준지도 학습 기반의 멀웨어 탐지 기법

전유란[1], 심혜연[1], 이일구[2]
[1]성신여자대학교 미래융합기술공학과 박사과정
[2]성신여자대학교 미래융합기술공학과, 융합보안공학과 교수
220247016@sungshin.ac.kr, 220237062@sungshin.ac.kr, iglee@sungshin.ac.kr

Semi-supervised learning based malware detection technique

Yu-Ran Jeon[1], Hye Yeon Shim[1], Il-Gu Lee[1,2]
[1]Dept. of Future Convergence Technology Engineering, Sungshin Women's University
[2]Dept. of Convergence Security Engineering, Sungshin Women's University

요 약

5G 통신과 인공지능 기술이 발전하고, 사물인터넷 기기의 수가 증가함에 따라 종래의 정보보호체계를 우회하는 지능적인 사이버 공격이 증가하고 있다. 그러나, 종래의 기계학습 기반 멀웨어 탐지 방식은 이미 알려진 멀웨어만 탐지할 수 있으며, 새로운 멀웨어는 탐지가 어렵거나, 기존의 알려진 멀웨어로 잘못 분류되는 문제가 있다. 본 연구에서는 비지도학습을 사용하여 알려지지 않은 멀웨어를 탐지하고, 새롭게 탐지된 멀웨어를 새로운 라벨로 분류하여 재학습하는 준지도 학습 기반의 멀웨어 탐지 기법을 제안한다. 다양한 데이터 환경에서 알려지지 않은 멀웨어 데이터가 탐지 모델로 입력될 때 제안한 방식의 성능을 평가했다. 실험 결과에 따르면 제안한 준지도 학습 기반의 멀웨어 탐지 방법은 종래의 방식 대비 정확도를 약 16% 개선했다.

1. 서론

최근 지능형 사이버 공격이 증가함에 따라 다양한 공격 대응 기법들이 연구되고 있다. 사이버 공격은 점차 지능화되고 있으며, 기존의 사이버 공격 대응 기법으로 탐지가 어렵게 우회하는 새로운 변종 멀웨어들이 증가하고 있다[1]. 이러한 알려지지 않은 새로운 멀웨어들은 종래의 사이버 공격 대응 기법으로 탐지가 어렵거나, 탐지하더라도 이미 알려진 멀웨어로 잘못 분류되는 문제가 발생한다. 따라서, 알려지지 않은 멀웨어 탐지를 위한 새로운 공격 대응 기법에 관한 연구가 필요하다.

종래의 사이버 공격 대응 기법인 기계학습 기반 공격 탐지 방식은 공격의 탐지 범위가 학습에 사용된 공격으로 한정된다. 따라서, 기계학습 모델의 학습에 사용되지 않은 공격 데이터가 기계학습 기반 탐지 모델에 입력되면 이미 알려진 공격 라벨로 잘못 분류되므로 탐지가 어렵다.

따라서, 본 연구에서는 준지도 학습 기반의 멀웨어 탐지 기법을 제안한다. 데이터의 라벨에 의존하지 않는 비지도학습을 활용하여 알려지지 않은 공격이 알려진 공격으로 잘못 분류되는 한계를 개선하며, 비지도학습을 통해 재구성된 데이터를 사용하여 지도학습함으로써 학습 성능을 개선한다.

본 논문의 주요 기여점은 다음과 같다.

- 데이터가 희소하고 노이지한 조건에 효과적인 준지도 학습 기반의 알려지지 않은 멀웨어 탐지 기법을 제안한다.
- 멀웨어 탐지 기법들의 알려지지 않은 멀웨어 탐지율을 비교 평가하는 프레임워크를 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 멀웨어 탐지 기법 관련 선행 연구를 분석하고, 3장에서는 제안하는 준지도 학습 기반의 멀웨어 탐지 기법을 설명한다. 4장에서 제안한 방식의 성능을 평가한 후, 5장에서 결론을 맺는다.

2. 관련 연구

F. wang 등[2]은 알려지지 않은 공격에 대한 탐지 정확도를 높이기 위해 알려진 공격의 분포를 이용하는 비지도학습 기반 도메인 적응(UDA, Unsupervised Domain Adaptation) 악성 코드 탐지 기법을 제안했다. 적대적 학습을 통해 알려지지 않은 공격의 분포를 알려진 공격의 분포와 유사하게 변경하고, 심층학습 알고리즘을 사용하여 알려지지 않은 공격에 대한 탐지 정확도를 개선했다. 그러나 알려진 공격

을 난독화하면 알려지지 않은 공격으로 분류하므로 기존 멀웨어와 다른 특성을 갖는 변종 멀웨어 탐지 성능을 확인하기 어렵다. G. Pitolli 등[3]은 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) 온라인 군집화 알고리즘을 사용하여 새로운 공격 식별과 공격 분류를 동시에 수행했다. 종래의 군집화 알고리즘에 비해 모델 재학습 및 업데이트 비용을 줄였고, 알려지지 않은 공격을 탐지했다. 그러나 군집 간 거리가 특정 임계치 이상인 경우에만 새로운 공격으로 분류하므로 알려지지 않은 공격이 알려진 공격과 유사하면 새로운 공격으로 분류하지 못하는 한계가 있다. J. Yang 등[4]은 알려지지 않은 공격을 식별하는 계층적 탐지 프레임워크를 제안했다. 알려진 공격 및 알려지지 않은 공격 분류 문제를 2개의 최소화 문제로 공식화함으로써 계층적으로 분류를 수행했다. CVAE(Conditioned Variational Auto-Encoder)와 EVT(Extreme Value Theory) 모델을 이용하여 알려지지 않은 공격을 식별하고, 라벨을 새롭게 지정했다. 그러나 제안하는 모델은 알려진 공격에 대한 사전 모델이 구축되어야 하므로 자원 제약적인 환경에 적용하기 어렵다. M. Soltani 등[5]은 군집화 알고리즘과 심층학습 알고리즘을 결합하여 알려지지 않은 공격을 탐지하는 프레임워크를 제안했다. 제안하는 프레임워크는 군집화 알고리즘을 사용하여 알려지지 않은 공격에 대한 라벨링을 수행하며, 침입 탐지 시스템에 새로운 라벨을 업데이트했다. 그러나 [3]과 마찬가지로 정확도 임계치를 이용해서 알려지지 않은 공격을 분류하므로 알려지지 않은 공격이 알려진 공격으로 잘못 분류되는 문제가 있다.

3. 준지도 학습 기반의 멀웨어 탐지 기법

그림 1은 제안하는 방식의 멀웨어 탐지 프레임워크를 나타낸다. 제안하는 방식은 비지도학습을 사용하여 알려지지 않은 멀웨어를 새로운 라벨로 분류하고, 새롭게 구성된 데이터를 활용하여 지도학습을 수행한다. 제안하는 방식의 동작 과정은 아래와 같다.

(1) 데이터 전처리

제안하는 방식의 분류 성능을 개선하기 위해 데이터 전처리를 수행했다. 먼저, 데이터셋의 분류 성능을 저하시키는 값들을 제거하였으며, 이 과정에서 허수 혹은 무한대 값을 갖거나, 소수의 데이터로 구성된 라벨의 데이터는 삭제하였다. 또한, 주성분 분석 (PCA, Principal Component Analysis) 기법을 사용하여 학습용 데이터의 차원을 축소함으로써 분류 성능을 개선했다.

(2) 알려지지 않은 멀웨어 탐지

제안하는 방식은 알려지지 않은 멀웨어를 탐지하기 위해 군집화를 수행한다. 기존 데이터셋의 일부를 선택하여 샘플 데이터를 생성하고, 샘플 데이터와 새롭게 들어오는 데이터를 병합해서 군집화 모델에 입력한다. 군집화 모델은 기존에 할당된 데이터 라벨에 의존하지 않으므로 새롭게 입력되는 데이터는 새로운 군집으로 분류된다. 따라서, 기존 데이터의 군집 수보다 더 많은 군집 수가 최적의 군집 수일 경우에는 알려지지 않은 멀웨어가 존재하는 것으로 간주한다.

(3) 데이터셋 재구성

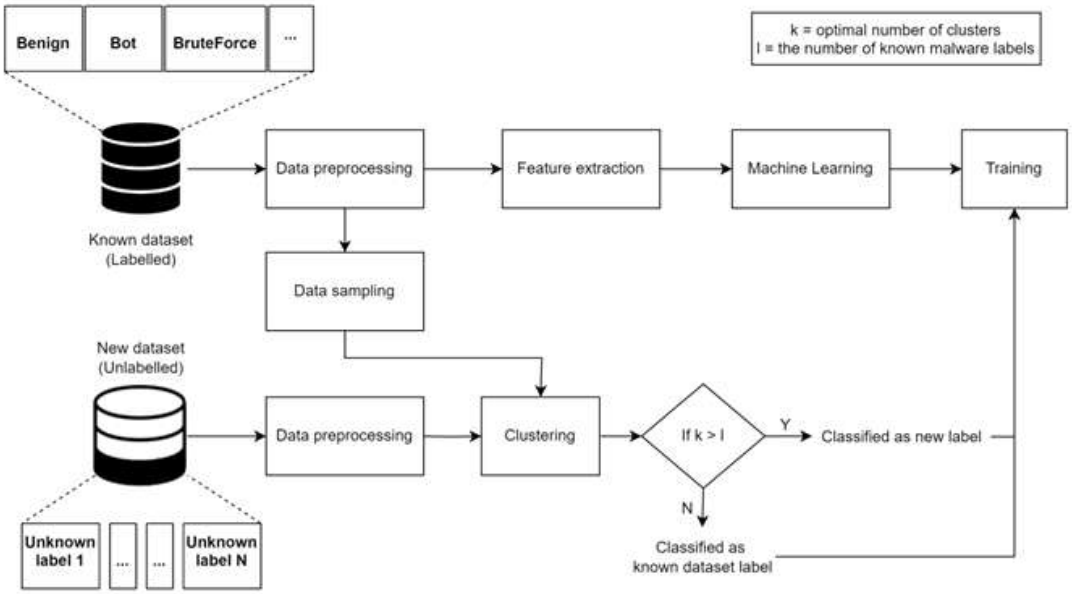
알려지지 않은 멀웨어가 존재하는 것으로 판단되면 새롭게 생성된 군집을 기존 데이터셋의 새로운 라벨로 분류한다. 이때, 새롭게 생성된 군집의 수는 1개 이상이며, 새로 생성된 군집 수만큼 새로운 라벨을 생성하여 데이터셋을 재구성한다. 본 연구에서는 알려지지 않은 멀웨어 데이터 군집이 1개인 경우를 가정했다.

(4) 모델 재학습 및 최종 분류

재구성된 데이터셋을 사용하여 모델을 재학습하고, 재학습된 모델을 사용하여 최종 분류를 수행한다. 알려지지 않은 멀웨어를 식별한 후, 모델을 재학습함으로써 알려지지 않은 멀웨어 탐지할 수 있으며, 이 과정은 알려지지 않은 멀웨어 데이터가 새롭게 입력될 때마다 수행한다.

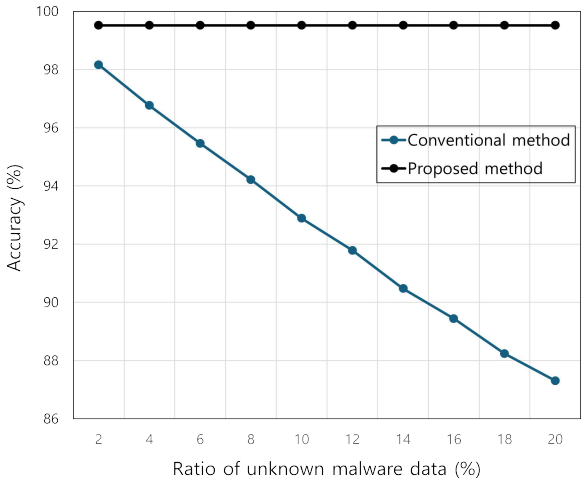
4. 성능 평가

본 장에서는 제안한 방식의 평가 환경을 설명하고 평가 결과를 분석한다. 비지도학습으로 가우시안 혼합 모델 기반 군집화 알고리즘을 사용하였고, 의사결정 트리 알고리즘으로 지도학습을 수행하였다. 데이터셋으로는 77개의 특성을 갖고, 7개의 라벨로 구성된 50,000개의 CIS-IDS 2017 데이터를 사용하였다[6]. CIC-IDS 2017 데이터셋은 Benign, Bot, BruteForce, DoS, Infiltration, PortScan, WebAttack 라벨을 가지며, 이중 WebAttack 라벨 데이터를 알려지지 않은 멀웨어 데이터로 가정하였다. 알려진 데이터셋을 학습용 데이터와 평가용 데이터로 분할



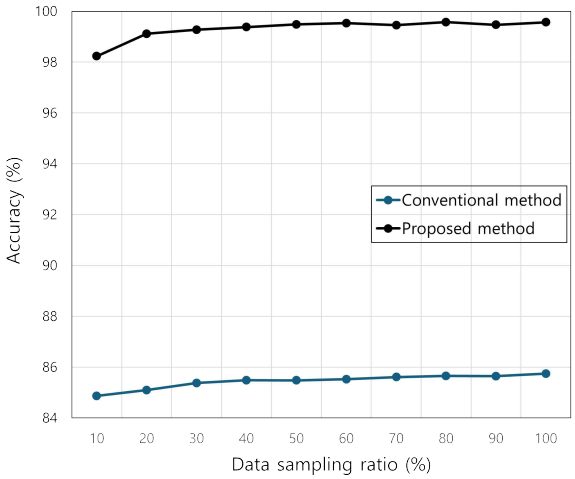
(그림 1) 알려지지 않은 멀웨어 탐지 프레임워크

하였으며, 그 비율은 4:1로 설정했다. 평가용 데이터에 알려지지 않은 멀웨어 데이터를 추가하여 알려진 데이터셋으로 학습된 모델의 성능을 평가했다. 알려지지 않은 멀웨어 데이터를 테스트 데이터에 추가할 때, 알려지지 않은 멀웨어 데이터의 라벨은 알려진 데이터의 라벨 중 하나를 무작위로 선택하였다. 또한, 군집화 기반 데이터 재구성을 수행하지 않는 종래의 방식과 군집화 기반 데이터 재구성을 수행하는 제안한 방식의 성능을 다양한 환경에서 비교하였다.



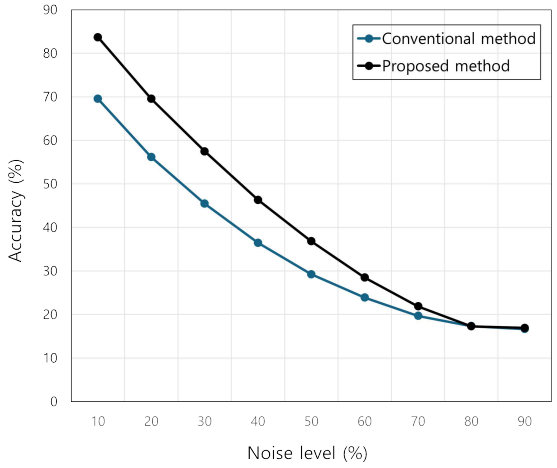
(그림 2) 알려지지 않은 멀웨어 데이터의 비율에 따른 정확도

그림 2는 테스트 데이터 중 알려지지 않은 멀웨어 데이터의 비율에 따른 정확도를 측정한 결과이다. 종래의 방식은 알려지지 않은 멀웨어가 새롭게 입력됐을 때 새로운 라벨이 아닌 기존의 라벨로 데이터를 분류하므로 제안한 방식보다 더 낮은 탐지 정확도를 보였다. 테스트 데이터 중 알려지지 않은 멀웨어 데이터의 비율이 증가할수록 종래의 방식과 제안한 방식의 정확도 차이가 더 커짐을 확인할 수 있었다.



(그림 3) 데이터 샘플링 비율에 따른 정확도

그림 3은 데이터셋의 크기가 증가함에 따른 제안한 방식과 종래 방식의 정확도를 비교한 결과이다. 전체 데이터 중 무작위로 데이터를 선택하여 샘플링했다. 제안한 방식과 종래의 방식 모두 데이터의 수가 증가함에 따라 정확도가 개선되었다. 데이터의 개수를 변화시켜도, 제안한 방식은 종래의 방식에 비해 평균 정확도를 약 16.2% 개선할 수 있었다.



(그림 4) 노이즈 수준에 따른 정확도

그림 4는 노이즈 수준에 따른 탐지 정확도를 나타낸다. 전체 데이터셋의 일부를 선택하여 데이터 라벨에 노이즈를 추가했다. 따라서, 노이즈 수준이 커질수록 노이즈가 삽입되는 데이터의 수가 증가한다. 노이즈가 증가할수록 제안한 방식과 종래 방식의 정확도는 감소했다. 제안한 방식은 종래의 방식에 비해 평균 정확도를 약 20.3% 개선했지만, 노이즈 수준이 매우 클 때는 제안한 방식과 종래방식의 정확도는 동일한 수준을 보였다.

5. 결론

사이버 공격의 수가 증가하고, 유형이 다양해짐에 따라 멀웨어 탐지를 위한 다양한 연구가 수행되고 있다. 그러나, 종래의 멀웨어 탐지 방식 중 하나인 지도학습 기반 멀웨어 탐지 방식은 이미 알려진 멀웨어만을 탐지할 수 있다는 한계를 갖는다. 본 연구에서는 비지도학습을 사용하여 알려지지 않은 멀웨어 데이터를 새로운 라벨로 분류하고, 재구성된 데이터를 사용하여 멀웨어 분류를 수행하는 준지도 학습 기반 탐지 방법을 제안했다. 다양한 데이터 환경에서 제안한 방식의 성능을 평가한 결과에 따르면 비지도학습 기반 데이터 재구성 방식을 사용하지 않는 종래의 방식 대비 개선된 정확도를 보였다. 본 연구에서는 공개 데이터셋을 사용하여 성능을 평가해서 제안한 방식이 전반적으로 높은 정확도를 보였다. 향후 연구에서는 테스트베드를 구축하여 데이터 수집한 후 성능을 평가하여, 실제 환경에서도 알려지지 않은 멀웨어를 효과적으로 탐지하는 방법을 연구할 계획이다.

ACKNOWLEDGEMENT

본 논문은 2024년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Aslan, Ömer Aslan, and Refik Samet. "A comprehensive review on malware detection approaches." IEEE access 8 (2020): 6249-6271.

[2] Wang, Fangwei, et al. "An efficient deep unsupervised domain adaptation for unknown malware detection." Symmetry 14.2 (2022): 296.

[3] Pitolli, Gregorio, et al. "MalFamAware: automatic family identification and malware classification through online clustering." International Journal of information security 20 (2021): 371-386.

[4] Yang, Jian, et al. "Conditional variational auto-encoder and extreme value theory aided two-stage learning approach for intelligent fine-grained known/unknown intrusion detection." IEEE Transactions on Information Forensics and Security 16 (2021): 3538-3553.

[5] Soltani, Mahdi, et al. "An adaptable deep learning-based intrusion detection system to zero-day attacks." Journal of Information Security and Applications 76 (2023): 103516.

[6] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization”, 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018