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.

- *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[]

TABLE I: Attributes used in out 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

B. Evaluation of existing techniques

We extensively ran existing techniques[] against our dataset. We ran supervised (Adaboost, Naive Bayes), unsupervised (K-Means using all attributes of the flow statistics) and Early Application Identification (K-Means using only Average Packet Size as the attribute for the machine learning algorithm).

We ran ACAS classification technique, which encodes initial *n-bytes* of the payload into space and uses supervised machine learning algorithm as the classification algorithm. We observe this technique is very effective in identifying applications with accuracy over 99%. Drawback of this technique is it requires completely labeled data. Which means we should know prior to the start classification it should know all the applications that classifier may encounter in future (which is not a feasible solution). It also requires the access to the payload, which is limited in many countries by privacy concerns, and also encryption plays important role in the classification accuracy.

We ran K-Means classification technique, where we use all the available 18 attributes of flow statistics for the classification. We ran experiments to study the effect of the number of clusters[]. We observe from the plot fig.1 accuracy of the technique increases as we increase the number of clusters. In fig.1 K-Means all Attrs is the accuracy of the existing technique [] and K-Means Selected is the accuracy when we consider only subset of total attributes, based on the selection strategy which we suggested later in the paper. We observe the accuracy of the existing technique is low in comparison with the accuracy of the clustering technique when we consider only subset of the attributes.

We ran Early Application Identification[] against our dataset. Instead of considering *Average Packet Size* for first 4 packets as suggested in the paper[], we considered whole packets *Average Packet Size* of the flow. It can observed from fig.2 that the accuracy increases slowly with increase in the percent of the training set.

III. Grouping of Attributes

Total attributes considered for the study are 18.

- *Flow Size* Total number of bytes transferred during the entire flow.

TABLE II: Accuracy based on IAT group

Avg IAT	Min IAT	Max IAT	Std Div IAT	Final Accuracy
✓	×	×	×	47.29%
×	✓	×	×	49.34%
×	×	✓	×	39.20%
×	×	×	✓	42.82%
✓	✓	×	×	45.86%
✓	×	✓	×	44.25%
✓	×	×	✓	47.78%
×	✓	✓	×	39.64%
×	✓	×	✓	42.91%
×	×	✓	✓	40.54%
✓	✓	✓	×	45.20%
✓	✓	×	✓	47.18%
✓	×	✓	✓	42.70%
×	✓	✓	✓	40.10%
✓	✓	✓	✓	42.75%

TABLE III: Accuracy based on RIAT group

Avg RIAT	Max RIAT	Std Div RIAT	Final Accuracy
✓	×	×	55.86%
×	✓	×	54.08%
×	×	✓	53.41%
✓	✓	×	55.21%
✓	×	✓	54.28%
×	✓	✓	55.25%
✓	✓	✓	56.41%

- *Flow Duration* Difference between time when last packet of the flow captured and first packet of flow captured
- *Packet Size and it's variations* Size of the each packet. We can get *minimum*, *maximum*, *Average* and *Standard deviation* by using all the packets of the flow.
- *Inter Arrival Time and it's variations* Inter Arrival Time(τ')[] is the time difference between τ_{next} - $\tau_{current}$.
- *Relative Inter Arrival Time and it's variations* [] Relative inter arrival time is defined as

$$\tau_i' = \frac{\tau_i}{\min\limits_{k=1..|f|-1} \tau_k}$$

Here, f is a flow with $|f|$ number of packets and τ_i is the time difference between i-th packet and $(i+1)$-th packet in that flow.[]

A. Preliminary Results

Our experiments results for each group of attributes are as following:

We observe from Table.II, Table.III, Table.V and Table.IV that accuracy is not very good when we consider all the

TABLE V: Accuracy based on Flow attribute group

Total Pkt Size	Flow Duration	Number of Packets	Avg Pkts Per Sec	Avg Bytes Per Sec	Payload Size	Accuracy
✓	×	×	×	×	×	89.05%
×	✓	×	×	×	×	40.72%
×	×	✓	×	×	×	59.54%
×	×	×	✓	×	×	44.98%
×	×	×	×	✓	×	50.48%
×	×	×	×	×	✓	89.43%
✓	✓	×	×	×	×	92.92%
✓	×	✓	×	×	×	91.66%
✓	×	×	✓	×	×	91.12%
✓	×	×	×	✓	×	49.91%
✓	×	×	×	×	✓	94.53%
×	✓	✓	×	×	×	60.99%
×	✓	×	✓	×	×	45.13%
×	✓	×	×	✓	×	49.29%
×	✓	×	×	×	✓	92.58%
×	×	✓	✓	×	×	50.00%
×	×	✓	×	✓	×	50.68%
×	×	✓	×	×	✓	93.80%
×	×	×	✓	✓	×	49.74%
×	×	×	✓	×	✓	87.07%
×	×	×	×	✓	✓	50.94%
✓	✓	✓	×	×	×	89.95%
✓	✓	×	✓	×	×	92.31%
✓	✓	×	×	✓	×	49.88%
✓	✓	×	×	×	✓	89.54%
✓	×	✓	✓	×	×	86.20%
✓	×	✓	×	✓	×	49.81%
✓	×	✓	×	×	✓	87.25%
✓	×	×	✓	✓	×	50.58%
✓	×	×	✓	×	✓	88.83%
✓	×	×	×	✓	✓	49.14%
×	✓	✓	✓	×	×	47.74%
×	✓	✓	×	✓	×	49.67%
×	✓	✓	×	×	✓	92.02%
×	✓	×	✓	✓	×	49.40%
×	✓	×	✓	×	✓	89.36%
×	✓	×	✓	×	✓	50.17%
×	×	✓	✓	✓	×	49.45%
×	×	✓	✓	×	✓	88.46%
×	×	✓	×	✓	✓	51.83%
×	×	×	✓	✓	✓	50.90%
✓	✓	✓	✓	×	×	91.86%
✓	✓	✓	×	✓	×	49.30%
✓	✓	✓	×	×	✓	81.47%
✓	✓	×	✓	✓	×	49.64%
✓	✓	×	✓	×	✓	85.54%
✓	✓	×	×	✓	✓	51.12%
✓	×	✓	✓	✓	×	49.76%
✓	×	✓	✓	×	✓	83.67%
✓	×	✓	×	✓	✓	50.40%
✓	×	×	✓	✓	✓	50.46%
×	✓	✓	✓	✓	×	50.33%
×	✓	✓	✓	×	✓	87.11%
×	✓	✓	×	✓	✓	51.16%
×	✓	×	✓	✓	✓	51.54%
×	×	✓	✓	✓	✓	51.18%
✓	✓	✓	✓	✓	×	49.20%
✓	✓	✓	✓	×	✓	80.33%
✓	✓	✓	×	✓	✓	49.41%
✓	✓	×	✓	✓	✓	50.08%
✓	×	✓	✓	✓	✓	51.12%
×	✓	✓	✓	✓	✓	51.92%
✓	✓	✓	✓	✓	✓	50.93%

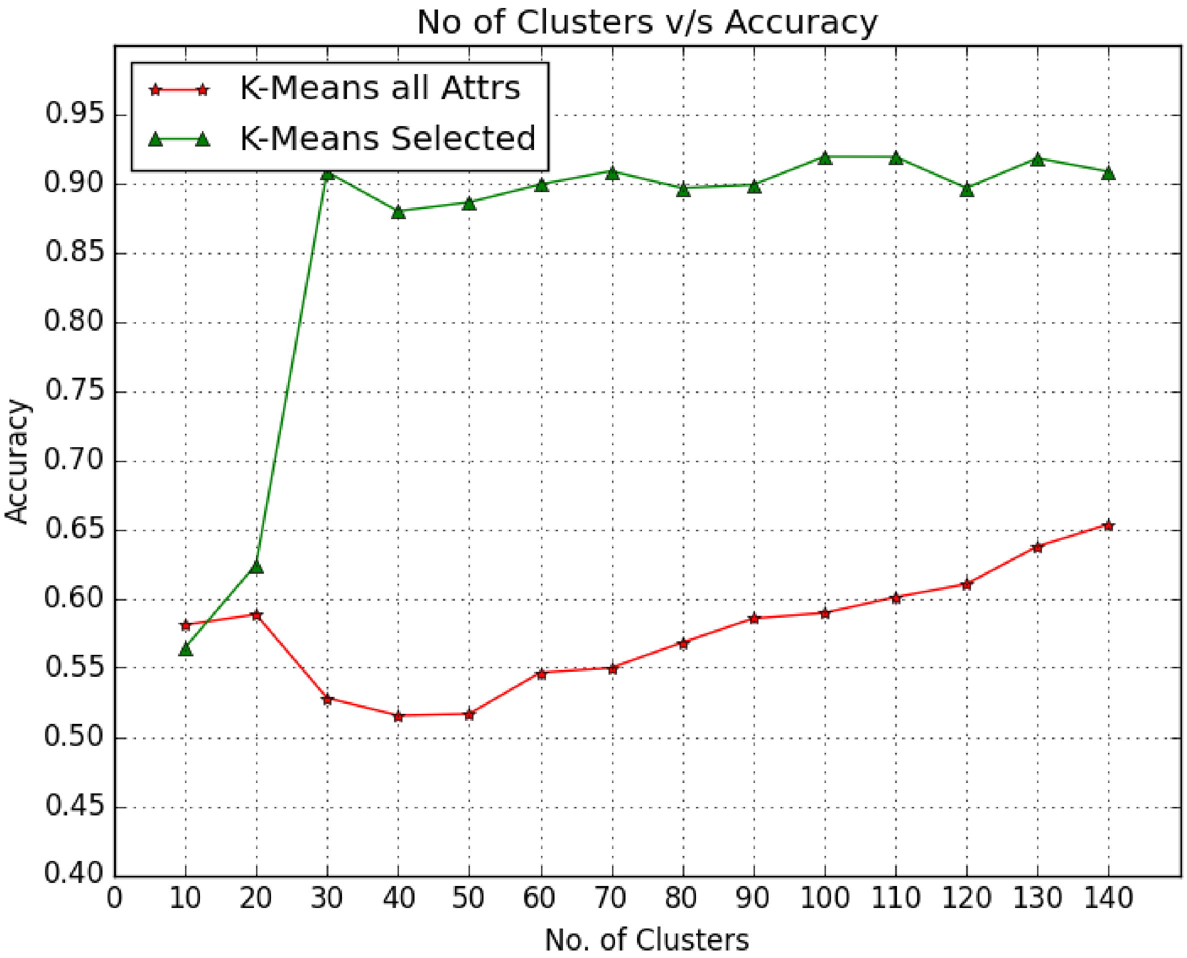

Fig. 1: Number of clusters v/s Accuracy for K-Means algorithm with All and Selected attributes

TABLE IV: Accuracy based on Packet Size group

Avg Pkt Size	Min Pkt Size	Max Pkt Size	Std Div Pkt Size	Final Accuracy
✓	×	×	×	92.43%
×	✓	×	×	75.95%
×	×	✓	×	92.22%
×	×	×	✓	91.82%
✓	✓	×	×	91.46%
✓	×	✓	×	92.98%
✓	×	×	✓	93.96%
×	✓	✓	×	91.46%
×	✓	×	✓	92.69%
×	×	✓	✓	92.76%
✓	✓	✓	×	93.98%
✓	✓	×	✓	93.62%
✓	×	✓	✓	92.76%
×	✓	✓	✓	91.73%
✓	✓	✓	✓	93.75%

attributes of flow in the classification. When subset of attributes in a group is considered then the accuracy is better than when all attributes in the group are considered. From Table.II, Table.III accuracy is very low, this group doesn't make much difference in the accuracy when considered in the classification. From Table.V and Table.IV we observe accuracy when considering all the attributes is not as good as when we consider subset of them. Table.IV we observe combination of *Avg Pkt Size* and *Std Div Pkt Size* gives the best accuracy of the group, similarly combination of *Min Pkt Size*, *Max Pkt Size* and *Avg Pkt Size* gives second best accuracy of the group. Similarly from Table.V combination of *Total Pkt Size*, *Payload Size* and also combination of *Payload size*, *Number of Packets in the flow* gives better accuracy in the group. From fig.3, we observe with only subset (7 out 18 attributes) is giving almost same accuracy(at some points even more), as when we consider total 18 attributes with Adaboost (Supervised) Algorithm[]. Extra attributes is not increasing accuracy at all. With less attributes we have faster classifier. From fig.1 we observe selection of subset of attributes increases accuracy

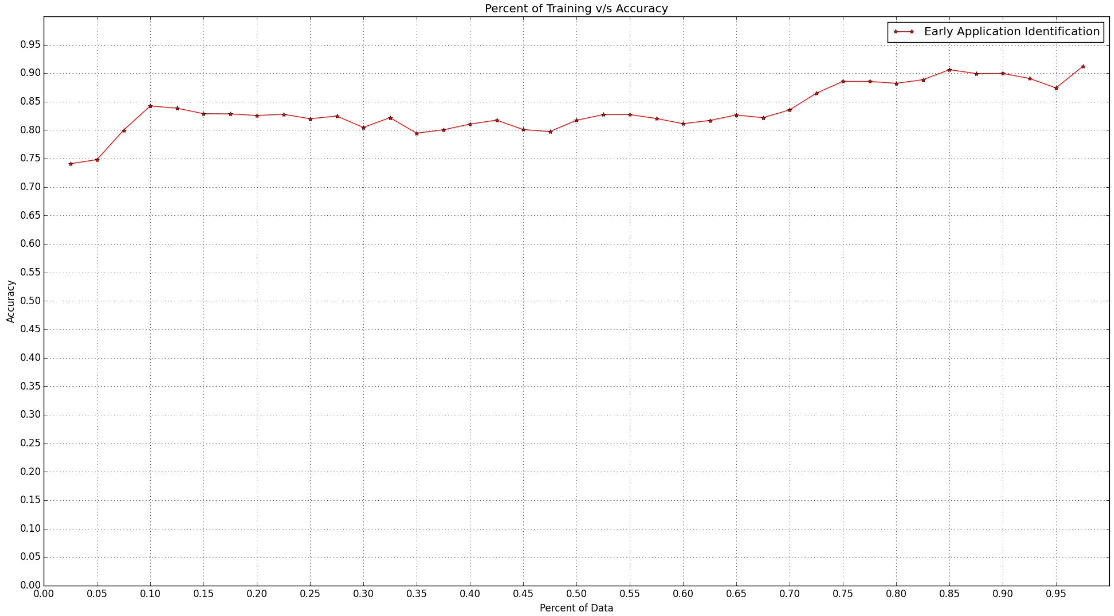

Fig. 2: Percentage of Training Data v/s Accuracy for Early Application Identification

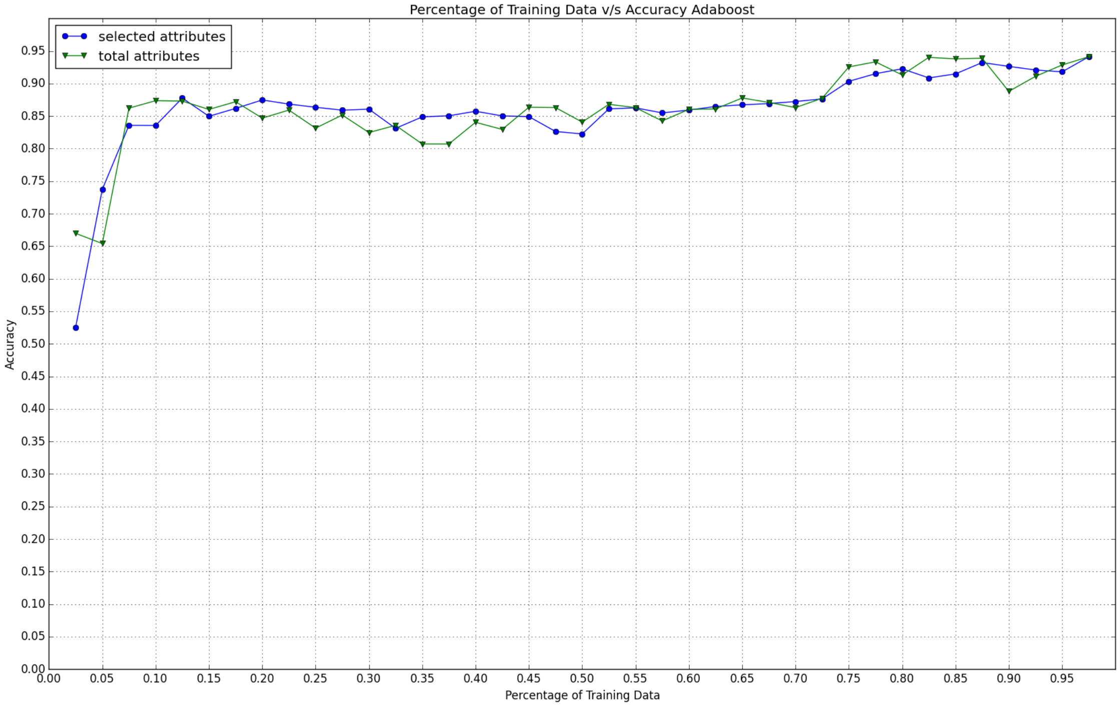

Fig. 3: Percentage of Training data v/s Accuracy for Adaboost

dramatically.

IV. Multiple Cluster Models

A. 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

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

B. Description of the technique used for classification

Based on V and IV we can observe particular combination of attributes results in better accuracy(like *Average Pkt Size* and *Std Div Pkt Size*). From fig.1 it can be observed that the accuracy of selected attributes is around 90%, which can be further improved by considering multiple trained models(with different combination of attributes) than with one trained model(which has superset of combination of attributes used in multiple models). In a single classifier, supervised or unsupervised, we cannot consider combination of attributes as we have to put every attribute to be considered through training model. So, we advised usage of four different models, each with particular combination of attributes.

- *Model 1* In this model we considered attributes, Average Pkt Size and Standard Pkt Size.
- *Model 2* We considered Average, Minimum, Maximum Packet size attributes in this model.
- *Model 3* Considered Total Flow Size, Payload Size as the attributes for this model
- *Model 4* Considered Number of Packets, Payload Size as the attributes for this model.

We use results from four models in determining the final classified result of the incoming flow. We tested different hypothesis. Following are the explanation of hypothesis considered in deciding the final classification result. *Considered Strategies*

- *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

V. Evaluation

A. Datasets

Data had been collected with full payload in early-2014. It has been gathered on various interfaces 1) Wired 2) WIFI 3) 3G and LTE. Data had been collected for individual application in isolation, by generating requests intentionally and capturing bidirectional data. Considered five tuples (i.e., Source IP, destination IP, source port, destination port and protocol). Used TCP flags to mark the start and end of flow(flow started before the capture, or flows terminating after the capture are not considered in construction of flows).

We selected 5 protocols (Skype, Bittorrent, Http, Edonkey and Gnutella) for further study, criteria for the selection of this protocols is number of flows.

B. Cleaning

We constructed flows from packets, which we read from *pcap* files using *scapy* a python library. Used community maintained signatures[] available for this protocols from L7 filters. We matched payload of each constructed flow with corresponding signatures from L7, if we find an unmatched flow then it is discarded otherwise labeled as the matched signature.

C. Experimental Setup

We used *sklearn* a python library for machine learning algorithms in our study. *sklearn* is the most used python library for data analysis in python. Used *Numpy* for numerical computations, it is a python library to handle numerical computation in efficient way. Used *Matplotlib* for plotting the results, it is a python library which lets us plot the results.

We divide our data into training and testing partitions. We used training data to train the model and testing data to test the accuracy of the trained model in correctly classifying protocols. During training phase we label clusters in each model(we have 4 of them) based on population fraction(majority based). Each test flow is fed into the trained model and we get the label for that test flow which is compared with the actual label for that flow. Finally

$$Accuracy = \frac{TP}{TotalFlows} \tag{2}$$

Where TP stands for True positive, meaning correctly identified flow during testing. Total flows are the total number of flows in the testing set.

D. Results

From fig.4 we notice classification with highest accuracy at almost all the ratios of training set is observed for Adaboost. Accuracy for Naive Bayes[] from fig.5 is very low when compared with Adaboost, it is not very effective in classifying the protocols.

From fig.6, it can noticed that accuracies are very high for *Quorum*, *Greatest* and *Unanimous Greatest* strategies.

We chose *Unanimous Greatest* for further study. Results for proposed classification technique, Supervised(Adaboost)

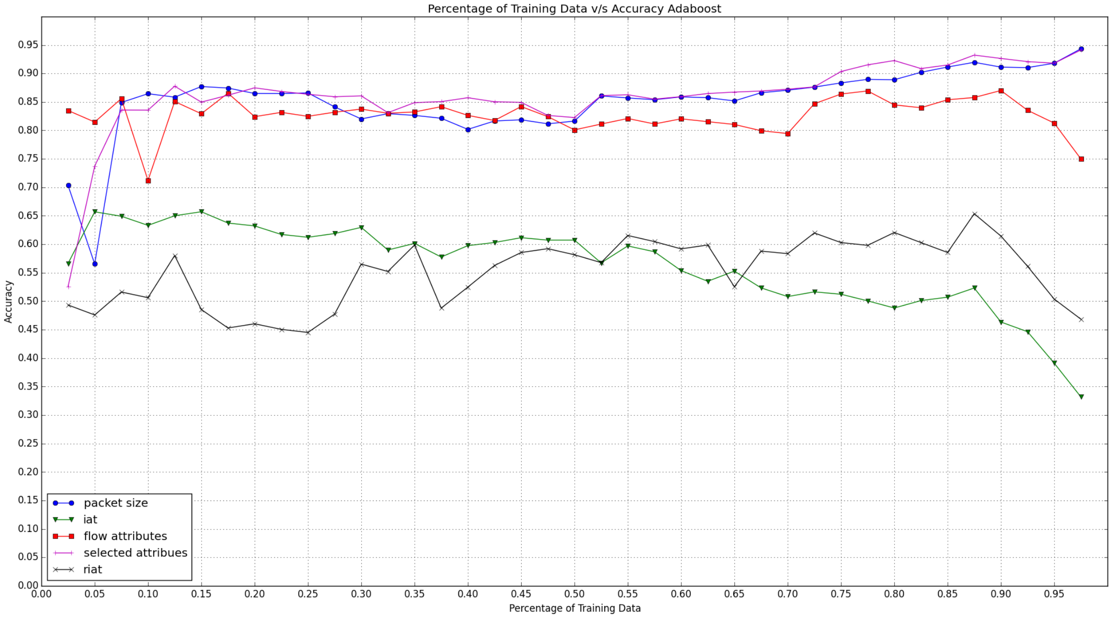

Fig. 4: Percent of training data v/s Accuracy for Adaboost with Groups of Attributes

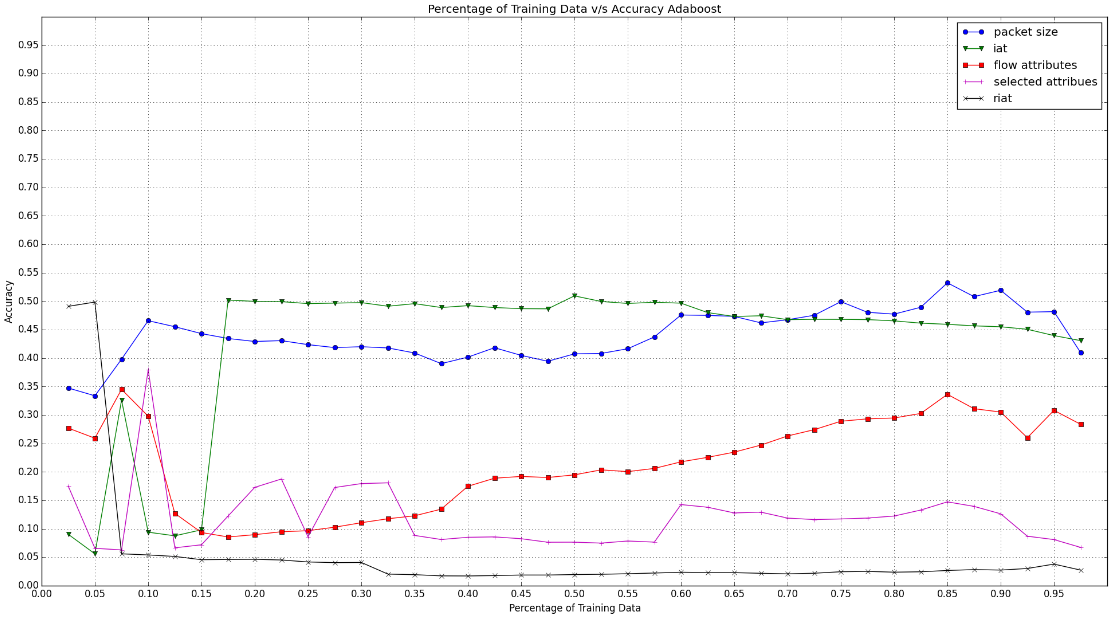

Fig. 5: Percent of training data v/s Accuracy for Naive Bayes with Groups of Attributes

and K-Means are used for comparison. From fig.7 we notice that the accuracy is higher than traditional K-Means clustering algorithm at almost all the ratios of the training set. Proposed technique accuracy is comparable to that of the supervised(Adaboost) technique, at times it is higher than supervised too.

VI. 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

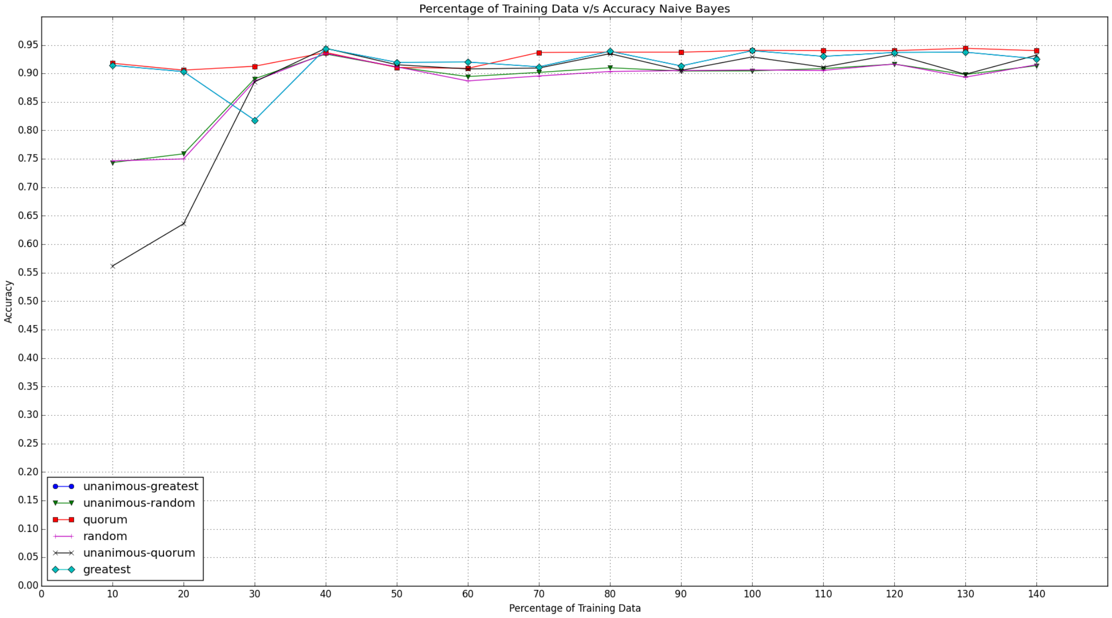

Fig. 6: Percent of training data v/s Accuracy. Comparison of various strategies

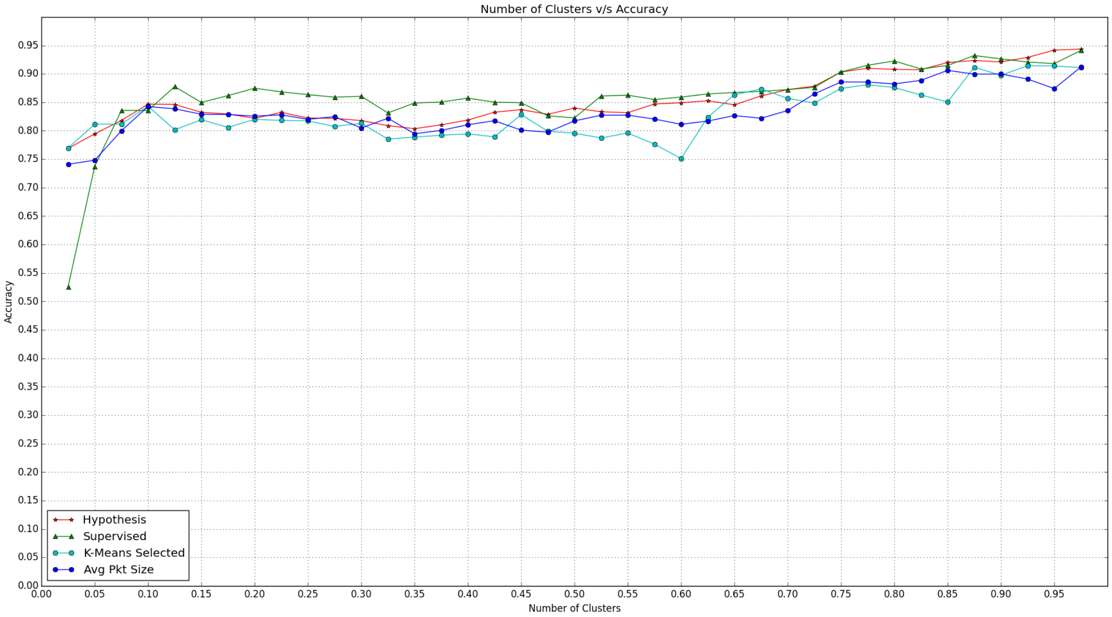

Fig. 7: Percent of training data v/s Accuracy Comparing all the existing and proposed technique

continually increases along with number of clusters, as we get better accuracy when we are around 40 clusters.

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.