머신러닝을 이용한 N-gram Opcode 기반 악성파일 탐지
N-gram Opcode based Malicious File Detection Using Machine Learning

이유진, 임정수
서울여자대학교 정보보호학과

Abstract

나날이 발전하는 악성코드를 탐지하기 위해 머신러닝 기반 탐지 연구가 중요해지고 있다. 악성코드는 주로 특정 Opcode 시퀀스를 포함하므로 특정 파일에 악성 Opcode 시퀀스가 존재한다면 악성 파일이라고 분류할 수 있다.
본 논문에서는 악성코드 실행파일로부터 추출한 Opcode 시퀀스를 N-gram 기반 전처리 방법인 Opcode Count, Opcode Frequency 및 TF-IDF을 사용하여 1-gram 및 2-gram을 기준으로 분석한다. 그리고 Gaussian Naive Bayes, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbor 머신러닝 모델을 사용하여 학습하고 정확도를 비교한다. 실험 결과 2-gram TF-IDF 방법으로 전처리해 Random Forest 모델로 학습하였을 때 가장 높은 정확도를 달성하였다.

연구 배경

악성코드의 위협이 나날이 발전하고 있다. 특히 코로나19 이후 재택근무를 하는 기업을 노리는 등, 악성코드는 여러 경로를 통해 기업과 개인을 위협하고 있다. 또한, 머신러닝과 결합된 신종 악성코드가 늘어남에 따라 더 이상 과거처럼 악성코드 분석가의 수작업 분석에 의존하는 것이 어려워지고 있다.
이러한 한계를 극복하기 위하여 머신러닝을 기반으로 악성코드를 탐지하고 분석하고, 나아가 예방하는 연구의 필요성이 더욱 대두되고 있다. 악성코드 실행파일에서 추출된 Opcode 시퀀스를 기반으로 악성코드를 분석하면 로우 데이터에서 피처를 학습할 수 있다는 장점이 있다.

Opcode 시퀀스로부터 학습 입력 데이터를 생성하는 기술로 N-gram 기법이 있다. N-gram은 많은 머신러닝 기반 악성코드 연구에서 특징으로 사용되었고, 특히 정적 분석에 있어 가장 일반적으로 사용되는 특징 유형 중 하나이다. N-gram을 기준으로 악성 파일 탐지하는 방법에 대해서는 이미 많은 연구가 이루어졌지만, 그러한 방법들 중 어느 방법이 가장 효율적인지 비교 및 분석하는 연구는 부재하여 필요성을 느꼈다.

연구 목표

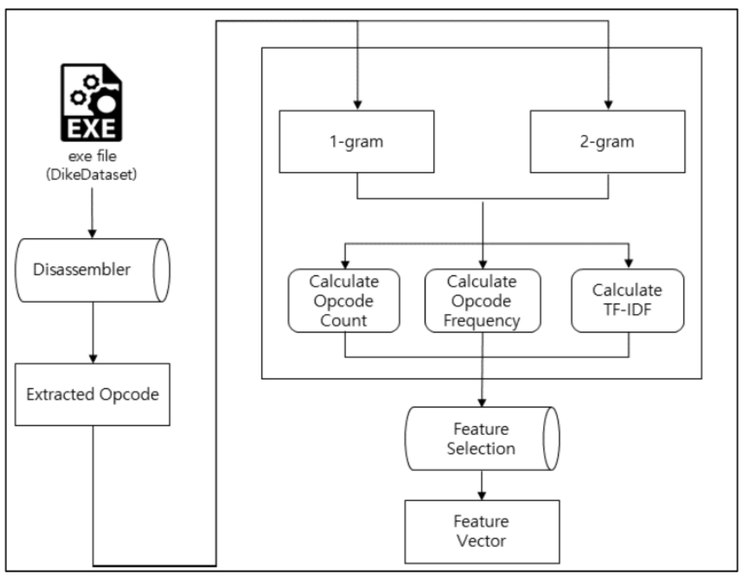
본 논문에서는 N-gram Opcode 기반 전처리 방법인 Opcode Count, Opcode Frequency, TF-IDF를 사용하여 1-gram 및 2-gram을 기준으로 분석하고, 어떤 전처리 모델이 악성코드를 분류하는 데 가장 좋은 방법인지 실험한다.

또한, Gaussian Naive Bayes, Support Vector Machine (linear kernel), Support Vector Machine (RBF kernel), Random Forest, Decision Tree, K-Nearest Neighbor의 6개 머신러닝 모델을 사용하여 Opcode의 N-gram을 적용한 전처리 방식의 악성파일 탐지에 있어 어떤 모델이 가장 좋은 성능을 보이는지 비교 및 분석한다.

N-gram Opcode 기반 악성파일 탐지 방법

데이터 전처리 과정

1) 악성 파일을 디스어셈블하여 실행 섹션에서 opcode 시퀀스를 추출한다.
2) 하나의 opcode 시퀀스는 각 코드 라인별로 Address, Hex Opcode, Opcode, Operand로 파싱하여 리스트로 구성한다.
3) 추출한 opcode 시퀀스를 연속된 1개 또는 2개 바이트(1-gram 및 2-gram)로 그룹화하여 워드를 생성한다.
4) 생성된 워드를 Opcode Count, Opcode Frequency, TF-IDF의 세 가지 방법으로 피처 벡터를 생성한다.



악성파일 탐지 방법

1. Opcode Count를 이용한 탐지 방법

- N-gram Opcode 시퀀스를 담은 각 파일에서 특정 Opcode가 등장한 횟수를 계산한다.

2. Opcode Frequency를 이용한 탐지 방법

- 파일에서 특정 Opcode가 발생한 횟수를 파일에서 Opcode가 발생한 총 횟수로 나눠서 계산한다.

$$Opcode\ \ Frequency = \frac{No.\ of\ Instances\ of\ the\ Opcode}{Total\ No.\ of\ Opcodes\ in\ the\ files}$$

3. TF-IDF를 이용한 탐지 방법

- 다양한 종류와 많은 개수로 이루어진 N-gram Opcode 중에서 의미 있는 N-gram Opcode를 추출하는 방법을 시도할 수 있다. 본 논문에서는 TF-IDF를 사용하여 파일 집합 전체에 대한 특정 N-gram Opcode 의 중요성을 분석하였다.
- TF(Term Frequency) 는 다음과 같이 계산하며, 특정 파일 d에서 주어진 N-gram Opcode t의 빈도를 나타낸다.

$$tf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

- 여러 파일에서 동일하게 자주 나타나는 N-gram Opcode는 특정 악성파일에서만 나타나는 특징으로 볼 수 없으므로, 유의미하지 않은 N-gram Opcode라고 할 수 있다. 따라서, TF를 역으로 계산한 IDF(Inverse Document Frequency)를 계산하여, 보편적으로 등장하는 N-gram Opcode를 제외하였다. $|D|$는 전체 파일의 수를 의미한다. 파일이 늘어날수록 idf의 값이 무한히 증가 하는 것을 방지하기 위하여 log를 취한다.

$$idf_{t,D} = log\ \frac{|D|}{|d \in D : t \in d|}$$

- tf에 idf를 곱한 값을 3.1절에서 구한 파일마다 계산한다.

$$tfidf_{t,d,D} = tf(t,d) \cdot idf(t,D)$$

악성파일 탐지 모델 구현 및 실험

1. 실험 대상 – DikeDataset

- 악성 파일 : 8,938 개
- 정상 파일 : 938 개
- Training 데이터 : (악성) 6,704 개, (정상) 703 개
- Test 데이터 : (악성) 2,234 개, (정상) 234 개

2. 실험 결과 및 분석

- 6개의 머신러닝 모델을 1-gram 및 2-gram을 적용한 Opcode Count(OC), Opcode Frequency(OF), TF-IDF 각각으로 전처리된 데이터셋으로 학습한 후 평가하여, 가장 좋은 성능의 악성파일 분류 방식을 비교하였다.

Model	1-gram			2-gram		
	OC	OF	TF-IDF	OC	OF	TF-IDF
NB	92.48	93.10	74.28	66.07	83.73	84.50
SVM (linear)	97.75	98.99	96.04	99.22	99.61	99.30
SVM (RBF)	90.31	98.14	98.14	93.88	99.38	99.30
RF	99.38	99.38	99.30	99.45	99.45	99.53
DT	97.52	98.21	98.14	98.91	99.14	98.76
KNN	97.67	98.21	98.37	97.98	98.68	98.60

- 전반적으로 Random Forest 모델이 높은 정확도를 보였다. 특히, 2-gram에 TF-IDF를 적용한 전처리 방식이 99.53%의 높은 수치를 기록하였다. 반면 Gaussian Naive Bayes 모델이 가장 낮은 정확도를 기록한 것을 확인할 수 있다.
- N-gram Opcode에 1-gram을 적용한 경우, Opcode Frequency 전처리 방식이 K-Nearest Neighbor 모델의 경우를 제외하고는 전체적으로 높은 정확도를 기록한 것을 확인할 수 있다. N-gram Opcode에 2-gram을 적용한 경우에도 대부분의 모델에서 Opcode Frequency를 사용한 경우에 높은 정확도를 보였다.
- 반면, Opcode Count는 대부분의 1-gram과 2-gram 적용 모두 모델에서 가장 낮은 정확도를 보여주였다. 특히, 1-gram을 적용한 RBF를 사용하는 Support Vector Machine 모델과 2-gram을 적용한 Gaussian Naive Bayes 모델에서는 다른 전처리 방식과 각각 약 8%, 약 18%의 큰 정확도 차이를 보였다.

결과

본 논문에서는 실제 악성 및 정상파일에서 추출한 N-gram Opcode를 기반으로 여러 전처리 방법을 통해 악성파일을 탐지하고 결과를 비교하였다. 이때, Opcode Count, Opcode Frequency, TF-IDF을 1-gram 및 2-gram을 기준으로 각각 적용하여 전처리를 하였으며, 이를 6가지의 머신러닝 모델에 학습시키고 정확도를 비교하여 가장 정확하게 악성파일을 탐지하는 모델을 비교 및 분석하였다.
그 결과 Opcode Frequency로 전처리하였을 때 1-gram 및 2-gram 전체에서 높은 정확도를 기록하였고, 머신러닝 모델의 경우 Random Forest 모델이 다른 모델에 비해 높은 정확도를 보였다. 특히, 2-gram에 TF-IDF를 적용한 전 처리 방식이 99.53%의 높은 수치를 기록하였다. 2-gram 이상의 경우 및 다른 머신러닝을 사용한 경우에도 동일한 결과를 도출하는지에 대하여 연구하는 것이 향후 과제이다.