DOI:10.21917/ijct.2015.0170

A NOVEL APPROACH FOR REAL TIME INTERNET TRAFFIC CLASSIFICATION

Rupesh Jaiswal[1] and Shashikant Lokhande[2]
[1]*Department of Electronics & Telecommunication Engineering, Pune Institute of Computer Technology, India*
E-mail: rcjaiswal@pict.edu
[2]*Department of Electronics & Telecommunication Engineering, Sinhgad College of Engineering, India*
E-mail: sdlokhande.scoe@sinhgad.edu

Abstract

Real time internet traffic classification is imperative for service discrimination, network security and network monitoring. Classification of traffic depends on initial first few network packets of full flows of captured IP traffic. Practically, the real world framework situation expects correct conclusion of classification well before a flow has ended even if the start of the Traffic flow is missed. This is achieved by calculating features from few N network packets, taken at any random time instant at any random point in the duration of flow. This research proposes a novel parameter Relative Uncertainty (RU) to estimate the level of diversity of internet traffic and can then be used for characterization of internet traffic. Small sub-flows from Full-flows are selected based on minimum RU value (MRUB-SFs: Minimum RU Based Sub Flows), and then features are calculated for training the C4.5 ML classifier. Experimentation is carried out with various standard datasets and results stable accuracy of 99.3167% for different classes of applications.

Keywords:

Classifier, Flows, Relative Uncertainty, Attributes, Packets

1. INTRODUCTION

The fast increase of both legal traffic and attack traffic needs ISPs and network administrators to examine and regulate Internet protocol traffic for both network safety as well as QoS control. Traditional approaches are port based method and payload signature based method. Port based method gives less accuracy and payload signature based method is slower because it consumes the resources heavily. Also payload signature based method is not fruitful for encrypted traffic.

Statistical features based approach [1-4] using machine learning algorithms are most accurate, faster and consistent, which can also be used for encrypted internet traffic classification. In real time classification of various applications, network packets are captured, flows (Unidirectional or bidirectional) are formed and flows based attributes are computed. Based on these attributes, ML classifier is trained and tested for newly captured packets of internet applications.

Most researchers use features computed from full flows along with various ML techniques for internet traffic classification. However, these schemes are not appropriate for real time classification because of non-acceptable processing delay occurs while processing full flows. Particularly, with high speed packet switching networks real time classifier must recognize the applications with restricted resources with finite time limitations. The only way is to obtain small sub-flows from full flows for quick classification. It is investigated that, accurate classification is achieved by selecting the Minimum Relative Uncertainty Based Sub Flows (MRUB-SFs).

For rapid and accurate classification decision, small sub-flows (derived from full flows) are selected based on minimum RU value and then attributes are calculated for training and testing of the ML classifier. Our results shows 100% of Precision and Recall values for gaming traffic and above 90% for other Internet protocols using MRUB-SFs method.

The rest of this paper is organized as follows. In section 2, related research work is described with specific focused on real time classifications of the internet applications using supervised learning techniques. Section 3 shows dataset preparation and methodology used. Proposed work of Minimum Relative Uncertainty Based Sub Flows (MRUB-SFs) classification, its implementation and analysis is explained in section 4. In section 5, the results of MRUB-SFs method and their implications are discussed. Section 6 summarizes the conclusion of paper and outlines potential future works.

2. RELATED WORK

Early traffic classification is proposed by Bernaille [5-7] using Simple K-Means unsupervised ML algorithm which identifies the internet applications using first few packets in a flow result 80% of accuracy but failed to classify the TCP flows if origin of flow is missing.

Shupeng Zhao [8] performs Online Traffic Classification Based on Few Sampled Packets and results accuracy less than 93%.

Online traffic classification Based on Multistage Classifier is experimented by DU Min [9] and results up to 98.15% accuracy using SVM models.

Early traffic classification is investigated by PENG Lizhi [10] by experimenting for the first few packets using various ML algorithms and concludes important observation that five to seven packets are required for early stage internet traffic identification. But unfortunately, notable problem is still unresolved what if origin of flow is missing?

In our paper, real time traffic identification method based on fundamental packet selection strategy is investigated by analyzing the information entropy and computing relative uncertainty for various internet applications. By implementing the MRUB-SFs based technique, quick and accurate classification is achieved with C4.5 ML classifier for all real time captured traffic. MRUB-SFs based approach and its efficacy is demonstrated for the online gaming and VoIP applications where delay cannot be tolerated.

Also, the internet applications like on line bit torrent sites are using HTTPS protocols that fools proxy server and get through the campus network to client desktop. Hence online gaming, VoIP, HTTP and HTTPS protocols are considered for the detail analysis. Our result shows 99.3167% of classification accuracy for these Internet applications.

3. DATASET PREPARATION AND METHODOLOGY

In this paper, C4.5 classifier is used for classification of various Internet protocols which is clearly explained by Witten and Frank [11] and Jaiswal[1]. Publicly available datasets composed of LANs as well as backbone networks traffic. Real time internet traffic is captured and proprietary datasets are developed for experimentation. The details are described in Table.1. The datasets used are of MAWI [12], UNIBS [13], IPLS [14], AUCKLAND [14] and Proprietary.

150 Mbps trans-pacific line since year 2006 provides the data traces files and used as MAWI standard data sets in our research work. Data using 100 Mbps uplink from University of Brescia LAN is collected from campus router-point with network of 20 computing stations to contribute UNIBS data traces.

The Waikato storage project collects the data sets which are the proprietary traces and ownership is of WAND Group. Freely and publicly available data traces are used in our experimentation from these WAND group data sets which are mentioned as IPLS and Auckland data traces on the same site.

Standard data traces of gaming traffic for ET, HL-CS, HL-2, Quake 3 and 4 games are used from the mentioned source [15]. All data trace files are available in pcap/tcpdump file format. In Table.2, the data instances comprising of attributes derived from flows along with its percentage for every class is given in detail.

Table.1. Standard and proprietary datasets used for MRUB-SFs based traffic classification

Datasets	Year	Volume of Traffic
PROPRIETERY	2014	7862 MB
MAWI	2014	6512 MB
UNIBS	2009	2297 MB
IPLS-I	2002	757 MB
IPLS-II	2002	732 MB
IPLS-III	2004	1402 MB
AUCKLAND IV	2001	597 MB
AUCKLAND VI	2001	202 MB

Table.2. Data Instances used for ML based traffic classification

Class	Flow%	Training Instances	Testing Instances	Total Instances
VoIP	22.82	3694	1903	5597
GAME	27.66	4479	2307	6786
HTTP	29.44	4767	2455	7222
HTTPS	20.08	3249	1673	4922

2.5GHz Intel Quad CPU workstation with 4GB of RAM equipped with Ubuntu 12.04 LTS Operating System is used. GCC, OCTAVE and Python are used for computations. Packet header is read using “C” program, Attributes are computed in Octave, WEKA [19] is used for classification and performance analysis is done in detail.

The performance is evaluated with well-known parameters like TP, TN, FP, FN, Precision, Recall, F-measure, ROC, and MAE etc.

Precision (P) =TP/(TP+FP),

Recall (R) = TP / (TP+FN),

Accuracy (A) = (TP+TN) / (TP+TN+FP+FN),

F-measure (FM) =2*(Precision x Recall)/(Precision + Recall),

ROC is known as Receiver operating characteristic graph with FP on the X axis and TP on the Y axis.

$$MAE = \frac{1}{2}\sum_{i=1}^{n}\left|\left(P_i - E_i\right)\right|$$

Mean absolute error, where Pi is the predicted value and Ei is the exact value of data instance calculated from flow features.

Flow is basically a sequence of network packets communicated between two workstations and share same five-tuples: IP address (source & destination), Port address (source & destination) and transport layer protocol with a timeout of ‘t’ seconds (64 seconds taken as standard value) with the reference of SYN and FIN flags. Flows classified based on size, duration, direction, burstiness and rate are described in published literatures [16-18].

Important features are calculated from flows like inter arrival time, duration of flow, length of packet, length of header etc. statistical calculations like SD and Mean are computed based on these basic features. Features sets are well prepared with various search methods and attribute evaluators [3-4] in WEKA tool. Best performing combination is selected for further experimentation and the performance is evaluated.

4. IMPLEMENTATION ANALYSIS FOR SUB-FLOWS

4.1 PROPOSED APPROACH

In proposed novel MRUB-SFs technique, simple sliding window methodology (SWM) is used. SWM use overlapping of windows of size ‘SZ’ (every window consist of ‘NCJ’ number of conjoint packets). Window size ‘SZ’ and number of windows WN are selected based on minimum RU. RU calculation for each window (sub-flow) is done based on IP packet length. Similarly RU is computed for all sub-flows.

Minimum RU based sub-flow is finalized for the flow considered and process is replicated for remaining flows. Attribute computation is done and classifier model is developed. Selection of sub-flows reduces the load on classifier which reduces training time. This makes real time traffic classification process faster. Theoretically, these MRUB-SFs are the members with most uneven distribution. Proposed novel approach is shown in Fig.1. Selected MRUB-SFs are used for computation of features. Training and testing is done as part of supervised learning process and then classification output results in four classes (Online Gaming, VoIP, HTTP and HTTPS) as shown in Fig.2.

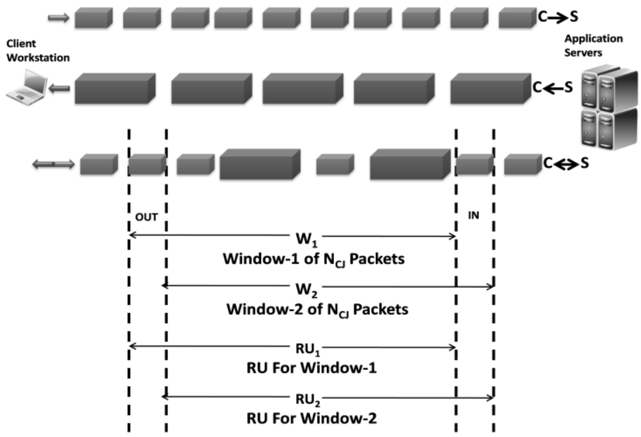

Fig.1. RU is calculated for each Window (Sub-flow)

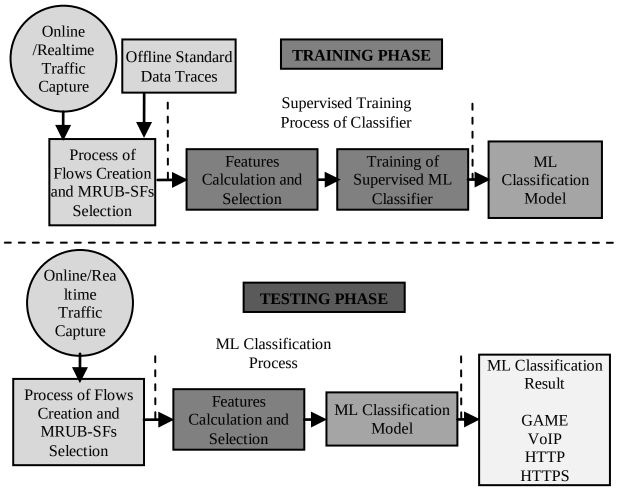
Online /Realtime Traffic Capture
Offline Standard Data Traces
TRAINING PHASE
Supervised Training Process of Classifier
Process of Flows Creation and MRUB-SFs Selection
Features Calculation and Selection
Training of Supervised ML Classifier
ML Classification Model
Online/Realtime Traffic Capture
TESTING PHASE
ML Classification Process
Process of Flows Creation and MRUB-SFs Selection
Features Calculation and Selection
ML Classification Model
ML Classification Result
GAME
VoIP
HTTP
HTTPS

Fig.2. Training and testing phase of ML classification

4.2 RELATIVE UNCERTAINTY COMPUTATION AND CHARACTERIZATION OF TRAFFIC

Consider a discrete random variable *A* which may have total possible outcomes *N* and represented as a_1, a_2, a_3,…a_N. The probability of each outcome is represented as,

$$p(a_i) = P(A = a_i)(\text{where } i = 1,2,3,\ldots,N)\text{and} \sum_{i=1}^{N} p(a_i) = 1$$

Information Entropy of random variable *A* is given as

$$H(A) = -\sum_{i=1}^{n} p(a_i)\log_2 p(a_i)$$

H(*A*) gets maximum value if the distribution of random variable *A* is uniform [21] in nature which is given as

$$p(a_i) = \frac{1}{N}(\text{where } i = 1,2,3,\ldots,N).$$

Thus relative uncertainty is given as

$$RU(A) = \frac{H(A)}{H_{\max}(A)} = \frac{H(A)}{\log(N)}$$

where, $H_{\max}(A)$ is known as maximum information Entropy of random variable *A*. The range of *RU* is between 0 and 1. *RU*(*A*) signifies the relative uncertainty of outcomes of random variable *A*.

RU(*A*) = 0, if random variable A takes single value with probability value = 1.

RU(*A*) = Maximum value, if random variable A results with uniform probability distribution.

Otherwise it takes values between 0 and 1. Smallest *RU* value of the flow based feature of internet traffic indicates maximum uneven nature of its distribution. Also maximum entropy value is observed if statistical parameters are uniformly distributed. Thus *RU* is computed based on Entropy and its value is within 0 to 1.

Theoretically, these MRUB-SFs are the members having similar statistical characteristics (in our case, minimum entropy results minimum RU. It means parameters are most unevenly distributed and hence more homogeneity and therefore observed statistical similarity is greater). Also, maximum entropy results maximum *RU* in our proposed approach. It means parameters are most uniformly distributed and hence more heterogeneity and therefore observed statistical dissimilarity is more. The various traffic parameters and their probability distributions are very well described by mathematical and simulation modeling of HTTP traffic in previous research work [20].

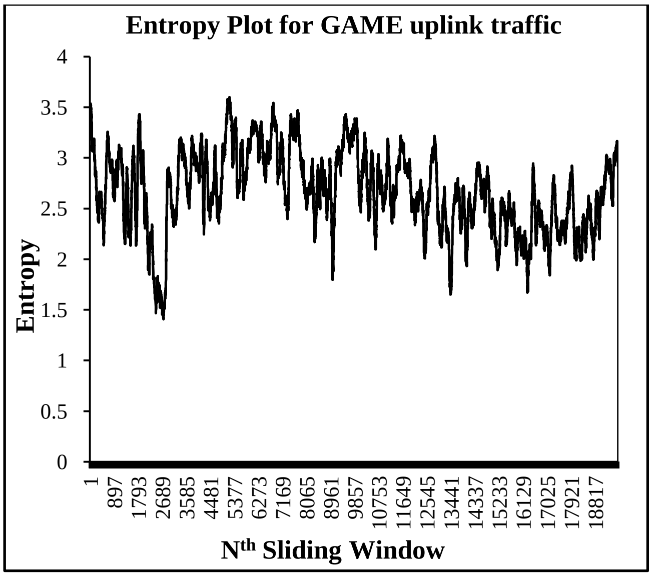

Fig.3. Entropy plot for GAME UL traffic

Following are the entropy versus Nth sliding window graphs drawn for various internet traffic as part of MRUB-SFs based real time internet traffic classification. Entropy plot for GAME Uplink traffic is shown in Fig.3. In this case, the traffic flow from various network game client users which are communicating with game servers are considered only and accordingly those many packets are processed and Entropy is calculated. Entropy plot for GAME Downlink traffic is shown in Fig.4. In this case, the traffic flow from various game servers to network game client users are considered only and accordingly those many packets are processed and Entropy is calculated. The same logic is used for other traffic plots discussed in Fig.5-10.

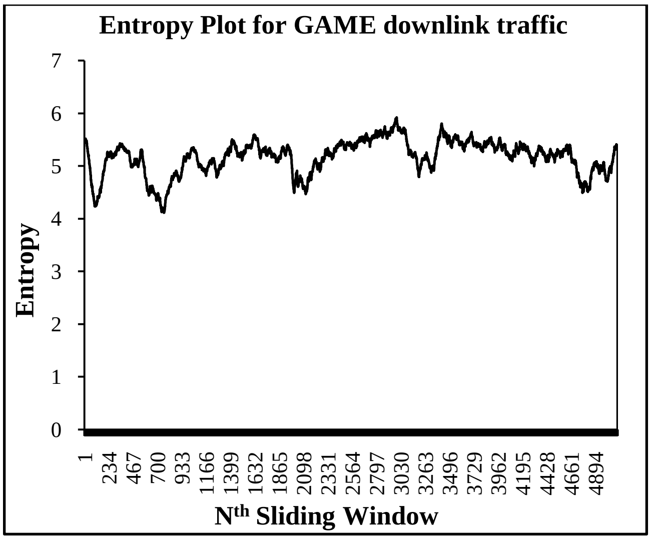

Fig.4. Entropy plot for GAME DL traffic

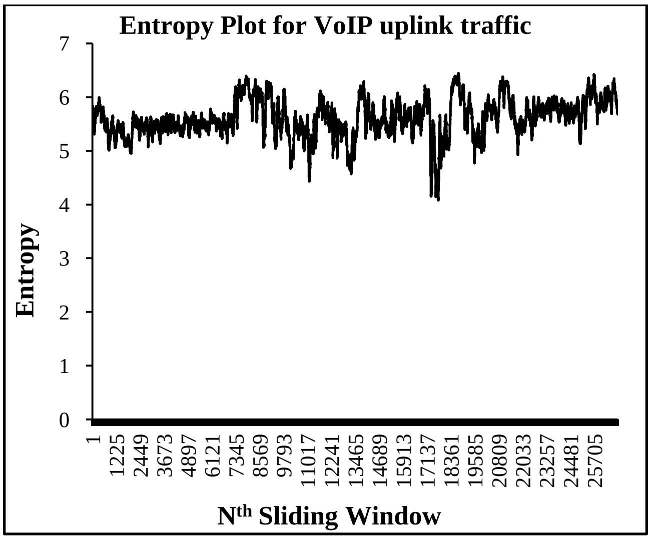

Fig.5. Entropy plot for VoIP UL traffic

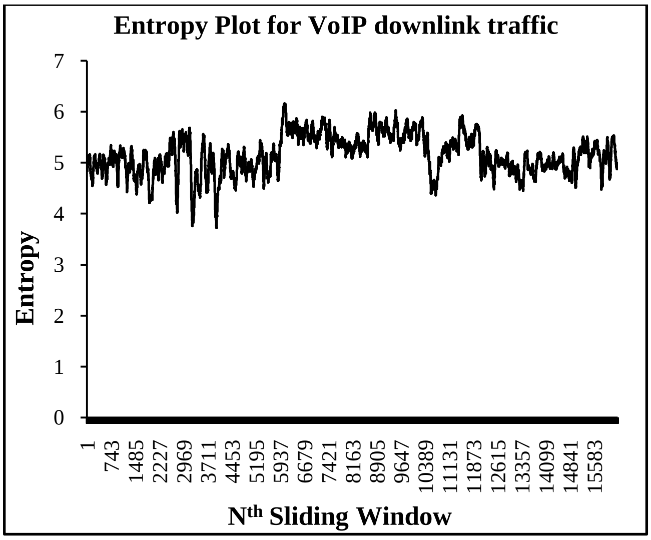

Fig.6. Entropy plot for VoIP DL traffic

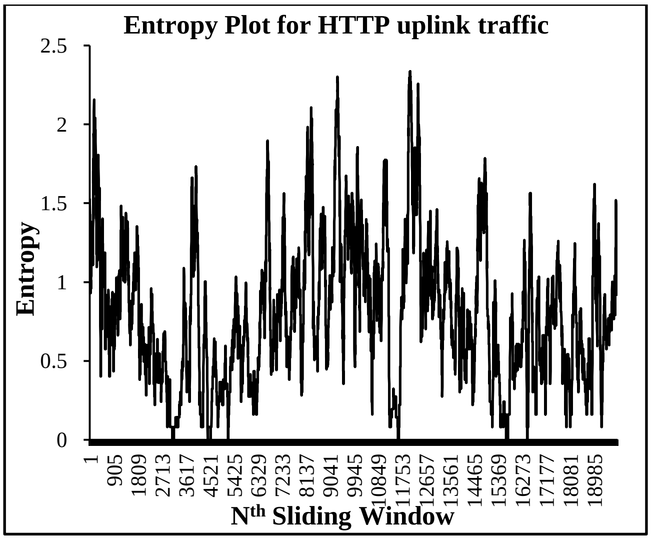

Fig.7. Entropy plot for HTTP UL traffic

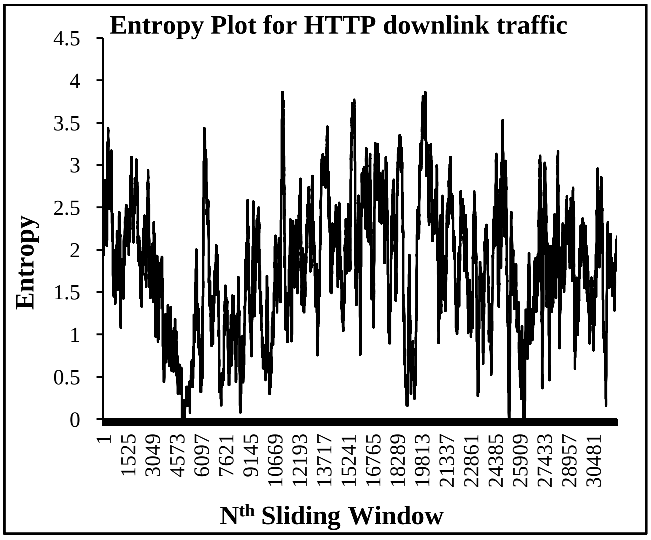

Fig.8. Entropy plot for HTTP DL traffic

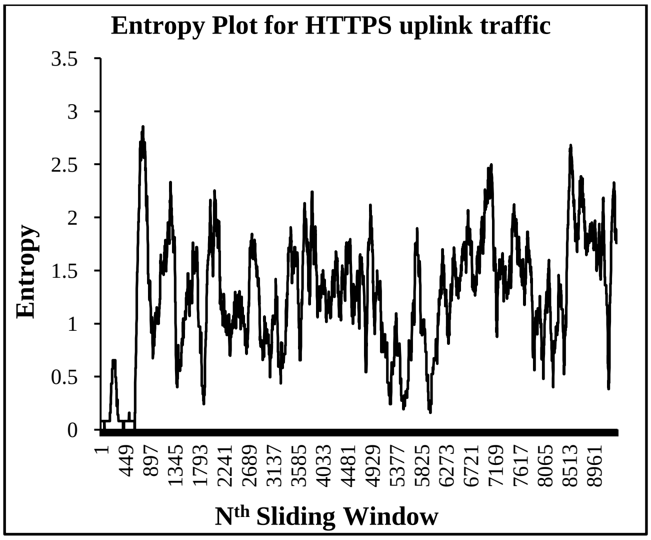

Fig.9. Entropy plot for HTTPS UL traffic

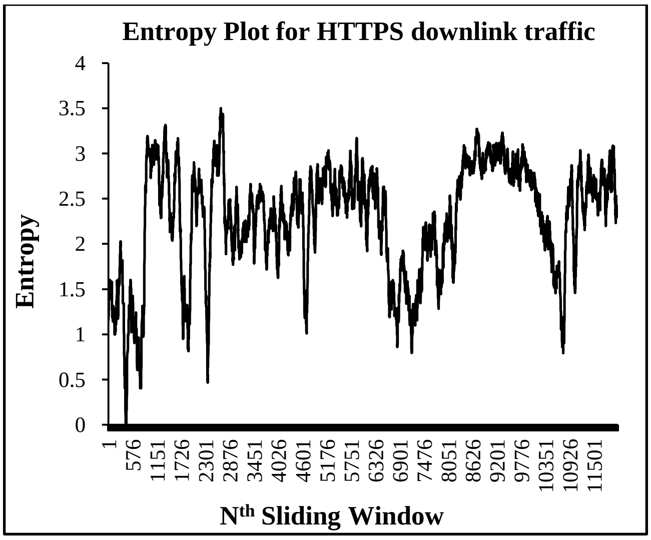

Fig.10. Entropy plot for HTTPS DL traffic

Entropy is calculated for uplink and download scenario separately for better understanding. VoIP traffic gives average entropy of 5.605843 for uplink packets and 5.159745 for downlink packets. Gaming traffic gives average entropy of 2.672886 for uplink packets and 5.173009 for downlink packets. HTTP traffic gives average entropy of 0.805681 for uplink packets and 1.789233 for downlink packets. HTTPS traffic gives average entropy of 1.261376 for uplink packets and 2.280995 for downlink packets.

4.3 RESULTS AND PERFORMANCE ANALYSIS OF ONLY SFS AND MRUB-SFS APPROACH

The impact of sub-flows selection on the performance of C4.5 ML classifier is demonstrated in Fig.11. It takes 35 packets for SFs-15, 7 packets for SFs-30 and 6 packets for SFs-100 to get the accuracy value closer to 99.3167%.Initially 2 packets are taken and attributes are computed based on it for SFs-15, SFs-30 and SFs-100. Same procedure is repeated upup to 100 packets in different steps for SFs-15, SFs-30 and SFs-100. C4.5 classifier is trained using computed attributes and graphs are plotted in Fig.11.

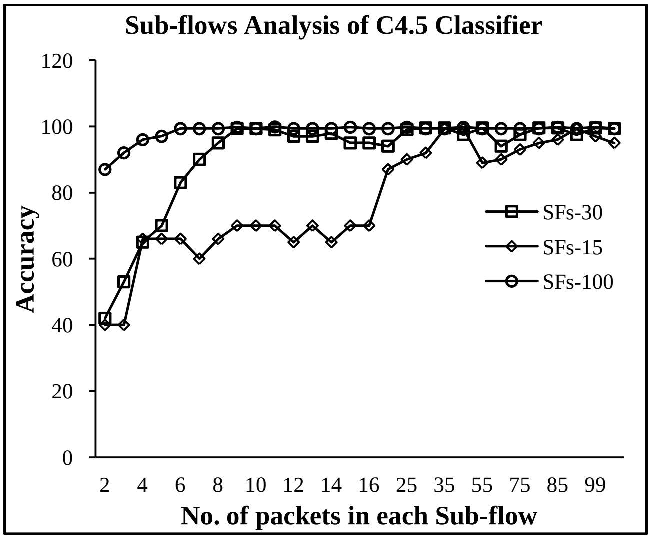

Fig.11. C4.5 with only SFs selection

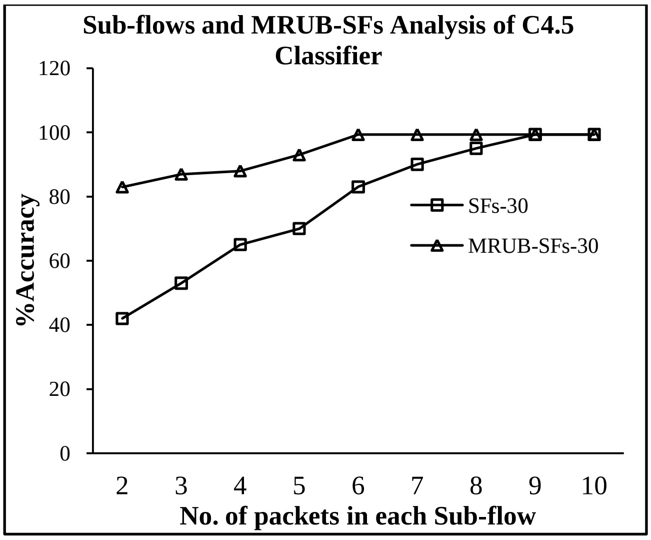

Fig.12. C4.5 with MRUB-SFs selection

The impact of selection of MRUB-SFs-30 on the performance of C4.5 classifier is compared with only SFs-30 in Fig.12. It indicates that for 2 packets the accuracy increased from 42.2 to 83.2%, for 3 packets it is 53.1 to 87.2%, for 4 packets it is 65.04 to 88.12%, for 5 packets it is 69.99 to 93.21%, for 6 packets it is 83.1 to 99.31% and remain stable thereafter. It indicates that MRUB-SFs-30 performance is much better than only SFs-30. This MRUB-SFs-30 accuracy of 99.3167% is same as compared to only SFs-100, which is noticeable investigation in our research experimentations and also assures that less number of MRUB-SFs-30 performances is same as compared to only SFs-100.

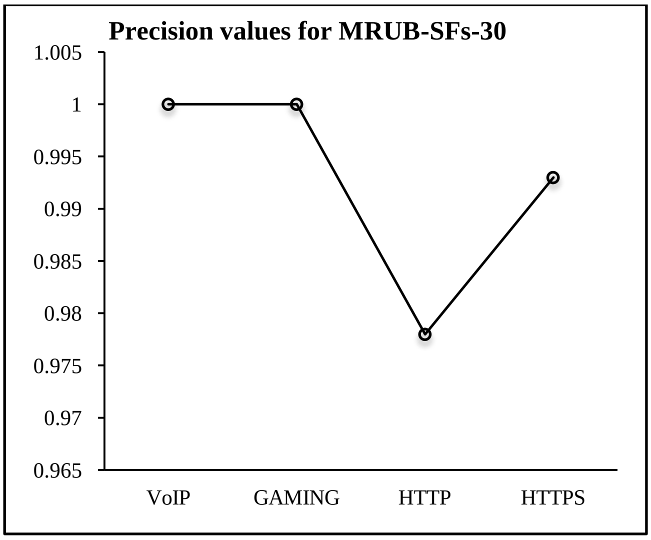

Fig.13. Precision analysis for C4.5 classifier

Detail analysis of MRUB-SFs-30 for C4.5 classifier in the terms of precision, recall, f-measure and ROC values are given in Fig.13 to Fig.16. Precision = TP/(TP+FP), indicates positive predictive value. Recall = TP/(TP+FN), indicates sensitivity characteristics.

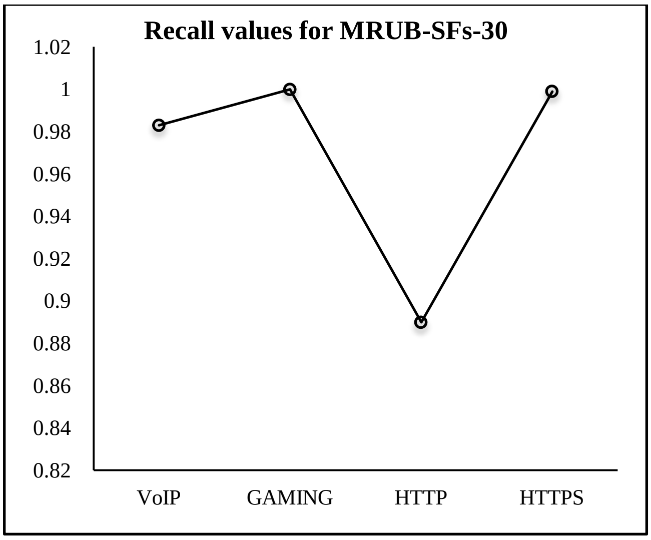

Fig.14. Recall analysis for C4.5 classifier

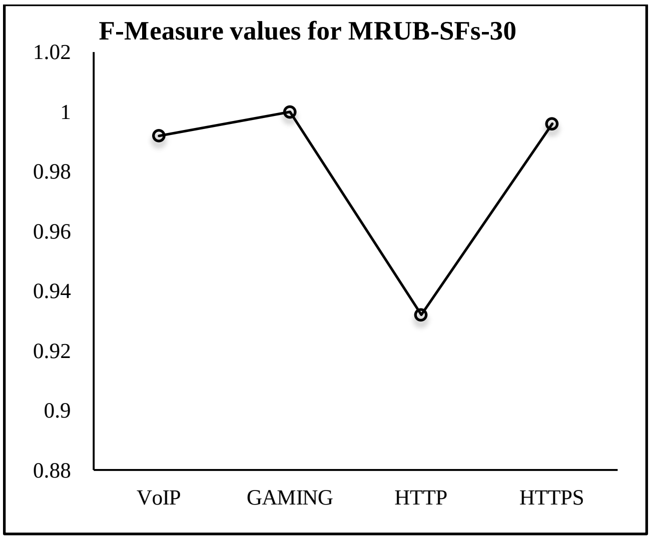

Fig.15. F-measure analysis for C4.5 classifier

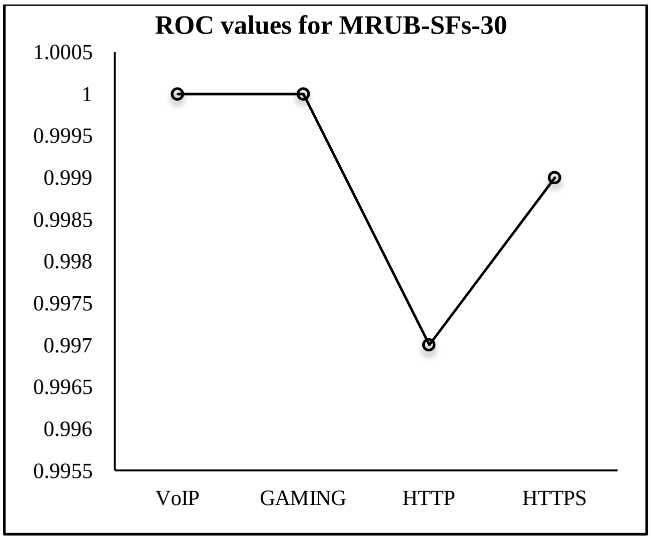

Fig.16. ROC analysis for C4.5 classifier

Table.3. Detail performance evaluation of C4.5 Classifier using MRUB-SFs approach

Class	Precision	Recall	F-measure	ROC
VoIP	1	0.983	0.992	1
GAME	1	1	1	1
HTTP	0.978	0.89	0.932	0.997
HTTPS	0.993	0.999	0.996	0.999

Detail analysis of MRUB-SFs-30 for C4.5 classifier is given in Fig.13 to Fig.16. 100% precision, Recall, F-measure and ROC values are obtained for Gaming traffic is remarkable achievement in this research which assures predictability of C4.5 ML classifier with proposed MRUB-SFs technique. For other traffics, more than 98% of precision Recall and F-measure is obtained. F-measure of 93.2% and Recall of 89% is observed for HTTP traffic which is due to improper filtering and sampling of standard data traces used. The performance evaluation parameters are given in Table.3.

5. RESEARCH DISCUSSIONS

There are many encouraging outcomes and few limitations about our novel MRUB-SFs technique. It works excellently for various standard as well as proprietary data traces. MRUB-SFs technique performs well for all length of flows we experimented in this research work but flows with insufficient packets are discarded here. Related experimentation and detail systematic analysis is left for future research if these discarded packets are taken into account for training the classifier. Also experimentation with other ML classifiers and other internet traffic protocols is our next step in the research work. Effect of forged packets on early traffic classifier makes ML classification unstable which is demonstrated already in our previous research work [3-4]. Also effect of random sampling of data on the performance of MRUB-SFs technique is not discussed and is our next immediate experimentation for all standard data traces. Also, actual deployment on campus gateway, related experimentation and performance analysis is in progress and related work is left for future research.

6. CONCLUSION AND FUTURE SCOPE

In this research paper we studied real time traffic classification by implementing MRUB-SFs technique and achieved 99.3167% accuracy for several traffic classes and provide powerful solution in the field of traffic recognition. 100% precision, Recall, F-measure and ROC values are obtained for Gaming traffic is encouraging achievement in this research which assures positive predictability of our novel approach with C4.5 ML classifier. This is outstanding outcome of our research and shows much better performance than previously published research work.

MRUB-SFs method is very productive and efficient in distinguishing and describing the behavior of various protocols. Simple and prominent features are computed based on MRUB-SFs approach and then appropriate features are selected using search techniques available in WEKA and performance is evaluated for real time traffic classification. MRUB-SFs-30 shows better performance compared to only SFs-30. Accuracy of 99.3167% obtained using MRUB-SFs-30 technique is same as compared to only SFs-100, which is remarkable investigation in our research and confirms that less number of MRUB-SFs-30 performances is challenging to the performance with only SFs-100. Our research has opened up several research issues and pointed towards new path of research. This research is motivating and can be very useful if deployed on the actual network security devices like IDS and IPS in future. Actual implementation is challenging for network operators and network experts for further improvement and deployment in today's fast growing networking world.

REFERENCES

[1] R. C. Jaiswal and S. D. Lokhande, "Machine Learning Based Internet Traffic Recognition with Statistical Approach", *Annual IEEE India Conference*, pp. 1-6, 2013.

[2] Rupesh Jaiswal and Shashikant. Lokhande, "Analysis of Early Traffic Processing and Comparison of Machine Learning Algorithms for Real Time Internet Traffic Identification Using Statistical Approach", *Advanced Computing, Networking and Informatics*, Vol. 2, pp. 577-587, 2014.

[3] Rupesh Jaiswal and Shashikant Lokhande, "Statistical Features Processing Based Real Time Internet Traffic Recognition and Comparative Study of Six Machine Learning Techniques", *8th International Conference on Communication Networks,* 2014.

[4] R.C. Jaiswal and S.D. Lokhande, "Comparative Analysis using Bagging, LogitBoost and Rotation Forest Machine Learning Algorithms for Real Time Internet Traffic Classification", *8th International Conference on Data Mining and Warehousing*, 2014.

[5] Laurent Bernaille, Renata Teixeira, Ismael Akodkenou, Augustin Soule and K. Salamatian, "Traffic classification on the fly", *ACM SIGCOMM Computer Communication Review*, Vol. 36, No. 2, pp. 23–26, 2006.

[6] Laurent Bernaille, Renata Teixeira and Kave Salamatian., "Early application identification", *2nd Conference on Future Networking Technologies*, pp. 1–12, 2006.

[7] Laurent Bernaille and Renata Teixeira, "Early recognition of encrypted applications", *Passive and Active Network Measurement*, Vol. 4427, pp. 165–175, 2007.

[8] Shupeng Zhao, Zhenxiang Chen, Lizhi Peng, Xiaomei Yu and Bo yang., "Online Traffic Classification Based on Few Sampled Packets", *IEEE 14th International Conference on Communication Technology*, pp. 1182-1187, 2012.

[9] DU Min, CHEN Xingshu and TAN Jun, "Online Internet Traffic Identification Algorithm Based on Multistage Classifier", *China Communications*, pp. 89-97, 2013.

[10] Peng Lizhp, Yang Bop, Chen Yuehup and Wu Tong, "How Many Packets Are Most Effective for Early Stage Traffic Identification", *China Communication*, Vol. 11, No. 9, pp. 183-193, 2014.

[11] Ian H. Witten and Eibe Frank, "*Data Mining-Practical machine learning tools and techniques*", Morgan Kaufmann publishers, 2005.

[12] MAWI Working Group Traffic Archive. Packet traces from wide backbone. http://mawi.wide.ad.jp/mawi/.

[13] UNIBS data traces are available at given site. http://www.ing.unibs.it/ntw/tools/traces/

[14] WAND network research group- Waikato Internet Traffic Storage from site http://wand.net.nz/wits/

[15] Tanja Lang Grenville, Tanja Lang, Grenville Armitage, Phillip Branch and Hwan-yi Choo, "A synthetic traffic model for Half-life", *In Australian Telecommunications Networks and Applications Conference,* 2003.

[16] N. Brownlee and K. Claffy, "Understanding Internet traffic streams: dragonflies and tortoises", *IEEE Communications Magazine*, Vol.40, No.10, pp. 110-117, 2002.

[17] Kun - Chan Lan and K. and John Heidemann, "A measurement study of correlations of internet flow characteristics", *International Journal of Computer and Telecommunications Networks*, Vol. 50, No. 1, pp. 46- 62, 2006.

[18] K. Papagiannaki, N. Taft, S. Bhattacharyya, P. Thiran, K. Salamatian and C. Diot, "A pragmatic definition of elephants in internet backbone traffic", *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement*, pp. 175-176, 2002.

[19] Weka website. Available at: http:// www.cs.waikato.ac.nz/ml/weka/

[20] R. C. Jaiswal and S. D. Lokhande, "Measurement, Modeling and Analysis of HTTP Web Traffic", *International Conference on Communication and Computing (ICCC-2014)*, 2014.

[21] Thomas M. Cover and Joy *A.* Thomas, *"Elements of Information Theory",* John Wiley & Sons Inc, 1991.