J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

오토인코더 기반의 외부망 적대적 사이버 활동 징후 감지☆

Detection of Signs of Hostile Cyber Activity against External Networks based on Autoencoder

박 한 솔[1,2] Hansol Park | 김 국 진[1,2] Kookjin Kim | 정 재 영[1,2] Jaeyeong Jeong | 장 지 수[1,2] jisu Jang | 윤 재 필[1] Jaepil Youn | 신 동 규[1,2*] Dongkyoo Shin

요 약

전 세계적으로 사이버 공격은 계속 증가해 왔으며 그 피해는 정부 시설을 넘어 민간인들에게 영향을 미치고 있다. 이러한 문제로 사이버 이상징후를 조기에 식별하여 탐지할 수 있는 시스템 개발의 중요성이 강조되었다. 위와 같이, 사이버 이상징후를 효과적으로 식별하기 위해 BGP(Border Gateway Protocol) 데이터를 머신러닝 모델을 통해 학습하고, 이를 이상징후로 식별하는 여러 연구가 진행되었다. 그러나 BGP 데이터는 이상 데이터가 정상 데이터보다 적은 불균형 데이터(Imbalanced data)이다. 이는, 모델에 학습이 편향된 결과를 가지게 되어 결과에 대한 신뢰성을 감소시킨다. 또한, 실제 사이버 상황에서 보안 담당자들이 머신러닝의 정형적인 결과로 사이버 상황을 인식시킬 수 없는 한계도 존재한다. 따라서 본 논문에서는 전 세계 네트워크 기록을 보관하는 BGP(Border Gateway Protocol)를 조사하고, SMOTE(Synthetic Minority Over-sampling Technique) 활용해 불균형 데이터 문제를 해결한다. 그 후, 사이버 공방(Cyber Range) 상황을 가정하여, 오토인코더를 통해 사이버 이상징후 분류하고 분류된 데이터를 가시화한다. 머신러닝 모델인 오토인코더는 정상 데이터의 패턴을 학습시켜 이상 데이터를 분류하는 성능을 92.4%의 정확도를 도출했고 보조 지표도 90%의 성능을 보여 결과에 대한 신뢰성을 확보한다. 또한, 혼잡한 사이버 공간을 가시화하여 효율적으로 상황을 인식할 수 있기에 사이버 공격에 효과적으로 방어할 수 있다고 전망된다.

☞ 주제어 : 이상 탐지, 오토인코더, BGP Archive Data

ABSTRACT

Cyberattacks around the world continue to increase, and their damage extends beyond government facilities and affects civilians. These issues emphasized the importance of developing a system that can identify and detect cyber anomalies early. As above, in order to effectively identify cyber anomalies, several studies have been conducted to learn BGP (Border Gateway Protocol) data through a machine learning model and identify them as anomalies. However, BGP data is unbalanced data in which abnormal data is less than normal data. This causes the model to have a learning biased result, reducing the reliability of the result. In addition, there is a limit in that security personnel cannot recognize the cyber situation as a typical result of machine learning in an actual cyber situation. Therefore, in this paper, we investigate BGP (Border Gateway Protocol) that keeps network records around the world and solve the problem of unbalanced data by using SMOTE. After that, assuming a cyber range situation, an autoencoder classifies cyber anomalies and visualizes the classified data. By learning the pattern of normal data, the performance of classifying abnormal data with 92.4% accuracy was derived, and the auxiliary index also showed 90% performance, ensuring reliability of the results. In addition, it is expected to be able to effectively defend against cyber attacks because it is possible to effectively recognize the situation by visualizing the congested cyber space.

☞ keyword : Anomaly Detection, AutoEncoder, BGP Archive Data

1. 서 론

전 세계적으로 사이버 공격은 계속 증가해 왔으며, 외화벌이, 군사 및 외교 기밀 유출 등을 목적으로, 정부 시설을 넘어 민간 기업과 병원까지 사이버 공격이 감행되었다. 위와 같이, 사이버 공격의 피해는 민간인들에게까지 미치고 있으며 그 피해는 계속 증가하고 있다 [1]. 이에, 미국 국방부(The United States Department of Defense)

1 Department of Computer Engineering, Sejong University, Seoul, 05006, Korea.
2 Department of Convergence Engineering for Intelligent Drones, Sejong University, Seoul, 05006, Korea.
* Corresponding author (shindk@sejong.ac.kr)
[Received 30 August 2022, Reviewed 17 September 2022(R2 11 October 2022), Accepted 17 October 2022]
☆ 본 연구는 2020년 국방과학연구소에서 주관하는 미래도전국방기술 연구개발사업(9129156)의 지원을 받아 수행되었습니다.

는 사이버 공격의 심각성을 인식하며, 사이버 공간에서 이루어지는 공격을 사이버전이라 부르며 사이버보안의 중요성을 강조하고 있다[2]. 이처럼 사이버 공격으로부터 민간인들의 피해를 최소화하고, 보안 담당자가 사이버 상황을 빠르게 인식 및 대처하는 것이 중요해졌다. 그러나, 현실적으로 혼잡한 사이버 상황 속에서 능숙하게 사이버 공격을 방어하기 어려우며 사이버 이상 징후를 식별하기 쉽지 않다. 따라서 본 논문에서는, 사이버 이상 징후를 탐지하기 위해 적은 양의 데이터로도 데이터의 특성을 파악할 수 있는 오토인코더를 활용하여 이상 징후를 식별한다. 그 후, 보안 담당자가 사이버 상황을 인식할 수 있도록 이상 데이터들을 가시화한다. 본 연구는 사이버 공간에서의 공격을 시각화하기 위한 연구이다. 이를 위해서 2장부터 5장까지 요약하면 아래와 같다.

2장은 라우팅 교환 프로콜인 BGP 데이터를 사이버 공간 및 이상 징후를 식별한 연구 사례를 조사한다. 그 후 이상 징후 데이터들의 공통적인 문제인 데이터 불균형 문제의 해결 방안을 모색한 후, 오토인코더로 이상 데이터를 분류하는 선행 연구를 조사한다. 마지막으로, 실제 사이버 공간을 시각화 사이버 공방 훈련(Cyber Range)을 조사한다. 3장에서는 2장에서 조사된 연구 기법들을 가지고 사이버 이상 데이터를 분류하기 위한 데이터 전처리 과정을 설명하고 오토인코더 모델의 학습 성능을 높이는 방법을 제안한다. 4장은 오토인코더로 분류 성능을 평가하기 위한 지표를 설명하고 실험 결과를 요약한다. 마지막으로 5장은 분류된 이상 데이터들을 시각화하고 6장은 본 논문의 결론을 도출한다.

2. 관련 연구

2.1 Border Gateway Protocol

BGP는 라우팅 정보를 교환하기 위한 프로토콜로, IP prefix 연결 정보이며 전 세계 게이트웨이 호스트의 기반이 되는 프로토콜이다. 해당 데이터로 데이터 패킷을 교환할 시, AS(Autonomous System) 번호가 기록되며, 이를 통해 데이터 패킷이 전송된 라우터 경로, IP, 국가를 유추할 수 있다. BGP에 수집된 데이터로는 표 1과 같이 다양한 Feature들을 얻을 수 있으며, 해당 Feature는 시간 정보, 근원지 ip 주소, AS의 번호, 해당 패킷이 지나쳐 온 라우팅 경로, 라우터의 위치 정보, 라우터 소유 국가, 네트워크 트래픽 등이 있다. BGP 데이터를 활용한 연구로는 Youn jaepil, et al. [3] 혼잡한 사이버 상황을 시각적으로 파악하기 위해, BGP 데이터을 수집한 후, 해당 데이터로 사이버 전장을 Di-Grahp(Directed graph) 형태로 구현하여 라우팅 경로를 시각화하는 방법을 제안했다.

Min Cheng, et al. [4]은 BGP에서 수집된 트래픽 데이터를 Change-Point Detection 방법을 사용하여 트래픽 변화량을 탐지하고 Multi-Scale LSTM을 사용하여 이상 징후를 식별하는 방법을 제안했다.

Biersack, et al. [5]은 BGP Archive Data를 활용하여 네트워크 시각화를 통해 라우팅 패턴을 분류하는 VIS-SENS 모델을 제안했다.

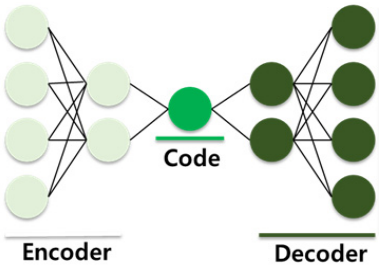
선행 연구들을 통해, BGP 데이터를 활용하여 사이버 공간과 라우팅 패턴을 시각화하며, 이상징후를 식별할 수 있다. 그러나, 기존의 방법으로는 사이버 공격을 조기에 탐지할 수 없다. 따라서, 악의적인 공격자가 평소 라우팅 경로 패턴을 벗어나며, 의심스러운 상황으로 판단하여 사이버 이상징후로 분류하도록 한다.

(표 1) BGP Archive Data
(Table 1) BGP Archive Data

데이터 명	설명
timestamp	시간 데이터
ip	라우터가 가지고 있는 ip 주소
asn	시작 AS 번호
path	AS의 경로
latitude	해당 AS의 위도 정보
longitude	해당 AS의 경도 정보
country	시작 AS(의 소유 국가
count	네트워크 트래픽 횟수

2.2 오토인코더(AutoEncoder)

오토인코더은 비지도 학습 알고리즘 중 하나로, 그림 1와 같이 인코더(Encoder), 코드(Code), 디코더(Decoder)로 구성되어 있으며 대칭 구조를 가진다[6]. 오토인코더는 데이터가 입력되면 인코더 부분에서 데이터의 핵심 정보들을 최대한 학습하고 나머지 데이터들은 손실시키며 데이터를 압축한다. 그 후, 압축된 데이터를 디코더 부분에서 입력된 데이터 형태로 다시 복원하는데 이를 재구성(Reconstruction)이라 하고 학습이 된 데이터라면 자연스럽게 복원이 되지만 학습이 되지 않은 데이터는 복원되지 않는다. 달리 말하면, 학습을 완료한 오토인코더는 정상 데이터를 그대로 복원하지만, 이상 데이터는 손실된다. 따라서, 복원되지 않은 이상 데이터를 추출하여 시각화하고, 사이버 상황인식 및 판단에 이바지할 수 있도록 한다.

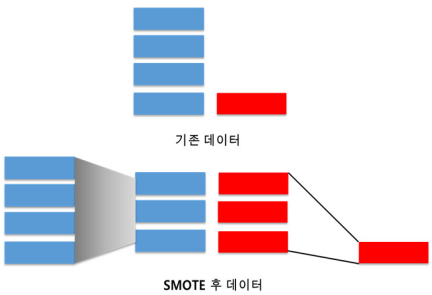


(그림 1) 오토인코더 구조
(Figure 1) Structure of AutoEncoder

2.3 데이터 불균형 문제

데이터 불균형 문제란 정상 데이터 수와 이상 데이터 수의 비율이 현저히 차이가 나는 데이터를 말한다. 데이터의 비율이 불균형하다면 머신러닝 모델을 학습할 때 정상 데이터를 이상 데이터로 판단하여 분류하는 문제점이 발생하게 된다. 데이터 불균형 문제를 해결하기 위해 다양한 연구가 활발히 이루어지고 있다. 대표적으로 다수의 정상 데이터를 이상 데이터의 수에 맞게 데이터를 감소시키는 undersampling과 소수의 이상 데이터를 정상 데이터 수만큼 증가시키는 oversampling 기법이 있다[7-8].

undersampling의 경우 정상 데이터가 감소하여 정보 손실이 발생한다. 이는 정상 패턴을 충분히 학습할 수 없어 새로운 패턴의 정상 데이터가 유입되면 해당 데이터를 이상 데이터로 구분하게 된다. 따라서, 본 연구에서는 그림 2와 같이 SMOTE(Synthetic Minority Over-sampling Technique) 기법이란 정상 데이터의 감소를 최소화하며 이상 데이터의 양을 증가시키는 기법이다. 본 연구에서는, SMOTE 기법을 활용해 정상 데이터를 양을 최소한으로 감소시켜, 해당 정보를 최대한 보존할 것이고 이상 데이터를 증가시켜 학습의 편향을 막으며 머신러닝 모델의 학습 효율을 높인다.



(그림 2) SMOTE 데이터 셋
(Figure 2) SMOTE data set

2.4 사이버 공방훈련

사이버 공방 훈련(Cyber Range)이란 사이버 공간 내에서 발생할 수 있는 다양한 공격 시나리오를 만들어 사이버 공간 내에서 공격과 방어를 수행하는 교육 및 훈련이다. 사이버 공방 훈련은 효과적으로 하기 위해 여러 연구가 진행되고 있다. Smyrlis 등 [9]은 훈련에서 사이버 공간을 자산, 모니터링, 침투, 취약성 파악 등, 4가지 요소를 파악하는 것이 중요하다고 설명한다. 또한, 사이버 공간에서 아군의 취약성과 자산을 파악하여 모니터를 시각화하는 것이 방어에 효과적이라 설명한다.

Gustafsson 등 [10]은 사이버 공방훈련을 위해 무작위로 팀을 2개로 나눈 후, VPN(Virtual Private Network)을 이용해 공방이 이루어진다. 훈련이 진행되는 동안 이 과정을 시각화한다. 사이버 공격을 보다 직관적으로 해당 훈련생들이 사이버 공격을 직관적으로 이해할 수 있게 된다고 설명한다.

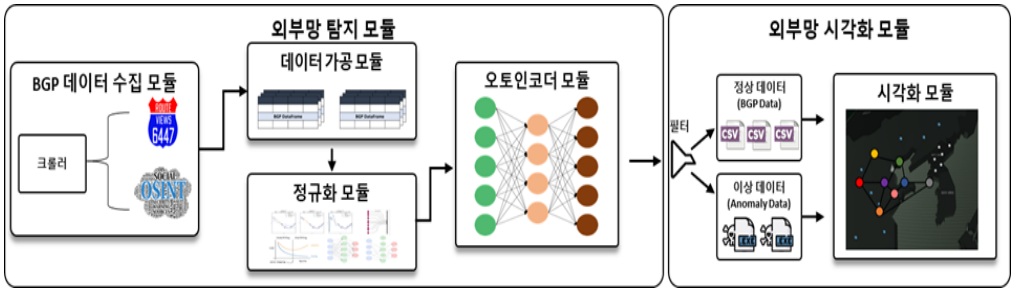
이를 통해, 실제 사이버 공방훈련에서 사이버 공간을 시각화하는 것이 매우 중요하며, 사이버 상황 인식에 큰 도움이 된다는 것을 알 수 있다.

3. 연구 방법

3.1 사이버 이상 데이터 가시화 설계 및 구현

본 논문은 방어자와 악의적인 공격자를 전 세계 국가 중 무작위로 선정한다. 임의로 선정된 국가의 BGP 데이터를 수집한 뒤, 오토인코더로 데이터의 패턴을 학습시키는 이상징후를 분류하는 모델을 제안한다. 그림 3은 사이버 이상징후를 효율적으로 식별하기 위해 설계된 프로세스이다. 해당 프로세스는 크게 외부망 탐지 모듈과 외부망 시각화 모듈로 나누어져 있으며, 외부망 탐지 모듈은 데이터 수집, 가공, 정규화, 오토인코더 모듈로 구성되어 있으며 사이버 이상징후를 식별하는 역할을 맡는다. 외부망 시각화 모듈은 오토인코더를 통해 분류된 결과를 정상 및 이상 데이터로 분류하여 데이터베이스에 저장한다. 그 후, 시각화 모듈을 통해 악의적인 사용자의 접근 경로를 시각화한다.

BGP 데이터 수집 모듈에서는 크롤러를 사용하여 Oragon 공식 홈페이지에 있는 BGP 데이터를 수집한다. 해당 데이터는 대부분 문자형 데이터이기 때문에 머신러닝 모델로 학습시킬 수 없다. 더불어 이상징후와 데이터 간의 상관관계 또한 알 수 없다. 따라서 데이터 가공 모듈에서 머신러닝 모델을 학습시킬 수 있는 학습 데이터로



(그림 3) 사이버 이상징후 식별 구조 및 프로세스
(Figure 3) Cyber Anomalies Identification Structure and Process

가공해야 한다.

데이터 가공 모듈은 앞서 언급한 BGP 데이터의 path 컬럼을 데이터를 정수 및 소수형 데이터로 전환한다. 해당 데이터는 AS의 경로를 기록한 문자형 데이터이다. 모델의 원활한 학습을 위해 소수 및 정수형 데이터로 전환한다. 그 후, AS의 경로를 추출하여 앞서 언급한 악의적인 공격자들이 방어자의 AS에 접근한다면 이상징후로 식별하도록 데이터 전처리 과정을 밟는다. 또한, 불균형 데이터 문제를 해결하고 학습의 편향을 막기 위하여, SMOTE 기법을 사용한다. 그 후, 네트워크 트래픽 및 네트워크 경로까지 추가되어 오토인코더 내부로 들어가게 된다.

정규화 모듈에서는 데이터들의 과도한 편향과 분산을 막기 위해 정규화(Normalization) 과정을 걸치게 된다. 만약 데이터들이 편향이나 분산되어 있다면 새로운 BGP 데이터가 유입될 시 이상 데이터를 식별하지 못할 경우가 발생하기 때문에 이를 방지하기 위해 정규화를 진행한다. 정규화를 진행하게 되면 데이터들은 0과 1 사이의 값을 가지게 된다.

오토인코더 모듈에서는 오토인코더가 전처리 된, BGP 데이터로 학습을 진행한다. 오토인코더의 학습방식은 데이터들의 주요 패턴들을 추출하여 압축한 후, 다시 압축된 데이터를 원래 형태로 복원하는 방식으로 학습을 진행한다. 학습을 진행할 때 방어자의 AS 경로를 가지고 진행한다. 여기서 방어자에 정상 AS 경로들을 학습이 되어 압축된 후 다시 복원되지만, 이상 AS 경로들은 복원되지 않는다. 복원되지 않은 이상 데이터들은 메모리에 모아서 시각화 모듈로 전송되게 된다.

마지막으로 시각화 모듈에서는 필터링 된 이상 데이터와 정상 데이터를 가지고 사이버 이상 징후와 특정 국가의 관계를 그래프 형식으로 시각화한다.

3.2 데이터 소개

3.2.1 사이버 정상 데이터의 정의

Chalapathy 및 D Gong은 [11-12]은 이상 데이터랑 데이터 셋에 존재하는 다수의 패턴 중 나타난 돌연변이 패턴 즉, 소수의 패턴을 이상 데이터로 정의 내렸다. 달리 말하면, 다수의 패턴에 속하는 데이터들은 정상 데이터로 정의 내릴 수 있다. 본 연구에서 사이버 공방이 이루어지고 있다는 상황을 가정하여 전 세계 국가 중 무작위로 악의적인 공격자와 방어자를 선정한다. 그 후, 해당 공격자와 방어자의 BGP 데이터를 크롤러를 통해 수집한다. 수집된 데이터를 토큰화(Tokenizer)을 이용해 변환한 뒤, AS 경로를 분석한다. 토큰화란 자연어 처리에서 많이 사용되는 기법으로 예를 들어, “경찰관이 인사한다”를 ‘경찰관’, ‘인사한다’로 나누고 반복적으로 쓰인 단어의 횟수를 기록하는 기법이다. 본 연구에서는 표 2와 같이 AS의 경로들은 하나의 단어로 취급하여 가장 빈번하게 사용된 경로부터 순차적으로 숫자를 부여한다. 위에 언급한 대로 정상 데이터는 다수의 패턴에 속해야 하므로 반복된 횟수가 50회 이상인 데이터들을 정상 데이터로 분류하여 총 123개의 패턴 중 118개를 정상 데이터로 분류되고 총 1만 개의 데이터를 가진다.

(표 2) 사이버 공간 정상 데이터
(Table 2) Cyberspace Normal Data

토큰 번호	AS(autonomous system) 경로	반복 횟수
1	‘7018 4837 134544 9957‘	326
2	‘701 4837 134544 4677’	234
3	’3257 4837 134544 9957’	197
4	’6939 4837 134544 4677”	122

3.2.2 사이버 이상 데이터

이상 데이터란 앞서 3.2.1 절에서 언급한 대로, 다수의 데이터 중에서 극소수의 데이터를 의미한다. 앞서 언급한 대로 사이버 공방을 가정하에 전 세계 국가 중 무작위로 선정된 국가를 악의적인 공격자로 선정하였다. 악의적인 공격자의 시나리오 및 공격 배경은 다음과 같다.

- 악의적인 공격자 조직 특성: 악의적인 공격자의 모든 사이버 활동은 개인이 아니라 특정한 1인에 의해 통제된다 [13].
- 악의적인 공격자의 목표: 악의적인 사용자의 공격 목적은 외화벌이이며, 그 목표는 국가 시설, 외교 문서 탈취, 민간 기업 기밀 노출 등이 있다.
- 공격 능력: 악의적인 사용자는 디도스, 랜섬웨어 등 다양한 공격 능력을 갖추고 있으며 평소 인터넷 및 라우터는 비활성화 상태이며 공격할 때 활성화된다.

시나리오를 바탕으로 무작위로 선정된 악의적인 사용자의 AS 번호는 표 3과 같다. 악의적인 사용자들의 BGP 데이터도 토큰화를 진행한다. 표 4는 악의적인 사용자가 방어자의 라우터에 접근한 횟수를 기록한 것이다. 본 논문에서는 악의적인 공격자가 대한민국 라우터에 접근하면 이상 데이터로 분류한다. 또한, 앞서 언급한 다수의 데이터 중 소수의 라우팅 패턴을 보여주기 때문에 이상 데이터의 조건으로도 적합하다. 해당 데이터는 총 1만 개의 정상 데이터 중에서 25개의 이상 데이터가 포함되어 있다. 그러나 이상 데이터는 불균형 데이터 문제가 발생하므로 SMOTE 기법을 활용해 정상 데이터와 이상 데이터의 비율이 같도록 전처리했다.

(표 3) 이상 ASN(autonomous system number)
(Table 3) Anomaly Autonomous System Number

토큰 번호	소유자	AS(autonomous system)
1	공격자 A	46288, 6484
2	공격자 B	10212, 131285, 131325, 132154
3	공격자 C	131279
4	공격자 D	1684

(표 4) 사이버 공간 이상 데이터
(Table 4) Cyberspace Abnormal Data

토큰 번호	AS(autonomous system) 경로	접근 횟수
1	'7018 131279 134544 9957'	12
2	'3303 6939 4837 6468 131285 4677'	6
3	'131285 20130 4677'	3
4	'46288 2497 53767 20130 6939 31325 9957'	2
5	'31019 18106 7660 4677'	2

3.2.3 데이터 전처리

사이버 이상 데이터를 식별하기 위한 모델로 오토인코더를 사용한다. 그러나 BGP 데이터는 문자형 데이터가 포함되어 있는데 이는 정수형과 실수형 데이터만으로 학습할 수 있는 머신러닝 모델에 적합하지 않다는 것을 의미한다. 따라서 모델이 학습할 수 있는 정수형과 실수형으로 변환하는 전처리 과정이 필요하다.

BGP 데이터의 문자형 데이터로는 앞서 표 2의 path 데이터가 있다. path 데이터는 토큰화를 이용하여 반복 횟수가 많은 순서대로 숫자를 부여한다. 따라서 전처리 된 패턴들의 숫자는 앞서 3.2.1에서 언급한 대로 1부터 118까지의 숫자로 변환되고 이상 데이터는 119부터 123으로 변화된다. 그 후, 날짜, 시간, 위치 정보, 토큰화된 AS 경로, 트래픽 발생 횟수를 범주형 데이터로 변환하기 위해 원-핫(One-hot)을 사용하였으며, 여기서 토큰화된 path 데이터는 정규화하여 0과 1 사이의 값으로 변형했다. 따라서 모델의 입력되는 Feature의 수는 32이다. 사이버 이상 데이터도 오토인코더를 통해 학습하기 위하여 원-핫 인코딩으로 전처리 과정을 걸친다. 그러나 해당 데이터는 오토인코더가 이상 징후로 식별하기 위해, 라벨(label)이 필요하다. 라벨(label)의 종류는 정상, 주의, 이상 대한 값으로 변환된 형태는 표 6과 같다

.

(표 5) 사이버 공간 정상 데이터 전처리 결과
(Table 5) Cyberspace Normal Data Preprocessing Results

구분	전처리 결과
1	0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0.03728581
2	0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0.0262378581
3	0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0.00939728581

(표 6) 사이버 공간 정상 데이터 전처리 결과
(Table 6) Cyberspace Abnormal Data Peprocessing Results

구분	전처리 결과
1	0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0.03728581 0 0 1
2	0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0.0262378581 0 0 1
3	0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0.00939728581 0 0 1

3.3 딥러닝 모델

3.2절에서 BGP 데이터를 딥러닝 모델에 학습시키기 위해서 전처리하는 과정을 다루었다. 표 5와 표 6처럼 가공이 완료된 데이터는 오토인코더 모델에 넣어 학습을 진행할 수 있다.

앞서 설명했던 것처럼 사이버 이상징후를 식별하기 위해서 오토인코더 모델로 학습을 완료한 후, 시각화 모듈로 이동하게 된다. 오토인코더의 학습과정을 설명하면, AS 경로 데이터가 인코더로 입력된다. 입력된 데이터의 주요 특징을 추출하면서 코드 영역까지 압축을 진행한다. 그 후, 디코더 부분에서 압축된 데이터를 다시 재구성하면서 데이터를 복원한 후, 처음 입력 형태와 복원된 데이터의 비교한다. 만약 이상 데이터라면 원래 형태로 복원되지 않는다. 본 연구에서는 오토인토더의 학습방식을 이용하여, 정상 데이터로만 학습을 진행한다. 이는 오토인코더의 학습 검증 과정에서 정상 데이터만을 복원시키고 이상 데이터를 복원하지 않기 위해서이다. 학습 검증이 끝난 후, 복원되지 않은 이상 데이터를 시각화하여 AS 경로를 화면에 나타내게 된다. 학습의 사용된 파라미터는 표 7과 같다.

(표 7) 실험에 사용된 파라미터
(Table 7) Experiment Parameters for Train

구분	오토인코더-학습 파라미터
Model Layer	32(input, output) 32 - 16 - ... - code (encoder) code - ... - 16 - 32 (decoder)
Activation Funcition	Relu (hidden) Relu (output)
Threshold (이상치 기준 값)	4.7
Batch size / Epoch	64/100
Optimizer / Learning rate	Adam / 0.001
Loss Function	Mean Squared Error

4. 실험 결과

4.1 성능 평가 지표

본 연구에서는 모델의 성능이 신뢰할 수 있는지 검증하기 위해서 성능 지표를 사용한다. 대표적인 성능 지표로 혼동 행렬(Confusion Matrix)과 보조 지표인, 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 스코어(F1-Score)가 있다. 혼동 행렬이란 모델의 성능을 평가할 때 예측값이 실제 값을 비교하여 얼마나 정확히 예측했는지를 보여주는 행렬이다. 혼동 행렬의 주요 요소로는 표 8과 같다.

혼동 행렬 요소 TP와 TN은 실제값을 맞게 예측한 것이며, FP와 FN은 실제값과 다르게 예측한 부분을 의미한다. 정확도는 식 1과 같으며, 모델이 얼마나 정확하게 예측하는지 나타내는 지표로 높을수록 좋은 수치이며 얼마나 정확하게 사이버 정상 데이터와 이상 데이터를 구분했는지를 의미한다. 실험에서 BGP 데이터는 정상 데이터와 이상 데이터의 비율이 다른 불균형 데이터이기 때문에 왜곡된 결과가 나타날 수 있다. 따라서 모델의 성능을 평가할 때, 해당 실험의 결과의 신뢰성을 검증하기 위하여 정확도를 포함하여 정밀도와 재현율 그리고 F1-Score를 보조 지표로 사용한다.

(표 8) 혼동 행렬
(Table 8) Confusion Matrix

	예상 값(예)	예상 값(아니오)
실제 값(예)	TP	FN
실제 값(아니오)	FP	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad (2)$$

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

$$F1\,Score = 2 * \frac{\mathrm{Precision} + Recall}{\mathrm{Precision} + Recall} \quad (4)$$

정밀도는 모델이 사이버 정상 데이터라고 분류한 것 중에서 정상 데이터로 분류한 비율을 말하며 식 2와 같다. 재현율은 실제 사이버 정상 데이터 중에서 모델이 사정상 데이터라 판단하여 분류한 데이터 비율을 의미하며 식 3과 같다. F1 스코어는 정밀도와 재현율의 조화평균이며 데이터 비율이 불균형할 때 모델의 성능을 정확하게 평가해 주는 지표이다. 해당 식 4와 같다.

4.2 실험 환경

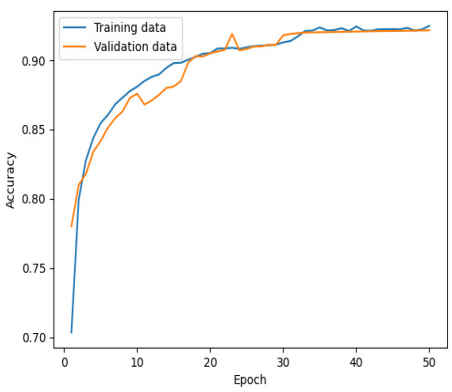
본 연구에서는 오토인코더를 활용하여 사이버 이상징후 데이터를 분류하는 실험은 표 9의 환경에서 진행했다.

(표 9) 실험 환경
(Table 9) Experimental Environment

구분	하드웨어 및 소프트웨어
운영체제	Windows 10 Pro
메모리	64GB
사용 언어	Python 3.10.1
그래픽카드	RTX 2060
라이브러리	tensorflow, keras, pandas, numpy, matplotlib

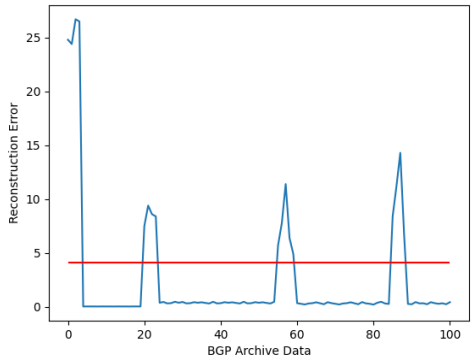
4.3 실험 결과

앞 장에서 언급하였듯이 사이버 이상 데이터 분류하기 위해 오토인코더를 사용하였다. 해당 모델의 성능은 압축된 데이터를 다시 복원했을 때 정상 데이터를 원래의 형태로 복원하는 것과 이상 데이터를 복원하지 못 하는 것으로 정확도의 성능을 평가할 수 있다. 그림 4는 오토인코더의 정확도를 나타낸 그래프이다. 해당 실험 결과 오토인코더는 약 92.4%의 근접하는 결과를 보여준다. 사이버 이상 데이터의 경우 정상 데이터가 포함되었기 때문에 그림 4의 그래프로 구분하기 어렵다. 따라서 그림 5는 이상 데이터의 재구축하여 과정을 그래프로 나타낸 것이다. y축은 재구축 값을 의미하며, 만약 빨간 선은 이상치 기준으로 선을 넘길 시, 이는 데이터가 복원되지 않았다는 것을 의미한다. 즉, 사이버 이상 데이터로 간주하여 분류한다.



(그림 4) 실험 결과
(Figure 4) Experimental results

오토인코더의 성능을 평가하기 위해 다른 모델들과 성능을 비교하였다. 실험은 같은 BGP 데이터를 사용하였으며, 같은 전처리 과정을 걸쳐 실험을 진행하였다. 결과는 표 10과 같다. 표 10은 보며 알 수 있듯이 다른 모델들을 평균 80%대의 성능을 보인다. 오토인코더는 가장 좋은 성능을 보이며 모든 성능에서 90%가 넘는 것을 보여준다.

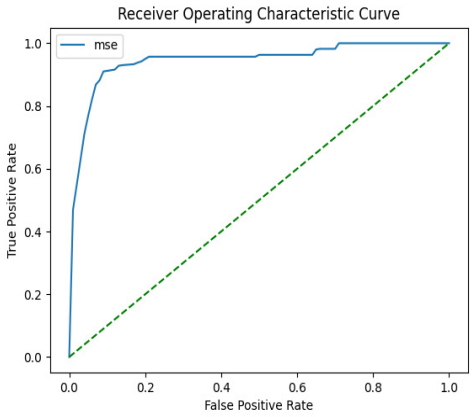


(그림 5) 사이버 이상 데이터의 손실 그래프
(Figure 5) Loss of Cyber Attack Data Graph

(표 10) 머신러닝 모델 성능 비교
(Table 10) Machine Learning Model Performance Comparison

Mode	Accuracy	Precision	Recall	F1-Score
Random Forest	0.82	0.8	0.81	0.8
Logistic Regression	0.84	0.84	0.83	0.83
SVM	0.85	0.81	0.81	0.8
One-class SVM	0.89	0.89	0.87	0.87
K-Neighbors	0.87	0.88	0.87	0.87
AutoEncoder	0.924	0.92	0.91	0.91

그림 6은 수신자 조작 특성 곡선(Receiver Operator Characteristic Curve; ROC Curve)이다. 수신자 조작 특성 곡선은 데이터 분류 모델을 평가하는 기법으로 자주 사용되며, 성능이 좋으면 좋을수록 외측 상단에 곡선 모양으로 나타난다. 정확한 평가를 위해 AUC(Area Under Curve)값을 사용하는데 이는 그림 5의 파란 선 아래에 면적을 의미하며, 그 면적을 AUROC라고 부른다. AUROC는 최대값으로 1을 가지며 본 실험에서는 오토인코더 모델의 AUROC 값으로 0.931의 값을 가지며, 이는 모델이 좋은 분류 성능을 가진 것을 의미한다.



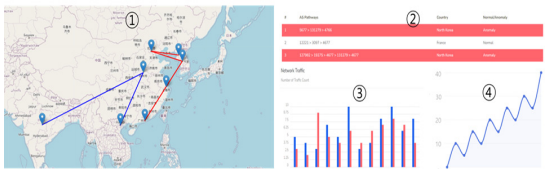
(그림 6) 수신자 조작 특성 곡선
(Figure 6) Receiver Operator Characteristic Curve; ROC Curve)

5. 사이버 이상징후 시각화

그림 7은 오토인코더로 구분된 BGP 데이터의 라우팅 경로를 시각화한 그림이다. 1번 화면은 사이버 이상 데이터의 라우팅 경로를 빨간색 선으로 정상 데이터의 라우팅 경로를 파란색 선으로 나타낸 것이다. 2번 화면은 1번 화면의 라우팅 경로의 정보를 나타낸 것으로 화면에 표시한 정보는 라우팅 경로, 해당 라우터의 소유자 정보, 사이버 이상징후의 유무를 식별 여부를 나타낸다. 만약 이상징후라면 빨간색 배경으로 강조하고, 이상징후가 아니라면 하얀색 배경으로 나타낸다. 3번 화면은 네트워크 트래픽을 비교한 것으로 파란색 선을 전월의 네트워크 트래픽을 빨간색은 현재 네트워크 트래픽을 나타낸다. 라우팅 경로뿐만 아니라 네트워크 트래픽의 변화로도, 보안 담당자가 이상징후를 효과적으로 식별할 수 있다. 4번 화면은 악의적인 사용자가 해당 라우터의 접근 증가량을 그래프로 표현한 것이다. 이 화면을 통해 보안 담당자는 악의적인 공격자의 접근 경로와 AS 소유자, 네트워크 트래픽, 접근 증가량은 직관적으로 인식할 수 있다.

6. 결 론

본 연구에서는 사이버 공간의 이상징후를 조기에 감지하여 관리자들의 상황 인식을 돕는 화면을 설계 및 구현했다. 라우팅 프로토콜 데이터인 BGP 데이터를 사용하였으며, 사이버 이상징후의 분류 성능을 높이기 위해 데이터를 토큰화를 활용하여 전처리하였다. 전처리 된 데이터를 오토인코더로 학습시킨 후 실험을 통해 분류된 결과로 92.4%를 보여주었다. 추가적으로 정밀도, 재현율, F1 스코어 등의 보조 지표들도 90%를 넘는 성능을 보이며 과적합을 예방할 수 있는 우수한 모델임을 알 수 있다. 또한, 기존의 이상징후 탐지 방식인 트래픽 및 패킷 분석이 아닌, 악의적인 사용자의 네트워크 공격 경로를 분석하는 새로운 방법론을 사용하였다. 이를 통해 사이버 공간에서 정상 데이터와 이상 데이터를 시각화하여 보안 담당자들이 사이버 상황을 보다 효과적으로 판단할 수 있다. 추후 에는 본 연구에서 사용한 오토인코더 모델을 더 발전시켜 개선된 사이버 성능의 모델로 가시화 시스템을 만들 예정이다. 또한, 사이버 이상 활동은 관리자들에게 소리나 경보음으로 이상을 알려주고 모니터링 화면에서 이벤트 기능을 추가하여, 효과적인 사이버 상황 인식 시스템을 만들 예정이다.



(그림 7) BGP Path 가시화 화면
(Figure 7) BGP Path Visualization Screen

참고문헌(Reference)

[1] Jakub Przetacznik, Russia's war on Ukraine: Timeline of cyber-attacks, 2022.

[2] S. dageet, "Quadrennial defense review report," Department of Defense., Virginia, USA, Feb. 2010.

[3] J. Youn. "Research on Cyber IPB Visualization Method based on BGP Archive Data for Cyber Situation Awareness," KSII Transactions on Internet and Information Systems(TIIS), 15(2), 749-766, 2021. https://doi.org/10.3837/tiis.2021.02.020

[4] M. Cheng and Q. XU, "MS-LSTM: A multi-scale LSTM model for BGP anomaly detection," 2016 IEEE 24th International Conference on Network Protocols (ICNP), 2016.

[5] P.A. Veriver, "Visual analytics for BGP monitoring and prefix hijacking identification," in IEEE Network, vol.

26, no. 6, pp. 33-39, November-December 2012. https://doi.org/10.1109/MNET.2012.6375891.

[6] P. Baldi, “Autoencoders, unsupervised learning, and deep architectures,” In Proceedings of ICML workshop on unsupervised and transfer learning, pp. 37-49, 2012. https://dl.acm.org/citation.cfm?id=3045796.3045801

[7] B.W. Yap, ″An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets.″ Proceedings of the first international conference on advanced data and information engineering (DaEng-2013). Springer, Singapore, December 2014. https://doi.org/10.1007/978-981-4585-18-7_2

[8] A. Fernández, ″SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary,″ Journal of artificial intelligence research 61, Apr 2018. https://doi.org/10.1613/jair.1.11192

[9] M. Smyrlis, ″CYRA: A model-driven CYber Range Assurance platform,″ Applied Sciences(MDPI), May 2021. https://doi.org/10.3390/app11115165

[10] T. Gustafsson and J. Almroth. ″Cyber range automation overview with a case study of CRATE.″ Nordic Conference on Secure IT Systems. Springer, Cham, March 2021.

[11] D. Freet and R. Agrawal, ″A virtual machine platform and methodology for network data analysis with IDS and security visualization.″ SoutheastCon 2017, pp. 1-8, 2017. https://doi.org/10.1109/SECON.2017.7925300.

[12] Kim, M., “North Korea’s cyber capabilities and their implications for international security,“ Sustainability, 14(3), 1744, February 2022. https://doi.org/10.3390/su14031744

[13] E. Chanlett-Avery, “North Korean Cyber Capabilities: In Brief,” Congressional Research Service, pp. 1-12, Washington, DC, USA, August 2017.

◐ 저 자 소 개 ◑

박 한 솔(Han-sol Park)
2021년 숭실대학교 컴퓨터공학과(학사)
2021년~현재 세종대학교 대학원 컴퓨터공학과(석사과정)
관심분야 : 사이버전, 이상탐지, 인공지능, etc.
E-mail : miro9303@sju.ac.kr

김 국 진(Kook-jin Kim)
2017년 서울호서전문학교 정보보호학과(학사)
2019년 (주)엠투소프트 전자문서사업부 주임
2019년~현재 세종대학교 대학원 컴퓨터공학과(석박사통합과정)
관심분야 : 사이버전, 사이버 지휘통제, 정보보호, 인공지능, etc.
E-mail : kjkim@sju.ac.kr

정 재 영(Jae-yeong Jeong)
2021년 숭실대학교 컴퓨터공학과(학사)
2021년~현재 세종대학교 대학원 컴퓨터공학(석사과정)
관심분야 : 사이버전, 네트워크, 정보보안, 인공지능, etc.
E-mail : jaeyeong@sju.ac.kr

장 지 수(Ji-su Jang)
2021년 호서전문학교 정보보호학과(학사)
2021년~현재 세종대학교 대학원 컴퓨터공학 지능형드론융합학과(공학석사)
관심분야 : 사이버전장, 소프트웨어 공학, 군사 공학, 기계학습, etc.
E-mail : wekki96@sju.ac.kr

윤 재 필(Jae-pil Youn)
2008년 육군3사관학교 전산정보처리학(학사)
2017년 아주대학교 정보통신대학원 사이버보안전공(석사)
2019~현재 세종대학교 대학원 컴퓨터공학과(박사과정)
2021~현재 육군사이버작전센터 사이버작전연습장교
관심분야 : 국방정보시스템, 사이버보안, etc.
E-mail : jpyoun@sju.ac.kr

신 동 규(Dong-kyoo Shin)
1986년 서울대학교 컴퓨터과학과(학사)
1992년 Illinois Institute of Technology 대학원 컴퓨터과학과(석사)
1997년 Texas A&M University 대학원 컴퓨터과학과(박사)
1998~현재 세종대학교 컴퓨터공학과 교수
관심분야 : 사이버전, 사이버보안, 사이버 지휘통제, 인공지능, 정보보호, etc.
E-mail : shindk@sejong.ac.kr