A Survey of Techniques for Internet Traffic Classification using Machine Learning

Author: Thuy T.T. Nguyen and Grenville Armitage

IEEE COMMUNICATIONS SURVEYS & TUTORIALS

Presenter: Huajie Shao

Background

- Real-time traffic classification may solve difficult network management problems for Internet service providers (ISPs) and their equipment vendors.

  E.g. DOS detection, QoS and lawful interception (LI)

- Traditional methods:
  1. Port based IP traffic classification: src and dest port and IP

     Problem: not register their ports such as P2P system, ports are dynamically allocated

  2. Payload based IP traffic classification: interpret payloads

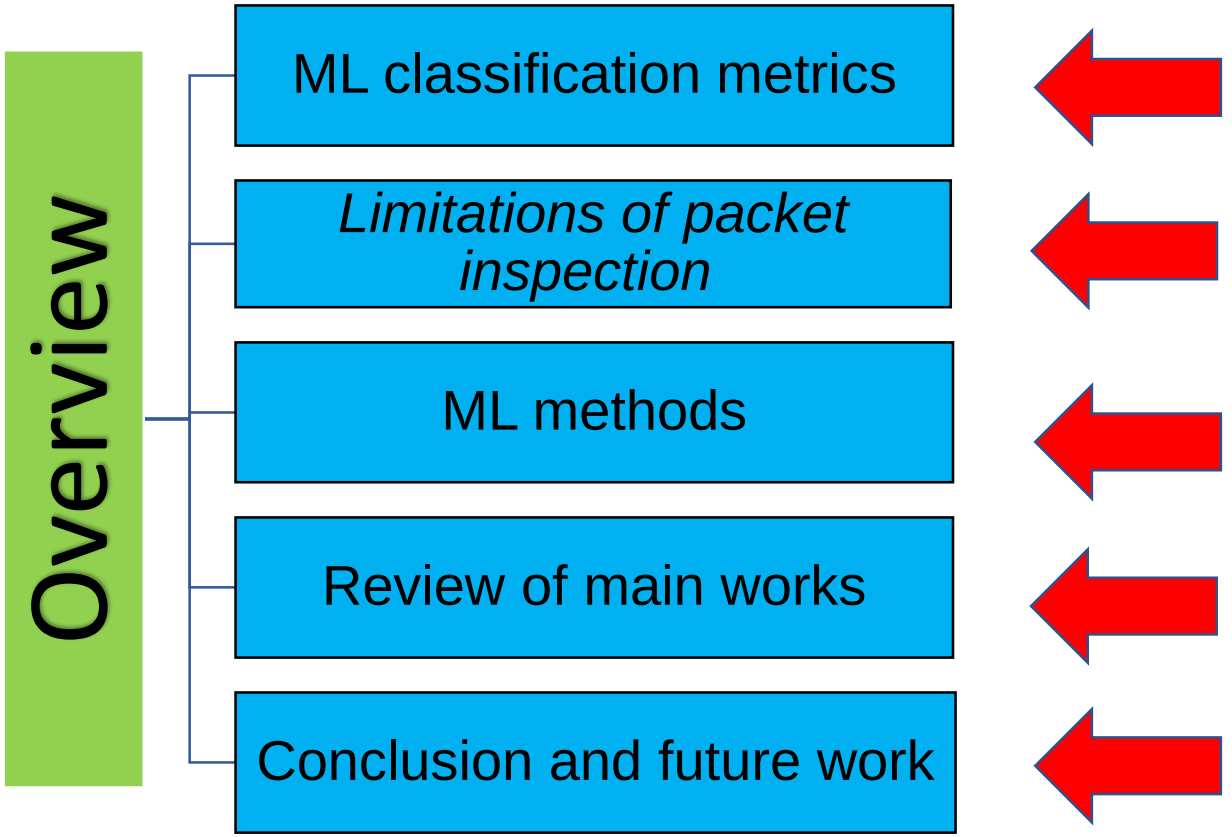
     Problem: privacy

- Recognize statistical patterns of traffic using machine learning

  E.g., packet lengths and inter-packet arrival times

Main Idea

- Review and critique emerging ML-based techniques for IP traffic classification.



ML Metrics

Confusion Matrix

Classified →	T	F
T	TP	FN
F	FP	TN

Classified	Cat	Dog
Cat	8	2
Dog	4	6

- Recall = TP/(TP+FN)
- Precision = TP/(TP+FP)
- Accuracy = (TP+TN)/All

Toy example of classification result of ML

- Recall = 8/10 = 0.8
- Precision = 8/12 = 0.67
- Accuracy = 0.7

Flow accuracy and byte accuracy

Limitation of traditional methods

- Port based IP traffic classification

  Limitation:

  - ❖ Some applications may not have their ports registered with IANA (P2P)
  - ❖ Server ports are dynamically allocated as needed (RealVideo)
  - ❖ Less than 70% accuracy for port based classification

- Payload based IP traffic classification

  Limitation:

  - ❖ Imposes significant complexity and processing load on the identification device
  - ❖ Impossible when dealing with encrypted traffic

Machine learning methods

- Two main ML methods
  - Supervised learning method: Native Bayes, Decision tree
    - Training and test. Need to label data
  - unsupervised learning method: EM, K means, AutoClass
    - Do not need to train the model, but the number of cluster is greater than application

- Features selection
  - Filter method: make independent assessment based on general characteristics of the data
  - Wrapper method: evaluate the performance of different subsets using the ML algorithm that will ultimately be employed for learning

ML framework—training and test

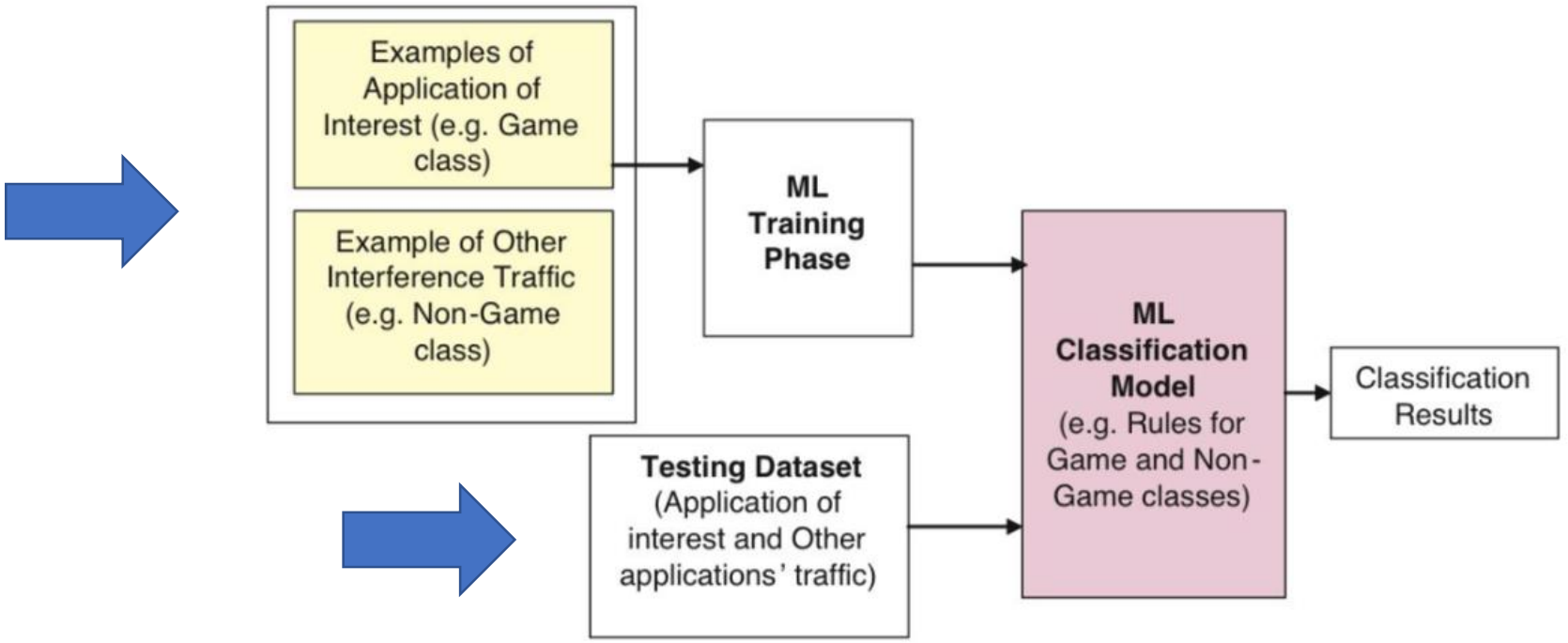


Fig. 2. Training and testing for a two-class supervised ML traffic classifier

ML framework--training

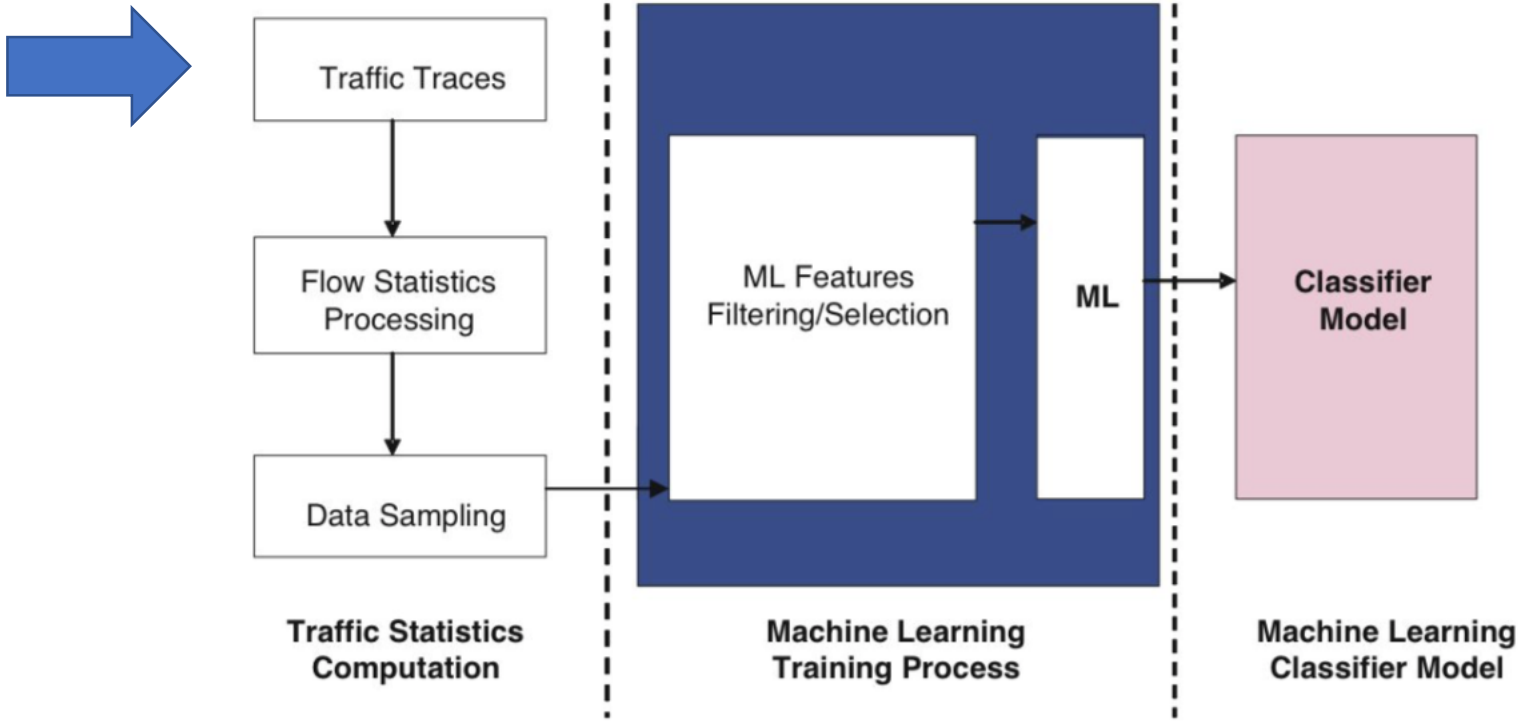


Fig. 3. Training the supervised ML traffic classifier

ML framework—operational traffic

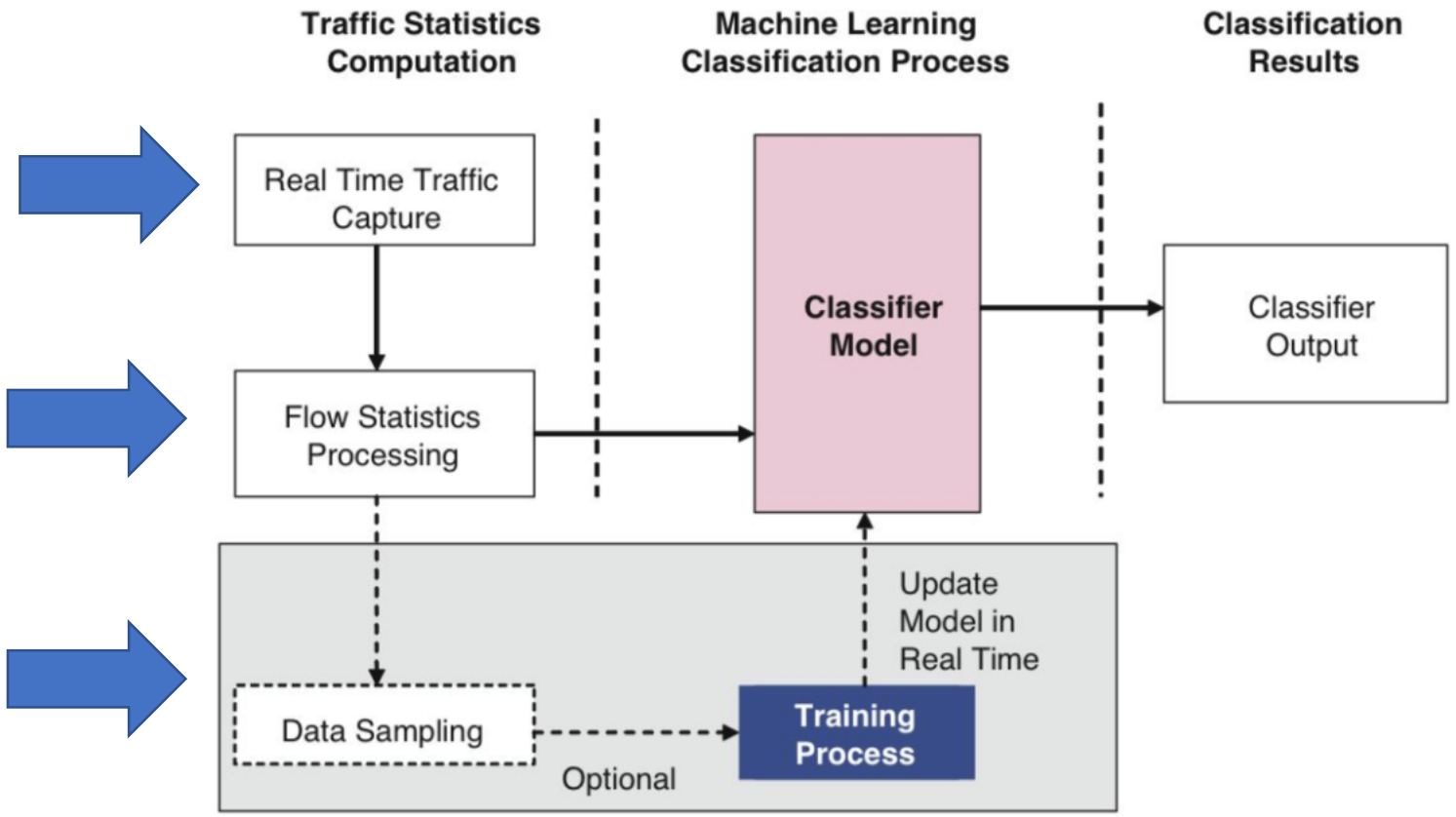


Fig. 4. Data flow within an operational supervised ML traffic classifier

Challenges for operational traffic

- *Timely and continuous classification*
  - ❖ Few packets as possible from each flow
- *Directional neutrality*
  - ❖ Application flows are often assumed to be bi-directional
- *Efficient use of memory and processors*
  - ❖ CPU cycles and memory consumption
- *Portability and Robustness*
  - ❖ A variety of network locations & consistent accuracy

Main works—EM algorithm

1. *Flow clustering using Expectation Maximization [McGregor 2014]*

- *HTTP, FTP, SMTP, IMAP, NTP and DNS traffic*
- *Features*
  - ❖ *Packet length statistics (min, max, quar- tiles, ...)*
  - ❖ *Inter-arrival statistics*
  - ❖ *Byte counts*
  - ❖ *Connection duration*
  - ❖ *Number of transitions between transaction mode and bulk transfer mode*
  - ❖ *Idle time*

Problem: EM is unsupervised method but it may suffer from local minimum

A. McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow clustering using machine learning techniques," in *Proc. Passive and Active Measurement Workshop (PAM2004)*, Antibes Juan-les-Pins, France, April 2004.

Main works--AutoClass

2. *Auto class:*
   - ❖ Repeats EM searches starting from pseudo-random points.
   - ❖ Preconfigured with the number of classes
   - ❖ Median accuracy is ≥ 80% for all applications across all traces

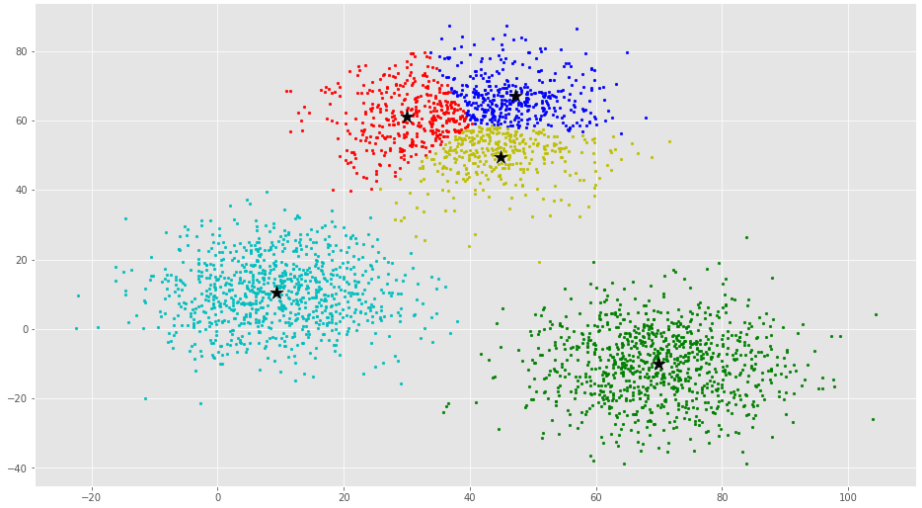
Problem: this work has not addressed the trade-offs between number of features used and their consequences of computational overhead

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

Main works--K-Means

3. K-Means for TCP-based application

- Allowed early detection of traffic flow by looking at only the first few packets
- Used Euclidean distance to measure similarity between flows
- More than 80% accuracy by using the first five packets of each TCP flow



Problem: Assumes that the classifier can always capture the start of each flow

L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly," *ACM Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review*, vol. 36, no. 2, 2006.

Main works—network core

4. Identifying Web and P2P traffic in the network core

- Used K-means to classify traffic at ***the core of the network***
- Flow and byte accuracies improved as k increased from 25 to 400
- Server-to-client dataset: 95% flow accuracy and 79% byte accuracy
- Random dataset: flow and byte accuracy is 91% and 67%
- Client-to-server dataset: 94% flow and 57% byte accuracy

Problem: it only works with TCP, not other transport protocol traffic

J. Erman, A. Mahanti, M. Arlitt, and C. Williamson, "Identifying and discriminating between web and peer-to-peer traffic in the network core," in *WWW '07: Proc. 16th international conference on World Wide Web*. Banff, Alberta, Canada: ACM Press, May 2007, pp. 883–892.

Main works--Bayesian analysis

5. Classification using Bayesian analysis techniques

- Adopted Naive Bayes method to categorize Internet traffic by application
- 248 full-flow based features were used to train the classifier
- 65% flow accuracy
- Extend it with Bayesian neural network and Naive Bayes Kernel Estimation (NBKE), 95% accuracy

Problem: Bayesian assumes that all the features are independent

T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: Multilevel traffic classification in the dark," in *Proc. of the Special Interest Group on Data Communication conference (SIGCOMM) 2005*, Philadelphia, PA, USA, August 2005.

Main works—sub-flows

6. Real-time traffic classification using Multiple Sub-Flows features

- Extract two or more sub-flows (of N packets) from every flow
- Each sub-flow should be taken from different places (start, middle of flow)
- Used Naive Bayes classifier to train these sub-flows
- Achieved 95% Recall and 98% Precision (flow accuracy)

Problem: it only identifies an online game application (UDP), may have interference with Web, P2P.

T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimise 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.

Main works--GA

7. GA-based classification techniques

- Used Genetic Algorithm (GA) to select features
- Adopted Naive Bayesian classifier with Kernel Estimation (NBKE), Decision Tree J48 and the Reduced Error Pruning Tree (REPTree) classifier

Problem: accuracy result is provided as overall result. It is not clear how it would be different for different types of Internet applications.

J. Park, H.-R. Tyan, and K. C.-C.J., "GA-Based Internet Traffic Classi- fication Technique for QoS Provisioning," in *Proc. 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Pasadena, California, December 2006.

Main works—Hybrid method

8. Semi- supervised traffic classification approach

- A training dataset consisting of labeled flows combined with unlabeled flows are fed into a clustering algorithm
- The available labeled flows are used to obtain a mapping from the clusters to the different known classes
- Achieved 94% flow accuracy

Problem: This paper did not mention the training time with labelled flows

J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Semi- supervised network traffic classification," *ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) Performance Evaluation Review*, vol. 35, no. 1, pp. 369–370, 2007.

Comparison of algorithms

- Unsupervised clustering algorithms

Algorithms	Auckland dataset	Calgary dataset
AutoClass	92.4%	88.7%
K-Means	79%	84%
DBSCAN	75.6%	72%

Dataset: University of Auckland and one self-collected trace from the University of Calgary.

Comparison of algorithms

- Unsupervised learning vs. supervised learning

Dataset: 72-hour data traces provided by the University of Auckland

Algorithms	Accuracy	Precision	Recall
AutoClass	91.2%	> 90% (6/9)	> 90% (6/9)
Naïve Bayes	82.5%	> 80% (6/9)	> 80% (6/9)

Comparison of algorithms

- *Supervised ML algorithms*

Algorithms	Accuracy
Naive Bayes with Discretisation (NBD)	>95%
Naive Bayes with Kernel Density Estimation (NBK)	> 80%
C4.5 Decision Tree	>95%
Bayesian Network	>95%
Naive Bayes Tree	>95%

Dataset: three public NLANR traces

Other methods

- Unsupervised approach for **protocol inference** using flow content
- BLINC: Multilevel traffic classification in the dark
  - ❖ Based on the **behaviors of the source** host at the transport layer
- Pearson's Chi-Square test and Naive Bayes classifier
  - ❖ Identify Skype traffic in real time

Conclusion and Future work

Conclusion:

- This paper reviewed the works about machine learning based IP traffic classification
- Compared the performance of different ML algorithms
- Summarized the advantages and disadvantages of different algorithms

Future work:

- Parallel processing for real- time classification for millions of concurrent flows
- Apply ML in skype video, intrusion detections, anomaly detection in user data and control, routing traffic and so on

Pros and Cons

Pros:

- Reviewed a large number of papers about Internet traffic classification
- Summarized the advantages and disadvantages of different algorithms in the published papers
- Gave some insights of future work for Internet traffic classification.

Cons:

- Did not discuss some advanced machine learning algorithms such as random forest and deep neural networks for classification.
- The background of introducing machine learning on Page 4 is a little bit redundant. For example, In 1983 Simon noted “Learning denotes changes in the system……”.
- Naïve Bayes assumes that the features are independent. But this paper did not comment it.

Thank you very much!!

Q&A