Applying Machine Learning Mechanism with Network Traffic
(draft-jiang-nmlrg-traffic-machine-learning)

Sheng Jiang
Bing Liu (Speaker)
Panagiotis Demestichas
Jerome Francois
Giovane C. M. Moura
Pere Barlet

@NMLRG, ietf96, July 2016
Reminder

• NMLRG meeting in IETF95
  – Focused on ML (Machine Learning) applied in network traffic handling
  – Some distinct use cases from different organizations were presented in the meeting

• draft-jiang-nmlrg-traffic-machine-learning
  – Analyzing/Summarizing the methodology of learning from traffic
  – Documenting use cases presented in ietf95
  – This presentation focuses on the draft itself: https://tools.ietf.org/html/draft-jiang-nmlrg-traffic-machine-learning
Motivation: Why Focused on Network Traffic?

- The user contents within traffic is becoming more diverse due to the development of various network services, and increasing use of encryption.
- It is more and more challenging for administrators to get aware of the network's running status (such as performance, failures, and security etc.) and efficiently manage the network traffic flows.
- It is natural to utilize machine learning technology to analyze the large mount of data regarding network traffic, to understand the network's status.
Methodology Analysis of Learning from Traffic

• Data of the Network Traffic
• Data Source and Storage
• Architecture Considerations
• Closed Control Loop
Data of the Network Traffic (1/2)

• Measurable properties
  – Latency, packet count, session duration etc.
  – These properties are very essential features, especially for use cases relevant to performance, QoS (Quality of Service), etc.

• Data within communication protocols
  – Protocol headers: e.g. source/dest IP, port number
  – Application protocols: e.g. FTP, HTTP(s), SMTP etc.
Data of the Network Traffic (2/2)

• Type of user content
  – e.g. file transferring/sharing, Web, Email, Video/Audio streaming etc.

• Data in network signaling protocols
  – Traffic flows are managed or indirectly influenced by various network signaling protocols:
    • Routing: IGP/BGP, MPLS-TE, Segment Routing etc.
    • P2P
Data Collection and Storage

• Data collection
  – Forwarding devices: in theory they could collect any kind of data in the traffic; however, mostly they’re suitable to collect data such as measurable properties, protocol information.
  – Source/dest nodes: especially for servers, they could provide session data, application data etc.

• The devices either collect data to a central repository for storage and learning, or collect and store the data by themselves for local learning.
Architecture Considerations (1/3)

• Global learning vs. local learning
  – Global learning refers to the tasks that are mostly network-level, so that they need to be done in a global viewpoint.
  – Local learning is more applicable to the tasks that are only relevant to one or a limited group of devices, and they could be done directly within that one node or that limited group of nodes.
  • In this case of grouped nodes, the data may also need to be transited from the data source entity to learning entity.
Architecture Considerations (2/3)

• Offline & online learning
  – Co-located mode: training (offline, based on historic data) and prediction (online, based on real-time data) are both done within the same entity.
  – De-coupled mode: training is done in the central repository, and prediction is made by the routers/switches/firewalls or other devices that directly process the network traffic.
• Central learning & distributed learning
  – Central learning means the learning process is done at a single entity, which is either a central repository or a node.
  – Distributed learning refer to ensemble learning that multiple entities do the learning simultaneously and ensemble the results together to sort out a final results.
    • Since network devices are naturally distributed, it could be foreseen that ensemble learning is a good approach for a certain of use cases.
Closed Control Loop

• Forming a closed control loop with the prediction/decision made by machine learning:
  – could be directly used on manipulating the network traffic
  – changing the device configuration, etc.

• Closed control loop might be suitable only for a small set of the use cases, due to the limited accuracy of machine learning technologies.
  – some critical usages simply cannot tolerate any false decision.
Next Step

• Contributions and reviews are needed
  – Is there something important missing?
  – Improvement of methodology analysis
  – Improvement of use case description
Comments?

Thank you!

IETF96, Berlin