모바일 앱 악성코드 분석을 위한 학습모델 제안

배세진* · 최영렬 · 이정수 · 백남균

부산외국어대학교

Proposal of a Learning Model for Mobile App Malicious Code Analysis

Se-jin Bae* · Young-ryul Choi · Jung-soo Rhee · Nam-kyun Baik

Busan University of Foreign Studies

E-mail : qotpqo@naver.com / dudfuf261@naver.com / rhee@naver.com / namkyun@bufs.ac.kr

요 약

앱(App) 또는 어플리케이션이라고 부르는 응용 프로그램은 스마트폰이나 스마트TV와 같은 스마트 기기에서 사용되고 있다. 당연하게도 앱에도 악성코드가 있는데, 악성코드의 유무에 따라 정상앱과 악성앱으로 나눌 수 있다. 악성코드는 많고 종류가 다양하기 때문에 사람이 직접 탐지하기 어렵다는 단점이 있어 AI를 활용하여 악성앱을 탐지하는 방안을 제안한다. 기존 방법에서는 악성앱에서 Feature를 추출하여 악성앱을 탐지하는 방법이 대부분이었다. 하지만 종류와 수가 기하급수적으로 늘어 일일이 탐지할 수도 없는 상황이다. 따라서 기존 대부분의 악성앱에서 Feature을 추출하여 악성앱을 탐지하는 방안 외에 두 가지를 더 제안하려 한다. 첫 번째 방안은 기존 악성앱 학습을 하여 악성앱을 탐지하는 방법과는 반대로 정상앱을 공부하여 Feature를 추출하여 학습한 후 정상에서 거리가 먼, 다시 말해 비정상(악성앱)을 찾는 것이다. 두 번째 제안하는 방안은 기존 방안과 첫 번째로 제안한 방안을 결합한 '앙상블 기법'이다. 이 두 기법은 향후 앱 환경에서 활용될 수 있도록 연구를 진행할 필요가 있다.

ABSTRACT

App is used on mobile devices such as smartphones and also has malicious code, which can be divided into normal and malicious depending on the presence or absence of hacking codes. Because there are many kind of malware, it is difficult to detect directly, we propose a method to detect malicious app using AI. Most of the existing methods are to detect malicious app by extracting features from malicious app. However, the number of types have increased exponentially, making it impossible to detect malicious code. Therefore, we would like to propose two more methods besides detecting malicious app by extracting features from most existing malicious app. The first method is to learn normal app to extract normal's features, as opposed to the existing method of learning malicious app and find abnormalities (malicious app). The second one is an 'ensemble technique' that combines the existing method with the first proposal. These two methods need to be studied so that they can be used in future mobile environment.

키워드

App, malware, malicious code, feature, ensemble technique

Ⅰ. 서 론

앱(App) 또는 어플리케이션이라고 부르는 응용 프로그램은 스마트폰이나 스마트TV와 같은 스마트 기기에서 사용되고 있다. 앱에는 악성코드가 있는데, AhnLab의 최근 3개월 간 탐지된 모바일 악성코드는 약 480,000건으로 그 수가 계속해서 증가하고 있는 추세이다. 특히, Android 앱은 오픈소스 기반이기 때문에 해커가 변형된 악성코드를 쉽게 제작할 수 있어 종류가 계속해서 증가하고 있고, 모바일 사용이 증가함에 따라 악성코드에 감염된 악성앱을 다운받는 횟수가 증가하고 있다. 이처럼 악성코드는 종류가 다양하고 많기 때문에 사람이 일일이 탐지하기 어렵다는 점이 있다. 따라서 AI를 이용하여 악성코드를 탐지하는 방안을 제안한다. 앱은 악성코드의 유무에 따라 정상앱과 악성앱으로 나눌 수 있다. 기존에는 악성앱에서 Feature를 추출하여 악성앱을 탐지하는 방법이 대부분이었다.

* speaker

하지만 악성코드는 그 종류와 수가 기하급수적으로 늘어 완벽히 탐지한다는 것은 사실상 불가능하다. 따라서 기존 대부분의 악성앱에서 Feature을 추출하여 악성앱을 탐지하는 방안 외에 두 가지를 더 제안하려 한다. 첫 번째 방안은 악성앱에서 Feature를 추출하여 학습한 후 악성앱을 탐지하는 방법과는 반대로, 정상앱에서 Feature를 추출하여 학습한 후 정상에서 거리가 먼 비정상(악성앱)을 찾는 것이다. 두 번째 제안하는 방안은 기존 방안과 첫 번째로 제안한 방안을 결합한 '앙상블 기법'이다[1].

본 논문에서는 두 가지 방안으로 악성앱을 탐지하는 기법을 제안한다. 2장에서는 관련된 기존 연구에 대해 살펴보고, 3장에서는 기존 방안을 개선한 새로운 탐지 방안을 소개하고, 마지막 4장에서는 결론을 맺는다.

II. 관련 연구

2.1 기존 연구

행위기반 탐지 및 머신러닝 기법 적용을 통한 악성코드 탐지, 빠르게 변화하는 악성 앱의 특성을 찾아내고 모델에 적용하여 악성앱을 탐지하는 등, 기존 대부분의 연구에서는 악성앱에서 특징(feature)을 추출하여 학습한 후 악성앱을 탐지하는 방법이 대부분이다[2][3].

2.2 이상 탐지 기법

이상 탐지(Anomaly Detection)란 대다수의 데이터(악성코드 앱)와 다른 양상을 보이는, 차이나는 특성을 찾아내는 기술을 의미한다. 이상탐지는 지도 학습, 반지도 학습, 비지도 학습 방식 모두와 연관이 있다. 학습 전에 전체 데이터를 정상과 이상으로 명확히 구분이 가능한 경우(레이블) 지도 학습 방식을, 정상 데이터만 명확히 구분이 가능한 경우 반지도 학습 방식을 사용한다. 정상과 이상 데이터 중 어떠한 데이터도 명확히 분류가 어려운 상황에서는 비지도 학습 방식을 사용한다. 비지도 학습 기반 이상 탐지 모델은 우선 정상 데이터의 패턴을 분석해 찾아내고, 그 패턴과 다른 양상을 보이거나 패턴으로부터 거리가 멀리 떨어진 데이터를 이상 데이터로 판별한다. '이상'을 주장하려면 먼저 '정상'에 대한 기준과 근거 데이터가 필요하다. 특정 데이터가 정확히 정상인지 혹은 이상인지 단순히 분류하는 것이 아니라 '이상 데이터일 가능성'을 예측하는 방식을 사용한다[4].

LOF (Local Outlier Factor) 알고리즘은 주어진 데이터와 이웃한 데이터들이 근처의 밀집 데이터 영역으로부터 얼마나 떨어져 있는지를 토대로 이상 데이터를 탐지하는 비지도 기반 학습 방법이다[5].

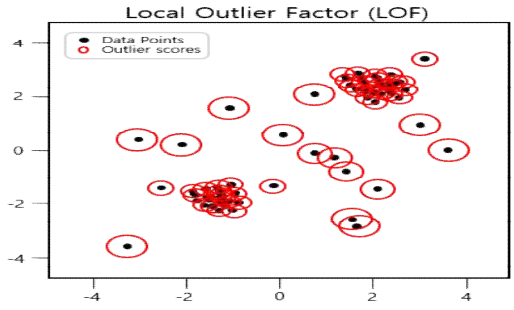

그림 1. LOF(Local Outlier Factor)

Isolation Forest는 소수 범주(이상치)는 개체수가 적고, 소수 범주 데이터는 정상 범주 데이터와는 특정 속성 값이 많이 다를 가능성이 있다는 특성을 이용한다. 이상치 탐지를 위해 먼저 하나의 객체를 고립(isolation) 시키는 tree를 생성한다. 특정 개체가 고립되는 말단 노드까지의 거리를 이상치 점수로 정의하고, 고립되는 말단 노드까지의 거리가 짧을수록 이상치 점수가 높아지도록 설정한다. 입력 데이터에 대한 이상치 점수가 특정 threshold 보다 높을 경우 이상치 데이터로 탐지한다[6].

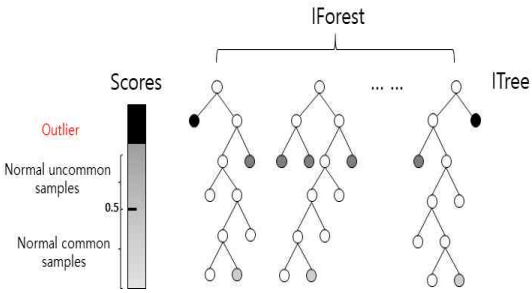

그림 2. Isolation Forest

III. 제안 방안

기존 연구에서 보았듯이, 대부분의 악성코드가 포함되어 있는 악성앱 탐지 방안은 악성앱에서 특징(feature)을 추출하여 학습한 후 악성앱을 탐지하는 방안이 대부분이었다. 행위기반 탐지 및 머신러닝 기법 적용을 통하여 악성코드를 탐지하는 기법은 변조 혹은 새로운 악성코드에 대한 탐지가 가능하다는 장점이 있지만, 빠른 속도로 종류와 수가 증가하는 악성코드의 모든 행위를 기반으로 탐지할 수 없다는 단점이 있다. 또한, 악성 안드로이드 앱 탐지를 위한 개선된 특성 선택을 통한 악성앱을 탐지하는 방안은 다양한 특성 선택 방법의 조합과 가중치를 부과하여 악성앱 탐지율을 높일 수 있다는 장점이 있지만, 특성이 변경된 신종 악성 앱을

탐지하기에는 어려움이 있다. 따라서 기존 방안의 단점을 보완할 수 있는 다른 두 가지 방안을 제시하고자 한다.[2][3]

첫 번째 방안은 기존 방안과는 반대로 정상앱에서 특징을 추출하여 학습한 후 정상에서 거리가 먼, 다시 말해 악성앱을 찾는 방안이다. 이 방안은 기존 악성앱을 기준으로 탐지하던 방안과는 다르게 정상앱을 기준으로 하여 악성앱을 탐지하는 기법이기 때문에 정상앱의 Feature를 추출하여 추출한 Feature를 바탕으로 이상 탐지 알고리즘을 이용하여 AI에 학습시켜 악성앱을 탐지한다면, 계속해서 증가하고 있는 악성코드를 기존보다 정확히 탐지함에 따라 탐지율이 높아질 것이다.

두 번째 방안은 악성앱에서 특징을 추출하여 학습한 후 악성앱을 탐지하는 기존 방안과 새롭게 제안한 정상앱에서 특징을 추출하여 학습한 후 정상에서 거리가 먼 악성앱을 찾는 두 번째 방안을 결합한 앙상블 기법(ensemble technique)이다. 앙상블 기법은 기존방안과 새롭게 제안한 방안의 장단점을 보완한 기법이다. 이 방안은 기존 방안과 새로운 방안을 서로 보완하기 때문에 이전보다 정확도와 탐지율 면에서 크게 증가할 것으로 예상한다.

그림 3. 악성앱 탐지 순서

악성앱의 Feature를 추출하여 학습한 후 악성앱을 탐지하는 기존 방법과는 다르게 정상앱 Feature를 추출하여 알고리즘을 이용하여 학습한 후 , 악성앱을 탐지한다. 사용 알고리즘은 Autoencoder, LOF(Local Outlier Factor), Isolation Forest 알고리즘 외 다른 알고리즘을 이용하여 학습해도 된다.

Ⅳ. 결 론

본 논문에서는 나날이 늘어가는 악성코드를 사람이 일일이 탐지하기 어려우므로 AI를 활용하여 탐지하는 방안을 제안했다. 기존 악성코드가 포함되어 있는 악성앱을 탐지하는 방안은 악성코드에서 Feature를 추출하여 악성코드를 탐지하는 방법이었다면, 새로 제안하는 방안은 반대로 정상앱의 Feature를 추출하고 학습한 후, 정상에서 거리가 먼 악성앱을 탐지하는 방안과, 기존 방안과 새로 제안한 방법을 결합한 앙상블 기법이다. 제안한 두 가지 기법은 Android 기반 앱에서만 해당한다.

새롭게 제안한 두 가지 기법은 기존 기법보다 악성앱을 탐지하는 과정에서 탐지율과 정확도를 높여주어 ROC Curve 가 우수하게 나올 것으로 예상된다. 또한 앙상블 기법은 기존 기법과 새로 제안한 기법 서로의 장단점을 보완해주기 때문에 더욱더 우수한 결과가 나올 것으로 기대된다.

Acknowledgement

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0사업의 연구결과로 수행되었음(IITP-2021-2020-0-01825)

References

[1] AhnLab Website [Internet]. Available : https://www.ahnlab.com/kr/site/securityinfo/statistics/security4.do

[2] S. W. Min, H. J. Cho, J. S. Shin, and J. C. Ryou, “Android Malware Detection Method Using Machine Learning*”, Journal of KIISE* 39.1C, Korea, pp. 280-282, 2012.

[3] J. H. Boo, K. H. Lee, “Advanced Feature Selection Method on Android Malware Detection by Machine Learning*”, Journal of The Korea Institute of Information Security & Cryptology, VOL.30, NO.3*, Korea, pp. 357-367, *2020 June.*

[4] J. S. Seo, *Learn Artificial Intelligence Security,* 1st ed. Korea, pp. 320-325, 2019.

[5] Github Website [Internet]. Available : https://jayhey.github.io/novelty%20detection/2017/11/10/Novelty_detection_LOF/

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” Proceedings of the 8th IEEE International Conference on Data Mining, pp. 413-422, Dec. 2008.