CLASSIFICATION OF IP TRAFFIC FLOWS USING MACHINE LEARNING METHOD

Nikitina L.O., Lysenko O.V.
National Technical University
«Kharkiv Polytechnic Institute»,
Kharkiv

Network traffic classification is very important problem for different network activities - Quality of Service, security and intrusion-detection, network monitoring. Identification of features of traffic flows greatly improves computational performance. Application of Machine Learning (ML) methods provides good results in network traffic classification.

ML methods for IP traffic classification can be divided into two groups - supervised and unsupervised. In supervised methods the knowledge about classes of each traffic flow are used before learning. A training set of example instances for every known class has to be preformed. The model of IP traffic flows can determine the class for new instances by identifying the feature values of unknown flows. Unsupervised algorithms provide clustering according to similar feature values. ML methods of this group determine the number and statistical nature of not pre-defined clusters. Supervised ML methods are Bayesian Network (acyclic graph of nodes and links, and conditional probability tables where nodes represent features or classes, links are the relationship between them, conditional probability tables determine the strength of the links), C4.5 Decision Tree (creates a tree structure where nodes represent features, branches represent possible values connecting features, leafs represent the class), Naive Bayes (estimates the probabilities of a feature having a certain feature value), Naive Bayes Tree (a hybrid of a decision tree classifier and a Naive Bayes classifier). In Deep Belief Network (DBN) method the number of flow features for classification is shortened. DBN is composed of multiple layers of hidden variables, with connections between the layers but not between units within each layer. Each sub-network's hidden layer is the visible layer for the next. This gives a fast, layer-by-layer unsupervised training procedure.

In our research, we have applied DBN method to classify instances of network traffic flows. For each flow we created vector of statistical attributes and associated feature values. A feature is a descriptive statistic that can be calculated from one or more packets – such as packet length or the deviation of arrival times, port numbers, time between consecutive flows. Each traffic flow is characterised by the same set of features, though each will exhibit different feature values depending on the network traffic class to which it belongs. Decision about the belonging of the traffic flow to the class was built on the basis of the full contents of the packets.