Classification of application based on multiple cluster models with selected attributes from Flow statistics

Anil Kumar
Computer Science & Information Systems
Texas A & M Commerce
Commerce, Texas
Email:akatta1@leomail.tamuc.edu

Dr.Jinoh Kim
Assitant Professor
Computer Science & Information Systems
Texas A & M Commerce
Commerce, Texas
Email:Jinoh.Kim@tamuc.edu

***Abstract*—Identifying applications are critical for a broad range of network related activities like bandwidth usage, security etc. Earlier, applications are identified based on port numbers, which proved to be not accurate anymore; based of payload signatures, which is proved to be accurate but has been limited in the real world implementation because of privacy concerns; based on flow statistics, which uses machine learning algorithms to find the patterns in the flow statistics and use it in classification, which has been widely used for many classification problems. In this research, we explore the importance of the attributes or a combination of flow attributes which can classify applications effectively. The idea is to combine clustering and using combinations of flow attributes and we measure accuracy of each combination. We are currently evaluating our model with real-world traffic traces indicating effectiveness of the selective attributes is effective than using the whole set of attributes.**

I. Introduction

Recently, there has been a lot of emphasis on using flow statistics with combination of Machine learning algorithms to determine network based applications. Accurately identifying network based application is of major interest for ISP. Which lets providing Quality based Service(limiting usage of bandwidth by unnecessary applications).

In this paper, we proposed a new technique which studies the relevance of flow statistics or combination of them to give us better classification. We compared our technique with previously published work. As with most machine learning algorithms, more attributes(not very important) may actually hurt the overall accuracy of the classification. So, determining the best combination of attributes and also combination of them helps in actually improving overall accuracy. We developed our own technique, where we used multiple trained models to classifying the incoming flow.

II. Related Work and Motivation

A. Related Study

Existing techniques based on port-numbers are proved to be very ineffective where accuracy is less than 70 [], as number of applications using random port numbers or non-standard port numbers are increasing day-by-day and also usage of tunneling makes identification of application more difficult just based on port numbers.

To counter drawbacks of port-based application identification, many payload-based techniques have been proposed[]. In payload-based technique, payload of the flow will be extracted and searched for known signature of the applications. Results indicate that this method is very effective with high accuracies. Drawbacks include encrypted traffic, and wide deployment of tools based on payload signature is a problem as many countries doesn't allow the extraction of full payload from packets due to privacy concerns

Drawbacks of both payload-based signature matching and port-based classification led us to use the transport layer characteristics of the application as the differentiator of the application. From various proposed techniques[] we can see combination of transport layer characteristics in combination with Machine learning techniques accurately identifies the applications, with accuracies comparable to that of payload-based signature matching.

Classification of network application by using machine learning algorithms is broadly divided into three categories.
Machine Learning Technqiues

Un-supervised
: Labeling information of the training data is not available at the time of training. We use various clustering algorithms for the classification of unlabeled data[].

Supervised
: We provide the labels for the flows when we train the model and then use this model to test each incoming flow whether it belongs to any of the application which is provided at the time of training[].

Semi-supervised
: We provide partial labeling information at the time of training and we use clustering algorithms to cluster the training data. We use the partially available labeling information to label each cluster. Heuristics have been proposed on how to label the cluster from partial training information[]

B. Evaluation of existing techniques

We extensively tested the currently existing techniques[].

III. Dataset and Cleaning

We used dataset provided by Sysmate. Dataset is 100GB of captured data with full payload. Dataset has been captured for isolated applications. We considered 5 applications for our study i.e. Skype, Http, Edonkey, Bittorrent, Gnutella. Used python scapy library to extract packet information from pcap files to built flows. Considered L7 signatures**??** to remove any discrepancies in the data by matching regular expressions provided by L7 with the payload for protocols. Used python to generate flow statistics for flow.

IV. Importance of attributes in classification

We have considered a total of 18 flow statistics for our studies.

Average IAT	Maximum of IAT
Minimum of IAT	Stadard Deviation of IAT
Average RIAT	Maximum of RIAT
Minimum of RIAT	Standard Deviation of RIAT
Average Packet Size	Minimum of Packet Size
Maximum of Packet Size	Standard Deviation of Packet Size
Total Packet Size	Flow Duration
Number of Packets	Average Packets/second
Average Bytes/second	Payload Size

We observe from fig.1 that combination of Total packet size and Payload size leads to better accuracy along with combination of Number of Packets and Payload size. From fig.2 and fig.3, we observe the accuracy is very low. fig.4 we observe that combination of Average, Minimum and Maximum packet size and also combination of Average and Standard deviation of packet size leads to better accuracy. Observing the tables we see accuracy is not at its best when we use all the attributes than when we use subset of them.

V. Population Fraction

In this paper, we introduced new term called population fraction, which stands for percentage of dominant application (application with most number of flows in the cluster in consideration.) flows to total flows in a cluster

$$P_{clus} = (\frac{flows_{dominant}}{flows_{total}})_{clus} \quad (1)$$

VI. Multiple cluster models used for classification

We built four cluster models (used simplest K-Means clustering algorithm) to build the classifier, selecting 2 combinations of attributes from flow attributes, and two combinations of attributes from Packet size attributes. We haven't considered IAT and RIAT in building cluster models as any combination of attributes is not giving higher accuracy. By considering combinations of subset of attributes rather than whole set, as they did in [] we get better accuracy.

Our technique, considers results from all the models in determining the final classification result. We considered six different hypothesis in determining the final classification result from the results from four different models.

Considered hypothesis:

Random
: Select classification of any of the results from the four models as the final classification result

Greatest
: Select classification result from model with highest population fraction as the final classification result

Quorum
: Select the majority result from trained models. If we have 3 models resulted in one classification result and other resulted in other application then we consider result of 3 as the final classification result. In case of tie we select application randomly

Unanimous
: Select the final classification only when all the results from the four models exactly results in the same result. Otherwise mark it as unknown

Unanimous Greatest
: Similar to unanimous but fallback hypothesis is greatest.

Unanimous Random
: Similar to unanimous but fallback hypothesis is random

Unanimous Quorum
: Unanimous with fallback hypothesis as Quorum

In fig.5 Observing the results in the 5 we can observe that better accuracies are observed for greatest, unanimous greatest and quorum. Accuracy increases with number of clusters for all the hypotheses, but sharp rise in accuracy with number of clusters for unanimous greatest. Out of them we observe the best accuracy for unanimous greatest with 40 clusters.

VII. Results

We compared our results with supervised, basic K-Means clustering algorithms. Fig.6 shows the result of various algorithms plotted against percent of training data used. In the fig.6 K-Means selected shows the accuracy of basic K-Means algorithm(we considered 40 clusters for all unsupervised algorithms in this figure) plotted when only considered the combination of attributes which is giving best accuracy(we considered all the attributes which were giving better accuracy, but different combination of subsets in our technique). All the plots follow similar kind of pattern, where accuracy increasing along with percentage of training data. Our technique gives better accuracy when compared with simple K-Means clustering almost all the time and also comparatively better accuracy than Adaboost (supervised machine learning algorithm) at some points and at other points it gives around same accuracy as supervised.

Total Pkt Size	Flow Duration	Number of Packets	Avg Pkts Per Sec	Avg Bytes Per Sec	Payload Size	Accuracy
Used	Not Used	Used	Not Used	Used	Not Used	49.81%
Used	Not Used	Used	Not Used	Not Used	Used	87.25%
Used	Not Used	Not Used	Used	Used	Not Used	50.58%
Used	Not Used	Not Used	Used	Not Used	Used	88.83%
Used	Not Used	Not Used	Not Used	Used	Used	49.14%
Not Used	Used	Used	Used	Not Used	Not Used	47.74%
Not Used	Used	Used	Not Used	Used	Not Used	49.67%
Not Used	Used	Used	Not Used	Not Used	Used	92.02%
Not Used	Used	Not Used	Used	Used	Not Used	49.40%
Not Used	Used	Not Used	Used	Not Used	Used	89.36%
Not Used	Used	Not Used	Used	Not Used	Used	50.17%
Not Used	Not Used	Used	Used	Used	Not Used	49.45%
Not Used	Not Used	Used	Used	Not Used	Used	88.46%
Not Used	Not Used	Used	Not Used	Used	Used	51.83%
Not Used	Not Used	Not Used	Used	Used	Used	50.90%
Used	Used	Used	Used	Not Used	Not Used	91.86%
Used	Used	Used	Not Used	Used	Not Used	49.30%
Used	Used	Used	Not Used	Not Used	Used	91.47%
Used	Used	Not Used	Used	Used	Not Used	49.64%
Used	Used	Not Used	Used	Not Used	Used	85.54%
Used	Used	Not Used	Not Used	Used	Used	51.12%
Used	Not Used	Used	Used	Used	Not Used	49.76%
Used	Not Used	Used	Used	Not Used	Used	83.67%
Used	Not Used	Used	Not Used	Used	Used	50.40%
Used	Not Used	Not Used	Used	Used	Used	50.46%
Not Used	Used	Used	Used	Used	Not Used	50.33%
Not Used	Used	Used	Used	Not Used	Used	87.11%
Not Used	Used	Used	Not Used	Used	Used	51.16%
Not Used	Used	Not Used	Used	Used	Used	51.54%
Not Used	Not Used	Used	Used	Used	Used	51.16%
Used	Used	Used	Used	Used	Not Used	49.20%
Used	Used	Used	Used	Not Used	Used	80.33%
Used	Used	Used	Not Used	Used	Used	49.41%
Used	Used	Not Used	Used	Used	Used	50.08%
Used	Not Used	Used	Used	Used	Used	51.12%
Not Used	Used	Used	Used	Used	Used	51.92%
Used	Used	Used	Used	Used	Used	50.93%

Fig. 1. Accuracy results based on Flow Attributes

Avg IAT	Min IAT	Max IAT	Std Div IAT	Final Accuracy
Used	Not Used	Not Used	Not Used	47.29%
Not Used	Used	Not Used	Not Used	49.34%
Not Used	Not Used	Used	Not Used	39.20%
Not Used	Not Used	Not Used	Used	42.82%
Used	Used	Not Used	Not Used	45.86%
Used	Not Used	Used	Not Used	44.25%
Used	Not Used	Not Used	Used	47.78%
Not Used	Used	Used	Not Used	39.64%
Not Used	Used	Not Used	Used	42.91%
Not Used	Not Used	Used	Used	40.54%
Used	Used	Used	Not Used	45.20%
Used	Used	Not Used	Used	47.18%
Used	Not Used	Used	Used	42.70%
Not Used	Used	Used	Used	40.10%
Used	Used	Used	Used	42.75%

Fig. 2. Accuracy results based on IAT Attributes

VIII. Conclusion

By considering all the attributes of flow in classifying the application doesn't give better results. Selected attributes with combinations gives better accuracy. Accuracy doesn't continually increases along with number of clusters, as we get better accuracy when we are around 40 clusters.

Avg IAT	Max IAT	Std Div IAT	Final Accuracy
Used	Not Used	Not Used	55.86%
Not Used	Used	Not Used	54.08%
Not Used	Not Used	Used	53.41%
Used	Used	Not Used	55.21%
Used	Not Used	Used	54.28%
Not Used	Used	Used	55.25%
Used	Used	Used	56.41%

Fig. 3. Accuracy results based on RIAT Attributes

Avg Pkt Size	Min Pkt Size	Max Pkt Size	Std Div Pkt Size	Final Accuracy
Used	Not Used	Not Used	Not Used	92.43%
Not Used	Used	Not Used	Not Used	75.95%
Not Used	Not Used	Used	Not Used	92.22%
Not Used	Not Used	Not Used	Used	91.82%
Used	Used	Not Used	Not Used	91.46%
Used	Not Used	Used	Not Used	92.98%
Used	Not Used	Not Used	Used	93.96%
Not Used	Used	Used	Not Used	91.46%
Not Used	Used	Not Used	Used	92.69%
Not Used	Not Used	Used	Used	92.76%
Used	Used	Used	Not Used	93.98%
Used	Used	Not Used	Used	93.62%
Used	Not Used	Used	Used	92.76%
Not Used	Used	Used	Used	91.73%
Used	Used	Used	Used	93.75%

Fig. 4. Accuracy results based on RIAT Attributes

Considering IAT and RIAT for classification gives us the least accuracy when compared with other attributes of the flow. Even considering all attributes of packet size and flow also doesn't give better accuracy.

Population fraction will be very good parameter in classification of flow using clustering techniques.

ACKNOWLEDGMENT

acknowledging Sysmate.

Avg Pkt Size	Min Pkt Size	Max Pkt Size	Std Div Pkt Size	Final Accuracy
Used	Not Used	Not Used	Not Used	92.43%
Not Used	Used	Not Used	Not Used	75.95%
Not Used	Not Used	Used	Not Used	92.22%
Not Used	Not Used	Not Used	Used	91.82%
Used	Used	Not Used	Not Used	91.46%
Used	Not Used	Used	Not Used	92.98%
Used	Not Used	Not Used	Used	93.96%
Not Used	Used	Used	Not Used	91.46%
Not Used	Used	Not Used	Used	92.69%
Not Used	Not Used	Used	Used	92.76%
Used	Used	Used	Not Used	93.98%
Used	Used	Not Used	Used	93.62%
Used	Not Used	Used	Used	92.76%
Not Used	Used	Used	Used	91.73%
Used	Used	Used	Used	93.75%

Fig. 5. Number of Clusters v/s Accuracy

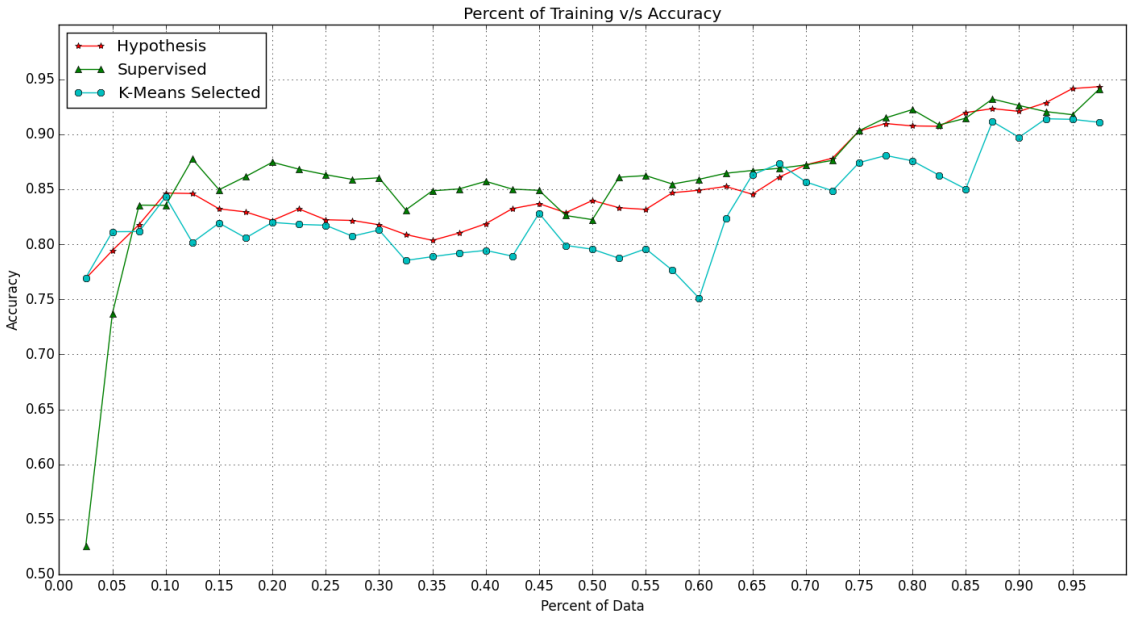

Fig. 6. Percent of training data v/s Accuracy