머신러닝 기반 응용 트래픽 분석을 위한 프로세스 로그 활용 데이터셋 수집 시스템

장윤성, 김주성, 박지태, 이윤서, 백의준, 김명섭*

고려대학교

{ brave1094, jsung0514, pjj5846, yoongbb, pb1069, tmskim* }@korea.ac.kr

Dataset Collecting System Using Process Logs for Machine Learning-Based Application Traffic Analysis.

Yoon-Seong Jang, Ju-Sung Kim, Jee-Tae Park, Yoon-Seo Lee,

Ui-Jun Baek, Myung-Sup Kim*

Korea Univ.

요 약

네트워크 트래픽 분류는 네트워크 관리 분야의 핵심 기술로 최근 머신러닝 기술을 적용하여 그 성능을 향상시키고 있으며, 이를 위해 다양한 공개 데이터셋이 활용되고 있다. 현재 사용 중인 대부분의 공개 데이터셋은 단일 호스트 환경에서 수집된 데이터셋이다. 이러한 데이터셋으로 학습된 모델은 해당 호스트의 패턴에 과적합되어, 일반화 성능이 저하될 수 있다. 따라서, 모델이 사용자의 패턴이 아닌 응용 프로그램 자체에 대한 특징을 더 잘 학습할 수 있도록 다양한 호스트의 패턴을 포함하는 다중 호스트 환경에서 수집된 데이터셋을 사용해야 한다. 본 논문에서는 프로세스 로그에서 추출한 네트워크 연결 정보를 통해 필터링하는 방법으로 데이터셋 수집 시스템을 구축함으로써, 기존 데이터셋의 단일 호스트 수집 특성이라는 제약을 해결하고 다양한 응용 트래픽 수집에 적용이 가능한 수집 방법을 제안하여 다중 호스트 환경의 데이터셋의 구축을 돕고자 한다.

Ⅰ. 서 론

트래픽 분류는 네트워크 보안, 성능 최적화, 서비스 품질 관리 등 다양한 네트워크 관리 및 보안 목적을 위해 중요한 역할을 하는 분야이다. 이러한 트래픽 분류는 최근 머신러닝 기술을 적용한 형태로, 복잡한 패턴을 스스로 학습하고 더 높은 정확도와 효율성을 제공할 수 있도록 연구되어 왔다. 이러한 상황에서 데이터셋은 머신러닝 기반 트래픽 분류 연구에 있어 핵심적인 구성 요소 중 하나로, 모델 학습 및 평가를 위한 핵심 자원이다.

Ⅱ. 관련 연구

현재, 모델을 평가하는 데이터셋으로 공개 데이터셋 [1] ISCX VPN-non VPN 2016, [2] Tor-nonTor 2016 등을 사용하는데, 이것들은 대부분 단일 호스트 환경에서 수집된 데이터셋(이하 단일 호스트 데이터셋)이다. [3]에서는 연구에서 이러한 단일 호스트 데이터셋보다 다중 호스트 환경에서 수집된 데이터셋(이하 다중 호스트 데이터셋)으로 학습된 모델의 더 높은 정확도를 결과로 보고하였다. 또한, [4]에서는 IDS(Intrusion Detection System) 분야에서의 단일 호스트 데이터셋의 한계점으로 '여러 좀비 호스트를 사용하는 분산 공격에 대한 고려가 없다'는 점을 제시하였다. 본 논문에서는 프로세스 로그를 활용한 데이터셋 수집 시스템을 구축함으로써, 각각의 응용 프로그램마다 다양한 호스트가 발생시킨 데이터를 수집하도록 하여, 기존 단일 호스트 데이터셋의 한계점을 극복하고, 효율적인 다중 호스트 데이터셋의 수집을 돕고자 한다.

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이며(No.RS-2023-00230661, 하이브리드 양자키분배 방법 및 망 관리 기술 표준개발), 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(00235509, ICT융합 공공 서비스●인프라의 암호화 사이버위협에 대한 네트워크 행위기반 보안관제 기술 개발)

Ⅲ. 본 론

데이터셋은 수집 환경에 따라, 단일 호스트 환경(Single-host)과 다중 호스트 환경(Multi-host)으로 구분할 수 있다. 단일 호스트 데이터셋으로 학습시킨 분류 모델은 해당 호스트에 특정되는 패턴을 학습할 가능성이 있는데, 이러한 분류 모델에서는 과적합이 발생하고, 일반화 성능이 떨어지게 된다. 따라서 호스트에 의존적인 특징이 아닌 응용 프로그램 자체에 대한 특징을 더 잘 학습할 수 있도록 다중 호스트 데이터셋을 사용해야 한다. 본 논문에서는 이러한 다중 호스트 환경 데이터셋을 확보하기 위해, 프로세스 로그 정보를 활용한 트래픽 데이터셋 수집 시스템을 제시한다. 그림 1은 전체 구성도를 나타내며, 시스템 운용을 위한 환경 설정과 시스템 작동의 구체적인 과정은 다음과 같다.

환경 설정 :

1) 수집할 호스트들을 연결하는 스위치에서 패킷 미러링을 통해 모든 호스트에서 발생하는 트래픽을 저장하는 Traffic Collection Server가 필요하다. 저자는 소속 연구팀에서 Flow Generator 도구를 사용하여, 미러링한 패킷을 Traffic Collection Server에 저장하도록 하였다. Flow Generator 도구는 패킷 미러링을 구현하기 위해 scapy 라이브러리를 활용하여 미러링한 패킷들을 .pcap 파일 형식으로 저장하도록하는 도구로, 파이썬 프로그램으로 개발되었다.
2) 특정 응용 프로그램에 대한 네트워크 연결 로그만을 저장하도록 하는 설정이 필요하다. 'Windows 이벤트 뷰어'는 모든 시스템, 보안 및 응용 프로그램 이벤트를 로깅하고 보여주는 관리 도구이지만, 본 시스템에서는 네트워크 연결 정보만을 추출하여 저장해야 한다. 따라서, 본 논문에서는 'Windows Sysmon'을 사용하였다. 이 도구의 'sysmonconfig-export.xml' 설정 파일에서 각 응용 프로그램의 실행파일의 경로

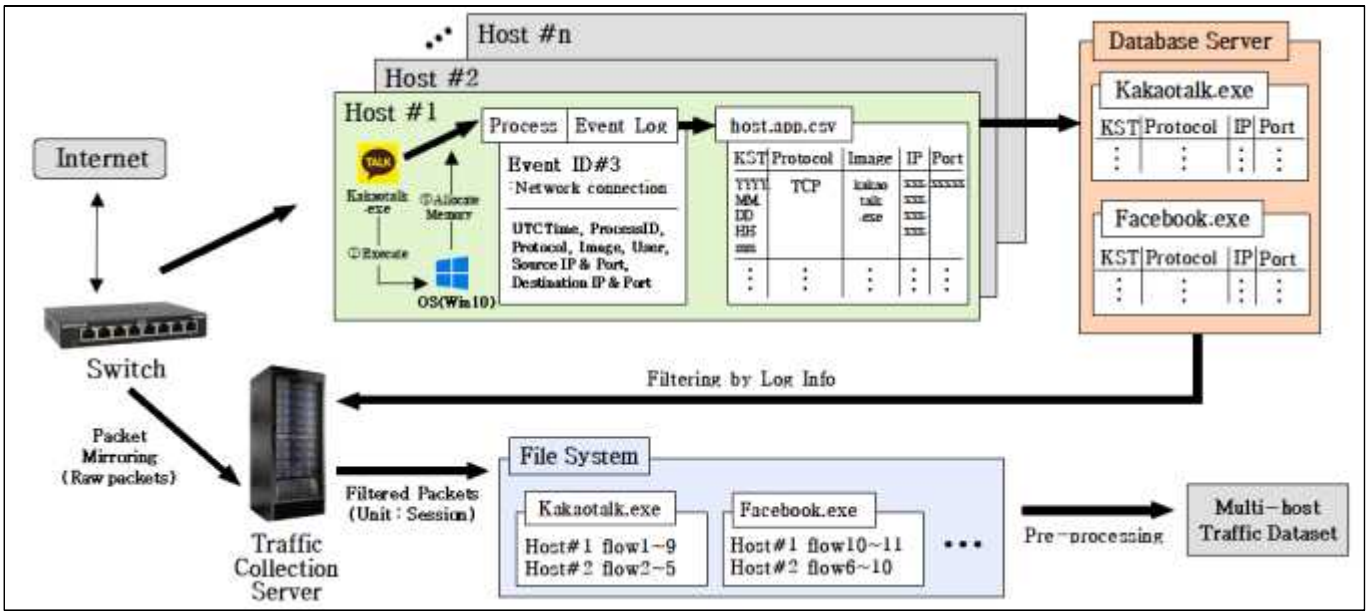

그림 1. 프로세스 로그 활용 다중 호스트 환경 트래픽 데이터셋 수집 시스템 구성도

를 설정하면, 해당 응용 프로그램이 프로세스로 등록되었을 때부터 해제될 때까지의 모든 네트워크 연결 정보를 저장할 수 있다.

수집 과정 :

1) 먼저, 스위치를 통해 연결된 여러 개의 호스트에서, 특정 응용 프로그램이 실행되면, 운영체제에 의해 메모리를 할당받아 프로세스가 생성된다. 이 때에, 'Windows Sysmon'을 사용하여 다양한 네트워크 연결 정보를 저장한다.
2) 주요 네트워크 연결 정보를 각 Host PC에 임시 저장한다. 이는 갑작스러운 통신 장애나 패킷 누락 시에 정상적으로 복구하는 데에 사용하기 위함이다. 본 시스템에서 사용하는 주요 정보는 UTC Time, Protocol, Image, IP, Port 5가지이며, UTC Time은 한국 표준시로 변환하여 사용하고, Image는 해당 응용 프로그램명을 식별하기 위해 사용한다. IP, Port, Protocol 정보는 실질적으로 패킷을 필터링하여 저장하기 위해 필요한 정보이다.
3) 과정 2에서 생성된 파일에 저장된 네트워크 연결 정보들은 DB Server에 저장되고, 이 때에 Image 정보에 의해 응용 프로그램명을 식별하고, 응용 프로그램별로 서로 다른 테이블에 저장된다. 저장된 테이블은 KST Time, Protocol, IP, Port 4가지 정보를 저장하게 된다.
4) Traffic Collection Server에 미러링 되어 저장된 패킷들 중에서 KST 정보와 일치하는 시간대부터 시작하여 저장된 Protocol, IP, Port정보로 필터링하여 File System에 저장한다.
5) 저장된 파일들에 사용 모델에 맞는 적절한 전처리 기법을 적용하여, 다중 호스트 데이터셋을 구축한다.

본 논문에서 제시하는 이 시스템은 2가지 장점을 가진다. 첫 번째로, 다양한 응용 프로그램에 대한 수집에 적용이 가능하다. 한 번에 한 가지의 응용 프로그램만을 사용하여 트래픽을 수집하던 기존의 방법과는 달리, 동시에 다양한 응용 프로그램에 대한 로그를 저장하는 것이 가능하다. 두 번째, 동시에 다중 호스트에서 발생한 트래픽 데이터를 수집할 수 있다. 다양한 호스트에서 발생한 네트워크 연결 로그를 DB 서버에 모두 저장하므로, 이를 통해 필터링 시에 모두 추출할 수 있다.

Ⅳ. 결 론

본 논문에서는 머신러닝 기반 응용 트래픽 분석을 위해 프로세스 로그를 활용한 데이터셋 수집 시스템을 제시한다. 이 시스템은 프로세스 별로 발생한 네트워크 연결 정보를 DB 서버에 저장한 뒤, 이를 바탕으로 스위치에서 미러링한 대량의 트래픽에서 필터링함으로써, 다중 호스트 환경에서 사용된 다양한 응용 트래픽을 동시에 수집하는 방법이다. 이를 통해, 모델이 사용자의 패턴보다 응용 프로그램 자체의 특징을 더 잘 학습하도록 하는 다중 호스트 데이터셋의 구축을 돕고자 한다. 본 연구에서 드러난 한계점은 수집과 필터링을 동시에 진행하지 못한다는 것과, 현재 캡슐화 분류, 서비스 분류, 사용자 행위나 기능 분류에 대해 자동 레이블링이 불가하다는 것이다. 따라서 수집과 필터링을 동시에 실행 가능한 새로운 트래픽 수집 기법에 대한 연구와 다양한 Task에 대한 레이블링 방법에 대한 연구를 후속 연구과제로 제시한다.

참 고 문 헌

[1] Gerard Drapper Gil, Arash Habibi Lashkari, ohammad Mamun, Ali A. Ghorbani, "Characterization of Encrypted and VPN Traffic Using Time-Related Features", In Proceedings of the 2nd International Conference on Information Systems Security and Privacy(ICIS SP 2016), pages 407-414, Rome, Italy.

[2] Arash Habibi Lashkari, Gerard Draper-Gil, Mohammad Saiful Islam Mamun and Ali A. Ghorbani, "Characterization of Tor Traffic Using Time Based Features", In the proceeding of the 3rd International Conference on Information System Security and Privacy,

[3] Kim, Taehoon, and Wooguil Pak. 2023. "Integrated Feature-Based Network Intrusion Detection System Using Incremental Feature Generation" Electronics 12, no. 7: 1657. (https://doi.org/10.3390/electronics12071657)

[4] Kwon, Junhyung, Byeonggil Jung, Hyungil Lee, and Sangkyun Lee. 2022. "Anomaly Detection in Multi-Host Environment Based on Federated Hypersphere Classifier" Electronics 11, no. 10: 1529. (https://doi.org/10.3390/electronics11101529)