International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 21 Issue 2 – APRIL 2016.

HYBRID ALGORITHM FOR THE CLASSIFICATION OF TRAFFIC FLOWS

T.Kalaiselvi, P.Shanmugaraja

PG scholar, Sona College of Technology, Department of Information Technology
Associate Professor, Sona College of Technology, Department of Information Technology

***Abstract*--For controlling and managing the network related tasks traffic classification is done. Traffic classification is done by using port-based, payload-based and machine learning techniques. Since the port-based and payload-based has several drawbacks due to dynamic port assignment and encryption techniques, machine learning techniques are preferred. Two types of Machine learning techniques are used for internet traffic classification they are supervised and unsupervised learning algorithm. Traffic can be classified using machine learning techniques by analyzing the statistics of the flows. C4.5 is an supervised algorithm which discretizes the features on its own. K-means is an unsupervised algorithm that provides best accuracy among all the algorithms used in unsupervised.In order to obtain the advantages of both the algorithms, a hybrid algorithm is introduced (combination of C4.5 and K-means).As the results show, the hybrid algorithm performs well in terms of accuracy for internet traffic classification.**

Key Words: Internet Traffic Classification, Machine Learning, K-means, C4.5

I. INTRODUCTION

Traffic classification is an important task for classifying the traffic available in the network. Traffic classification is the fundamental functionality for providing network management and security related tasks. Traditional traffic classification methods like port-based and payload-based approach has several drawbacks due to the advent of dynamic port assignment and some of the traffic encryption techniques. Nowadays several machine learning methods are available for classifying the traffic. For identifying and classifying the traffic using machine learning techniques several features like packet size, flow duration, arrival time of the packet, etc., are used.

Traffic classification process is divided into two stages. They are training and classification. In the training stage dataset is given as an input to the algorithm. Output of this stage will be a probabilistic model or a tree structure. In the classification stage application type (i.e.) real time or non-real time is predicted by using classifier. Real time packets are sensitive to delay whereas non-real time packets are insensitive to delay.

Supervised learning and unsupervised learning are the two types of machine learning techniques used to classify the internet traffic. This classification is based on whether the training data set is provided or not. Supervised learning classification deals with the labelled data. Whereas unsupervised learning (clustering) divides the unlabeled data based on the similarity among the data. Practically clustering approach is easy because clusters are formed from unlabeled data, so there is not necessary for acquiring labelled data, but clusteringapproach is time consuming. Clustering algorithms for classifying the internet traffic is K-means, Expectation Maximization (EM), DBSCAN, etc., Some of the commonly used supervised machine learning algorithms are Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, C4.5 decision tree, etc.,

Port number and size of the packets are an important features for classifying the traffic. Feature set is important and necessary for improving the accuracy of traffic classification[1]. Features are statistical characteristics obtained from packet length, flow length, etc., Some of the features are obtained only after the completion of the flow. Most of the machine learning algorithms are trained and testing by using full flows. Whereas in real networks we have to make quick decisions before the end of the flow so, in that case only first few packets are investigated. However, investigating only first few packets is not sufficient because malicious attacks may come without any signature during earlier packet flows. Feature discretization can enhance the accuracy of the internet traffic classification. Discretization is the process of transferring the continuous function into discrete counterparts. Supervised algorithm C4.5 performs well in any situation because during the process of classification the features are discretized without any changes in the algorithm we can achieve greater than 93% of accuracy due to discretization of features [2].

The contribution of this paper is as follows:

- In order to overcome the problem of class imbalance and concept drift (problem due to classifying only majority of classes and dynamic changes in flow) some of the features are selected based on the study. The selected features are port number at server side, minimum segment size, total number of bytes sent in initial window.
- We present a hybrid algorithm for improving the accuracy of the traffic classification.

This paper is organized as related work in section 2. Section 3 presents traffic flow features. Section 4 presents the classification algorithms used for traffic classification. Section 5 presents our proposed hybrid algorithm and methodology. Classification results are given in section 6. In section 7 the paper is concluded.

II. RELATED WORK

In machine learning based traffic classification there are three main challenges to consider. They are identifying the

feature, identifying the best algorithm for classification and obtaining the training datasets.

In [3] an unsupervised learning approach Expectation Maximization for classifying the internet traffic which are represented by a set of flow attribute such as size of the packet, duration of connection, count of bytes and inter-arrival time.

K-means online approach for classifying the traffic by using only first five packets of the flow is presented in [4]. Better results were obtained with the first five packets that is more than 80%. But this result can't be achieved if any of the packet from the first five packets are missed.

Some classifiers like Bayesian network, Naïve Bayes, C4.5, Naïve Bayes tree are compared in [5]. From their analysis they have found that C4.5 algorithm performs well in terms of accuracy than other algorithms.

An approach for classifying the traffic into different type of services is presented in [6]. Here along with the kernel estimation, a Naïve Bayes classifier is combined for classifying offline TCP traces. It achieves 96% accuracy. Here features are selected using correlation based filter algorithm.

Accuracy of decision tree algorithms like C4.5 and random forest are always stable than other algorithms. They also stated that build time of the real time features are low. So decision tree algorithms are also well suited for online traffic classification [7].

III. TRAFFIC FLOW FEATURES

IP flows are the basic units in internet traffic classification. IP flows are bidirectional packets send in every direction can be found by IP address of source, IP address of destination, source port number, destination port number, protocol of the transport layer. For classifying the traffic feature value has to be calculated from the IP flows. Selecting the feature is important because if large feature set is used then there is possibility for computational overhead so for avoiding this simple features has to be preferred.

For selecting the best feature several feature selection algorithms are like Filter model and Wrapper model are used. In Filter model for identifying the importance and the relevance of the features, training data characteristics is used. Wrapper model uses the results of the classifier by using different combination of features. In internet traffic the applications are categorized as bulk transfer, streaming, instant messaging, peer to peer, web browsing [8].

From the available 248 features [9] some most important features are identified in order to avoid the class imbalance and concept drift problem. The selected features performs well in terms of byte accuracy. Identified features are listed in the table 1.

TABLE 1

FLOW FEATURE SET

Feature	***Description***
Server port	Port number at server
Minimum segment size	Minimum data transfer during the connection (client→ server)
Initial window bytes	Total number of bytes sent in initial window (client→ server and server→ client)

IV. DATASETS

The packet traces are collected using Wireshark tool. Full packet payloads are preserved without any loss. Categories of application are bulk transfer, mail, streaming, web browser, instant messaging, peer to peer. The categories of application are listed in table 2.

TABLE 2

CATEGORIES OF APPLICATION

Category	***Application***
Bulk	FTP
Mail	POP3,SMTP,IMAP
Web	http, https
Chat	Yahoo messenger, MSN, Talk
P2P	BitTorrent, Gnutella, Napster
Streaming	Youtube, Quicktime, Real

V. CLASSIFICATION ALGORITHMS

A. K-means

K-means is an unsupervised clustering method. During the training phase this K-means will generate K random cluster. Input to this K-means algorithm is flow features and number of cluster. During the classification phase the generated random clusters are classified. Here the clusters are formed based on the similarity among the data. This similarity is identified by calculating the Euclidian distance.

Major advantage of K-means algorithm is if the K value (i.e.) number of cluster is small then the computational speed will be faster and overall accuracy of algorithm is good.

B. C4.5

C4.5 is a decision tree based machine learning algorithm. C4.5 algorithm constructs decision tree from the training data traces by using information entropy concept. Decision tree is constructed based on normalized information gain of features.

VI. HYBRID ALGORITHM

In this section hybrid algorithm which is a combination of supervised C4.5 algorithm and unsupervised K-means algorithm. The hybrid algorithm is fed with small amount of labelled data and large amount of unlabeled data. Our hybrid algorithm comes under semi-supervised approach as it works with the unlabeled and labelled data. Advantage of this hybrid algorithm is some samples which cannot be grouped along with their respective clusters can be grouped based on the some of the labelled information available.

A. *Methodology*

Figure 1 shows how the classification of internet traffic is done. In this research network traces are collected using packet capturing tool. Collected packet traces contain several types of application. After collecting the traces, features are selected in order to avoid class imbalance and concept drift problem the selected features are server port, minimum segment size and initial window byte by using Weighted Symmetrical Uncertainty algorithm. After analyzing the feature it is applied to hybrid algorithm. Analysis is done on the results of traffic classification on the basis of accuracy.

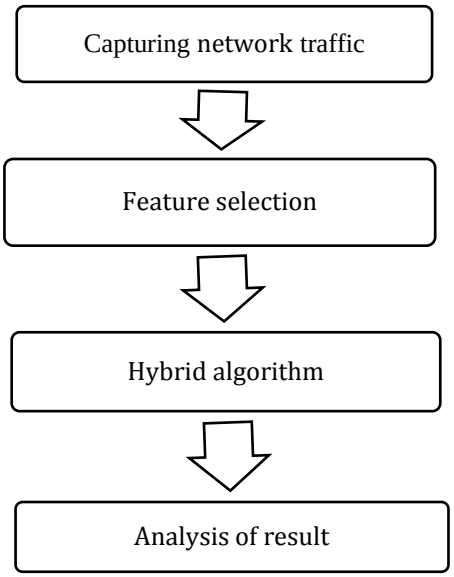
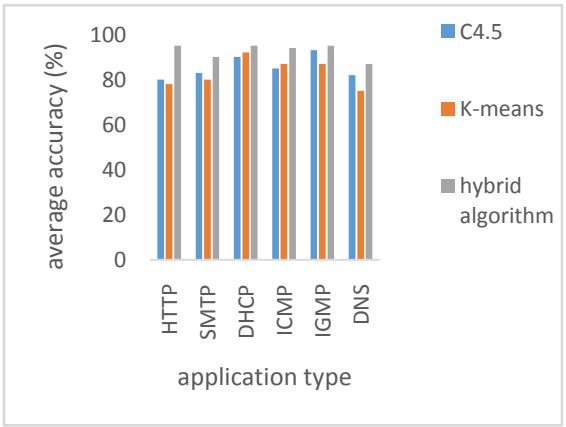


Figure 1 Methodology

VII. CLASSIFICATION RESULT

For evaluating the performance of the machine learning algorithms overall accuracy is used.



A ccuracy: It is the ratio of number of traffic flows classified correctly to the total number of flows classified correctly to the total number of flows available in the dataset. The number of traffic flows classified correctly is known as True Positive (TP). It is given as

$$\text{Accuracy}=\sum\frac{\text{TP for every cluster}}{\text{total number of flows available}}$$

Accuracy of different application types are shown in the figure (2). The figure(2) shows the performance gap between C4.5 algorithm, K-means and hybrid algorithm.From figure (2) we can state that accuracy of hybrid algorithm have higher accuracy for all application types.

Figure (3) shows the accuracy of different algorithms like C4.5, K-means and algorithm in accordance with number of clusters.

Figure 2 Comparison of accuracy in accordance with application type

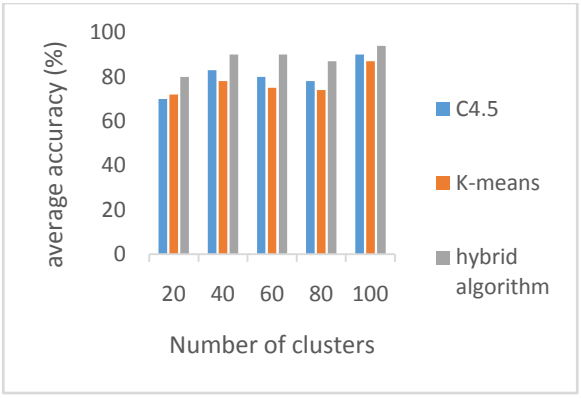


Figure 3 Comparison of accuracy in accordance with number of clusters

VIII. CONCLUSION

In this work for packet classification a new hybrid algorithm is presented. For improving the performance the features server port, minimum segment size and initial window bytes are used. As the result shows, the presented hybrid algorithm can able to achieve good accuracy when it is used with features like server port, minimum segment size and initial window bytes.

REFERENCES

[1] Nguyen T T T, Armitage G.” Training on multiple sub-flows to optimizethe use of machine learning classifiers in real-world IP networks” IEEELCN, 2006, pp. 369–376.
[2] Y. Lim, H. Kim, J. Jeong, C. Kim, T. Kwon, and Y. Choi, ‘‘Internet Traffic Classification Demystified: On the Sources of the Discriminative Power‘‘, Proc. ACM CoNEXT, 2010, p. 9.

[3] McGregor, M. Hall, P. Lorier, and J. Brunskill. Flow clustering using machine learning techniques. *LectureNotes in Computer Science*, 2004, pp.205–214.
[4] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian. “Traffic classification on the fly”*SIGCOMMComput. Commun.* 2006.
[5] Williams N., Zander S., Armitage G., “A Prelimenary Performanc Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Comparison”, ACM SIGCOMM Computer Communication Review, Vol. 36, No. 5, 2006, pp. 5-16.
[6] Andrew W. Moore and Denis Zuev. Internet traffic classification using bayesian analysis techniques. In *Proceedings of ACM Sigmetrics*,2005, pp.50-60.

[7] ZHAO Jing-jing, HUANG Xiao-hong, SUN Qiong, MA Yan, "Real-time feature selection in traffic classification", ELSEVIER, 2008.
[8] Y. Lim, H. Kim, J. Jeong, C. Kim, T. Kwon, and Y. Choi, ''Internet Traffic Classification Demystified: On the Sources of the Discriminative Power,'' Proc. ACM CoNEXT, 2010, p. 9.
Traffic Classification Using Machine Learning"
[9] Shijun Huang Kai Chen Chao Liu Alei Liang Haibing Guan "A Statistical-Feature-Based Approach to Internet