정 보 보 호 학 회 지
제 29 권 제3호, 2019. 6

딥러닝 기반의 IDPS 탐지 데이터의 정/오탐 분류

임 종 혁*, 김 진*, 김 건 우*, 유 진 상*

요 약

딥러닝 기법이 영상 분야를 시작으로 여러 분야에서 빠르게 적용되고 있고, 관련된 다양한 연구도 함께 같이 발전하고 있다. 정보보안 분야 역시 악성코드를 위주로 다양한 데이터에 대해서 딥러닝 기법을 적용하기 위한 많은 연구들이 진행되고 있지만, 본 논문에서는 IDPS에서 탐지된 이벤트들에 대해서 정/오탐을 자동으로 식별할 수 있는 딥러닝 기반의 분류 방법을 소개 하고자 한다.

Ⅰ. 서 론

IBM은 체스 에서 나올 수 있는 모든 경우의 수를 시간 내에 계산할 수 있는 슈퍼컴퓨터인 '딥 블루'를 만들었고, 1997년 정식 체스 게임에서 인간 최강자를 꺾으면서 그 능력을 공식적으로 증명하게 된다. 하지만 이러한 슈퍼컴퓨터조차도 체스보다 더 많은 경우의 수를 계산해야 하는 바둑만큼은 정복하지 못했고, 앞으로도 정복 하기는 쉽지 않을것이라고 모두 예측했었다. 하지만, 2016년 이세돌 9단과의 공개대국을 통해 등장한 '알파고'는 모두가 불가능하다고 생각했던 바둑조차도 전 세계인이 지켜보는 앞에서 인간 최강자를 꺾으면서 정복되었음을 증명 하였다. 이러한 알파고의 뒤에는 '딥러닝' 이라는 기술이 존재하였고, 알파고를 통해 계산 능력을 인정받은 딥러닝 기술이 각광을 받는 건 당연한 일이었을 것이다. 이를 증명이라도 하듯이 산업 전반에 걸쳐서 딥러닝 기술이 적용되기 시작 하고, 많은 연구들이 더해져 영상, 의학, 게임, 번역 등 다양한 분야에서 성과들이 나오기 시작했다. 보안 분야에서는 대표적으로 악성코드의 복잡한 패턴을 탐지하기 위해 딥러닝 기술을 적용한 사례들이 있다. 본 논문에서는 IDPS(IDS/IPS)와 같은 보안 장비에서 발생한 탐지 이벤트를 이용해 이벤트의 정탐과 오탐을 판별할 수 있는 방법을 제시하고자 한다. 정탐과 오탐을 정확하게 구별해 낼 수 있다면 탐지 이벤트가 발생했을 경우 탐지 이벤트에 대한 검증을 하기 위해 과도하게 인력이 소모되는 부분을 줄일 수 있으며, 더 나아가서는 IDPS에서 사용되는 시그니처에 딥러닝 모델을 적용하여 공격에 대한 탐지 율을 높일 수도 있을 것이다.

Ⅱ. 관련 연구

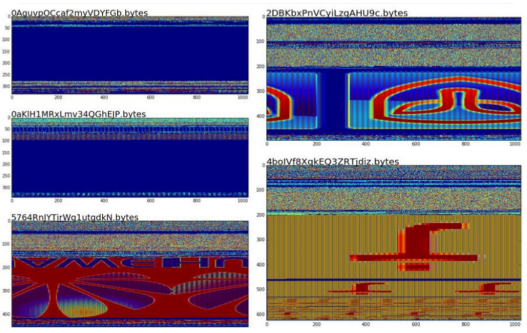
본 논문에서 제안하는 딥러닝 모델은 IDPS의 탐지 이벤트에 대한 정/오탐 분류 모델이다. 정/오탐을 분류하기 위한 데이터 셋은 CSIC 2010 Dataset을 활용하였다. 데이터 셋에 대한 자세한 분석을 하기 전에 기존의 악성코드를 분류하기 위해 제안되었던 딥러닝 기법을 살펴보고 IDPS의 탐지된 이벤트에 대해서도 적용가능한지 살펴본다. 또한, 본 논문 이전에 제안되었던 웹 공격을 탐지하기 위한 방법론을 통해 데이터 전 처리 방법에 대한 새로운 아이디어를 얻어 올 것이다.

2.1. 악성코드 분류

악성코드를 분류하기 위해 기존에 제안된 방법들로는 악성코드가 사용하는 API나 어셈블리 명령어를 이용하는 등 기본적으로 탐지하기 전에 바이너리 분석이나 샌드박스 등을 통한 동적 분석이 선행되지 않으면 사용할 수 없기 때문에 실시간 탐지에 바로 활용될 수 없었다는 단점이 있었다. 본 논문에서는 가장 최근에 연구된 방법으로 기존의 단점을 보완할 수 있는 의미 있는 논문을 살펴보고자 한다.

해당 논문에서 제안하고 있는 방법은 악성코드를 이

* 시큐레이어 기술연구본부 (jonghyuk.im@seculayer.com, jinkim@seculayer.com, kwkim@seculayer.com, jsyu@seculayer.com)



〔그림 1〕 악성코드 시각화

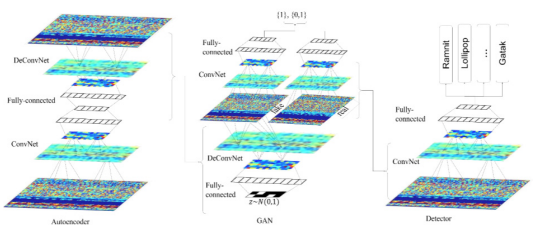
미지로 바꿔서 시각화 하였을 경우 다수의 악성코드에서 개발자의 서명을 확인할 수 있었다는 것이다[1][2].

악성코드를 시각화 하는 방법은 어렵지 않다. 바이너리 파일을 로딩 하여 각 byte를 integer로 환산하여 픽셀 값으로 사용하기만 하면 된다. 이를 이용하면 [그림 1]에서 보이는 것과 같이 시각화된 형태의 악성코드를 볼 수 있다. 복잡하게 동적인 분석을 통해 학습 데이터를 뽑아내지 않아도 되기 때문에 충분히 실시간 탐지에도 활용될 수 있을 것이다.

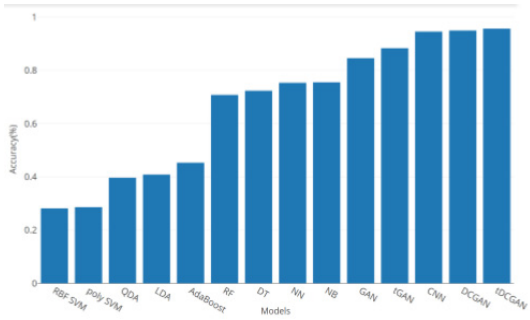
[그림 2]에서처럼 CNN을 이용해 이미지 변환된 악성코드의 특징을 추출하여 사용한다. 그리고 정확도를 높이기 위해 GAN을 이용하여 실제 이미지와 유사한 다양한 이미지를 생성해 학습에 사용하는 방법으로 정확도를 높이는 방법을 제안하고 있다.

해당 논문에서 제안하는 방법은 [그림 3]과 [그림 4]에서 보는 것처럼 DCGAN과 tDGAN으로 다른 알고리즘과 비교하여 정확도가 제일 높은 것을 확인할 수 있다.

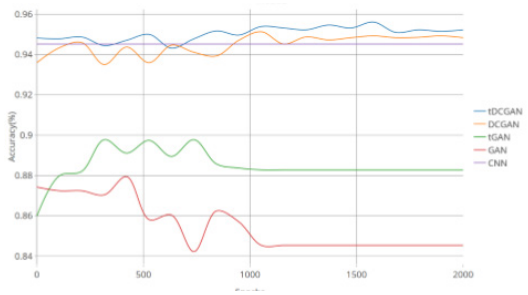
해당 논문에서 사용한 데이터는 Kaggle Microsoft Malware Classification Challenge에서 제공된 데이터 셋을 사용하였고 내용은 [표 1]과 같다.



〔그림 2〕 전이학습기반 악성코드 탐지용 GAN



〔그림 3〕 기존 알고리즘과의 비교



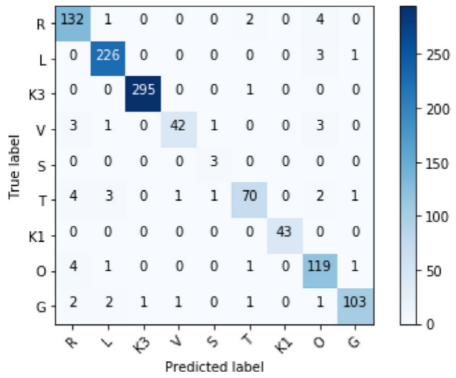
〔그림 4〕 딥러닝 알고리즘과 비교

〔표 1〕 실험에 사용된 악성코드 데이터 셋

Type	Samples
Ramnit(R)	1539
Vundo(V)	451
Kelihos_ver1(K1)	391
Lollipop(L)	2459
Simda(S)	42
Obfuscator.ACY(O)	391
Kelihos_ver2(K3)	2942
Tracur(T)	744
Gatak(G)	1011

[그림 5]에서 simda와 같은 경우 데이터의 개수가 충분치 않기 때문에 정확한 판단은 할 수 없었다. 트레이닝 데이터와 테스트 데이터의 비율은 90:10 이었고 정확도는 96.39%의 정확도를 가지는 것으로 나타났다.

악성코드가 가지는 가장 큰 특징들은 일반적으로 사용하는 API나 명령어들이 일반적인 프로그램들과는 크게 다르다는 것이다. 때문에 많은 논문들에서 이러한 API나 명령어들의 순서적인 특징을 잡아내기 위해 많



〔그림 5〕 Confusion matrix for malware detection

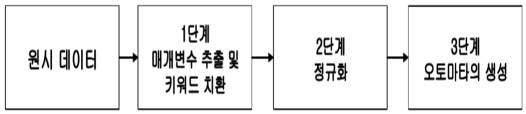
은 연구들이 되어 오고 있었다.

하지만 해당 논문에서는 그러한 순서적인 특징 외에도 공간적인 특징들만으로도 충분히 분류 및 탐지가 가능할 수 있음을 보였다는 것에 매우 큰 의미가 있다고 할 수 있다. 본 논문에서는 GAN과 같은 알고리즘 까지는 사용하지는 않을 것이지만 IDPS가 탐지하는 이벤트들에 대해서도 순서적인 특징 보다 공간적인 특징들을 보기 위하여 데이터들을 이미지 형태로 바꾸어 탐지에 충분히 활용할 수 있음을 보일 것이다.

2.2. 공격 프로파일링

IDPS의 탐지 이벤트의 형태는 탐지하는 제조사별로 다양한 형태로 기록이 된다. 본 논문에서는 탐지된 이벤트 로그들에 대한 분석을 수행해 공격 데이터를 공격 데이터로 탐지(정탐) 하였는지, 정상적인 데이터를 공격으로 탐지(오탐) 하였는지를 분류하는 방법을 제안할 것이기 때문에 로그에 기록되는 여러 항목들 중에 서도 페이로드를 사용해 분류하는 방법을 제안할 것이다. 이유는 사람이 탐지된 이벤트 로그를 보고 정/오탐을 판단한다고 했을 때에도 페이로드 의외의 항목은 판단의 기준을 가져가기 적절치 않다고 판단했기 때문이다. 페이로드의 형태도 다양한데 본 논문에서는 웹 공격 이벤트를 기준으로 HTTP 형태의 문자 기반의 페이로드로 한정해서 사용할 것이다.

기존의 웹 공격을 탐지하기 위한 방법으로 프로파일링 기반의 오토마타 접근 방법이 있었다[3]. 프로파일링 된 데이터들에 대해서 탐지하기 위한 방법으로 오토마타를 제안한 방법이었는데 이 부분에 오토마타 대신에



〔그림 6〕 오토마타 생성 과정

딥러닝 기반의 알고리즘을 적용하는 방법을 제안하고자 한다. 해당 논문에서 제안한 방법은 [그림 6]과 같다.

해당 논문에서 제안하는 방법은 수집 기간 중에 수집된 데이터들 중에 정상적인 데이터만 프로파일링 하여 오토마타를 생성하고 입력 값과 비교하여 일치 하지 않으면 비정상 행위로 간주하는 정상행위 기반의 프로파일링 방법을 제안하였다.

해당 논문에서 사용한 데이터는 수집 기간 중에 발생한 웹 URL 요청을 사용하였는데, 이 데이터를 그대로 사용하지 않고 특정 부분만을 추출하여 [그림 7]의 치환 테이블을 사용하여 특정 키워드로 치환하게 된다. [그림 8]에서처럼 정상 URL과 공격 URL에서 사용되는 특수 문자들의 형태가 다름을 볼 수 있다. 즉, 정상 URL의 특수문자 패턴과 공격 URL의 특수문자 패턴이 다르기 때문에, 특수 문자들의 패턴을 프로파일링 하여 탐지에 사용하기 위한 방법을 제안한 것이다. 이를 위해 각 URL에서 특수문자만을 추출하고 특정 키워드로 치환하여 프로파일링과 탐지에 사용한다.

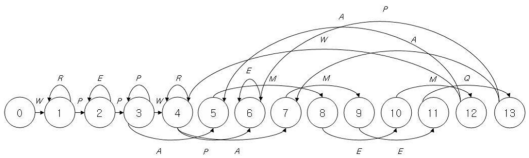
수집 기간 중에 수집된 정상 URL 데이터들은 전부 치환되어서 중복을 제거하고 정규화를 거쳐 [그림 9]와 같은 상태 전이 다이어그램을 볼 수 있고 이를 탐지에

문자	키워드	문자	키워드	문자	키워드	문자	키워드
write	WR	php	PP	.	PE		SP
login	LG	admin	AD	?	QE	'	AP
index	IN	delete	DE	=	EQ	or	OR
down load	DW	view	VW	&	AM	-	HY
upload	UP	update	UD	_	VS	"	BP

〔그림 7〕 키워드 치환 테이블

매개변수	Normal	login.php?id=pds&referer=&user_id=web_admin&password=1234
	Attack	login.php?id=pds&referer=&user_id='or 1 = 1 --&password=' or 1=1--
추출되는 키워드	Normal	login . php ? = & = & _ = _ & =
	Attack	login . php ? = & = & _ = ' or = - - & = ' or = - -
키워드 치환	Normal	LG PE PP QE EQ AM EQ AM VS EQ VS AM EQ
	Attack	LG PE PP QE EQ AM EQ AM VS EQ AP SP OR SP EQ HY HY AM EQ AP OR SP OR SP EQ HY HY

〔그림 8〕 키워드 치환 과정



〔그림 9〕 생성된 정상행위 프로파일에 대한 상태 전이도

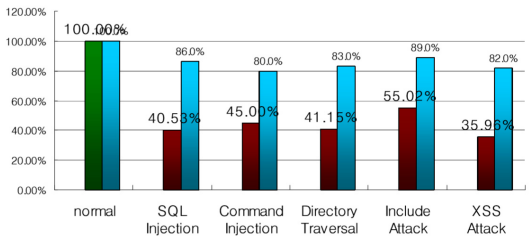
적용하기 위해 오토마타를 생성하여 URL요청에 대해 비교하여 정상과 비정상 행위를 탐지하게 된다.

해당 논문에서는 snort와의 비교를 통해 탐지 율을 보였는데 시그니처 기반의 알려진 공격을 탐지하는 snort와는 다르게 해당 논문에서 제안한 방법은 수집된 정상행위 이외의 요청에 대해서는 전부 공격으로 탐지하기 때문에 탐지 율에 대한 차이가 많이 날 수 밖에 없는 것으로 보인다. [그림 10]을 보면 snort와 비교하여 탐지 율이 높게 나온 것을 볼 수 있다.

하지만 정상행위 또한 정상행위로 인식하지 않을 확률이 높아져 오탐의 발생이 높은 것으로 나왔다. 이러한 문제는 기존의 딥러닝 기반의 학습 방법에서도 동일하게 발생하는 문제로 아무리 좋은 알고리즘을 이용한다 해도 양질의 많은 데이터들이 확보가 되지 않으면 소용이 없다.

2.1절의 악성코드 분류에서는 순서적인 데이터가 아니더라도 공간적인 특징을 이용하여 분류가 가능함을 확인할 수 있었고, 2.2절 에서는 웹 공격 데이터에 대한 특수 문자들의 패턴을 이용하여 효율적으로 공격을 탐지할 수 있음을 확인할 수 있었다.

본 논문에서는 살펴본 두 가지 연구를 중심으로 IDPS에서 발생한 탐지 데이터 중에 웹 공격 탐지 이벤트에서 발생한 페이로드를 활용하여 정/오탐을 분류하는 방법을 제시한다. 또한 이러한 방법은 WAF와 같은 다른 보안 장비에도 적용이 가능하며, 정/오탐 분류만이 아니라 IDPS의 공격 탐지에도 적용이 가능할 것이다.



〔그림 10〕 공격유형에 따른 snort와의 탐지 율 비교

Ⅲ. 데이터 전 처리

본 논문에서 사용하는 데이터 셋은 CSIC 2010 Dataset(36000개의 정상 요청, 25000개의 비정상 요청)을 사용하였다. 25000개의 비정상 요청에는 잘 알려진 SQL Injection, 버퍼 오버플로우, 정보 수집, 파일 노출, CSRF, XSS, SSI, 매개변수 조작과 같은 공격을 포함하고 있다.

3.1. 염기서열을 이용한 유사도 측정 방법

유전자 형식의 대표 형상 문자열 기법으로 구분하는 염기서열 방식을 응용하여 전 처리에서 페이로드의 문장에서 특수문자가 공격이 이루어지는 위협 데이터와 정상적인 데이터의 패턴이 다르다는 것에 착안하여 특수문자를 추출하여 응용하는 전 처리 방법을 적용했다 [4][5].

[그림 11]은 DNA 구조를 이루는 단백질의 정보이다. 단백질을 이루는 4가지의 아미노산 정보(A:아데닌, T:티민, G:구아닌, C: 사이토산)을 이용한 배열 정보이고, 이러한 아미노산의 배열 정보를 이용하여 각 DNA들 간의 유사성이나 역학적인 정보를 찾는데 활용한다.

[그림 12]에서처럼 특수문자들의 패턴을 이용하면 정상 데이터와 공격 형태의 분류가 가능함을 육안으로도 확인이 가능하다. 각 문자별 아스키로 대표되는 숫자들이 있기 때문에 실제 사용할 때는 특정 문자로 치환하지는 않는다.

[그림 11]을 보면 정렬을 했을 경우 육안으로도 공간

〔그림 11〕 염기서열을 이용한 DNA 유사도 측정

〔그림 12〕 페이로드 내 특수 문자의 패턴 추출

적인 특징이 있음을 잘 확인할 수 있다. 물론 순서적인 정보도 매우 중요한 정보일 것이다. 다만 공간적인 특징 또한 비정상 행위를 구분하는데 있어서 충분히 활용 가능함을 보이려고 한다.

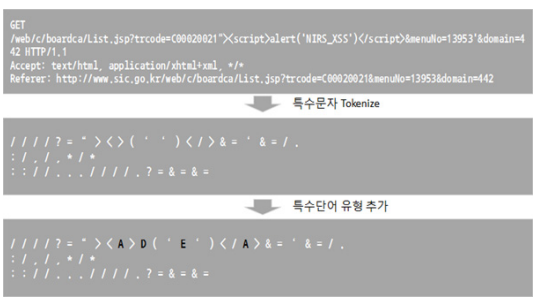
3.2. 공격 단어 적용

특수 문자의 구조를 통해 패턴화 하는 내용 외에도 특수 문자의 특징이 잘 드러나지 않는 공격 유형들도 존재한다. 정보 노출이나, 관리자 페이지에 대한 접근과 같은 것들이 그러한 것들이다. 이러한 유형에 대해서는 특수 문자의 패턴들이 잘 드러나지 않으므로 공격 유형별로 특별한 단어 사전을 만들어 학습 시에 사용할 수 있도록 하였다.

[그림 13]은 선정된 단어들의 사전을 미리 만들어 특징점 추출에 사용하게 되고, 최종적으로 학습에 필요한

A: html 태그	B: 자바스크립트 함수	C: 자바스크립트 함수	D: 실제 공격/변조 구	E: 문자열
innerHTML	onload	eval	alert	XSS
script	onmouseover	find	naviqator	Hello
svg	onsubmit	top	vibrate	fuzzelement
contenteditable	onfocus	source	document	test
x	onblur	toString	domain	INJECTX
src	onclick	URL	message	netsparker
iframe	oncopy	slice	write	openbugbounty
javascript	oncontextmenu	location	cookie	baiduspider
embed	oncut	hash	echo	csrf
math	ondblclick	setInterval	exec	
brute	ondrag	function	cmd	
href	oninput	appendChild	MSGBOX	

〔그림 13〕 선정된 공격 단어들



〔그림 14〕 Feature 추출 과정

Feature들을 추출하게 된다.

전체 추출 과정은 [그림 14]에서 보이는 것처럼 진행이 된다. 이러한 전 처리 과정을 적용하였을 경우 분석과 상관이 없는 데이터들을 제거함으로써, 정확도를 더 높일 수 있고, 딥러닝 기법이 크고 복잡한 데이터에 효율적이라고 하더라도 무한히 큰 데이터를 처리할 수는 없다. 즉, 데이터의 크기가 크고 많아질수록 학습에 걸리는 시간과 메모리의 양이 커지는 문제는 딥러닝 기법을 적용하여도 여전히 해결할 수 없는 문제이다. 때문에, 이러한 전 처리 기법을 사용하였을 경우 데이터의 크기가 줄이는 효과도 같이 가져올 수 있기 때문에 학습 성능이 개선될 수 있다.

[그림 15]는 레이블링(정탐:1, 오탐:0)까진 마친 학습 직전의 데이터의 모습이다.

1	[0, 0, 34, 47, 47, 95, 47, 46, 63, 61, 47, 38, 75, 61, ···	0
2	[0, 0, 47, 46, 63, 61, 38, 61, 38, 73, 61, 99, 47, 46, ···	0
3	[0, 0, 47, 46, 63, 73, 61, 38, 61, 38, 61, 39, 41, 70, ···	1
4	[0, 0, 47, 46, 63, 73, 61, 70, 65, 70, 80, 44, 80, 44, ···	1
5	[0, 0, 47, 46, 63, 73, 61, 95, 38, 61, 38, 61, 38, 61, ···	0
6	[0, 0, 47, 46, 63, 73, 61, 38, 95, 61, 95, 73, 38, 95, ···	1
7	[0, 0, 47, 46, 63, 61, 38, 73, 61, 95, 99, 47, 46, 34, ···	0
8	[0, 0, 47, 45, 47, 47, 99, 47, 46, 35, 35, 100, 58, 58, ···	0
9	[0, 0, 47, 46, 63, 61, 38, 61, 95, 70, 70, 65, 40, 44, ···	1
10	[0, 0, 47, 47, 63, 68, 61, 38, 61, 38, 61, 47, 38, 88, ···	1

〔그림 15〕 학습에 사용될 최종적인 데이터의 모습

Ⅳ. 알고리즘 선정

알고리즘을 선정한 중요 기준은 학습 데이터의 복잡도와 데이터의 양, 그리고 앞서 언급했던 학습 데이터의 특징, 여기서는 공간적인 특징을 보고자 했다.

그리고 전 처리를 통해서 데이터의 크기를 줄였다고는 하지만 페이로드의 크기가 기본적으로 아주 큰 크기를 가진 경우, 또는 공격 데이터에 특수문자가 많이 사용되는 경우에는 전 처리 과정을 통해도 사이즈는 여전히 클 수 있다. 때문에 GPU를 이용한 분산 처리가 필요하고, 데이터의 복잡도 또한 크기 때문에 딥러닝 기법에 대한 적용을 우선으로 고려하였다.

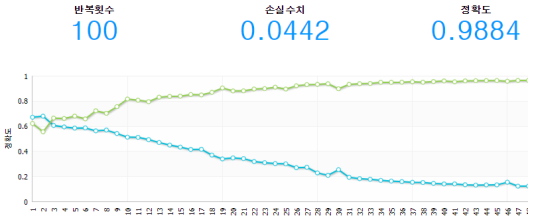
4.1. 모델 적용

본 논문에서 적용한 딥러닝 알고리즘은 영상 분야에서 많이 사용되는 CNN 알고리즘을 사용하였다. 순서적인 특징들도 특징이지만 공간적인 특징을 보려 하는게 첫 번째 이유이고, CNN 알고리즘의 특징인

Convolution과 Pooling과정을 거치면서 학습 데이터의 크기 또한 줄어들기 때문에 CNN을 우선 고려하여 실험을 진행 하였다.

4.2. 실험 결과

트레이닝 데이터와 테스트 데이터의 비율은 80:20으로 진행하였다. [그림 16]에서처럼 최종적인 정확도는 98.84%의 수치를 기록했고, 정밀도(precision)은 98.79%, 재현율(recall)은 98.80%의 수치를 기록하였다. 본 논문에서 제안한 내용과 같이 순서적인 특징이 아니더라도 공간적인 특징으로도 충분히 높은 정확도로 분류가 가능함을 알 수 있다. 또한, 문자 형태의 데이터임에도 불구하고 NLP와 같은 복잡한 자연어 형태의 분석을 하지 않아도 충분히 학습 데이터로서 사용이 가능함을 알 수 있다.



〔그림 16〕 최종 실험 결과

V. 결 론

본 논문에서는 IDPS 이벤트 로그들 중에서 문자 형태의 페이로드를 가지고 기존의 딥러닝 기법이 적용 가능함을 보이려고 했다. 그 결과 99%에 가까운 정확도로 분류가 가능함을 보였다. 또한 제안한 전 처리 기법을 적용하여 데이터를 크기를 줄일 수 있었고, 학습 시간이나 탐지 시간적인 측면에서 많은 향상을 이뤄낼 수 있었다. 앞으로 남은 과제들은 다른 알고리즘들과의 성능 비교를 통해 더 정확하게 성능에 대한 지표를 확인하는 것과, 정/오탐 뿐만 아니라 공격 형태에 따른 패턴을 학습할 수 있다면 공격 형태의 분류 또한 가능할 것으로 보인다. 그렇다면 기존 시그니처 기반의 탐지 방법보다 더 유연한 탐지 방법을 제안할 수 있을것으로 보이고, 아직 이 논문에서는 다루지 않았던 순서적인 특징까지 고려한다면 성능 향상을 위해 아직 많은 과제들을 남겨두고 있다고 할 수 있다.

참 고 문 헌

[1] Jin-Young Kim, Seok-Jun Bu, Sung-Bae Cho, “Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders”, Information Sciences 460-461, 83-102 2018.

[2] Jin-Young Kim, Seok-Jun Bu, Sung-Bae Cho, “Malware Detection Using Deep Transferred Generative Adversarial Networks”, ICONIP 2017, Part 1, LNCS 10634, pp. 556-564, 2017.

[3] 임종혁, “웹 공격 탐지를 위한 오토마타 프로파일링 접근 방법” 학위논문(석사), 전남대학교 대학원 일반대학원: 정보보호협동과정 2008. 2, 광주, 46p, 2008.

[4] Needlman, S, B. Wunsch C. D. "A general method applicable to the search for similarities in the amino acid sequence of two proteins", J. Mol. Biol, 48:443-453, 1970.

[5] Waterman, M. S. Smith T. F. Beyer W. A, "Some biological sequence metrics", Adv, Math, 20:367-387, 1976.

〈저 자 소 개〉

임 종 혁 (Im, Jong Hyuk)

2008년 2월 : 전남대학교 일반대학원 정보보호협동과정 석사과정 졸업
2018년 3월~현재 : 시큐레이어 재직
<관심분야> 머신러닝, 정보보호

김 진 (Kim, Jin)

2014년 2월 : 중앙대학교 컴퓨터공학부 학사 졸업
2016년 8월 : 중앙대학교 일반대학원 컴퓨터공학과 응용소프트웨어 석사과정 졸업
2016년 7월~현재 : 시큐레이어 재직
<관심분야> 영상처리, 머신러닝

김 건 우 (Kim, Kun Woo)

2002년 2월 : 명지전문대학 컴퓨터과 공업전문학사 과정 졸업
2014년 4월~현재 : 시큐레이어 재직
<관심분야> 머신러닝, 정보보호, 보안관제

유 진 상 (You, Jin Sang)

2000년 2월 : 조선대학교 정치외교학사 과정 졸업
2012년 2월~현재 : 시큐레이어 재직
<관심분야> 빅데이터, 검색엔진, 머신러닝, 정보보호, 보안관제