JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 10, pp. 19-26, October 2021
https://doi.org/10.9708/jksci.2021.26.10.019

CNN-based Android Malware Detection Using Reduced Feature Set

Dong-Min Kim*, Soo-jin Lee*

*Student, Dept. of Computer Science and Engineering, Korea National Defense University, Nonsan, Korea
*Professor, Dept. of Computer Science and Engineering, Korea National Defense University, Nonsan, Korea

[Abstract]

The performance of deep learning-based malware detection and classification models depends largely on how to construct a feature set to be applied to training. In this paper, we propose an approach to select the optimal feature set to maximize detection performance for CNN-based Android malware detection. The features to be included in the feature set were selected through the Chi-Square test algorithm, which is widely used for feature selection in machine learning and deep learning. To validate the proposed approach, the CNN model was trained using 36 characteristics selected for the CICANDMAL2017 dataset and then the malware detection performance was measured. As a result, 99.99% of Accuracy was achieved in binary classification and 98.55% in multiclass classification.

▸ **Key words**: CNN, Android, Malware, Feature selection. Binary classification, Multiclass classification

[요 약]

딥러닝 기반 악성코드 탐지 및 분류모델의 성능은 특성집합을 어떻게 구성하느냐에 따라 크게 좌우된다. 본 논문에서는 CNN 기반의 안드로이드 악성코드 탐지 시 탐지성능을 극대화할 수 있는 최적의 특성집합(feature set)을 선정하는 방법을 제안한다. 특성집합에 포함될 특성은 기계학습 및 딥러닝에서 특성추출을 위해 널리 사용되는 Chi-Square test 알고리즘을 사용하여 선정하였다. CICANDMAL2017 데이터세트를 대상으로 선정된 36개의 특성을 이용하여 CNN 모델을 학습시킨 후 악성코드 탐지성능을 측정한 결과 이진분류에서는 99.99%, 다중분류에서는 98.55%의 Accuracy를 달성하였다.

▸ **주제어**: CNN, 안드로이드, 악성코드, 특성추출, 이진분류, 다중분류

• First Author: Dong-Min Kim, Corresponding Author: Soo-jin Lee
*Dong-Min Kim (kdm891107@gmail.com), Dept. of Computer Science and Engineering, Korea National Defense University
*Soo-jin Lee (cyberkma@gmail.com), Dept. of Computer Science and Engineering, Korea National Defense University
• Received: 2021. 08. 24, Revised: 2021. 10. 06, Accepted: 2021. 10. 06.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

코로나 19로 인해 비대면, 원격근무 등이 일상화되면서 모바일 장비의 활용은 예전에 비해 획기적으로 증가했고, 그에 맞춰 스마트폰 등 모바일 기기 내에 저장된 개인정보 및 중요 정보 탈취를 목적으로 한 악성코드 공격이 증가 하고 있다. 2021년 4월 발표된 Check Point사 위협분석보고서[1]에 의하면 거의 모든 조직들이 모바일 악성코드 공격을 경험하였으며, 조상 대상 조직의 46%에서 최소 한 명 이상의 직원이 조직의 네트워크와 데이터를 위협하는 악성코드를 다운로드하였다. 그리고 Kaspersky사 조사에 의하면, 2021년 1분기에만 전년 동기 대비 298,998건이 증가한 총 1,451,660건의 악성코드 설치 프로그램이 탐지되었다[2].

급증하는 모바일 악성코드에 효율적으로 대응하기 위해 다양한 기법들이 연구되고 있으며, 최근에는 기계학습과 딥러닝을 기반으로 악성코드를 탐지 또는 식별하는 연구가 활발하게 진행되고 있다. 특히 지도학습 기법의 일종인 심층신경망(DNN, Deep Neural Network)과 합성곱 신경망(CNN, Convolutional Neural Network, 이하 CNN)은 모바일 악성코드를 신속·정확하게 탐지하기 위한 대안으로 부상하고 있다[3].

대규모 네트워크 트래픽에 포함된 악성코드를 정확하고 신속하게 탐지 또는 분류하기 위해서는 모델 학습에 적용할 악성코드의 대표 특성(feature)을 정확하게 식별하는 특성추출(feature selection)과정이 무엇보다 중요하다[4]. 그러나 특성추출은 악성코드에 대한 전문적인 지식을 필요로 하기 때문에 쉽지 않은 작업이며, 대부분 전문가에 의해 수행된다. 이러한 이유로 최근에는 특성추출을 자동으로 수행하는 CNN이 많은 관심을 받고 있다.

한편, CNN을 기반으로 악성코드 탐지를 시도할 경우 특성추출에 대한 전문지식 없이도 탐지모델을 생성할 수 있지만, 어떤 특성이 악성코드 탐지에 영향을 미치는지는 알 수 없다. 따라서 생성된 탐지모델의 성능을 개선하기 위해 학습에 적용할 특성을 최적화할 수 없으며, 조직의 보안관리자 입장에서는 악성코드 차단을 위해 네트워크 차원의 보안대책을 수립하는 것도 불가능해진다.

이에 본 연구에서는 CNN을 기반으로 모바일 악성코드를 탐지함에 있어 어떤 특성들이 영향을 미치는지와 추출된 일부 특성만으로도 최상의 탐지성능을 달성할 수 있는지를 분석하고자 한다. 특성추출은 가장 널리 사용되는 방법 중 Filter method[5]를 선정하고, ‘Chi-Square test’[6], ‘ANOVA f-test’[7], ‘Mutual information’[8] 3가지 알고리즘을 적용한다. 이어서 추출된 핵심 특성들을 이용하여 만족할 수 있는 수준의 탐지성능이 달성되는지를 CNN 기반의 모델을 이용하여 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 동일한 데이터세트를 대상으로 특성을 추출하여 모바일 악성코드 탐지를 시도했던 기존 연구를 정리한다. 3장에서는 제안하는 특성추출 방법과 추출된 특성 및 CNN을 기반으로 모바일 악성코드를 탐지하는 절차에 대해 설명한다. 4장에서는 실험결과를 정리하고, 마지막으로 5장에서 연구결과를 요약한 후 결론을 맺는다.

II. Preliminaries

1. Related works

A. Arora와 S. Peddoju(2017)는 특성추출 기법 중 ‘Information Gain’과 ‘Chi-Square test’ 두 기법을 이용하여 안드로이드 악성코드 탐지에 영향을 미치는 네트워크 트래픽 특성을 추출하였다. 우선순위가 지정된 특성들 중 상위 22개의 특성과 9개의 특성을 이용하여 탐지성능을 측정한 결과 9개의 특성만을 이용한 모델이 더 높은 정확도로 악성코드를 탐지하면서, 학습 및 테스트에 소요되는 시간을 단축시킴을 확인하였다[9].

A. H. Lashkari 등(2018)은 CICANDMAL2017 데이터세트가 가지는 85개 특성 중 특성추출을 통해 식별된 상위 9개의 특성만을 활용하여, RF(Random Forest), KNN(K-Nearest Neighbors), DT(Decision Tree)를 활용하여 악성코드 분류성능을 분석하였다. 그 결과 이진 분류(Benign, Malware)에서는 Precision 85.80%, Recall 88.30%를 달성하였으나, 악성코드 카테고리를 분류하는 다중분류에서는 Precision과 Recall 모두 50% 미만의 저조한 성능을 보였다[10].

Chen Ret 등(2019)은 CICANDMAL2017 데이터세트에 포함된 각 악성코드 패밀리에서 무작위로 하나의 PCAP 파일을 선정하여 2단계를 걸친 특성추출을 실시하였다. 총 15개의 특성을 추출하여 이진분류 (Benign, Malware) 및 3가지 카테고리(Adware, Ransomware, Scareware)에 대한 다중분류를 시도하였고, 학습은 [10]에서와 동일하게 RF, KNN 및 DT를 활용하였다. 이진분류는 RF 알고리즘을 사용한 경우 분류성능이 가장 우수하게 나타났으며, F1-score, Precision 및 Recall 모두 95% 이상이었다[11].

A. Akhunzada 등(2019)은 랜섬웨어 유형의 악성코드 탐지를 위해 LSTM 기반 딥러닝 프레임워크를 사용하였다. CICANDMAL2017 데이터세트에서 Benign과

Ransomware만 선택하여 이진분류 실험을 진행하였고 Chi-Square test를 포함한 총 8개의 특성추출 기법을 적용하여 상위 19개의 특성을 선정하였다. 제안된 모델의 Accuracy, Recall 및 F1-Score는 모두 97%로 나타났다[12].

A. Mohammad와 M. Khaled(2020)는 RF, RFE (Recursive Feature Elimination) 및 LightGBM (Light Gradient Boosting Machine)의 3가지 특성추출 기법을 결합한 앙상블 학습 기법을 적용하였다. 그 결과 이진분류, 카테고리 및 패밀리 분류에서 각각 87.75%, 79.97%, 66.71%의 Accuracy를 달성하였다[13].

III. The Proposed Scheme

본 논문에서 제안하는 접근방법은 Fig. 1에서 보는 바와 같다. 우선 CICANDMAL2017 데이터세트[10]를 대상으로 Jupyter Notebook 환경에서 작성한 Python Script를 통해 특성추출을 실시하여 36가지의 특성을 선정한다. 이어서 선정된 특성들만을 이용해 16비트 그레이스케일(grey scale) 이미지를 생성하고, 자체적으로 구성한 CNN 기반의 모델을 활용하여 학습을 실시한 후 악성코드를 탐지 및 분류한다.

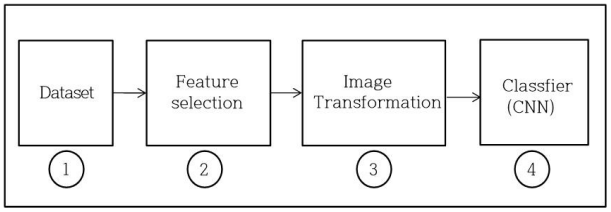


Fig. 1. Proposed System Overview

1. Dataset

본 논문에서 사용한 데이터세트는 A. H. Lashkari 등이 생산하고 발표한 안드로이드 모바일 트래픽의 데이터세트인 CICANDMAL2017이다. 해당 데이터세트는 구글마켓에서 다운로드한 Benign 및 Malware 애플리케이션을 이용하여 85개의 네트워크 트래픽 특성을 가지는 Benign 및 Malware 데이터를 수집하고 정리한 것이다. Malware는 Adware, Ransomware, Scareware 및 Smsware 4개의 카테고리로 구분되며, 각 카테고리는 42개의(각각 10개 또는 11개의 패밀리 포함) 패밀리로 구성되어 있다. 카테고리별 샘플의 수는 Table 1, 카테고리별 패밀리의 구성은 Table 2에서 보는 바와 같다.

Table 1. Total number of samples in CICANDMAL2017 Dataset

Class	Total number of samples
Benign	1,205,117
Adware	424,147
Ransomware	348,943
Scareware	391,644
Smsware	210,998

Table 2. CICANDMAL2017 Malware Classes and Families

Categories(4)	Families (42)
Adware (10)	Ewind, Koodus, Kemoge, Dowgin, Feiwo, Youmi, Selfmite, Shuanet, Mobidash, Gooligan
Ransomware (10)	Charger, Pletor, LockerPin, Jisut, Svpeng, Koler, RansomBO, PornDroid WannaLocker, Simplocker
Scareware (11)	Android, Defender, AndroidSpy, FakeAV, Penetho AVforAndroid,AVpass,,FakeApp, FakeJobOffer, FakeTaoBao, FakeApp.AL, VirusShield
Smsware (11)	Beanbot, Biige, Fakeinst, FakeMart, Jifake, Mazarbot, Nandrobox, Zsone, Plankton, FakeNotify, SMSsniffe

2. Feature selection

기계학습과 딥러닝을 기반으로 예측 및 분류 모델을 만들 때 모델 개발 및 학습을 느리게 만들거나 시스템 메모리를 많이 필요로 하는 특성들이 존재한다. 또한 대상 특성과는 관련성이 떨어지는 특성들을 함께 포함하여 학습을 진행할 경우 생성되는 모델의 성능이 심각하게 저하 될 수 있다. 따라서 대상이 되어야 할 특성을 정확하게 추출하기 위해서 가장 영향력을 미치지 않는 특성의 수를 줄이는 과정이 필수적이다. 그러한 측면에서 본다면 특성추출은 모델 설계에 있어 가장 중요하고 우선적으로 실행되어야 하는 단계이다.

특성추출의 주된 목적은 독립 변수 중, 중복되거나 종속변수와 관련이 없는 변수들을 제거함으로써 종속변수를 가장 잘 예측하는 변수들의 조합을 찾아내는 것이기 때문에, 최적화 문제로도 정의할 수 있다[14]. 특성추출이 제공하는 장점은 학습 시간 단축, 모델의 분산 감소를 통한 원활한 학습 진행 보장, 모델 간소화 및 결과 해석의 용이성 보장 등을 들 수 있다. 그리고 조직의 보안관리자 입장에서는 특성추출을 통해 식별된 특성을 이용하면 네트워크 차원에서의 보안대책 수립이 보다 용이해질 수 있다.

특성추출에 많이 사용되는 3가지 방법은 Wrapper method, Filter method 및 Embedded method이다. 본 연구에서는 3가지 방법 중 불필요한 특성을 제거하고 세트화시키는 Filter method 방법을 적용하였다. 이 방법은

통계적인 계산에 의해 각 특성들을 순위화한 후 그 순위에 따라 데이터세트에 그대로 남거나 제거되도록 한다. Filter method의 대표적인 알고리즘은 Chi-Square test, ANOVA f-test, Mutual information 등이 있다[15].

본 연구에서는 일차로 상기 3가지 알고리즘을 모두 적용하여 특성추출을 시도한 후 특성들이 가지는 가중치를 순위화하였다. 그 결과 가중치 총합 90%를 기준으로 특성을 추출했을 때, Chi-Square test는 36개, ANOVA f-test는 38개, 그리고 Mutual information은 35개 특성이 추출됨을 확인하였다. 이어서 추출된 특성집합을 이용하여 CNN 모델을 학습시키고 탐지성능을 측정한 결과 Chi-Square test 알고리즘을 통해 추출된 36개의 특성을 이용한 경우가 가장 높은 탐지성능을 보여주었다. 가중치 총합을 90%로 설정한 이유는 가중치 총합을 50%에서 95%까지 변화시켜 가면서 탐지성능을 측정했을 때 90%에서 가장 높은 탐지성능이 확인되었기 때문이다.

특성추출에 적용한 서브 데이터세트의 구성은 Table 3에서 보는 바와 같다. 이진분류 실험을 위해 사용한 서브 데이터세트는 'Benign'에서 100,000개, 4개의 'Malware' 카테고리에서 각각 25,000개씩 샘플을 추출하여 총 200,000개로 구성하였다. 다중분류 실험을 위해 사용한 서브 데이터세트는 'Benign'을 제외한 4개의 'Malware' 카테고리에서 각각 100,000개씩 추출하여 총 400,000만개로 구성하였다. 이상과 같이 구성된 서브 데이터세트는 특성추출 과정에서만 사용하였으며, CNN 기반의 탐지 및 분류 모델의 성능평가를 위한 실험에서는 특성추출에 적용된 서브 데이터세트를 제외한 나머지 데이터만 이용하여 서브 데이터세트를 재구성한 후 실험을 진행하였다.

이후에서는 Chi-Square test 알고리즘을 통해 추출된 특성을 이용하여 진행된 실험 결과를 중심으로 기술하며, CICANDMAL2017 데이터세트에서 선정된 이진분류와 다중분류를 위한 특성 36개는 Table 4에서 보는 바와 같다. 1번부터 36번까지의 번호는 가중치 순위를 의미한다.

Table 3. Subdataset for Feature Selection

Experiment	Subdataset	
	Benign	Malware
Binary Classification	100,000	100,000
Multiclass Classification	0	400,000

Table 4. Features extracted by Chi-square Test

	Binary classification	Multi classification
1	Fwd IAT Total	Bwd Header Length
2	Bwd Header Length	Fwd Header Length
3	Fwd IAT Max	Fwd Header Length.(2)
4	Flow Duration	Idle Max
5	Idle Min	Idle Mean
6	min_seg_size_forward	Bwd IAT Max
7	Idle Mean	Flow IAT Max
8	Idle Max	Idle Min
9	Fwd Header Length	Flow Duration
10	Fwd Header Length (2)	Bwd IAT Total
11	Fwd IAT Std	Fwd IAT Max
12	Active Max	Flow IAT Min
13	Active Mean	Fwd IAT Min
14	Active Min	Fwd IAT Total
15	Bwd IAT Max	Flow IAT Mean
16	Bwd IAT Mean	Flow IAT Std
17	Bwd IAT Min	Fwd IAT Std
18	Flow IAT Max	Fwd IAT Mean
19	Fwd IAT Mean	Bwd IAT Std
20	Bwd IAT Std	Bwd IAT Mean
21	Bwd IAT Total	Idle Std
22	Active Std	Bwd IAT Min
23	Flow IAT Std	Subflow Bwd Bytes
24	Idle Std	Total Length of Bwd Packets
25	Fwd Packets/s	Active Max
26	Flow Packets/s	Flow Bytes/s
27	Init_Win_bytes_forward	Active Mean
28	Packet Length Variance	Active Min
29	Flow IAT Min	Active Std
30	Init_Win_bytes_backward	Packet Length Variance
31	Bwd Packets/s	min_seg_size_forward
32	Fwd IAT Min	Flow Packets/s
33	Flow IAT Mean	Fwd Packets/s
34	Total Length of Fwd Packets	Bwd Packets/s
35	Subflow Fwd Bytes	Init_Win_bytes_backward
36	Total Length of Bwd Packets	Total Length of Fwd Packets

3. Image Transformation

특성추출을 통해 선정된 36개의 특성에 해당하는 값들을 기반으로 CNN 모델에 입력될 이미지를 생성한다. 이미지를 생성하기 전 각 특성의 값은 0~1 사이의 값으로 정규화하여 6×6 크기의 2차 행렬로 변환하였다. 이어서 행렬을 1채널 16비트 PNG 이미지로 변환하였다.(Fig. 2.)

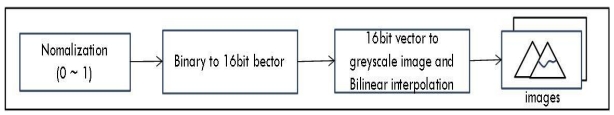


Fig. 2. Image Transformation Process

본 논문에는 사용하는 데이터 값은 1자리 수부터 크게는 10자리 이상의 실수형 데이터로 구성되어 있음을 고려하여 각 데이터의 특성정보 손실을 최대한 줄이기 위해 16비트 이미지로 생성하였다.(Fig. 3.)

	1	2	3	4	5	6
1	146	0	0	65535	123	56
2	0	0	56	45	29856	45596
3	0	7	0	0	45	0
4	32	65535	9	65535	24	0
5	78	65535	0	0	0	0
6	65535	0	0	15698	4526	0

Fig. 3. 6×6, 16bit Image Matrix

4. Clssifier (CNN)

생성된 이미지를 CNN 신경망에 입력 데이터로 하여 실험에 적용하였으며, 모델의 신경망 구조는 Table 5에서 보는 바와 같다. 입력되는 이미지의 크기는 (28×28×1)이며, 3개의 컨볼루션 계층(Convolution layer)과 3개의 맥스풀링 계층(Maxpooling layer), 3개의 완전연결 계층(Fully-connected layer)를 배치하였다. 출력층의 활성화 함수는 소프트맥스(Softmax)를 사용하였고, Optimizer는 Adam을 사용하였다.

IV. Experimental Results

본 장에서는 특성추출 과정을 통해 추출된 36개의 특성들을 이용하여 CICANDMAL2017 데이터세트를 새롭게 구성한 후 진행한 실험 결과를 정리한다.

Table 5. CNN Model Structure

#	Layer	Option	Activation Function
1	Convolution 1	Filter = 64, Size = 3 Batchnormalization	Relu
2	Maxpooling 1	Size = 2, Stride = 2	/
3	Convolution 2	Filter = 128, Size = 3 Batchnormalization	Relu
4	Maxpooling 2	Size = 2, Stride = 2	/
5	Convolution 3	Filter = 256, Size = 3 Batchnormalization	Relu
6	Maxpooling 3	Size = 2, Stride = 2	/
7	Flatten	-	-
9	FC 1	Neurons = 2,048 Dropout = 0.5	Relu
10	FC 2	Neurons = 2,048 Dropout = 0.5	Relu
11	FC 3	Neurons = The number of class	Softmax

1. Environment

본 논문의 모든 실험은 Windows 10 Home 64bit 운영체제, AMD Ryzen 7 3700U CPU, RAM 16G 사양의 PC에서 진행하였다. 개발언어는 Python 3.8.5버전을 사용하였고, 특성추출 알고리즘은 Phython 기반의 라이브러리 Scikit Learn을 사용하였다.

2. Experimental dataset

실험은 'Benign'과 'Malware'를 구분하는 이진분류와 4개의 Malware 카테고리를 분류하는 다중분류로 구분하여 진행하였다. 서브 데이터세트는 앞서 3장 2절에서 설명한 것처럼 특성추출에서 사용한 데이터를 제외한 나머지 데이터만을 사용하여 구성하였다. 이진분류 실험에서는 'Bengin'은 특성추출에서 사용된 100,000개를 제외한 약 110만개, 'Malware'는 특성추출에서 사용된 500,000개를 제외한 약 87만개의 데이터에서 무작위 추출하였다. 다중분류 실험에서는 이진분류 실험에 사용된 48,000개를 제외한 데이터에서 무작위 추출하여 중복성을 완전히 제거한 서브 데이터세트를 구성한 후 실험을 진행하였다.

이진분류 실험에 사용한 학습(Train) 서브 데이터세트 구성은 다음과 같다. 'Benign'은 160,000개, 'Malware'는 4개 카테고리에서 각각 10,000개씩의 샘플을 추출하여 총 40,000개로 구성하였다, Validation과 Test 서브 데이터세트는 'Benign' 16,000개, 'Malware' 카테고리에는 각각 1,000개씩 총 4,000개의 샘플을 추출하였다. Train, Validation 및 Test 서브 데이터세트를 구성함에 있어 Benign과 Malware 클래스의 비율은 현실세계에서 정상파일과 악성파일의 비율이 8:2로 관찰된다는 과거 연구결과를 참조하여 8:2 비율로 구성하였다[16].

다중분류 실험에 사용된 Train 서브 데이터세트 구성은 다음과 같다. 4개의 Malware 카테고리에서 각각 20,000개씩 추출하여 총 80,000개로 구성하였으며, Validation 및 Test 서브 데이터세트는 각각 2,500개의 샘플을 추출하여 10,000개로 구성하였다. 이상과 같은 과정을 거쳐서 총 10개의 서브 데이터세트를 생성하였으며, 세부 현황은 Table 6에서 보는 바와 같다.

Table 6. Experimental Subdataset

* B : Benign, M : Malware

Experiment	Train		Validation		Test	
	B	M	B	M	B	M
Binary Classification	160,000	40,000	20,000	4,000	20,000	4,000
Multiclass Classification	0	80,000	0	10,000	0	10,000

3. Binary Classification

'Benign'과 'Malware'를 분류하는 이진분류 실험의 결과는 Fig. 4. 와 Table 7에서 보는 바와 같다. Chi-Square test 알고리즘을 이용하여 추출한 36개 특성들로 신경망 모델을 학습시키고 테스트를 진행한 결과 Accuracy, Precision, Recall 및 F1-score 모두 99.99% 이상을 달성하였다. ANOVA f-test 와 Mutual information 알고리즘을 통해 추출된 특성을 이용한 경우에도 이진분류 성능은 동일하게 나타났다.

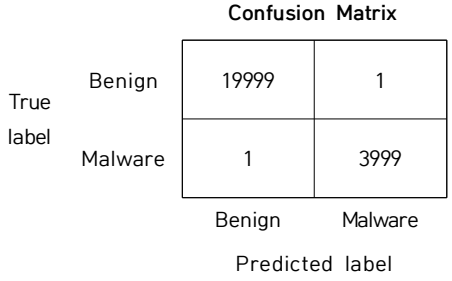


Fig. 4. Confusion Matrix of Binary Classification

Table 7. Result of Binary Classification

#	Accuracy	Precision	Recall	F-1 Score
Proposed	99.99%	99.99%	99.99%	99.99%

4. Multiclass Classification

'Benign'이 제외된 서브 데이터세트를 대상으로 36개의 특성을 이용하여 'Adware', 'Ransomware', 'Scareware' 및 'Smsware' 4개 카테고리로 분류하는 다중분류 실험의 결과는 Table 8에서 보는 바와 같다. 평균 Accuracy는 98.55%로 확인되었으며, ANOVA f-test 및 Mutual information 알고리즘으로 선정한 특성을 이용할 경우의 평균 Accuracy는 각각 85% 및 87%로 확인되었다.

Table 8. Result of Multiclass Classification

Category	Precision	Recall	F-1 Score
Adware	99.28%	99.52%	99.40%
Ransomware	100%	99.24%	99.62%
Scareware	98.39%	97.76%	98.07%
Smsware	98.13%	98.88%	98.51%
Average	98.95%	98.85%	98.90%
Accuracy	98.55%		

5. Comparison with Previous studies

제안하는 접근방법에서 Chi-Square test 알고리즘을 통해 선정한 36개의 특성이 탐지성능 향상에 기여하는 최적화된 특성집합인지를 살펴보기 위해 기존 연구들과의 비교를 실시하였다. CICANDMAL2017 데이터세트를 이용하여 특성추출을 진행한 후 이진분류 및 다중분류를 시도했던 연구들과의 성능지표 비교 결과는 Table 9와 Table 10에서 확인할 수 있다.

Table 9에서 보는 바와 같이 이진분류에서는 제안하는 접근방법을 적용했을 때 모든 성능평가 지표가 우수하게 나타났고, 패킷 시퀀스(sequence) 생성을 위한 특성 8개를 제외한 77개의 특성을 사용한 접근방법[16]과는 대등한 성능을 보이는 것으로 확인되었다. 다중분류의 경우에도 제안하는 접근방법이 모든 지표에서 기존 연구보다 월등하게 성능이 향상되었음을 확인하였다.

Table 9. Comparison of Binary Classification Performance

Model	# of Used Features	Precision	Recall	F-1 score	Accuracy
Proposed	36	99.99%	99.99%	99.99%	99.99%
[10]	9	85.80%	88.30%	-	-
[11]	15	95.00%	95.00%	95.00%	-
[13]	9	89.35%	85.33%	-	87.75%
[16]	77	99.97%	99.99%	97.97%	

Table 10. Comparison of Multiclass Classification Performance

Model	# of Used Features	Precision	Recall	F-1 score	Accuracy
Proposed	36	98.95%	98.55%	98.90%	98.55%
[10]	9	49.90%	48.50%	-	-
[11]	15	86.00%	85.00%	86.00%	-
[13]	9	80.20%	79.91%	-	79.91%
[16]	77	96.20%	95.98%	95.97%	-

탐지성능 비교에 추가하여 축소된 특성집합이 탐지모델 생성에 미치는 영향을 분석하기 위해 생성되는 이미지의 크기를 확인하고 모델의 학습에 소요되는 시간을 측정하였다. 특성집합의 크기가 작아질수록 생성되는 이미지의 크기 역시 작아졌으며, 77개의 특성을 사용했던 접근방법[16]과 비교하면 파일 크기가 평균 22% 정도 축소되었다.

학습소요시간 역시 특성집합이 작을수록 단축됨을 확인하였다. CNN 모델을 생성하고 Table 6에 제시된 데이터세트를 기반으로 77개 및 36개의 특성을 이용하여 학습을 10회 반복 진행한 결과 36개의 특성만을 학습시킨 경우가 평균 38% 정도 학습소요시간이 단축되었다.

물론 이러한 결과는 실험환경이 달라지면 격차가 줄어들거나 늘어날 수도 있다. 그러나 특성집합이 축소될 경우 CNN 모델에 입력되는 이미지의 크기, 즉 차원이 축소되어 학습소요시간이 단축된다는 사실은 명확하다. 따라서 정교

한 특성추출을 통해 학습에 적용한 특성집합을 최적화한다면 모델의 탐지성능 향상과 신속한 탐지모델 생성이 동시에 가능해진다.

Ⅴ. Conclusions

본 논문에서는 안드로이드 네트워크 트래픽 데이터에서 탐지에 영향을 적게 미치는 특성을 제거하여 축소된 특성집합을 구성한 후 CNN을 기반으로 신속하게 학습모델을 구축하고 악성코드 탐지성능을 향상시킬 수 있는 방법을 제안하였다.

Filter method의 일종인 Chi-Square test, ANOVA f-test 및 Mutual information 3가지의 특성추출 알고리즘을 통해 특성집합을 구성하고 1차적인 탐지성능 비교를 실시하였다. 그 결과 Chi-Square test 알고리즘을 통해 선정된 36가지 특성이 가장 높은 탐지성능을 보여 학습에 사용할 특성집합으로 채택하였다. 선정된 36개의 특성에 해당하는 값을 추출하여 1차원(1×36)에서 2차원 행렬(6×6)로 변환한 후, 1채널 16비트 그레이스케일 PNG 이미지를 생성하였다. 이어서 생성된 이미지를 CNN 모델에 학습시키고 탐지성능 확인을 위한 테스트를 진행하였다.

Ⅳ장 실험결과에서 확인한 바와 같이 본 논문에서 제안하는 접근방법은 동일한 데이터세트를 사용한 기존 연구들에 비해 이진분류 및 다중분류 모두에서 향상된 탐지성능을 보여주었다. 그리고 77개의 특성집합을 적용하여 우수한 탐지성능을 달성했던 접근방법[16]과 비교했을 때는 학습모델 생성에 소요되는 시간도 크게 단축됨을 확인하였다.

향후에는 본 연구를 확장하여 4개의 카테고리에 속하는 42개의 악성코드 패밀리를 정확하게 분류하기 위한 특성집합을 선정하는 연구를 진행할 예정이다.

REFERENCES

[1] Check Point, “MOBILE SECURITY REPORT 2021", U.S. Headquarters959 Skyway Road, Suite 300, San Carlos, CA 94070, Apr. 2021.

[2] Securelist. by Kaspersky, “IT threat evolution Q1 2021. Mobile statistics”, https://securelist.com/it-threat-evolution -q1-2021-mobile- statistics/102547/

[3] Byeon, J. Y., Kim, D. H., Kim, H. C., and Choi, S. Y. “RFA : Recursive Feature Addition Algorithm for Machine Learning-Based Malware Classification”, Vol. 26, No. 2, pp.61-68, February. 2021. DOI 10.9708/ jksci.2021.26.02.061.

[4] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. “Feature selection: A data perspective", ACM Computing Surveys (CSUR) Vol.50. No. 6 pp.1-45, December 2017. DOI 10.1145/3136625

[5] Chandrashekar, G. and Sahin, F. “A survey on feature selection methods", Computers & Electrical Engineering vol 40. No. 1 pp. 16-28 January 2014. DOI 10.1016/ j.compeleceng.2013.11.024

[6] Thaseen, I. S., Kumar, C. A., and Ahmad, A. "Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers", Arabian Journal for Science and Engineering Vol 44, No 4 pp.3357-3368, August 2018.

[7] Kumar, B. J., Naveen, H., Kumar, B. P., Sharma, S. S., and Villegas, J. "Logistic regression for polymorphic malware detection using ANOVA F-test", 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). IEEE, pp.1-5, March 2017. DOI 10.1109 /ICIIECS.2017.8275880

[8] Hoque, N., Bhattacharyya, D. K. and Kalita, J. K. "MIFS-ND: A mutual information-based feature selection method", Expert Systems with Applications VOl 41 No. 14 pp.6371-6385. October 2014. DOI 10.1016/j.eswa.2014. 04.019

[9] Arora, A. and Peddoju, S. K. "Minimizing network traffic features for android mobile malware detection", Proceedings of the 18th International Conference on Distributed Computing and Networking. pp.1-10, January 2017. DOI:10.1145/3007748.3007763

[10] Lashkari, A. H., Kadir, A. F. A., Taheri, L. and Ghorbani, A. A. “Toward Developing a Systematic Approach to Generate Benchmark Android malware Datasets and Classification", Proc. of the 2018 International Carnahan Conference on Security Technology, pp. 1-7, 2018. DOI 10.1109/CCST.2018.8585560

[11] Chen, R., Li, Y. and Fang, W. “Android malware identification based on traffic analysis”, International Conference on Artificial Intelligence and Security. Springer, Cham, pp. 293-303. 2019. DOI 10.1007/978-3-030-24274-9_26

[12] Bibi, I., Akhunzada, A., Malik, J., Ahmed, G. and Raza, M. “An effective Android ransomware detection through multi-factor Feature filtration and recurrent neural network”, 2019 UK/China Emerging Technologies (UCET). IEEE, pp 1-4, August 2019. DOI 10.1109/UCET. 2019.8881884

[13] Abuthawabeh, M. and Mahmoud, K. “Enhanced android malware detection and family classification, using conversation-level network traffic Features”, The International Arab Journal of Information Technology, Vol 17, No.4A pp.607-614. June 2020. DOI:10.34028/iajit/17/4A/4

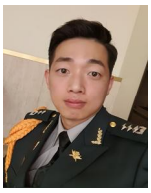
[14] Guyon, I. and Elisseeff, A. “An introduction to variable and feature selection.", Journal of machine learning research 3 Vol 3. No. March pp. 1157-1182. March 2003.

[15] Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D. and Saeed,

J. “A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction.", Journal of Applied Science and Technology Trends Vol 1, No. 2, pp.56-70, May 2020. DOI 10.38094/ jastt1224

[16] J. Kang, S. Lee “Android Malware Detection Through the Conversion of Network Traffic to Images”, Journal of KIISE, Vol. 47, No. 8, pp. 761-768, August 2020. DOI : 10.5626/JOK. 2020.47.8.761

Authors



Dong-Min Kim received the B.A. degree in Military history from Korea Army Academy at Yeongcheon, Korea, in 2013. He is now working toward a M.S. degree at Department of Computer Science and Engineering, Korea National Defense University Dong-Min Kim is interested in malware and machine learning among cyber security fields

Soo-jin Lee received the B.S. degree from Korea Military Academy in 1992, M.S. degree in Computer Science from Yonsei University in 1996, and Ph.D. degree in Computer Science and Engineering from Korea Advanced Institute of Science and Technology in 2006. Dr. Lee is currently a Professor in the Department of Computer Science and Engineering at Korea National Defense University, Nonsan, Korea, from 2006. He is interested in Cyber Warfare and Cyber Security Policy, Intrusion Detection System, Mobile Network Security, Encryption theory and applications