암호화 트래픽 분류를 위한 통계적 특징 추출 방법

김주성, 유경민, 박지태, 백의준, 김명섭

고려대학교

{jsung0514, rudals271, pjj5846, pb1069, tsmkim}@korea.ac.kr

Statistical feature extraction method for classifying encrypted traffic

Ju-Sung Kim, Gyeong-Min Yu, Jee-Tae Park, Ui-Jun Baek, Myung-Sup Kim

Korea University

요 약

인터넷의 성장과 다양한 서비스로 인해 트래픽 분류의 중요성이 증가했고, 정보 보호가 중요해지면서 암호화 통신의 발달이 이루어졌다. 이에 따라 기존 트래픽 분류 방법의 한계가 드러나면서 데이터 외부의 특성을 통해 분류하는 머신러닝 기반의 분류 방법이 중요해졌다. 본 논문은 머신러닝을 통한 암호화 트래픽의 분류를 위해 데이터셋에서 통계적 특징을 추출하는 방법과 해당 특징 중 트래픽 분류에 중요한 역할을 하는 특징을 선택하는 방법을 제안하였다.

Ⅰ. 서 론

인터넷의 발전과 서비스 다양화로 인해 네트워크 트래픽의 양상이 점점 복잡해지고 있다. 이러한 변화는 네트워크 자원의 효율적인 사용과 트래픽 관리 전략에 새로운 도전을 제시하며, 특히 보안 및 서비스 품질의 유지를 위해 정밀한 트래픽 분류 기술의 필요성을 강조한다. 이와 동시에 정보 보호의 중요성이 증가하면서 암호화 통신의 사용이 대중화되고 있으나, 기존의 트래픽 분류 방식은 암호화 트래픽의 분석에 한계를 보이고 있다. 이에 본 논문은 암호화된 응용 트래픽을 대상으로 응용을 구분하기 위한 특징을 추출하고, 이를 효과적으로 선별하여 머신러닝을 기반으로 한 트래픽 탐지 기술에 적용하기 위한 특징 추출 방법을 제안한다.

본 논문은 1장 서론에 이어 2장에서 관련 연구를 검토하고, 3장에서 통계적 특징 추출 및 선택 방법을 설명한다. 4장에서는 실험 결과를 설명하고 5장에서 결론 및 향후 연구를 제시하는 것으로 본 논문을 마친다.

Ⅱ. 관련 연구

본 장에서는 통계적 특징을 학습하여 분류 모델을 생성하는 머신러닝 기반의 트래픽 분류 방법을 설명한다. 암호화 통신의 발달로 기존 트래픽 분류 방법은 한계를 보였다. 이를 극복하기 위해 통계적 특징을 학습한 머신러닝 기반 분류 방법을 사용한 다양한 연구들이 있다.

[1]에서는 Auckland-vi-20010611, Auckalnd-vi-20010612, Leipzig-ii-20030221, NZIX-ii-20000706 데이터셋에서 6가지의 특징 선택 알고리즘으로 각각 22개의 통계적 특징을 추출하였고 AdaBoost와 C4.5를 조합하여 99.2%의 분류 정확도를 나타내었다. [2]에서는 Auckland IV, Calgary 데이터셋에서 각각 11개의 통계적 특징을 선택하였고 AutoClass 클러스터링을 사용하여 92.4%의 분류 정확도를 나타내었다. [3]에서는 Wide 데이터셋을 앙상블 기법을 사용하여 17개의 통계적 특징을 추출하였고 K-Means 클러스터링을 사용하여 95%의 분류 정확도를 나타내었다.

본 논문은 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과 (2021RIS-004) 이며 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(00235509, ICT융합 공공 서비스•인프라의 암호화 사이버위협에 대한 네트워크 행위기반 보안관제 기술 개발)

카테고리	세션 분포		
	전처리 전	노이즈	전처리 후
Chat	13,955	13,519	436
Email	8,059	7,489	570
File Transfer	70,297	67,329	2,968
P2P	477	117	360
Streaming	2,535	1,785	750
VoIP	214,434	210,820	3,614
총합	**309,757**	**301,059**	**8,698**

표 1. 카테고리 별 전처리 전후 세션 데이터 분포

[4]에서는 자체적으로 수집한 데이터셋에서 11개의 통계적 특징을 선택하였고 WFNP를 사용하여 96%의 분류 정확도를 나타내었다.

Ⅲ. 본 론

본 장에서는 데이터셋에서 추출 및 선택한 특징을 학습하여 응용 트래픽을 분류하는 머신러닝 기반의 시스템을 설명한다.

A. 데이터셋

데이터셋은 ISCX 2016 VPN-nonVPN 데이터셋을 사용하였으며 [5]에서 사용한 전처리 과정을 거쳐 관련 없는 프로토콜과 TCP 3 handshake가 없는 세션 등을 제거하였다. 이후, 전처리 과정을 거친 데이터셋을 대상으로 6개의 응용 카테고리 분류 실험을 진행한다. 표 1은 응용 카테고리별로 전처리 과정 전후의 세션 데이터 분포를 정리한 것이다.

B. 특징 추출

전처리 과정을 거친 응용 트래픽을 입력으로 세션 내 Packet Size, Inter Arrival Time, Payload에 대한 15가지의 통계적 특징을 추출하고, 이를 패킷 방향성, 사용 패킷 개수 등을 고려하여 945개의 통계적 특징을 추출하였다. 그리고 시간과 관련된 특징, 헤더 필드, 벡터 값 등의 특징을 추가로 추출하여 최종적으로 총 1,204개의 특징을 추출하였다. 표 2는 추출한 특징들을 정리한 것이다.

C. 특징 선택

본 연구에서는 특징 선택 알고리즘으로 RFECV를 사용하였다. 특징 추출의 결과를 교차 검증을 통해 중요도가 낮은 특징을 반복적으로 제거하

Section	Feature	Direction	Num of Packets	Statistic Items	Count
Field	Header Field	-	-	-	29
	Flag Count				9
Scalar1	Packet Size	All Forward Backward	All 5^{th}, 10^{th}, 15^{th} 20^{th}, 25^{th}, 30^{th}	Sum, Max, Min, Variance Arithmetic Mean, Geometric Mean, Sample Standard, Population Standard Inter Quantile Range, Skewness Kurtosis, Range	945 (3×3×7×15)
	Inter Arrival Time				
	Payload				
Scalar2	Packets per sec	All Forward Backward	All	-	21 (7×3)
	Bytes per sec				
	Ratio of Packets				
	Ratio of Bytes				
	Packet Count				
	Duration				
	Mean Packet Arrived Time				
Vector	Packet Size Distribution	-	100^{th}	-	200 (2×100)
	IAT Distribution				
Total Number of Features					**1,204**

표 2. 추출 특징 구조

고 모델 성능 향상에 기여한 특징을 선별하였고, 최종적으로 총 406개의 특징을 선택하였다.

D.분류

분류 모델은 XGBoost를 사용하였다. 추출 및 선택한 특징을 입력으로 모델의 하이퍼 파라미터는 표 3과 같이 사용하였고 정확도가 가장 높게 나온 파라미터를 사용하였다.

Parameter	범위	단위
Min Child Weight	5 ~ 10	1
Max Depth	1 ~ 20	1
Colsample by Tree	0.1 ~ 0.9	0.1
Colsample by Level	0.1 ~ 0.9	0.1
N_estimators	5 ~ 50	5

표 2. 카테고리 별 전처리 전후 세션 데이터 분포

Ⅳ. 실험 결과

표 4는 동일한 데이터셋과 분류 모델을 적용하였을 때의 분류 정확도를 정리한 것이다. 다른 연구들에서 사용한 특징들보다 본 연구에서 제안한 방법으로 추출한 특징을 사용하였을 때 더 높은 성능을 보이는 것을 확인하였다.

Ref. No.	# of Feat	Accuracy	Rank
[2]	16	0.841	5
[3]	11	0.848	4
[4]	**17**	**0.89**	**2**
[5]	16	0.841	6
Proposed	1,204	0.86	3
	406	**0.917**	**1**

표 4. 동일 조건에서 다른 연구와의 성능 비교

그림 1은 특징 카테고리 별로 트래픽 분류에 기여한 정도를 평균으로 정리한 결과이다. 특히 Vector의 경우 PSD에서 99번째 값이 중요도가 높게 나왔는데, 이는 VoIP와 같은 elephant flow를 구분할 수 있기 때문에 유의미하다고 분석된다.

Ⅴ. 결 론

본 논문은 전처리 되지 않은 데이터셋을 사용하여 데이터 불균형을 반영하지 못하거나 일부 특징만 사용하여 모델 과적합이 발생하는 머신러

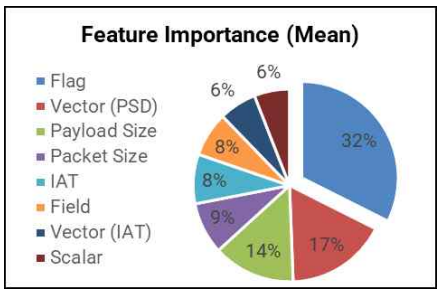

그림 1. 평균에 따른 특징 중요도 비율

닝 기반의 트래픽 분류 방법의 문제를 해결하기 위해 데이터셋에서 노이즈를 제거하고 다양한 통계적 특징을 추출하는 방법을 제안하였고 높은 분류 정확도를 이끌어 내었다. 향후 연구로 제안한 방법을 딥러닝 기반의 분류 방법과 조합하여 활용할 수 있는 방법을 연구할 계획이다.

참 고 문 헌

[1] Williams, Nigel, and Sebastian Zander. "Evaluating machine learning algorithms for automated network application identification." (2006).

[2] Erman, Jeffrey, Martin Arlitt, and Anirban Mahanti. "Traffic classification using clustering algorithms." Proceedings of the 2006 SIGCOMM workshop on Mining network data. 2006.

[3] Glennan, Timothy, Christopher Leckie, and Sarah M. Erfani. "Improved classification of known and unknown network traffic flows using semi-supervised machine learning." Information Security and Privacy: 21st Australasian Conference, ACISP 2016, Melbourne, VIC, Australia, July 4-6, 2016, Proceedings, Part II 21. Springer International Publishing, 2016.

[4] Liu, Yang, et al. "A novel algorithm for encrypted traffic classification based on sliding window of flow's first N packets." 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). IEEE, 2017.

[5] 백의준, et al. "응용 트래픽 분류 공개 데이터셋의 전처리 방법." 한국통신학회 학술대회논문집 (2023): 63-64.