SDN 을 위한 머신러닝 기반 트래픽 분류 기법 분석 연구

엄원주*, 송영준, 김건환, 조유제
경북대학교

dnjswn9612@knu.ac.kr, syj5385@knu.ac.kr, {kgh76, yzcho}@ee.knu.ac.kr

An Analysis of Machine Learning based Traffic Classification Technique for Software Defined Network

Won-Ju Eom*, Yeong-Jun Song, Geon-Hwan Kim, You-Ze Cho
School of Electronics Engineering, Kyungpook National University

요 약

다양한 애플리케이션의 등장과 유무선 네트워크에서의 급격한 트래픽 증가로 인해 정확한 트래픽 분류와 애플리케이션 특성화의 중요성이 부각되고 있으며, 복잡한 네트워크 인프라를 효율적으로 관리하기 위해 소프트웨어 정의 네트워크 (Software Defined Networking, SDN)가 등장하게 되었다. SDN 의 제어기는 전체적인 네트워크 정보를 수집할 수 있기 때문에 SDN 은 트래픽 분류 기법의 구현에 큰 이점을 갖는다. 특히 머신러닝 기반 트래픽 분류 기법은 payload 와 독립적인 트래픽 통계 feature 를 사용하여 애플리케이션 특성화 측면에서 기존의 트래픽 분류 기법보다 더 나은 선택권을 제공한다. 본 논문에서는 대표적인 머신러닝 알고리즘을 이용한 트래픽 분류 결과를 비교 분석하고, SDN 환경에서 더욱 효율적인 트래픽 분류가 가능한 머신러닝 알고리즘을 제시한다.

Ⅰ. 서론

시스코는 2017 년부터 2022 년까지 전 세계 IP 트래픽은 26%의 연평균복합성장률을 기록할 것이며, 2017 년에 16GBytes 였던 일인당 월 평균 IP 트래픽이 2022 년에 50GBytes 에 이를 것으로 예측한다 [1]. 폭발적으로 증가하는 네트워크 트래픽에 적합한 Quality of Service (QoS)를 제공하기 위해서는 트래픽 분류가 필수적이다.

소프트웨어 정의 네트워크 (Software Defined Networking, SDN)는 제어 평면 (Control plane)과 데이터 평면 (Data plane)의 분리를 통해 네트워크의 라우팅과 제어 및 복잡한 운용관리를 효율적으로 처리할 수 있도록 하는 혁신적인 네트워킹 기술이다. SDN 의 중앙집중식 구조로 인해 SDN 의 제어기는 전체적인 네트워크 정보를 수집할 수 있어, 스위치에서 트래픽의 통계적인 feature 를 비교적 간단하게 추출할 수 있다. 따라서 SDN 은 트래픽 분류 기법의 구현에 매우 적합할 것으로 기대된다.

널리 사용되는 트래픽 분류 기법에는 포트 번호 기반 접근법, 페이로드 기반 접근법, 머신러닝 기반 접근법 등이 있다 [2]. 포트 번호를 기반으로 하는 트래픽 분류는 애플리케이션에서 사용하는 well-known 포트 번호를 분석하는 기법이다. 하지만 현재 대부분의 애플리케이션은 동적으로 할당된 포트 번호를 사용하므로, 포트 번호 기반 접근법은 더 이상 효과적이지 않다. 페이로드 기반 접근법은 애플리케이션의 페이로드를 분석하여 특정 signature 를 추출하고, 이를 통해 애플리케이션을 식별한다. 이 기법은 분류 정확도가 높다는 장점을 가지지만, 최근 암호화된 페이로드를 발생하는 애플리케이션의 비율이 증가하고 있어서 활용에 제약이 있다.

따라서 기존 트래픽 분류 기법의 한계를 넘어서기 위해 머신러닝 기반의 트래픽 분류 기법에 대한 많은 연구가 진행되고 있다 [2]. 머신러닝 기반 트래픽 분류 기법은 동적으로 변하는 포트 번호의 패턴을 알아낼 수 있고, 페이로드와 독립적인 통계 feature 를 이용하기 때문에 애플리케이션의 변화에 보다 능동적으로 대처할 수 있다 [3].

본 논문에서는 LightGBM 과 XGBoost 와 같은 앙상블 알고리즘을 포함하여 대표적인 머신러닝 알고리즘을 이용한 트래픽 분류 결과를 비교 분석하고, SDN 환경에서 효율적인 트래픽 분류가 가능한 머신러닝 알고리즘을 제시한다.

Ⅱ. 본론

A. 실험 방법

본 논문에서는 트래픽 분류 모델에 어떤 머신러닝 알고리즘의 적용이 가장 적합한지 알아보기 위해 여러 개의 모델을 이용하여 실험을 진행하였다.

이 실험에서는 Kaggle 데이터베이스의 “Labeled Network Traffic Flows - 141 Applications” 데이터셋을 사용했다 [4]. 해당 데이터셋은 Universidad Del Cauca, Popayn, Colombia 에서 여러 패킷 캡처 도구와 데이터 추출 도구를 이용해 네트워크 데이터를 수집하여 생성되었으며, 2,704,839 개의 인스턴스와 50 개의 feature 로 구성되어 있다.

트래픽 플로우의 feature 는 관측 수준에 따라 flow-level feature 와 packet-level feature 로 구분된다. Flow-level feature 는 일반적으로 플로우가 완료된 후에 계산된다. 반면 packet-level feature 는 플로우의 초기 단계에 계산되며, 초기 트래픽 식별을 위해서는 5~7 개의 패킷이 적절하다 [5]. 실제 네트워크

환경에서는 빠른 트래픽 분류가 중요하므로, 플로우가 완료될 때까지 기다리기보다는 가능한 적은 수의 패킷을 사용하여 분류 결과를 도출하는 것이 유리하다. 이를 고려하여, 데이터셋의 49 개의 feature 중 11 개의 feature 를 선정하였다.

데이터 전처리 과정을 거친 후 데이터셋은 8:2 의 비율로 학습 데이터셋과 테스트 데이터셋으로 분리된다. 모든 트래픽 분류 모델은 학습 데이터셋을 이용해 학습되고, 테스트 데이터셋을 통해 각 모델의 정확도와 분류 속도가 계산된다. 그림 1 은 본 연구의 전체 실험 과정을 나타낸다.

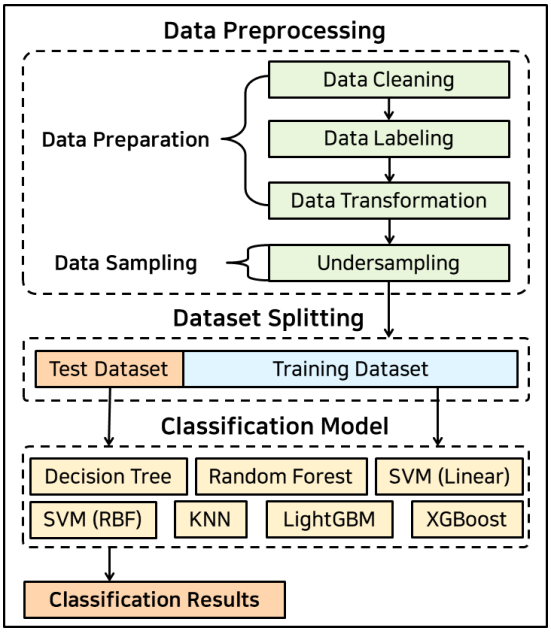

그림 1. 실험 과정

B. 실험 결과

그림 2 는 전체 feature 세트를 사용할 때와 선정된 11 개의 feature 세트를 사용할 때 각 머신러닝 알고리즘을 이용한 모델의 분류 정확도를 나타낸다. k-NN 알고리즘을 제외한 모든 알고리즘은 전체 feature 세트를 사용할 때 더 높은 정확도를 달성하지만, Random Forest, XGBoost, LightGBM 알고리즘은 선정된 feature 세트를 사용하여도 높은 정확도를 보이는 것을 확인할 수 있다.

그림 3 는 각 머신러닝 알고리즘을 이용한 모델이 전체 feature 세트와 선정된 feature 세트를 사용하여 11,256 개의 플로우를 분류하는데 걸리는 시간을 나타낸다. k-NN, SVM, XGBoost 알고리즘은 선정된 feature 를 사용하는 경우 더 빠르게 분류를 완료한다. 하지만 전체 feature 세트를 사용하여도 빠른 분류 속도를 보이고, 동시에 어떤 feature 세트를 사용하여도 가장 높은 정확도를 갖는 LightGBM 이 실험에서 사용된 6 개의 알고리즘 중 SDN 환경에서의 트래픽 분류를 위한 머신러닝 알고리즘으로 가장 적합하다.

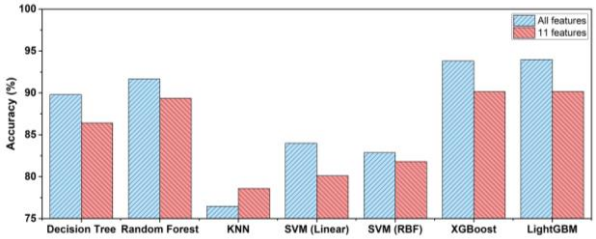

그림 2. 분류 정확도

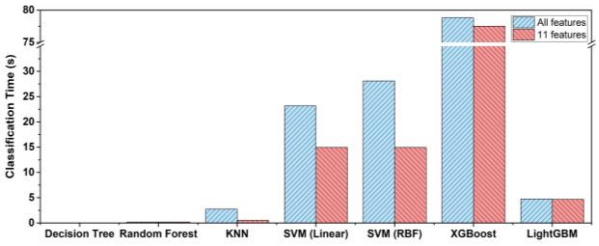

그림 3. 분류 시간 (초)

Ⅲ. 결론

머신러닝 기반 트래픽 분류는 기존의 트래픽 분류 방식의 한계를 극복하는 더욱 효과적인 분류 기법이다. 본 논문에서는 6 가지의 머신러닝 알고리즘을 이용한 트래픽 분류 결과를 비교 분석하고, 향후 연구를 위해 SDN 환경에서 효율적인 트래픽 분류가 가능한 머신러닝 알고리즘을 제시하였다. 추후에는 실제 SDN 환경에서 트래픽 분류 실험을 수행하여, 본 연구에서 가장 뛰어난 성능을 보였던 LightGBM 과 다른 머신러닝 알고리즘을 비교 분석할 예정이다.

ACKNOWLEDGMENT

This research was supported in part by the Next-Generation Information Computing Development Program through NRF funded by the Ministry of Science and ICT (no. NRF-2017M3C4A7083676) and in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2018R1A6A1A03025109) and in part by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1A2C1006249).

참 고 문 헌

[1] V. N. Index, C. Vni, C. Vni, V. N. I. Forecast, H. Tool, and R. Rate, “Cisco 비주얼 네트워킹 인덱스 2017 ~ 2022 년 전망 및 추세,” 2017.

[2] J. Yan and J. Yuan, “A Survey of Traffic Classification in Software Defined Networks,” in *Proceedings of 2018 1st IEEE International Conference on Hot Information-Centric Networking, HotICN 2018*, Jan. 2019, pp. 200–206, doi: 10.1109/HOTICN.2018.8606038.

[3] L. Jun, Z. Shunyi, L. Yanqing, and Z. Zailong, “Internet traffic classification using machine learning,” *Proc. Second Int. Conf. Commun. Netw. China, ChinaCom 2007*, vol. 10, no. 2, pp. 239–243, 2007, doi: 10.1109/CHINACOM.2007.4469372.

[4] “labeled-network-traffic-flows-114-applications @ www.kaggle.com.” [Online]. Available:https://www.kaggle.com/jsrojas/labeled-network-traffic-flows-114-applications.

[5] L. Peng, B. Yang, and Y. Chen, “Effective packet number for early stage internet traffic identification,” *Neurocomputing*, vol. 156, pp. 252–267, 2015, doi: https://doi.org/10.1016/j.neucom.2014.12.053.