機械学習に基づくネットワークトラフィックの分析手法

和田見　一真

ネットワークトラフィックの分類はネットワーク管理やセキュリティ監視、通信の最適化などネットワーク関連の様々な活動に有用である。近年、情報化社会の進展によって、インターネット上のネットワークやアプリケーションのデータ量が急激な増大を続け、トラフィック分類の重要性がより高まっている。2017 年の総務省の発表によれば、日本におけるトラフィックは 2004 年から 2016 年にかけて約 40 倍も増えている。しかし、最近は SSL に代表される暗号化された通信が主流なので、DPI(ディープ・パケット・インスペクション)などのパケット全体を解析する必要がある従来の手法では十分な分析が難しくなりつつある。そこで最新の研究では従来の手法が抱える問題に対処する有望な方法として機械学習が用いられるようになってきた。機械学習はパケットの中身を分析するだけでなく、その統計的な特性をもとに分類することもできる。そこで本研究ではネットワークフローに注目して機械学習によるネットワークトラフィックの分類とその分類に適したネットワークの属性の分析を行う。

本研究では、機械学習ソフトウェア Weka を用いて教師なし学習アルゴリズムの一つである k-means 法で、ネットワークフローをいくつかのクラスタにグループ化した。研究で使用したデータセットはデルカウア大学の学内ネットワークでパケットキャプチャにより、合計 3,577,296 件のネットワークフローを収集したものである。そのデータセットでは、レイヤー7 として SSL を含む HTTP を用いたアプリケーションのラベルがあらかじめ付与されている。そこから 5 つの主要なアプリケーションごとに約 2 万件になるようにランダムサンプリングし、計約 10 万件としたものを 4 セット準備して実験を行った。クラスタの数は 3 から 7 の範囲で指定して、それぞれの数におけるクラスタリングの精度の変化を確かめた。それらのクラスタごとに振り分けられたプロトコルとフローごとのラベルを照合して、その適合率を求めた。トラフィック分類に大きな影響を与えるネットワークフローの因子を特定するために、一回のクラスタリングごとに利用する属性を必要に応じて変えながら、クラスタリングとその適合率の算出を繰り返した。

実験の結果、クラスタの数に関係なく順方向フローを含んだモデルと逆方向フローを含んだモデルでは、明らかに順方向フローの方が優れた精度をもたらすことがわかった。データセットのプロトコル数とクラスタ数が一致する場合では、平均して 4.1 ポイント順方向フローの精度が高かった。より精度を高めるためには異なるクラスタリングアルゴリズムの使用や教師あり学習アルゴリズムとの併用が考えられる。

（指導教員　阪口哲男）