near these hyper-planes are called extreme points. The algorithm will optimize these decision boundaries by setting up margins that separate hyper-planes. Several kernels that uses to optimize these decision boundaries. Linear, RBF, Polynomial and Sigmoid are the most commonly used kernels. Real-world data can be one dimensional or multidimensional. And these data sets are not always linear separable. The linear kernel can handle datasets that can linear separable and for nonlinear datasets, can use other kernels that can transform nonlinear datasets into linear datasets and classify. SVM is effective in multi-dimensional datasets and it is a memory-efficient model.

b. **Decision Tree** is another supervised learning model that classifies data based on information gains by calculating the entropy of the dataset. It is a graphical representation of all the conditions and decisions of the dataset. The root node will be calculated using entropy with the highest information gain among the dataset. This process will continue to split branches and complete the tree. Each internal node is a test on attribute and branches represent the outcome. Leaf represents a class label. The decision tree can use numeric and categorical data for the classification problems. It also supports nonlinear relationships between features.

c. **Random Forest** is one of the powerful supervised learning algorithm, which can perform both regression and classification problems. This is a combination of multiple decision tree algorithms and higher the number of trees, higher the accuracy. It works as same as the decision tree, which based on information gain. In classification, each decision tree will classify the same problem and the overall decision will be calculated by considering the majority vote of the results. The most important advantage of this model is that it can handle missing values and able to handle large datasets.

d. **KNN** is an instance based supervised learning algorithm. In the KNN model, the value k represents the number of neighbors needs to consider for the classification. The model will check the labels of those neighbors and select the label of the majority. The value k should be an odd number to avoid drawing the decision. It is a robust model that can work with noisy data and perform better if the training data set is large.

However, it is not performing well in multidimensional datasets and could reduce efficiency, accuracy, etc.

V. Traffic classification metrics

A key criterion on which to differentiate between classification techniques is predictive accuracy (i.e., how accurately the technique or model makes decisions when presented with previously unseen data). A number of metrics exist with which to express predictive accuracy.

1) Accuracy, precision and recall: Assume there is a traffic class X in which we are interested, mixed in with a broader set of IP traffic. A traffic classifier is being used to identify (classify) packets (or flows of packets) belonging to class X when presented with a mixture of previously unseen traffic. The classifier is presumed to give one of two outputs - a flow (or packet) is believed to be a member of class X, or it is not. A common way to characterize a classifier's accuracy is through metrics known as False Positives, False Negatives, True Positives and True Negatives. These metrics are defined as follows:

• **False Negatives (FN):** Percentage of members of class X incorrectly classified as not belonging to class X.

• **False Positives (FP):** Percentage of members of other classes incorrectly classified as belonging to class X.

• **True Positives (TP):** Percentage of members of class X correctly classified as belonging to class X (equivalent to 100% - FN).

• **True Negatives (TN):** Percentage of members of other classes correctly classified as not belonging to class X (equivalent to 100% - FP).

VI. CONCLUSION

This paper studies the classification of network traffic, proposes the establishment rules of network traffic topology graph structure, and proposes a network traffic classification method based on graph convolution and LSTM. This method first processes the data with the graph convolution layer, extracts its spatial features, and then combines the LSTM model to extract its potential temporal features. On the sampled UNSW-NB15 data set, it is compared with feature selection and other commonly used deep learning methods (such as CNN, BiDLSTM and CNN-LSTM) to verify the performance and effectiveness of the proposed method. There are also some shortcomings and areas to be optimized in the experiment. When building a topological graph for network traffic data, the more the number of nodes, the more undirected edges are established, and the greater the amount of matrix operations involved, which is a big challenge to the memory size and computing power of the machine. This article provides an idea for using graph convolution model in network traffic environment, exploring the relationship between normal and abnormal traffic flows, and the correlations between traffic flows can be further explored in the future.

VII. REFERENCES

[1] Y. Dhote, S. Agrawal, A.J. Deen, A survey on feature selection techniques for internet traffic classification, in: Proceedings of International Conference on Computational Intelligence and Communication Networks, 2016, pp. 1375–1380.

[2] I. Inza, P. Larra, Aga, R. Etxeberria, B. Sierra, Feature subset selection by Bayesian network-based optimization, Artificial Intelligence 123 (1–2) (2000) 157–184.

[3] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[4] J. Yan, A survey of traffic classification validation and ground truth collection, in: Proceedings of the 8th International Conference on Electronics Information and

Emergency Communication, ICEIEC, 2018, pp. 255–259.

[5] I.T. Jolliffe, Principal component analysis, J. Mark. Res. 87 (4) (2002) 513.

[6] Tongaonkar, A.; Keralapura, R.; Nucci, A. Challenges in Network Application Identification. In Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats, San Jose, CA, USA, 25–27 April 2012; p. 1.

[7] Salman, O.; Elhajj, I.; Kayssi, A.; Chehab, A. A Review on Machine Learning Based Approaches for Internet Traffic Classification. Ann. Telecommun. 2020, 673–710. [CrossRef]

[8] Alqudah, N.; Yaseen, Q. Machine Learning for Traffic Analysis: A Review. Procedia Comput. Sci. 2020, 170, 911–916. [CrossRef]

[9] Xie, J.; Yu, F.R.; Huang, T.; Xie, R.; Liu, J.; Wang, C.; Liu, Y. A Survey of Machine Learning Techniques Applied to Software Defined Networking (SDN): Research Issues and Challenges. IEEE Commun. Surv. Tutor. 2019, 21, 393–430. [CrossRef]

[10] Aureli, D.; Cianfrani, A.; Diamanti, A.; Sanchez Vilchez, J.M.; Secci, S. Going Beyond DiffServ in IP Traffic Classification. In Proceedings of the NOMS 2020—2020 IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary, 20–24 April 2020; pp. 1–6.

[11] Zhongsheng, W.; Jianguo, W.; Sen, Y.; Jiaqiong, G. Traffic identification and traffic analysis based on support vector machine. Concurr. Comput. Pract. Exp. 2020, 32, e5292. [CrossRef]

[12] Al-Turjman, F. Smart-city medium access for smart mobility applications in Internet of Things. Trans. Emerg. Telecommun. Technol. 2020, e3723. [CrossRef]

[13] Yao, H.; Gao, P.; Wang, J.; Zhang, P.; Jiang, C.; Han, Z. Capsule Network Assisted IoT Traffic Classification Mechanism for Smart Cities. IEEE Internet Things J. 2019, 6, 7515–7525. [CrossRef]

[14] Miao, Y.; Ruan, Z.; Pan, L.; Zhang, J.; Xiang, Y. Comprehensive analysis of network traffic data. Concurr. Comput. Pract. Exp. 2018,

[15] Perera, P.; Tian, Y.C.; Fidge, C.; Kelly, W. A Comparison of Supervised Machine Learning Algorithms for Classification of Communications Network Traffic. In Neural Information Processing; Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S.M., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 445–454.

[16] Rahman, A.; Jin, J.; Cricenti, A.; Rahman, A.; Yuan, D. A Cloud Robotics Framework of Optimal Task Offloading for Smart City Applications. In Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 4–8 December 2016; pp. 1–7.

Cite this article as :

Namita Parati, Dr Salim Y. Amdani, "Analysis of Various Network Traffic Classification Techniques for Cyber Security", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 5, pp. 115-120, September-October 2021.

Journal URL : https://ijsrcseit.com/CSEIT217683