IDS 검출 성능 개선을 위한 효과적인 라벨 노이즈 검출 및 정정 기술 연구

전유하 [1], 이경호 [2], 백명선 [2]

1 세종대학교 전자정보통신공학과, 2 세종대학교 지능정보융합학과

yuhaj613@naver.com, gyeongho@sejong.ac.kr, msbaek@sejong.ac.kr

Study on Effective Label Noise Detection and Correction Techniques to improve IDS Performance.

Yuha Jeon[1], Gyeong Ho Lee[2], Myung-Sun Baek[2]
1 Department of Information and Communication Engineering
2 Department of Artificial Intelligence and Information Technology

요 약

침입 탐지 시스템(IDS)의 성능은 학습 데이터의 라벨 품질에 크게 의존한다. 그러나 실제 네트워크 환경에서 수집된 IDS 데이터는 자동화된 규칙 기반 라벨링이나 제한된 전문가 개입에 의해 생성되는 경우가 많아 라벨 노이즈 문제가 빈번하게 발생한다. 특히 암호화 트래픽 환경에서는 페이로드 접근이 제한되어 라벨 오류를 사전에 식별하고 정정하는 과정이 더욱 중요해진다. 본 연구에서는 노이즈가 포함된 라벨 데이터를 대상으로 멀티모달 기반 NoiseGPT 와 구조 기반 라벨 정정 기법인 RAPIER 를 결합한 데이터 정제 파이프라인을 제안한다. 제안한 방법은 Flow 단위 특징 데이터를 시각화하여 NoiseGPT 를 통해 라벨 노이즈를 탐지하고, 노이즈로 판단된 샘플에 대해서만 RAPIER 를 적용함으로써 라벨 정정을 수행한다. 제안된 방식은 라벨 노이즈 정정을 통해 신뢰도 높은 학습 데이터셋을 구축할 수 있다.

Ⅰ. 서 론

네트워크 이상 탐지 분야에서 트래픽 분석은 전통적으로 페이로드(payload) 기반 기법에 의존해 왔다. 그러나 HTTPS 및 TLS 와 같은 암호화 기술의 확산으로 인해 실제 네트워크 환경은 점차 암호화 트래픽 중심으로 전환되고 있으며, 이에 따라 페이로드 접근이 제한되는 상황이 일반화되고 있다. 이런 환경에서는 이상탐지 시스템 (IDS: Intrusion Detection System) 모델이 활용할 수 있는 정보가 제한되며, 학습 데이터의 라벨 품질이 탐지 성능에 미치는 영향이 더욱 커진다.

실제 IDS 데이터는 자동화된 규칙 기반 라벨링이나 제한된 전문가 판단에 의해 생성되는 경우가 많아, 라벨 노이즈 (라벨 오류)가 포함될 가능성이 높다. 이러한 라벨 노이즈는 학습 과정에서 성능 저하를 유발할 뿐만 아니라, 모델의 일반화 능력을 크게 저해할 수 있다. 이를 해결하기 위해 기존 연구인 RAPIER[1]는 정상 및 악성 트래픽의 확률 분포 차이를 활용하여 라벨 노이즈를 정정하는 구조 기반 프레임워크를 제안하였다. RAPIER는 라벨 노이즈가 존재하는 환경에서도 안정적인 탐지 성능을 유지하는 데 효과적임을 보였다.

최근 제안된 NoiseGPT[2]는 원본 데이터에 대해 Mixture-of-Feature (MoF) 기법을 적용하여 변형된 입력을 생성하고, 이에 대한 멀티모달 대형 언어 모델(MLLM: Multi-Modal Large Language Model)의 확률 반응을 분석함으로써 라벨 노이즈를 탐지하는 방법이다. NoiseGPT 는 단일 모델 기반 라벨 노이즈 탐지 기법에 비해, 데이터와 라벨 간의 불일치를 보다 안정적으로 포착할 수 있다는 장점을 가진다. 그러나 NoiseGPT 는 각 샘플을 독립적으로 처리하며, 샘플 간 구조적 관계를 활용하지 않는다.

이에 본 연구에서는 NoiseGPT 와 RAPIER 를 결합하여, 라벨 노이즈 탐지와 라벨 정정을 단계적으로 수행하는 IDS 데이터 정제 파이프라인을 제안한다. NoiseGPT 를 통해 라벨 노이즈를 사전에 탐지하고, 노이즈로 판단된 샘플에 대해서만 RAPIER 를 적용함으로써, 구조 기반 정정 과정에서 발생할 수 있는 오류 전파를 완화하는 것을 목표로 한다.

Ⅱ. 제안된 데이터 정제 파이프라인 구조

본 연구에서 제안하는 데이터 정제 파이프라인은 라벨이 존재하지만 일부 라벨에 노이즈가 포함된 라벨이 존재하는 IDS 데이터를 입력으로 활용한다. 전체 파이프라인은 라벨 노이즈 탐지 단계와 라벨 정정 단계로 구성되며, 각 단계는 NoiseGPT 와 RAPIER 가 담당한다.

먼저, 노이즈가 포함된 라벨 데이터를 대상으로 Feature Extraction 을 수행한 후, Flow 단위 데이터 시각화를 통해 NoiseGPT 의 입력 형식을 구성한다. NoiseGPT 는 각 샘플과 해당 라벨을 입력으로 받아, 모델의 확률 반응 변화를 분석함으로써 라벨 노이즈 여부를 판단한다. 이 단계에서 각 샘플은 Clean 또는 Noisy 로 분류된다.

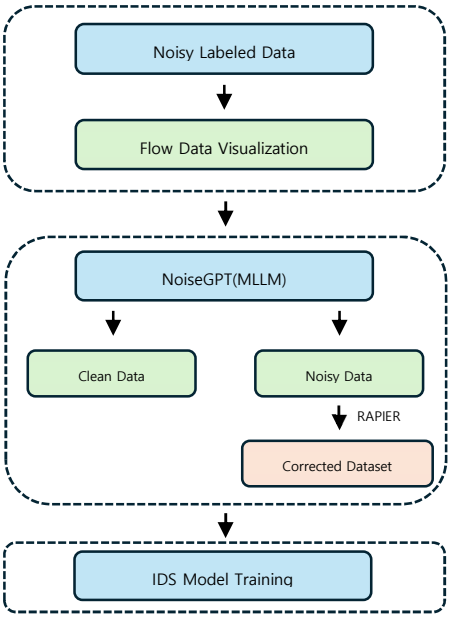

그림 1. 제안된 기술의 파이프라인

이후 Noisy 로 판단된 샘플에 대해서만 RAPIER 를 적용하여 라벨 정정을 수행한다. 반면, Clean 으로 판단된 샘플은 추가적인 수정 없이 그대로 IDS 학습 모델로 입력된다. 이러한 선택적 적용 방식은 불필요한 라벨 수정으로 인한 성능 저하를 방지하고, 구조 기반 정정의 안정성을 높이는 데 목적이 있다. 최종적으로 정제된 데이터셋을 이용하여 IDS 모델 학습을 수행한다.

Ⅲ. 플로우 단위 데이터의 visualization

기존 RAPIER 는 네트워크 트래픽을 Flow 단위의 수치 특징 벡터로 표현하고, 이를 그래프 구조 상에서 처리하여 라벨 정정을 수행한다. 반면, NoiseGPT 는 MLLM 을 기반으로 데이터와 라벨 간의 관계를 판단하므로, 수치 벡터 형태의 입력을 직접적으로 활용하기 어렵다. 따라서 본 연구에서는 Feature Extraction 이 완료된 Flow 단위 데이터를 시각적으로 변환하는 과정을 추가한다. 본 과정에서 i번째 Flow $\mathbf{x}_i$는 고정 차원의 특징 벡터로 아래와 같이 나타낼 수 있다.

$$\mathbf{x}_i = [x_{i,1},\ x_{i,2},\ \cdots,\ x_{i,F}] \in \mathbb{R}^F \quad (1)$$

위의 식에서 F 는 Flow 의 길이를 나타낸다. 이후 각 Flow 단위로 정규화를 수행한다.

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} \quad (2)$$

위의 식에서 $\mu(\mathbf{x}_i)$와 $\sigma(\mathbf{x}_i)$ 는 각각 $\mathbf{x}_i$의 평균과 분산을 의미한다. 이후 정규화된 Feature 벡터를 세로 방향으로 H번 반복하여 2 차원 행렬을 구성한다. 이때 열 방향은 feature index 를, 색상 값은 해당 feature 의 상대적 크기를 나타낸다.

$$\mathbf{H}_i = \begin{bmatrix} \hat{\mathbf{x}}_i \\ \hat{\mathbf{x}}_i \\ \vdots \\ \hat{\mathbf{x}}_i \end{bmatrix} \in \mathbb{R}^{H \times F} \quad (3)$$

이 행렬 $\mathbf{H}_i$를 heatmap 이미지로 변환하여 NoiseGPT 의 입력으로 사용한다. 이러한 시각화 방식은 Feature 간 상대적 분포를 유지하면서도, MLLM 기반 분석이 가능하도록 데이터를 변환하는 역할을 수행한다.

Ⅳ 제안된 데이터 정제 파이프라인 동작 방식

본 연구는 라벨 노이즈의 효과적 정정을 위해 NoiseGPT 기반의 라벨 노이즈 탐지 단계를 선행시키는 데 있다. NoiseGPT 는 각 샘플과 해당 라벨을 입력으로 받아, MoF 기반 변형에 대한 모델의 확률 반응 변화를 분석함으로써 라벨 노이즈 여부를 판단한다. 이 과정에서 데이터 자체를 수정하지 않고, 라벨의 신뢰성만을 평가한다.

NoiseGPT 의 출력은 각 샘플에 대한 Clean 또는 Noisy 의 이진 분류 결과이다. Noisy 로 판단된 샘플에 대해서만 RAPIER 를 적용하여 구조 기반 라벨 정정을 수행함으로써, 샘플 간 관계 정보를 활용한 추가적인 검증과 보정을 수행한다. 반면, Clean 으로 판단된 샘플은 RAPIER 적용 대상에서 제외된다.

이와 같은 결합 방식은 NoiseGPT 의 노이즈 탐지 능력과 RAPIER 의 구조 기반 정정 능력을 상호 보완적으로 활용함으로써, 단일 기법 적용 시 발생할 수 있는 한계를 완화한다. 결과적으로 제안하는 파이프라인은 라벨 노이즈로 인한 성능 저하를 줄이면서도, IDS 환경에 적합한 안정적인 데이터 정제 과정을 제공한다.

Ⅳ. 성능검증 및 논의

표 1 은 RAPIER 단독 방식과 NoiseGPT-RAPIER 결합 방식의 IDS 성능을 비교한 결과이다. 제안 방식은 Precision 을 0.280 에서 0.719 로 크게 향상시켰으며, F1-score 또한 0.435 에서 0.736 으로 개선되었다. 이는 NoiseGPT 기반 노이즈 선별을 통해 오탐을 줄이고, 공격으로 판단한 샘플의 신뢰도를 높였기 때문이다.

반면 Recall 은 0.970 에서 0.754 으로 감소하였는데, 이는 제안 방식이 보다 보수적으로 공격을 판별하면서 탐지 누락이 증가한 결과로 해석할 수 있다. 따라서 운영 환경에서는 목표에 따라 임계값 조정이 필요하다.

표 1. 기존 RAPIER 방식과 제안된 방식의 성능 비교

Method	Recall	Precision	F1-score
RAPIER-only	0.970	0.280	0.435
NoiseGPT+ RAPIER	0.754	0.719	0.736

ACKNOWLEDGMENT

이 논문은 2025 년도 정부(국방부)의 재원을 받아 정보통신기획평가원의 국방 ICT 혁신기술사업으로 수행된 연구성과입니다[No. RS-2025-02363049, 다중 다계층 네트워크의 동적 신뢰 연결 및 지능적 관제기술 개발].

참 고 문 헌

[1] Yuqi Qing, “Low-Quality Training Data Only? A Robust Framework for Detecting Encrypted Malicious Network Traffic”, Sep. 2023

[2] Haoyu Wang, “NoiseGPT: Label Noise Detection and Rectification through Probability Curvature”, Dec. 2024