2026년 한국전자파학회 동계종합학술대회 논문집 Vol. 8, No. 1 2026. 2. 25~28

메타데이터 기반 신규 특징을 활용한 암호화 네트워크 트래픽 분류

°이승원*, 하준서*, 강정민*, 최양서**
*고려대학교 세종캠퍼스 인공지능사이버보안학과, **한국전자통신연구원
sangcsy@korea.ac.kr, cujs1106@korea.ac.kr, jmkang@korea.ac.kr, yschoi92@etri.re.kr

I. 서론

최근 네트워크 트래픽의 암호화 비중이 급격히 증가함에 따라, 패킷의 페이로드 내용을 직접 분석하는 심층 패킷 검사(Deep Packet Inspection, DPI) 기반 트래픽 분류 기법은 적용상의 한계에 직면하고 있다. TLS 및 VPN과 같은 암호화 기술은 사용자 프라이버시 보호 측면에서는 효과적이지만, 네트워크 보안 관점에서는 트래픽의 의미적 분석을 어렵게 만드는 요인으로 작용한다. 이러한 환경 변화로 인해, 암호화된 트래픽에서도 활용 가능한 메타데이터 기반 트래픽 분석 기법이 대안으로 주목받고 있다[1].

메타데이터 기반 접근 방식은 패킷의 길이, 전송 방향, 시간 간격 등 암호화 이후에도 관측 가능한 정보만을 활용하여 트래픽을 분석한다. 기존 연구들에서는 이러한 메타데이터를 기반으로 다양한 통계적 특징을 정의하고, 머신러닝 기법을 적용하여 트래픽 분류 성능을 평가하였다. 그러나 다수의 연구에서 비교적 많은 수의 특징을 사용함으로써, 특징 간 중복성 증가 및 모델 복잡도 상승이라는 한계가 존재한다[2].

본 논문에서는 암호화 네트워크 트래픽 환경에서 소수의 핵심 메타데이터 특징만으로도 효과적인 트래픽 분류가 가능한지를 검증하는 데 초점을 둔다. 이를 위해 기존 연구에서 공통적으로 사용된 특징 집합을 정리하고, 플로우의 동작 특성을 보다 직접적으로 반영할 수 있는 신규 특징을 정의한다. 제안한 특징셋의 유효성은 CIC-IDS 2017(이하 CIC-IDS) 데이터셋과 ISCX VPN-nonVPN 2016(이하 VPN-nonVPN) 데이터셋을 대상으로 기존 특징셋과의 분류 성능 비교 및 특징 중요도 분석을 통해 검증한다.

II. 본론

2.1 데이터셋

본 연구에서는 제안한 특징셋의 일반성과 유효성을 검증하기 위해 성격이 상이한 두 개의 공개 암호화 트래픽 데이터셋을 사용하였다. CIC-IDS 2017 데이터셋은 실제 네트워크 환경을 모사하여 수집된 침입 탐지용 데이터셋으로, 정상 트래픽과 다양한 공격 시나리오(DDoS, DoS, Brute Force, Botnet 등)를 포함한다. 해당 데이터셋은 공격 트래픽의 동작 특성과 초기 연결 패턴이 비교적 뚜렷하게 나타난다는 특징을 가진다.

반면, ISCX VPN-nonVPN 2016 데이터셋은 동일한 애플리케이션 트래픽을 VPN과 non-VPN 환경에서 수집한 데이터셋으로, 암호화 수준과 터널링 여부에 따른 트래픽 특성 차이를 분석하는 데 적합하다. 이 데이터셋은 명시적인 공격 트래픽보다는 암호화 방식에 따른 트래픽 패턴 차이를 중심으로 구성되어 있어, 특징의 일반화 성능을 평가하는 데 유용하다.

이와 같이 서로 다른 목적과 구조를 가진 두 데이터셋을 활용함으로써, 본 연구에서는 제안한 특징셋이 특정 데이터셋에 종속되지 않고 다양한 암호화 트래픽 환경에서도 효과적으로 작동하는지를 검증하고자 하였다.

2.2 기존 메타데이터 특징

기존 암호화 트래픽 분류 연구에서는 주로 플로우 단위의 통계적 특징을 활용해 왔다. 본 연구에서는 선행 연구들에서 공통적으로 높은 사용 빈도를 보인 다음의 7개 특징을 기준 특징(baseline feature)으로 선정하였다.

- **Flow Duration**: 플로우의 시작부터 종료까지의 전체 지속 시간
- **Total Fwd Packets**: 송신 방향 패킷의 총 개수
- **Total Backward Packets**: 수신 방향 패킷의 총 개수
- **Flow Bytes/s**: 플로우 전체 바이트 수를 지속 시간으로 나눈 값
- **Flow Packets/s**: 플로우 전체 패킷 수를 지속 시간으로 나눈 값
- **Fwd IAT Mean**: 송신 방향 패킷 간 평균 시간 간격
- **Bwd IAT Mean**: 수신 방향 패킷 간 평균 시간 간격

이들 특징은 트래픽의 전송량, 속도, 방향성 및 시간적 분포를 요약하는 데 효과적이며 암호화 여부와 무관하게 계산 가능하다는 장점이 있다. 그러나 플로우 내부의 동적 변화, 비대칭성, 또는 간헐적 통신 패턴과 같은 세부적인 동작 특성을 충분히 반영하지 못하는 한계가 존재한다.

2.3 신규 메타데이터 특징 정의

본 연구에서는 플로우의 동작 특성을 보다 정밀하게 표현하기 위해 총 6개의 신규 특징을 정의하였다.

(1) Burstiness Index (BI)

$$BI = \frac{\sigma_{IAT}}{\mu_{IAT}}$$

여기서 σ_{IAT}와 μ_{IAT}는 각각 인터패킷 시간의 표준편차와 평균값을 의미한다. BI는 패킷 전송이 일정한지 또는 특정 구간에 집중되는지를 정량적으로 나타낸다.

(2) Flow Direction Ratio (FDR)

$$FDR = \frac{N_{fwd}}{N_{bwd} + \epsilon}$$

N_{fwd}와 N_{bwd}는 각각 송·수신 패킷 수이며, ϵ은 분모가 0이 되는 상황을 방지하기 위한 작은 상수이다. FDR은 플로우의 방향성 비대칭성을 표현한다.

(3) Active/Idle Time Ratio (AIR)

$$AIR = \frac{T_{active}}{T_{idle} + \epsilon}$$

AIR는 활성 구간(active) 시간의 평균을 비활성 구간(idle) 시간의 평균으로 나눈 값으로, 연결의 지속성 및 간헐적 통신 특성을 반영한다.

(4) Burstiness Ratio (BR)

$$BR = \frac{\max(T_{active})}{\mu(T_{active})}$$

BR은 활성 구간들의 지속 시간 분포에서 최대 활성 구간이 평균 대비 얼마나 큰지를 나타내며, 순간적인 트래픽 집중 현상을 정량화한다.

(5) Handshake Symmetry (HS)

$$HS = \log\left(\frac{W_{fwd}}{W_{bwd} + \epsilon}\right)$$

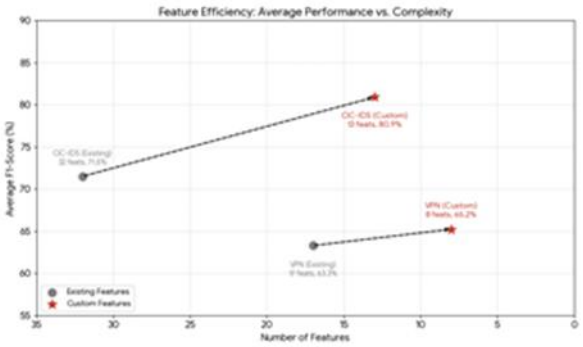
W_{fwd}와 W_{bwd}는 TCP 세션 초기 핸드셰이크 구간에서 관측된 송·수신 윈도우 크기의 합을 의미한다. HS는 초기 연결 단계에서의 비대칭성을 표현한다.

(6) Payload Shape Score (PSS)

$$PSS = \frac{\sigma_{len}}{\mu_{len}}$$

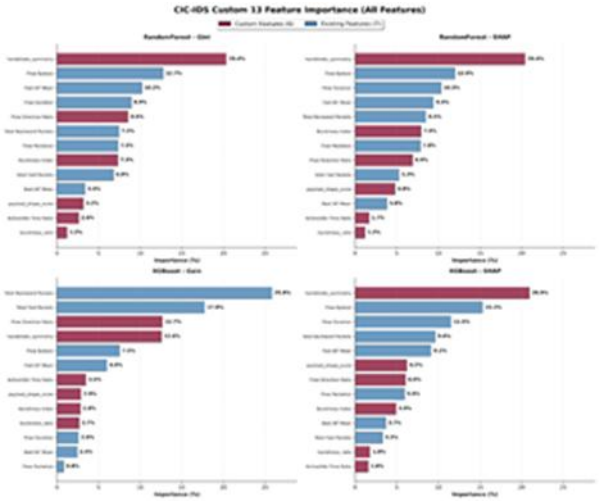
여기서 σ_{len}과 μ_{len}은 패킷 길이의 표준편차와 평균값이다. PSS는 암호화 및 패딩으로 인해 발생하는 패킷 길이 분산 특성을 간접적으로 반영한다. 이는 애플리케이션 특성이나 암호화 방식에 따라 나타나는 패킷 길이의 규칙성 차이를 반영한다.

제안한 특징셋의 유효성을 검증하기 위해 CIC-IDS 및 VPN-nonVPN 데이터셋을 대상으로 기존 특징셋과의 분류 성능을 비교하였다. CIC-IDS 데이터셋에는 Random Forest 모델과 XGBoost 모델을 사용하였고 VPN-nonVPN 데이터셋에는 Decision Tree, KNN, Random Forest와 XGBoost 모델을 사용하여 실험을 수행하였으며, 평가 지표로 평균 F1-score를 사용하였다.



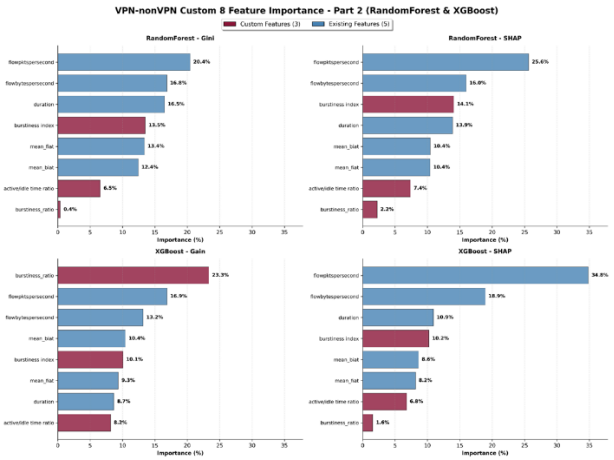
[그림 1] 특징셋 효율성 비교

그림 1은 실험 결과로 특징 수 대비 평균 F1-score를 나타낸 그래프이다. CIC-IDS 데이터셋에서 13개의 커스텀 특징셋은 기존 32개 특징셋 대비 평균 F1-score를 71.5%에서 80.9%로 향상시켰다. VPN-nonVPN 데이터셋에서도 8개의 커스텀 특징셋이 기존 17개 특징 대비 평균 F1-score를 63.3%에서 65.2%로 개선하였다.



[그림 2] 특징 중요도 분석(CIC-IDS)

그림 2는 CIC-IDS 데이터셋을 대상으로 커스텀 특징셋 13개에 대해 Random Forest 및 XGBoost 모델을 적용한 특징 중요도 분석 결과를 나타낸다. 신규 제안 특징 중 Handshake Symmetry가 두 모델에서 공통적으로 가장 높은 중요도를 보였고 Burstiness Index, Flow Direction Ratio 등이 상대적으로 높은 중요도를 보였다. 기존 특징 중 Flow Bytes/s, Total Backward Packets, Fwd IAT Mean, Flow Duration 등이 높은 중요도를 보였다.



[그림 3] 특징 중요도 분석(VPN-nonVPN)

그림 3은 VPN-nonVPN 데이터셋을 대상으로 커스텀 특징셋 8개에 대해 Random Forest 및 XGBoost 모델을 적용한 특징 중요도 분석 결과를 나타낸다. 신규 제안 특징 중 Burstiness Index가 높은 중요도를 보였다. 기존 특징 중 Flow Packets/s가 두 모델에서 공통적으로 가장 높은 중요도를 보였고 Flow Bytes/s, Flow Duration 등이 상대적으로 높은 중요도를 보였다.

2.4 실험 결과 및 분석

실험 결과, 두 데이터셋 모두에서 신규로 정의한 특징들이 기존 통계적 특징들과 함께 높은 중요도를 보이며 분류 성능 향상에 기여함을 확인하였다. 특히 CIC-IDS 데이터셋에서는 Handshake Symmetry가 높은 중요도를 보였는데, 이는 공격 트래픽이 초기 연결 단계 및 비정상적인 패킷 전송 패턴에서 뚜렷한 차이를 보이기 때문으로 해석된다. VPN-nonVPN 데이터셋에서는 신규 특징 중 Burstiness Index가 상위 중요도로 나타나, 암호화 및 터널링 환경에서도 플로우 동작 특성을 효과적으로 포착할 수 있음을 보여주었다.

이러한 결과는 제안한 신규 메타데이터 특징들이 단순히 기존 특징을 보조하는 역할에 그치지 않고, 암호화 트래픽 분류에서 핵심적인 판별 정보로 기능할 수 있음을 실험적으로 입증한다. 나아가, 데이터셋의 특성에 따라 중요하게 작용하는 특징이 달라질 수 있음을 확인함으로써, 암호화 트래픽 분석에서는 도메인 특성을 고려한 맞춤형 특징 설계가 필요함을 시사한다.

III. 결론

본 논문에서는 암호화 네트워크 트래픽 환경에서 메타데이터 기반 분석을 적용하여, 기존 연구 대비 적은 수의 커스텀 특징셋을 설계하고 그 유효성을 검증하였다. 기존 공통 특징과 신규로 정의한 6개의 특징을 결합한 특징셋은 CIC-IDS 및 VPN-nonVPN 데이터셋 실험에서 평균 F1-score 기준으로 기존 특징셋 대비 향상된 분류 성능을 보였다. 이는 암호화 네트워크 트래픽 분류에서 단순 통계 기반 특징을 넘어, 플로우 동작 특성을 직접 반영한 메타데이터 특징 설계가 효과적인 대안이 될 수 있음을 시사한다.

본 연구는 공개 데이터셋 기반 실험에 한정되어 있으며, 실제 운영 환경 트래픽에 대한 검증은 향후 연구로 남긴다.

Acknowledgements

본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2026-RS-2022-00164800). 또한 2025년도 정부(교육부, 과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. 25411243, No. NRF-00252157)이며, 2025년도 산업통상자원부 및 한국산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임(RS-2025-02317769).

참고문헌

[1] 최양서 외, "네트워크 이상행위 탐지를 위한 암호 트래픽 분석기술 동향", 전자통신동향분석, 제38권 제5호, 71-80쪽, 2023년 10월

[2] Bruno Reis, Eva Maia, Isabel Praça, "Selection and Performance Analysis of CICIDS2017 Features Importance", *Foundations and Practice of Security*, pp 56-71, Apr. 2020.