한국정보보호학회
Korea Institute of Information Security & Cryptology

2019 한국정보보호학회 하계학술대회

CISC-S'19

Conference on Information Security and Cryptography-Summer 2019

2019. 6. 20(목)~22(토)

부산 동명대학교 (제1정보통신관, 중앙도서관)

Proceedings

주최 한국정보보호학회 Korea Institute of Information Security & Cryptology

주관 동명대학교 TONGMYONG UNIVERSITY

후원 국가정보원 National Intelligence Service Korea, 과학기술정보통신부, 행정안전부, KISA 한국인터넷진흥원, ETRI 한국전자통신연구원 Electronics and Telecommunications Research Institute, NSR 국가보안기술연구소 National Security Research Institute, IITP 정보통신기획평가원, AhnLab

소프트웨어 취약점 분석을 위한 다수의 머신러닝 알고리즘과 데이터셋 조합 플랫폼 설계에 관한 연구

임민훈*, 송준환*, 유지현*, 장은태**, 윤주범**, 박기웅†

*세종대학교 정보보호학과, **세종대학교 시스템보안 연구실, † 세종대학교 정보보호학과

mhiunn09@naver.com, wnsghks30@naver.com, yjhy8783@naver.com, euntaejang@gmail.com, jbyun@sejong.ac.kr, woongbak@sejong.ac.kr

A Study on the Design of Multiple Machine Learning Algorithms and Datasets Combination Platform for Software Vulnerability Analysis

Minhoon Lim*, Joonhwan Song* ,Jihyeon Yu*, Euntae Jang**, Joobeom Yun**, Kiwoong Park†

*Department of computer and information security, Sejong University.

요 약

최근 4차 산업혁명 시대를 맞이함에 따라 다양한 소프트웨어가 발전하고 있지만, 그에 따라 소프트웨어의 보안적인 취약점도 기하급수적으로 증가하고 있다. 이를 해결하기 위해 인공지능 기술을 활용한 취약점 분석에 대한 연구가 활발히 진행 중이나, 기존의 단일 머신러닝 알고리즘과 데이터셋을 통한 분석은 과적합(Overfitting), 분석할 소프트웨어 특성에 따른 정확도 하락 문제를 가지고 있다. 본 논문에서는 이러한 문제점들에 대한 해결책으로 다수의 머신러닝 알고리즘과 데이터셋들을 사용자의 기호에 따라 하나의 세트로 조합하여 학습 및 예측 결과를 산출할 수 있는 조합 플랫폼 설계에 관한 연구 내용을 기술하였다.

I. 서론

최근 4차 산업혁명 시대를 맞이함에 따라, 사회 전반적인 분야에서 기존에 인간이 수동으로 하는 번거로운 일들을 소프트웨어가 대체할 수 있도록 발전하고 있다. 다양한 분야에 접목된 형태로 연구수준에 머물던 기술들도 인공지능 기술 등의 발전에 따라 실현가능한 형태로 구현되고 있다.

보안적인 관점에서는 다양한 소프트웨어의 발전의 긍정적인 영향만을 고려할 수 없다. 각 기업들은 우선 출시를 통한 시장 점거에 초점을 두고 있기 때문에 소프트웨어에 보안적인 취약점이 존재하는 것이 비일비재하다. 보안업체 Risk Based Security가 2018년에 발표한 보고서에 의하면 해당 년도에 22000개가 넘는 취약점 데이터를 수집하였고 이는 2016년 대비 35%이상의 증가 추세를 보이고 있다고 한다.[1] 이를 해결하기 위한 소프트웨어 출시 전 보안성 검증을 위해 최근에는 머신 러닝, 딥 러닝 등의 인공지능 기술을 사용한 취약점 분석에 대한 연구가 활발히 진행 중이다. 본 논문에서는 단일 머신러닝 알고리즘과 데이터셋 사용으로 인한 과적합, 소프트웨어 특성에 따른 정확도 하락 문제를 해소하기 위해 다수의 알고리즘과 데이터셋들을 사용자의 기호에 따라 하나의 세트로 조합하여 결과를 산출할 수 있는 플랫폼 설계에 관한 연구 내용을 기술하고자한다.

본 논문의 전체 구성은 다음과 같다. 2장은 제안하는 플랫폼의 사전 구성요소인 머신러닝 알고리즘의 특징 및 이용 동향을 관련 연구로서 설명하였고, 3장은 제안하는 플랫폼의 전체 구조를 소개하고 추가 알고리즘과 데이터셋의 규격, 플랫폼의 기능 및 지향하는 구현방법을 설명하였다. 4장 결론은 제안된 모델이 가지는 이점 및 향후 연구 발전 방향으로 익스플로잇 기능과 패치 기능 추가를 두고 있음을 보였다.

II. 관련연구

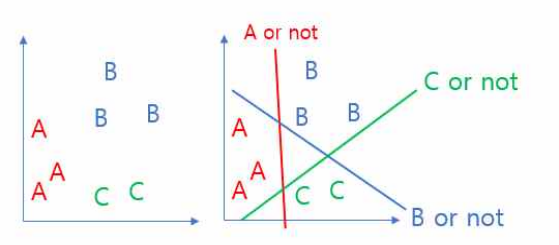
머신러닝 지도학습은 데이터에 대한 레이블(Label : 명시적인 정답)이 주어진 상태에서 컴퓨터를 학습시키는 방법이다. 훈련 데이터셋으로 알고리즘이 데이터들을 해석할 수 있는 임의의 모델을 만들고, 그 것을 바탕으로 새로운 데이터를 예측하는 것을 말한다. 이러한 특징 때문에 머신러닝 지도학습은 데이터 분류 및 회귀 작업에 특화되어있고 본 논문에서의 소프트웨어 취약점 분석을 위한 분류 작업에 적합하다. 머신러닝 알고리즘은 사용 프레임워크와 환경에 따라 연산속도, 성능 면에서 차이를 보이기 때문에 컨테이너 기반 가상화 기술을 사용하여 자원 소모를 최소화하는 작업이 추세로 떠오르고 있다.

2.1 Random Forest 알고리즘

Random Forest 알고리즘은 무작위(Random)로 추출한 사례의 집합들을 이용하여 많은 수의 의사결정트리를 생성하고 생성된 여러 의사결정트리의 판별 클래스들을 가중 투표하여 최종 클래스를 결정하는 머신러닝 분류 지도학습 기법이다. Random Forest 기법은 부트스트랩(Bootstrap)기법을 이용하여 학습에 필요한 훈련 데이터를 생성하므로 적은 수의 데이터셋만으로도 일정 수준이상의 정확성을 가지는 분류를 실행할 수 있다. 또한 학습 과정에서 무작위로 추출된 학습데이터로 많은 수의 의사 결정 트리를 생성하여 다양한 패턴을 포괄하기 때문에 훈련 데이터가 아닌 예측할 새로운 데이터가 판별을 위하여 분류기에 입력되었을 경우에도 높은 수준의 정확성을 가질 수 있다. 이외에도 과적합 문제가 발생하지 않고, 결측값(missing value)을 제공한다.

2.2 Softmax Regression 알고리즘

Softmax Regression 알고리즘은 데이터를 2개로 나누는 이진 분류(Binary classification)알고리즘인 Logistic Regression 알고리즘을 확장시킨 분류 알고리즘이다.[5] Logistic 알고리즘은 시그모이드 함수(Sigmoid)를 사용하여 데이터를 0과 1로 구분하는 하나의 직선(decision boundary)을 만들어 낸다. 우리가 원하는 시스템에서는 이진 분류가 아닌 다항 분류(Multinomical Classification)가 필요하기 때문에 직선을 여러 개 그어서 각 항목 별 예상 확률이 도출 가능한 Softmax Regression 알고리즘을 사용한다. <그림 1>은 Softmax Regression을 설명하는 예시이다. 좌측 그래프에서는 좌표 상 A,B,C 세 그룹이 존재하고 Softmax Regression을 알고리즘을 적용할 경우 우측 그림처럼 각 그룹을 구분하는 Decision boundary가 생성된다. 이 직선이 각 그룹을 구분하는 기준이 된다. 이 알고리즘은 구조가 간단하여 쉽게 구현할 수 있다는 장점과 행렬의 곱셈으로 연산하여 그룹끼리 서로 영향을 주기 때문에 각 그룹의 예상 확률이 결과값으로 나온다는 장점이 있다.

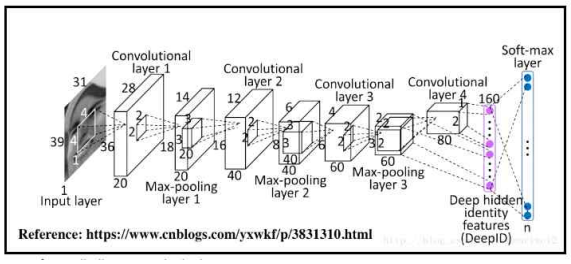


<그림 1> Softmax Regression 구조

2.3 Convolutional Neural Network 알고리즘

Convolutional Neural Network(CNN) 알고리즘은 인간의 뇌가 패턴을 인식하는 방식을 모사한 알고리즘으로 모델이 직접 데이터를 분류하는 딥러닝에 가장 많이 사용되는 알고리즘 중 하나이다[2]. CNN은 데이터에서 특정 부분 찾아내기 위한 패턴을 찾는 것에 특화되어있다. CNN의 구성은 <그림 2>와 같다. 일반적인 인공신경망 구조 이전에 합성곱 계층(Convolutional layer)과 풀링 계층(pooling layer)이라는 계층을 추가함으로써 원본 데이터에 필터링 기법을 적용한 뒤에 필터링된 데이터에 대해 분류 연산이 수행되도록 구성된다. 이러한 필터링 기법은 공간 정보를 유지하면서 인접 데이터와의 특징을 효과적으로 인식하고, 필터를 공유 파라미터로 사용하기 때문에, 일반

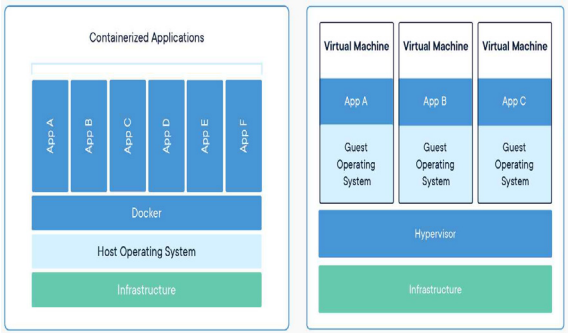
인공신경망과 비교하여 학습 파라미터가 매우 적다.



<그림 2> CNN 계층구조

2.4 컨테이너 기반 가상화 기술(Docker)

<그림 3>는 기존 하이퍼바이저 기반 가상머신과 도커 가상화 컨테이너의 구조이다.

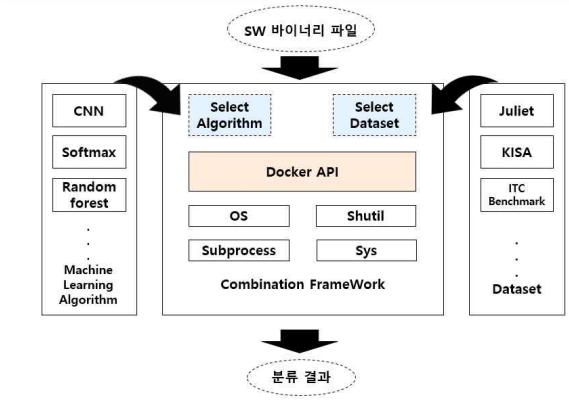


<그림 3> 도커 컨테이너와 기존 하이퍼바이저 기반 가상머신의 구조

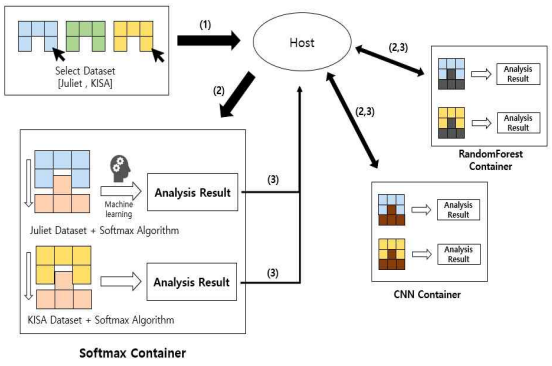
기존 운영체제 수준의 가상화 기술은 자원의 소모가 심하고, 하이퍼바이저 기반 체계로 성능적인 면에서 오버헤드가 생기는 문제점이 있었다. 이를 해결하기 위해, 도커는 리눅스 실행 환경을 컨테이너로 가상화하고 서로 격리된 상태를 실행하는 기능을 제공한다[3][4]. 컨테이너 개념은 사용자 및 필요한 자원을 최소한의 프로세스만으로 독립시켜 더 가볍고 원시적인 사용 환경을 제공한다는 장점을 가진다.

III. 머신러닝 알고리즘과 데이터셋 조합 플랫폼 설계

본 논문에서는 다양한 머신러닝 알고리즘과 여러 데이터셋의 교차검증을 통해 소프트웨어의 보안 취약점의 종류를 분류 하는 플랫폼 설계를 <그림 4>과 같이 제시한다. 해당 플랫폼의 근본적인 목적은 바이너리 파일이 가지고 있는 취약점 종류 분류이므로 입력 데이터로 리눅스 바이너리 파일을 받도록 설계한다. 이 입력 파일을 다수의 머신러닝 알고리즘에 동일하게 적용하기 위해 일관성 있는 데이터 전처리 작업을 거친다. 본 플랫폼에서의 입력데이터는 다중 분류, CNN 알고리즘의 이미지에 특화된 특징 등을 고려하여 256 x 256 크기의 이미지 데이터로 바이너리 파일을 변환시키도록 설계한다. 조합 시스템은 도커(Docker)로 시스템을 구성하여 각각의 알고리즘을 컨테이너(Container)로 구현한다.



<그림 4> 제안 플랫폼 구성도



<그림 5> Combination Framework 구동 방식

Combination Framework의 세부 구동 방식은 <그림 5>와 같다. 각 컨테이너마다 위에서 제시하였던 다중 분류에 적절한 Random Forest 알고리즘, Softmax Regreesion 알고리즘, CNN 알고리즘들을 컨테이너 프로세스로서 활용하고 사용자가 원할 경우 새로운 알고리즘을 컨테이너로서 추가하거나 삭제할 수 있다.

새로운 알고리즘을 추가할 경우 처리될 데이터의 일관성을 위해 256 x 256 크기의 이미지 데이터로서 입력을 받고 학습 및 예측 할 수 있도록 알고리즘을 모델링한다. 데이터셋들에 관해 머신러닝 학습을 진행하고 해당 학습 내용을 저장한다. 사용자가 새로운 바이너리 파일을 분석하고자 할 경우 어떤 알고리즘과 어떤 데이터셋을 사용하여 조사할지 선택하고 기존에 학습하였던 데이터를 바탕으로 해당 바이너리 파일이 어떠한 종류의 취약점을 가지고 있는지 판단한다.

데이터셋도 알고리즘과 마찬가지로 원하는 데이터셋을 추가하거나 삭제할 수 있다. 각 데이터셋들은 하위 디렉토리에 클래스별로 분류되어 클래스에 대한 정보를 담는 라벨데이터로 활용할 수 있게 규격화하여 추가한다. 데이터셋을 사전에 규격화 하여 새로운 알고리즘이 추가된다 하더라도 이 규격에 맞춰 학습을 진행할 수 있도록 하기 위함이다.

해당 프로그램은 GUI(Graphic User Interface)로 구현하여 머신러닝 기술이나 취약점 분석 도구 사용에 익숙하지 않은 사용자도 쉽게 사용할 수 있도록 설계한다. 사용자가 선택한 알고리즘별 학습 결과를 그래프 이미지를 통해 보여줌으로써 선택한 알고리즘과 데이터셋이 어떻게 학습되었는지 알 수 있게 설계한다. 만약 학습 결과가 적절하지 않을 경우, 각 알고리즘과 데이터셋의 학습결과 이미지를 보고 알고리즘의 learning rate를 다르게 하거나 데이터셋이 과적합(Overfitting)될 경우 일반화(Regulation)을 통해 데이터셋을 수정하도록 한다.[5] 위와 같은 플랫폼을 통해 사용자에게 직관적이고 수정 용이한 환경을 제공한다.

IV. 결론

본 논문에서는 기존의 인공지능 기술을 도입한 보안 테스팅 도구들은 단일 머신러닝 알고리즘과 데이터셋의 사용으로 인한 과적합, 점검할 소프트웨어 특성에 따른 정확도 하락 등의 한계를 보일 수 있다는 것을 인지하였고[6], 이를 해결하기 위해 다수의 머신러닝 알고리즘과 데이터셋들을 사용자의 입력에 따라 조합하여 테스팅을 할 수 있는 플랫폼 설계에 대해 기술하였다. 본 플랫폼은 사용적인 면에서 pluggability, user-friendly, 가독성 등의 장점을 보일 것으로 기대한다. 향후 플랫폼의 예측 정확도와 실용성을 높이기 위해 예측한 취약점을 검증하는 익스플로잇 기능과 취약점을 보완하는 패치 기능 등을 추가한 플랫폼 설계 및 구현 작업을 발전 방향으로 두고 있다.

[ACKNOWLEDGEMENT]

본 연구는 2019년도 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원 (No.2019-0-00426) 및 한국연구재단 연구과제 (NRF-2017R1C1B2003957)의 지원을 받아 수행된 연구임.

[참고문헌]

[1] Risk Based Security, “More Than 22,000 Vulnerabilities Disclosed In 2018”,

https://www.riskbasedsecurity.com/

[2] 김상우, 정진곤 (2018). “딥 러닝 기반 MNIDT Classification에 대한 softmax Regression과 CNN 성능 비교”. 한국통신학회 학술 대회논문집, 871-872

[2] Docker io

https://www.docker.com/

[4] 윤준원, 송의성 (2018). “컨테이너 기반 가상화를 통한 교육 실습환경 구축”. 한국디지털콘텐츠학회 논문지, 19(3), 453-460

[5] Github

http://hunkim.github.io/ml/

[6]Toward Large-Scale Vulnerability Discovery using Machine Learning

https://dl.acm.org/citation.cfm?id=2857705.2857720