IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

A Survey on: Finding Network Traffic Classification Methods based on C5.0 Machine Learning Algorithm

Amit Kumar[1*], Daya Shankar Pandey[2], Varsha Namdeo[3]

[1,2,3]RKDF IST College, SRK University, India

Corresponding Author: amitkumar1989.niec@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.788791 | Available online at: www.ijcseonline.org

Accepted: 16/Apr/2019, Published: 30/Apr/2019

Abstract**—** Classifying traffic in a residential area is always a challenging task in the high-speed network. The analysis and quality of service require more specific network control, which generates network traffic. The existing network has many disadvantages because of that the network was unable to detect the traffic in a network. This survey is based on the machine learning algorithm which will work accordingly to the generated traffic information that will be get from the client for that the boosted classifier contain high accuracy has been generated. So, this network will be used for the classification of the applications like- FTP, Skype, TCP etc. , This type of paper demonstrates that the Machine Learning Algorithm and the use of this algorithm are used to classify network traffic.

Keywords**—** Traffic Classification, Computer Networks, C5.0, Machine Learning Algorithms (MLAs), Performance Monitoring.

I. Introduction

Each network conveys information for various applications, which have diverse utilization. Network traffic grouping make challenge for entire system, arrange traffic increment the heap of system, in light of that one can't do work in appropriate or in quick way. Accordingly giving data about the quality dimension requires learning of what sort of information is flowing in the network right now. By and large, Traffic circle strategies use a stream idea to characterize them as a collection of bundles with a similar end of IP addresses, using a similar transport agreement and port number [2]. Posts are considered bidirectional packets from the TV channel to the current server, and the remote server to the station computer are parts of a similar stream. Using application pairs for wrong grouping is a highly targeted idea that network administrators often use to unwantedly control group-generated traffic. This strategy is fast and can be combined with virtually all available modifications. This strategy is a great idea to work on a number of fixed port number conventions. However, this method is not suitable for dynamic ports such as Skype, games, etc. [4]. Since working on a unique port, this application cannot now identify. Compulsory Package Inspection Programs (DPI) are highly measurable. In addition, they were treated when reviewing customer data, and because of these rules, security and confidentiality issues can indicate that customer data is treated privately. Existing techniques such as C4.5, J48, Random wales are very common. They can be used on any type of network, namely, exceptionally fast measurement identification of the application where the traffic is located. Feasible location rate rightness is more than 95. The objective of AI is to structure and create algorithms that enable frameworks to utilize experimental information, experience, and preparing to develop and adjust to changes that happen in their condition.

A noteworthy focal point of AI look into is to consequently initiate models, for example, guidelines and examples, from the preparation information it investigations. Remembering that objective, this paper depicts past related work and after that centers around usage and assessment of new AI algorithmi.e.C5.0 algorithm which has more precision than past [5-9].

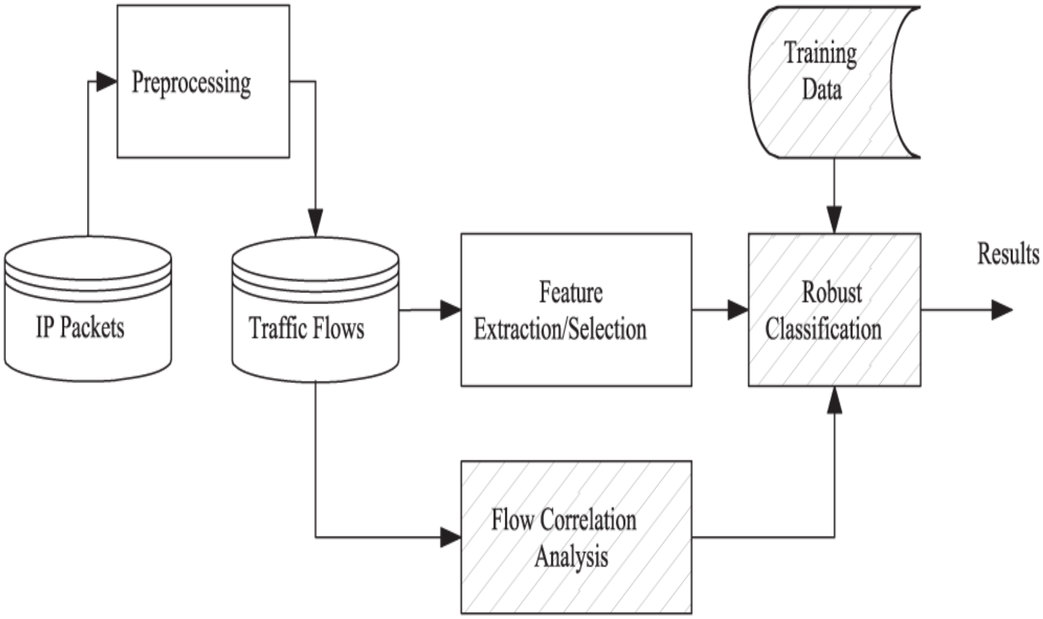

Figure 1: Traffic Classification.

The above figure 1 states the process involves in the classificato of the traffic.

II. RELATED WORK

In [1], the authors rely on some basic basic algorithms that depend on Restreet and C4.5. You can group P2P traffic with the first 5 bundles of the stream. Their strategy, which depended on C4.5, was done just enough (97% of P2P traffic was well-ordered), but was not visited when new bundles of the stream were lost. Therefore, the incremental overhead for grouping source and destination port numbers was set that would determine the classifier with the current task of port numbers for particular applications in the preparation information.

Another way to handle P2P applications for jobs was provided in [3] by a Java implementation of C4.5 called J48 to detect 5 different applications. The authors try to distinguish parts that are between 10 and 1000 after starting the stream, and they were only slightly unstable in the implementation, with a plan accuracy of over 96%. In [10], it was determined that the first C4.5 and J48 are unique only in secure small or stretched transform files (the precision of J48 and C5.0 was virtually identical and harder than C4.5 in the proven cases). J48 uses estimates in [11] for the disclosure of BitTorrent and FTP traffic made for large subjects.

This means that this is exactly 98%. This production demonstrated that conduct of information parameters contained in encoded and decoded traffic created by a similar application appears to be identical. Besides it was appeared zero-payload bundles (ACK) can contort measurements dependent on sizes. In [12] diverse systems of grouping of the system traffic were assessed, including C5.0. The overall accuracy was 88 - 97% in traffic with a location of 14 different application classes.

These extremely complete arrays were most likely incomplete, as both experimentation and experimentation were planned, thereby taking over the choice (name of application) through DPIs (PACE, OpenDPI and L7 channel). These DPI arrangements use many calculations to obtain the names of the application, including fact finding. Along these lines, both preparing and test information were in certain degrees off base, what caused likewise more blunders from the side of C5.0.

III. MODULES DESCRIPTION

Supervised Methods

Supervised techniques, otherwise called arrangement or classification strategies, separate information structures to arrange new cases in pre-characterized classes. It is essential to note that is called managed in light of the fact that the yield classes are pre-characterized. The procedure of a regulated ML techniques begin with a preparation dataset TS characterized as,

T S = < x1, y1 >, < x2, y1 >, ..., < xN , yM >, where xi is the vector of estimations of the highlights relating to the I th example, and yi is its yield class esteem. It finds the distinctive relations between the occurrences and yields a structure, normally a choice tree or order governs, that will characterize the cases in a discrete set y1, y2, ..., yM. There is a great deal of related work that utilization administered procedures [12– 19] with a promising outcomes. The accompanying traffic group strategies enable the managed preparation information and create a partial capacity that the reindeer class can propose for each test track. In the normal rush hour characteristic, sufficient information for creating is a general representation. For example, to solve the problems that arise from a useful traffic career, compulsory applications and customer information are required.

Unsupervised Methods

The non-existent methods (or clustering) attempt to find the cluster structure in different traffic data and pass each test stream to the application-based class of the next cluster. The intent is to increase group traffic in a small number of expectancy-maximizing (EM) clusters and to manipulate each cluster for an application.

A Traffic Classification Approach with Flow Correlation

It represents a new framework that we call incorrect classification with correlation information, or TCC for short. A romantic, non-parametric approach has also been proposed to effectively influence correlation correlation information in the classification process.

Correlation Analysis

The connected floods, which have a similar three-tuple, are generated by a similar application. For example, some bottles started by multiple hosts are usually connected to a similar host on the TCP port 80 of each short period of time. These streams are likely to be created in all respects by a similar application, such as an Internet browser. The three-tuple heuristic over current limit was seen as a few inclusive rush-hour shutdown plans that proposed a cached bundle strategy for conventional belief in which they fought in equal groups with heuristics. Tried the rightness of the three-tuple heuristic with certifiable follows [13].

Network Flexibility

The proposed system show is available to include extraction and relationship investigation. Initially, any sorts of flow measurable highlights can be connected in our system display. In this work, we extricate unidirectional factual highlights from full flows. The measurable highlights separated from parts of flows can likewise be utilized to speak to traffic flows in our system display. Second, any new relationship examination technique can be implanted into our system demonstrate. We acquaint flow relationship investigation with find connection data in rush hour gridlock flows to improve the strength of grouping [14] [15].

IV. RESULTS AND DISCUSSION

The C5.0 algorithm is another era of Machine Learning Algorithm (MLA) in the face of decision trees [16]. This means that thoughtful features and case histories were taken into account in the decision making process. Subsequently, the trees can be used to group experiments. C5.0 has emerged as an improved form of the safely referenced and widely used C4.5 classifier and has several important advantages over its ancestor [17]. The generated rules are constantly increasing and the time taken to create them is lower (some 360 times even for some data). In C5.0 several new techniques have been introduced:

- Boosting: Multiple decision sources are generated and combined to improve predictions.
- Increased expense allowance: This helps to avoid mistakes that can cause harm.
- New attributes: date, time, timestamp, sorted discrete attributes.
- Works may be marked as degraded or unusable in certain cases.
- Plot sampling and cross validation. The C5.0 classification includes a simple command line interface that controls the trees and rules and tests the classification.

V. CONCLUSION AND FUTURE SCOPE

The work supports a novel technique that relies on C5.0 MLA to detect different types of traffic in the PC system. It has been shown that our strategy with precise informal collections is possible both for the preparation and testing of the supported digits. Our results have shown that the classifier can detect traffic that gives the impression of a similar impression, similar to Internet browser traffic and radio traffic through a website. The numbers have no problems with the intuitive traffic: Skype, Game and SSH. However, we see that ftp and torrent document boxes are exactly the opposite of flow attributes, and then a critical number of packages between these two lessons was explained. Our strategy is a field for more research and further improvements. In this study, however, both the preparation and testing of information collections were collectively collected by the same customers. Next, we look at different customers for research

ACKNOWLEDGMENT

I would like to thank my guide (Daya Shankar Pandey) & Head of Department (Dr. Varsha Namdeo) from RKDF IST College, SRK University, India for supporting this work.

REFERENCES

[1]. A. Vlăduțu, D. Comăneci, and C. Dobre, ‘‘Internet traffic classification based on flows’ statistical properties with machine learning,’’ Int. J. Netw. Manage., **vol. 27, no. 3, p. e1929**, May 2017.

[2]. Y. Yu, J. Long, and Z. Cai, ‘‘Session-based network intrusion detection using a deep learning architecture,’’ in Modeling Decisions for Artificial Intelligence. V. Torra, Y. Narukawa, A. Honda, and S. Inoue, Eds. Cham, Switzerland: Springer, 2017, **pp. 144–155.**

[3]. D. J. Weller-Fahy, B. J. Borghetti, and A. A. Sodemann, ‘‘A survey of distance and similarity measures used within network intrusion anomaly detection,’’ IEEE Commun. Surveys Tuts., **vol. 17, no. 1, pp. 70–91**, Jan. 2015.

[4]. M. Ahmed, A. N. Mahmood, and J. Hu, ‘‘A survey of network anomaly detection techniques,’’ J. Netw. Comput. Appl., **vol. 60, pp. 19–31**, Jan. 2016.

[5]. Y. Li, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” 2018.

[6]. R. Alvizu, S. Troia, G. Maier, and A. Pattavina, “Matheuristic with machine-learning-based prediction for software-defined mobile metrocore networks,” Journal of Optical Communications and Networking, vol. 9, no. 9, pp. D19–D30, 2017.

[7]. A. Azzouni and et al, “Neutm: A neural network-based framework for traffic matrix prediction in sdn,” CoRR, vol. abs/1710.06799, 2017.

[8]. Y. Liu and et al, “Short-term traffic flow prediction with conv-lstm,” in Wireless Communications and Signal Processing (WCSP), 2017. IEEE, 2017, pp. 1–6.

[9]. M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang, “Machine Learning for Networking: Workflow, Advances and Opportunities,” IEEE Network, vol. 32, no. 2, pp. 92–99, Mar. 2018.

[10]. “Fault Tolerance in TCAM-limited Software Defined Networks,” Computer Networks, vol. 116, no. C, pp. 47–62, Apr. 2017.

[11]. D. Kreutz, F. M. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, “Software-defined Networking: A Comprehensive Survey,” Proceedings of the IEEE, vol. 103, no. 1, pp. 14–76, 2015

[12]. Y.-D. Lin, H.-Y. Teng, C.-R. Hsu, C.-C. Liao, and Y.-C. Lai, “Fast Failover and Switchover for Link Failures and Congestion in Software Defined Networks,” in IEEE ICC 2016, KL, Malaysia, May 2016.

[13]. I. Butun and S. Morgera, “A Survey of Intrusion Detection Systems in Wireless Sensor Networks,” IEEE, 2014.

[14]. Janice Ca˜nedo, Anthony Skjellum, “Using machine learning to secure IoT systems,” IEEE International Conference on Privacy, Security and Trust (PST), 2016.

[15]. Information on See5/C5.0 - RuleQuest Research Data Mining Tools, 2011. [Online]. Accessible: http://www.rulequest.com/see5-info.html

[16]. Anshul Vishwakarma1* , Amit Khare2, “Vehicle Detection and Tracking for Traffic Surveillance Applications: A Review Paper”, **Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693.**

[17]. K. Thyagarajan 1* , N. Vaishnavi 2, “Performance Study on Malicious Program Prediction Using Classification Techniques”, **Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693.**

Authors Profile

Mr. Amit kumar had completed B.Tech in CSE Branch from GGSIPU New Delhi in 2011, and currently pursuing M.Tech from RKDF IST College, SRK University, India.

Dayashankar pandey is Assistant professor in the Department of Computer Science & Applications in Sarvepalli Radhakrishnan University, Bhopal, India. He is a teacher in the field of computer science and information technology.

Dr. Varsha Namdeo is Professor in the Department of Computer Science & Applications in Sarvepalli Radhakrishnan University, Bhopal, India. She is a teacher and researcher in the field of computer science and information technology. She earned her Master degree in Computer Application from Barkatullah University Bhopal (M.P.) in 2000 and in Computer Science and Engineering from Barkatullah University Bhopal (M.P.) in 2009, and PhD degree from Maulana Azad National Institute of Technology, Bhopal (M.P.). Currently, she is guiding several PhD research scholars.