RelationMesh: 설명 가능한 비지도 이상탐지 모델

김병록, 이현우*

한국에너지공과대학교 (대학원생), *한국에너지공과대학교 (교수)

RelationMesh: An Explainable and Unsupervised Anomaly Detection

Byung-Rok Kim, Hyunwoo Lee*

KENTECH (Graduate Student), *KENTECH (Faculty)

요 약

AI가 급격히 발전함에 따라 침입탐지 보안 분야에서도 인공지능 기술을 활용하려는 움직임이 많아지고 있다. 제로데이 공격을 탐지할 수 없다는 서명 기반 시스템의 단점을 보완하기 위해 AI기반 비지도 이상 탐지를 활용하려는 시도가 활발하다. 다만, 이 방식은 탐지한 공격이 어떤 규칙을 위반했는지 설명하기 쉽지 않다. 본 논문에서는 개별 샘플들의 특징들 사이의 관계성에 초점을 맞춘 비지도 이상탐지 기법인 RelationMesh를 제안한다. 이는 기존 기법에 비해 더 세밀하게 특징에 초점을 맞춘 탐지 기법이며, 이 모델 자체가 탐지의 원인이 되는 특징을 바로 보고하는 설명가능성도 제공한다. 우리는 실험을 통해 RelationMesh가 3가지 벤치마킹 데이터셋에서 IsolationForest 대비 F1-score를 29%와 오토인코더 대비 9% 향상시키는 것을 보였다.

I. 서론

침입탐지시스템은 인공지능이 널리 보급되기 이전부터 오랫동안 연구되어 왔다. 특히나 직접 입력된 규칙을 기반으로 무엇이 악성인지 판단하는 서명 기반 침입탐지시스템은 1985년에 개발된 Dorothy Denning의 Intrusion Detection Expert System (IDES) [1]를 시초로 하여 현재 널리 쓰이는 V3 같은 백신 소프트웨어에 이르기까지 지속적으로 사용되어 왔다. 하지만 이 방식은 해당 규칙에 나열되지 않은 새로운 공격(혹은 제로데이 공격)은 탐지가 불가능하다는 단점을 가진다 [2].

이러한 한계를 극복하고자 이상 탐지 시스템이 연구되어 왔다. 이는 정상 데이터만으로 정상 분포를 학습하고 이 분포에서 벗어나는 샘플을 이상이라고 판단하는 방식이다. 이러한 방식은 사전에 모르는 대상이라도 정상 분포에서 떨어져 있다면 탐지가 가능하다. 그래서 모르는 악성 공격을 탐지할 수 있는 가능성이 생긴다. 그러나 모든 이상 패턴이 악성은 아닐 수 있다는 점에서 오탐이 많다는 단점을 가진다 [2].

본 논문에서는 데이터 내의 특징들 간의 상호 관계성에 주목하여 이상 탐지 기법의 정확도를 높이는 RelationMesh (RM)를 제안한다. RM은 데이터를 구성하는 개개의 특징별로 다른 특징들과의 관계를 기반으로 학습한 의사결정나무들의 앙상블이다. 기존 이상 탐지 기법들이 전체 데이터들의 분포에 초점을 맞췄다면, RM은 개별 데이터 내의 특징들 간의 관계성까지 집중하게 되면서 정확도를 보다 높이고자 하였다. 나아가, RM은 특징별로 개별 모델이 생성되기 때문에, 개별 모델의 결과를 기반으로 설명 가능성도 제공한다.

II. 비지도 이상탐지

비지도 혹은 보다 엄밀히 반지도 이상 탐지 모델은 기본적으로 정상 데이터만을 입력받아 학습한다. 이렇게 '정상'이라는 것의 패턴을 학

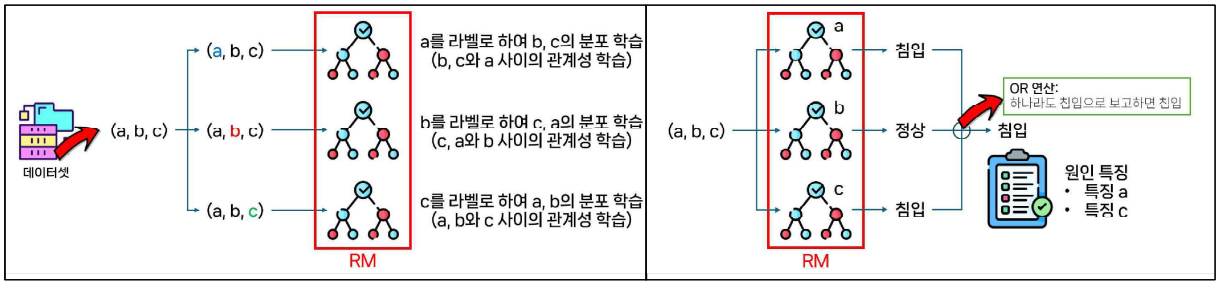

그림 1 RelationMesh 학습, 탐지 및 원인특징 보고 과정

습한 모델은 이상한 데이터가 들어왔을 때 모델은 다른 반응을 보이고, 이 반응을 분석해 이상치를 탐지하는 과정을 '비지도 이상탐지'라고 부른다. 여기서 정상성(normality)을 어떻게 정의하는지에 따라 대표적으로 분할 기반과 신경망 기반의 이상탐지 모델이 있다.

분할 기반 이상 탐지에서는 주어진 샘플을 여러 기준으로 고립시킨다. 어떤 공간에 무작위로 칸막이를 놓아 분리해 나가면서 샘플을 고립시키는데, 이상의 경우, 정상인 경우에 비해 보다 적은 수의 칸막이로 고립이 된다. 이 방식의 시초가 Isolation Forest [3]이며, 이는 여전히 현업에서 활발히 사용되고 있다.

신경망 기반 이상 탐지에서는 신경망을 통해 정상 데이터의 분포를 학습한다. 이 분류의 대표적인 기술로 오토인코더(Autoencoder) [4]가 있으며, 이는 정상 데이터를 더 작은 크기로 압축하고 다시 복원하도록 신경망을 학습시킨다. 이 신경망에 새롭게 들어오는 정상 데이터는 문제없이 복원하지만 이상 데이터는 복원 오류가 커지게 된다. 이러한 신경망 기반 이상탐지는 비지도 이상탐지의 주류가 되어 끊임없이 성능이 발전하고 있다.

위 두 가지 이상탐지 기법은 블랙박스 기법에 해당하여 모델 자체가 설명가능성을 제공하지는 않는다. 침입 탐지에서 침입의 원인을 식별하면 정확한 대응까지 나아갈 수 있다는 점에서, 설명가능성은 침입 대응을 위한 중요한 요구사항이 된다. 그러하기에 기존 기법들은 LIME이나 Shapley 같은 설명을 위한 추가 방식을 활용하여 탐지의 원인을 분석한다.

III. RelationMesh 개관

본 논문에서 우리는 의사 결정 트리에 기반한 새로운 비지도 이상탐지 기법인 RelationMesh (RM)를 제안한다.

3.1. 설계 목적

RM은 다음을 목적으로 하여 설계되었다.

1) **탐지 성능 개선**: RM은 기존 기술에 비해 탐지 성능을 높여야 한다.

2) **설명가능성**: 탐지 기법은 그 자체가 탐지 원인을 제공할 수 있어야 한다.

우리는 위 두 가지 목적을 달성하기 위하여 데이터 샘플을 구성하는 개별 특징에 집중하였다. 개별 특징이 다른 특징에 대해 갖는 관계성을 기반으로 특징별로 모델을 만들고 이들의 앙상블을 구성하여 침입을 탐지한다. 이는 기존 기술이 데이터 샘플들의 분포에 초점을 맞춘 것과 달리, 우리는 보다 세밀하게 특징들 간의 관계도 초점을 맞춤으로써 탐지 성능을 높이고자 하였고, 개별 특징에 종속된 모델 결과를 분석하여 탐지의 원인을 쉽게 식별할 수 있게 함으로써 설명가능성도 제공하고자 하였다.

3.2. 개관

그림 1은 RM의 학습 수행과 탐지를 수행하고 탐지의 원인이 되는 특징을 보고한다.

학습. 주어진 데이터셋의 n개의 특징을 가진 개별 샘플에 대해 특징 하나씩을 라벨로 하여 나머지 특징들에 대한 관계를 의사결정트리로 학습시킨다. 이렇게 하면 n개의 의사결정트리가 만들어지고, 이들의 앙상블이 RM이 된다.

탐지 및 원인특징 보고. 이상 여부를 식별하기 위한 샘플에 대해, 각 샘플이 RM 내 n개의 개별 모델에 입력을 들어가면, 개별 모델이 입력에 대한 이상 여부를 판단하여 출력한다. 이

렇게 출력된 n개의 판단 중에 하나라도 침입으로 판단된 경우가 있다면, 침입으로 결론 내리고, 그렇지 않으면 정상으로 판단한다. 동시에, n개의 모델 중 침입으로 판단한 모델이 대표하는 특징을 침입으로 판단한 결과에 대한 원인 특징으로 보고한다.

IV. RelationMesh에 대한 분석

특성이 n개인 데이터는 보통 n차원 공간을 균일하게 채우지 못한다. 일반적으로 훨씬 낮은 차원의 다양체에 존재하고 이 가설을 '다양체 가설'(Manifold Hypothesis)이라 한다. 오토인코더가 주어진 데이터를 더 낮은 차원으로 압축할 수 있는 이유도 여기에 있다.

다양체 가설은 다른 관점으로 보면, 주어진 입력 데이터는 다차원 공간을 균일하게 채우지 못한다고 말하는 것과 같다. n차원 데이터가 n차원 공간을 균일하게 채운다는 것은 모든 특성이 서로 독립적이라는 것이다. 하지만 현실의 데이터는 그렇지 않으므로 반드시 특성 간 종속관계가 있다는 것이고, 우리는 그 종속관계를 함수로 모델링한다.

특징 x_1과 x_2가 서로 종속되어 있다고 하자. 둘 중 하나는 나머지의 함수가 되고, x_2가 x_1의 함수라고 하면 다음과 같이 쓸 수 있다.

$$x_2 = f(x_1)$$

10차원 데이터라고 가정하면 $(x_1, ..., x_{10})$로 표현할 수 있다. 이때, 우리는 개별 특징들을 나머지 특징들에 대한 함수로 아래와 같이 표현할 수 있다.

$$x_2 = f(x_1, x_3, x_4, x_5, ..., x_{10})$$

n개의 특성을 가진 정상 데이터에 대해, 각 특성을 나머지로 예측하는 n개의 모델을 만들고 학습시킨다면, 나머지 정보로 쉽게 결정되지 않는 특징을 예측하는 모델은 자연스레 학습이 안 될 것이다. 결과적으로 예측이 잘 되는 특징의 예측오차는 좁고 작게 분포할 것이고 나머지로 결정되지 않는 특징을 예측하는 모델의 오차 분포는 크고 넓게 분포할 것이다.

우리는 이러한 관계들을 부스팅 모델을 활용한 특징별 예측(feature-wise prediction)을 통해 정상 데이터 내부에 형성된 종속관계들을 학습한다. 그 후, 오토인코더의 복원 오차가 큰 데이터를 이상치로 간주하는 것처럼 관계를 학습한 모델의 예측 오차가 충분히 커지는 데이터를 이상치로 간주한다.

본 논문에서 제안하는 RM은 입력 데이터에 대해 여러 모델의 예측오차가 존재한다. 이때 단 하나의 모델이라도 임계값을 넘으면 이상이라고 판단한다. 우리는 모든 특성의 예측오차 임계값을 하나로 통일하여 Bonferroni식 보정을 수행한다. n개의 모델이 존재하고 예측오차가 적어도 하나의 모델에서 훈련 오차의 분위수 q를 넘으면 이상이라고 판단할 때, 입력 데이터의 예측 오차가 모델 적어도 하나에서 q를 넘길 확률은 1에서 모든 모델에서 q보다 작을 확률을 뺀 것과 같다.

$$1 - Q = 1 - q^n$$

여기서 최종 분위수를 나머지 두 모델과 같이 훈련 오차의 0.9로 설정하면 훈련 데이터 중 10%만이 이상으로 판단되므로

$$0.1 = 1 - q^n$$

세 데이터의 차원 n에 맞춰 모든 모델의 예측오차 임계분위수는 아래와 같이 된다.

$$q = 0.9^{\frac{1}{n}}$$

V. 실험

우리는 RM을 구현하고 이에 대한 성능 검증을 3가지 벤치마킹 데이터셋(ToN-IoT, UNSW-NB15, NSL-KDD)에 대해 수행하였다. 우리는 개별 데이터셋을 훈련셋과 평가셋으로 분할하면서 평가셋의 공격 비율을 10%로 고정하였다. 이는 현실의 정상 데이터가 공격 데이터에 비해 많다는 것을 고려한 것이다.

실험에 사용한 RM의 의사결정트리는 더 정교한 관계까지 포착하기 위해 범주형 처리에 강한 CatBoost를 사용했다. 대조군 이상탐지 모델로 IsolationForest와 오토인코더를 택했다.

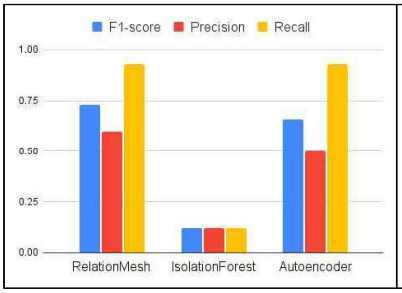

그림 2 ToN-IoT 결과

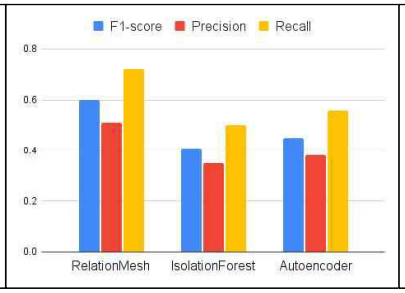

그림 3 UNSW-NB15 결과

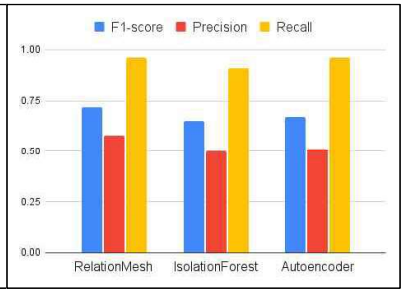

그림 4 NSL-KDD 결과

ToN-IoT	UNSW-NB15	NSL-KDD
목적지 포트 (0.87)	서비스 (0.92)	서버 에러율 (0.96)
송신지 포트 (0.78)	목적지 IP와 송신 포트 (0.91)	동일 연결 여부 (0.96)
TCP 플래그 (0.59)	프로토콜 (0.89)	호스트 서버 에러율 (0.96)
L7 프로토콜 (0.46)	송신 IP와 목적지 포트 (0.85)	목적지 바이트 수 (0.96)
들어오는 패킷 수 (0.01)	상태 (0.70)	서비스 서버 에러율 (0.95)

표 1 설명가능성: 정밀도가 높은 상위 5개 특징

오토인코더와 IsolationForest는 범주형 특성을 처리하는 기능이 없기에 공정한 평가 비교를 위해 데이터셋 내의 범주형 특징은 모두 onehot 인코딩으로 전처리하였다. 평가 지표는 F1-score와 정밀도, 재현율을 활용하였다.

5.1. 탐지 성능 분석

그림 2-그림 4는 데이터셋 별로 성능 평가를 수행한 결과이다. 여기서 RM은 IsolationForest, 오토인코더와 비교하여 최소 5%로부터 최대 15%만큼 더 높은 F1-score를 보였다. 주목할 점은 오토인코더 결과와 비교하면, ToN-IoT와 NSL-KDD에서 재현율은 같은데 RM의 정밀도가 각각 10%와 7%가 오른 것을 확인할 수 있다. 이는 샘플들의 분포에 초점을 맞추는 오토인코더에 비해 샘플 내 특징 간의 관계까지 보다 세밀하게 고려한 RM의 정밀도가 더 높게 나온 것을 보인다.

5.2. 설명가능성

표 1은 데이터셋 별로 탐지 결과의 원인이 되는 특징들을 계수하고 해당 특징들이 실제 정답의 원인일 비율을 계산한 것이다. ToN-IoT는 상위 4가지 특징의 비중이 높은 것으로 확인된다. UNSW-NB15는 주로 영향을 끼친 특징들이 IP와 포트 번호 사이의 관계라는 사실을 확인할 수 있다. NSL-KDD의 경우, 에러율이 침입 탐지에 큰 역할을 했다는 사실을 알 수 있다.

VI. 결론

본 논문에서 우리는 데이터 샘플의 개별 특징과 나머지 특징간의 관계성에 초점을 맞춘 비지도 이상탐지 기법인 RelationMesh를 제안하였다. 이 모델은 기존 기법에 비해 보다 세밀한 영역에 초점을 맞춰 탐지 성능을 높이고자 하였고, 탐지 원인에 해당하는 특징들을 식별할 수 있게 하여 설명 가능성도 제공한다. 우리는 RelationMesh의 설계를 고도화하여 탐지 성능을 보다 강화하고자 한다.

사사(謝辭)

이 연구는 2025년도 산업통상자원부 및 한국산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임. (과제번호: RS-2025-02653102)

[참고문헌]

[1] Lunt, Teresa F., et al. A real-time intrusion-detection expert system (IDES). Computer Science Laboratory, 1992.

[2] Lee, Hyunwoo, et al. "An infection-identifying and self-evolving system for IoT early defense from multi-step attacks." ESORICS, 2022.

[3] Liu et. al., "Isolation forest." 2008 eighth