TRUSTEE 를 이용한 설명가능한 네트워크 트래픽 분류 모델 분석

윤성주, *이연준
한양대학교 컴퓨터공학과 바이오인공지능융합전공, *한양대학교 컴퓨터공학과

tjdwn77777@hanyang.ac.kr, yeonjoonlee@hanyang.ac.kr

Analysis of explainable Network Traffic Classification Model using TRUSTEE

Yoon Sung Ju, Lee Yeon Joon
Major in Bio Artificial Intelligence, Department of Computer Science and Engineering, Hanyang University, *Department of Computer Science and Engineering, Hanyang University

요 약

본 논문에서는 네트워크 트래픽을 분류하는 머신러닝 모델을 구현하고 TRUSTEE 를 이용하여 머신러닝 모델을 분석하였다. TRUSTEE 를 이용한 Decision Tree(DT)는 원본 모델과 0.99 이상의 유사도를 보였고 합리적인 깊이와 노드 수로 모델을 설명하였다.

Ⅰ. 서 론

다양한 네트워크 문제를 풀기위해 많은 네트워크 트래픽 분류 연구들이 진행되었다. 또한 최근 머신러닝과 딥러닝을 이용한 모델도 네트워크 분류 분야에서 높은 정확도를 보이고 있다. 하지만 기존 연구의 머신러닝 모델들은 어떻게 결정을 내리는지에 대한 설명을 가지고 있지 않다. 이러한 모델에 대한 설명의 부재는 실제 환경에 사용되었을 때 머신러닝 모델의 오작동을 불러일으킬 수 있고 네트워크 관리자가 이런 black-box 모델의 결정에 대한 이해 없이 모델을 사용하는 것은 위험하다. 본 논문에서는 Kaggle open dataset[1]에 K-nearest Neighbors (KNN)모델을 이용하여 분류하고 TRUSTEE[2]를 이용하여 설명 가능한 결정 트리를 만들어 어떤 피처가 머신러닝 모델 결정에 영향을 주었는지 분석하고자 한다.

Ⅱ. 본론

1. KNN 모델을 이용한 네트워크 트래픽 분류

기존 머신러닝 기반 다양한 트래픽 분류 모델 (KNN, SVM, 2D-CNN, etc.) 중 KNN 을 사용했다.

1.1 Dataset

머신러닝 및 딥러닝에 사용되는 다양한 데이터와 모델을 공유하는 Kaggle 에서 Labeled Network Traffic [1] 데이터셋을 다운로드 받아서 사용했다. 총 50 가지 feature 로 이루어져 있고 2,704,839 개의 data 로 이루어져 있다.

1.2 Preprocessing

데이터셋의 웹서비스 행의 분포가 굉장히 불균형 하여 상위 4 개의 값만 사용하였고 데이터셋이 충분히 많기 때문에 NearMiss 알고리즘을 이용하여 언더샘플링을 진행하였다. IP, 카테고리 등의 정보는 이산형 변수로 바꾸어 주기 위하여 Label Encoding 을 진행하였다.

데이터셋의 피처는 크게 네트워크 헤더정보와 통계적 흐름 정보로 나눌 수 있다. 모든 피처를 이용하여 모델을 학습시킨 결과 0.48 정도의 낮은 정확도를 보였다. 따라서 [그림 1]처럼 총 세개의 데이터셋으로 나누어 모델 학습을 진행하였다 (상관계수 계산을 통하여 피처를 선택한 경우, 통계적 흐름 피처를 이용한 경우, 그 외 피처를 이용한 경우).

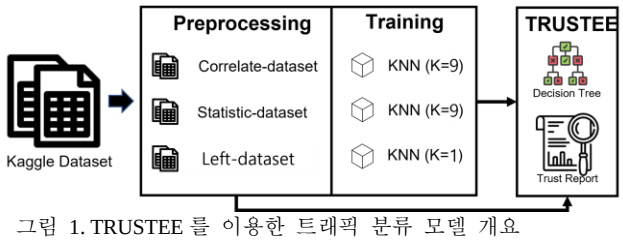


그림 1. TRUSTEE 를 이용한 트래픽 분류 모델 개요

1.3 Training

모델 학습에는 KNN 모델을 사용했다. 각 데이터셋에 대하여 k=1~10 인 KNN 을 학습하여 가장 높은 정확도를 보인 모델을 사용하였다. 상관관계 데이터셋에서는 k=9, 통계적 흐름 데이터셋은 k=9, 나머지 데이터셋에서는 k=1 에서 높은 성능을 보였다.

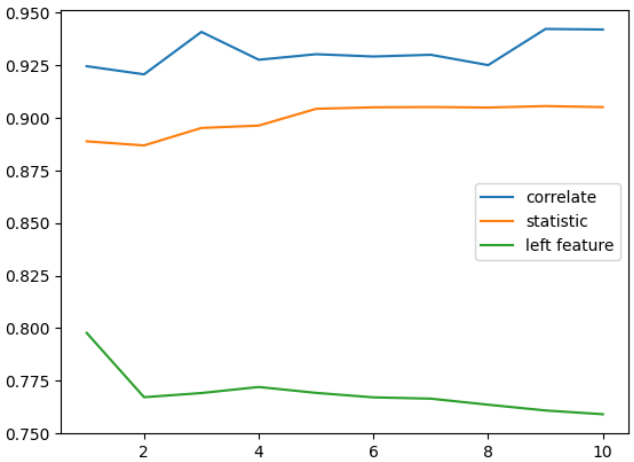


그림 2. KNN 의 K=1~10 범위에서 정확도 그래프

2.TRUSTEE 를 이용한 설명가능한 결정 트리 생성

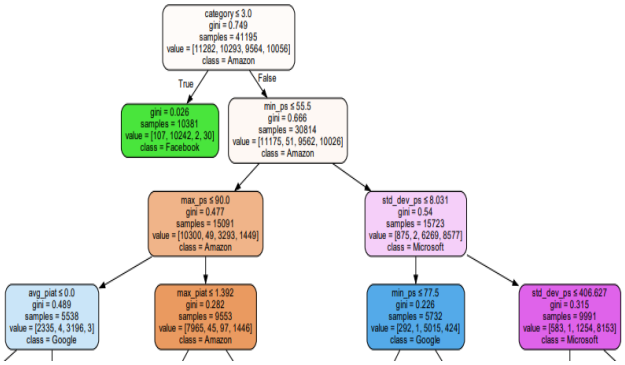
2.1 TRUSTEE

TRUSTEE[2]는 모든 black-box 모델에 적용 가능하고 기존 모델과의 높은 유사도를 가지면서 안정적이고 이해하기 쉬운 DT 를 만드는 것을 목표로 한다. 이를 위해 입력 데이터를 여러 번 샘플링하고 CART 알고리즘을 이용하여 DT 를 훈련시킨다. 기존 모델과의 유사도를 R-squared value 로 계산하여 이들 중 가장 높은 유사도를 갖는 DT 를 선택한다.

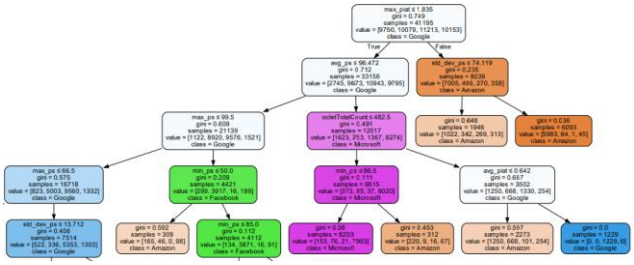
CART 알고리즘은 DT 를 생성하여도 오버피팅이 발생하기 쉽다. 이는 높은 유사도를 보장하지만 DT 의 노드가 많아질수록 트리의 크기가 커지고 이해하기 쉬운 DT 를 만들기 힘들게 한다. TRUSTEE 에서는 Top-k pruning 을 통해 이를 개선하였다.

2.2 TRUSTEE 의 적용 결과

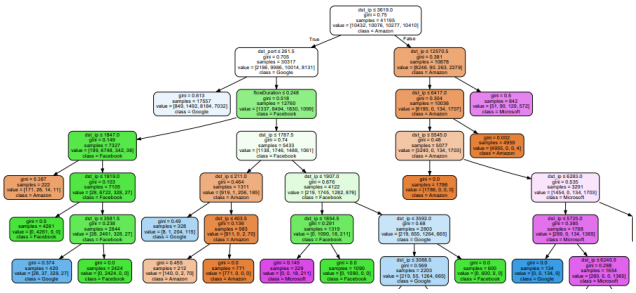
2.2.1 상관관계 데이터셋



2.2.2 통계적 흐름 데이터셋



2.2.3 나머지 데이터셋



총 4 개의 웹서비스를 분류하였고 색이 진할수록 한 웹서비스만 분류되었음을 의미한다. 위 결과는 전체 DT 중 일부를 캡쳐 한 결과이다. 유사도는 차례대로 0.96, 0.97, 0.75 이었으며 상관관계 데이터셋과 통계적 흐름 데이터셋은 DT 의 깊이와 노드 수가 적고 설명이 용이했으나 나머지 데이터셋은 DT 의 깊이가 깊고 노드 수가 많아 모델 이해가 어려웠다.

3. 결정에 주요한 영향을 끼친 feature

3.1 상관관계 데이터셋

카테고리, 어플리케이션 프로토콜, 최대/최소 패킷 사이즈, 패킷 사이 시간 등이 주요한 영향을 끼쳤다. 카테고리 값이 SocialNetwork 인 경우 Facebook 으로 분류되었다. 최소 패킷 크기 55 바이트 보다 작은 경우 Amazon 으로 분류되었고 패킷 크기 표준편차가 8.031 보다 작으면 Google, 크면 Microsoft 로 분류하는 경향을 보였다. DT 의 깊이가 깊지 않아도 낮은 gini 값을 보였고 각 웹서비스의 특성을 잘 보여주었다.

3.2 통계적 흐름 데이터셋

최대 패킷 사이 시간, 패킷 크기 표준편차, 가 큰 영향을 끼쳤다. Amazon 의 패킷 사이 시간이 큰 경향을 보였고 평균 패킷 크기가 96.432 바이트보다 크면 Microsoft 로 분류하는 경향을 보였다. Google 과 Facebook 은 최소 패킷 크기가 작은 경우 Google 최대 패킷 크기가 큰 경우 Facebook 으로 분류하는 경향을 보였다.

Google 과 Microsoft, Amazon 을 구분하는 과정에서 DT 의 깊이가 깊어 졌지만 sample 수가 작은 구분이었고 다른 node 의 gini 값이 작았기 때문에 충분히 모델을 잘 설명한 DT 라고 할 수 있다.

3.3 Left feature

도작 포트, 도작 ip 등의 정보가 DT 에 영향을 주었지만 DT 의 깊이가 깊어져도 gini 값이 어느정도 높은 값을 유지했고 도착 포트나 도착 ip 등의 정보로는 DT 를 만들 때 나누어야 할 범위가 넓어 DT 가 복잡하게 만들어졌다.

Ⅲ. 결론

본논문에서는 악의적 트래픽 발견모델을 설명하기위해 사용되었던 TRUSTEE 를 트래픽 분류 모델에 적용시켰다. 상관관계 데이터셋 기반 분류 모델과 통계적 흐름 데이터셋 기반 분류모델의 기준을 잘 설명하였다.

TRUSTEE 를 이용하여 분류 모델의 피쳐와 피쳐 간의 경계 값을 알아낼 수 있고 이를 이용하여 모델의 분류 방식 설명이 가능해 black-model 을 실제 사용하는데 있어 큰 도움이 될 것으로 생각된다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2022R1F1A1074999)을 받아 수행된 연구임

참 고 문 헌

[1] Kaggle Network Traffic Dataset, https://www.kaggle.com/datasets/jsrojas/labeled-network-traffic-flows-114-applications

[2] Jacobs, Arthur S., et al. "AI/ML for Network Security: The Emperor has no Clothes." Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022.