ApriSig: A-priori 알고리즘 활용 데이터 기반 침입 탐지용 서명 자동 생성 프레임워크

박서현*, 황명하, 권유진, 이현우**

*KENTECH (학부생), **KENTECH (교수), 한국전력공사 전력연구원

ApriSig: Data-based Automated Signature Generation Framework Leveraging A-priori Algorithm for Intrusion Detection

Seohyeon Park*, Myeong-Ha Hwang, YooJin Kwon, Hyunwoo Lee**

*KENTECH (Undergraduate student), **KENTECH (Faculty), KEPRI

요 약

서명 기반 네트워크 침입탐지시스템은 전문가들이 침입에 대한 분석을 바탕으로 생성한 서명을 이용하여 네트워크 패킷으로부터 침입을 탐지한다. 서명 기반 탐지는 일반적으로 높은 정밀도를 보이지만, 새로운 침입에 대한 서명을 생성하기 위해서는 많은 인력과 시간이 투입되어야 한다는 단점이 있다. 본 논문에서는 데이터를 기반으로 자동으로 서명을 생성하는 프레임워크인 ApriSig을 제안한다. ApriSig은 여러 특징들로 이루어진 테이블 데이터셋에서 A-priori 알고리즘을 통해 특징들을 클러스터링하고, 이를 바탕으로 서명을 자동 생성한다. 미라이 봇넷 데이터셋을 바탕으로 실험한 결과 신뢰도 90%의 A-priori 알고리즘을 바탕으로 생성한 서명 중 100%의 정밀도를 보이는 서명 5개로 100%의 재현율을 달성함으로써 ApriSig의 유용성을 보였다.

I. 서론

사이버 위협이 나날이 증대하면서 네트워크 침입탐지시스템은 외부의 침입으로부터 네트워크를 보호하는데 유용한 도구로 사용되고 있다[1, 2]. 현업에서 침입탐지시스템은 주로 서명 기반으로 침입을 탐지하는데, 여기서 서명이란 침입과 관련된 패킷을 탐지하기 위한 규칙이다. 예를 들어, 탐지 규칙은 “HTTP 헤더의 길이가 99999이면 침입”이라는 식으로 서술된다. 이러한 서명기반 시스템은 높은 정확도를 보이지만, 새로운 서명을 생성하는데 많은 인력과 시간이 들어간다는 단점이 있다[3].

이러한 단점을 극복하기 위해 본 논문에서는 ApriSig이라는 A-priori 알고리즘[4]을 활용하여 침입 데이터를 분석하고 이를 바탕으로 자동으로 탐지용 서명을 생성하는 프레임워크를 제안한다. ApriSig은 여러 특징(feature)들로 구성된 레이블이 있는 데이터셋을 입력으로 받아, 클러스터링 알고리즘인 A-priori를 사용하여 레이블별로 연관된 특징들을 모아낸다. 그리고 공격과 관련된 클러스터의 특징들을 활용하여 서명을 생성해 낸다. 우리는 미라이 봇넷 데이터셋[5, 6]에 대해 ApriSig을 수행하였다. 그 결과로 생성된 서명 중 100%의 정밀도를 보이는 서명 5개로 100%의 재현율을 달성함으로써 프레임워크의 유용성을 입증하였다.

II. 배경 이론 - A-priori 알고리즘

A-priori 알고리즘은 클러스터링 알고리즘으로 특징들 간의 연관성을 만들어낸다. 예를 들어, 소비자의 소비 패턴을 분석한다고 할 때, 품목 A를 사는 사람이 품목 B도 함께 “유의미하게” 구매하는지를 판단하여 연관성을 찾는 알고리즘이다. 여기서 유의미의 기준이 되는 값을 신뢰도라고 한다. 알고리즘은 신뢰도가 보장되는 하에서 3개 이상의 품목도 묶어낼 수 있다.

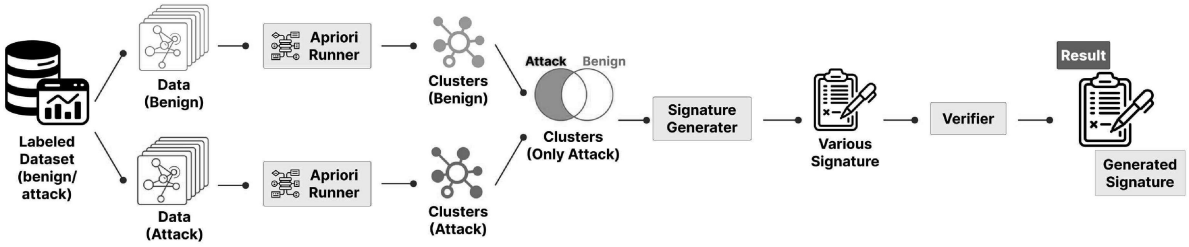

그림 1 ApriSig 프레임워크 전체 구조

III. ApriSig: 자동 서명 생성

본 논문에서는 서명 생성을 머신러닝을 활용하여 자동화하는 프레임워크인 ApriSig을 제안한다(그림 1 참조).

ApriSig은 “침입”과 “정상”이라는 레이블을 가진 데이터셋을 입력으로 받아 서명 리스트를 출력한다. 여기서 입력 데이터셋은 여러 개의 특징으로 이루어진 테이블 데이터셋이다. 이를 통해 서명을 생성하는 과정은 아래와 같다.

① 입력 데이터셋을 침입과 관련된 데이터셋과 정상과 관련된 데이터셋으로 분류한다.

② Apriori Runner는 각 데이터셋 별로 A-priori 알고리즘을 돌려서 서로 연관성이 있는 특징들을 묶어낸다.

③ 침입 데이터셋의 특징 묶음들 중 정상 데이터셋의 특징 묶음에 속한 묶음들은 걸러낸다.

④ Signature Generator는 모든 묶음에 대해 서명을 생성한다.

⑤ Verifier는 최초 데이터셋에 대해 서명의 성능을 평가하면서 특정 기준 이상의 성능을 보이는 서명들을 모아 최종 서명들을 출력한다.

이렇게 생성한 서명들은 침입탐지시스템에 적용되어 네트워크 패킷들을 분석하는데 사용하게 된다.

IV. 실험

우리는 ApriSig을 구현하여 미라이 봇넷 침입 데이터셋[5, 6]을 바탕으로 ApriSig를 평가하였다. 실험에 사용된 데이터셋은 미라이 봇넷 pcap 파일에 대해 각 패킷 별로 레이블을 달고, 매 1초 동안의 패킷들 간의 관계(예를 들어, 초당 바이트 수)를 정량화한 41개의 플로우 특징들로 구성되어 있다.

Itemset	Accuracy	Precision	Recall
1	45.97%	100%	44.61%
2	45.97%	100%	44.61%
3	45.97%	100%	44.61%
4	43.51%	100%	43.52%
5	43.51%	100%	43.52%
Average		100%	

표 1 ApriSig가 생성한 상위 5개 서명

우리는 A-priori 알고리즘의 신뢰도를 90%로 설정하여 돌렸고, Signature Generator는 41개의 서명을 생성하였다. 생성된 서명 각각에 대해 정확도(accuracy), 정밀도(precision), 재현율(recall)을 살펴보았다. 서명은 침입을 정확히 탐지해야 하기 때문에 재현율보다 정밀도가 더 중요하다. 부족한 재현율은 이를 보완하기 위한 추가적인 서명으로 채울 수 있지만, 정밀도가 떨어진다면 오탐으로 인해 가용성이 떨어지게 되기 때문이다.

표 1은 생성된 서명들 각각을 이용하여 데이터셋에 적용한 결과 중 정밀도 기준으로 상위 5개에 대한 성능을 보여준다. 각 서명에서 A-priori 알고리즘을 통해 묶여진 특징들은 다음과 같다(번호는 표의 Itemset 번호에 해당).

① 패킷 길이의 평균, 초당 양방향 패킷의 개수, 패킷 길이의 최소값, 패킷 길이의 최대값, 전송 계층 프로토콜

② 전체 패킷 개수, 패킷 길이의 평균, 초당 일방향 패킷 수, 초당 양방향 패킷 수, 패킷 길이의 최소값, 패킷 길이의 최대값, 전체 헤더 길이, 전송 계층 프로토콜

③ 패킷 길이의 평균값, 초당 양방향 패킷 개수, 패킷 길이의 최대값, 전송 계층 프로토콜

④ 일방향 패킷 간 거리의 평균, 양방향 패킷 간 거리의 평균, 양방향 패킷간 거리의 표준편차, 일방향 패킷간 거리의 표준편차

⑤ 초당 양방향 패킷 수

위 표에서 볼 수 있듯이, 상위 5개의 서명은 모두 정밀도를 100%를 보였으나 재현율은 43.52% 혹은 44.61% 수준이었다. 생성된 서명들이 상호 보완적인지 살피기 위해 우리는 이 5개 모두를 활용하여 성능을 평가하였고, 그 결과 재현율이 100%가 되는 것을 확인할 수 있었다. 이는 ApriSig으로 생성한 서명들이 유용하다는 것을 보여준다.

V. 결론 및 후속 연구 방향

본 논문에서 우리는 A-priori 알고리즘을 이용하여 주어진 데이터셋에서 자동으로 서명을 생성하는 ApriSig을 제안하였다. 실험을 통해 실현 가능성을 보였지만, ApriSig은 다음 두 가지의 한계를 가진다.

첫째, ApriSig은 레이블이 있는 데이터셋을 입력으로 받기 때문에 인력과 시간을 완전히 줄이지 못한다. 침입 데이터에 대한 레이블을 갖추는 작업은 여전히 대량의 인력과 시간을 요구한다. 이에 따라 ApriSig은 원인 분석에 필요한 시간만 단축한다. 후속 연구로서, 레이블이 없는 데이터셋에 대해서 자동으로 서명을 생성하는 프레임워크로 발전시키고자 한다.

둘째, A-priori 알고리즘은 각 벡터의 특징들이 있다/없다만 판정하기 때문에 여러 값들 혹은 실수 범위의 특징에 대해서는 동작하지 않는다. 후속 연구로서, 보다 엄밀한 서명을 생성하기 위해서, 원핫(one-hot)인코딩으로 테이블 데이터셋의 벡터를 확장하거나 범위를 기준으로 벡터를 확장하여 서명을 생성하고자 한다. 예를 들어, “데이터 크기”라는 특징이 있다면, 현재 설계에서는 이를 데이터 크기가 0인 경우와 데이터 크기가 있는 경우로만 나눈다. 우리의 계획은 이를 “0이상 200 미만”, “200 이상 400 미만” 등으로 값을 기준으로 세분화하여 A-priori 알고리즘을 적용하여 보다 엄밀한 서명을 생성하는 것이다.

본 연구에서 제안한 프레임워크는 사이버 위협 헌팅[7]을 수행하는 기관들에 매우 유용한 도구로 사용될 수 있을 것으로 기대된다.

Acknowledgment

본 연구는 한국전력공사의 2022년 착수 기초연구개발 과제 연구비에 의해 지원되었음(과제번호: R24XO01-3)

[참고문헌]

[1] Ahmad, Zeeshan, et al. "Network intrusion detection system: A systematic study of machine learning and deep learning approaches." Transactions on Emerging Telecommunications Technologies 32.1 (2021): e4150.

[2] Lee, Hyunwoo, et al. "An infection-identifying and self-evolving system for iot early defense from multi-step attacks." European Symposium on Research in Computer Security. Cham: Springer Nature Switzerland, 2022.

[3] Aulia Novi, et al. “Automated Detection of Network Intrusions Using Machine Learning in Real-Time Systems”, International Journal of Computer Technology and Science, Volume. 1, No. 2, April 2024

[4] Agrawal, R. "Fast Algorithms for Mining Association Rules." VLDB, 1994.

[5] Kang, H., et al. “IoT Network Intrusion Dataset,” https://ieee-dataport.org/open-access/iot-network-intrusion-dataset, 2019

[6] IoTEDef Dataset, https://github.com/iotedef/iotedef-dataset, 2022

[7] Sqrrl Data, I: A framework for cyber threat hunting, Whitepaper, 2012