iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 Issue: IX Month of publication: September 2017

DOI:

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

# A Compound Statistical Network Traffic Classification for VoIP Traffic

A.Jenefa[1], Dr.M.BalaSingh Moses[2]
[1]Teaching Faculty, Department of CSE/IT, Anna University, Trichy
[2]Assistant Professor, HOD, Department of EEE, Anna University, Trichy

***Abstract: Human interaction has changed rapidly in the past few years. Nowadays users are shifting from traditional phone calls to Voice over Internet Protocol (VoIP) applications, especially Skype, Gtalk, Yahoo Messenger, etc., by the reason of developing trends and technologies. A VoIP traffic identification is necessary for many fields, because of generating a large amount of VoIP traffic. Conventional traffic classification methods include dynamic port numbers and deep packet inspection of payload-based methods do not work properly in an encrypted environment. So this paper, we prompt a newfangled scheme of the Machine Learning systems are appropriate for the unique pattern characteristics. So, the classification algorithm using for network traffic flow is an entry to examine the network status. To challenge the problem of vital situation where supervised information and considerable unknown applications are present, a new novel approach called semi-supervised machine learning algorithm is proposed in this research to classify the enormous traffic data. Novel techniques of combining Incremental K-means algorithm and C5.0 Machine Learning algorithm is intended in our work. Furthermore, our proposed scheme exhibits the experimental results show that the algorithm to meritoriously classify the VoIP network traffic in network backbone using machine learning algorithms.***

***Keywords: Semi Supervised Approach, Known Application Identification, Statistical traffic boundaries***

## I. INTRODUCTION

Network traffic classification is one of the most important challenging tasks in most recent few years. The Aspire of network traffic classification is to detect which kind of applications are run by the end user and what is the share of the traffic spawn by the fusion of heterogeneous traffic. The chore of network engineers include the network design, network planning, gathering bandwidth requirement of customers, managing bandwidth consumption, etc., In contemplation of attaining all these chores, it is crucial to empathize network traffic properties which would facilitate to improve network performance.

In past few years, there are an enormous amount of network traffic from various voice established applications communicated via the Peer-to-Peer (P2P) Voice over Internet Protocol (VoIP) over the Internet. A traditional Public Switch Telephone Network (PSTN) normally uses a per-minute charge for long distance. But the VoIP takes the low free cost per call in obstructive environments. There are many VoIP products that are gifted to[i] present high call quality such as Microsoft Messenger [1], Skype [2], Gtalk [3], Yahoo! Messenger (YMSG) [4]etc.,. In order to classify the VoIP based traffic is very essential, because of a volume of applications employed on the network is swelling. Therefore an efficient classification of network traffic denotes ultimate dispute for network organization task such as managing bandwidth budget and certify the quality of service objectives.

A number of traffic classification techniques have been anticipated for categorizing the traffic. The various traffic classification methods are port-based, payload-based and flow statistics-based [5].The conventional port based method relies on read-through reserved ports used by the eminent application. But the P2P Voice over Internet Protocol(VoIP) applications obfuscate themselves by issuing dynamic ports including the port numbers registered for well-known protocols by IANA (Internet Assigned Network Authentication) [6].So it is difficult to classify such type of application using a port-based technique.

A substitute technique for the port-based technique was the inspection of the packet payloads [7] [8] [9].The Payload based technique avoids dependency for a port number. In this technique, they are matching payload of packets with well-known signature. In this technique setup the constraints according to different types of payload matching. But this method fails for two drawbacks: 1) It cannot deal with encrypted packet because we cannot apply Deep Packet Inspection technique with encrypted packets.2) It takes low processing efficiency, too much time to classify the packets.

Therefore several types of research treat learning techniques using the statistical flow features estimate from network flow traffic [10] [11] [12]. An efficient way of identifying traffic streaming approach is to use machine learning technique to estimate the

classifier is used to identify those traffic according to packet statistical features like maximum, minimum, standard deviation packet length. Machine learning techniques mainly based on supervised and unsupervised learning.
The classification accuracies rely on supervised machine learning algorithms are evaluated by applying them to test data sets. Machine learning requires training data to characterize the different application. Unsupervised learning algorithm [13] is an arrangement of a sample that has similar way to cluster, with no prior knowledge. Supervised learning algorithm [14] learn a classifier from the data set labeled training samples traffic based on pre-defined classes. The supervised learning goal is to identify a mapping from input feature to an output class. Two major phase in supervised learning.1) Training phase (training dataset) 2) Testing phase (classification).
In this paper, a new novel approach called semi-supervised machine learning algorithm is proposed in this research to classify the enormous traffic data.Novel techniques of combining Incremental K-means algorithm and C5.0 Supervised learning algorithm is intended in our work. The Incremental K-means algorithm used for clutch the new data and the previous cluster is updated into a new cluster. A C 5.0 supervised learning algorithm is used to generate the decision tree. Decision tree generated by C5.0 algorithm is used for classification. A C5.0 algorithm will be applying on the dataset would allow predicting the target variable of a new dataset record. Furthermore, our scheme demonstrates the experimental result show that the algorithm to meritoriously classify the network traffic flow in network backbone using machine learning algorithm. Table I represents the VoIP traffic mix involved in our work

## II. RELATED WORK

A conventional network traffic classification is trust on port based or signature based or connection patterns based classification. These methods are suffered from more than a few limitations are discussed below. The Related Experimental research work is shown in Table 2.

TABLE I: VOIP TRAFFIC MIX IN THE INTERNET BACKBONE

| SI.NO | Application Group | Transport Layer Protocol |
|---|---|---|
| 1 | Skype | TCP/UDP |
| 2 | Facebook | TCP/UDP |
| 3 | Google Talk | TCP/UDP |
| 4 | MSN Messenger | TCP/UDP |
| 5 | Yahoo Messenger | TCP/UDP |
| 6 | Hangouts Voice calls | TCP/UDP |
| 7 | YouTube | TCP/UDP |
| 8 | BitTorrant | TCP/UDP |
| 9 | eDonkey | TCP/UDP |
| 10 | Games | TCP/UDP |

### *A. Port-based VoIP traffic classification*

Previously, a network traffic flows are classified by using port numbers. Essentially, port numbers include the five basic ranges of port numbers. In this approach, traffic classification is based on relating a well-defined port number in TCP or UDP packet headers which are reserved by IANA (Interne Assigned Network Authority) which are used in most of the applications to which other hosts pledge the communication. Then a classifier which is placed in the middle of the network aspects for SYN packets which are TCP packets utilized during 3-way handshake process, to classify the server side of the communication. And then, the packet also includes the target port number to identify the IP traffic. In a similar way, UDP also uses port numbers, although there is neither connection establishment nor preservation of connection state. Moore et al [1] exhibit 70% of the time, a classification is accurate based on IANA port list.
Presently, A newer application that includes P2P VoIP application may not register well-defined port no by IANA to avoid being detected or applications such as FTP in passive mode, their ports are rehabilitated dynamically [2]. Williamson et al [3] substantiate 30-70% of their network flows misclassify using IANA port list. Because there are generate the dynamic port numbers automatically instead of the well-known port number. Sen et al [4] preserved that only 30% of the total network traffic in bytes for Kazaa P2P protocol could be found using registered port no. Hence this method is declined for VoIP traffic classification.

### *B. Signature based VoIP traffic classification*

To overcome the port based VoIP traffic classification we innovate the signature based classification. In this technique the network traffic packets having Packet header information and Payload information. Generally, every application in a network have the statistical characteristics and it's created a reference database. The categorization mechanism compares the traffic again to its reference to identify the exact application. The packet header includes source and destination Address and the payload based methods utilize the Deep Packet Inspection (DPI) scheme to inspect the application in network traffic. It is able to perform classification accurately. For example, web traffic can be recognized with '\GET' string. EDonkey P2P can include 'xe3\x38' string and etc., P2P traffic detection [4] and intrusion detection [5] is customarily used in this approach.

This method takes long time processing and impenetrability. The payload information is encrypted by the purpose of user privacy protection. Therefore, it does not work well with the encrypted environment. So, it is difficult to classify the network traffic on the internet.

### *C. Connection Pattern based VoIP traffic Classification*

It categorizes the traffic based on observing and identifying the configuration of host behavior at the transport layer. The foremost benefits of this technique there is no requirement for the payload of packets and not essential for port numbers. Karagiannis et al.[6]apply a host behavior to classify the P2P traffic using various levels which are functional level, social level, and application level. The functional level gives the information about whether intended host provides or consumes particular service. The social level inspects the status of the host. Finally, the application level anticipated discovering the identification of application of the starting point is considered.

This approach cannot classify network traffic accurately because of using the same behavior to classify the different group and also which takes the enormous period to categorize the network traffic applications. Additionally, this method can just contract with extensive classes of applications which cannot discriminate between individual traffic in the same group. Furthermore, encrypted header information of the host and real-time classification is not appropriate by this method.

### *D. Statistical Approach for VoIP Traffic Classification*

Owing to the number of limitations of the traditional technique, we introduce a machine learning approach which is the capability of computer ability to learn about the environment without explicitly programmed. These approach consist of various categories are Supervised, Unsupervised and semi-supervised machine learning approach. It is unique feature approach to analysis the network status using traffic flows for packet size, packet inter-arrival time etc.

### *E. Unsupervised Machine Learning Approach*

Unsupervised machine learning technique is a responsibility of inferring a function to illustrate the innovative structure from the unlabelled flows. This method is not handled by any training samples further it can produce a training data using clustering method. And also this method to clutch the similar data items from the unlabeled testing data. Zander et al [7] proposed auto class unsupervised clustering to identify the traffic in the network using the parameters include flow size, packet length in TELNET, FTP and SMTP traffic. Bernallie et al[8] apply unsupervised clustering algorithm includes K-means and cluster analysis tool to identify the peer to peer network traffic.

### *F. Supervised Machine Learning Approach*

The supervised traffic classification analyses the labeled training data in network traffic. Supervised learning divided samples into classes of application. In this method, all data is labeled and it produces the inferred function the algorithm learn to predict the output from the input data. It is essential to note that it is termed as supervised because the output classes are predefined. Bonfiglio et al [9] apply supervised learning techniques for classifying Skype traffic. They achieve the best performance by using Pearson's Chi-Square and Naïve Bayesian classifier in P2P VoIP traffic. To anticipated for identifying the application names in network traffic using testing data set of a packet length and mean inter-packet gap pre-labelled data trained data model. Also, supervised learning applied to payload based traffic classification. Nguyen and Armitage[10] proposed supervised machine learning techniques to identify multiple applications such as P2P, HTTP, HTTPS, DNS, NTP, SMTP,etc.,

### *G. Semi-Supervised Traffic Classification*

Semi-supervised machine learning is the combination of Supervised (labeled data) and Unsupervised (unlabelled flows) learning approaches**.** T.Bakhshi et al.[11] apply a semi-supervised classification algorithm includes the combination of K-means and C5.0 decision tree to identify various traffic such as video streaming, P2P, games and etc., using statistical features includes packet and

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887*
*Volume 5 Issue IX, September 2017- Available at www.ijraset.com*

data rate, port number and labels, protocol(TCP and UDP), flow duration and packet counts. Valentin et al[12] apply both DPI and C5.0 decision tree algorithm to identify P2P, VoIP, Multimedia and etc., to achieve high performance.

### *H. Our Innovative MSC for VoIP Traffic*

In this paper, our extension is based on Valentine et al [12], in addition to publicized for handling with the newer applications in the University Campus Network. To challenge the achievement of new-fangled application, the semi-supervised approach effectively identifies the VoIP traffic.

*1)* The traffic flows are collected from the University Campus Network and extract the features from collected dataset.
*2)* In network traffic, the incoming packets of the unlabelled data are converted into labeled data (VoIP) using Incremental K-means clustering.
*3)* The unknown traffic flows are compared with training dataset using C5.0 classifier and then gain the VoIP application

Table II: Related Work In Network Traffic Schedule Scheme

| S.No | Environment | Feature used | Nature | ML algorithms | Evaluated traffic | Granularity |
|---|---|---|---|---|---|---|
| 1 | Bonfiglio et.al [9] | 1. length captured<br>2. Mean inter-packet gap | Supervised learning | Pearson's Chi-Square and Naïve Bayesian classifier | Skype application | Fine-grained. |
| 2 | Erman et al. [13] | 1.Total number of packets<br>2.Mean packet length<br>3.mean payload length excluding headers<br>4.Number of bytes transferred | Semi-supervised learning | Naïve Bayes and Auto class | Web, P2P, FTP, Others | Coarse-grained. |
| 3 | Nguyen and Armitage [10] | 1.Packet lengths (min, max, mean, standard deviation)<br>2. Inter-Packet lengths statistics (min, max, mean, standard deviation)<br>3.Packet Inter-arrival times statistics (min, max, mean, std dev | Supervised classification | Naïve Bayes | P2P, HTTP, HTTPS, DNS, NTP, SMTP, Telnet, SSH | Coarse-grained. |
| 4 | Zander et al [7] | 1.Flow Size and Duration,<br>2.Packet length statistics<br>3.Inter-Arrival time statistics | Unsupervised Clustering | Auto class | DNS,SMTP,TELNET, FTP, NAPSTER | Coarse-grained. |
| 5 | Alshammari [14] | 1.Forward packet inter-arrival time (min, max, mean, std dev)<br>2.Backward packet inter-arrival time (min, max, mean, std dev)<br>3.Forward Packet length (min, max, std dev) | Supervised classification | C4.5, AdaBoost and Genetic Programming (GP) | Skype, Gtalk | Fine-grained |
| 6 | T.Bakhshi et al.[11] | 1.Port Number and Labels<br>2.Protocol(TCP and UDP)<br>3.Packets Count<br>4.Packet Rate and Data Rate<br>5.Flow Duration | Semi-supervised machine learning | | VoIP, P2P, VIDEO STREAMING, GAMES and others | Fine-grained |

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | Valentin et al [12] | 1.Average Packet Size<br>2.protocol(TCP and UDP)<br>3 Flow Time and Rate.<br>5 Inter-Arrival Time | Semi-supervised learning | DPI and C5.0 decision tree algorithm | P2P, VoIP, Multimedia and others | Coarse-grained |
| 8 | ThomasKaragiannis et al. [6] | Various level for Social level, functional level, and application level are used to capture the behavior from a host. | Connection pattern-based classification | Behaviour based Classification | P2P | Coarse-grained |
| 9 | Bernaille et al.[8] | 1.Packet Size of TCP flow (First few packets) | Unsupervised clustering | K-means and cluster analysis tool | SMTP,POP3,FTP,HTTP, KAZAA, SSH AND eDONKEY | Fine Grained |
| 10 | J.Zhang et.al [15] | 1.3-Tuple Data.<br>2 Number of Packets and Bytes-Volume<br>3. Inter-arrival time among packets (minimum, mean, maximum and standard deviation) | Semi-supervised | Flow Statistical K-Means Clustering + Compound Classification | BITTORRENT, EDONKEY, AND OTHERS | Fine Grained. |

## III.THE COMPOUND CLASSIFICATION SCHEME

In this section, established on clustering and classification towards examining the network traffic using labeled and unlabeled flow. The detection process can be segmented into two phases which are 1) Offline phase clustering and 2) Online phase classification to classify the VoIP traffic. Phase-I is used to cluster the VoIP application using K-means algorithm and Phase-II is used to classify the VoIP traffic using C5.0 algorithm. Fig 1 illustrates the System architecture of Compound Classification scheme.

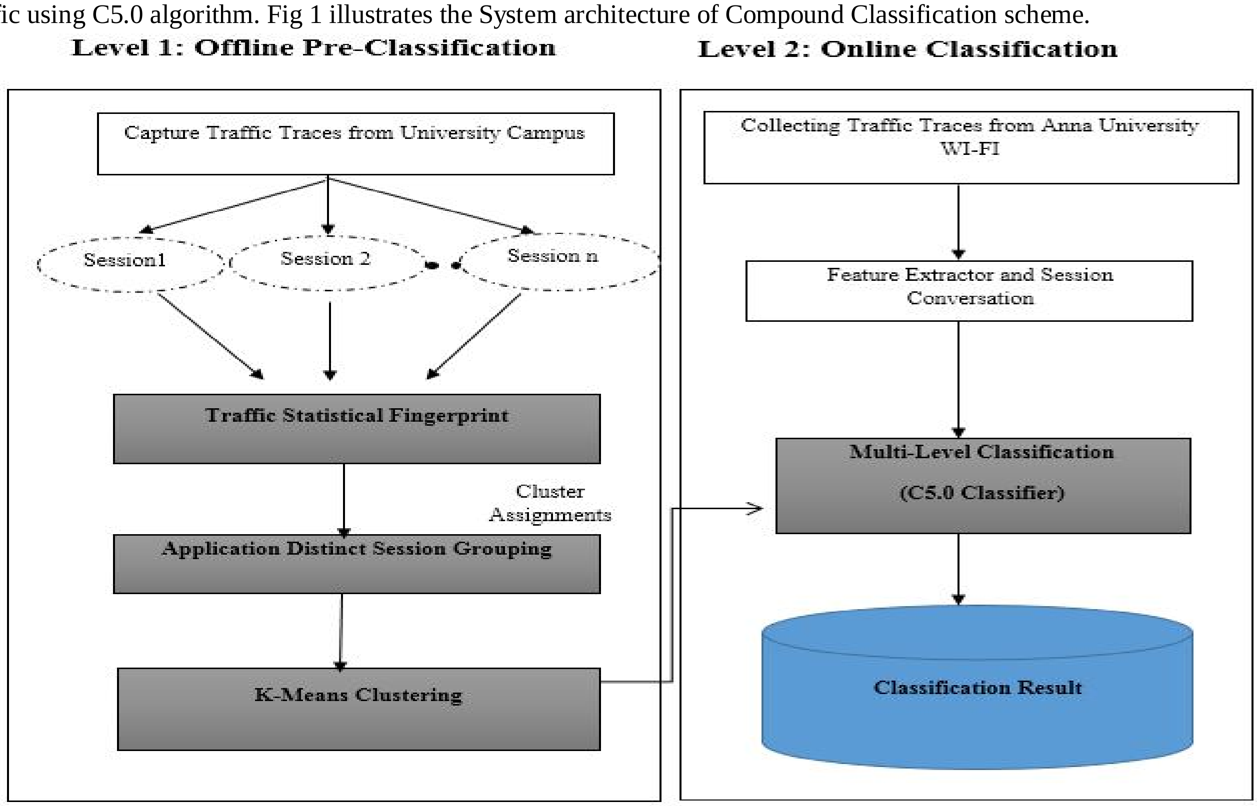



Fig. 1 System Architecture of Compound Classification Scheme

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887*
*Volume 5 Issue IX, September 2017- Available at www.ijraset.com*

### *A. Offline Phase Clustering*

The following subsections describes the methodology involved in the offline pre-classification

### *B. Traffic Statistical Fingerprint*

By identifying the characteristics of traffic applications, each application can be distinguished by their size of the packets. The flows of the identical traffic class have similar Packet Size distributions. By hand the traffic traces for each specific application are collected from Wi-Fi in our campus is shown in Fig 2.

Ethernet · 37 | IPv4 · 35 | IPv6 · 4 | TCP · 3 | UDP · 134

| Address A | Port A | Address B | Port B | Packets | Bytes | Packets A → B | Bytes A → B | Packets B → A | Bytes B → A | Rel Start | Duration | Bits/s A → B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0.0.0 | 68 | 255.255.255.255 | 67 | 7 | 2426 | 7 | 2426 | 0 | 0 | 21.533491 | 72.7843 | 266 |
| 10.1.173.130 | 137 | 10.1.173.255 | 137 | 1 | 92 | 1 | 92 | 0 | 0 | 203.845719 | 0.0000 | — |
| 10.1.173.130 | 138 | 10.1.173.255 | 138 | 1 | 243 | 1 | 243 | 0 | 0 | 600.782597 | 0.0000 | — |
| 10.1.173.152 | 137 | 10.1.173.237 | 137 | 1 | 104 | 1 | 104 | 0 | 0 | 37.669351 | 0.0000 | — |
| 10.1.173.152 | 138 | 10.1.173.255 | 138 | 2 | 495 | 2 | 495 | 0 | 0 | 112.900031 | 255.2192 | 15 |
| 10.1.173.152 | 137 | 10.1.173.255 | 137 | 133 | 12 k | 133 | 12 k | 0 | 0 | 292.763323 | 379.5322 | 257 |
| 10.1.173.152 | 137 | 10.1.173.159 | 137 | 1 | 104 | 0 | 0 | 1 | 104 | 292.764555 | 0.0000 | — |
| 10.1.173.152 | 138 | 10.1.173.159 | 138 | 1 | 236 | 1 | 236 | 0 | 0 | 292.765023 | 0.0000 | — |
| 10.1.173.152 | 68 | 255.255.255.255 | 67 | 4 | 1368 | 4 | 1368 | 0 | 0 | 530.156178 | 76.7518 | 142 |
| 10.1.173.152 | 61471 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 541.169744 | 0.1000 | 10 k |
| 10.1.173.152 | 55526 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 543.753139 | 0.0997 | 10 k |
| 10.1.173.152 | 54499 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 546.344285 | 0.0997 | 10 k |
| 10.1.173.152 | 50211 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 549.061570 | 0.1004 | 10 k |
| 10.1.173.152 | 61136 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 551.640558 | 0.0996 | 10 k |
| 10.1.173.152 | 50705 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 556.943521 | 0.0998 | 10 k |
| 10.1.173.152 | 64255 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 559.516832 | 0.0995 | 10 k |
| 10.1.173.152 | 64513 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 562.103949 | 0.0996 | 10 k |
| 10.1.173.152 | 63357 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 564.715854 | 0.0997 | 10 k |
| 10.1.173.152 | 59151 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 567.324661 | 0.1000 | 10 k |
| 10.1.173.152 | 64035 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 569.964669 | 0.1004 | 10 k |
| 10.1.173.152 | 58213 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 572.535213 | 0.0995 | 10 k |
| 10.1.173.152 | 65049 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 577.813024 | 0.1030 | 9940 |
| 10.1.173.152 | 58024 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 580.400089 | 0.1002 | 10 k |
| 10.1.173.152 | 64443 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 582.977483 | 0.0998 | 10 k |
| 10.1.173.152 | 58796 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 585.616877 | 0.1004 | 10 k |
| 10.1.173.152 | 53831 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 588.191929 | 0.0995 | 10 k |
| 10.1.173.152 | 62473 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 590.771408 | 0.1001 | 10 k |
| 10.1.173.152 | 51938 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 593.414410 | 0.1002 | 10 k |
| 10.1.173.152 | 52709 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 596.057307 | 0.1005 | 10 k |
| 10.1.173.152 | 59709 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 598.732831 | 0.1000 | 10 k |
| 10.1.173.152 | 51679 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 601.335396 | 0.0995 | 10 k |
| 10.1.173.152 | 60601 | 224.0.0.252 | 5355 | 2 | 128 | 2 | 128 | 0 | 0 | 614.923620 | 0.1001 | 10 k |

Fig. 2 Statistical Parameters

The packets are grouped together to produce high-quality clustering trained data samples. The packet sizes and quantity is used for building up the Statistical protocol for each application traffic samples. Traffic flows are collected from the network and patent with application labels based on the Statistical Protocol. But the Statistical protocol has to be updated recurrently for better performance. Then the class labeled flows are then passed on to application distinct flow cluster.

### *C. Application Distinct Session Grouping:*

In this section, we extracted the feature from the collected dataset of VoIP application using TCP stat. Initially, as to group the individual traffic flows in an application, the flow based statistical features is used. The collected traffic is extracted for the necessary parameters of 5-tuple information of the packet including the statistical characters of Source IP address, Destination IP addresses, Source Port Number, Destination Port number and Application Protocol used is shown in Table III. To avoid deep packet inspection, the flow features are examined by the packet header. The different traffic flows are grouped by the statistical feature pick up from the IP packet header. The Source IP address and Destination IP address of two different individual flows are the same and assign successive port numbers, then the simultaneous streams will be grouped as clusters.

Table III: Application Distinct Session Grouping

| Source IP_ address | Destination IP_ address | Source Port Number | Destination Port Number | Grouping of Session (Session_id) |
|---|---|---|---|---|
| 10.0.0.1 | 192.168.2.51 | 43321 | 80 | X |
| 10.0.0.1 | 192.168.2.51 | 43321 | 80 | X |
| 10.0.0.1 | 192.168.2.51 | 43321 | 80 | X |
| 10.0.0.1 | 192.168.2.51 | 43321 | 80 | X+1 |
| 10.0.0.1 | 192.168.2.51 | 43321 | 80 | X+1 |
| 10.0.0.1 | 192.168.2.51 | 43321 | 80 | X+1 |
| 10.0.0.1 | 74.125.236.181 | 54467 | 80 | Y |

### D. Unsupervised Clustering

The packets that are captured were cluster analysed independently for each application using the computationally efficient implementation of -means in*R*. Since the value of*k* directly influences the number of flow clusters (classes) per application. In network traffic, the incoming packets are grouped together based on Euclidean distance. Initially, the clusters are randomly selected and the clusters contain the cluster centroid. The incoming packets are compared with cluster centroid and the data points are partitioned based on minimum distance to generate the new cluster. This process continues until the clusters are stabilized. The Segregated different flows of application traffic are grouped into 10 (k) clusters of the corresponding class as depicted in Figure 3.


Fig. 3 The Segregated App Sessions Plotted Using Fviz_Cluster

In general, we have *n* data points **a***t*, *t*=1...*n*th at having to be subdivided into *k* clusters. The goal is to dispense a cluster to each data point. K-means is a clustering method that aims to find the points $\mu t$,*t*=1...*k* of the clusters that minimize the *distance* from the data points to the cluster. K-means clustering solves,

$$\arg\min \sum_{t=1}^{k} \sum_{a \in ct} d(a, \mu t) = \arg\min \sum_{t=1}^{k} \sum_{a \in ct} \|a - \mu t\|_2^2$$

Where **c***t* is the set of points that belong to cluster *t*. The K-means clustering uses the square of the Euclidean distance *d* (**a**,$\mu t$)=‖**a**−$\mu t$‖22. This problem is not trifling (in fact it is NP-hard), so the K-means algorithm only hopes to find the inclusive least, possibly getting obstructed in a different solution.

### E. Online Phase Classification

C4.5 is a decision tree Machine Learning algorithm used to develop Univariate decision tree. It is an augmentation of Iterative Dichotomise 3 (ID3) algorithm which is used to an invention of simple decision trees.C4.5 algorithm using the concept of information entropy to make a decision tree from a set of training data samples. The training data set encompasses of a countless number of training samples which are regarded by different aspects and it also consists of the objective class. C4.5 pick out a particular attribute of the tree which is used to apportion its set of data samples into subsets in one or another class. It is used for the principle of normalized information gain that is attained by selecting an attribute for excruciating the data. The attribute with the maximum normalized information gain is preferred and made a decision and this process repeats until the smaller subsets. C4.5 has made various enhancements to ID3 like it can handle both continuous attributes and discrete attributes, it can handle training data with missing elements values, and it can also handle attributes with conflicting costs etc resulting with the greatest accuracy for large datasets.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

This portion deals with the experimental handling of the proposed idea with its results and discussions. The software system necessities for the trial work include the mainstream system Intel core 3 Duo Processor 2.20GHz, 4.00 GB RAM, Windows 7, Windows 10 and Ubuntu 14.04 operating system which is upright to run the proposed idea. We used R programming language for the execution of the proposed idea. The application traffic is assembled using packet sniffer tools like using Wireshark and the features are extracted using GCC compiler. The accuracy rate can be improved by using a number of cases for training and testing phases. Figure 4 shows the misclassification table using Feature Set collected for training and classifying online traffic in R.

```
Evaluation on training data (10000 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	    23   14( 0.1%)   <<


	   (a)   (b)   (c)   (d)   (e)   (f)   (g)   (h)   (i)   (j)    <-classified as
	  ----  ----  ----  ----  ----  ----  ----  ----  ----  ----
	  1005                                                          (a): class BitTorrant
	     3  1025                                                    (b): class eDonkey
	               992                                              (c): class Facebook
	                     996                                        (d): class GAME
	                          1019                                  (e): class GoogleTalk
	                                1002                            (f): class Hangouts
	                                       958                      (g): class MSN Messenger
	           5           5                     974                (h): class skype
	                                                  1018          (i): class Yahoo Messenger
	                       1                                   997  (j): class YouTube
```

Fig. 4 Misclassification Table Using C5.0 Classification Algorithm

## V. CONCLUSION

This paper used a dual machine learning approach for traffic identification on a per-flow basis by uniquely using the Statistical features. In the offline phase, the flows for all applications were collected and cluster scrutinized consequence in 5 unique flow application. The online phase used the statistical set of elements from the derived per-flow classes to test and train the C5.0 decision tree classifier. Consideration factor of the classifier was also enormously great fluctuating above 90%.The consistent required factor, a transfer for classifier flow perception competence series for all applications. In addition, the elementary exactness of the present approach is achieving excessive granular flow application discover and the estimated efficiency in associate with other machine learning organization manner for forthcoming exertion in encompassing this technique to incorporate additional solicitation for real-time based classification.

## VI. ACKNOWLEDGMENT

This research paper is made possible through the help and support from everyone. First and foremost, I would like to acknowledge and extend my heartfelt gratitude to my Ph.D. guide Mr. BalaSingh Moses who has helped me to complete this research possible. Second, I would like to thank my department colic's, for their vital encouragement and support. Most especially to God, who made all things possible. At last, the authors would like to thank our Anna University Campus admin for allowing us to monitor the VoIP traffic

## REFERENCES

[1] Moore, A.W., Papagiannaki, K., 2005. Toward the accurate identification of network applications. In: Passive and Active Network Measurement: Proceedings of the Passive & Active MeasurementWorkshop. pp. 41–54.

[2] Abuagla Babiker Mohd and Sulaiman bin Mohd Nor. Towards a Flow-based IP traffic classification for Bandwidth Optimization, International Journal of Computer Science and Security (IJCSS), Vol. 3, Issue 2, pp. 146-153

[3] Madhukar, A., Williamson, C., 2006. A longitudinal study of p2p traffic classification. In: Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2006. 14th IEEEInternational Symposium on MASCOTS 2006. pp. 179–188.

[4] S. Sen, C. Spatscheck, D. Wang, "Accurate, scalable in network identification of P2P traffic using application signature", in 13th International Conference on World Wide Web, 2004.

[5] .K. Wang, S.J. Stolfo, "Anomalous Payload-based network intrusion detection", in Lecture Notes in Computer Science, Springer, Berlin, 2004.

[6] Thomas Karagiannis, Konstantina Papagiannaki, and Michalis Faloutsos, BLINC: Multilevel Traffic Classification in the Dark, in SIGCOMM'05, August 21–26, 2005, Philadelphia, Pennsylvania, USA.

[7] S. Zander, T. Nguyen, and G. Armitage, "Automated traffic classification and application identification using machine learning," in IEEE 30th Conference on Local Computer Networks (LCN 2005), Sydney, Australia, November 2005.

[8] . L. Bernaille, I. Akodkenou, R. Teixeira, A. Soule, and K. Salamatian, Traffic classification on the fly, SIGCOMM Comput. Commun. Rev., vol. 36, pp. 2326, Apr. 2006.

[9] Bonfiglio, D., Mellia, M., Meo, M., Rossi, D., Tofanelli, P., 2007. Revealing Skype traffic: when randomness plays with you. SIGCOMM Comput. Commun. Rev. 37 (4), 37–48.

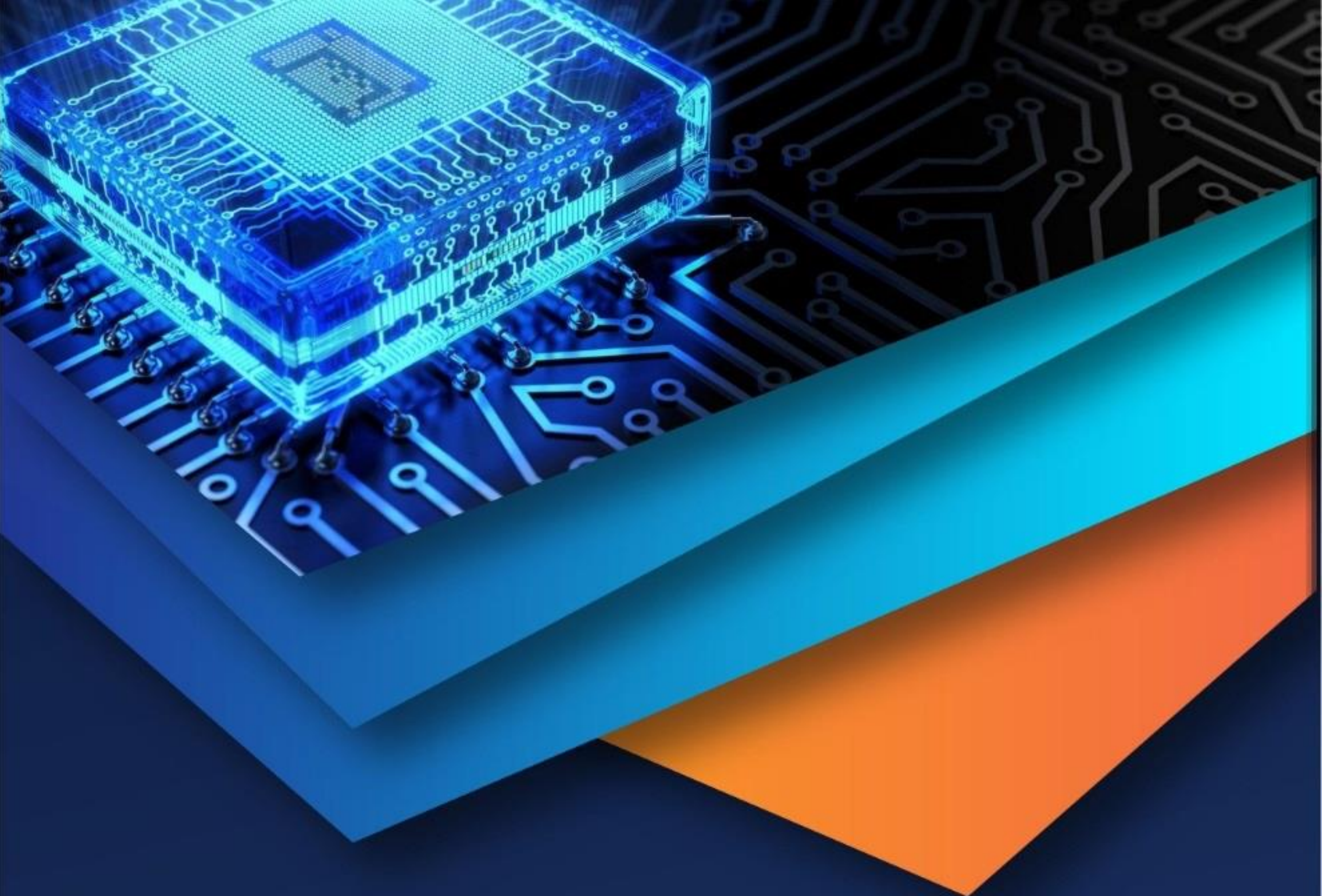
[10] T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimize the use of Machine Learning classifiers in real-world IP networks," in Proc. IEEE 31st Conference on Local Computer Networks, Tampa, Florida, USA, November 2006

[11] T.Bakhshi and B.Ghita, On Internet traffic Classification: A Two-Phased Machine Learning Approach, Journal of Computer Networks and Communications, vol.2016

[12] Valentin carela-Espanol, Pere Barlet-Ros, Oriol Mula-Valls, Josep Sole- Pareta, An Automatic Traffic Classification System for network operation and Management, Springer, October 2013

[13] J. Erman, A. Mahanti, and M. Arlitt, "Internet traffic identification using machine learning techniques," in Proc. of 49th IEEE Global Telecommunications Conference (GLOBECOM 2006), San Francisco, USA, December 2006.
[14] Alshammari, R., Zincir-Heywood, A.N., 2011. Can encrypted traffic be identified without port numbers, IP addresses, and payload inspection? Comput. Networks 55 (6), 1326–1350
[15] J. Zhang, X. Chen, Y. Xiang, W. Zhou, and J. Wu, Robust network traffic classification, IEEE/ACM Transactions on Networking vol. 23, no. 4, pp. 12571270, 2015
[16] Tcpdump, http://www.tcpdump.org, Accessed 13 February 2009.




10.22214/IJRASET

INDEX COPERNICUS 45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429


# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY