주성분 분석을 활용한 안드로이드 악성코드 분류 성능 향상 방안

전동하O, 이수진(교신저자)*
O국방대학교 국방과학학과,
*국방대학교 국방과학학과
e-mail: acenoma@naver.comO, cyberkma@gmail.com*

Performance Enhancement of Android Malware Classification using PCA

Dong-Ha JeonO, Soo-Jin Lee(Corresponding Author)*
ODept. of Defence Science, Korea National Defense University,
*Dept. of Defence Science, Korea National Defense University

● 요 약 ●

최근 API Call을 기반으로 하는 악성코드 탐지 및 분류에 대한 연구가 활발히 진행되고 있다. 그러나 API Call 기반의 데이터는 방대한 양과 다양한 차원의 특성으로 인해 분석과 학습 모델 구축 측면에서 비효율적인 한계가 있다. 이에 본 연구에서는 방대한 API Call 정보를 포함하고 있는 CICAndMal2020 데이터세트를 대상으로 기존의 특성 선택 기법이 아닌 주성분 분석(Principal Component Analysis)을 사용하여 차원을 대폭 축소 시킨 후 머신러닝 기법을 적용하여 분류를 시도하였다. 실험 결과 전체 9,503개의 특성을 25개의 주성분(전체 대비 약 0.26% 수준)으로 축소시키고 다중 분류 기준 약 84%의 정확도를 나타냈다. 결과적으로 기존 연구에서의 탐지 모델 대비 정확도, F1-score 등의 성능 향상은 물론 차원 축소 측면에서 매우 향상된 결과를 달성하였다.

키워드: API-Call, 주성분분석(PCA), 차원축소(Dimensionality reduction), Malware Classification

I. Introduction

2022년 5월 기준 안드로이드의 글로벌 모바일 운영체제 시장 점유율은 71.45%(아시아 기준 81.36%), 한국에서의 점유율은 72.19%이다. 안드로이드 운영체제는 특유의 개방성으로 접근이 용이하기 때문에 안드로이드를 활용한 악성코드 유포는 그 수와 방법이 더욱 다양해지고 있다.

최근 API Call 정보를 활용한 악성코드 침입 탐지 모델 연구가 많은 관심을 받고 있다. 그러나 API Call 기반 데이터세트는 데이터양이 방대하고 차원이 크기 때문에 악성코드 분류 및 탐지 정확도에도 영향을 미치고, 시간과 비용 측면에서 비효율적인 한계가 있다.

본 연구에서는 데이터 특성의 연관성을 고려한 주성분을 추출하여 데이터 차원을 축소하는 주성분 분석(PCA, Principal Component Analysis)과 머신러닝 기법을 이용하여 API Call 기반의 악성코드 분류 성능을 향상시킬 수 있는 방안을 제안한다.

실험은 캐나다 사이버보안센터와 보안연구소가 제작한 대표적인 안드로이드 악성코드인 CICAndMal2020을 이용하였다[1].

II. Related works

CICAndMal2020 데이터세트에 다양한 특성 선택 및 추출 기법을 사용하여 효율적인 악성코드 탐지 및 분류 방안을 찾는 연구가 다양하게 이루어지고 있다. [2]의 연구에서는 다양한 특성 선택 방법을 적용하여 API Call 정보에 대한 차원을 축소시킨 후, 핵심 특성 집합을 추출하는 방안을 제시하였다.

[3]의 연구에서는 저사양 IoT 디바이스 영역에 주성분 분석(PCA)을 이용하는 선형연산 기반의 저복잡도 이상탐지 기술을 제안하였고 기존 연구보다 향상된 탐지 성능을 달성하였다.

III. The Proposed Scheme

3.1 Data

CICAndMal2020은 195,623개의 악성코드, 14개의 카테고리(Malware)로 구분되어 있고 총 9,503개의 특성정보를 가지고 있다. 정상파일(Benign)은 악성코드 데이터 수를 고려하여 Androzoo 데이

터 162,901개(5개 카테고리)를 수집하였다. 실험에 사용된 학습 및 테스트 서브 데이터세트는 Malware(7,200개)과 Benign(7,000개)을 8:2의 비율로 10번 반복 샘플링하였다.

3.2 Method

총 9,503개의 특성 중 주성분 분석을 통해 새로운 주성분을 추출하여 전체 데이터세트의 차원을 축소시켰다. 학습 데이터세트를 대상으로 추출된 주성분을 적용시키고 Random forest 등 다양한 머신러닝 기법을 활용하여 Benign과 Malware를 분류하였다. 전체적인 흐름은 Fig. 1과 같다.

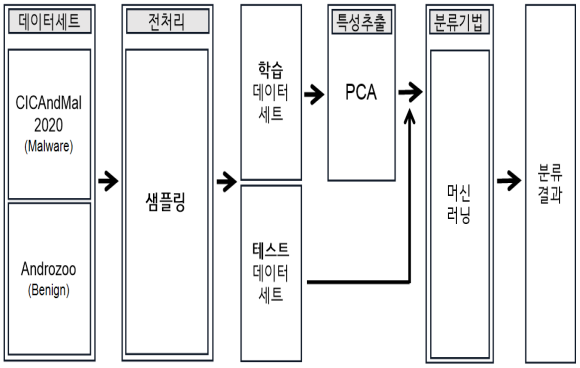

Fig. 1. Flowchart of the proposed Method

3.3 Results

실험결과는 Table 1에서 보는 바와 같으며, 전체 9,503개의 특성을 주성분(PC) 25개로 축소했을 때 99.99%의 분산도를 보였다. 해당 결과값은 LightGBM 기준, 전체 9,503개의 특성으로 실험한 결과값(정확도 0.8507, F1-score 0.8493) 대비 정확도는 0.0104, F1-score는 0.0090 낮은 값이다.

CICAndMal2020 데이터세트에서 최대 8%의 특성을 사용한 기존 연구[4]의 결과값은 정확도 0.8340, F1-score 0.8226이다. 본 연구에서 제안한 기법이 데이터 특성을 전체 대비 0.26%로 대폭 줄이면서 더욱 향상된 결과값을 보였다.

Table 1. Experimental Results

구 분		PC=5	PC=10	PC=15	PC=25
분산도(%)		99.91	99.98	99.98	99.99
Light GBM	정확도	0.7764	0.8097	0.8201	0.8403
	F1-score	0.7763	0.8088	0.8194	0.8397
Random Forest	정확도	0.7903	0.8194	0.8313	0.8417
	F1-score	0.7902	0.8187	0.8313	0.8419
KNN	정확도	0.7438	0.7458	0.7854	0.7931
	F1-score	0.7422	0.7448	0.7849	0.7924

IV. Conclusions

CICAndMal2020 데이터세트를 대상으로 주성분 분석을 활용하여 새롭게 추출된 특성에 다양한 머신러닝 기법을 적용하여 악성코드를 분류하는 기법을 시도하였다.

주성분을 25개로 감소시켜 해당 데이터세트의 전체 특성 차원을 0.26%로 줄이면서도 기존연구 대비 정확도와 F1-score 측면에서 우수한 성능 향상을 보였다. 이러한 결과는 주성분 분석이 많은 계산 비용과 처리시간을 필요로 하는 API Call 기반의 악성코드 탐지 및 분류에 있어 유용한 기법임을 확인시켜준다.

REFERENCES

[1] David Sean Keyes, Beiqi Li, Gurdip Kaur, Arash Habibi Lashkari, Francois Gagnon, Frederic Massicotte, "EntropLyzer: Android Malware Classification and Characterization Using Entropy Analysis of Dynamic Characteristics", Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge (RDAAPS), IEEE, Canada, ON, McMaster University, 2021

[2] Heejin Whang, and Soojin Lee, "Dimensionality Reduction of Feature Set for API Call Based Android Malware Classification" The journal of The Korea Society of Computer and Information, Vol. 26, No. 11, pp. 41-49, November. 2021.

[3] Hyoseon Kye, and Minhae Kwon, "PCA-Based Low-Complexity Anomaly Detection", The Journal of Korean Institue of Communications and Information Science, Vol. 46, No. 06, pp. 941-955, June. 2021.