Journal of KIIT. Vol. 22, No. 10, pp. 27-34, Oct. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.10.27

패킷 페이로드 분석과 오버샘플링을 적용한 트랜스포머 기반 침입탐지 모델

이태화*, 이수진**

Transformer-based Intrusion Detection Model with Packet Payload Analysis and Oversampling

Taehwa Lee*, Soojin Lee**

요 약

인공지능을 기반으로 하는 침입탐지 모델들은 합성 데이터인 메타데이터를 학습하기 때문에 실제 네트워크에서 발생하는 패킷을 이용해 침입을 탐지하는 것이 제한된다. 그리고 학습 과정에서 데이터의 불균형 문제를 해결하지 못할 경우 다중분류에서 특정 클래스에 대한 탐지성능이 저하될 수 있다. 이러한 문제를 해결하기 위해 본 논문에서는 패킷 페이로드를 트랜스포머 기반의 언어모델을 통해 전처리 없이 분석하고, 오버샘플링을 적용하여 희소 클래스에 대한 탐지성능을 향상시킨 침입탐지 모델을 제안하였다. UNSW-NB15 데이터세트를 사용해 성능평가를 수행하였으며, 학습 모델은 RoBERTa 모델을 사용하였다. 실험 결과 ADASYN으로 오버샘플링을 수행했을 때 다중분류에서 가장 높은 정확도인 87.15%를 달성하였다.

Abstract

Most intrusion detection models based on artificial intelligence use meta-data, which is synthetic data generated through packet analysis. Therefore, it is limited to detect intrusion using packets occurring in real networks. In addition, failure to solve the data imbalance problem in the learning process can significantly degrade detection performance for specific classes with a small number of data in multi-class classifications. To address these problems, we propose an intrusion detection model that analyzes packet payloads without preprocessing through a transformer-based natural language processing model and improves detection performance for rare classes through oversampling. We used UNSW-NB15 dataset for performance evaluation, and the RoBERTa model was used as the training model. As a result of the experiment, 87.15% of accuracy was achieved in multi-class classifications when oversampled with ADASYN.

Keywords

intrusion detection, packet payload, oversampling, transformer, natural language processing

* 국방대학교 컴퓨터공학과 박사과정
- ORCID: https://orcid.org/0009-0005-2462-1519
** 국방대학교 컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4117-407X

· Received: Jul. 16, 2024, Revised: Aug. 02, 2024, Accepted: Aug. 05, 2024

· Corresponding Author: Soojin Lee
Dept. of Computer Engineering, Korea National Defense University, 1040, Hwangsanbul-ro, Nonsan-si, Chungcheongnam-do, Republic of Korea
Tel.: +82-41-831-5378, Email: cyberkma@korea.kr

Ⅰ. 서 론

기계학습 또는 딥러닝 모델은 거대한 데이터세트를 신속하게 처리하고 분석하여 의미있는 패턴을 정확하게 찾아낸다. 이러한 장점으로 인해 대규모 위협정보에 대한 분석을 통해 잠재적 위협과 취약점을 신속하게 찾아내는 것이 매우 중요한 사이버 보안 분야에서 인공지능 기술의 활용은 날로 증가하고 있다.

특히 침입탐지 분야에서는 이전의 위협정보를 기반으로 의심스러운 공격 활동을 식별하거나 잠재적 위협을 예측하는 연구들이 매우 활발하게 수행되고 있다. 그리고 학술적인 차원에서 제시된 인공지능 기반 탐지 모델들이 보여준 각종 성능평가 지표는 거의 100%에 가까워지고 있다. 그러나 선행연구를 통해 제시된 인공지능 기반의 침입탐지 모델들은 대부분 합성데이터(Synthetic data)를 이용해 학습을 수행했고, 학습 데이터와 통계적 특성이 동일한 테스트 데이터에 대해서만 우수한 분류 및 탐지성능을 발휘할 수 있다는 문제점을 안고 있다[1].

실험 네트워크 혹은 실제 네트워크에서 수집된 패킷을 분석하여 가공함으로써 생성되는 합성데이터인 메타데이터(Meta-data)를 학습한 모델은 그 자체로 실제 네트워크에서 생성된 패킷을 이용하여 침입탐지를 수행하는 것이 불가능하다. 모델이 침입탐지를 정상적으로 수행하기 위해서는 입력 패킷을 학습 데이터와 동일한 통계적 특성을 가진 메타데이터로 변환하는 과정이 반드시 선행되어야 하며, 이로 인해 실시간 침입탐지가 제한된다.

이러한 문제 인식하에 본 연구진은 실제 네트워크에서 동작 가능한 침입탐지 모델을 개발하기 위해 메타데이터가 아닌 패킷 페이로드(Payload) 기반으로 학습 모델을 구축하는 방안을 제시하였다[2]. 침입탐지 분야에서 활용되는 대표적인 데이터세트인 UNSW-NB15[3]에 포함된 PCAP 파일에 대해 먼저 라벨링을 실시한 후 학습 데이터로 활용하였다. 후속연구[4]에서는 동일한 데이터세트에 트랜스포머(Transformer) 모델을 적용하여 패킷 페이로드를 문장으로 학습하는 자연어 처리(Natural language processing) 기반의 침입탐지 모델을 제시하였으며, 두 가지 접근방법 모두 기존 연구 대비 우수한 탐지성능이 보임을 확인하였다.

그러나 공격 클래스 간 불균형 문제를 간과하여 다중분류에서 특정 클래스에 대한 탐지성능이 현저하게 저하되는 현상도 확인하였다. 이에 본 연구에서는 선행연구들에서 해결하지 못한 데이터 불균형 문제를 극복하기 위해 오버샘플링(Over-sampling) 기법을 추가적으로 적용하고 자연어 처리 성능이 보다 개선된 트랜스포머 모델을 활용하여 침입탐지 모델을 구축하는 방안을 제안한다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서 본 연구에 활용된 트랜스포머 모델과 성능평가 실험에 사용한 UNSW-NB15 데이터세트 및 오버샘플링 기법에 대해 설명하고, 관련된 기존 연구를 정리한다. 3장에서는 제안하는 침입탐지 모델의 구축방안과 실험 방법을 설명하고 실험결과를 분석한다. 마지막으로 4장에서 연구를 요약하고 결론을 맺는다.

Ⅱ. 관련 연구

2.1 RoBERTa

RoBERTa(A Robustly Optimized BERT Pretraining Approach)[5]는 BERT(Bidirectional Encoder Representations from Transformers)[6]를 개선한 언어모델로 다양한 자연어 처리 모델 중 상당히 뛰어난 성능을 보여주고 있다.

RoBERTa는 BERT의 사전 학습 과정에서 크게 4가지 사항이 개선되었다. 우선, BERT보다 더 다양하고 복잡한 패턴을 학습해 언어 이해능력이 향상되었으며, 동적 마스킹을 적용해 학습 Epoch마다 마스킹하는 단어를 변경해 새로운 표현을 학습하는 효과를 나타내었다. 그리고 배치 크기를 증가시켜 더 빠르고 효율적인 학습이 가능하도록 하였으며, NSP(Next Sentence Prediction) 태스크를 제거해 문맥 내 단어들 간 관계를 좀 더 집중적으로 학습시켰다.

2.2 데이터세트

UNSW-NB15 데이터세트는 호주 사이버보안센터의 UNSW Cyber Range Lab에서 패킷을 수집하고 정상 및 비정상 트래픽을 분류한 공공 데이터세트

로서, 라벨링이 되어 있지 않은 PCAP 파일과 라벨링이 되어 있는 메타데이터 2가지 형태로 제공된다. 총 175,341개의 학습 데이터와 82,232개의 테스트 데이터를 포함하고 있고, 9개의 유형으로 분류되어 있다.

본 연구에서는 패킷 페이로드를 통한 침입탐지를 수행하기 위해 먼저 Payload-Byte[7] 기법을 적용해 메타데이터와 비교하면서 PCAP 파일에 라벨링을 실시하였다. 라벨링된 데이터에 대해서는 결측치와 중복 제거 등의 전처리를 거쳤으며, 최종적으로 59,168개의 데이터만 활용하여 실험을 진행하였다. 표 1은 본 연구의 실험에서 활용한 데이터세트의 세부 구성을 보여주고 있다.

표 1. 실험에 활용한 데이터세트의 구성
Table 1. Experimental dataset configuration

Class	Label	Train data	Test data
Analysis	0	538	134
Backdoor	1	620	155
DoS	2	2,210	552
Exploits	3	9,794	2,449
Fuzzers	4	8,476	2,119
Generic	5	9,457	2,364
Normal	6	9,849	2,463
Reconnaissance	7	5,644	1,411
Shellcode	8	680	170
Worms	9	66	17
Total	-	47,334	11,834
		59,168	

2.3 오버샘플링

오버샘플링은 데이터의 불균형 문제를 해결하기 위해 다른 클래스에 비해 소수인 클래스의 데이터를 증가시키는 방법이다. 본 연구에서는 대표적인 오버샘플링 기법 중 랜덤샘플링(Random sampling), ADASYN(Adaptive Synthetic Sampling Approach) 및 SMOTE(Synthetic Minority Over-sampling Technique) [8] 3가지 기법을 활용하였다.

랜덤샘플링은 소수 클래스의 데이터를 무작위로 복제하는 기법으로, 단순하면서도 구현이 빠르지만 동일한 데이터를 복제하기 때문에 과적합의 위험이 있다. ADASYN은 k-NN(k-nearest neighbor) 알고리즘을 이용해 데이터로부터 거리가 가까운 k개의 이웃을 찾은 후 학습 난이도에 따라 가중치를 부여하고 학습이 어려운 샘플 근처에 더 많은 데이터를 생성한다. 이 기법은 모델의 학습 효율을 향상시킬 수는 있지만, 복잡하고 가중치 계산에 더 많은 시간이 소요된다. SMOTE는 소수 클래스를 k-NN 알고리즘을 사용해 이웃을 찾고, 무작위로 이웃을 선택하여 두 데이터 사이에 새로운 데이터를 생성한다. 새로운 데이터를 생성하여 다양성을 높이지만 고차원 데이터의 경우 실제 데이터의 분포를 제대로 반영하지 못할 수도 있다.

2.4 선행연구 고찰

인공지능 기술을 기반으로 한 침입탐지 연구는 매우 다양하게 진행되었다. 그러나 본 절에서는 동일한 데이터세트를 사용한 연구, 불균형한 데이터 문제를 해결하기 위해 샘플링 기법을 적용한 연구, 그리고 트랜스포머 모델을 사용한 연구들을 중점적으로 고찰한다.

먼저 UNSW-NB15 데이터세트를 사용해 모델을 학습시키고 성능을 평가했던 연구들은 다음과 같다.

G. N. Kim et al.[2]은 데이터세트의 샘플을 줄이는 GOSS (Gradient-based One-Side Sampling)와 데이터세트의 특성 수를 줄이는 EFB(Exclusive Feature Bundling) 알고리즘을 활용하여 LightGBM 모델에 학습시킨 결과, 이진분류 99.33% 및 다중분류 85.63%의 정확도를 달성하였다. W.-S. Park et al.[4]은 BERT를 활용하여 UNSW-NB15 데이터세트에서 제공하는 패킷 페이로드를 16진수 형태의 문장처럼 학습시켰으며, 다중분류에서 86.63%의 정확도를 달성하였다.

D. Jing et al.[9]은 비선형 로그 함수 스케일링으로 데이터를 전처리 후 SVM(Support Vector Machine)으로 분류를 시도하여 Random Forest나 Naive Bayes 모델보다 높은 정확도를 달성하였다.

S. Meftah et al.[10]은 Random Forest 모델을 통해 중요 특성을 선택하고 결정트리와 다른 분류 기법들을 결합하여 적용하는 2단계 하이브리드 접근방법을 제시하였다.

실험결과에서 결정트리와 다중 클래스 SVM 모델을 결합한 하이브리드 모델의 다중분류 정확도가 86.04%로 가장 높게 나타났다.

S. M. Kasongo et al.[11]은 XGBoost 모델을 통해 특징들의 중요도를 계산하고 19개의 중요 특징만을 선택하여 적용하였다. SVM, 로지스틱 회귀, 인공신경망, 의사결정 트리, k-NN 알고리즘 등을 통해 학습 및 테스트를 실시한 결과, 인공신경망 모델이 77.51%로 가장 높은 성능을 보였다.

다음으로 불균형한 데이터로 인해 모델의 탐지성능이 저하되는 문제를 해결하기 위해 샘플링 기법을 적용한 연구들은 다음과 같다.

S. Bagui et al.[12]은 불균형한 데이터들에 대해 샘플링을 적용하여 클래스 간 비율을 조정하고 균형 잡힌 데이터세트를 생성하였다. 랜덤 언더샘플링, 랜덤 오버샘플링, SMOTE 및 ADASYN 등 다양한 기법들을 NSL-KDD, UNSW-NB15 등에 적용하였으며, 인공신경망 기반 모델에 학습시켜 샘플링 전보다 더 향상된 성능을 가진 모델을 구현하였다.

H. A. Ahmed et al.[13]은 데이터를 주성분 분석(Principle component analysis) 기법을 통해 차원을 축소하고, 클래스 간 불균형 문제를 해결하기 위해 SMOTE 기법으로 오버샘플링을 실시하였다. 전처리한 데이터를 인공신경망 모델로 학습하였으며, 정확도는 77.6%로 나타났다.

마지막으로 본 연구와 유사하게 트랜스포머 모델을 활용해 침입탐지를 시도했던 연구들은 다음과 같다.

Z. Wu et al.[14]은 트랜스포머 모델의 Self-attention 메커니즘을 통해 데이터의 위치 임베딩으로 정보들 간의 관계와 문맥를 학습한 침입탐지 모델을 구현하였다. CICDS2017와 CIC-DDoS 2019 데이터세트를 이용해 학습을 실시하고 성능을 평가한 결과 이진분류에서 각각 99.98%와 99.65%의 정확도를 달성하였다.

Y.-G. Yang et al.[15]은 이미지 분류에서 우수한 성능을 보인 Vision Transformer(ViT)를 이용하여 침입탐지 모델을 구축하였다. NSL-KDD 데이터세트를 학습하였으며, 이진분류에서 ResNet-18 모델보다 0.35%p 향상된 99.68%의 정확도를 달성하였다.

III. 오버샘플링을 적용한 트랜스포머 기반 침입탐지 모델

3.1 제안 방법

본 연구진의 선행연구에서는 BERT와 DistilBERT 모델을 사용하였다. 그러나 본 연구에서는 다중분류 성능을 보다 개선하기 위해 BERT보다 성능이 크게 향상된 RoBERTa 모델을 사용한다. 또한, 데이터 불균형 문제를 극복하기 위해 오버샘플링 기법을 활용하여 희소 클래스의 데이터를 샘플링한다.

침입탐지 모델 구축 및 성능평가 절차는 그림 1에서 보는 바와 같다.

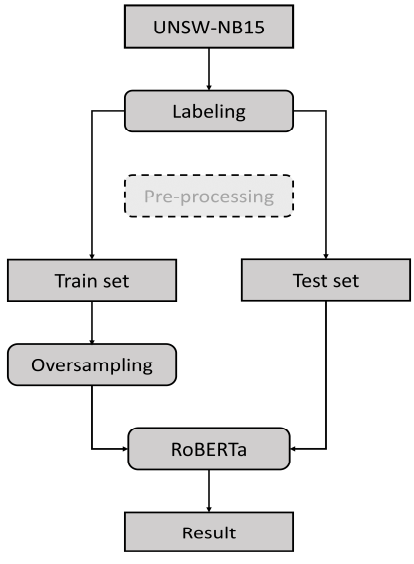

그림 1. 개요
Fig. 1. Overview

UNSW-NB15 데이터세트에 대한 라벨링을 먼저 수행한 후 학습 데이터세트와 테스트 데이터세트로 분할한다. 이어서 학습 데이터세트에 대해서는 오버샘플링을 적용하여 희소 클래스 5개의 데이터 수를 증가시켜 RoBERTa 모델로 학습을 실시하고, 테스트 데이터세트를 통해 모델의 성능을 평가한다. 보다 자세한 절차는 3.2절에서 설명한다.

오버샘플링이 적용된 학습 데이터세트의 세부적인 구성은 표 2에서 보는 바와 같다.

표 2. 오버샘플링 후 학습 데이터세트의 구성
Table 2. Configuration of training data after oversampling

Class	Before oversampling		After oversampling		
	#	%	#	%	Sampling
Normal	9,849	20.81	9,849	17.63	-
Exploits	9,794	20.69	9,794	17.53	-
Generic	9,457	19.98	9,457	16.93	-
Fuzzers	8,476	17.91	8,476	15.17	-
Reconnaissance	5,644	11.92	5,644	10.1	-
DDoS	2,210	4.67	3,315	5.93	× 1.5
Shellcode	680	1.44	2,720	4.87	× 4
Backdoor	620	1.31	2,480	4.44	× 4
Analysis	538	1.14	2,152	3.85	× 4
Worms	66	0.14	1,980	3.54	× 30
Total	47,334	-	55,867	-	-

3.2 실험 방법

먼저, 라벨링되어 있지 않은 데이터세트의 PCAP 파일을 Payload-Byte 기법을 통해 라벨링하고 실제 패킷 페이로드와 동일한 형태인 16진수로 변환한다. 그리고 8:2 비율로 학습 데이터와 테스트 데이터를 분리한다.

성능평가 실험은 오버샘플링에 의한 효과를 검증하기 위해 2단계로 구분하여 진행한다. 우선 오버샘플링 전 데이터세트를 RoBERTa 모델에 학습시켜 결과를 도출한다. 이후, 학습 데이터에 랜덤샘플링, SMOTE 및 ADASYN을 적용하여 희소 클래스의 데이터를 증가시킨다.

오버샘플링 적용 시에는 클래스 중 데이터 수가 가장 많은 클래스에 맞추지 않고, Bagui 등[12]이 제시한 방법처럼 전체 데이터 중 5% 미만에 해당하는 하위 5개 클래스에 대해서만 각각의 비율이 최소 3%가 넘도록 조절하였다. 표 2에서 보는 바와 같이 데이터를 1.5배에서 최대 30배 증가시켜 모든 클래스가 전체 데이터세트에서 차지하는 비중이 3% 이상이 되도록 재구성하였다.

오버샘플링을 완료한 학습 데이터를 RoBERTa 모델에 학습시켜 선행연구 및 오버샘플링 수행 전 학습 데이터를 학습한 RoBERTa 모델의 결과들과 비교하였다. 하이퍼파라미터는 선행연구와 동일하게 3 epochs, batch size 32, learning rate 2e-5로 설정하였으며, 모든 수치는 10회 반복 실험을 통해 확인된 결과의 평균값이다.

3.3 성능평가

제안하는 침입탐지 모델의 성능평가 결과는 표 3에서 보는 바와 같다. 오버샘플링을 적용하지 않은 모델의 탐지 정확도는 86.88%로 나타났다. 그리고 오버샘플링을 적용한 데이터세트를 학습한 모델은 ADASYN을 적용한 모델이 87.15%로 가장 높은 정확도를 보여주었다.

표 3. 다중분류 결과
Table 3. Results of multi-class classification

	Proposed			
	RoBERTa	Random sampling	SMOTE	ADASYN
Accuracy	86.88	87.08	87.15	87.03

ADASYN으로 오버샘플링된 데이터세트를 학습한 RoBERTa 모델의 성능을 선행연구에서 언급한 동일한 데이터세트를 사용했던 연구들과 비교한 결과는 표 4에서 보는 바와 같다.

표 4. 다중분류 성능 비교 결과
Table 4. Performance comparison of multi-class classification

	RoBERTa ADASYN	[2]	[4]	[9]	[10]	[11]
Accuracy	87.15	85.63	86.63	75.77	86.04	77.51
F1-score	87.15	85.68	-	-	-	77.28

그림 2는 오버샘플링이 적용되지 않은 데이터세트를 학습한 RoBERTa 모델의 다중분류 혼동행렬을 보여주고 있으며, 그림 3은 ADASYN으로 오버샘플링된 데이터세트를 학습한 RoBERTa 모델의 혼동행렬을 보여주고 있다.

혼동행렬에서 주목할 점은 오버샘플링 적용 전과 후 오분류가 증가하는 클래스가 변경된다는 점이다. 그림 2에서 보는 바와 같이 오버샘플링이 적용되지 않은 데이터세트를 학습한 모델에서는 데이터 수가 많은 3번(Exploits) 및 4번(Fuzzers) 클래스로 오분류하는 경우가 상대적으로 많았다. 데이터 수가 적은 편에 속하는 0번(Analysis) 클래스로 오분류되는 경우는 발생하지 않았다. 그러나 오버샘플링이 적용된 데이터세트를 학습한 모델은 그림 3에서 보는 바와 같이 0번 클래스로의 오분류가 상당히 많이 발생하였다.

이러한 결과는 오버샘플링 적용하는 과정에서 희소 클래스 내 데이터의 특성을 완벽하게 반영한 데이터를 생성하지 못하여 발생했다고 볼 수 있다. 즉, 오버샘플링을 통해 데이터를 증가시키면 해당 클래스에 대한 탐지성능은 향상되지만, 해당 클래스로의 오분류도 함께 증가할 수 있다.

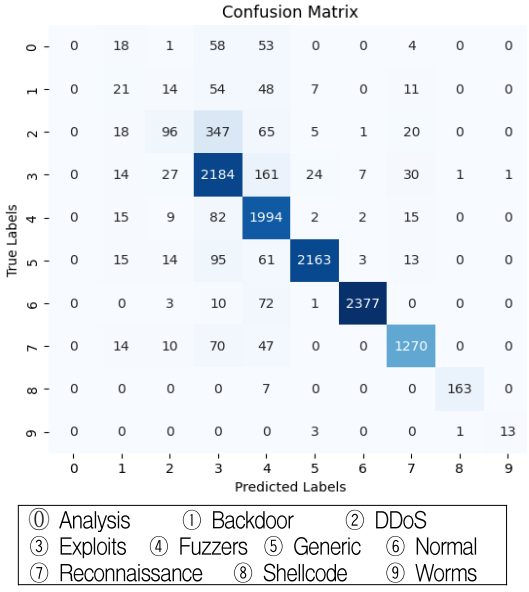

그림 2. RoBERTa 모델의 혼동행렬
Fig. 2. Confusion matrix of RoBERTa model

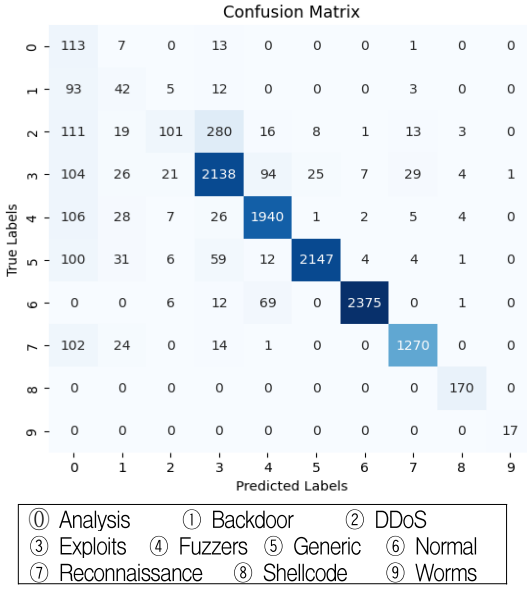

그림 3. ADASYN 방법의 혼동행렬
Fig. 3. Confusion matrix of ADASYN & RoBERTa model

IV. 결론 및 향후 과제

본 연구에서는 합성 데이터를 학습하는 인공지능 기반 침입탐지 모델의 한계를 극복하기 위해 패킷 페이로드를 직접 학습하는 트랜스포머 기반의 침입탐지 모델을 제안하였다. 그리고 희소 클래스 탐지 성능 저하 문제를 해결하기 위해 전체 데이터세트에서 차지하는 비중이 5% 미만인 희소 클래스에 대해 오버샘플링 기법을 추가로 적용하여 침입탐지 성능을 개선하고자 시도하였다.

제안 모델에 대한 성능평가는 UNSW-NB15 데이터세트를 활용하였으며, 학습 모델은 자연어 처리 성능이 개선된 트랜스포머 모델인 RoBERTa를 사용하였다. 먼저 UNSW-NB15 데이터세트에 포함된 PCAP 파일에 대해 Payload-Byte 기법으로 라벨링을 실시한 후 결측치와 중복을 제거하여 실험 데이터세트를 구성하였다. 이어서 8:2의 비율로 학습 데이터세트와 테스트 데이터세트로 분할하고, 학습 데이터세트에 대해서는 대표적인 오버샘플링 기법인 랜덤샘플링, ADASYN 및 SMOTE 3종을 적용하여 5개 희소 클래스에 대한 오버샘플링을 수행하였다. 실험 결과 ADASYN으로 오버샘플링을 수행한 데이터세트를 학습한 모델이 다중분류에서 가장 높은 정확도와 F1-score을 달성하였다.

한편 오버샘플링을 통해 희소 클래스의 데이터를 증가시키면 해당 클래스에 대한 다중분류 성능이 향상되어 전체적인 탐지 정확도가 향상되지만 오버샘플링이 적용된 클래스로의 오분류도 동시에 증가하는 현상이 확인되었다. 이는 오버샘플링을 통해 추가 생성한 데이터가 해당 클래스의 특성을 완벽하게 반영하지 못함을 의미한다.

따라서 향후 연구에서는 희소 클래스의 특성이 명확하게 반영되도록 오버샘플링을 정교하게 수행하는 방안을 모색하고, 보다 다양한 기법의 적용을 통해 트랜스포머 기반 모델에 최적화된 오버샘플링 기법을 찾아나갈 예정이다. 또한, 트랜스포머 기반 모델들의 자연어 처리 성능은 지속적으로 개선되고 있음을 고려하여 최신 모델들의 적용을 통해 패킷 페이로드 기반 침입탐지 모델의 성능을 개선해 나갈 것이다.

References

[1] S. Layeghy, M. Gallagher and M. Portmann, "Benchmarking the benchmark — Comparing synthetic and real-world Network IDS datasets", Journal of Information Security and Applications, Vol. 80, pp. 103689, Feb. 2024. https://doi.org/10.1016/j.jisa.2023.103689.

[2] G.-N. Kim, H.-S. Kim, and S.-J. Lee, "Intrusion Detection System based on Packet Payload Analysis using LightGBM", Journal of the Korea Society of Computer and Information, Vol. 28, No. 6, pp. 47-54, Jun. 2023. http://doi.org/10.9708/jksci.2023.28.06.047.

[3] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive dataset for network intrusion detection systems (UNSW-NB15 network data set)", Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, pp. 1-6, Nov. 2015. https://doi.org/10.1109/milcis.2015.7348942.

[4] W.-S. Park, G.-N. Kim, and S.-J. Lee, "Intrusion Detection System based on Packet Payload Analysis using Transformer", Journal of the Korea Society of Computer and Information, Vol. 28, No. 11, pp. 81-87, Nov. 2023. https://doi.org/10.9708/jksci.2023.28.11.081.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv:1907.11692, Jul. 2019. https://doi.org/10.48550/arXiv.1907.11692.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, Vol. 1, pp. 4171-4186, Jun. 2019. https://doi.org/10.18653/v1%2FN19-1423.

[7] Y. A. Farrukh, I. Khan, S. Wali, D. Bierbrauer, J. A. Pavlik, and N. D. Bastian, "Payload-Byte: A Tool for Extracting and Labeling Packet Capture Files of Modern Network Intrusion Detection Datasets", IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, Vancouver, WA, USA, pp. 58-67, Dec. 2022. https://doi.org/10.1109/BDCAT56447.2022.00015.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357, Jun. 2002. https://doi.org/10.1613/jair.953.

[9] D. Jing and H. B. Chen, "SVM Based Network Intrusion Detection for the UNSW-NB15 Dataset", IEEE International Conference on ASIC (ASICON), Chongqing, China, pp. 1-4, Oct. 2019. https://doi.org/10.1109/ASICON47005.2019.8983598.

[10] S. Meftah, T. Rachidi, and N. Assem, "Network Based Intrusion Detection Using the UNSW-NB15 Dataset", International Journal of Computing and Digital Systems, Vol. 8, No. 5, Sep. 2019. http://dx.doi.org/10.12785/ijcds/080505.

[11] S. M. Kasongo and Y. Sun, "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset", Journal of Big Data, Vol. 7, No. 105, Nov. 2020. https://doi.org/10.1186/s40537-020-00379-6.

[12] S. Bagui and K. Li, "Resampling imbalanced data for network intrusion detection datasets", Journal of Big Data, Vol. 8, No. 6, Jan. 2021. https://doi.org/10.1186/s40537-020-00390-x.

[13] H. A. Ahmed, A. Hameed, and N. Z. Bawany, "Network intrusion detection using oversampling technique and machine learning algorithms", PeerJ Computer Science, Vol. 8, Jan. 2022. https://doi.org/10.7717/peerj-cs.820.

[14] Z. Wu, H. Zhang, P. Wang and Z. Sun, "RTIDS: A Robust Transformer-Based Approach for Intrusion Detection System", IEEE Access,

Vol. 10, pp. 64375-64387, Jun. 2022. https://doi.org/10.1109/ACCESS.2022.3182333.

[15] Y-G. Yang, H-M. Fu, S. Gao, Y-H. Zhou, and W-M. Shi, "Intrusion detection: A model based on the improved vision transformer", Transactions on Emerging Telecommunications Technologies, Vol. 33,No. 9, Apr. 2022. https://doi.org/10.1002/ett.4522.

저자소개

이 태 화 (Taehwa Lee)

2015년 3월 : 육군사관학교 전자공학과(공학사)
2022년 8월 : 한국과학기술원 정보보호대학원(공학석사)
2023년 3월 ~ 현재 : 국방대학교 컴퓨터공학과 박사과정
관심분야 : 머신러닝, 정보보호, 컴퓨터 보안, 침입탐지시스템

이 수 진 (Soojin Lee)

1992년 3월 : 육군사관학교 전산학과(이학사)
1996년 2월 : 연세대학교 컴퓨터과학과(공학석사)
2006년 2월 : 한국과학기술원 전산학과(공학박사)
2006년 3월 ~ 현재 : 국방대학교 사이버·컴퓨터공학과 교수
관심분야 : 국방 사이버 보안 정책, 침입탐지시스템, 모바일 네트워크 보안, 머신러닝, 암호 이론 및 응용