A Survey of Techniques for Internet Traffic Classification using Machine Learning

Author: Thuy T.T. Nguyen and Grenville Armitage

IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 10, NO. 4, FOURTH QUARTER 2008

Reviewer: Huajie Shao

Main idea:

This paper reviewed the emerging techniques for Internet traffic classification based on machine learning (ML). There are two main ML methods: supervised learning and semi-supervised learning. For supervised learning, this paper mainly discussed several important algorithms such as Naïve Bayes and Decision Tree. For unsupervised learning, Expectation Maximization (EM) and AutoClass were present in this paper. The authors compared the advantages and disadvantages of each algorithm from different papers. Finally, the authors gave some suggestions for the future work.

Pros:

- The authors reviewed a large number of papers about Internet traffic classification in the past few years.
- This paper summarized the advantages and disadvantages of different algorithms in the published papers.
- This paper gave some insights of future work for Internet traffic classification.
- This paper outlined a number of critical operational requirements for real-time classifiers.

Cons:

- This paper 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……".
- In general, supervised learning has higher accuracy than the unsupervised learning method. But this paper mentioned that the accuracy of Naïve Bayes is lower than the AutoClass. This paper did not explain the reason.
- Naïve Bayes assumes that the features are independent. But this paper did not comment it.

Comments:

There are two folds to improve the paper.

- This paper could introduce some advanced machine learning methods such as deep learning to improve the classification accuracy.
- In addition, the authors need to reduce the redundancy of background introduction of ML algorithm.