International Journal of Pure and Applied Mathematics
Volume 117 No. 22 2017, 107-111
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

A SURVEY ON NETWORK TRAFFIC CLASSIFICATION TECHNIQUES

Dr R Suguna[1], Suriya Prakash J[2]
[1]Professor,Department of Computer Science and Engineering
Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology
Avadi, Chennai, TamilNadu, INDIA 600062
drsuguna@veltechuniv.edu.in
[2]Research Scholar, Department of Computer Science and Engineering,
SKR Engineering College, Chennai, Tamil Nadu, INDIA.
suriya.engineer@gmail.com

Abstract*— Network traffic classification process deals with various parameters such as port and protocol based that are used to automatically identify the traffic classes. This type of classification is used to provide security in network level as well as system level in a complex classification environment where data is in encrypted form. It also addresses the issues related to Encryption of data, Security in Modern Network Architecture and its Management, Controlling QoS for products, Identifying Intruders in the Network and Privacy protection of users among the network. This paper lists the problems faced by Traffic Classification while handling the network traffic. Most of the traffic classification methods are not able to satisfy the special requirements of individual datasets. There are massive amount of network traffic datasets and limited numbers of resources are available to produce classification analysis. The survey reveals that traffic classification need to be updated regularly to maintain the accuracy and should be able to adapt the dynamic behaviour of network flow.

Keywords*— Network Traffic, Network Traffic Class, Network Features, Statistical features, Classification

1. Introduction

Internet has become an unavoidable information hub in our daily life and in the work place. Internet today has created a great evolution in network technology and interconnection of networks. Newly emerging network architectures, network protocols and the applications are becoming complex to understand and research community spurred a lot to invent a noble research work in complex networks. Network traffic classification can be adopted in the important applications such as network operators, Twitter, Facebook, Bit Torrent, WhatsApp, Skype, Youtube (i.e., live video streaming) or Uploading and Downloading Videos. Network traffic classification help for planning and designing new infrastructures. Through this accurate classification, Internet Service Providers (ISPs) can be able to provide appropriate billing based on user's actual usage and recommend for Quality of Service (QoS) based on the application needs. Research Community has invented various methodologies of traffic classification from real time network traffic. Internet Assigned Number Authority (IANA) assigns port numbers for TCP or UDP in Port based attributes and also assigns source port and the destination port for every packet in the IP traffic. All the applications in the network traffic do not have registered port numbers, hence it's very difficult to identify the unknown application using port based methods. Some applications dealing with online games and peer to peer networks are using dynamic port numbers so that it's difficult to classify such applications using port based techniques. In Payload based approach, attributes are based on application layer level traffic signatures. Statistical based attributes related to traffic such as duration between the flow, packet ideal time, Length of the packets and it's inter arrival time also play an important role in traffic classification. Payload based uses the technique named deep packet inspection that match both the payload of the packet and known traffic signature but this method will not produce good classification accuracy in encrypted packets.

Assigned Port	Application
20	FTP Data
21	FTP Control
22	SSH
23	Telnet
25	SMTP
53	DNS
80	HTTP
110	POP3
123	NTP
161	SNMP
3724	WoW

Basically there are two types of flows in the network traffic: unidirectional and bidirectional. The unidirectional flow shares information such as source and destination ports, IP and Transport Protocol. In bidirectional the analysis of flow between source and destination starts from the establishment of connection to end of the network connection. Flow Directional Neutrality is calculated from the forward and backward direction of individual statistical features. IP traffic indicated by the Traffic classes can be caused by a single application or multiple applications. Features are in the form of numerical attributes and usually more number of packets belong to same flow

P2P Protocol	String	Trans. Prot.
eDonkey 2000	0xe319010000	TCP/UDP
	0xe53f010000	
Fasttrack	"Get /.hash"	TCP
	0x2700000002980	UDP
BitTorrent	"0x13Bit"	TCP
Gnutella	"GNUT" "GIV"	TCP
	"GND"	UDP
Ares	"GET hash:"	TCP
	"Get sha1:"	

Network Feature selection play a major role in providing accurate results. It is necessary to identify unique attributes in the network traffic flow and in the flow observation taken

between source and destination of the network in a particular time period. In Filter and Wrapper Methods are used to do classification work in Feature selection algorithm. Filter method finds the independent features based on unique and general characteristics. Different subset of machine learning algorithms are examined in Wrapper method and the results obtained can be used for further learning. The features used in packet level categories are length of the packet, Mean and Variance of the packet length, Square of root Mean. The features at flow level deals with duration of the flow, volume of data and number of packets per flow. Fig. 3 lists the features used in different levels.

Level	Protocol	Feature
Transaction	HTTP	Hostname Referrer Cookie User Agent Content type
	SSL	Server name SSL version Certificate date expired
	DNS	Query name Alexa 1M rank Number of canonical names Response flags Time-to-live
Session	TCP	Destination port Packet size Number of packets with the PUSH bit set Number of out-of-order packets
Flow	TCP	Quantity of keep-alive packets in flow Packet inter-arrival time Number of port reusing packets
	IP	Destination IP IP Geo-location IP Autonomous System number
Conv. Win.	UDP	Ratio between sent and received packets
	DNS	Number of non-existent domain responses
		Number of sessions in flow Total amount of data transmitted

2. **Network Traffic Classification Techniques and its Applications**

Kwitt et al [1] and Shen et al. [2] have used Principal Component Analysis (PCA) to identify anomaly detection and analyzing behavioral metrics of network. Karasa et al. [3] have identified botnets through network classification. Jin et al works identified network intrusion detection using pattern recognition method [4]. Sang et al [5] have used (ARMA) auto regressive moving average to identify traffic in the network. Cho et al. [6] used policy of LRU and Patricia tree to measure and filter the traffic on the network in real time. Pang et al. [7], have contributed on reducing the system load by examine the known anomalies and filtering data. Feldman et al. [8] concentrated on identifying multi behavioral traffic on network using cascade-based approach and also their system is able to identify several problems in the network traffic. Li et al [9] and Plonka et al [10] have worked on efficient monitoring the network flow using FlowScans and NetFlows. The works of Gong et al [11], [12] focus on identifying intrusion detection and also finding worms using NetFlow. Alarcon et al have proposed Wavelet approach that detects unknown flow of traffic in the Ethernet [13]. In the works of Gu and Lakina et al preventing unknown flows in the network traffic using Entropy measurement are addressed [14], [15]. Visual dependencies are conjunct with flows and structure of traffic to produce network alerts [16]. Entropy in network traffic and its statistics are used in many approaches [17][20]. Eimann, et al.[17] have identified events in the network using entropy based approach. Harringtons et al [19] identified changes in the network using second order distribution and cross entropy. Lall, et al. [18] monitored the network using entropy on distribution of network traffic. Gu, et al.[14] identified the unknown application by measuring the entropy and its utilization. Gianve et al and Wang et al [20] worked on identifying the dynamic changing channels using entropy based approach. Kim, et al. [21], have worked on deep analysis of packet header using discrete wavelet transform and correlation analysis. For the same analysis Celenk et al have applied recognition of patterns theory and statistical data analysis to improve the accuracy in results [22]. Fu et al [23] proposed a technique considering overall statistics of load on the network and identifying the attack using supervised statistical pattern recognition. Wagner and Plattner et al [24] have identified the change of IP addresses and Port numbers in the network traffic but not discussed on suspicious flow in traffic. Thott and Ji et al [25] works contribute on identifying the change of signals from multiple functionalities with different properties. Statistical data analysis helped to identify intrusion detection through signal processing technique. In Hajjis et al [26], have detected unknown application in local area network by observing the change in the port and IP of the network traffic. In [27], K-Means algorithm was used to produce unique clusters with training dataset with their similar instances. ID3 decision tree has been constructed with k-means cluster and unknown flow detected using score matrix. Ziviani et al [28] have identified unknown traffic using entropy that reduced false negatives. Kim and Reddy [29] analyzed different packet header in real time using time series analysis and their signal series could identify the attacks efficiently than using traffic volume. Androulidakis, et al. [30] work on port entropy and proposed an approach for anomaly detection. A two-phased machine learning classification mechanism with NetFlow as input has been proposed by Taimur Bakhshi [33]. The individual flow classes are derived per application through k-means and are further used to train a C5.0 decision tree classifier. A recurrent neural network (RNN) combined with a convolutional neural network (CNN) for network traffic classification has been experimented by Manuel Lopez-Martin et al [34].

3. **MACHINE LEARNING**

A. Types of Machine Learning Techniques

Machine Learning techniques are broadly categorized as supervised and unsupervised learning. Supervised Learning works on the dataset with known examples and a model is built with the training dataset to classify the unknown sample. Unsupervised classification builds a model by analysing the similarity between the data.

B. Supervised (classification) Methods

Bayes Net is a probabilistic graphical model that represents a set of random variables and their conditional dependencies. It produces better accuracy than other classification methods such as RBF and C4.5. Naïve Bayes classifier has demonstrated high classification speed and good performance using the discretized statistical features in traffic classification. It is easy for naive Bayes classifier to produce the posterior probability that a testing flow belongs to a traffic class. The key point is to estimate the posterior probability that a testing flow belongs to a traffic class. The Decision tree algorithm in the classifier C4.5 and enhanced ID3 algorithm are also used in some network classification approaches. This is a statistical classifier and it produces accurate classification performance but requires huge number of training samples and deals with various types of attributes. Radial Basis Function Neural Network are also suggested for network classification which is a feed forward network and the output depends on weighted linear basis function.

C. *Unsupervised Methods*

Clustering can combine similar flow attributes using unsupervised learning. DBSCAN algorithm identifies the corresponding nodes density distribution and its cluster count based on two parameters density reachability and density-connectivity. In Expectation Maximization the clusters are produced through maximum likelihood based on iterative method. In Expectation step the parameters are identified using random numbers. In second step mean and variance are used to re-iterate and this process continues until local maximum is reached. K-Means will partition the objects with K disjoined subsets and it's a partition based clustering algorithm. It calculates the distance between each object center and mean called as square error.

The combination of unsupervised and supervised is said to be Hybrid or semi supervised approach. If the training dataset is minimal then supervised learning methods will not produce good classification results. If the new application sends the data traffic that not known to the classifier then prediction is not possible. Supervised techniques can identify only the known flows.

The challenge is to build a classifier that take the decision in a fraction of seconds by analyzing few packets from each flow. Minimal feature analysis will reduce memory for buffering the packets in network traffic classification.

A brief survey on various supervised and unsupervised machine learning techniques to solve internet traffic classification has been discussed in [35]. A comparative analysis of machine learning algorithms for network classification is discussed in [36].

D. *Performance Metrics*

Overall accuracy, Precision and Recall are used to evaluate the network traffic class predictions. Fig 4. lists the metrics for measuring the classification performance.

$$Overall\ accuracy = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Precision\ by\ application\ group = \frac{TruePositives(X)}{TruePositives(X) + FalsePositives(X)}$$

$$Recall\ by\ application\ group = \frac{TruePositives(X)}{TruePositives(X) + FalseNegatives(X)}$$

4. Conclusion

This survey paper has presented different techniques involved in network traffic classification. The drawbacks of existing approaches are discussed. Features used in various levels for identifying network class are listed. Analysis of Techniques and applications of network traffic classification are also presented. Statistical features in network flow provide appreciable results in identifying type of network flow. Supervised and unsupervised algorithm suitable for predicting the network class are also suggested. The weakness of this method is, we define certain parameters and threshold values experimentally since it does not follow any systematic approach for gesture recognition, and maximum parameters taken in this approach are based on the assumption made after testing a number of images.

5. References

[1] R. Kwitt and U. Hofmann, Unsupervised anomaly detection in network traffic by means of robust PCA, in Proc. Int. Conf. Computing in the Global Information Technology (ICCGI), 2007, pp. 3737.

[2] G. Shen, D. Chen, and Z. Qin, Anomaly detection based on aggregated network behavior metrics, in Proc. Wireless Communications, Network- ing and Mobile Computing, 2007 (WiCom 2007), pp. 22102213.

[3] A. Karasaridis, B. Rexroad, and D. Hoeflin, Wide-scale botnet detection and characterization, in Proc. First Conf. First Workshop on Hot Topics in Understanding Botnets (HotBots07), Berkeley, CA, 2007, p. 7, USENIX Association.

[4] S. Jin, D. S. Yeung, and X. Wang, Network intrusion detection in covariance feature space, Pattern Recognit., vol. 40, no. 8, pp. 21852197, 2007.

[5] A. Sang and S. Li, A predictability analysis of network traffic, in Proc. INFOCOM (1), 2000, pp. 342351.

[6] K. Cho, R. Kaizaki, and A. Kato, An aggregation technique for traffic monitoring, in Symp. Applications and the Internet (SAINT) Workshops, 2002, p. 74.

[7] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, Characteristics of internet background radiation, in Proc. ACM Internet Measurement Conf., Oct. 2004, pp. 2740.

[8] A. Feldmann, A. C. Gilbert, and W. Willinger, Data networks as cascades: Investigating the multifractal nature of internetWAN traffic, in Proc. SIGCOMM, 1998, pp. 4255.

[9] W. Yurcik and Y. Li, Internet security visualization case study: Instru- menting a network for netflow security visualization tools, in Proc. Annual Computer Security Applications Conf. (ACSAC 05), Tucson, AZ, Dec. 59, 2005.

[10] D. Plonka, A network traffic flow reporting and visualization tool, in Proc. 14th USENIX Conf. System Administration, New Orleans, LA, 2000, pp. 305318.

[11] Y. Gong, Security Focus Article: DetectingWorms and Abnormal Ac- tivities With NetFlows, Part 1 Aug. 2004 [Online]. Available: http:// www.securityfocus.com/infocus/1796

[12] Y. Gong, Security Focus Article: DetectingWorms and Abnormal Ac- tivities With NetFlows, Part 2 Sep. 2004 [Online]. Available: http:// www.securityfocus.com/infocus/1796

[13] V. Alarcon-Aquino and J. A. Barria, Multiresolution FIR neural- network- based learning algorithm applied to network traffic prediction, IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 36, no. 2, pp. 208220, Mar. 2006.

[14] Y. Gu, A. McCallum, and D. Towsley, Detecting anomalies in network traffic using maximum entropy estimation, in Proc. 5th ACM SIGCOMM Conf. Internet Measurement (IMC 05), New York, 2005, pp. 16, ACM.

[15] A. Lakhina, M. Crovella, and C. Diot, Mining Anomalies Using Traffic Distributions CS Department, Boston University, Tech. Rep. 2005-002, Feb. 2005.

[16] S. Foresti, J. Agutter, Y. Livnat, S. Moon, and R. F. Erbacher, Visual correlation of network alerts, IEEE Comput. Graphics Applicat., vol. 26, no. 2, pp. 4859, Mar./Apr. 2006.

[17] R. Ellaithi, O. Spelder, N. Brownlee, and J. Pang, Network Event Detection With T-Entropy Centre for Discrete Mathematics and Theo- retical Computer Science, University of Auckland, New Zealand, Rep. CDMTCS-266, May 2005.
[18] A. Lall, V. Sekar, M. Ogihara, J. J. Xu, and H. Zhang, Data Streaming Algorithms for Estimating Entropy of Network Traffic Computer Science Department, University of Rochester, Tech. Rep. TR886, Nov. 2005.
[19] E. F. Harrington, Measuring network change: Rnyi cross entropy and the second order degree distribution, in Proc. Passive and Active Mea- surement (PAM) Conf., Adelaide, Australia, Mar. 2006.
[20] S. Gianvecchio and H. Wang, Detecting covert timing channels: An entropy-based approach, in Proc. ACM Conf. Computer and Communi- cations Security, 2007, pp. 307316.
[21] S. S. Kim, A. L. N. Reddy, and M. Vannucci, Detecting traffic anomalies through aggregate analysis of packet header data, in Networking. New York: Springer, 2004, vol. 3042, pp. 10471059.
[22] M. Celenk, T. Conley, J. Graham, and J.Willis, Anomaly prediction in network traffic using adaptive Wiener filtering and ARMA modeling, in Proc. IEEE Int. Conf. Systems, Man, and Cybernetics, Oct. 2008, pp. 35483553.
[23] X. Fu, B. Graham, R. Bettati, and W. Zhao, On effectiveness of link padding for statistical traffic analysis attacks, in Proc. 23rd IEEE Int. Conf. Distributed Computing Systems (ICDCS 03), Washington, DC, 2003, p. 340.
[24] A. Wagner and B. Plattner, Entropy based worm and anomaly detection in fast IP networks, in Proc. 14th IEEE Int. Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise, Washington, DC, 2005, pp. 172177.
[25] M. Thottan and C. Ji, Anomaly detection in IP networks, IEEE Trans. Signal Process., vol. 51, no. 8, pp. 21912204, Aug. 2003.
[26] H. Hajji, Statistical analysis of network traffic for adaptive faults detection, IEEE Trans. Neural Netw., vol. 16, no. 5, pp. 10531063, Sep. 2005.
[27] S. R. Gaddam and K. S. Balagani, 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., vol. 19, no. 3, pp. 345354, Mar. 2007.
[28] A. Ziviani, M. L. Monsores, P. S. S. Rodrigues, and A. T. A. Gomes, Network anomaly detection using nonextensive entropy, IEEE Commun. Lett., vol. 11, no. 12, pp. 10341036, Dec. 2007.
[29] S. S. Kim and A. L. N. Reddy, Statistical techniques for detecting traffic anomalies through packet header data, IEEE/ACM Trans. Netw., vol. 16, no. 3, pp. 562575, Jun. 2008.
[30] G. Androulidakis, V. Chatzigiannakis, and S. Papavassiliou, Network anomaly detection and classification via opportunistic sampling, IEEE Network, vol. 23, no. 1, pp. 612, Jan./Feb. 2009.
[31] R. Beverly, S. Bauer, and A. Berger, \The internet is not a big truck: Toward quantifying network neutrality," in Proceedings of the 8th International Conference on Passive and Active Network Measurement, ser. PAM'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 135-144. [Online]. Available: http://dl.acm.org/citation.cfm?id=1762888.1762907.
[32] Service Name and Transport Protocol Port Number Registry (IANA), [Online]. Available: http://www.iana.org/assignments/port-numbers, as of October 17, 2014.
[33] Taimur Bakhshi and Bogdan Ghita : On Internet Traffic Classification: A Two-Phased Machine Learning Approach, Journal of Computer Networks and Communications, Vol. 2016 (2016), Article ID 2048302, 2016.
[34] Manuel Lopez-Martin, Belen Carro, Antonio Sanchez-Esguevillas, Jaime Lloret: Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things, IEEE Access, Vol. 5, **pp:** 18042 – 18050, 2017.
[35] Neeraj Namdeva,, Shikha Agrawala, Sanjay Silkaria, Recent Advancement in Machine Learning based Internet Traffic Classification, 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, 2015.
[36] Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari Lu Yao, Nabin Kumar Karn, Foudil Abdessamia : Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms, 2nd IEEE International Conference on Computer and Communications (ICCC), 2016.