https://doi.org/10.33778/kcsa.2023.23.1.069

혼합샘플링 기법을 사용한 랜섬웨어탐지 성능향상에 관한 연구

김 수 철*, 이 형 동**, 변 경 근**, 신 용 태***

요 약

최근 아일랜드 보건당국, 美 송유관 등 全 세계적으로 랜섬웨어 피해가 급증하고 있으며, 사회 모든 분야에 피해를 입히고 있다. 특히, 랜섬웨어 탐지 및 대응에 기존의 탐지방법뿐 아니라 머신러닝 등을 이용한 연구가 늘어나고 있다. 하지만, 전통적인 머신러닝은 모델이 데이터가 많은 쪽으로 예측하는 경향이 강해 정확한 예측값을 추출하기 어려운 문제점이 있다. 이에 다수(Majority)의 Non-Ransomware(정상코드 또는 멀웨어)와 소수의(Minority) Ransomware로 구성된 불균형(Imbalance) 클래스에서 샘플링 기법을 통해 불균형을 해소하고 랜섬웨어탐지 성능을 향상시키는 기법을 제안하였다. 본 실험에서는 두가지 시나리오(Binary, Multi Classification)을 사용하여 샘플링 기법이 다수 클래스의 탐지 성능을 유지하면서 소수 클래스의 탐지 성능을 개선함을 확인하였다. 특히, 제안된 혼합샘플링 기법(SMOTE+ENN)이 10% 이상의 성능(G-mean, F1-score) 향상을 도출했다.

A study on the improvement ransomware detection performance using combine sampling methods

Kim Soo Chul*, Lee Hyung Dong**, Byun Kyung Keun**, Shin Yong Tae***

ABSTRACT

Recently, ransomware damage has been increasing rapidly around the world, including Irish health authorities and U.S. oil pipelines, and is causing damage to all sectors of society. In particular, research using machine learning as well as existing detection methods is increasing for ransomware detection and response. However, traditional machine learning has a problem in that it is difficult to extract accurate predictions because the model tends to predict in the direction where there is a lot of data. Accordingly, in an imbalance class consisting of a large number of non-Ransomware (normal code or malware) and a small number of Ransomware, a technique for resolving the imbalance and improving ransomware detection performance is proposed. In this experiment, we use two scenarios (Binary, Multi Classification) to confirm that the sampling technique improves the detection performance of a small number of classes while maintaining the detection performance of a large number of classes. In particular, the proposed mixed sampling technique (SMOTE+ENN) resulted in a performance(G-mean, F1-score) improvement of more than 10%.

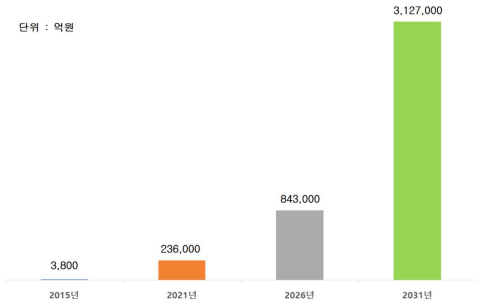
Key words : ransomware detection, sampling, imbalanced data, classification, machine learning

접수일(2023년 02월 28일), 수정일(2023년 03월 14일), 게재확정일(2023년 03월 31일)

* 숭실대학교/IT정책경영학과 박사과정(주저자)
** 숭실대학교/IT정책경영학과 박사과정(공동저자)
*** 숭실대학교/IT정책경영학과 교수(교신저자)

1. 서 론

IT 기술의 발전으로 사이버 위협은 더욱더 고도화, 지능화 되고 있다. 더구나 랜섬웨어 피해 규모는 (그림 1)과 같이 2015년 3천8백억원, 2021년 23조6천억원에서 2031년에는 311조7천억원으로 급증할 것으로 예상된다[1]. 따라서 기존의 보안 접근 방식으로는 모든 위협을 탐지하고 대응 하기에는 한계에 이르렀다. 이에 사이버 보안 분야에 머신러닝을 적용하는 사례가 증가하고 있다[2],[3].



(그림 1) 랜섬웨어 피해규모[1]

랜섬웨어 데이터셋은 다수의 Non-Ransomware (정상코드 또는 멀웨어)와 소수의 Ransomware로 구성되어 있다. 랜섬웨어탐지는 다수의 Non-Ransomware 가운데서 소수의 Ransomware를 찾는 것이다. 이러한 소수의 랜섬웨어는 매우 불균형한 데이터가 중복된 형태로 구성되어 있다. 때문에 머신러닝 학습에 사용되는 랜섬웨어 데이터셋은 불균형 클래스로 구성되어 있다.

이러한 불균형 클래스는 랜섬웨어탐지 성능에 부정적인 영향을 주게 된다. 많은 전통적인 분류 방법에서 다수 클래스에 비해 소수 클래스는 무시되거나 잘 분류되지 않은 경향이 있다. 사실 의사결정 트리, 부스팅 등 다수의 분류 알고리즘이 균형 잡힌 클래스의 데이터셋에 성능이 최적화 되도록 설계되어 있다[4],[5].

때문에 불균형 데이터 처리는 머신러닝 분야에서 해결되야 할 문제 중 하나였다. 이에 많은 연구에서 불균형 데이터 문제를 해결하기 위해 데이터 수준의 접근과 알고리즘 수준의 접근 등 다양한 방법이 제안되고 있다. 데이터 수준의 접근 방식은 샘플링을 통해 훈련 데이터셋의 균형을 맞추는 것이고 알고리즘 수준의 접근 방식은 불균형한 데이터에 대응하기 위해 새로운 알고리즘을 개발하는 것이다.

이에 본 논문에서는 랜섬웨어탐지 분야의 불균형 데이터 문제를 해결하기 위해 데이터 수준의 접근 방식으로 혼합샘플링을 제안한다. 샘플링은 데이터의 분포와 전처리, 알고리즘에 따라 최적의 기법이 다르다. 이에 본 실험에서는 원본 데이터를 크게 변경하지 않으면서도 성능을 보장 받을 수 있는 방법을 제안한다.

이를 위해 다음 사항을 고려하였다.

첫째, 이번 연구의 대상이 되는 데이터셋은 윈도우용 램섬웨어로 한정하였으며, 공격에 대한 근거가 되는 프로세스 및 네트워크 명령어를 중점으로 feature를 추출하였다. 둘째, 불균형 클래스 처리를 위해 데이터 기반의 샘플링 기법을 사용하였다. 샘플링 기법은 오버샘플링, 언더샘플링, 혼합샘플링으로 나눌 수 있으며 데이터셋의 특성과 분포, 적용 알고리즘에 따라 머신러닝의 성능이 달라진다. 오버샘플링은 분포가 작은 클래스 값을 분포가 큰 클래스로 맞춰주는 샘플링 방법이다[6],[7]. 대표적으로 ROS, SMOTE, ADASYN 등이 있다. 언더샘플링은 데이터의 분포가 큰 값을 낮은 값으로 맞춰주는 작업을 거치는 것을 말한다[8],[9]. 대표적으로 RUS, ENN 등이 있다. 또한, 오버샘플링과 언더샘플링의 장점을 활용한 혼합샘플링이 있는데 SMOTE를 적용하여 오버샘플링 한 이후 ENN을 이용하여 언더샘플링 하는 방법(SMOTE+ENN) 등이 있다. 셋째, 분류 알고리즘으로 의사결정 트리 기반의 Random Forest를 사용하였다.

우리는 본 연구를 통해 다음과 같은 기여(Contribute)를 하고자 한다.

첫째, 불균형 클래스 처리에 대한 명확한 효과를 도출하기 위해 클래스 불균형한 CICMalMem2022 데이터셋을 선택하였다. 이를 통해 클래스 불균형 처리가 머신러닝 성능 향상에 도움이 됨을 검증하였다. 둘째, 불균형 데이터 처리를 위한 샘플링 기법으로 기존의 6가지 방법을 사용하였으며, 성능 비교를 통해 가장 최적화된 샘플링 기법을 추출할

수 있다. 셋째, Binary Classification(이중 분류)와 Multi Classsification(다중 분류)를 통해 불균형 처리 및 머신러닝 모델 사용이 랜섬웨어탐지 성능 향상에 도움을 줌을 증명하였다.

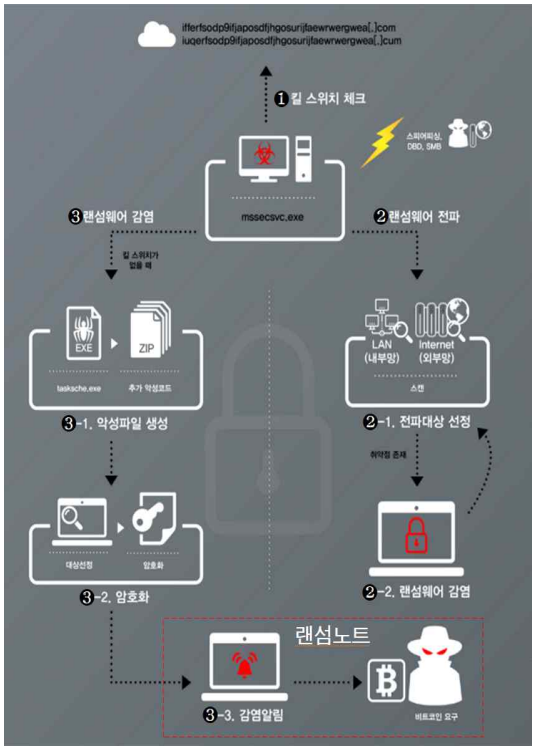
이 논문의 구성은 다음과 같다. 2장은 랜섬웨어 동작원리와 연구 관련 부분에 대해 설명한다. 3장에서는 샘플링 기법을 설명하고 접근방법을 제안한다. 4장에서는 이에 대한 실험 및 평가를 제공하고, 5장에서는 결론을 설명한다.

2. 관련 연구

2.1 랜섬웨어 동작방식

랜섬웨어는 시스템을 암호화하고 금전을 요구하는 악성 프로그램이다. 랜섬(몸값)과 소프트웨어의 합성어이다.

2017년 全 세계 30만대 이상의 컴퓨터에 피해를 입힌 워너크라이(WannaCry)를 통해 랜섬웨어의 동작방식을 파악하였다.



(그림 2) 랜섬웨어 동작방식[10]

(그림 2)와 같이 워너크라이 악성코드에 감염되면 킬 스위치(특정 도메인으로 연결되면 종료)를 체크하고, 동시에 SMB 취약점(이터널 블루, CVE-2017-0144)을 이용하여 랜섬웨어를 전파하며, 특정파일 암호화 및 랜섬노트(감염 공지 및 비트코인 요구)를 실행한다.

2.2 랜섬웨어탐지에 대한 관련 연구

랜섬웨어탐지 기술은 알려진 정보 기반, 미끼파일 기반, 랜섬웨어 행위 분석 기반, 네트워크 패킷 분석 기반, 머신러닝 기반기술 등이 있다.

연구에 사용되는 랩섬웨어탐지 기술은 머신러닝 기반이며 공개된 데이터셋을 찾기가 매우 어렵다. 그러나 CICMalMem2022 등과 같이 사용 가능한 데이터셋이 있으며 비교적 널리 사용된다.

2.3 머신러닝을 사용한 랜섬웨어 탐지에 대한 관련 연구

랜섬웨어탐지에 적용되는 머신러닝 기술은 지도학습, 비지도학습, 앙상블학습 등이 있다.

본 논문에서 사용하는 앙상블학습은 약 분류기들을 결합하여 강 분류기를 만드는 것으로 Bagging, Boosting 등으로 나눌 수 있다. Bagging은 여러개의 분류기를 만들어 Voting(투표)로 최종 결정하는 것으로 Random Forest가 이에 속한다. Random Forest는 의사결정기반(Decision Tree)의 예측 결정 알고리즘이다. Boosting은 잘못 예측한 데이터에 가중치를 부여해서 오류를 개선하는 것으로 AdaBoost(Adaptive Boosting), GBM(Gradient Boosting Machine), XGBoost(eXtra Gradient Boosting), LGBM(Light GBM), CatBoost(Categorical Boosting) 등이 있다. AdaBoost는 가중치를 높이면서 순차적으로 학습한다. GBM은 데이터별 오류를 예측하고, XGBoost는 과적합의 문제를 해결하였다. LGBM은 2016년 마이크로소프트에서 개발한 약 분류기를 수직적으로 확장한 것이고, CatBoost는 2017년 Yandex에서 개발한 범주형 변수를 처리하는데 중점을 둔 알고리즘이다.

샘플링과 머신러닝을 이용한 랜섬웨어탐지 연구는 <표 1>과 같이 진행되고 있다.

<표 1> 샘플링과 머신러닝을 이용한 랜섬웨어탐지 연구

년도	저 자	샘플링 또는 알고리즘
2003	Chawla et al.[6]	SOMTE
2005	Han et al.[11]	B-SMOTE
2009	Haibo et al.[12]	ADASYN
2019	Lachtar et al.[13]	SVM
2019	Alzahrani et al.[2]	XGBoost
2019	Scalas et al.[4]	Random Forest
2019	Singh et al.[14]	KNN
2020	Faris et al.[3]	AdaBoost
2021	Jung et al.[8]	LGBM

3. 혼합샘플링을 사용한 랜섬웨어탐지

3.1 탐지 성능 향상을 위한 샘플링 기법

사이버 공간에서는 Non-Ransomware가 대다수를 차지하므로 대부분의 실행 데이터는 Non-Ransomware이다. 랜섬웨어탐지는 다수의 Non-Ransomware 가운데서 소수의 Ransomware를 탐지하는 것이다. 이러한 소수의 랜섬웨어탐지는 매우 불균형한 데이터로 구성되어 있다.

이에 본 논문에서는 불균형 클래스로 구성된 랜섬웨어 데이터셋에 대해서 샘플링을 사용한 머신러닝 모델을 통해 탐지 성능을 향상시키고자 한다. 샘플링 기법은 크게 오버샘플링, 언더샘플링, 혼합샘플링으로 나눌 수 있으며, 데이터셋의 특성과 분포, 적용 알고리즘에 따라 머신러닝의 성능이 달라진다.

오버샘플링(OverSampling)은 분포가 작은 클래스 값을 분포가 큰 클래스로 맞춰주는 샘플링 방법이다[6],[7]. 대표적으로 ROS, SMOTE, ADASYN 등이 있다.

오버샘플링의 장점으로는 정보의 손실을 막을 수 있으며, 언더샘플링에 비해 높은 분류 정확도를 보인다. 하지만, 과적합(Overfitting) 문제가 발생할 수 있으며, 노이즈 또는 이상치에 민감한 편이다.

언더샘플링(UnderSampling)은 다운샘플링으로도 불리며, 데이터의 분포가 큰 값을 낮은 값으로 맞춰주는 작업을 거치는 것을 말한다[8],[9]. 대표적으로 RUS, ENN 등이 있다.

언더샘플링의 장점은 불필요한 데이터가 삭제되고, 유의미한 데이터만 남을 수 있다는 것이다. 데이터 크기가 줄어들어 메모리 사용이나 처리 속도 측면에서 시스템 부하를 적게 주어 유리하다.

하지만, 정보 유실의 문제가 생길 수 있으며, 학습에 사용되는 전체 데이터 수를 급격하게 감소시켜 성능이 저하될 수 있다.

또한, 오버샘플링과 언더샘플링의 장점을 활용한 혼합샘플링(Combine Sampling)이 있는데 SMOTE를 적용하여 오버샘플링 한 이후 ENN을 이용하여 언더샘플링 하는 방법(SMOTE+ENN) 등이 있다[8].

이러한 방법은 첫째, 학습 데이터셋의 불균형을 효과적으로 줄여 머신러닝 성능을 향상시킨다. 둘째, 다수의 클래스로 인해 분류되기 어려운 소수 클래스에 대해서 더 효과적으로 분류할 수 있는 모델을 만들게 된다. 셋째, 이러한 샘플링과 머신러닝 모델을 사용하여 랜섬웨어 탐지 성능을 향상시킨다.

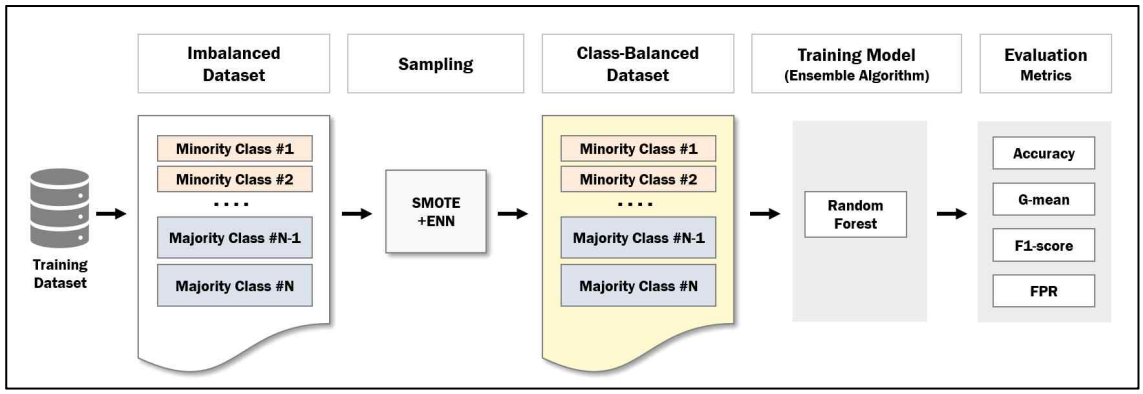
3.2 제안된 접근방법

이 장에서는 학습 데이터셋에 대해서 혼합샘플링 기법을 통해 랜섬웨어탐지 성능을 강화하는 방안을 제시한다. 랜섬웨어탐지 이벤트는 매우 불균형한 데이터 기반이며, 중복된 형태로 구성되어 있다.

머신러닝 알고리즘은 소수의 클래스에 대해서 완전히 학습할 수 없으며, 잘못된 분류를 하기 쉽다. 또한, 최근에는 배깅(Bagging)과 부스팅(Boosting) 기반의 알고리즘을 사용하는 사례가 증가하고 있다.

아래 (그림 3)은 혼합샘플링을 사용한 랜섬웨어탐지 프로세스를 설명한 것이다. 혼합샘플링을 사용한 랜섬웨어탐지 프로세스는 다음과 같다.

첫째, 랜섬웨어 데이터셋을 테스트와 훈련 데이터셋으로 분류한다. 둘째, 불균형 클래스로 구성된 훈련 데이터에 대해 랜섬웨어 유/무를 나타내는 Binary Class와 악성코드 유형별로 분류된 Multi Class로 구분한다. 셋째, 오버샘플링과 언더샘플링이 모두 섞인 혼합샘플링(SMOTE+ENN) 기법 수행을 통해 데이터를 생성한다. 넷째, 분류 알고리즘으로 Random Forest를 사용하여 가장 좋은 모델을 생성한다.



(그림 3) 혼합샘플링을 사용한 랜섬웨어탐지 프로세스

이렇게 랜섬웨어탐지 성능향상을 위해 혼합 샘플링과 머신러닝 모델을 사용할 것을 제안한다. 이 방법은 첫째, 학습 데이터셋의 불균형을 효과적으로 줄여 머신러닝 성능을 향상시킨다. 둘째, 다수의 클래스로 인해 분류되기 어려운 소수 클래스에 대해서 분류 성능이 향상되는 모델을 만들게 된다. 결국 랜섬웨어탐지 성능을 향상시키는 것이다.

이때 사용되는 평가지표로는 G-mean, F1-score 등을 사용한다.

4. 실험 및 평가

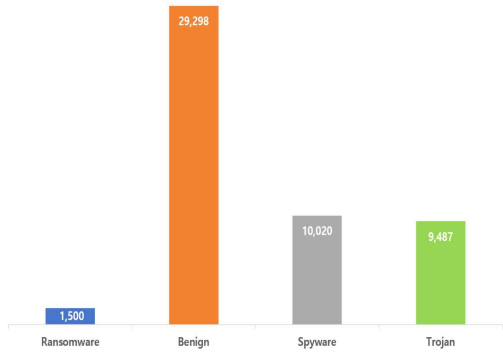
불균형한 랜섬웨어 데이터셋에서 샘플링 기법을 통해 머신러닝의 성능이 얼마나 개선될 수 있는지를 알아본다. 또한, 다음과 같은 물음을 해결하고자 한다. 첫째, 불균형한 랜섬웨어 데이터셋에 샘플링 기법을 통해 머신러닝 성능을 향상 시킬 수 있는가? 둘째, 샘플링 기법 중 가장 적절한 것은 무엇인가? 셋째, Multi Classification에서 다수 클래스의 탐지 성능을 유지하면서 소수의 클래스에 대한 탐지 성능을 높일 수 있을까?

본 실험을 위해 사용된 데이터셋과 샘플링 기법, 알고리즘, 평가방법은 다음과 같다. 실험 데이터셋으로 윈도우 환경 랜섬웨어탐지 데이터셋인 CIC MalMem2022를 활용하였다. 샘플링 기법으로 RUS, ENN, ROS, SMOTE, ADASYN, SMOTE +ENN을 사용하였다. 또한, 알고리즘으로는 Random Forest를 사용하였다. 평가 방법으로는 랜섬웨어탐지 성능과 데이터 불균형에 대한 평가를 위해 G-mean, F1-score를 사용하였다.

4.1 데이터셋

랜섬웨어탐지 분야의 다양한 공용 데이터셋 중 CIC MalMem2022을 사용하였다. 해당 데이터셋은 Non-Ransomware와 Ransomware로 Binary Classficication, 멀웨어 유형별로 Multi Classification을 구분하여 실험이 가능하며, 데이터 불균형이 크다. 또한, 데이터셋이 모두 전처리 된 데이터 형태로 구성되어 있다

CICMalMem2022 데이터셋은 2022년 캐나다 뉴브런즈윅대학교(UNB)의 사이버보안연구소(CIC)에서 제공하였다. (그림 4)와 같이 정상코드(Benign) 29,298개와 멀웨어(Malware) 19,507개 및 랜섬웨어 1,500개로 구성되었다. 멀웨어는 Spyware(10,020개), Trojan(9,487개)로 구성되었다.



(그림 4) CICMalMem2022 데이터셋 구성

4.2 평가 방법

평가 지표를 계산하기 위해 Confusion Matrix를 사용하였다.

Accuracy(정확도)는 Non-Ransomware와 Ransomware가 잘 분류되었는지의 비율로 정의된다. Precision(정밀도)는 랜섬웨어라고 예측한 것 중 실제 랜섬웨어라고 분류한 비율을 말한다. Recall(재현율, 민감도)는 실제 랜섬웨어 중 랜섬웨어라고 예측한 비율을 말한다.

이에 FPR, F1-score, G-mean을 추가하였다. FPR(오탐율, False Positive Rate)는 Non-Ransomware를 Ransomware라고 예측한 비율을 말한다. F1-score는 Precision과 Recall간의 조화평균을 의미한다. G-mean 값은 민감도와 특이도의 기하평균으로 계산된다[15],[16].

즉, 소수 클래스의 정확성이 높을수록 G-mean 값도 좋아지게 된다.

본 연구에서는 불균형에 대한 모델 성능에 대한 평가지표로서 G-mean, F1-score 를 사용하였다.

4.3 실험 결과

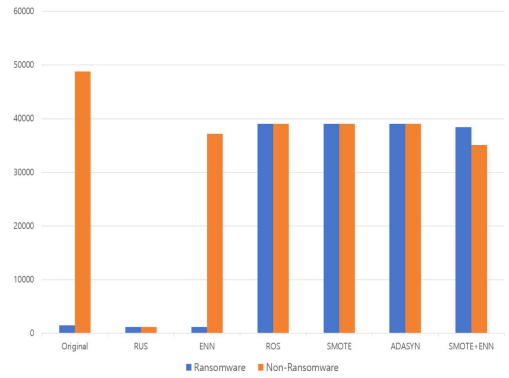
불균형한 클래스로 구성된 랜섬웨어 데이터셋에서 샘플링 기법을 통한 성능 향상을 나타내기 위해 Binary Classification 및 Multi Classification 실험을 수행하였다.

데이터셋의 레이블이 Ransomware, Non-Ransomware 두 가지로 분류된 Binary Classification 경우와 레이블이 Benign, Ransomware, Spyware, Trojan으로 분류된 Multi Classification의 경우이다. 위의 데이터셋(CICMalMem2022)에 대해서 샘플링을 적용하기 전인 original과 6가지의 샘플링 기법인(RUS, ENN, ROS, SMOTE, ADASYN, SMOTE+ENN)을 포함하여 클래스 불균형 처리 기술을 비교하였다.

불균형 처리 후 Random Forest 알고리즘을 사용한 머신러닝 모델을 적용하였다.

4.3.1 Binary Classification(이진 분류)

랜섬웨어탐지 데이터셋의 클래스가 Binary Classification인 경우이다.



(그림 5) 이진 분류에서 샘플링 적용 전/후 데이터 수

(그림 5)는 CICMalMem2022에 대한 샘플링 적용 전/후 데이터 수를 나타낸 결과이다.

Original은 샘플링 전 데이터 수이며, 언더샘플링(RUS)은 소수 클래스인 Ransomware 1,500건에 맞추어져 다운되었으며, 오버샘플링(ROS, SMOTE, ADASYN) 및 혼합샘플링(SMOTE+ENN)은 다수 클래스인 Non-Ransomware 48,805건에 근접하게 생겨났다. 특히, ENN의 경우 특성에 따라 데이터 수가 결정되었다.

<표 2>는 불균형이 심한 CICMalMem2022 데이터셋에 대해 샘플링 후 모델 성능 결과를 나타낸 것이다. 샘플링 기법으로는 RUS, ENN, ROS, SMOTE, ADASYN, SMOTE+ENN을 사용하였고 알고리즘으로는 Random Forest를 사용하였다. 평가지표는 Accuracy, Precision, Recall, F1-score, G-mean을 사용하였다.

<표 2> 샘플링 후 모델(Random Forest) 성능 결과

평가지표	Original	RUS	ENN	ROS	SMOTE	ADASYN	SMOTE+ENN
Accuracy	0.97465	0.80389	0.97356	0.97554	0.97097	0.97057	0.95328
Precision	0.74615	0.13047	0.62441	0.68341	0.54873	0.54275	0.35834
Recall	0.30407	0.91536	0.41692	0.42633	0.47648	0.45768	0.59874
F1-score	**0.43207**	0.22839	0.500000	**0.52509**	0.51006	0.49659	0.44835
G-mean	**0.55049**	0.85587	0.64304	0.65082	0.68583	0.67223	**0.76008**

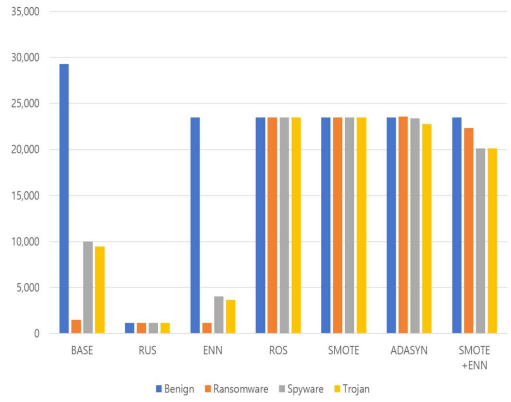
<표 2>와 같이 F1-score는 오버샘플링(ROS, SMOTE, ADASYN)과 혼합샘플링(SMOTE+ENN)에서 좋은 성능을 보이고 있다. 특히, ROS에서 20% 이상 성능이 향상되었다. G-mean은 혼합샘플링에서 좋은 성능을 보이고 있으며, SMOTE+ENN에서 30%이상 성능이 향상되었다. 결론적으로 불균형 처리는 데이터셋의 머신러닝 성능을 향상시킴을 확인할 수 있었다.

4.3.2 Multi Classification(다중 분류)

랜섬웨어탐지 데이터셋의 클래스가 Multi Classification인 경우이다. 본 실험에서의 레이블은 각 멀웨어 유형(Malware Type)이 된다. 클래스별 불균형이 심한 CICMalMem2022 데이터셋을 실험하였다. 샘플링 기법으로는 샘플링을 적용하기 전인 Base와 6가지의 샘플링 기술(RUS, ENN, ROS, SMOTE, ADASYN, SMOTE+ENN)을 적용하였다.

본 실험에서의 목적은 불균형한 클래스로 구성된 랜섬웨어탐지 데이터셋에서 다수 클래스의 탐지율을 유지하면서, 소수 클래스의 탐지율을 향상시키는 것이다.

(그림 6)을 통해 데이터셋을 보면 다수 클래스로 Benign, Spyware, Trojan을 확인할 수 있고, 소수 클래스로는 Ransomware를 확인할 수 있다.



(그림 6) 다중 분류에서 샘플링 적용 전/후 데이터 수

다음은 Random Forest 모델에 대한 평가이다. 해당 평가 지표인 G-mean, F1-score을 통해 확인할 수 있다.

<표 3> 샘플링 후 모델 성능 결과(G-mean)

Class	Base	RUS	ENN	ROS	SMOTE	ADASYN	SMOTE+ENN
Benign	0.99988	0.99889	0.99929	1.00000	1.00000	1.00000	1.00000
Ransomware	**0.53927**	0.82422	0.72473	0.63641	0.69994	0.69127	**0.74240**
Spyware	0.90222	0.80070	0.85714	0.90275	0.89657	0.90269	0.88628
Trojan	0.90049	0.80817	0.85985	0.90056	0.89322	0.89315	0.88308

<표 4> 샘플링 후 모델 성능 결과(F1-score)

Class	Base	RUS	ENN	ROS	SMOTE	ADASYN	SMOTE+ENN
Benign	0.99991	0.99914	0.99948	1.00000	1.00000	1.00000	1.00000
Ransomware	**0.42661**	0.41367	0.46299	0.51081	0.54202	0.53846	**0.53353**
Spyware	0.84699	0.73344	0.80302	0.85153	0.84527	0.84999	0.83506
Trojan	0.81685	0.71405	0.76731	0.82157	0.81631	0.81960	0.80587

<표 3> 및 <표 4>와 같이 Multi Classification의 경우, 소수 클래스인 Ransomware의 F1-Score, G-mean이 증가함을 알 수 있으며, 다수 클래스인 Benign, Spyware, Trojan의 탐지 성능이 유지됨을 알 수 있다.

랜섬웨어탐지 데이터셋에 대한 샘플링 기법과 모델링 알고리즘은 탐지성능을 향상시켰으며, 혼합 샘플링 기법인 SMOTE+ENN이 가장 좋은 결과를 도출하였다.

5. 결론

본 논문에서는 불균형한 랜섬웨어탐지 데이터셋에 대해서 혼합샘플링(SMOTE+ENN)과 머신러닝 모델을 통해 탐지성능을 향상시킬 수 있음을 확인하였다. 이는 다수의 Non-Ransomware에 대한 탐지율을 유지하면서, 소수의 랜섬웨어 탐지율을 향상시킬 수 있다는 측면에서 본 실험은 유용한 결과이다. 향후에는 B-SMOTE, TOMEK과 같은 샘플링 기법을 추가하고 LGBM, CatBoost 등 최신 Boosting 모델 연구를 수행하고자 한다[17],[18],[19].

참고문헌

[1] Cyber Crime Magazine, 2023,3, https://cybersecurityventures.com/global-ransomware-damage-costs-predicted-to-reach-250-billion-usd-by-2031

[2] A. Alzahrani, H. Alshahrani, A. Alshehri, and H. Fu, "An intelligent behavior-based ransomware detection system for Android platform," in Proc. 1st IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS ISA), Dec. 2019

[3] H. Faris, M. Habib, I. Almomani, M. Eshtay, and I. Aljarah, "Optimizing extreme learning machines using chains of salps for efficient Android ransomware detection," Appl. Sci., vol. 10, no. 11, p. 3706, May 2020.

[4] M. Scalas, D. Maiorca, F. Mercaldo, C. A. Visaggio, F. Martinelli, and G. Giacinto, "On the effectiveness of system API-related information for Android ransomware detection," Comput. Secur., vol. 86, pp. 168 - 182, Sep. 2019.

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

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

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

[8] I. Jung, J. Ji, C. Cho, EmSM: Ensemble Mixed Sampling Method for Classifying Imbalanced Intrusion Detection Data. Electronics 2022, 11, 1346. https://doi.org/10.3390/electronics11091346.

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

[10] KISA(한국인터넷진흥원), 워너크라이 분석 스페셜 리포트, 2017.10.13.

[11] H. HAN, W. Y. WANG, B. H. MAO, 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.

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

[13] N. Lachtar, D. Ibdah, and A. Bacha, "The case for native instructions in the detection of mobile ransomware," IEEE Lett. Comput. Soc., vol. 2, no. 2, pp. 16 - 19, Jun. 2019.

[14] A. K. Singh, G. Wadhwa, M. Ahuja, K. Soni, and K. Sharma, "Android malware detection using LSI-based reduced opcode feature vector," Procedia Comput. Sci., vol. 173, pp. 291 - 298, 2020.

[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] B. Yan, G. Han, M. Sun, and S. Ye, A Novel Region Adaptive SMOTE Algorithm f

or 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.

[18] 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.

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

〔저자소개〕

김 수 철 (Soo-chul Kim)
2008년 고려대학교 컴퓨터공학과 석사
2022년 숭실대학교 IT정책경영학과 박사수료
email : kscfuture@naver.com

이 형 동 (Hyung-Dong Lee)
1990년 서울시립대학교 전자공학과 학사
2017년 건국대학교 정보보호학과 석사
2021년 숭실대학교 IT정책경영학과 박사과정
email : hdtiger77@gmail.com

변 경 근 (Kyung-Keun Byun)
1996년 홍익대학교 컴퓨터공학과 학사
1998년 홍익대학교 전산학과 석사
2023년 숭실대학교 IT정책경영학과 박사과정
email : kkbyun@hanmail.net

신 용 태 (Yong-Tae Shin)
1985년 한양대학교 산업공학과 학사
1990년 Univ. of lowa, 컴퓨터학과 석사
1994년 Univ. of lowa, 컴퓨터학과 박사
1995년 숭실대학교 컴퓨터학부 교수
email : shin@ssu.ac.kr