https://doi.org/10.33778/kcsa.2020.20.4.081

소프트맥스 함수 특성을 활용한 침입탐지 모델의 공격 트래픽 분류성능 향상 방안

김 영 원*, 이 수 진**

요 약

현실 세계에서는 기존에 알려지지 않은 새로운 유형의 변종 공격이 끊임없이 등장하고 있지만, 인공신경망과 지도학습을 통해 개발된 공격 트래픽 분류모델은 학습을 실시하지 않은 새로운 유형의 공격을 제대로 탐지하지 못한다. 기존 연구들 대부분은 이러한 문제점을 간과하고 인공신경망의 구조 개선에만 집중한 결과, 다수의 새로운 공격을 정상 트래픽으로 분류하는 현상이 빈번하게 발생하여 공격 트래픽 분류성능이 심각하게 저하되었다. 한편, 다중분류 문제에서 각 클래스에 대한 분류가 정답일 확률을 결과값으로 출력하는 소프트맥스(softmax) 함수도 학습하지 않은 새로운 유형의 공격 트래픽에 대해서는 소프트맥스 점수를 제대로 산출하지 못하여 분류성능의 신뢰도 또는 정확도를 제고하는데 한계를 노출하고 있다 . 이에 본 논문에서는 소프트맥스 함수의 이러한 특성을 활용하여 모델이 일정 수준 이하의 확률로 판단한 트래픽을 공격으로 분류함으로써 새로운 유형의 공격에 대한 탐지성능을 향상시키는 방안을 제안하고, 실험을 통해 효율성을 입증한다.

Improvement of Attack Traffic Classification Performance of Intrusion Detection Model Using the Characteristics of Softmax Function

Young-won Kim*, Soo-jin Lee**

ABSTRACT

In the real world, new types of attacks or variants are constantly emerging, but attack traffic classification models developed through artificial neural networks and supervised learning do not properly detect new types of attacks that have not been trained. Most of the previous studies overlooked this problem and focused only on improving the structure of their artificial neural networks. As a result, a number of new attacks were frequently classified as normal traffic, and attack traffic classification performance was severly degraded. On the other hand, the softmax function, which outputs the probability that each class is correctly classified in the multi-class classification as a result, also has a significant impact on the classification performance because it fails to calculate the softmax score properly for a new type of attack traffic that has not been trained. In this paper, based on this characteristic of softmax function, we propose an efficient method to improve the classification performance against new types of attacks by classifying traffic with a probability below a certain level as attacks, and demonstrate the efficiency of our approach through experiments.

Key words : IDS, AI, ML, DL, Activation Function, Softmax, Multiclass classification

접수일(2020년 09월 22일), 게재확정일(2020년 10월 12일)

* 국방대학교 국방과학학과 석사과정(주저자)
** 국방대학교 국방과학학과 교수(교신저자)

1. 서 론

과학기술정보통신부 통계에 따르면, 2020년 7월 기준 무선인터넷 트래픽은 2019년 7월 대비 29.2%P 증가한 687,348TB에 이르렀으며[1], 보안관제센터에서는 초당 15만 건, 하루에 10억 건에 달하는 보안관제 이벤트가 발생하고 있다[2]. 이처럼 매일 폭증하고 있는 유무선 네트워크 트래픽에 대해 이상징후를 기존 시그니쳐 기반의 방식으로 탐지하기란 매우 어려우며[3], 지능화되고 있는 보안 위협에 대한 관련지식 등을 검색하여 패턴화하는 것도 쉽지 않다[2].

침입탐지시스템은 악성 트래픽을 구분하는 방법에 따라 오용탐지 기반 침입탐지시스템(signature based IDS)과 비정상행위 기반 침입탐지시스템(anomaly based IDS)으로 구분되고, 침입탐지 데이터 소스의 출처에 따라 호스트기반 침입탐지시스템(host based IDS) 또는 네트워크 기반 침입탐지시스템(network based IDS)으로 구분할 수 있다[4]. 대부분의 침입탐지시스템은 이 가운데 오용탐지 및 네트워크를 기반으로 작동하며, 각종 공격을 분석하여 탐지에 필요한 패턴(시그너처)을 만들기 위해서는 특별한 전문가가 필요하고 지속적인 유지보수를 위한 비용도 발생한다. 그리고 기존 패턴을 우회하거나 변형된 공격을 감지하기 어렵다는 단점이 있다[2].

따라서 이러한 어려움을 극복하면서 침입탐지 성능을 개선하기 위해 최근에는 비정상행위 기반 침입탐지시스템에 인공지능을 접목하려는 시도가 활발하게 진행 중이다. 그중에서도 딥러닝(deep learning)을 활용한 방법이 가장 광범위하게 적용되고 있는데 이를 활용하면 수집된 네트워크 패킷을 미리 학습된 모델에 입력하여 정상 또는 공격 트래픽 여부를 실시간으로 확인할 수 있다. 그러나 인공신경망이 사용된 모델의 경우 학습 데이터의 구성에 따라 희소 클래스에 해당하는 공격은 그 특성을 올바르게 학습하지 못하기 때문에 탐지 정확도가 떨어지는 문제가 발생한다. 더욱 심각한 문제는 현실 세계에서는 기존에 알려지지 않은 새로운 공격 또는 변종 공격이 끊임없이 등장하지만, 지도학습을 통해 개발된 악성트래픽 분류모델은 새로운 유형의 공격을 학습한 적이 없으므로 이를 제대로 탐지하지 못한다는 사실이다. 그러나 대부분의 선행연구들은 이러한 문제점을 간과하고 인공신경망 구조 개선에만 집중하였기 때문에, 새로운 유형의 공격을 정상으로 분류하는 문제는 여전히 해결되지 않았고 전체적인 탐지성능 역시도 크게 개선되지 않고 있다.

한편 인공신경망에 기반한 다중분류 문제에서 출력층의 활성화 함수로 주로 사용되는 소프트맥스(softmax) 함수는 모델이 예측한 각 클래스가 정답일 확률을 결과값으로 출력한다. 그러나 학습하지 않은 새로운 유형의 공격 트래픽에 대해서는 클래스 간 소프트맥스 점수의 차이가 이미 학습한 유형의 트래픽에 비해 크지 않다. 이러한 현상은 NSL-KDD 데이터세트를 대상으로 한 식별성능 검증실험에서 명확하게 확인하였다. 즉 모델이 학습하지 않은 유형의 공격 트래픽은 모델 자신도 어떤 클래스인지를 확신할 수 없는 경우가 확연하게 증가하는 것이다. 이 경우 클래스 분류가 모호한 트래픽을 공격 트래픽으로 분류하면 침입탐지시스템을 회피하여 내부로 침투하는 새로운 유형의 악성 트래픽을 획기적으로 줄일 수 있다. 이러한 점에 착안하여, 본 논문에서는 소프트맥스 함수의 특성을 활용하여 정답일 확률을 일정 수준 이하로 판단한 샘플을 공격 트래픽으로 분류함으로써 침입탐지 성능을 향상시킬 수 있는 방안을 제안한다.

본 논문의 이후 구성은 다음과 같다. 2장에서는 딥러닝을 활용하여 침입탐지시스템의 성능향상을 시도했던 기존 연구들을 정리하고, 3장에서는 침입탐지시스템 관련 연구의 벤치마크 데이터세트로 알려진 NSL-KDD를 살펴본다. 4장에서는 제안하는 방안을 설명하고 실험 및 분석 결과를 기술하며, 마지막으로 5장에서 연구 결과를 요약하고 결론을 맺는다.

2. 관련 연구

Y.Lecun 등[5]이 그 개념을 제안한 이후 딥러닝은 시각 및 음성 인식, 자연어 처리 등에 폭넓게 적용되고 있으며 이를 활용하여 침입탐지 성능을 향상하기 위한 시도도 활발하게 진행 중이다[6].

Salama 등[7]은 침입탐지에 DBN(Deep Believe Network)을 처음 도입했는데 이때 특성추출을 위해

서 DNN(Deep Neural Network)을, 트래픽 분류를 위해 SVM(Support Vector Machine)을 사용하였다.

U. Fiore 등[8] 은 DRBM(Discriminative Restricted Boltzman Machine)을 사용하여 네트워크 환경에서 수집한 불완전한 데이터 세트로부터 유용한 정보를 얻는데 성공하였고, J. Kim 등[9] 과 Le 등[10] 은 RNN(Recurrent Neural Network)의 한 유형인 LSTM(Long-Short Term Memory)을 침입탐지 시스템에 접목하여 이전 연구에 비해 개선을 이루었다.

CNN(Convolutional Neural Network)은 딥러닝의 대표적인 알고리즘으로 이미지 분류에 뛰어난 성능을 보이며 침입탐지 분야에서는 대체로 트래픽을 이미지로 변환한 후 CNN을 활용하여 이미지의 클래스를 분류하는 방법 등으로 사용된다. 이 경우 초기에 별도로 특성을 추출하는 과정(feature selection)을 수행하지 않아도 되는 장점이 있다. 별도의 특성추출을 거치면 알고리즘의 연산량을 줄일 수는 있지만, 샘플의 특성 중 일부를 불가피하게 배제하게 되므로 손실이 발생한다. 반면 CNN을 사용하면 샘플의 특성을 모두 사용하면서 클래스를 효과적으로 분류할 수 있다[11]-[15].

그러나 위의 연구들은 단순히 이미지 분류를 위해 CNN을 사용하였을 뿐 데이터 세트의 불균형과 희소 클래스의 영향을 무시하였다. 이를 해결하기 위해 각 클래스의 비용함수 가중치 계수를 조정하는 방법[6], [16], 언더샘플링 등의 방법[17] 이 제안되었으나 분류모델 성능은 일정 수준(NSL-KDD 기준 ACC 80% 초반)을 넘기지 못하고 있다.

딥러닝을 침입탐지에 적용한 연구 중 가장 뛰어난 지표를 보인 연구 중 하나는 Y. Chuanlong 등[18]이 SGAN(Semi-supervised Generative Adversarial Network)을 침입탐지 모델 생성에 적용한 것으로, NSL-KDD 데이터세트에 대해 84.75%의 높은 이진분류 정확도를 달성하였다. 원래 GAN의 구분모델은 샘플이 실제 데이터세트에서 나온 것인지 아닌지만 판단하는 이진 분류기로 샘플의 클래스를 구분하는 능력은 없다[19]. 반면에 O. Augustus 등[20]이 제안한 SGAN의 구분모델은 분류모델 역할을 수행하여 샘플의 클래스까지도 구별할 수 있다.

침입탐지 모델의 성능이 일정 수준을 넘지 못하는 중요한 이유는 데이터 세트의 훈련세트와 시험세트의 세부 구성이 서로 다르다는 사실을 간과하고 있기 때문이다. 침입탐지 관련 연구의 벤치마크 데이터 세트인 NSL-KDD의 경우에도, 현실 세계에서 기존에 알려지지 않은 새로운 유형의 공격 또는 변종 공격이 끊임없이 등장한다는 경향을 반영하기 위해 시험세트에 훈련세트에는 등장하지 않는 새로운 유형의 공격이 다수 등장한다. 다음 장에서는 NSL-KDD데이터세트에 대해 보다 자세히 살펴본다.

3. NSL-KDD

NSL-KDD 데이터세트는 KDD cup'99의 개선된 버전이다. KDD cup'99에는 중복되는 데이터 샘플이 다수(78%)가 포함되어 있어 분류모델의 학습을 방해하므로 NSL-KDD에서는 이를 제거하였다. 각 트래픽 샘플은 41개의 특성정보를 가지며, 크게 5개 클래스(Normal, DoS, Probe, U2R, R2L)로 구분된다. 훈련세트와 시험세트의 구성은 <표 1>에서 보는 바와 같다[21].

<표 1> NSL_KDD 클래스 구성

Class	Nor-mal	DoS	Probe	U2R	R2L
KDD Train+	67,343	45,927	11,656	995	52
KDD Tset+	9,711	7,460	2,421	2,885	67

한편 각 공격 클래스는 세부 공격 타입의 집합으로 구성되는데, <표 2>에서 보는 바와 같이 시험세트에서는 훈련세트에서 등장하지 않는 17개 타입 3,750개의 샘플이 새롭게 등장한다.

4. 제안 방안

제안하는 침입탐지 모델 성능 향상 방안의 전반적인 개념은 다음과 같다. 먼저 NSL-KDD의 원시데이터를 인공신경망의 입력에 적합하도록 이미지로 변환 후 SGAN의 분류모델을 학습시킨다. 여기서 SGAN의

<표 2> NSL_KDD 공격 클래스 구성

Attack Classes	Total number of instances in the training set	Total number of instances in the test set
DoS	45,927	7,460
	back (956), land (18), neptune(41,214), pod (201), smurf (2,646), teardrop (892)	back (359), land (7), neptune(4,657), pod (41), smurf (665),teardrop (12)
		Additional attacks
		apache2 (737), udpstorm (2), worm (2), processtable (685), mailbomb (293)
Probe	11,656	2,421
	satan (3,633), ipsweep (3,599), nmap (1,493), portsweep (2,931)	satan (735), ipsweep (141), nmap(73), portsweep (157)
		Additional attacks
		mscan (996), saint (319)
R2L	995	2,885
	guess_passwd (53), ftp_write (8), imap (11), phf (4), multihop (7), warezmaster (20), warezclient(890), spy (2)	guess_passwd (1,231), ftp_write (3), imap (1), phf (2), multihop (18), warezmaster (944)
		Additional attacks
		xlock (9), xsnoop (4), snmpguess(331), snmpgetattack (178), httptunnel (133), sendmail (14), named (17)
U2R	52	67
	buffer_overflow (30), perl(3), loadmodule (9), rootkit (10)	buffer_overflow (20), loadmodule(2), rootkit (13), perl (2)
		Additional attacks
		sqlattack (2), xterm (13), ps (15)
Total	59,277	13,139

분류모델을 선택한 이유는 인공신경망을 이용하여 NSL-KDD 분류를 시도한 연구중 SGAN의 분류모델이 가장 뛰어난 성능을 보였기 때문이다. 이어서 시험 데이터세트의 각 샘플 별 소프트맥스 점수를 확인하고 이 점수가 일정 수준을 넘지 못하면 분류모델이 모호하게 분류한 것으로 간주하여 해당 샘플을 공격 트래픽으로 분류한다.

침입탐지시스템의 가장 중요한 목적은 정상 트래픽으로 가장하여 시스템 내부로 침투하는 공격 트래픽을 차단하는 것이다. 따라서 공격의 클래스를 세부적으로 구분하는 것에 앞서 해당 트래픽이 정상 트래픽인지 공격 트래픽 인지를 구분하고 공격 트래픽은 내부로 침투하지 못하도록 차단하는 것이 더 중요하다.

본 논문에서는 NSL-KDD 데이터세트에서 공격에 해당하는 4개의 클래스(DoS, Probe, U2R, R2L)를 모두 'Attack' 클래스로 재구성하고, 'Normal' 클래스와 함께 이진분류를 시도한다.

4.1 실험 환경

본 논문의 모든 실험은 Window 10 Home 64비트 운영체제, Intel Core i7-9700F CPU, 16.0GB RAM, NVIDIA GeForce RTX 2070 SUPER 8GB GPU 사양의 PC에서 Python 및 Keras를 이용하여 진행하였다.

4.2 원시데이터 전처리 및 이미지 변환

NSL-KDD 데이터세트의 42개의 속성(feature) 중 'num_outbound-cmds'은 모든 데이터가 '0'의 값을 가지기 때문에 모델의 분류성능에 영향을 미치지 않는

속성으로 판단하여 삭제하였다. ‘protocol_type’, ‘flag’, ‘service’ 3개 속성은 숫자가 아닌 범주형 데이터이므로 원-핫-인코딩(one hot encoding)을 통해 기호를 숫자로 매핑하였다. 예를 들어, ‘protocol_type’ 속성의 ‘TCP’, ‘UDP’, ‘ICMP’ 3가지 범주는 원-핫-인코딩을 통해 각각 (1, 0 ,0). (0, 1 ,0). (0, 0 ,1)로 매핑된다. NSL-KDD 데이터세트의 41차원 특성은 위의 과정을 거쳐 121차원 형상으로 변형된다.

NSL-KDD 데이터세트의 각 특성은 그 값의 분포가 서로 크게 다르므로 특정 특성이 분류모델의 학습에 과도하게 영향을 주지 않도록 정규화가 필요하다. 본 논문에서는 min-max 정규화를 통해 모든 값을 [0, 1] 범위로 변경하며 계산식은 다음과 같다.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad (1)$$

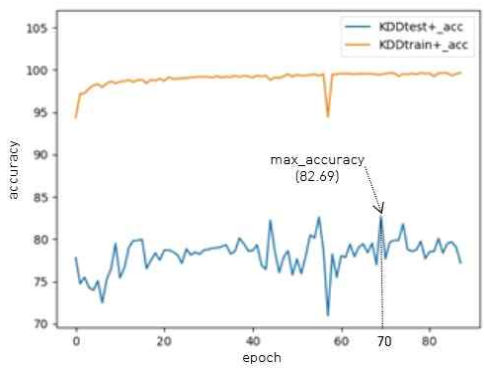
이어서 SGAN의 입력에 적절하도록 11×11 크기의 16비트 그레이 스케일 이미지로 변환한다.

4.3 분류모델 학습

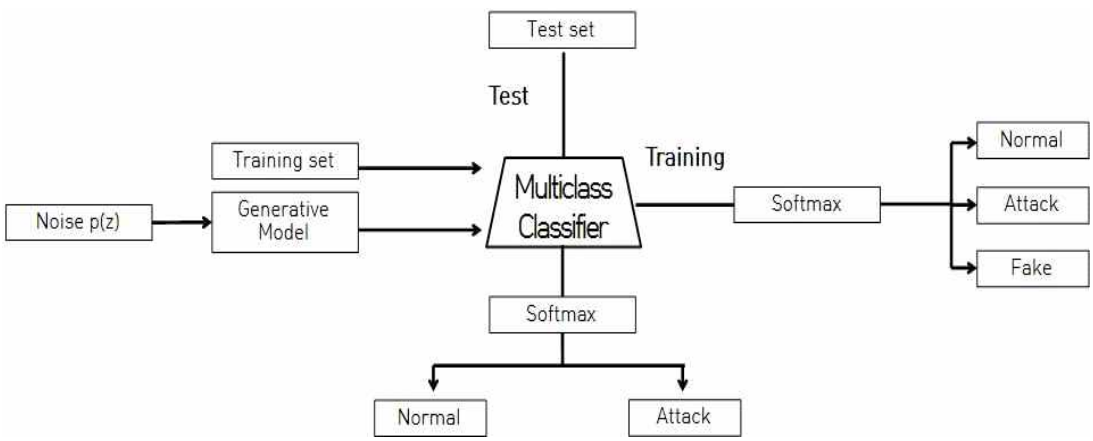
분류모델의 학습에는 (그림 1)에서 보는 바와 같이 학습세트의 정상 및 공격 트래픽 이미지와 SGAN의 생성모델이 만들어낸 가짜 이미지가 함께 입력으로 사용된다.

분류모델은 8개의 CNN 계층(CNN layer), 1개의 완전연결 계층(fully-connected layer), 출력층으로 구성되며 다중분류를 위해 출력층에서는 소프트맥스 함수가 사용된다. CNN 계층의 활성화 함수로 LeakyReLU를 사용하고, 과적합 방지를 위해 완전연결 계층 이후 Drop-Out 계층(0.4)을 배치하였다.

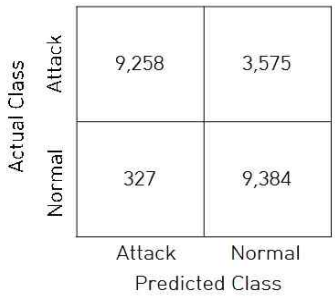
학습 결과는 (그림 2)에서 보는 바와 같이 Epoch 70에서 학습세트 99.45%, 시험세트 82.69%의 가장 높은 정확도를 달성하였으며, 해당 모델로 이후 실험을 진행하였다. 분류모델의 KDDTest+에 대한 오차행렬(그림 3)을 보면, 정상 트래픽은 대부분 정상으로 정확하게 분류하는 반면, 공격 트래픽은 총 12,833개 중 3,575개를 정상 트래픽으로 잘못 분류한 것을 확인할 수 있다.



(그림 2) SGAN 분류모델의 정확도



(그림 1) SGAN을 활용한 분류모델 학습



(그림 3) KDDTest+ 오차행렬

4.4 시험 데이터세트의 소프트맥스 점수 확인

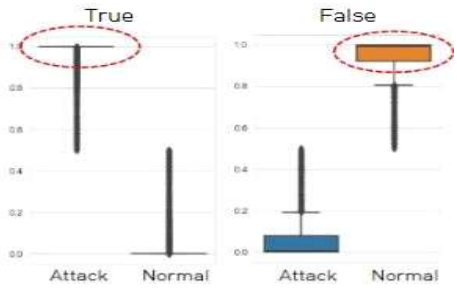
<표 3>은 분류모델의 KDDTest+에 대한 소프트맥스 점수 중 일부이다. 예를 들어, 1번 샘플의 경우 공격 트래픽일 확률 100%, 정상 트래픽일 확률 0%로 예측을 하였다. 29번 샘플의 경우 실제로는 공격 트래픽인 샘플을 70.097%의 확률로 정상 트래픽으로 잘못 분류하였는데 공격 트래픽일 확률과의 차이가 다른 샘플에 비해 낮은 것을 확인할 수 있다. 본 논문의 핵심 아이디어는 이와 같이 모호한 확률로 분류된 샘플을 공격 트래픽으로 재분류하는 것이다.

<표 3> KDDTest+ 소프트맥스 점수

Real class	Sample number	Softmax score		Predicted class
		Attack	Normal	
Attack	1	1.00000	0.00000	Attack
	10	0.99988	0.00012	Attack
	29	0.29903	0.70097	Normal
Normal	22541	0.99997	0.00003	Attack
	22542	0.00003	0.99997	Normal
	22543	0.00002	0.99998	Normal

연구의 목적을 달성하기 위해서는 (그림 3)의 오차행렬 중 실제 클래스가 공격 트래픽임에도 정상 트래픽으로 분류된 샘플 3,575개를 최대한 공격 트래픽으로 재분류해야 한다.

(그림 4)와 <표 4>에서 확인할 수 있는 바와 같이 모델이 실제 공격 클래스인 샘플을 공격으로 제대로 분류한 경우 최대 소프트맥스 점수는 평균 0.97, 표준편차 0.08의 좁은 구역에 집중되어 있다. 이는 모델이 높은 확률로 정답을 맞춘 샘플이 많음을 의미한다. 반면 실제로는 공격 트래픽인데 정상 트래픽으로 잘못 분류한 경우의 최대 소프트맥스 점수는 평균 0.93, 표준편차는 0.12로 앞의 경우보다 낮은 구역에 넓게 분포되어 있다. 이는 모델이 비교적 모호하게 잘못 분류한 샘플이 다수 존재함을 의미한다. 이러한 분석 결과를 통해 소프트맥스 점수를 이용하여 샘플을 재분류할 경우 모델의 탐지 정확도를 높일 수 있다는 가능성을 확인할 수 있다.

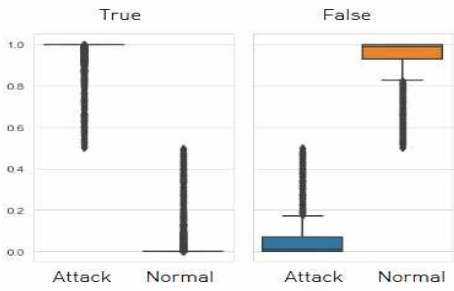


(그림 4) 공격 클래스 소프트맥스 점수 박스플롯

<표 4> 공격 클래스의 소프트맥스 점수 통계

Pred class	True		False	
	Attack	Normal	Attack	Normal
Count	9258		3575	
Mean	0.97	0.03	0.07	0.93
Std	0.08	0.08	0.12	0.12
Min	0.50	0.00	0.00	0.50
25%	0.99	0.00	0.00	0.92
50%	1.00	0.00	0.01	0.99
75%	1.00	0.00	0.08	0.99
Max	1.00	0.50	0.50	1.00

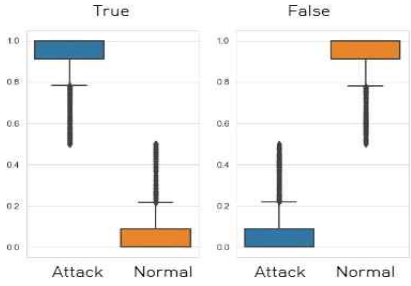
같은 방법으로 학습세트에서부터 존재했던 공격유형과 시험세트에 새롭게 등장하는 공격유형을 비교하면 (그림 5), (그림 6), <표 5>, <표 6>에서 보는 바와 같이 새로운 공격유형의 최대 소프트맥스 점수가 더 낮은 구역에 넓게 분포하고 있다. 이러한 사실을 통해 모델이 이미 알고 있는 유형의 샘플은 더 확실하게, 학습에 사용되지 않은 새로운 유형의 샘플은 낮은 확률로 분류를 하고 있다는 것을 알 수 있다.



(그림 5) 기존 공격 클래스의 소프트맥스 점수 박스플롯

<표 5> 기존 공격 클래스의 소프트맥스 점수 통계

Pred class	True		False	
	Attack	Normal	Attack	Normal
Count	7241		1842	
Mean	0.98	0.02	0.06	0.94
Std	0.06	0.06	0.10	0.10
Min	0.50	0.00	0.00	0.50
25%	1.00	0.00	0.00	0.93
50%	1.00	0.00	0.01	0.99
75%	1.00	0.00	0.07	1.00
Max	1.00	0.50	0.50	1.00



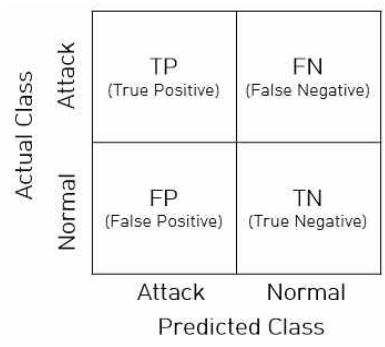
(그림 6) 새로운 공격 클래스의 소프트맥스 점수 박스플롯

<표 6> 새로운 공격 클래스의 소프트맥스 점수 통계

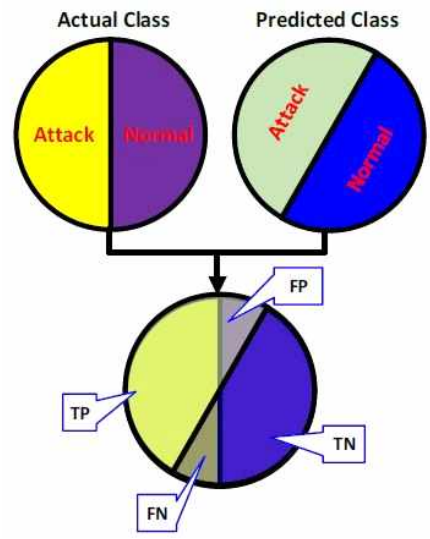
Pred class	True		False	
	Attack	Normal	Attack	Normal
Count	2017		1733	
Mean	0.93	0.67	0.08	0.92
Std	0.12	0.12	0.13	0.13
Min	0.50	0.00	0.00	0.50
25%	0.91	0.00	0.00	0.91
50%	1.00	0.00	0.00	1.00
75%	1.00	0.87	0.09	1.00
Max	1.00	0.50	0.50	1.00

4.5 성능평가지표

분류모델의 성능을 평가하기 위해서는 일반적으로 Accuracy, Precision, Recall, F1-score 등의 지표를 사용하지만, 침입탐지시스템 관련 연구에서는 직관적인 이해를 위해 AC(accuracy), DR(detection rate), FAR(false alarm rate) 등의 지표를 사용하기도 한다. 본 논문에서도 성능평가지표로 AC, DR, FAR을 사용하며, 이를 설명하기 위해 먼저 이진 분류에서 오차행렬과 요소들의 관계를 살펴보면 (그림 7)과 (그림 8)에서 보는 바와 같다[6].



(그림 7) 오차행렬



(그림 8) TP, FP, FN, TN의 관계[6]

세 가지 평가지표는 다음과 같이 정의할 수 있고, 모델의 성능을 향상시키기 위해서는 FAR을 과도하게 증가시키지 않으면서 AC와 DR을 향상시켜야 한다.

$$AC = \frac{TP + TN}{TP + FN + FP + TN} \quad (2)$$

$$DR = \frac{TP}{TP + FN} \quad (3)$$

$$FAR = \frac{FP}{FP + TN} \quad (4)$$

4.6 소프트맥스 점수를 이용한 클래스 재분류

<표 7>은 기준 소프트맥스 점수를 변경하면서 해당 점수 이하로 판단한 샘플을 공격 트래픽으로 재분류한 결과이다.

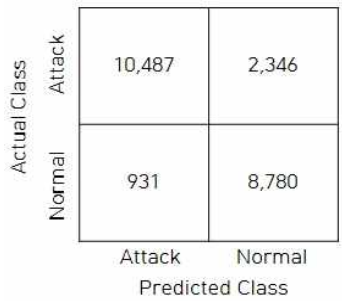
<표 7> 공격트래픽 재분류 결과

Max softmax score	AC (%)	DR (%)	FAR (%)
Original	82.69	72.14	3.37
< 0.60	83.29	73.28	3.49
< 0.70	83.89	74.46	3.64
< 0.80	84.48	75.74	3.96
< 0.90	84.97	78.13	6.00
< 0.91	84.83	78.52	6.83
< 0.92	84.77	78.98	7.58
< 0.93	84.83	79.46	8.06
< 0.94	84.95	79.99	8.50
< 0.95	85.08	80.46	8.80
< 0.96	85.09	80.89	9.36
< 0.97	85.46	81.47	9.59
< 0.98	86.03	83.07	10.06
< 0.99	86.48	84.68	11.14

예를 들어 최대 소프트맥스 점수가 0.90 이하인 샘플을 공격 트래픽으로 재분류할 경우 원래의 모델에 비해 AC는 2.28%P, DR은 5.99%P 증가하며 공격 트래픽으로 재분류하는 최대 소프트맥스 점수를 높일수록 AC, DR, FAR이 모두 상승한다.

FAR이 어느 정도 상승하더라도 AC와 DR을 높이는 것이 더 중요한 이유는 FAR이 높아지면 관리자가 해당 트래픽을 분석해야 하는 수고로움이 따를 뿐이지만 공격트래픽을 제대로 차단하지 못하면 내부 시스템이 공격받을 수 있기 때문이다.

FAR이 과도하게 증가하지 않는 기준을 10% 이하로 가정하면 최대 소프트맥스 점수 0.97 이하인 샘플을 공격 트래픽으로 재분류할 경우, AC 2.77%P, DR 9.33%P 가 상승한 효과를 얻을 수 있다. 제안 모델의 오차행렬은 (그림 9)에서 보는 바와 같다.



(그림 9) 제안 모델의 오차행렬

4.7 기존 연구와의 비교

<표 8>은 본 연구에서 제안하는 방법을 관련연구에서 언급한 기존 연구들과 비교한 결과를 보여주고 있다. 제안하는 방법이 기존 연구에 비해 AC는 0.68~3.77%P, DR은 5.72~8.93%P 까지 전반적으로 높으므로 충분히 효용성을 가진다고 할 수 있다. 특히 기존 연구 중 가장 뛰어난 성능을 보인 [18]의 모델과 비교하여 DR이 5.72%P 높다는 것은 그 만큼 더 많은 공격 트래픽을 정확히 분류했다는 것을 의미한다.

<표 8> 기존 연구와의 비교 결과

Model	AC (%)	DR (%)
Proposed	85.46	81.47
[6]	81.69	72.54
[18]	84.75	75.75

5. 결 론

본 논문에서는 인공신경망을 기반으로 생성한 침입탐지 모델의 공격 트래픽 분류성능을 향상하기 위해 모델의 시험세트에 대한 최대 소프트맥스 점수를 확인하고, 일정 수준 이하로 판단한 샘플을 공격트래픽으로 재분류하는 기법을 제안하였다. 제안된 기법을

구현하기 위해 NSL-KDD 데이터세트을 16비트 그레이 스케일 이미지로 변환하고, SGAN의 분류모델을 학습 후 사용하였다. 실험을 통해 분류성능을 확인한 결과, 정상 트래픽에 비해 공격 트래픽이, 학습세트에 존재하던 기존의 공격 유형보다 시험세트에서 새롭게 등장한 공격유형의 최대 소프트맥스 점수가 낮고 넓은 범위에 분포한다는 사실을 확인하였다. 그리고 정답일 확률을 일정 수준 이하로 판단한 샘플을 공격 트래픽으로 재분류함으로써 AC는 2.77%P, DR은 9.33%P 상승한 효과를 거두었다. 이러한 결과는 제안된 침입탐지 모델이 탐지를 회피하여 네트워크 내부로 침투하는 공격트래픽을 획기적으로 줄일 수 있음을 확인시켜 준다. 또한 제안된 기법은 SGAN의 분류모델뿐만 아니라 출력층에 소프트맥스를 사용하는 다른 형태의 신경망으로 구현된 침입탐지 모델이나 NSL-KDD 이외의 다른 데이터세트에도 적용 할 수 있는 방법이다.

향후에는 SGAN 이외의 신경망과 데이터세트를 활용하고 이진분류 이상의 다중분류 문제에 적용 가능한 방법으로 연구를 확장할 예정이다.

참 고 문 헌

[1] 과학기술정보통신부, “무선데이터 트래픽 통계”, Online, 2020Available:https://msit.go.kr/web/msipContents/contentsView.do?cateId=_status&artId=3067561

[2] 국경완, 공병철, “인공지능을 활용한 보안기술 개발 동향”, 정보통신기획평가원 주간기술동향, 1913호, pp 5, 2019.

[3] 박형근, “정보보안에서의 인공지능 도입 분야와 주요 사업자”, 시큐리티플러스, pp.3-9, 2018.

[4] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, “Survey of intrusion detection systems: techniques, datasets and challenges”, Cybersecurity, pp.1-22, 2019.

[5] Y. Lecun, Y. Bengio, G. Hinton, "Deep learning", Nature, vol.521, no.7553, pp.436-444, 2015.

[6] W. Kehe, C. Zuge, L. Wei, “A Novel Intrusion Detection Model for a Massive Network Using Convolutional Neural Networks”, IEEE Access, vol.6, pp.50850-50859, 2018.

[7] M. A. Salama, H. F. Eid, R. A. Ramadan, A. Darwish, A. E. Hassanien, "Hybrid intelligent intrusion detection scheme", Soft Computing in Industrial Applications, Berlin, Germany:Springer, pp.293-303, 2011.

[8] U. Fiore, F. Palmieri, A. Castiglione, A. De Santis, "Network anomaly detection with the restricted Boltzmann machine", Neurocomputing, vol.122, pp.13-23, Dec, 2013.

[9] J. Kim, J. Kim, H. L. T. Thu, H. Kim, "Long short term memory recurrent neural network classifier for intrusion detection", Proc. Int. Conf. Platform Service, pp.1-5, 2016.

[10] T. T. H. Le, J. Kim, H. Kim, "An effective intrusion detection classifier using long short-term memory with gradient descent optimization", Proc. Int. Conf. Platform Technol. Service, pp.1-6, 2017.

[11] R. Vinayakumar, K. P. Soman, P. Poornachandran, "Applying convolutional neural network for network intrusion detection", Proc. Int. Conf. Adv. Comput. Commun. Inform., pp.1222-1228, 2017.

[12] W. Wang, M. Zhu, X. Zeng, X. Ye, Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning", Proc. Int. Conf. Inf. Netw., pp.712-717, Jan. 2017.

[13] M. Wang, J. Li, "Network intrusion detection model based on convolutional neural network", J. Inf. Secur. Res., vol.3, pp.990-994, 2017.

[14] E. Min, J. Long, Q. Liu, J. Cui, W. Chen, "TR-IDS: Anomaly-based intrusion detection through text-convolutional neural network and random forest", Secur. Commun. Netw., vol. 2018, Jul. 2018.

[15] W. Wang, "HAST-IDS: Learning hierarchical spatial-temporal features using deep neural

networks to improve intrusion detection", IEEE Access, vol.6, pp.1792-1806, 2018.

[16] L. Peng, H. Zhang, Y. Chen, B. Yang, "Imbalanced traffic identification using an imbalanced data gravitation-based classification model", Comput. Commun., vol. 102, pp. 177-189, 2016.

[17] Y. Liu, S. Liu, X. Zhao, "Intrusion detection algorithm based on convolutional neural network", Beijing Ligong Daxue Xuebao/Trans. Beijing Inst. Technol., vol.37, no.12, pp.1271-1275, 2017.

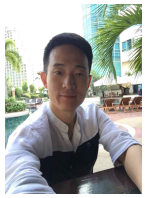
[18] Chuanlong Y., Yuefei Z., Shengli L., Jinlong F., Hetong Z, "Enhancing network intrusion detection classifiers using supervised adversarial training", The Journal of Supercomputing, pp. 6690-6719, 2020.

[19] G. Ian, P. Jean, M. Mehdi, X. Bing, W. David, O. Sherjil, C. Aaron, B. Yoshua, "Generative Adversarial Nets", Neural Information Processing Systems, pp.2672-2680, 2014.

[20] O. Augustus, "Semi-Supervised Learning with Generative Adversarial Networks", arXiv preprint arXiv:1606.01583, 2016.

[21] T. Mahbod, B. Ebrahim, L. Wei, A. AlI, "A Detailed Analysis of the KDD CUP 99 Data Set", 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.

〔저자소개〕



김 영 원 (Young-won Kim)

2009년 3월 해군사관학교 전산학과 학사
2020년 9월 국방대학교 국방과학학과 석사

email : headsun21@gmail.com

이 수 진 (Soo-jin Lee)

1992년 3월 육군사관학교 전산학과 학사
1996년 2월 연세대학교 컴퓨터과학과 석사
2006년 2월 한국과학기술원 전산학과 박사
2006년 ~ 현재 국방대학교 국방과학학과 교수

email : cyberkma@gmail.com