Taxonomy of Machine Learning Algorithms to classify real-time Interactive applications

Hamza Awad Hamza Ibrahim
Universiti Teknologi Malaysia-UTM
Faculty of Electrical Engineering

Sulaiman Mohd Nor
Universiti Teknologi Malaysia-UTM
Faculty of Electrical Engineering

Aliyu Mohammed
Universiti Teknologi Malaysia-UTM
Faculty of Electrical Engineering

Abuagla Babiker Mohammed
Universiti Teknologi Malaysia-UTM
Faculty of Electrical Engineering

***Abstract*— the needs of Internet applications QoS guarantee increased the demand of internet traffic classification, especially for interactive real time applications. Therefore, several classification methods were developed. Machine Learning (ML) classification is one of the most modern techniques, which solves the problem of traditional port base method. This paper compared experimentally the accuracy of ten ML algorithms, that when it's used to classify interactive applications. The technique applied by collecting of real data from UTM. The result shows that *Tree.RandomForest* algorithm provided optimal results of 99.8% accuracy, compared with other algorithms.**

Keywords-Classification; Mashine Learning; Interactive application;

I. Introduction

One of the most popular Internet applications is interactive applications, which are consisting of large packets actively initiated by users.[1]. This includes (but not limited to), online conference, online lectures, Skype, MSN, online game, online TV or any type of online communication. When observing these applications we found that it needs specific treatment because of the following reasons:

- Interactive application is more sensitive than others
- Users that work in real time suffer from delay, data lose, and retransmission
- The QoS for interactive applications depend on the rapid response[2]
- Interactive applications, like online meetings, online lectures are very important to process [3]

Since the future of Internet depends largely on real time application, the motivation is to improve service quality for these applications. Interactive real-time multimedia applications such as Video over IP, Voice over IP, web learning among others, require guaranteed and predictable QoS. Moreover, several interactive applications need continuous user interaction. Furthermore, there are no strict delay requirements for interactive traffic because these types of packets do not likely provide for data loses. For all that interactive application needs specific methodology in classification and prioritization so as to reach acceptable level of applications QoS.

To date, several VOIP applications were in existence, such as Skype, yahoo Messenger, MSN, Google Talk, iCall, etc. VOIP gain succeed significantly because of several reasons; 1) VOIP includes different useful services such as voice call, video call, SMS, share screen, etc. 2) VOIP is more cheaper than public Switched Telephone Networks (PSTNs) [4] 3) Pay credit as you go 4) There are enhancement in voice and video quality. 5) The ability to transmit more than one telephone call over a single broadband connection. With the rapid increase in Internet usage, the bandwidth not sufficient to satisfy QoS needed by Internet applications.[5]. Thus, the demand for Internet traffic classification increases day by day.

Although ISPs have used various approaches to satisfy the needed QoS, nevertheless the rapid increase in Internet users coupled with the need for real time interactive applications have caused this task to be even more challenging. In a campus environment, for example, the need to use on-line video for teaching and learning has become popular. As general Internet traffic classification has reaps several benefits such as:

- Internet traffic classification answered the operator's question, what types of packets are swimming in ours networks?
- The first step for ISPs to manage network traffic is to provide accurate and fast mechanism for applications classification.[6]
- Traffic classification is essential issue to applying strong network security mechanism.
- To improve interactive applications QoS,

Simple classification assumes that, most applications used well-known Port, and the classifier used this port number to identify the application type. However most Internet applications used unknown port number or more than one application used the same port number, which indicates to failure of port base classification[7]. Another classification method is payload base (deep packet inspection), which is individual packet inspection, looking for unique signatures. However, using of this technique faced by two problems; first, it impossible to detect non-standard port by using packet inspection, because these packets were encrypted. Second, deep packet inspection touches users' privacy. Other classifications works used mixed techniques [8], and [9]] , which are evaluated based on the work itself not based on mechanisms that are applied. In particular, to classify interactive applications, rapid and accurate mechanism can be provided. In order of that, several methods and techniques are proposed. This research paper compares ten ML algorithms, that when used to distinguish between two of popular real time applications which are online TV and Skype. The aim of that is to go forward of building a method for real time interactive application classifications.

VOIP traffic classification appears as a big challenge because of the following; i) most (if not all) of VOIP applications are P2P ii) no well-known port iii) encrypted packets payload iv) with the presence of high speed of Internet data transmission (1-10Gbps), the rapid classification for real time application attracts attention[10]. In this research paper, ML algorithms comparison was obtained fot the purpose of forwarding progress to build a mechanism of interactive applications classification.

The remaining of this paper is organized as following; section two talks about ML in traffic classification, section three outlining of five modern related works, section four explains our experimental work stages, whereas section five discusses the network capturing environment, section six discusses by graph the paper results, and finally section seven concludes this work by results summary and future work.

II. ML CLASSIFICATION

One of the modern application classification techniques is Machine Learning (ML), which use Artificial Intelligence to classify IP traffic. In order to solve the problem of past classification methods (base port and payload inspection), ML provide a great solution which is extracting real information from application features.[11]. Moreover, some of ML algorithms are suitable for Internet traffic flow classification at a high speed.[5] .

ML technique is performed in several steps; firstly, selection of dataset which are some of features values. These features are attributes of traffic flow, such as, packet length, inter arrival time, protocol, idle time, etc. Secondly, applied training for ML to establish classification rules. That based on statistical computation extracting from the features. Lastly, apply ML classification for unknown packets using training rules. Because of the rapid nature of real time applications, the important issue when classifying interactive applications is the time of collecting the statistical values (build rules), which is assumed to be very short. ML consist of a different algorithms, which are categorized into two main types, i) supervised learning, its inferring a function (rules) from labeled training data. The classifier rules should predict the correct output value for any new valid input object. ii) Unsupervised learning, its classifier method trying to find hidden structure in unlabeled data. There are different approaches in unsupervised ML such as, clustering (K-mean, mixture model, hierarchical clustering), blind separation (Principal component analysis, Independent component analysis)

When comparing ML works, several approaches can be considered such as, features, algorithms, value of training data, value of test data, scope of data collection, and accuracy. ML classifiers can be evaluated by different metrics, which are illustrated as following:

- True positive (TP): Percentage of instances of specific class correctly classified as belonging to this specific class
- True Negative (TN):): Percentage of instances of other classes correctly classified as not belonging to these classes
- False Positive (FP):): Percentage of instances of other classes incorrectly classified as belonging to specific class
- False Negative (FN): Percentage of instances of specific classes incorrectly classified as belonging to other classes
- Recall , $\mathrm{Recall} = \left(\frac{TP}{TP+FN} * 100\right) \%$, The recall assumed to be equal to TP (in perfect accuracy)
- Precision, $\mathrm{precision} = \left(\frac{TP}{TP+ FP} * 100\right) \%$, The precision assumed to be equal to TP (in perfect accuracy)
- Byte accuracy, which is define as, how many bytes packets are carried by the correctly classified flows. Byte accuracy is important in the traffic classification.[12]
- The techniques assumed to be, high number of flows and each flow include small size of byte.

III. RELATED WORKS

Several research papers in different mechanism have used ML classifier. [13] , compare between a five ML algorithms (MLP, RBF, C 4.5, Bayes Net and Naïve Bayes). The authors develop real time internet traffic dataset to classify seven applications, which are www, e-mail, web media, P2P, FTP data, instant messaging and VoIP. The work used Wireshark as capturing tool. Several features such as minimum, maximum, mean no. of packets, average packets per second, packet size, and duration, were obtained, and considered in two ways dataset (full feature dataset and reduced feature dataset). The result shows that, in case of full features dataset, Bayes Net classifier provides the better accuracy which is 85.33 %, and the approach of reduced features, C4.5 provides the higher accuracy which is 93.66%. The similarity between this paper and our work is the analysis and comparison, however, the first result in this paper shows high training time (14 seconds), furthermore, capturing duration is short (2 minutes).

In [9], the authors proposed re-sampling methods for network traffic classification, and consider by comparing three types of samples (data), stratified sampling, uniform sampling, and tuning sampling. The dataset was divided into different classes such as WWW, MAIL, BULK, ATTACK, CHAT, P2P, MULTIMEDIA, VOIP, and INTERACTIVE GAMES, and each class include some of Internet applications. The introduced methods is tuning sampling to maintain the accuracy, in other words, re-sampling of training data to decrease the data skew. The goodness of this work is the high number of applications and features was considered, but this methodology is very difficult to use as online classification.

In another related work [14], semi-supervised methodology was used. The authors developed method using statistical values from labeled and unlabeled flows, that can be summarized into three step 1) the input will be labeled and unlabeled data 2) then, dividing whole data into clusters by using K-means algorithm (training step), 3) then, label the clusters. Three types of dataset are considered; KDD Cup 99, IRIS, and GLASS, also, connection and symbolic features are selected. The proposed method was tested by classifying five applications Normal, Probe, Denial of Service, User to root, and Remote to Local. The accuracy result was compared with SVM to show preference with KDD (90.65%) and IRIS (96.66%). However, the data's selected are not real data; also they are very strange type of classes.

The authors of [4] provide work aims to distinguish between Gtalk/Skype traffic traces and other traffic traces. The work conducted more than 100 experiments of more than 25 hours VOIP capturing, which is equal to 6 GB data. The authors carried out different scenarios, such as, switching between using firewall in some of computers, using wireless and wire-line, blocking of all UDP connections, and blocking of all TCP connections. Furthermore, the firewall was adapted by different ways such as permit access, or do not permit anything, or limit permit port number. The authors generate different traffics such as Yahoo messenger traffic (encrypted with Zfone), Primus VoIP, Zfone (encrypted), primus Session Initiation Protocol (SIP) (non-encrypted), also from the university traces the work generate DNS, FTP, SSH, MAIL, HTTP, HTTPS and MSN traffic. Two metrics was used to evaluate the result, Detection Rate (DR)(means Recall) and False Positive Rate (FP). 3 algorithms within Weka, C4.5, AdaBoost and Genetic Programming (GP) are applied. The result shows that C4.5 5 provides more than 99% of DR and less than 1% FP. The only weakness here is that the work considers only TCP and UDP, and ignored the other protocols.

[15] is very close to our paper, which is ML work that explored Genetic Algorithm (GA) based on Random Forest (RF) to identify Skype traffic. The process starts by capturing packets, and aggregate them into traffic flows according to the 5 tuples. Then, calculate statistics flow features (inter-arrival of packets, and packet length). Finally, three ML algorithms were considered to classify Skype and non-Skype, Random Frost, C4.5, and SVM. The classification considers about 15 applications including Skype flows. The results show that RF provides the better accuracy 96%. However, the time taken to build the model is too long (34.14 seconds), which is a bad indicator for real time classification.

IV. Experimental work

In order to find interactive applications classification mechanism, ten of ML algorithms were used to classify OnlineTV and Skype. Another objective is to build a method that is valid for online classification. Figure 1 below shows experimental process steps, which start by using Wireshark to captures only two interactive applications (OnlineTV and Skype), then, the two application files will be edited by Excel to remove unneeded parameters. All samples were collected from UTM (CICT and Perdana collage) for the purpose of network traffic classification.

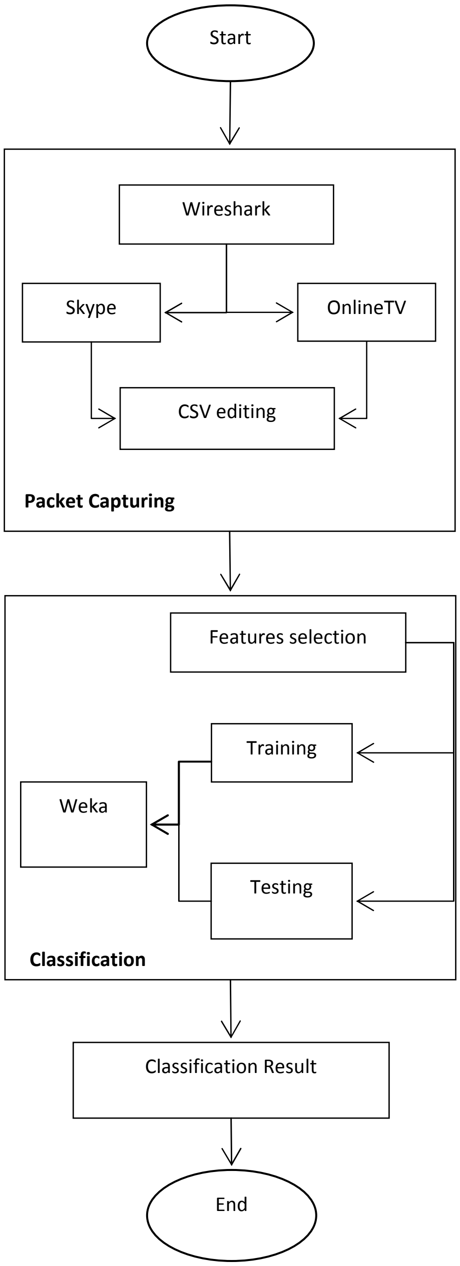

Figure1. Experimental process steps

In the first step of classification, Packet interarrival time and packet length were selected as features for machine learning classifier. The advantage of selecting only two features is reducing of classification complexity, particularly we deals with real time application. Each capturing file was divided into two parts. The upper portions (T1 and S1) of the files were used as training dataset and the lower portions (T2 and S2) of the same files were used as testing dataset.

The total number of training instances are19, 995 packets, which include the upper portion of Skype capturing files (S1) and the upper portion of OnlineTV capturing file (T1). 2,002 packets (S2 + T2) were used as testing dataset. In the same manner, the total numbers of testing are 2002 packets, which include the lower portion of Skype file (S2) and the lower portion of OnlineTV files. Three benefits were gained when collecting data by this way, first, ensuring of similarity between training and testing data, second; reducing of classification errors rate (this type of error discussed in [7]), third; easy in data processing.

V. NETWORK CAPTURING ENVIRONMENT

The network traffic composed of several application traces, therefore pre-capturing was performed. Two capturing devices was provided, the first device equipped with Intel core (TM) i3-2330M CPU 2.2 GHz 2 core and memory 6.00 GB, and the second device is Server of 4 CPUs, Intel(R) XEON(TM) 2.00 GHz . Our aim is to study and analysis interactive applications. To do so, we capture only two interactive applications, Skype and OnlineTV packets, that is applying Pre-capturing mechanism.

VI. RESULT AND ANALYZING

To classify two of most popular real time applications (OnlineTV and Skype), ten algorithms within Weka Explorer were used. These algorithms were applied to both training and the testing, which are ZeroR, PART, DecisionStump, J48, J48graft, LADTree, NBTree, RandomForest, RandomTree, and REPTree. Figure 2 and 3 below illustrates training and testing accuracy results and time taken to build models. It's clear that Tree.RandomForest algorithm provided optimal results of 99.8% accuracy, compared with other algorithms. In addition to that, DecisionStump provides the shortest time (0.05 Seconds) when compared with other algorithms. Because we were dealing with interactive applications, time to build a model is very important factor. Figure 4 below analysis the ML metrics True Positive, False Positive, Precision, and Recall, which are used to evaluate algorithms. Likewise, Tree.RandomForest obtained very high values in TP, Precision, and Recall.

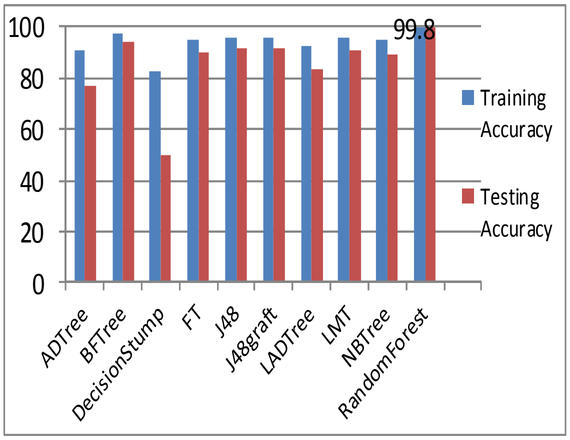

Figure 2. ML algorithms accuracy

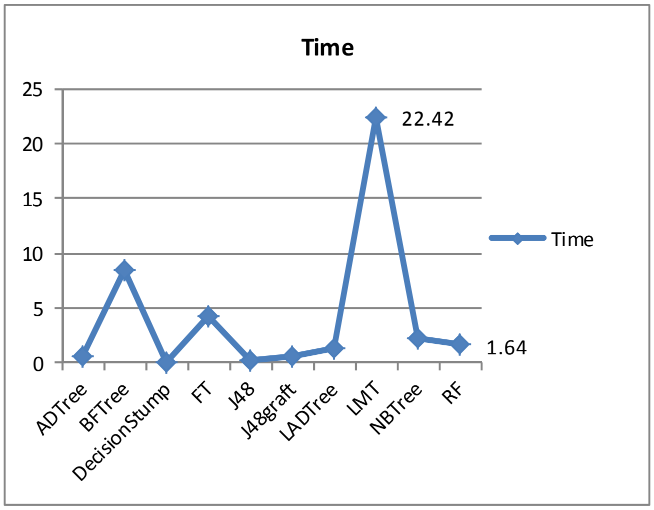

Figure 3. ML algorithms duration

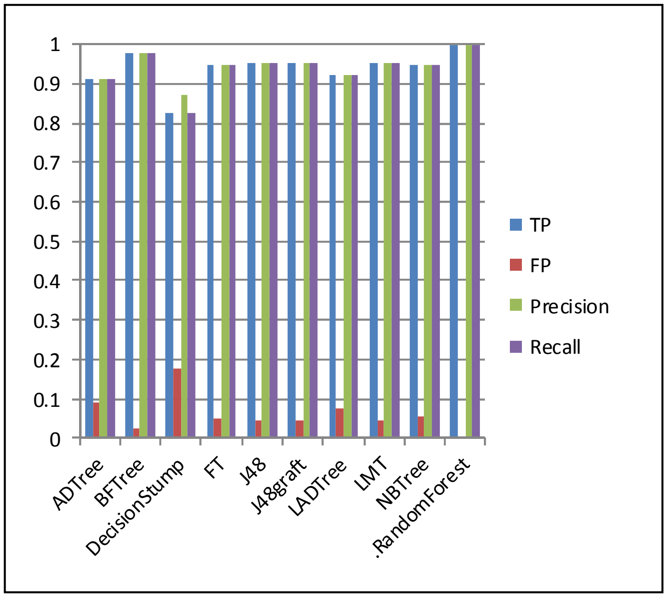

Figure 4. ML algorithms metrics values

VII. CONCLUSION AND FUTURE WORK

The classification of interactive real time applications was sensitive to detection duration. In this paper we compared between ten ML algorithms to distinguish between two of real time internet application, Skype and OnlineTV. Real dataset was collected from campus environment to give inputs to classifier. The comparison result shows, Tree.RandomForest algorithm gave high accuracy result of 99.8%. Moreover, and from classification time point of view, the work has acquired that some of ML algorithms were suitable to classify interactive real time applications.

In future work, we intend to focus on the interactive applications classification with the view to increase the scope of the applications and number of features. It is also our desire to build a very short time online classification for real time applications.

References

1. Dimopoulos, P., P. Zeephongsekul, and ZahirTari, *A dynamic priority approach to reducing delay in interactive TCP connections.* Telecommun Syst, 2007).
2. Mehrotra, S., J. Li, and Y.Z. Huang, *Optimizing FEC Transmission Strategy for Minimizing Delay in Lossless Sequential Streaming.* Ieee Transactions on Multimedia, 2011. **13**(5): p. 1066-1076.
3. Elshaikh, M.A., et al., *A new fair marker algorithm for DiffServ networks.* Computer Communications, 2008. **31**(14): p. 3064-3070.
4. Alshammari, R. and A.N. Zincir-Heywood. *An investigation on the identification of VoIP traffic: Case study on Gtalk and Skype*. in *Network and Service Management (CNSM), 2010 International Conference on*. 2010.
5. Soysal, M. and E.G. Schmidt, *Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison.* Performance Evaluation, 2010. **67**(6): p. 451-467.
6. Hays, C.B., *Bypassing internet service provider traffic shaping with peer-to-peer file sharing through deliberate false positives.* Iet Communications, 2011. **5**(11): p. 1540-1543.
7. Nguyen, T.T.T. and G. Armitage, *A Survey of Techniques for Internet Traffic Classification using Machine Learning.* Ieee Communications Surveys and Tutorials, 2008. **10**(4): p. 56-76.
8. Wang, Y., Y. Xiang, and S. Yu, *Internet Traffic Classification Using Machine Learning: A Token-based Approach.* Computational Science and Engineering (CSE), 2011 IEEE 14th International Conference on, 2011.
9. Ruoyu, W., L. Zhen, and Z. Ling, *A New Re-sampling Method for Network Traffic Classification Using SML.* 2010.
10. Santiago del Rio, P.M., et al. *On the processing time for detection of Skype traffic*. in *Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International*. 2011.
11. Yu, J., et al., *Real-time Classification of Internet Application Traffic using a Hierarchical Multi-class SVM.* Ksii Transactions on Internet and Information Systems, 2010. **4**(5): p. 859-876.
12. Erman, J., A. Mahanti, and M. Arlit, *Byte Me: A Case for Byte Accuracy in Traffic Classification.* 2007.
13. Singh, K. and S. Agrawal, *Comparative Analysis of Five Machine Learning Algorithms for IP Traffic Classification* 2011.
14. Shrivastav, A. and A. Tiwari. *Network Traffic Classification Using Semi-Supervised Approach*. in *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*. 2010.
15. Li, J., et al., *Identifying Skype Traffic by Random Forest.* 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Vols 1-15, 2007: p. 2841-2844.