악성코드 탐지를 위한 시스템 측정 메트릭 커버리지의 한계점 분석

Analyze the Limitations of System Measurement Metric Coverage for Malware Detection

구인회
SysCore Lab
세종대학교
서울시, 대한민국
kuinchang@gmail.com

최상훈
SysCore Lab
세종대학교
서울시, 대한민국
csh0052@gmail.com

박기웅*
정보보호학과
세종대학교
서울시, 대한민국
woongbak@sejong.ac.kr

요 약

악성코드는 인터넷에서 공격자가 공격을 수행하기 위한 소프트웨어이다. 최근에는 전통적인 유형의 악성코드 행동 패턴에서 벗어난 악성코드가 등장함에 따라 대응의 어려움을 겪고 있다. 따라서, 새로운 유형의 악성코드 탐지와 분석을 위해 다양한 유형의 악성코드 분석 기법 및 탐지 기법이 요구된다. 본 논문에서는 머신러닝 기반 멀웨어 탐지 연구의 동향과 악성코드 탐지향상을 위한 시스템 측정 메트릭 수집기술의 한계점을 분석한다.

키워드: 악성코드, 이상행위 탐지, 머신러닝, 데이터 마이닝

1. 서론

인터넷 사용량의 증가와 네트워크 접속 방법이 다양해지면서 새로운 유형의 악성코드 탐지와 분석을 위해 다양한 분석기술이 연구되어지고 있다. 그러나 일반적으로 사용되는 악성코드의 시스템 측정 메트릭을 미리 저장, 수집 및 탐지 작업을 수행하는 방식으로는 최근 빠르게 변화하고 출현하는 변종 악성코드에 대한 대응이 어려워진다는 단점이 있다[1]. 이와 같은 단점을 극복하기 위해 악성코드의 시스템상에서 하는 행동들을 시간순으로 수집하는 시계열 기반 다양한 방법론들은 있지만, 런타임으로 시스템 측정 메트릭을 수집하기 위해서는 많은 오버헤드를 발생시키는 한계점이 있다. 이러한 한계점을 극복하기 위해 시스템상에서 악성코드의 행동을 시간순으로 수집하는 시계열 기반 시스템 측정 메트릭을 통한 악성코드의 탐지모델을 확립하여 예측하는 방법을 시도는 하고 있지만, 시간으로 주기를 설정할 경우 변화와 패턴 파악의 어려움으로 시스템 측정 메트릭을 수집하기가 어렵다는 추가적인 한계점이 도출되고 있다[2]. 따라서 현재 진화하고 있는 변종들을 수집하기 위해서는 적은 오버헤드를 통한 런타임에 대량의 데이터들을 수집하는 역량을 갖춰야 한다. 이를 통해 다양한 변종들의 악성코드를 판별·탐지·분석·대응을 위한 시스템 측정 메트릭 분석 속도를 향상시키면, 적은 용량으로 시스템에 오버헤드를 낮추는 시스템 퍼포먼스를 향상시키는 이점을 가질 수 있다.

본 논문에서는 악성코드 탐지의 정확도를 향상시키고, 분석 시간을 감소시키는 머신러닝 기반 정적 분석 방법과 동적 분석 방법의 최근 연구 동향을 서술하고, 악성코드 탐지 분석 속도를 향상시키고, 적은 용량으로 시스템의 오버헤드를 낮추기 위해 악성코드 탐지 향상을 위한 런타임 시스템 측정 메트릭 수집 기술의 한계점을 살펴보고자 한다.

* 교신저자

2. 머신러닝 기반 악성코드 탐지 동향

머신러닝 기반 악성코드 탐지를 위한 정적 분석(Static Analysis)과 동적 분석(Dynamic Analysis)을 구분하여 연구 동향을 살펴보았다.

기존의 정적 분석 방법의 경우 넓은 커버리지와 오버헤드가 적다는 장점을 지니고는 있지만 멀웨어 프로그램에서는 오탐지가 발생할 위험이 존재한다. 이러한 문제를 해결하기 위해 탐지의 정확도와 분석 시간을 감소시키는 멀웨어 탐지 연구가 진행되고 있다. D. Ö. Şahın의 경우 멀웨어를 효과적으로 탐지하기 위하여 최소한의 권한만을 사용하여 탐지 정확도 향상 및 분석 속도를 감소시키고자 했다[3]. 이는 기존의 권한 기반 악성코드 탐지 머신러닝 모델인 KNN(K-Nearest Neighbor), SVM(Support Vector Machine)와 비교하여 정확도 향상 및 분석 속도를 감소시킬 수 있었다.

머신러닝 기반 동적 분석 방법은 멀웨어의 다양한 동적 데이터의 특징을 추출하여 탐지 정확도를 향상시키기 위한 연구를 진행하였다.

Yahye Abukar Ahmed은 윈도우 플랫폼에서 랜섬웨어가 악의적인 동작을 숨기기 위해 관련 없는 중복 호출을 포함하는 방대한 양의 시스템 콜로 탐지 회피하는 문제가 있었다. 이를 해결하고자 동적 분석 로그에서 수집한 Windows API 추적 및 시스템 호출 시스템 측정 메트릭의 정제 프로세스를 도입하여 머신러닝 지도 학습 방식을 통해 Windows API 호출 시퀀스에서 향상된 최대 관련성 및 최소 중복성(EmRmR) 방법을 사용, 기존의 mRmR 방식 보다 빠른 탐지의 정확도를 보였다[4].

3. 악성코드 탐지 향상을 위한 시스템 측정 메트릭 수집 기술 동향

본 논문에서는 그림 1과 같이 상용 시스템 측정 메트릭 수집 프로그램의 범위에는 시스템에서 제공하는 자원인 프로세스, 메모리, 이더넷, 파일을 이용하여 3.1 네트워크 측정 메트릭 유형을 수집, 3.2 시스템 측정 메트릭 유형을 수집한다. 이를 통해 머신러닝 기반 멀웨어 탐지 연구에서 요구하는 측정 메트릭을 활용하여 악성코드를 판별한다. 다음 표 1은 시스템 측정 메트릭 수집 연구로 멀웨어 탐지를 위한 머신러닝 연구동향을 정리한 표로서 메트릭 유형, 타겟 플랫폼, 대응되는 데이터 수집 프로그램을 나타낸다.

3.1. 네트워크 측정 메트릭 유형 수집

T. N. Nguyen의 경우 IoT 봇넷 탐지 및 분류를 위해 PSI-rooted 서브 그래픽 특징을 이용하였다. 이를 통해 네트워크 측정 메트릭을 포함한 정적, 동적 분석을 통합한 새로운 프레임워크를 제안하여, 8,330개의 실행 가능한 샘플데이터로 전체 알고리즘의 구현시간을 30%에서 50%까지 단축하였다. 그러나 실행파일이 전체 악성행위를 표시하지 않거나, 모니터링 중에 동적 문자열을 제대로 수집하지 못한다는 한계점이 있다[7].

Mahdi Rabbani의 경우 클라우드 컴퓨팅 환경에서 사용자의 네트워크 트래픽으로 악성행위 식별을 하기 위해 입자 군집 최적화 기반 확률 신

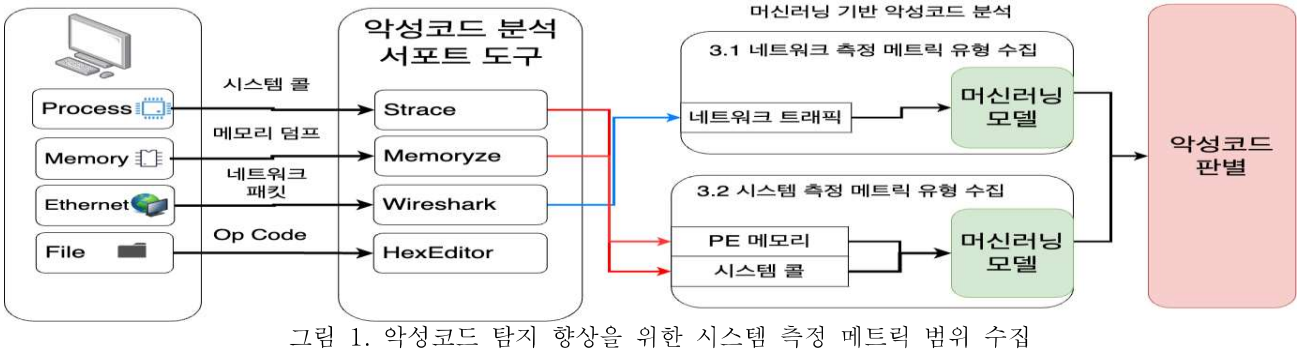

그림 1. 악성코드 탐지 향상을 위한 시스템 측정 메트릭 범위 수집

경망(PSO-PNN)을 제안하였다. 그리고 실시간 사용자 네트워크 트래픽 측정 데이터 셋인 UNSW-NB15[5]를 활용하여 데이터 디도스, 백도어 같은 악성 행위들의 패턴들을 구분하여 높은 정확도를 얻었다. 이를 통해 네트워크 패킷 수집 및 트래픽 모니터링 상용 프로그램인 WireShark나 Tcpdump를 활용해서 런타임으로 수집은 가능하나, 수집 시 네트워크 트래픽 데이터가 메모리에 저장하는 용량이 크고 빠르게 저장된다는 한계를 갖고 있다[8].

3.2. 시스템 측정 메트릭 유형 수집

Çağatay Yücel의 경우 멀웨어 정적, 동적 분석에 대한 패킹, 난독화, 데드코드 삽입, 안티디버깅 등 한계를 해결하고자 악성파일의 메모리 이미징과 패턴추출을 위한 새로운 메모리 연산을 기반으로 메모리 접근 이미지를 추출하는 방법론을 제시하여 머신러닝 모델중 하나인 CNN(Convolutional Neural Networks)을 통해 유사한 악성코드 샘플 탐지에 성공했다[9].

Amir Namavar Jahromi의 경우 알려지지 않은 멀웨어와 그 변종을 탐지하기 위해 윈도우 환경에서 악성코드의 원시 시퀀스 (Op Code 및 시스템 콜) 시스템 측정 메트릭을 활용한 머신러닝 탐지 방식인 2계층 극한 학습 머신(TELM)을 제안하여 순차적 데이터 봇넷 멀웨어를 탐지를 할 수 있다. 해당 연구에서는 실제 시스템에 적용을 위해 런타임 시스템 측정 메트릭을 수집하기 위한 방법으로 동적 바이너리 계측 프레임워크인 Pin[6]을 활용할 수 있다[10].

그러나, 프로세스는 끊임없이 데이터가 변화하는 특징을 갖고 있기 때문에 일반적인 시스템 메트릭 수집도구를 사용했을 때 데이터 수집으로 인한 연산적, 공간적 오버헤드가 발생한다는 한계점이 있다[8],[9],[10]. 따라서 머신러닝 기반 악성코드 탐지에 필요한 데이터의 커버리지에 비해 시스템 측정 메트릭을 수집할 수 있는 커버리지가 작다는 한계점을 가진다. 시스템 측정 메트릭을 수집하기 위해서는 많은 오버헤드를 발생시키는 한계점이 있으므로, 현재 진화하고 있는 변종들을 탐지하기 위해서는 적은 오버헤드를 통한 런타임으로 대량의 시스템 측정 메트릭들을 수집하는 역량을 갖춰야한다.

표 1. 머신러닝 기반 악성코드 탐지에 필요한 시스템 메트릭 유형에 매칭한 데이터 수집 프로그램

연구	메트릭 유형	타겟 플랫폼	대응되는 데이터 수집 프로그램
[7]	Network Traffic 시스템 콜, Op Code	Linux	Strace[11] WireShark[12]
[8]	Network Traffic	Universal	WireShark, TCPDump[13]
[9]	PE Memory	Windows	Memoryze[14]
[10]	시스템 콜, Op Code	Windows	PinTool[6]

4. 결론

본 논문에서는 머신러닝 기반 멀웨어 탐지 연구의 동향과 악성코드 탐지 향상을 위한 시스템 측정 메트릭 수집기술의 한계점에 대해 분석하였다. 악성코드 탐지 향상을 위한 시스템 측정 메트릭 수집 기술 한계점은 첫째, 시스템 메트릭 데이터 수집할 때 성능상의 오버헤드가 될 수 있다. 둘째, 제안한 탐지법이 시스템 측정 메트릭을 수집하기에는 많은 용량이 빠르게 채워지는 한계점이 있다. 그러므로 현재 진화하고 있는 변종들을 수집하기 위해서는 런타임에 적은 오버헤드로 대량의 데이터를 수집하는 추가적인 연구가 필요하다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송기술국제공동연구사업

(Project No. RS-2022-00165794, 40%), 국방ICT융합사업(Project No. Project No.2022-0-00701, 10%), 실감콘텐츠핵심기술개발사업(Project No. RS-2023-00228996, 10%), 대학ICT연구센터 육성지원사업(Project No. 2023-2021-0-01816, 10%) 및 한국연구재단 개인기초연구과제(Project No. RS-2023-00208460, 30%)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] 오성택, 신삼신, "악성코드 유포사이트 탐지 기술 동향 조사," 정보보호학회지, vol.33, no.1, pp. 77-88, 2023.

[2] 김기현, 최미정, "안드로이드 환경에서 시계열 기반의 악성코드 탐지 기법과 행동 기반 악성코드 탐지 기법의 비교 분석," 한국통신학회, vol.2015, no.1, pp. 149-150 2015.

[3] D. Ö. Şahın, S. Akleylek and E. Kiliç, "LinRegDroid: Detection of Android Malware Using Multiple Linear Regression Models-Based Classifiers," IEEE Access, vol. 10, pp. 14246-14259, 2022.

[4] Yahye Abukar Ahmed, Barış Koçer, Shamsul Huda, Bander Ali Saleh Al-rimy and Mohammad Mehedi Hassan, "A system call refinement-based enhanced Minimum Redundancy Maximum Relevance method for ransomware early detection," Journal of Network and Computer Applications, vol.167, 2020.

[5] USNW Sydney, UNSW-NB15 Dataset. https://research.unsw.edu.au/projects/unsw-nb15-dataset

[6] Intel, Pin. https://www.intel.com/content/www/us/en/developer/articles/tool/pin-a-dynamic-binary-instrumentation-tool.html

[7] T. N. Nguyen, Q. -D. Ngo, H. -T. Nguyen and G. L. Nguyen, "An Advanced Computing Approach for IoT-Botnet Detection in Industrial Internet of Things," IEEE Transactions on Industrial Informatics, vol. 18, no.11, pp. 8298-8306, 2022.

[8] Mahdi Rabbani, Yong Li Wang, Reza Khoshkangini, Hamed Jelodar, Ruxin Zhao, Peng Hu, "A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing," Journal of Network and Computer Applications, vol.151, 2020.

[9] Çağatay Yücel, Ahmet Koltuksuz, "Imaging and evaluating the memory access for malware", Forensic Science International: Digital Investigation, vol.31, 2020.

[10] Amir Namavar Jahromi, Sattar Hashemi, Ali Dehghantanha, Kim-Kwang Raymond Choo, Hadis Karimipour, David Ellis Newton, Reza M. Parizi, "An improved two-hidden-layer extreme learning machine for malware hunting," Computers & Security, vol.89, 2020.

[11] Linux, Strace. https://strace.io

[12] WireShark Foundation, Wireshark. https://www.wireshark.org/

[13] TCPDump Group, TCPDump. https://www.tcpdump.org/index.html

[14] MANDIANT, Memoryze. https://fireeye.market/apps/211368