섀넌 엔트로피를 통한 트래픽 암호화 판별

김주성[1], 장윤성[1], 박재원[1], 유경민[1], 백의준[2], 김명섭[3]
[1]고려대학교 컴퓨터정보학과 석사과정
[2]고려대학교 컴퓨터정보학과 박사과정
[3]고려대학교 세종캠퍼스 컴퓨터융합소프트웨어학과 교수
jsung0514@korea.ac.kr, brave1094@korea.ac.kr, 2018270614@korea.ac.kr,
rudals2710@korea.ac.kr, pb1069@korea.ac.kr, tmskim@korea.ac.kr

Determining Traffic Encryption with Shannon Entropy

Ju-Sung Kim[1], Yoon-Seong Jang[1], Jae-Won Park[1], Gyeong-Min Yu[1],
Ui-Jun Baek[1], Myung-Sup Kim[2]
[1]Dept. of Computer and Information Science, Korea University
[2]Dept. of Computer Convergence Software, Korea University

요 약

이 논문에서는 네트워크 관리와 보안에서 중요한 응용 트래픽 분류 기술의 한계를 해결하기 위해 섀넌 엔트로피를 활용한 암호화 여부 판단 방법을 제안한다. 기존의 포트 기반 및 페이로드 기반 분류 방식은 암호화된 트래픽 증가로 인해 정확도가 떨어지며, 머신러닝/딥러닝 기법 또한 암호화된 데이터 학습에 어려움을 겪고 있다. 이러한 문제를 극복하기 위해 섀넌 엔트로피를 활용한 암호화 판단 방법을 제안하였고, 프로토콜 분석 없이도 암호화 구간을 95%의 높은 정확도로 식별하였다.

1. 서론

응용 트래픽 분류는 네트워크상에서 발생하는 다양한 유형의 트래픽을 특정 응용 프로그램 또는 서비스에 따라 분류하는 기술로, 네트워크 관리와 보안에 있어 매우 중요한 역할을 한다. 네트워크 관리자들은 트래픽을 정확히 분류함으로써 자원을 효율적으로 할당하고, 서비스 품질을 유지하며, 잠재적인 보안 위협을 미리 감지할 수 있다. 특히, 인터넷의 급속한 발전과 함께 온라인 서비스와 응용 프로그램의 다양성이 증가함에 따라 응용 트래픽 분류는 더욱 복잡하고 중요한 과제로 떠오르고 있다.

응용 트래픽 분류 방법에는 포트 기반, 페이로드 기반, 그리고 머신러닝/딥러닝 기반의 방법이 있다[1]. 포트 기반 방법은 각 응용 프로그램에 할당된 고유 포트를 통해 트래픽을 분류하지만, 현대의 응용 프로그램들은 임의의 포트를 사용하거나 포트 번호를 위장하는 경우가 많아 정확도가 떨어진다. 페이로드 기반 방법은 패킷의 내용을 직접 분석하여 트래픽을 분류하는 방식이지만, 정보보호의 중요성이 높아짐에 따라 대부분의 네트워크 트래픽이 암호화 프로토콜을 사용해 페이로드가 암호화됨으로써 더 이상 유효하지 않게 되었다. 이러한 한계를 극복하기 위해 머신러닝과 딥러닝 기법이 도입되어, 암호화된 트래픽의 패턴을 학습하여 더 정교한 분류가 가능하게 되었으나, 무작위의 난수로 이루어진 암호화 데이터는 머신러닝이나 딥러닝 모델이 데이터셋의 특징을 학습하는데 악영향을 주기 때문에 여전히 암호화된 트래픽 환경에서는 기존 분류 방법들이 한계를 보이고 있어 새로운 접근 방식이 필요하다. 이에 본 논문에서는 섀넌 엔트로피를 통하여 네트워크 트래픽 데이터의 암호화 여부를 판단하는 방법을 제안한다.

2. 관련연구

2.1 암호화 트래픽

TLS(Transport Layer Security)는 인터넷상에서 데이터 전송의 기밀성과 무결성을 보장하기 위해 널리 사용되는 암호화 프로토콜이다[2]. TLS는 HTTP, SMTP, FTP와 같은 애플리케이션 계층 프로토콜 위에서 작동하며, 전송되는 데이터를 암호화하여 도청, 변조, 위조를 방지한다. TLS는 다양한 버전이 있으며, 각 버전은 이전 버전에서 발견된 보안 취약점을 해결하고 데이터 보호 기능을 강화한다. TLS는 대칭 암호화, 비대칭 암호화, 그리고 해시 함수와 같은 다양한 암호화 기법을 결합하여 데이터의 기밀성과 무결성을 유지하며, 이를 통해 네트워크상에서 안전한

<표 1> 전처리 전후 데이터셋 통계 비교

Method	# of Packets			# of Flows			Capacity
	TCP	UDP	Total	TCP	UDP	Total	
Raw-datasets	8,770,334	12,629,084	21,399,418	7,300	302,889	310,189	28 GB
Pre-processed	343,871	11,896,803	12,240,674	5,086	5,077	10,163	19.8 GB

통신을 보장한다.

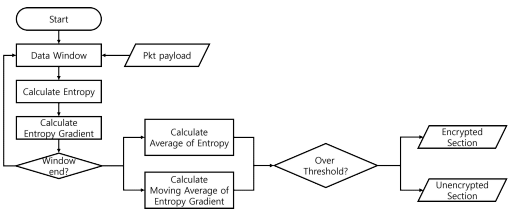
2.2 섀넌 엔트로피

섀넌 엔트로피(Shannon Entropy)는 정보 이론에서 데이터를 분석하는 데 중요한 개념으로, 데이터의 불확실성이나 예측 불가능성을 측정하는 데 사용된다. Claude Shannon이 제안한 엔트로피는 통계적 특성을 기반으로 한 데이터의 정보량을 나타내며, 이를 통해 데이터의 무작위성이나 복잡성을 정량적으로 평가할 수 있다[3]. 섀넌 엔트로피는 데이터에서 각 값의 발생 확률을 기반으로 계산되며, 확률 분포가 고르게 분포된 데이터는 높은 엔트로피를 가지며, 특정 값에 집중된 데이터는 낮은 엔트로피를 가진다. 다음은 섀넌 엔트로피를 계산하는 수식이다.

$$H(x) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \quad (1)$$

3. 본론

본 장에서는 섀넌 엔트로피를 활용하여 네트워크 트래픽 데이터의 암호화 여부를 판단하는 방법을 설명한다. 그림1은 제안하는 방법론의 알고리즘을 순서도로 표현한 것이다.



(그림 1) 방법론 알고리즘 순서도

먼저, 각 패킷에 대한 엔트로피를 계산하는 절차를 도입하였다. 섀넌 엔트로피는 데이터의 불확실성을 측정하는 중요한 지표로, 값이 높을수록 해당 데이터가 무작위적임을 나타낸다. 주어진 문자열 데이터에 대해 각 문자의 빈도를 기반으로 엔트로피를 계산하며, 이를 통해 패킷의 통계적 특성을 정량적으로 평가할 수 있다. 슬라이딩 윈도우 기법을 사용하여 패킷 데이터를 일정한 길이의 구간으로 나누고, 각 구간에 대해 섀넌 엔트로피를 계산한다. 이러한 방식은 패킷의 전체 데이터를 단일 값으로 분석하는 대신, 데이터의 시간적 흐름에 따른 변화를 파악할 수 있도록 한다. 이 과정에서 이동평균을 계산하여 엔트로피의 변화량을 더욱 명확히 분석하였다. 이동 평균은 각 구간의 엔트로피 변화율을 부드럽게 처리하여, 단순한 변동이 아닌 실제적인 패턴 변화를 포착하는 데 도움을 준다. 이후, 각 구간의 엔트로피 값과 이동평균 값의 차이가 설정된 임계값을 초과하는 경우 이를 암호화된 구간으로 판단하고 플래그 처리하도록 하였다.

4. 실험

4.1 데이터셋

실험에 사용된 데이터셋은 ISCX VPN 2016으로 암호화된 응용 트래픽을 모아둔 공공 데이터셋이다[4]. 해당 데이터셋은 패킷 단위로 구성된 파일 형태이며, 16개의 응용 프로그램, 6개의 응용 카테고리, 2개의 VPN/NonVPN으로 구성되어 있다.

4.2 전처리

공개 데이터셋 전처리 방법에 따라 패킷 단위의 pcap 파일을 입력으로 패킷 헤더의 5-tuples 정보가 같은 플로우 단위의 트래픽 파일로 변환한다[5]. TCP 플로우의 경우 핸드셰이크 과정이 무결점인 플로우만, UDP 플로우의 경우 패킷 개수가 5개 이상인 플로우만 사용하였고. 최종적으로 10,163개의 플로우, 12,240,674개의 패킷을 얻었다. 표 1은 전처리 전후로 데이터셋의 통계 정보를 비교한 표이다.

4.3 평가 지표

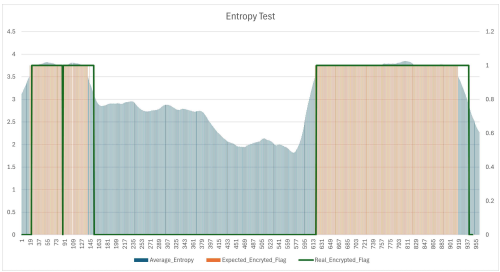
해당 실험에서는 결과의 평가를 위해 정확도(Accuracy)를 사용한다. 다음은 평가 지표를 계산하는 수식이다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

True Positive(TP)는 실제 True인 정답을 True라고 예측(정답)하는 것이다. False Positive(FP)는

실제 False인 정답을 True라고 예측(오답)하는 것이다. False Negative(FN)는 실제 True인 정답을 False라고 예측(오답)하는 것이다. 마지막으로 True Negative(TN)는 실제 False인 정답을 False라고 예측(정답)하는 것이다. 이 실험에서 True는 암호화된 구간, False는 암호화되지 않은 구간을 말한다.

4.3 실험 결과



(그림 2) 엔트로피 테스트

그림2는 특정 패킷을 대상으로 제안한 방법론으로 찾은 암호화 구간과 실제 암호화 구간을 비교한 그래프이다. 푸른색 그래프는 평균 엔트로피값, 주황색 그래프는 제안한 방법론으로 찾은 암호화 구간, 초록색 그래프는 실제 암호화 구간을 나타낸다. 데이터셋 전체 패킷을 대상으로 정확도 비교 실험을 진행한 결과 평균 95%의 정확도를 기록하였고, 제안한 방법론으로 실제 난수로 이루어진 필드를 높은 정확도로 찾아낼 수 있음을 확인할 수 있었다.

5. 결론

응용 트래픽 분류는 네트워크 관리와 보안에서 필수적인 기술로, 네트워크상에서 발생하는 다양한 트래픽을 특정 응용 프로그램이나 서비스에 따라 분류하는 역할을 한다. 이를 통해 네트워크 관리자는 자원을 효율적으로 배분하고 서비스 품질을 유지하며, 잠재적인 보안 위협을 조기에 감지할 수 있다. 인터넷의 발전과 함께 다양한 온라인 서비스와 응용 프로그램이 등장하면서 응용 트래픽 분류는 더욱 복잡해졌으며, 특히 암호화된 트래픽의 증가로 인해 기존의 분류 방법들은 한계를 보이고 있다.

본 논문은 머신러닝/딥러닝 기반의 네트워크 트래픽 분류 방법이 무작위 난수로 이루어진 암호화 데이터로 데이터셋의 특징을 학습하는데 악영향을 받기 때문에 보이는 한계를 극복하고자 섀넌 엔트로피를 활용하여 네트워크 트래픽 데이터의 암호화 여부를 판단하는 방법을 제안하였다. 이를 통해 프로토콜 분석 없이 95%의 높은 정확도로 암호화 구간을 찾아낼 수 있었다. 향후 연구로는 제안한 섀넌 엔트로피 기반 암호화 여부 판단 방법의 성능을 다양한 암호화 프로토콜과 네트워크 환경에서 더욱 확장하여 검증하는 연구를 진행할 예정이다.

Acknowledgement

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이며 (00235509, ICT융합 공공 서비스●인프라의 암호화 사이버위협에 대한 네트워크 행위기반 보안관제 기술 개발) 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과이고 (2021RIS-004) 본 논문은 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력 기반 지역혁신 사업 (2021RIS-004)과 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원(P0024177, 2023년 지역혁신클러스터육성)을 받아 수행된 연구입니다.

참고문헌

[1] 이민성, et al. "순차적인 데이터 처리를 통한 딥러닝 기반 트래픽 분류속도 개선." 한국통신학회논문지 47.12 (2022): 2096-2103.

[2] Apostolopoulos, George, Vinod Peris, and Debanjan Saha. "Transport Layer Security: How much does it really cost?." IEEE INFOCOM'99. Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. The Future is Now (Cat. No. 99CH36320). Vol. 2. IEEE, 1999.

[3] Shannon, Claude Elwood. "A mathematical theory of communication." The Bell system technical journal 27.3 (1948): 379-423.

[4] Gerard Draper-Gil, Arash Habibi Lashkari, Mohammad Saiful Islam Mamun, and Ali A. Ghorbani. 2016. Characterization of Encrypted and VPN Traffic using Time-related Features. In the International Conference on Information Systems Security and Privacy. 407-414.

[5] Baek, Ui-Jun, et al. "Preprocessing and Analysis of an Open Dataset in Application Traffic Classification." 2023 24st Asia-Pacific Network Operations and Management Symposium (APNOMS). IEEE, 2023.