https://doi.org/10.33778/kcsa.2021.21.3.057

불균형 데이터 처리를 통한 침입탐지 성능향상에 관한 연구★

정 일 옥*, 지 재 원**, 이 규 환**, 김 묘 정**

요 약

침입탐지 분야에서 딥러닝과 머신러닝을 이용한 탐지성능이 검증되면서 이를 활용한 사례가 나날이 증가하고 있다. 하지만, 학습에 필요한 데이터 수집이 어렵고, 수집된 데이터의 불균형으로 인해 머신러닝 성능이 현실에 적용되는데 어려움이 있다. 본 논문에서는 이에 대한 해결책으로 불균형 데이터 처리를 위해 t-SNE 시각화를 이용한 혼합샘플링 기법을 제안한다. 이를 위해 먼저, 페이로드를 포함한 침입탐지 이벤트에 대해서 특성에 맞게 필드를 분리한다. 분리된 필드에 대해 TF-IDF 기반의 피처를 추출한다. 추출된 피처를 기반으로 혼합샘플링 기법을 적용 후 t-SNE를 이용한 데이터 시각화를 통해 불균형 데이터가 처리된 침입탐지에 최적화된 데이터셋을 얻게 된다. 공개 침입탐지 데이터셋 CSIC2012를 통해 9가지 샘플링 기법을 적용하였으며, 제안한 샘플링 기법이 F-score, G-mean 평가 지표를 통해 탐지성능이 향상됨을 검증하였다.

A study on intrusion detection performance improvement through imbalanced data processing

Jung Il Ok*, Jae-Won Ji**, Gyu-Hwan Lee**, Myo-Jeong Kim**

ABSTRACT

As the detection performance using deep learning and machine learning of the intrusion detection field has been verified, the cases of using it are increasing day by day. However, it is difficult to collect the data required for learning, and it is difficult to apply the machine learning performance to reality due to the imbalance of the collected data. Therefore, in this paper, A mixed sampling technique using t-SNE visualization for imbalanced data processing is proposed as a solution to this problem. To do this, separate fields according to characteristics for intrusion detection events, including payload. Extracts TF-IDF-based features for separated fields. After applying the mixed sampling technique based on the extracted features, a data set optimized for intrusion detection with imbalanced data is obtained through data visualization using t-SNE. Nine sampling techniques were applied through the open intrusion detection dataset CSIC2012, and it was verified that the proposed sampling technique improves detection performance through F-score and G-mean evaluation indicators.

Key words : imbalanced data, intrusion detection, machine learning, sampling

접수일(2021년 09월 01일), 게재확정일(2021년 09월 17일)

★ 본 논문은 2021년 정부(국토교통부)의 재원으로 국토교통과학기술진흥원(KAIA)의 지원을 받아 연구가 수행된 연구임 (21TLRP-B152768-03, 자율협력주행 도로교통체계 통합보안 시스템 운영을 위한 기술 및 제도개발).

* 고려대학교/정보보호학과(주저자)
** 이글루시큐리티(공동저자)

1. 서 론

인터넷 사용이 증가하면서 사이버 위협 또한 증가하고 있다. 이에 침입탐지 분야에서는 기존의 시그니처 기반의 탐지 기법에 대한 한계를 넘을 수 있는 방안에 관해 지속적인 연구가 진행되고 있다. 침입탐지 분야에서 딥러닝과 머신러닝을 이용한 탐지성능에 대한 검증으로 인해 활용 사례가 나날이 증가하고 있다. 하지만, 학습에 필요한 데이터 수집이 어렵고, 수집된 데이터 또한 불균형으로 인해 머신러닝 성능에 영향을 주고 있다. 특히, 침입탐지 분야에서 학습 데이터 수집은 보안상 비밀리에 다루어지기도 하며, 수집이 되었다고 하더라도 자주 탐지되는 공격에 대한 데이터가 다수이며, 자주 발생하지 않은 공격에 대해서는 수집하기가 어렵기 때문이다.

침입탐지에서의 분류 문제는 입력 데이터가 주어졌을 때 해당 데이터의 클래스를 예측하는 문제를 말한다. 일반적으로 이러한 문제를 해결하기 위해 머신러닝 알고리즘을 주어진 데이터셋으로 학습시켜 모델을 만든다. 이때 이상적인 데이터셋은 분류하고자 하는 클래스의 데이터 분포가 균일해야 한다. 그러나 데이터셋 대부분은 클래스마다 데이터 수 차이가 존재하며 심할 때는 클래스 하나에만 데이터가 편중되기도 한다. 머신러닝 알고리즘은 각 클래스의 비율이 비슷한 상황을 가정하기 때문에, 클래스가 불균형한 데이터셋의 경우 전체적인 데이터에 대해 제대로 학습하지 못하고 큰 비중을 차지하는 클래스에 편향되어 학습된다. 그 결과 전체적인 정확도는 높으나 정작 원하는 항목에 대해서는 분류해 내지 못하는 클래스 불균형 현상이 발생한다. 이러한 불균형 데이터셋은 네트워크 침입 감지[1], 스팸 감지[2], 텍스트 분류[3], 의료 애플리케이션[4]과 같은 많은 실제 도메인에 존재한다. 이 부분에서 우리가 정말로 관심을 두는 것은 다수 계급이 아닌 소수 계급이다.

이러한 클래스 불균형을 해결하기 위해 주로 사용되는 방법으로 데이터 샘플링(Data Sampling) 기법이 있다. 데이터 샘플링은 불균형한 데이터셋에서 대부분을 차지하는 클래스인 다수 클래스(Majority Class)와 반대로 적은 부분만 차지하는 소수 클래스(Minority Class)의 샘플 개수를 조정하여 균형 있는 데이터 집합으로 만드는 기법으로, 두 클래스 중 어느 클래스의 샘플 개수를 조절하느냐에 따라 언더샘플링(Under-sampling) 기법과 오버샘플링(Over-sampling) 기법으로 분류된다[5], [13].

언더샘플링은 소수 클래스의 샘플 수에 맞도록 다수 클래스의 샘플을 제거하는 기법이다. 언더샘플링 기법으로는 RUS(Random Under-sampling)과 ENN(Edited Nearest Neighbor), Tomek Links 등의 기법이 제안되고 있다. 그러나 언더샘플링 기법은 데이터를 제거하기 때문에 정보의 손실을 유발하게 된다는 문제점이 있다.

오버샘플링은 언더샘플링과는 반대로 다수 클래스 샘플 개수에 맞춰 소수 클래스를 위한 샘플을 생성하는 기법으로, 정보 손실을 피할 수 있다. 오버샘플링 기법에는 ROS(Random Over-sampling), SMOTE(Synthetic Minority Over-sampling Technique)[5], ADASYN(Adaptive synthetic sampling)[6], B-SMOTE(Borderline SMOTE)[7] 등 다양한 기법이 존재한다. 하지만 오버샘플링 기법들 역시 데이터 생성 과정을 통해 분류 모델이 학습 데이터에 과적합(Overfitting)이 발생할 수 있다.

본 논문에서는 불균형한 침입탐지 데이터셋에 대한 탐지성능을 향상하기 위해 페이로드 특성에 맞는 필드 분류와 TF-IDF를 이용한 피처 추출 후 샘플링 기법을 적용한다. 이렇게 생성된 데이터셋에 대해서 t-SNE 시각화 기법을 통해 최적의 데이터셋을 추출하게 된다. 이때, 침입탐지 이벤트를 분류하는데 최적의 샘플링 기법을 찾기 위해 언더샘플링, 오버샘플링, 혼합(오버샘플링+ 언더샘플링) 샘플링을 적용하여 비교 분석하며, 분류 알고리즘으로는 XGBoost를 사용한다.

2. 관련 연구

침입탐지 데이터는 기본적으로 정상적인 네트워크 트래픽에서 악의적인 트래픽을 분류해 내는 과

정이기 때문에 자연스럽게 매우 불균형한 데이터로 구성되어 있다. 특히, 공격에 대한 파급효과 및 위험성이 크지만, 자주 발생하지 않은 침입 공격 유형은 학습 데이터 자체가 적기 때문에 다른 공격을 탐지하는 것보다 탐지에 대한 정확성이 적어진다. 이 때문에 침입탐지 분야에서도 불균형 데이터에 대한 문제를 해결하기 위해 <표 1>과 같이 다양한 연구가 진행되고 있다.

<표 1> 침입탐지 분야의 불균형 데이터 처리 관련 연구

년도	저자	데이터셋
2016	S. Rodda and U. S. R. Erothi	NSL-KDD
2016	Yong Sun, Feng Liu	KDD CUP 99
2016	Reza, Mohammad & Miri Rostami, et al.	NSL-KDD
2017	B. Yan, G. Han, M. Sun and S. Ye	NSL-KDD
2018	Peihuang Su, Yanhua Liu, and Xiang Song	KDD CUP 99
2018	Seo, Jae-Hyun & Kim, Yong-Hyuk.	KDD CUP 99
2019	Tripathi, Priyanka & Makwana, Rajni	KDD CUP 99
2019	JooHwa Lee & KeeHyun Park	CICIDS2017
2019	Ibrahim Yilmaz, Rahat Masum	UGR16
2020	Zhang, Hongpo & Huang, Lulu & Wu, Chase & Li, Zhanbo.	UNSW-NB15, CICIDS2017
2020	Bedi, Punam & Gupta, Neha et al.	NSL-KDD
2020	Zhang, Jie & Zhang, Yong & Li, Kexin	NSL-KDD

Sun. Y. et al.(2016)[10]은 SMOTE 기반의 개선된 SMOTE-NCL을 제안했다. SMOTE-NCL은 각 클래스의 비율과 이를 기반으로 계산된 평균 비율, 클래스 비율의 표준 편차, 표준 편차를 클래스 비율로 나누어 얻은 불균형 척도를 계산하고, SMOTE를 사용하여 소수 클래스 데이터를 샘플링 하였다. 또한, 샘플링 후 이웃 청소 규칙을 통해 노이즈로 간주하는 데이터를 처리하는 방법을 제안했다. 사용된 데이터셋은 KDD Cup 99로 SMOTE-NCL을 통해 희귀 클래스뿐만 아니라 다른 클래스의 AUC도 향상됨을 보였다.

B. et al.(2017)은 기존의 SMOTE의 문제점인 과적합을 해결하고자 제시된 B-SMOTE와 SMOTE-ENN 방법이 침입탐지에서는 적절하지 않다고 이야기하면서 Region Adaptive 기반의 SMOTE를 제시하였다. 제안한 RA-SMOTE 알고리즘을 통해 희귀 클래스의 탐지율을 효과적으로 향상하게 시킬 수 있음을 보였다. 사용된 데이터셋은 NSL-KDD로 U2L 및 R2L과 같은 희귀 클래스에 대한 탐지율을 향상할 수 있었다.

Tripathi et al.(2019)은 KDD Cup 99 데이터셋에 존재하는 클래스 불균형 문제를 완화하기 위해서 알고리즘 방식인 AdaBoost와 Random Forest 분류기 조합을 제안하였다. 이를 위해 희귀 클래스 U2R 및 R2L에 대해 50% 1,000%까지 다양한 비율로 활용되었으며, 이때 앙상블 분류를 사용하였다.

Zhang. et al.(2020)은 소수 클래스 탐지율을 높이기 위해 SGM이라는 대규모 데이터셋에 대한 클래스 불균형 처리 기술을 제안한다. SGM은 SMOTE라는 오버샘플링과 GMM의 언더샘플링을 결합한 모델이다. 데이터셋으로 UNSW-NB15, CICIDS 2017을 사용하였다.

Bedi, Punam & Gupta, Neha et al.(2020)은 클래스 불균형 문제를 처리하기 위해 입력 쌍 간의 유사성 점수를 계산하여 클래스 간의 동질성을 식별하는 Siam-IDS 방법을 제안하였다. 이를 통해 DNN 및 CNN 기반의 IDS(침입탐지시스템)에서 높은 재현율 값을 얻었지만, 정밀도 부분에서는 기존의 IDS(침입탐지시스템)보다 낮은 성능을 보이는 단점이 있었다. 사용한 데이터셋은 NSL-KDD를 사용하였다.

JooHwa Lee, KeeHyun Park(2019)는 침입탐지에 최근 딥러닝을 사용한 연구가 증가하면서 클래스 불균형 문제가 기존의 전통적인 머신러닝 기반의 알고리즘을 적용할 때보다 미치는 영향이 크다고 했다. 또한, 불균형 문제를 해결하기 위한 기존의 연구가 데이터 손실 또는 과적합에 대한 약점을 지니고 있다고 지적했다. 이를 위해 기존의 데이터와 유사하면서 새로운 가상 데이터를 생성하는 Generative Adversarial Network(GAN) 모델을 제안하였다. 이를 통해 기존의 SMOTE 모델보다 G

AN을 이용한 모델의 성능이 우수함을 보였다. 이때 사용한 데이터셋은 CICIDS2017로 0.1% 미만으로 구성된 희귀한 클래스인 Bot, Infiltration, Heartbleed에 GAN을 이용하여 오버샘플링을 수행하였다.

침입탐지 분야의 불균형에 대한 접근은 데이터 분야 접근과 알고리즘 접근이 혼합되는 방식의 형태로 변하고 있으며, 최근에는 GAN 등을 활용하여 불균형 문제를 해결하려는 연구가 활발히 진행되고 있다.

3. 샘플링 기법을 통한 불균형 학습 데이터 처리

3.1 제안된 접근법

이 장에서는 수집된 데이터를 통해 만들어진 학습 데이터셋에 대해서 샘플링 기법을 통해 침입탐지 성능을 강화하는 방안을 제시한다.

실 사이버 공간에서는 정상적인 활동이 대다수를 차지하므로 대부분의 트래픽 데이터는 정상적인 이벤트이다. 침입탐지 이벤트는 이러한 대다수의 정상적인 트래픽 가운데서 소수의 침입 데이터를 탐지하게 된다. 이러한 소수의 침입 데이터를 탐지하는 것으로 사이버 침입탐지 이벤트는 매우 불균형한 데이터 기반이며, 중복된 형태로 구성이 되어 있다.

따라서 머신러닝 기반의 알고리즘은 소수의 클래스에 대해서 완전히 학습할 수 없으며, 잘못된 분류를 하기 쉽다. 또한, 최근에는 부스팅(Boosting)과 딥러닝 기반의 알고리즘을 사용하는 사례가 증가하고 있어서 이들 알고리즘 성능 측정 시 학습 데이터의 질과 양에 대한 부분이 더욱더 중요하다.

불균형 데이터 처리에 대한 샘플링 프로세스는 다음과 같다.

첫째, 페이로드 기반의 불균형 침입탐지 데이터셋에 필드 분류 후 TF-IDF 피처 추출을 통한 전처리를 수행한다.

둘째, 전처리 수행 후 테스트와 훈련 데이터셋으로 분류한다.

셋째, 클래스로 구분된 훈련 데이터에 대해 시각화 t-SNE를 통해 분석한다. 여기에서 클래스는 공격 유/무를 나타내는 Binary Class와 공격 유형별로 클래스가 분류된 Multi Class로 구분된다.

넷째, 원본 데이터 및 각 샘플링이 적용된 후의 데이터에서 20% 데이터만 추출한다. 이때 적용된 샘플링 방법은 Binary와 Multi Class로 구분될 수 있으며, CSIC2012 원본 데이터를 기반으로 오버샘플링과 언더샘플링이 모두 섞인 혼합샘플링(SMOTE+ ENN, SMOTE+ Tomek) 기법 수행을 통해 데이터를 생성한다.

다섯째, Binary, Multi Class 비율을 유지하되 무작위로 샘플링 추출이 이루어진다.

여섯째, 추출된 데이터에 대해서 t-SNE 시각화를 구현한다.

마지막으로 이렇게 추출된 데이터에 대해서 분류 알고리즘 XGBoost를 사용하여 가장 좋은 성능이 도출된 모델을 선택한다.

3.2 불균형 데이터에 대한 샘플링 기법

다음은 본 논문에서 불균형 클래스 처리를 위해 사용된 샘플링 기법에 대해서 알아보도록 한다. 침입탐지 이벤트를 분류하는데 가장 최적의 샘플링 기법을 찾기 위해 <표 2>에서와 같이 오버샘플링, 언더샘플링, 혼합(오버+ 언더)샘플링으로 크게 3가지로 분류된 9가지 샘플링 기법을 비교 분석하였다.

<표 2> 3가지 분류에 따른 샘플링 기법

분류	기법
언더샘플링	RUS, ENN, Tomek
오버샘플링	ROS, SMOTE, B-SMOTE, ADASYN
혼합샘플링	SMOTE+ENN, SMOTE+Tomek

4. 실험 및 평가

본 실험에서는 침입탐지 기반의 데이터셋에서 불균형이 머신러닝 성능에 얼마나 영향을 미치는지를 알아보고, 이를 해결하기 위한 다양한 데이터 샘플링 기법을 적용하여 최적의 효율적인 샘플링 기법

을 평가하도록 한다.

우리는 본 실험을 통해 다음과 같은 질문을 해결하고자 한다. 첫째, 불균형한 침입탐지 데이터셋에 샘플링 기법을 통해 머신러닝 성능을 향상할 수 있는가? 둘째, 가장 적절한 샘플링 기법은 무엇인가? 셋째, 바이너리 클래스와 멀티 클래스 형식의 불균형 침입탐지 데이터셋에서 소수 클래스 탐지성능에 영향을 미치는가?

본 실험에서 실험 데이터셋으로 웹어플리케이션 침입탐지 데이터셋인 CSIC2012를 활용하였다.

해당 데이터셋의 불균형 문제를 해결하기 위해 샘플링 기법으로 RUS, ENN, Tomek, ROS, SMOTE, B-SMOTE, ADASYN, SMOTE+ENN, SMOTE+Tomek을 사용하였다.

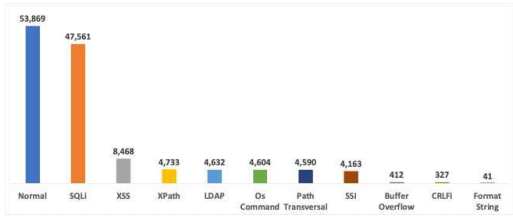
또한, 알고리즘으로는 XGBoost를 사용하였다.

4.1 데이터셋

본 실험을 위해 우리는 다양한 침입탐지 분야의 공용 데이터셋 가운데 CSIC2012 (HTTP CSIC Torpeda 2012 데이터셋)을 사용했다. 대부분의 침입탐지 분야의 데이터셋은 피처화 되어 있어 현실 데이터와 차이가 있다. 이 때문에 HTTP 요청 헤더와 페이로드가 포함된 데이터셋을 선택하였다.

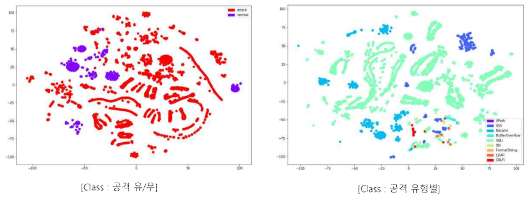
HTTP 요청에 헤더와 페이로드가 모두 포함된 데이터셋을 선택한 이유는 같은 시그니쳐를 가지고 탐지하는 방법과 보호하는 웹서비스의 개발언어와 웹구조에 따라 다른 도메인의 특징을 가지는 페이로드를 통해 같은 피처 공간을 가지고 있지만, 분포가 다른 구성을 위해 선택하였다.

CSIC2012 데이터셋은 2012년 RECSI2012에서 TORPEDA Framework를 통해 제시되었다. TORPEDA 프레임워크는 웹 공격 탐지 시스템의 평가 및 테스트를 위해 레이블이 지정된 웹 트래픽을 개발하는 데 사용된다[17],[18],[19]. 데이터 구성은 10개의 클래스로 구성되어 있으며, 정상 (Normal)으로 분류된 8,363건의 요청과 Anomalous로 분류된 16,456건, 공격 별로 분류된 40,948건의 요청이 포함되어 있다. 10가지 공격 유형은 Normal, XSS, SQLi, Buffer Overflow, LDAP, XPath, Format String, SSI, CRLFi, Anomalous로 (그림 1)과 같이 구성되어 있다. 데이터셋은 XML 파일 형식으로 되어 있으며, label, request로 구성되어 있으며, request는 method, protocol, path, headers, body로 구분되어 있다.



(그림 1) CSIC2012 데이터셋 클래스별 현황 그래프

또한, 해당 데이터셋에 대해서 (그림 2)와 같이 공격 유/무와 공격 유형별 레이블에 대해 t-SNE를 통해 시각화로 표현하면 다음과 같다.



(그림 2) CSIC2012 데이터셋에 대한 t-SNE 시각화 결과

4.2 데이터 전처리

본 단계는 수집된 데이터셋 CSIC2012를 머신러닝에 적용할 수 있도록 데이터를 정형화 및 전처리를 수행하는 단계이다.

전처리 단계는 수집되는 데이터 소스에 대해서 Normalization, Field Selection, Feature Extractor & Selection 순서로 진행된다.

4.2.1 Normalization

수집된 데이터셋 CSIC2012는 비정형화된 XML 형식의 데이터로 구성되어 있다. 우선, 해당 데이터셋을 동일하게 Method, Version, Uri, Query, Body로 Normalization을 수행한다. 특히, 사용자 정보 등을 나타내는 값에 대해서는 Body 필드로 포함한다. 그리고 Uri, Query, Body에 대해서

'₩n' 문자를 제거하고, Uri Decoding을 적용한다.

4.2.2 Field Selection

데이터셋에서 실험에 사용할 필드에 대해 선택한다. CSIC2012에서는 Category Type은 Class, Method, Version으로 구분할 수 있으며, Text Type은 Uri, Query, Body를 선택한다. 이때, 생성된 데이터에서 Query가 존재하지 않았을 때 결측치가 발생한다. 따라서 그 필드에 대해 '?'로 결측치를 한꺼번에 처리한다.

4.2.3. Feature Extractor & Selection

분리된 http_url, http_query, http_body 필드에 대해서 키워드별로 TF-IDF를 적용하여 벡터화를 수행하였다.

4.3 평가 방법

본 실험 환경은 Ubuntu 18.04.2 LTS 에서 Python을 사용하여 구현되었다. 사용된 머신러닝 라이브러리는 Scikit-learn 0.20.4를 사용하였다. 하드웨어 사양은 GPU는 NVidia Geforce RTX 2060 * 2 이었으며 128GB RAM, 8TB 하드 디스크, AMD Ryzen Threadripper 1900X 8-Core Processor 환경이다.

본 실험의 목적은 불균형한 클래스로 구성된 침입탐지 데이터셋에서 다수의 클래스에 대한 탐지율을 유지하면서, 소수 클래스에 대한 탐지율을 향상하는 데 있다.

이 때문에 이에 대한 평가 지표로는 일반적으로 사용되는 Confusion Matrix를 기반으로 기본적으로 사용되는 ACC, Precision, Recall, F-score 뿐만 아니라, 침입탐지 및 불균형 시스템에서 사용되는 지표인 FPR, TNR, G-mean, AUC를 추가하여 고려하였다.

평가 지표에 대한 설명은 다음과 같다. Accuracy는 모든 샘플 중에서 정상과 공격이 올바르게 분류된 항목의 비율로 정의된다. Precision은 공격이라고 예측한 것 중 실제 공격이라고 분류한 비율을 말한다. Recall은 실제 공격 중 공격이라고 예측한 비율을 말한다. (침입탐지 데이터셋에서 말하는 DR (Detection Ratio) 와 같은 의미이다) F-score는 Precision과 Recall 간의 조화평균(harmonic mean)을 의미한다. 이러한 보편적인 평가 지표에 침입탐지 및 불균형 데이터에서 사용하는 지표인 FRP, FNR, G-mean을 추가하였다. FPR(False Positive rate)는 정상 트래픽을 공격이라고 예측한 비율을 말한다. TNR(True Negative Rate)는 정상 트래픽을 정상이라고 예측한 비율을 말한다. G-mean 값은 Kubat 와 Matwin(1997)[15] 이 제안한 이 값은 민감도와 특이도의 기하평균(geometric mean)으로 계산된다. 따라서 다수 집단이 정확하게 분류되었지만, 소수 집단에 대한 예측값이 낮다면 이들의 기하평균 값인 G-mean 값이 낮게 된다. 즉, 소수 집단의 정분류율이 좋을수록 G-mean 값도 커지는 경향이 있다. AUC(Area Under Curve) 값은 ROC(Receiver Operating Characteristic) 곡선의 아래 영역을 정의하고 이 영역이 넓을수록 분류 결과가 좋다는 것을 나타낸다 (Huang, Ling, 2005)[16]

멀티 클래스 분류 문제에서는 불균형 데이터 세트에서 모델의 검출 성능을 보다 합리적으로 평가하기 위해 각 클래스 샘플 수에 따라 가중 평균 방법을 사용하여 각 지수를 계산한다.

4.4 실험 결과

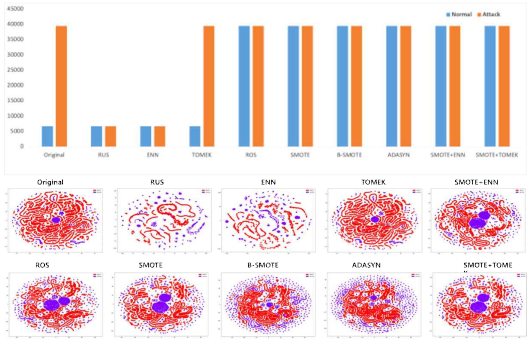
4.4.1 Binary Classification

침입탐지 데이터셋의 클래스가 Binary Classification인 경우이다. 데이터셋 CSIC2012에 대해서 샘플링을 적용하기 전인 original과 9가지의 샘플링 기술인 (RUS, ENN, Tomek, ROS, SMOTE, B-SMOTE, ADASYN, SMOTE+ENN, SMOTE+Tomek)을 포함하여 클래스 불균형 처리 기술을 비교하였다.

불균형 처리 기술 후 모델을 평가하기 위해 대표적인 알고리즘인 XGBoost를 적용하였다.

샘플링 적용 후 데이터 분포에 대해 (그림 3)과 같이 나타난다. 언더샘플링은 소수 클래스인 normal 6,691건에 맞추어져 다운되었으며, 오버샘플링은 다수 클래스인 attack 39,449건에 근접하게 생겨났다. 특히, ENN, Tomek의 경우 RUS와 같이 소수 클래스에 정확히 일치하는 게 아니라 각 샘플

링 특성에 맞게 데이터 수가 결정되었다. 오버샘플링은 ROS, SMOTE, B-SMOTE는 다수 클래스인 attack 39,449건과 같이 생성되었지만, ADASYN은 다른 수로 증가하였다.



(그림 3) Binary Classification 샘플링 적용 후 데이터 수와 t-SNE를 통해 본 데이터 분포도

<표 3>은 CSIC2012 데이터셋에서 XGBoost 알고리즘 기반으로 모델을 생성한 결과이다. 정확성 측면에서는 오히려 원본 데이터셋이 언더샘플링보다 약간 좋지만, 오버샘플링, 혼합샘플링 결과보다는 좋지 않음을 알 수 있다. 특히, 이 부분에 대해서는 혼합샘플링의 결과가 가장 좋음을 알 수 있다.

FPR는 정상 트래픽을 공격이라고 잘못 측정한 오탐 비율이다. FPR는 낮을수록 좋은 결과이며, 원본 데이터에서는 오탐의 비율 0.0012%가 있었지만, 샘플링된 모든 경우에 대해서는 모두 0으로 좋은 결과임을 확인할 수 있다. TPR은 공격 트래픽을 공격으로 예측한 결과로 보통 탐지율로 불리는 것으로 침입탐지에서 중요한 지표이다. 언더샘플링 ENN의 경우 좋지 않은 결과를 보이지만, 다른 샘플링의 경우 원본 데이터와 비슷한 결과를 보인다. TNR은 정상 트래픽을 정상으로 예측한 비율로 원본 데이터보다 모두 좋은 결과를 보인다.

G-mean은 민감도와 특이도의 기하평균을 계산한 것으로 다수 집단이 정확하게 분류되었지만, 소수 집단의 예측력이 낮다면 G-mean값은 낮아지게 된다. 이 때문에 클래스 불균형을 보일 때 중요하게 보는 지표 중 하나이다. 이곳에서 언더샘플링 중 ENN을 제외하고는 원본보다 더 좋은 결과를 보이고 있으며, 오버샘플링이나 혼합샘플링의 경우 모두 좋은 결과를 내고 있다. 또한, Recall 값이나 F-score 부분은 원본과 혼합샘플링의 경우가 좋은 결과를 내고 있으며, 언더나 오버샘플링 같은 경우 약간 낮은 성능을 보이고 있음을 알 수 있다.

4.4.2 Multi Classification

침입탐지 데이터셋의 클래스가 Multi Classification인 경우이다. 데이터셋 CSIC2012에 대해서 샘플링을 적용하기 전인 original과 9가지의 샘플링 기술인(RUS, ENN, Tomek, ROS, SMOTE, B-SMOTE, ADASYN, SMOTE+ENN, SMOTE+Tomek)을 포함하여 클래스 불균형 처리 기술을 비교하였다.

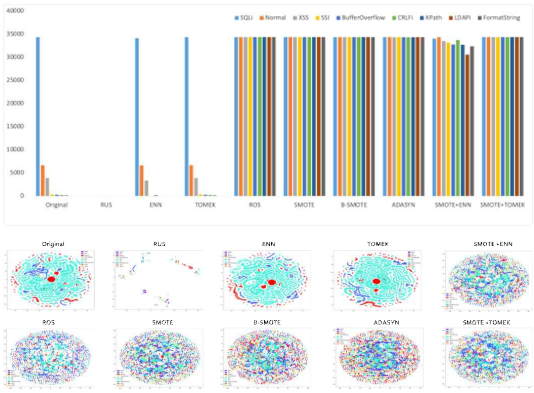
<표 3> Binary classification 평가 결과 CSIC2012 데이터셋 (Algorithm: XGBoost)

XGBoost	Original	RUS	ENN	Tomek	ROS	SMOTE	B-SMOTE	ADASYN	SMOTE +ENN	SMOTE +Tomek
Accuracy	0.99853	0.99974	0.79097	0.99853	0.99870	0.99974	0.99827	0.99991	0.99870	0.99974
FPR	0.00120	0.00000	0.00000	0.00120	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
TPR	0.99848	0.99970	0.75553	0.99848	0.99848	0.99970	0.99797	0.99990	0.99848	0.99970
TNR	0.99880	1.00000	1.00000	0.99880	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
G-mean	0.99864	0.99985	0.86921	0.99864	0.99924	0.99985	0.99899	0.99995	0.99924	0.99985
Precision	0.99980	1.00000	1.00000	0.99980	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Recall	0.99848	0.99970	0.75553	0.99848	0.99848	0.99970	0.99797	0.99990	0.99848	0.99970
F-score	0.99914	0.99985	0.86074	0.99914	0.99924	0.99985	0.99898	0.99995	0.99924	0.99985

<표 4> 샘플링 적용 후 데이터셋 분포 (알고리즘: XGBoost, 평가 지표: G-mean, F-score)

XGBoost		Original	RUS	ENN	Tomek	ROS	SMOTE	B-SMOTE	ADASYN	SMOTE +ENN	SMOTE +Tomek
G-mean	CRLFi	0.45714	0.34361	0.44643	0.32143	0.58182	0.53659	0.35106	0.43919	0.74860	0.60377
	LDAPi	0.55556	0.16129	0.00000	0.00000	0.00000	0.53659	0.28571	0.15385	0.46512	0.51429
	SQLi	0.99569	0.91409	0.99301	0.99078	0.99466	0.98518	0.97588	0.99264	0.98619	0.98774
	XSS	0.93935	0.83584	0.91027	0.91531	0.92678	0.93700	0.90288	0.93961	0.96292	0.94212
F-score	CRLFi	0.45714	0.34361	0.44643	0.32143	0.58182	0.53659	0.35106	0.43919	0.74860	0.60377
	LDAPi	0.55556	0.16129	0.00000	0.00000	0.00000	0.53659	0.28571	0.15385	0.46512	0.51429
	SQLi	0.99569	0.91409	0.99301	0.99078	0.99466	0.98518	0.97588	0.99264	0.98619	0.98774
	XSS	0.93935	0.83584	0.91027	0.91531	0.92678	0.93700	0.90288	0.93961	0.96292	0.94212

(그림 4)를 통해 샘플링이 적용된 데이터셋에서 다수 클래스로는 SQLi, XSS, 소수 클래스로는 CRLFi, LDAPi로 도출하였다. 또한, 해당 데이터에 대한 분포를 t-SNE를 통해 살펴보았다.



(그림 4) Multi Classification 샘플링 적용 후 데이터 수와 t-SNE를 통해 본 데이터 분포도

<표 4>는 각 공격 별로 샘플링 후 알고리즘 XGBoost에 의해 생성된 모델에 대한 공격 별 G-mean, F-score 값이다. G-mean 값은 다수 클래스인 SQLi, XSS는 원본 데이터셋 뿐만 아니라 다른(언더, 오버, 혼합) 샘플링 기법들 또한 큰 값을 내고 있다. 소수 클래스에 속해 있는 CRLFi, LDAPi는 원본 데이터셋보다 전체적으로 큰 값을 내고 있으며, 혼합샘플링 > 오버샘플링 > 언더샘플링 순으로 좋은 결과값을 내고 있다.

F-score 값은 다수 클래스인 SQLi, XSS는 원본 데이터셋이나 다른(언더, 오버, 혼합) 샘플링 기법들과 같이 큰 값을 내고 있다. 소수 클래스에 속해 있는 CRLFi, LDAPi는 전체적으로 크게 변화는 없음을 알 수 있다.

5. 결 론

지금까지 본 논문에서는 실험을 통해 불균형 데이터에 대해 제안한 t-SNE 시각화를 이용한 혼합 샘플링 기법을 적용한 데이터셋으로 생성된 침입탐지 모델이 다른 샘플링 기법보다 긍정적인 결과를 도출함을 확인하였다. 이는 다수의 침입탐지 공격에 대한 탐지율을 유지하면서, 수집하기 어려운 침입탐지 공격에 대한 탐지율을 향상할 수 있다는 측면에서 본 실험은 유용한 결과이다.

하지만, 혼합샘플링 기법도 데이터의 특성에 따라 결과의 차이가 발생하므로 앞으로는 자동으로 최적의 샘플링 기법을 찾아 적용할 수 있는 앙상블 기반의 샘플링 기법에 관한 연구를 수행하고자 한다.

참고문헌

[1] TAN X P, SU S J, HUANG Z P, et al. Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm. Sensors, 2019, 19(1): 203.

[2] LI C L, LIU S G. A comparative study of the class imbalance problem in Twitter spam detection. Concurrency and Computation: Practice and Experience, 2017, 30(5): e4281.

[3] LI Y L, SUN G S, ZHU Y H. Data imbalance problem in text classification. Proc. of the 3rd International Symposium on Information Processing, 2010: 301 - 305.

[4] ZHU M, XIA J, JIN X Q, et al. Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access, 2018, 6: 4641- 4652.

[5] Yan, B.; Han, G.; Sun, M.; Ye, S. A Novel Region Adaptive SMOTE Algorithm for Intrusion Detection on Imbalanced Problem. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC); IEEE: Chengdu, December 2017; pp. 1281-1286.

[6] H. Haibo , A. Garcia, E.: "Learning from Imbalanced Data", IEEE Transactions On Knowledge And Data Engineering, Vol.2, No.9, September (2009).

[7] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Proc. of the International Conference on Advances in Intelligent Computing, 2005: 878 - 887.

[8] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTE- Boost: improving prediction of the minority class in boosting. Proc. of the 7th European Conference on Principles and Prac- tice of Knowledge Discovery in Databases, 2003: 107 - 119.

[9] FREUND Y. Experiment with a new boosting algorithm. Proc. of the 13th International Conference on Machine Learning, 1996: 148 - 156.

[10] Yong Sun; Feng Liu SMOTE-NCL: A Re-Sampling Method with Filter for Network Intrusion Detection. In Proceedings of the 2016 2nd IEEE International Conference on Computer and Communications (ICCC); IEEE: Chengdu, China, October 2016; pp. 1157-1161.

[11] 정일옥, 전이학습과 불균형 데이터 처리를 통한 침입탐지 성능향상에 관한 연구, 박사학위논문, 고려대학교 2021. 8

[12] Leea, H.J.; Lee, S. 데이터 전처리와 앙상블 기법을 통한 불균형 데이터의 분류모형 비교 연구. 응용통계연구 2014, 27, 357-371, doi:10.5351/KJAS.2014.27.3.357.

[13] Son, M.J.; Jung, S.W.; Hwang, E.J. 불균형 데이터 분류를 위한 딥러닝 기반 오버샘플링 기법. 정보처리학회논문지:소프트웨어 및 데이터공학 2019, 8, 311-316, doi:10.3745/KTSDE.2019.8.7.311.

[14] Kim, D.; Kang, S.; Song, J. 불균형 자료에 대한 분류분석. 응용통계연구 2015, 28, 495 -509, doi:10.5351/KJAS.2015.28.3.495.

[15] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in Proceedings of the International Conference on Machine Learning, pp. 179-186, Nashville, Tenn, USA, 1997.View at: Google Scholar

[16] Y. Liu, X. H. Yu, J. X. Huang, and A. J. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," Information Processing & Management, vol. 47, no. 4, pp. 617-631, 2011.View at: Publisher Site | Google Scholar

[17] Csic torpeda 2012, http data sets, July 20, 2021. [Online]. Available: http://www.tic.itefi.csic.es/torpeda.

[18] Carmen Torrano-Gimenez, Alejandro Perez-Villegas, and Gonzalo Alvarez. "TORPEDA: Una Especificacion Abierta de Conjuntos de Datos para la Evaluacion de Cortafuegos de Aplicaciones Web." 2012. TIN2011-29709-C0201.

[19] Web Attacks Detection based on CNN - Csic torpedo 2012 http data sets - GitHub, July 20, 2021. [Online]. Available: https://github.com/DuckDuckBug/cnn_waf.

[저자소개]

정 일 옥 (Il-ok Jung)
2001년 2월 전남대학교 물리학과 학사
2008년 8월 고려대학교 컴퓨터공학과 석사
2021년 8월 고려대학교 정보보호학과 박사
email : okkida@korea.ac.kr

지 재 원 (Jae-Won Ji)
2010년 2월 한남대학교 컴퓨터공학과 학사
2012년 2월 한남대학교 컴퓨터공학과 석사
2012년 3월 ~ 현재 이글루시큐리티 재직
email : jaewon.ji@igloosec.com

이 규 환 (Gyu-Hwan Lee)
2003년 2월 명지대학교 컴퓨터공학과 학사
2013년 8월 ~ 현재 이글루시큐리티 재직

email : gubooki@naver.com

김 묘 정 (Myo-Jeong Kim)
2021년 2월 이화여자대학교 사이버보안전공 학사
2021년 4월 ~ 현재 이글루시큐리티 재직

email : myojeong126@gmail.com