Internet Traffic Classification Techniques

Tanjila Ahmed

Friday Group Meeting

Nov 9, 2017

UCDAVIS
UNIVERSITY OF CALIFORNIA

Agenda

1. What is Traffic Classification?
2. Why it is necessary?
3. Traffic Classification Types
4. Port-based
5. Payload-based
6. Flow-feature-based
7. Machine-Learning : Supervised Learning
   Unsupervised Learning
8. Reference

What is Traffic Classification?

Traffic classification of internet traffic means categorizing the traffic according to various application type.

It is necessary to achieve for network management, intrusion detection, and network monitoring goals for ISPs and their equipment vendors.

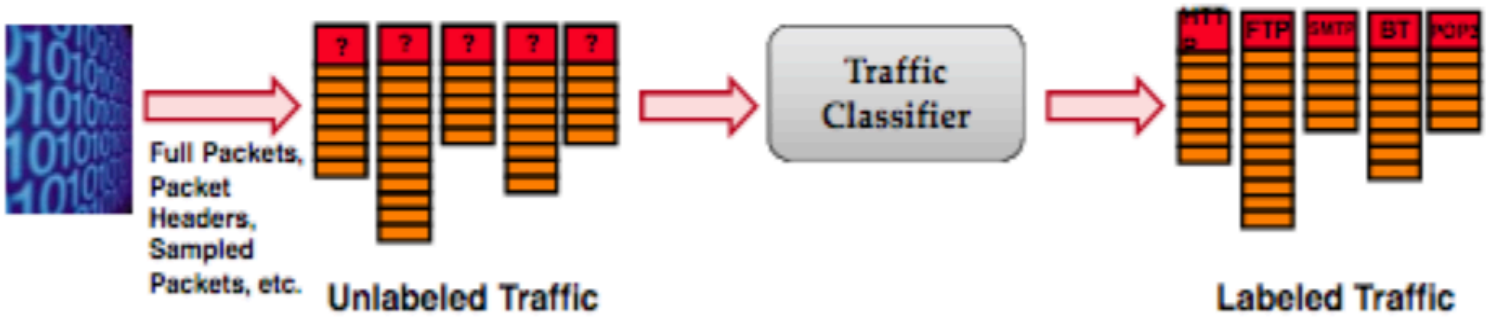
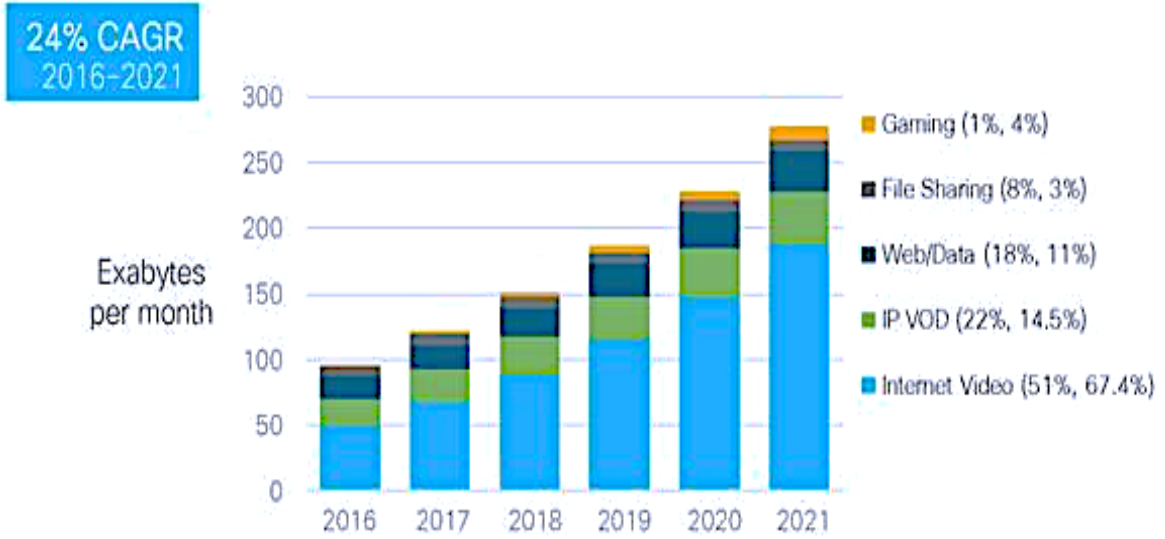


Image courtesy: [1]

UCDAVIS

Why it is necessary?[2]

- **Growth of smartphones** as the "communications hub" for social media, video consumption, tracking IoE applications, as well as traditional voice. This trend demonstrates the effect that smartphones have on how consumers and businesses users access and use the Internet and IP networks.
- **Internet gaming:** ISPs have observed a pronounced increase in traffic associated with gaming downloads. Newer consoles such as the Xbox One and PlayStation 4 have sufficient onboard storage to enable gamers to download new games(large graphically intense files) rather than buy them on disc.



Figures (n) refer to 2016, 2021 traffic shares.
Source: Cisco VNI Global IP Traffic Forecast, 2016-2021.

UCDAVIS

Why it is necessary?[2]

- **Virtual reality and augmented reality**: This growth stems mainly from the download of large virtual reality content files and applications, virtual reality streaming.
- **Immersive video**: Social media platforms such as Facebook have launched support for spherical, or immersive video that integrates multiple camera angles to form a single video stream and can be watched from the viewer's preferred perspective. It can generate bit rates 3 to 10 times greater than non-immersive HD bit rates.
- **Video surveillance**: New Internet-connected video surveillance cameras upload a constant video stream to the cloud for remote viewing. Internet-enabled cameras can produce up to 300 GB per camera per month for full HD-resolution monitoring of high-activity areas.

Traffic Classification Approaches

- Port number based
- Payload based
- Flow feature based
- Host behavior based
- Host interaction based

Port number based Approach

- Internet Assigned Number Authority (IANA) assigns port number for different applications. For example:

  TCP Port 80- HTTP

  UDP Port 53- DNS

  TCP Port 25-SMTP

- A classifier sitting in the middle of a network look for TCP SYN packets to know the server side of a new client-server TCP connection. It infers the application by looking up the TCP SYN packets' target port number.

Port number based Approach

Advantages:

- Simple to implementation
- Very efficient even in large networks
- The first packet can be used for classification

Disadvantages:

- Some applications may not have their ports registered (Napster, Kazaa)
- Applications may use ports other than its well-known ports (non-privileged users)
- Dynamic allocated port number (RealVideo streamer)
- IP layer encryption may obfuscate TCP/UDP header

Payload based Approach

- Also known as ‘Deep Packet Inspection (DPI)
- Classifier matches ‘Signatures’ in a payload to a known application
  - Regular expression for HTTP signature
  - Regular expression for FTP signature
- It avoids reliance on fixed port numbers

Disadvantages:

1. Significant complexity and processing load on the classifier
2. Difficult to implement on proprietary protocols
3. Privacy policy breaching
4. Cannot identify new applications
5. Application payload encryption

UCDAVIS

Flow-Feature Based Approach

- Classification without looking into the payload:

  Newer approaches rely on traffic's statistical characteristics to identify the application.

Assumption: *traffic at the network layer has statistical properties(such as distribution of flow duration, flow idle time, packet inter-arrival time and packet lengths) that are unique for certain classes of applications and enable different source applications to be distinguished from each other.*

- Which feature is best suited to classify the applications? How well these features works for all other applications.

Machine-Learning Approach

- *ML has historically been known as a collection of powerful techniques for data mining and knowledge discovery, which search for and describe useful structural patterns in data* [3].
- A network traffic controller using ML techniques was proposed in 1990, aiming to maximize call completion in a circuit-switched telecommunications network [4];
- In 1994 ML was first utilized for Internet flow classification in the context of intrusion detection [5].

ML Based Approach

Different types of learning:

1. Supervised (Classification) : learn from a set of known flow
2. Unsupervised (clustering) : Group flows based on measure of similarity
3. Association: find association between features; group flows with the same feature
4. Numeric Prediction: the outcome is not a discrete class but a numeric quantity

Supervised Learning

- Supervised learning creates knowledge structure that supports the task classifying new instances into pre-defined classes. The knowledge learnt can be presented as a flowchart, a decision tree, classification rules, etc.

Two major steps:

1. **Training** : the learning phase that examines the provided data and constructs a classification model.
2. **Testing**: The model that has been built in the training step is used to classify new unseen instances.

Supervised Learning

Challenges:

1. Lack of large labeled training and independent testing datasets
2. Labeling the testing and training dataset(pre-labeled dataset is scarce)

Hold out/N-fold cross-validation:

-Each dataset is partitioned into N approximately equal parts.

-Each partition in turn is used for testing, while the remainder N-1 is used for training

-This process is repeated N times.

Examples: Naïve Bayes, Decision Tree etc.

UCDAVIS

Unsupervised Learning

- It discovers natural clusters in the dataset
- It clusters instances with similar properties defined by a specific distance metric such as Euclidean space into groups.
- This groups can be exclusive or overlapping. Overlapping means one instance falls into several groups. They can be probabilistic as well; that is an instance belongs to a group with certain probability. They can be hierarchical as well.
- Example: k-means, incremental and probability based clustering

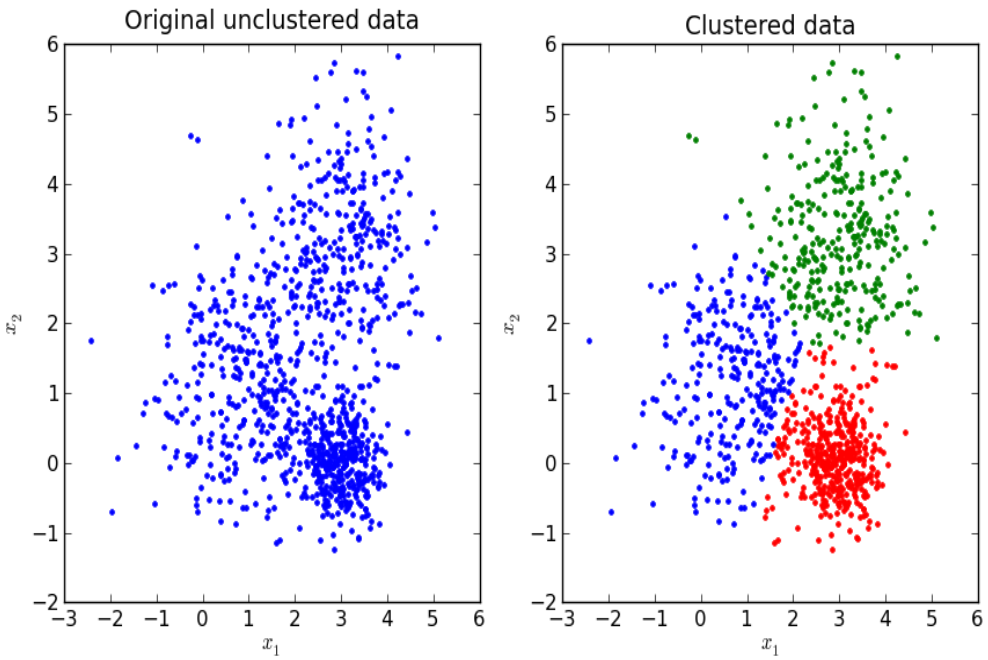


Image courtesy: [7]

Unsupervised Learning

Advantages:

Unsupervised learning method is not dependent on any training dataset, so it is able to identify any new type of application.

Challenges:

Clusters do not have a 1:1 mapping with applications. Usually there is more clusters so mapping back from cluster to application might be challenging.

It requires human input to make sure the clusters make sense

ML based Traffic Classification Techniques

- Clustering approach
- Supervised learning
- Semi-supervised: data whose classes are known and data whose classes are unknown co-exists together.
- Comparisons: works that compares different ML algorithms , or considers non-ML approaches in conjunction with ML approaches.

ML based Traffic Classification Techniques

1. **Flow clustering using Expectation Maximization:** based on flow features(packet length, inter arrival time, byte count etc) EM algorithm groups the traffic into a small number of clusters. Then manually classify this clusters.
2. **AutoClass:** Unsupervised Bayesian classifier using EM algorithm to select best clusters from a set of training data. To achieve global maxima it repeats EM searches multiple times.
3. **K-mean:** Unsupervised ML using first few packets of traffic flow. It was assumed that the first few packets captures the application negotiation phase which is distinct among applications.

ML based Traffic Classification Techniques

Flows are grouped into clusters based on the values of their first P packets. This flows are then represented by a point in a P dimensional space. The Euclidean distance between new flow and predefined cluster is computed. New flow belongs to the cluster for which the distance is minimum.

4. 3-Tuple heuristic: Assumption is, flows sharing same destination ip, destination port and transport layer protocol are generated by same application in a short period of time.

5. Extreme Learning Machine (ELM) algorithm is used to classify applications thousand times faster than conventional feedforward neural network. With randomly chosen input weights and hidden layer biases can exactly learn *N* distinct observations.

UCDAVIS

ML based Traffic Classification Techniques

5. Profiling by association: PBA takes as input an IP-to-IP connectivity graph and information about a small subset of IP-hosts and produces a prediction about the class of all the flows (edges) in the graph.

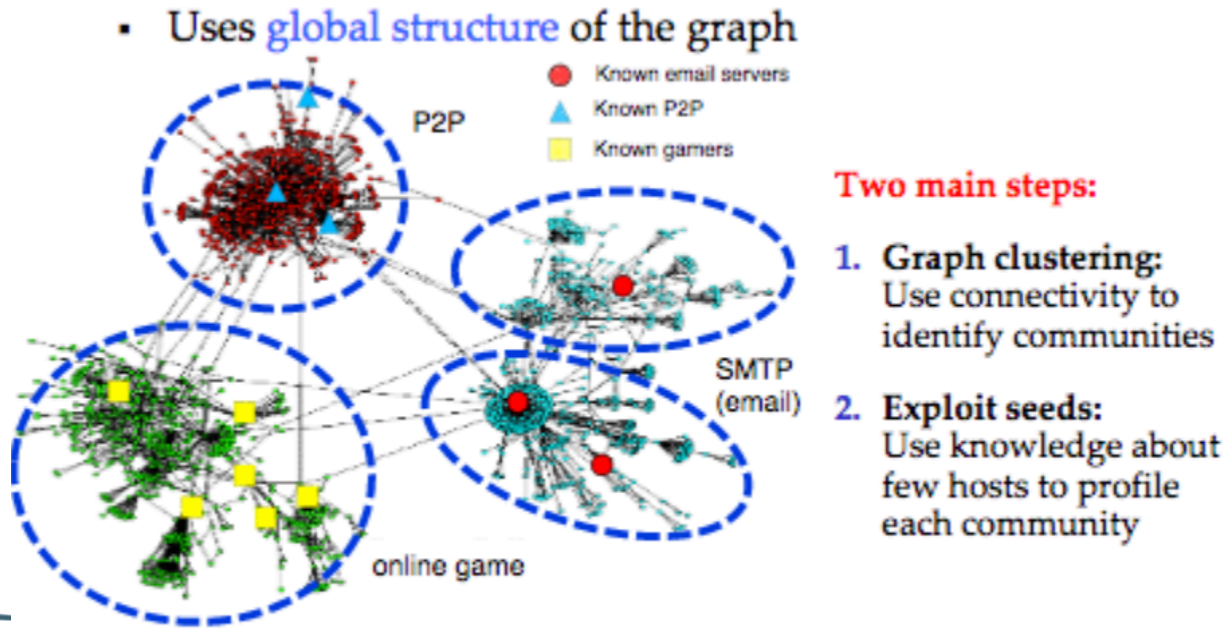


Image courtesy: [1]

Reference

[1] EEC 274 lecture notes from Prof. Chen-neh Chuah.

[2]https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html#_Toc484556821

[3] Nguyen, Thuy TT, and Grenville Armitage. "A survey of techniques for internet traffic classification using machine learning." *IEEE Communications Surveys & Tutorials* 10, no. 4 (2008): 56-76.

[4] B. Silver, “Netman: A learning network traffic controller,” in Proc. Third International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Association for Computing Machinery, 1990.

[5] J. Frank, “Machine learning and intrusion detection: Current and future directions,” in Proc. National 17th Computer Security Conference, Washington,D.C., October 1994.

[6] Ertam, Fatih, and Engin Avci. "Classification with intelligent systems for internet traffic in enterprise networks." *Int. J. Comput. Commun. Instrum. Eng* 3 (2016): 1469-2349.

[7] https://stackoverflow.com/questions/24645068/k-means-clustering-major-understanding-issue