안드로이드 랜섬웨어 탐지를 위한 Chi-square test 기반 특징 추출 방안

정재환, 정인웅, 이한진, 최석환*

연세대학교

jjh021101@yonsei.ac.kr, jiw4281@yonsei.ac.kr, 2020253046@yonsei.ac.kr, *sh.choi@yonsei.ac.kr

Chi-square test-based feature extraction method for Android ransomware detection

Jeong Jae Hwan, Jeong In Woong, Lee Han Jin, Choi Seok-Hwan*

Yonsei Univ.

요 약

최근 딥러닝 기술을 적용한 안드로이드 랜섬웨어 탐지 방법이 활발하게 연구되면서, 랜섬웨어 탐지 모델의 실시간 처리 능력과 정확도 향상을 위한 효과적인 특징 추출 기법에 대한 연구 필요성 또한 증가하고 있다. 따라서, 본 논문에서는 Chi-square test 알고리즘 기반의 랜섬웨어 특징 추출 기법을 제안한다. 또한, CIC-AndMal2017 데이터셋을 사용한 실험을 통해 제안하는 특징 추출 기법이 안드로이드 랜섬웨어 탐지에 있어 응답시간을 줄이고 탐지 정확도를 유지함을 검증하였다.

Ⅰ. 서 론

스마트폰의 보편화로 안드로이드 랜섬웨어에 대한 위협이 증가하고 있으며, 이에 대응하는 효과적인 탐지 기법 개발이 중요시되고 있다. 최근에는 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network) 등 딥러닝 기술을 적용한 랜섬웨어 탐지 방법이 활발하게 연구되고 있다. 이러한 딥러닝 기반의 랜섬웨어 탐지 방법들의 실시간 처리능력과 정확도 향상을 위해서는 효과적인 특징 추출 기법에 대한 연구가 필요하다. 따라서, 본 논문에서는 안드로이드 랜섬웨어 탐지의 효율성을 높이기 위한 Chi-square test 알고리즘 기반 특징 추출 기법을 제안한다.

Ⅱ. 본론

2.1 Feature extraction

본 논문에서는 랜섬웨어 탐지의 효율성을 높이기 위해 Chi-square test 알고리즘을 사용하여 랜섬웨어 탐지에 가장 영향력 있는 특징을 추출한다. Chi-square test 알고리즘은 두 범주형 변수 사이의 연관성을 확인하는 통계적 방법으로, 관찰된 빈도가 기대되는 빈도와 의미 있게 다른지를 검정한다[1].

본 논문에서는 Chi-square test 알고리즘의 독립성 검정을 사용하여 양성과 랜섬웨어 샘플 간의 특징이 독립적인지, 연관성이 있는지를 평가하였다. 먼저, 수집된 패킷에서 비수치형 데이터의 영향을 배제하고 순수하게 수치적 분석에 초점을 맞추기 위해 Flow ID, Source IP, Destination IP, Timestamp를 제외한다. 이후, 각 특징이 양성과 랜섬웨어 샘플을 구분하는 데 얼마나 중요한지를 Chi-square test 알고리즘을 통해 수치화하고 가장 유의미한 정보를 제공하는 특징들을 추출하였다. Chi-square test 알고리즘은 식 (1)과 같다. 여기서 O_i는 관찰된 빈도수를 나타내며, 실제 데이터에서 각 범주에 속하는 사례의 수이다. E_i는 기대 빈도수를 나타내며, 귀무가설 하에서 해당 범주에 속할 것으로 기대되는 사례의 수이다. X^2값은 각 범주에 대해 계산된 차이를 제곱하여 기대 빈도수로 나눈 값들의 합이다.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad (1)$$

표 1 Chi-square test 기반 추출된 특징

카테고리	기존 개수	추출 개수
Packet Identification	7	2
Packet Flow	9	6
Packet Size	18	3
Packet Interval	19	16
Packet Flag	12	6
기타	20	3

표 1은 Scikit-learn 라이브러리의 SelectKBest 클래스를 사용하여 각 특징의 Chi-square 값을 기준으로 추출한 전체 85개 중 상위 36개의 특징을 분류한 결과를 나타낸다. 패킷 식별(Packet Identification) 관련 특징 2개, 패킷 흐름(Packet Flow) 관련 특징 6개, 패킷 크기(Packet Size) 관련 특징 3개, 패킷 간격(Packet Interval) 관련 특징 16개, 패킷 플래그(Packet Flags) 관련 특징 6개, 기타 특징 3개가 추출되었으며, 이를 통해 데이터의 복잡성을 줄였다.

2.2 Classification models

본 논문에서는 제안하는 특징 추출 기법의 성능을 평가하기 위해 Decision Tree(DT)[2], k-Nearest Neighbor(k-NN)[3], Convolution Neural Network(CNN)[4], Residual Neural Network(ResNet)[5]을 안드

표 2 모델 성능 비교 결과

		DT	k-NN	CNN	ResNet50
36_Feature	Accuracy	0.6803	0.6045	0.6676	0.6203
	Time(s)	10.56	19.29	4.7135	77.2344
All_Feature	Accuracy	0.6973	0.6157	0.6258	0.6170
	Time(s)	16.61	24.58	4.8541	86.5650

로이드 랜섬웨어 분류 모델로 사용하였다. DT의 경우, 입력된 랜섬웨어 샘플의 특징을 바탕으로 결정 경로를 생성하여 랜섬웨어를 탐지한다. k-NN의 경우, 입력된 랜섬웨어 샘플에 대해 거리를 계산하여 k값에 따라 가장 인접한 샘플의 유형을 결정한다. CNN의 경우, 입력된 랜섬웨어 샘플의 이미지화된 특징에 대해 컨볼루션 연산을 수행하여 랜섬웨어 여부를 판단한다. ResNet의 경우, 깊은 층의 네트워크를 통해 랜섬웨어 샘플의 복잡한 특징을 학습하며 랜섬웨어를 분류한다.

2.3 Experimental setup

본 논문에서는 캐나다 사이버 보안 연구소가 제공하는 CIC-AndMal2017 데이터셋을 실험에 사용하였다. CIC-AndMal2017 데이터셋은 안드로이드 애플리케이션 관련 트래픽 데이터와 다양한 악성 및 정상 행위 패턴을 포함한다[6]. 실험을 위해 양성 및 랜섬웨어 샘플을 각각 200,000개를 선택하였으며, 훈련, 검증, 테스트 데이터는 모델의 학습 능력 평가 및 과적합 방지를 위해 8:1:1 비율로 분할하여 사용하였다.

또한, 본 논문에서는 AMD Ryzen 5 5600X CPU, RAM 32G, NVIDIA GeForce RTX 3060 Ti GPU, Windows 10 Home 64bit 운영체제 사양의 PC에서 실험을 진행하였다. 개발 언어는 Python 3.10.9버전을 사용하였고, 특징 추출 알고리즘은 Python 기반의 라이브러리 Scikit Learn을 사용하였다.

2.4 Experimental result

본 논문에서는 Chi-square test 알고리즘을 사용하여 선별된 특징 집합이 랜섬웨어 탐지 모델의 성능과 응답시간에 미치는 영향을 평가하였다. 여기서 모델의 응답 시간은 실시간 탐지 환경에서의 적용 가능성을 결정하는 중요한 요소로 간주된다.

표 2는 Chi-square test 알고리즘을 통해 추출된 특징을 사용한 머신러닝, 인공신경망 모델들의 성능을 전체 특징을 사용했을 때와 비교한 결과를 나타낸다. 먼저, Chi-square test 알고리즘을 적용한 모델들은 전체 특징을 사용한 모델들보다 더 빠른 응답시간을 보였다. 예를 들어, 제안하는 특징 추출 기법을 적용했을 경우, 40,000개의 테스트 데이터에 대해 ResNet50 모델의 응답시간이 77.2344초로, 전체 특징을 사용했을 때의 86.5650초에 비해 현저히 단축되었다. 이는 특징 추출이 모델의 시간 효율성을 개선하는 데 많은 영향을 미치는 것을 확인할 수 있다. 또한, 제안하는 특징 추출 기법은 응답시간을 줄임에도 불구하고 탐지 랜섬웨어 정확도를 유지함을 확인할 수 있다. 예를 들어, DT 모델과 k-NN 모델은 소폭의 정확도 감소에도 불구하고 응답시간이 줄었으며, CNN과 ResNet50 모델은 선택된 특징을 사용했을 때 더 높은 정확도를 달성하여 전체 특징 대비 성능이 개선되었다.

Ⅲ. 결론

본 논문에서는 Chi-square test 알고리즘 기반의 안드로이드 랜섬웨어 특징 추출 기법을 제안하고 CIC-AndMal2017 데이터셋을 사용해 제안하는 특징 추출 기법의 성능을 평가하였다. Chi-square test 알고리즘을 사용한 특징 추출을 통해, 제안하는 특징 추출 기법은 모델의 응답시간을 감소함과 동시에 탐지 정확도를 유지할 수 있었다. 하지만, 본 연구에서 안드로이드 랜섬웨어 탐지 모델의 평균 정확도가 약 0.6인 것을 고려할 때, 향후 연구에서는 이를 개선하는 데 중점을 둘 필요가 있다. 이를 위해, 더욱 정교한 데이터 전처리 기법, 다양한 특징 선택 알고리즘, 앙상블 학습 방법을 적용하여 모델의 정확도 향상을 위한 연구를 수행하고자 한다.

ACKNOWLEDGMENT

본 연구는 정부(과학기술정보통신부)의 재원으로 한국 연구재단(RS-2023-00243075) 및 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업 (IITP-2023-RS-2023-00259967)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] McHugh, Mary L. "The chi-square test of independence." Biochemia medica 23.2 (2013): 143-149.pp. 129-132.

[2] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." Shanghai archives of psychiatry 27.2 (2015): 130.

[3] Liu, Liu, et al. "Automatic malware classification and new malware detection using machine learning." Frontiers of Information Technology & Electronic Engineering 18.9 (2017): 1336-1347.

[4] Ganesh, Meenu, et al. "CNN-based android malware detection." 2017 international conference on software security and assurance (ICSSA). IEEE, 2017.

[5] Khan, Riaz Ullah, Xiaosong Zhang, and Rajesh Kumar. "Analysis of ResNet and GoogleNet models for malware detection." Journal of Computer Virology and Hacking Techniques 15 (2019): 29-37.

[6] Lashkari, Arash Habibi, et al. "Toward developing a systematic approach to generate benchmark android malware datasets and classification." 2018 International Carnahan conference on security technology (ICCST). IEEE, 2018.