그래프 기반 내부자 위협 탐지 기법 동향 조사

김주영*, 오찬석**, 최석환***

*연세대학교 소프트웨어학부(학부생)

**연세대학교 전산학과(석사과정)

***연세대학교 소프트웨어학부(지도교수)

A Survey on Graph-Based Insider Threat Detection Techniques

Ju-Young Kim*, Chan-Seok Oh**, Seok-Hwan Choi***

*Yonsei University Dept. of Software(Undergraduate Student)

**Yonsei University Dept. of Computer Science(Graduate Student)

***Yonsei University Dept. of Software(Professor)

요 약

내부자 위협은 정당한 접근 권한을 가진 사용자의 정상 행위와 구분하기 어렵고, 기존 방법은 수작업 특징 설계와 다량의 정답 레이블을 요구하여 실제 환경에 적용하기 어렵다. 이에 최근에는 다중 도메인 로그를 융합하고 개체 간 관계를 그래프로 모델링하여 레이블 의존도를 낮추는 방법들이 제안되고 있다. 본 논문에서는 이러한 동향을 특징 추출, 관계 표현, 학습 전략의 세 가지 측면에서 분석한다. 구체적으로 의미적 시퀀스 임베딩 기반의 자동 특징 추출, 행위 유사도를 활용한 다중 관점 그래프 구성, 레이블 가용성에 따른 학습 전략의 분화를 중심으로 각 기법의 방법론과 의의를 분석한다. 이를 통해 실환경에 적용 가능한 내부자 위협 탐지 연구를 위한 기초 자료를 제공하고자 한다.

Keywords: Insider Threat Detection, Anomaly Detection, Graph Neural Networks (GNN), User Behavior Modeling, Unsupervised Learning

I. 서론

최근 조직 시스템의 복잡성에 따라 내부자가 취약점을 악용하여 기업에 심각한 피해를 입히는 사례가 증가하고 있다. 기존의 보안 체계는 외부로부터 침입을 차단하는 데 집중되어 있어, 이미 합법적인 권한을 부여받은 내부자의 이상 행위를 식별하는 데에는 한계를 보인다. 특히 내부자의 악의적 행위는 정상적인 업무 패턴과 교묘하게 섞여 나타나므로, 개별 사용자의 접속 기록만으로는 정확한 위협을 가려내기 어렵다. 따라서, 최근에는 사용자 행위를 그래프 구조로 변환하여 개체 간 관계를 함께 학습하는 그래프 기반 내부자 위협 탐지 방법이 제안되고 있다. 그래프 기반 접근은 사용자 간 통신과 기기 공유 등 구조적 관계를 학습하여 단일 로그 분석으로 포착하기 어려운 위협 패턴을 탐지할 수 있다. 이러한 그래프 기반의 탐지 성능은 그래프를 구성하는 방식뿐 아니라, 어떠한 데이터를 특징으로 삼고 어떤 방식으로 학습하는지에 따라 크게 좌우된다. 이에 본 논문은 다중 도메인 로그 기반 내부자 위협 탐지 연구 동향을 특징 추출, 관계 표현, 학습 전략의 세 관점에서 분석하고자 한다.

II. 본론

2.1 다중 도메인 로그 데이터 특징 추출 동향

초기 연구는 이메일 통신과 같이 사용자 간 직접적인 상호작용이 드러나는 데이터에 크게 의존하였다[1]. 그러나 실제 조직에서 사용자 간의 직접적인 교류(관계)가 없는 경우에는 단일 데이터만으로는 행위를 충분히 특징화하기 어렵다는 한계가 지적되었다. 이를 보완하기 위해

최근 연구들은 Logon, 기기 사용, 웹 브라우징, 파일 접근, 이메일 등 다양한 도메인 로그를 융합하여 개체 간의 잠재적 상호작용과 구조적 관계를 탐지할 수 있는 데이터 특징을 추출하여 사용하고 있다. 다중 도메인 로그에서 특징을 추출하는 방식은 크게 세 가지로 구분되며, 이를 병행하거나 선택적으로 활용하는 추세를 보인다.

첫째, 시간적 변화에 따른 통계 및 시계열 특징 추출이다[2]. 사용자 행위는 시간에 따라 변화하므로 전체 기간을 단순 누적하는 대신 일(Day) 또는 주(Week) 단위 윈도우를 설정한다. 절댓값 대신 윈도우 내 분포에서의 백분위수나 평균값의 차이를 계산하여 변화 자체를 인코딩함으로써, 정상 패턴에서 벗어나는 변화를 포착한다.

둘째, 원시 로그의 자동 수치화이다[4]. 집계된 통계 피처를 만드는 대신 로그 레코드 자체를 필드 단위로 수치화한다. 날짜 및 시간은 Unix 타임스탬프로, 활동 유형은 Factorization을 통해 정수로 자동 매핑하며, 파일 경로, URL과 같은 텍스트 필드는 구분자 단위로 분절한 뒤 임베딩을 거쳐 수치화한다. 이후 값의 스케일을 통일하기 위해 Min-Max 정규화를 수행한다. 이 방식은 수작업 특징 설계를 최소화한다는 이점이 있다.

셋째, 자연어 처리 및 의미 기반 시퀀스 임베딩이다[5]. 가장 최근 연구 동향으로 사용자의 일일 활동 전체를 하나의 문장으로, 개별 행위를 단어로 간주하여 Word2Vec 등의 임베딩 기법으로 벡터화한다. 이러한 임베딩 기법은 레이블을 요구하지 않기 때문에 신뢰할 만한 레이블 확보가 어려운 환경에 적합하다[4,5]. 또한 행동의 단순 빈도를 넘어 발생 순서와 문맥적 흐름까지 종합적으로 학습할 수 있게 되었다.

2.2 그래프 기반 데이터 모델링 및 관계 표현 기법 동향

다중 도메인 로그로부터 추출한 특징만으로는 개체 간 상호작용이나 행위 간 문맥을 반영할 수 없다. 최근 연구들은 로그 데이터를 그래프 구조로 변환하여 관계 정보를 함께 학습하며, 엣지를 정의하는 방식은 크게 세 가지로 정리된다.

첫째, 행동 유사성을 결합한 가중치 기반 인접 행렬 구성이다[1,3]. 단순한 직접 연결의 한계를 극복하기 위해 사용자의 직접적인 통신 관계와 행동 특징 간 유사성(코사인 유사도 등)을 조합한 가중치 함수를 설계하여 인접 행렬을 구성한다. 이를 통해 직접 연결로 엣지를 구성했을 때의 고립 노드 문제를 완화할 수 있다.

둘째, 엣지 표현을 다각화하는 다중 그래프 구성 방식이다[3,4]. 하나의 관계 정의에 모든 관계를 융합하는 대신 서로 다른 관점의 그래프를 구성한다. DD-GCN은 Topology domain과 Feature domain을 분리한 뒤 어텐션으로 융합하며[3], MEWRGNN은 개별 활동 로그를 노드로 두어 시간, 거리, 어텐션 기반 다중 그래프를 구성한다[4]. 이를 통해 단일 엣지로 포착하기 어려운 관계를 표현한다.

셋째, 공통 문맥 기반의 상호작용 그래프 구축이다[5]. 단순한 행동 특징의 수치적 유사성을 넘어, 구체적인 공유 상황을 기준으로 노드를 연결하기 때문에 유사도를 별도로 계산하지 않는다. 예를 들어 동일 기기를 같은 기간 사용하거나 동일 도메인에 같은 날 접속하는 등 물리적·논리적 자원을 공유하는 사용자들 사이에 엣지를 생성한다. 이러한 방식은 개별 사용자의 독립적인 이상 행동이 아닌 집단적 이상 행동과 패턴까지 그래프로 표현할 수 있다.

2.3 이상 탐지 모델 및 학습 전략 동향

내부자 위협 데이터는 악의적인 행위가 극히 희소하다는 한계를 지니며, 정답 레이블의 활용 여부에 따라 학습 전략은 크게 세 가지로 분류된다.

첫째, 탐지 정확도와 해석 가능성에 중점을 둔 지도 학습 기반 탐지이다[4]. MEWRGNN은 정상 및 악성 로그를 활용하여 사용자별 탐지 모델을 구축하므로 계정 도용 탐지에 유리하지만, 악성 레이블이 없는 신규 사용자에게 적용하기 어려울 것으로 판단된다.

둘째, 제한된 레이블의 활용을 극대화하는 준지도 학습 기반 탐지이다[1,3]. GCN은 학습을

위해 그래프의 모든 노드에 레이블을 요구하지 않기 때문에 이러한 환경에 적합하다[1]. 대표적으로 DD-GCN 모델은 Topology domain과 Feature domain 간의 정보를 융합할 때 어텐션 메커니즘을 사용한다[3]. 일부 위협 레이블을 활용하여 두 도메인의 중요도를 적응적으로 학습하므로 전체 노드의 레이블 없이도 탐지가 가능하다.

셋째, 레이블이 전혀 없다는 것을 전제로 한 비지도 학습 기반 탐지이다[2,5]. 레이블이 전혀 없는 실환경에 가장 적합하지만, 정답이 없으므로 학습 데이터 자체에 악성 행위가 섞여 모델이 오염될 수 있는 구조적 위험성이 존재한다. 이를 극복하기 위해 다수의 비지도 모델을 앙상블하여 강건성을 높이거나[2], GCN을 완전 비지도 방식으로 응용하는 기법이 제안되었다. 최근 연구에서는 정답 레이블 없이 그래프 구조와 의미론적 특징만으로 GCN을 학습시킨 뒤, 문맥이 강화된 임베딩을 k-Means 클러스터링과 결합하여 소규모 군집을 이상치로 분리해 낸다[5]. 이 경우 GCN은 임베딩 보강 도구로 사용되므로 비지도로 유지된다.

III. 결론

본 논문은 다중 도메인 로그 기반 내부자 위협 탐지 연구를 특징 추출, 관계 표현, 학습 전략의 세 관점에서 분석하였다. 분석 결과, 특징 추출은 수작업 피처에서 자동 임베딩으로, 관계 표현은 직접 연결에서 유사도 기반 다중 관점 그래프로, 학습 전략은 지도 학습에서 비지도 학습으로 이동하고 있다. 이는 레이블 확보가 어려운 실제 운영 환경에서 정답 의존도를 낮추면서도 개체 간 맥락을 확보하려는 방향으로 수렴한다. 다만, 대상 연구들의 검증이 대부분 합성 데이터셋에 국한되고 평가 조건이 상이하다는 점은 공통된 한계로 남는다. 향후 본 기술이 산업 현장에 정착되기 위해서는 다음의 후속 연구가 요구된다. (1) 다중 엣지 및 도메인 세분화로 인해 증가한 연산 복잡도를 낮추는 모델 경량화 및 최적화 (2) 부서 이동 등 시간의 흐름에 따라 변화하는 조직 내 관계를 반영하기 위한 동적 그래프 신경망의 적용이다. 이러한 과제가 해결된다면 내부자 위협 탐지 기술은 제한된 실험 환경을 넘어 실제 조직의 보안 운영 체계에 통합될 수 있을 것으로 기대된다.

[Acknowledgement]

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2026년도 의료AI반도체 전·후방 엔지니어 육성사업의 결과로 수행되었음 (2024-0-0096)

[참고문헌]

[1] J. Jiang et al., "Anomaly detection with graph convolutional networks for insider threat and fraud detection," in Proc. MILCOM 2019 - IEEE Military Communications Conf., Norfolk, VA, USA, pp. 109 - 114, 2019.

[2] D. C. Le and N. Zincir-Heywood, "Anomaly detection for insider threats using unsupervised ensembles," IEEE Trans. Netw. Serv. Manag., vol. 18, no. 2, pp. 1152 - 1164, Jun. 2021.

[3] X. Li et al., "A high accuracy and adaptive anomaly detection model with dual-domain graph convolutional network for insider threat detection," IEEE Trans. Inf. Forensics Security, vol. 18, pp. 1638 - 1652, 2023.

[4] Xiao, Junchao, et al. "Robust anomaly-based insider threat detection using graph neural network." IEEE Transactions on Network and Service Management 20.3 (2023): 3717-3733.

[5] Baghalizadeh-Moghadam, Neda, et al. "Semantic and graph-based unsupervised learning for insider threat detection using user activity sequences." 2025 22nd Annual International Conference on Privacy, Security, and Trust (PST). IEEE, 2025.