논문 18-43-01-12 The Journal of Korean Institute of Communications and Information Sciences '18-01 Vol.43 No.01
https://doi.org/10.7840/kics.2018.43.1.77

적대적 데이터를 이용한 기계학습 기반의 악성코드 분류기 공격

윤 성 국◆, 김 창 훈°

Machine Learning Based Malwae Code Classifier Attack Using Hostile Data

Sung-koock Yoon◆, Chang hoon Kim°

요 약

인공신경망을 이용한 악성코드 분류기법의 취약점을 제시하기 위해, 본 논문에서는 악성코드 분류모델의 정확도를 낮출 수 있는 공격기법을 제안한다. 공격의 절차는 1) 블랙박스 공격기법을 이용하여 모방모델을 학습시킨 후, 2) 적대적 데이터를 선택적으로 수집하고, 3) 수집된 데이터를 기반으로 공격을 실시한다. Microsoft사에서 제공하는 데이터를 이용하여 77.82%의 악성코드 분류 정확도를 보이는 기존 연구결과에 적용한 결과 그 정확도가 53.21%까지 감소함을 보인다.

Key Words : Malware Classifier, Machine Learning, Black-box Attack, Convolutional Neural Network, Vulnerability Analysis

ABSTRACT

In order to find vulnerabilities of malware classification method, this paper proposes a new attack scheme that can reduce the accuracy of malware classification. To realize the proposed concept, we use the following procedures; 1) After learning an imitation model by using black-box attack strategy, 2) Crafting adversarial samples from the imitation model, and 3) attacking target model. An implementation by using the Microsoft's data and experimental analysis show that the previously proposed malware classification accuracy can be forced to decrease from 77.82% to 53.21%.

Ⅰ. 서 론

최근 인공신경망 기반의 악성코드 분류기법에 관한 연구가 활발히 진행되고 있다[1,2]. 특히, 2015년 Microsoft에서 개최한 "Malware Classification Challenge"에서 악성코드를 이미지로 인식하여 분류하는 컨볼루션 신경망(CNN : Convolution Neural Network) 기반의 악성코드 분류 기법이 제안되었으며, 99.72%의 정확도를 보인다[1]. 그러나 CNN과 같은 인공신경망 모델의 경우 학습용 데이터의 분류 정확도는 상당히 높지만, 현실세계의 데이터에 적용 시 정확도가 낮아지는 과적합현상(Overfitting)이 발생하며, 이는 아주 중요한 문제로 여겨지고 있다. 특히, 통계학적으로 표본 집단은 모집단의 모든 특징을 나타낼 수 없기 때문에, 회귀분석 시, 과적합현상이 발생할 가능성이 있다. 인공신경망은 학습용 데이터를 표본 집단, 현실세계 데이터를 모집단으로 하는 통계학적 회귀분석을 기반으로 하고 있으며, 보다 더 유연한 형태가 가능하므로 과적합현상이 일어날 가능성이 더욱더 높다 할 수 있다. 따라서 모든 인공신경망의 경우 과적합현상을 완전히 해결하기는 매우 어려운 실정이며[3], 이러한 과적합현상을 모든 인공신경망 기반의 기계학습 모델에 적용할 경우 적대적 데이터 생성의 도구로 악용 될 가능성이 있다[3].

본 논문에서는 과적합현상을 이용하여 적대적 데이터를 수집한다. 여기서 적대적 데이터는 블랙박스 공격기법[4]을 이용하여 수집하며, 공격 대상이 되는 CNN 기반 악성코드 분류모델의 내부적인 구조에 대한 정보 없이 이루어진다. 수집된 적대적 데이터를 이용하여, 기존에 제안된 인공신경망 기반의 악성코드 분류모델 중 하나인 [2]에 적용함으로써, 취약점이 존재하고 공격이 가능함을 보인다. 검증을 위해 Microsoft에서 제공하는 데이터를 이용하여 제안된 기법을 [2]에서 제안한 규격과 동일한 모델에 공격하는 실험을 수행한 결과 모방모델을 통해 추출한 적대적 데이터 1,630개 중 753개의 데이터가 우회 가능하고, 공격받은 모델의 분류정확도가 77.82%에서 53.21%

※ 이 논문은 대구대학교 DU-리더스 학부생 연구지원 사업에 의해 진행된 연구임.
◆ First Author : School of Computer and Information Engineering, Daegu University, koock1994@gmail.com, 학생회원
° Corresponding Author : School of Computer and Information Engineering, Daegu University, kimch@daegu.ac.kr, 종신회원
논문번호 : KICS201711-355, Received November 21, 2017; Revised December 11, 2017; Accepted December 22, 2017

로 낮아짐을 보인다.

Ⅱ. 본 론

2.1 공격을 위한 전체적인 절차

본 연구에서 진행한 공격의 전체적인 과정은 그림 1과 같다. 그림 1에 기술된 바와 같이 1) 블랙박스 공격기법을 이용하여 모방모델의 학습을 통해 공격 대상과 최대한 동일한 환경을 구성하고, 2) 모방 모델로부터 우회 가능한 적대적 데이터를 추출하여, 3) 공격 대상 모델을 공격한다.

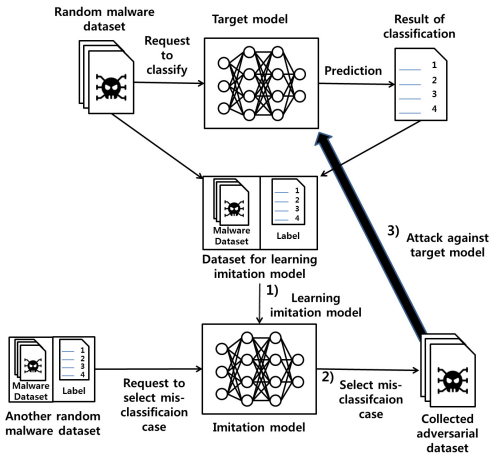

그림 1. 분류모델의 공격과정
Fig. 1. Attacking Procedure against Classification Model

2.2 공격 대상 모델

본 연구의 공격 대상 모델은 어떤 신경망구조를 이용한 분류모델인지, 어떤 학습 알고리즘으로 평가 되었는지에 대한 사전 지식이 없다는 가정 하에 실험을 진행한다. 실험을 위해 최근에 제안된 기존 연구[2]에서 제안한 규격과 동일하게 신경망을 구성하여 학습시킨 악성코드 분류모델을 공격대상으로 삼는다. 본 논문에서는 이 모델을 모델 A라고 명명한다.

2.3 블랙박스 공격기법 통한 모방모델 구현

모델 A의 공격을 위한 데이터를 생성하거나 수집하기 위해서는 모델 A의 취약점을 발견해야 한다. 모델 A의 취약점을 발견하기 위해 신경망 구조를 알고 역공학을 통해 적대적 데이터를 생성하거나 수집해야 한다. 실제 서비스 되는 분류모델의 경우 입력에 따른 출력만 알 수 있는 블랙박스 형태이기 때문에 내부 구조 파악은 매우 어렵지만 모델 A와 유사한 모방모델(모델 B)을 만드는 것은 가능하다[4]. 블랙박스 형태의 인공신경망 모델은 입력과 출력을 근사시킬 경우 그 성능 또한 유사한 것으로 알려져 있다. 이러한 점을 이용하여 악성코드 데이터와 그에 따른 모델 A의 분류결과를 학습용 데이터로 삼아 모델 B를 학습시킨다. 학습이 진행되면서 동일 악성코드 데이터에 대해 모델 B의 분류결과는 모델 A의 분류결과에 점차 근사하게 된다. 이렇게 근사시킨 모델 B는 모델 A의 취약점과 유사한 취약점을 가질 가능성이 크다. 본 연구에서는 [1]에서 실험한 세 가지 모델 중 정확도가 가장 높은 "CNN B:2C 1D"와 동일하게 모델 B를 구성한다.

2.4 악성코드 분류모델 공격

2.4.1 적대적 데이터 수집

모델 A를 공격하는 가장 간단한 방법은 바로 많은 양의 악성코드를 주입하는 무작위 공격이다. 그러나 만 건 이상의 많은 데이터를 한 번에 주입하게 되면 서비스 제공자가 무작위 공격을 감지할 수 있다. 이에 따라 모델 A대신 2.2에서 학습시킨 모델 B에 다량의 악성코드 데이터를 무작위로 주입하여 모델 B가 오 분류한 데이터만을 별도로 수집한다. 이렇게 수집된 데이터집합은 모델 B의 취약점에 따라 도출된 결과이다. 이는 피 모방 모델인 모델 A에서도 오 분류될 가능성이 높음을 의미한다. 뿐만 아니라, 소량의 데이터가 수집되기 때문에 해당 데이터로 공격을 시도하더라도 서비스 제공자는 공격을 감지하기 매우 어렵다.

2.4.2 모델 A 공격

2.2에 기술한 모방모델의 구현과정과 2.3.1에서 수집된 적대적 데이터를 이용하여 공격을 진행한다. 이 과정은 전체적인 공격과정이 기술되어 있는 그림 1에서의 3)에 해당하는 부분이다. 여기서 평가방식은 데이터를 모델 A에 주입하여 모델 A가 오 분류를 일으키게 되면 해당 데이터는 공격에 성공한 것으로 간주한다.

2.5 실험 및 성능평가

본 연구의 공격 가설에 대한 입증과 성능을 평가를 위한 실험 환경은 표 1과 같다.

2.5.1 실험

본 논문에서 진행하는 실험의 전체적인 과정은 그림 2에 기술된 바와 같이 1) 공격 대상인 모델 A를 준비하고 2) 해당 모델을 모방하여 모델 B를 구성한 후,

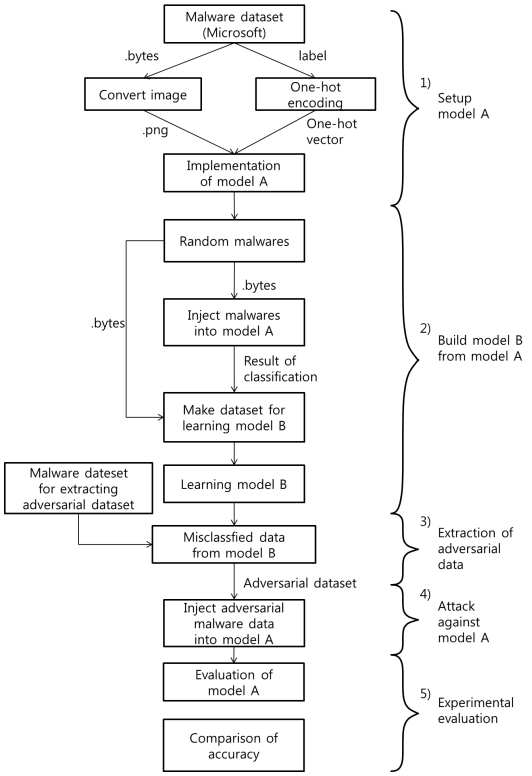

그림 2. 실험의 세부적 절차
Fig. 2. Detailed Experiment Procedures

표 1. 실험 및 성능 평가 환경
Table 1. Environment of Implementation and Performance Evaluation

Name	Spec
CPU	Intel Core i5-2300 2.80GHz
GPU	GTX Titan X
RAM	8G
OS	Ubuntu 16.04 LTS
GPGPU Platform	Cuda, 8.0.61
Machine Learning Platform	CuDNN 5.1 Tensorflow 1.0
Language	Python3.5

3) 이를 통해 적대적 데이터를 추출하고 4) 모델 A를 공격하여 5) 실험의 결과를 평가한다.

2.5.2 성능평가

Microsoft에서 제공하는 10,869개의 학습용 데이터 집합[5]으로 모델 A를 학습시킨 결과, 77.82%의 정확도를 보인다. 학습된 모델 A의 출력 값을 Label로 하는 학습용 데이터집합 10,000개로 모델 B를 학습시킨 결과, 87.14%의 정확도가 나왔으며, 이는 모델 A와 모델 B의 유사도가 87.14%임을 의미한다. 또한, Microsoft에서 제공하는 테스트용 데이터집합[5] 10,000개를 모델 B에 입력한 결과 오 분류되는 데이터는 1,630개로 16.3%이다. 이 데이터를 모델 A 공격을 위한 적대적 데이터로 사용하였으며, 이 중 753개의 데이터가 우회되었다. 정확도 측정결과, 델 A의 정확도는 기존의 77.82%보다 24.61%만큼 낮게 측정되었다. 여기서 53.21%는 아래 그림 3에 기술된 모든 악성코드 분류군의 정확도에 대한 평균값이다.

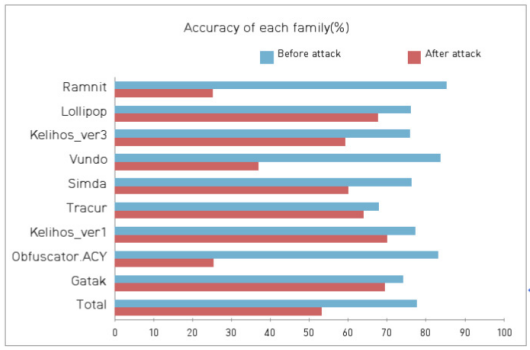

그림 3. Result of Attack Experiment
Fig. 3. 공격에 대한 실험 결과

Ⅲ. 결 론

본 논문에서는 인공신경망 기반 학습 모델의 단점인 과적합현상을 이용한 우회 공격이 가능함을 보였다. 이를 위해 블랙박스 공격기법을 이용, 모방모델을 학습시켰으며, 이로부터 적대적 데이터를 추출하여 공격모델에 대한 공격을 실시하였다. Microsoft에서 제공하는 데이터를 이용하여 실험한 결과 기존의 77.82%의 정확도를 보이는 모델 A를 공격하였으며, 정확도를 53.21%로 24.61%만큼 낮출 수 있었다. 이는 인공신경망 기반 악성코드 분류 시스템의 새로운 취약점으로 인식 되어야 함을 의미한다.

References

[1] D. Gibert “Convolutional neural networks for malware classification,” M.S. Thesis, Dept. of Computer Science, UPC, 2016.

[2] S. Seok and H. Kim, “Visualized malware classification based on convolutional neural network,” *J. KIISC*, vol. 26, no. 1, pp. 197-

208, Feb. 2016.

[3] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Comput. Sci.,* vol. 44, no. 1, pp. 1-12, 2004.

[4] N. Papernot, P. McDaniel, l. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *ASIA CCS '17*, pp. 506-519, Abu Dhabi, UAE, Apr. 2017.

[5] Microsoft Malware Classification Challenge (BIG 2015), from https://www.kaggle.com/c/malware-classification