오버샘플링 기법을 적용한 악성 패킷 분류 모델의 리콜 지표 최적화

김성일[1], 유헌창[2]
[1]고려대학교 빅데이터융합학과 석사과정
[2]고려대학교 정보대학 컴퓨터학과 교수
2023511036@korea.ac.kr, yuhc@korea.ac.kr

Optimization of Recall in Malicious Packet Classification Models Using Oversampling Techniques

Seongil Kim[1], Heonchang Yu[2]
[1]Dept. Big Data Convergence, Korea University
[2]Dept. of Computer Science & Engineering, Korea University

요 약

최근 사이버 공격의 지능화와 다양화로 인해 네트워크 보안의 중요성이 더욱 부각되고 있다. 특히, 악성코드를 포함한 악성 패킷은 시스템 감염 및 정보 유출과 같은 심각한 피해를 초래할 수 있으므로 이를 효과적으로 탐지하고 차단할 수 있는 기술 개발이 필수적이다. 기존의 인공지능 기반 침입 탐지 시스템은 다양한 성능 지표(정확도, 정밀도, 재현율 등)의 균형을 맞추기 위해 단일 분류 모델을 기반으로 구축되어 왔다. 본 연구에서는 모든 악성 패킷을 놓치지 않고 탐지하기 위해, 특히 리콜(Recall) 지표를 극대화하는 것을 목표로 하여 오버샘플링 기법을 적용하였다. 이를 통해 기존 시스템의 한계를 보완하고, 모든 사이버 공격에도 효과적으로 대응할 수 있는 새로운 성능 평가 기준의 필요성을 제시하고자 한다.

1. 서론

최근 사이버 보안 위협이 급증함에 따라 이를 방어하는 시스템의 중요성이 더욱 강조되고 있다. 사이버 공격은 갈수록 지능화되고 복잡해지며, 악성코드를 포함한 악성 패킷을 탐지하고 차단하지 못할 경우 정보 유출, 시스템 마비 등의 심각한 피해를 초래할 수 있다. 이러한 위협에 대응하기 위해 전 세계적으로 많은 자원과 투자가 사이버 보안 분야에 집중되고 있으며, 보안 전문가들은 보다 완벽한 방어 체계를 구축하는 데 초점을 맞추고 있다.

이처럼 사이버 보안에 대한 투자가 증가하면서, 이제는 보다 완벽한 방어 시스템을 설계하는 것이 필수적인 요소로 자리 잡고 있다. 기존 머신러닝 기반 침입 탐지 시스템에 적용되는 모델들은 정확도(Accuracy)를 최적화하는 데 주로 초점을 맞추어 개발되었다. 하지만 사이버 공격의 치명적 피해를 고려할 때, 단순한 정확도 향상만으로는 충분하지 못하며, 오탐(False Positive)을 다소 허용하더라도 악성 패킷을 100% 탐지하는 것을 목표로 하는 모델이 필요하다.

이러한 관점에서, 사이버 보안분야에서의 리콜(Recall) 지표는 악성 공격을 놓치지 않고 탐지하는 데 중요한 역할을 한다. 리콜을 극대화하는 시스템은 사이버 공격을 방어하는 데 있어 핵심적인 성능 지표로 작용하며, 이를 통해 인력과 비용이 추가되더라도 악성 패킷 탐지의 실패를 최소화하는 것이 목표가 되어야 한다.

따라서 본 연구는 리콜 지표의 향상을 최우선으로 하여 오버샘플링 기법을 적용해, 기존 시스템의 한계를 극복하고 보다 진보된 방어 체계를 위한 새로운 성능 평가 기준의 필요성을 제시하고자 한다.

2. 관련연구

리콜 지표는 모델의 성능 평가에서 중요한 요소 중 하나이며, 높은 정확성보다는 탐지해야 할 모든 객체를 빠짐없이 탐지하는 것이 더 중요한 분야에서 우선시되고 있다. 이러한 분야에서는 거짓 음성(False Negative)의 비용이나 위험이 매우 크기 때문에, 이를 최소화하는 것이 매우 중요하다.

높은 리콜값이 요구되는 검색 분야에서는 방대한

결과 집합에서 사용자의 검색값과 관련된 문서의 모든 집합을 찾아내는 것이 중요한데, 이는 특허, 법률, 의료문서와 관련된 검색 등에 해당한다. 특히, 기업에 적용하고자 하는 기술의 관련 특허를 검색하는 것은 대규모 투자 결정을 내리기 전에 매우 중요한 문제이며, 누락되거나 침해된 특허에 대한 대가를 치르는 것은 가늠할 수 없을 정도로 피해가 클 수 있다. 이러한 문제를 해결하기 위해, 정밀도를 희생하지 않으면서 높은 재현율 검색 작업에 적합한 동적 순위 검색 방법에 대한 연구가 진행되었다.[1]

암과 같은 중대한 질병을 조기 발견하는 진단 시스템에서도 리콜 기반 접근 방식에 중점을 두고 연구되고 있다.[2] 금융분야에서도 사기 행위를 탐지하지 못하는 경우 큰 금전적 손실로 이어질 수 있으므로, 리콜 향상 측면에서 신용카드 사기 탐지를 위한 개선된 알고리즘 연구가 진행되고 있다.[3]

한편, 기계학습에서 데이터 불균형(Data Imbalance)은 데이터셋 내에서 클래스 간의 분포가 불균등할 때 발생하며, 이는 분류작업에서 모델의 성능 저하로 이어질 수 있다. 이러한 문제를 해결하기 위한 가장 일반적인 방법은 리샘플링(resampling) 기법으로, 소수 클래스의 데이터를 증가시키는 오버샘플링(oversampling)과 다수 클래스의 데이터를 감소시키는 언더샘플링(undersampling)이 있다. Winkler et al.(2019)는 kaggle의 공개된 불균형 데이터셋 'Santander Customer Transaction Prediction'을 활용하여, 잘 알려진 기계학습 알고리즘들과 다양한 하이퍼파라미터를 적용하여 실험하였으며, 오버샘플링이 언더샘플링보다 다양한 분류기에서 더 나은 성능을 보임을 발표하였다.[4]

3. 분류모델과 오버샘플링 기법

3.1 이진분류모델

3.1.1 Logistic Regression

로지스틱 회귀는 이진 분류 문제를 해결하기 위한 선형 모델로, 입력 변수와 결과 변수 사이의 관계를 모델링하여 결과를 두 가지 범주 중 하나로 예측한다. 시그모이드 함수를 사용하여 출력값을 0과 1 사이의 확률로 변환하며, 로그 손실 함수를 통해 모델의 예측 오류를 최소화한다.

3.1.2 SVM

SVM은 분류와 회귀 분석에 사용되는 지도 학습 알고리즘으로, 데이터 포인트를 고차원 공간으로 매핑하여 두 클래스를 구분하는 최적의 초평면(Hyperplane)을 찾는다. 마진을 최대화하여 일반화 성능을 높이며, 커널 함수를 사용하여 비선형 분류 문제도 효과적으로 해결할 수 있다.

3.1.3 Decision Tree

Decision Tree(의사결정나무) 모델은 데이터의 특징을 기반으로 트리 구조의 분기점을 만들어 예측을 수행한다. 각 노드는 특정 특징에 대한 조건을 나타내며, 리프 노드에 도달하면 예측 결과를 제공한다. Decision Tree는 그 구조적 특성상 해석이 용이하고, 계산 속도가 빠르다는 장점이 있어 대규모 데이터의 처리에 적합하다. 따라서 다양한 조건에서 반복적인 테스트를 빠르게 수행할 수 있는 환경을 조성하기 위해, 본 연구에서는 Decision Tree를 분류모델로 적용하였다.

3.2 오버샘플링 기법

기존 데이터셋에서 타겟 클래스를 가지는 레코드들을 1%씩 단계적으로 오버샘플링하여 모델을 구축하였다. 각 오버샘플링 기법마다 최대 24%까지 오버샘플링을 적용하였으며, 이를 위해 각 기법당 24번씩 반복하여 오버샘플링하지 않은 원본 데이터 포함 적용, 총 100개의 모델을 생성하였다. 이렇게 생성된 모델들의 성능을 분석하고 그 결과를 상세히 기록하였다.

3.2.1 RandomOverSampler

RandomOverSampler는 타겟 클래스의 데이터를 무작위로 복제하여 다수 클래스와 데이터 균형을 맞추는 기법이다. 데이터의 분포를 고려하지 않고 단순 복제하는 방식이기 때문에 구현이 간단하지만, 과적합의 위험이 있을 수 있다. 주로 소수 클래스의 탐지 성능을 높이는 데 사용된다.

3.2.2 SMOTE

SMOTE는 타겟 클래스의 인접한 샘플들을 선형 보간하여 새로운 합성 데이터를 생성하는 기법이다. 무작위 복제가 아닌 인접 샘플 간의 관계를 고려하여 새로운 데이터를 만들기 때문에 과적합의 위험을 줄인다. 불균형 데이터 문제에서 소수 클래스의 다양성을 향상시키는 데 효과적이다.

3.2.3 ADASYN

ADASYN은 소수 클래스 중에서도 결정 경계에 가까워 잘못 분류되기 쉬운 어려운 샘플들에 더 많은 합성 데이터를 생성하는 방법이다. 데이터 분포의 복잡성을 반영하여, 잘 분류되지 않는 샘플을 집중적으로 보강하는 것이 특징이다. 이를 통해 소수 클래스의 탐지 성능을 보다 세밀하게 개선한다.

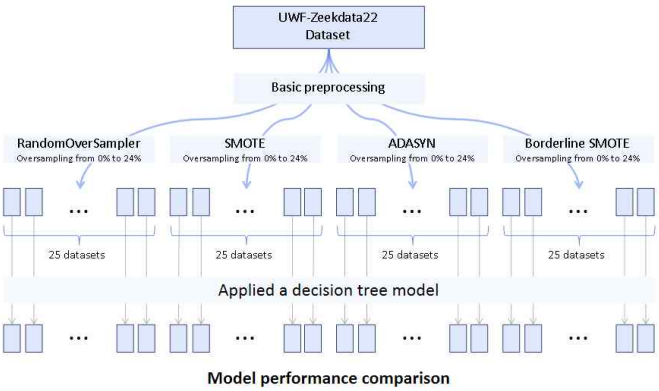
3.2.4 Borderline SMOTE

Borderline-SMOTE는 소수 클래스 중 결정 경계에 위치한 샘플을 집중적으로 오버샘플링하여 모델의 경계 학습을 강화하는 기법이다. 경계에 위치한 샘플들은 분류기가 오분류할 가능성이 높으므로, 이를 보강함으로써 성능을 개선할 수 있다. 주로 복잡한 경계가 존재하는 불균형 데이터에서 사용된다.

4. 리콜 지표 개선을 위한 오버샘플링 기법 적용

4.1 실험 순서

그림 1에 나타난 바와 같이, 하나의 데이터셋에 4가지 오버샘플링 기법을 0%부터 24%까지 적용하여 총 100개의 데이터셋으로 전처리하였다. 이후, 동일한 분류 모델을 적용해 오버샘플링 비율이 증가함에 따라 분류 모델 성능의 추이 변화를 확인하였다.



(그림 1) 실험 프로세스 다이어그램

4.2 데이터셋

본 연구에서 사용된 데이터셋은 UWF-ZeekData22로, 웨스트 플로리다 대학교(UWF)의 사이버 시뮬레이션 환경에서 생성되었다. 이 데이터셋은 사이버보안 분야의 연구 목적으로 공개된 데이터셋 중 가장 최근에 공개되었다. 본 데이터셋은 MITRE 공격 전술, 테크닉 및 일반 지식(ATT&CK) 프레임워크를 사용하여 라벨링되었다. MITRE ATT&CK 프레임워크는 14개의 전술과 다양한 테크닉 및 그 하위 카테고리의 테크닉을 포함하고 있다. UWF-ZeekData22는 12개의 전술을 포함하고 있으며, 표 1에 나타난 바와 같이 총 9,280,869개의 공격 전술 기록과 9,281,599개의 정상 기록을 가지고 있다.[5]

<표 1> 전술별 개수: UWF-ZeekData22

Tactic	Count
none	9281599
Reconnaissance	9278722
Discovery	2086
Credential Access	31
Privilege Escalation	13
Exfiltration	7
Lateral Movement	4
Resource Development	3
Persistence	1
Defense Evasion	1
Initial Access	1

4.3 전처리

본 연구의 핵심 목표는 오버샘플링 기법을 활용하여 리콜 지표를 극대화하는 것이다. 이를 보다 명확하게 분석하기 위해, 데이터 전처리 단계에서 불필요한 요소들을 제거하는 작업을 진행하였다. 특히, 저빈도 전술 레코드는 분류 작업에서 혼란을 야기할 수 있으므로 삭제하여, 정상 패킷과 Reconnaissance 패킷을 대상으로 이진 분류하도록 데이터 구성을 단순화하였다.

또한, 숫자형 데이터는 Min-Max 정규화를 통해 0과 1 사이로 스케일링하여 모델 학습 과정에서 각 변수의 중요도를 균형 있게 유지하였으며, 문자형 데이터는 Label Encoding을 통해 수치형으로 변환하여 처리했다.

타겟 클래스와 지나치게 높은 상관관계를 보이는 feature들을 사전에 제거함으로써 과적합을 방지하였다. 과적합은 모델이 학습 데이터에 지나치게 의존해 실제 예측 능력이 떨어지는 문제를 초래할 수 있다. 또한 이미 과적합된 모델을 통해 오버샘플링 기법을 적용했을 때, 그 효과를 명확하게 확인할 수 없는 문제가 생긴다. 따라서 앞선 전처리 과정을 통해 모델의 일반화 능력을 향상시켰으며, 결과적으로 2.6G에서 1.5G로 데이터의 용량을 축소시켜 연구에 사용하였다.

5. 실험 결과

특정 클래스를 분류하기 위해 오버샘플링을 적용했을 때, 리콜값 향상에 도움이 되는 모습을 확인할

수 있다.

<표 2> 오버샘플링 기법별 성능지표값

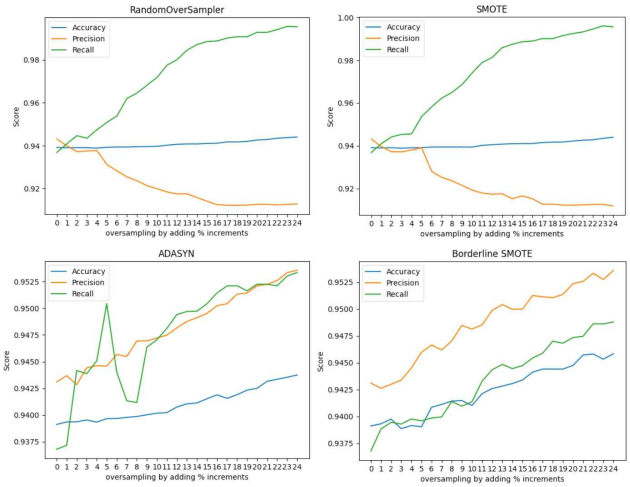
		0	1	2	3	...	12	13	...	21	22	23	24
RandomOverSampler	Accuracy	0.939	0.939	0.939	0.939		0.941	0.941		0.943	0.943	0.944	0.944
	Precision	0.943	0.940	0.937	0.935		0.917	0.916		0.913	0.912	0.913	0.913
	Recall	0.937	0.941	0.945	0.947		0.977	0.980		0.993	0.994	0.995	0.995
SMOTE	Accuracy	0.939	0.939	0.939	0.939		0.941	0.941		0.943	0.943	0.944	0.944
	Precision	0.943	0.940	0.937	0.935		0.917	0.916		0.912	0.912	0.913	0.913
	Recall	0.937	0.941	0.945	0.948		0.977	0.980		0.993	0.994	0.995	0.996
ADASYN	Accuracy	0.939	0.939	0.940	0.940		0.942	0.943		0.945	0.945	0.945	0.945
	Precision	0.943	0.944	0.943	0.944		0.948	0.949		0.952	0.953	0.953	0.953
	Recall	0.937	0.937	0.939	0.939		0.945	0.944		0.948	0.948	0.949	0.950
Borderline SMOTE	Accuracy	0.939	0.939	0.940	0.940		0.943	0.943		0.945	0.945	0.945	0.946
	Precision	0.943	0.943	0.943	0.944		0.949	0.950		0.953	0.953	0.954	0.954
	Recall	0.937	0.938	0.939	0.939		0.944	0.944		0.948	0.949	0.949	0.949

표 2를 확인한 결과, RandomOverSampler와 SMOTE에서 약 12% 오버샘플링 이후 리콜이 99%를 초과하며, 24% 오버샘플링에서 99.5%에 도달한다. RandomOverSampler에서 오버샘플링 비율이 증가함에 따라 정밀도가 점차적으로 감소하고, SMOTE에서는 초반에 빠르게 하락하고 이후 거의 일정하게 유지되는 것을 볼 수 있는데, 이는 거짓 양성(False Positive)의 증가로 정밀도가 떨어지는 것을 의미한다. 두 기법에서는 분류하고자 하는 데이터를 과도하게 증식하여 모델이 해당 클래스를 무분별하게 예측하게 만듦으로써 recall은 극대화하였지만, 정확도나 정밀도는 다소 저하되는 모습을 보인다.

그림 2에 나타난 바와 같이, RandomOverSampler와 SMOTE를 사용한 오버샘플링에서 리콜 값이 빠르게 99.5%에 도달한 반면, 정밀도는 91.2%까지 하락하는 경향을 보였다. 반면, ADASYN과 Borderline-SMOTE를 적용했을 때는 리콜 값이 각각 94.8%와 94.9%로 향상되었으나, 그 이후에는 추가적인 성능 향상이 나타나지 않았다. 그러나 이 두 기법을 통해 정확도와 정밀도 또한 함께 향상되는 모습을 확인할 수 있었다. 이는 두 기법이 결정 경계를 더 정확히 학습하도록 도와 과적합을 방지하고 일반화 능력을 향상시킨 것으로 보이며, 리콜값이 95%에서 멈춘 것은 모델이 데이터의 복잡성을 이미 충분히 학습하여 불필요한 추가 증식 없이 최적의 성능을 달성했음을 나타낸 것으로 보인다.

6. 결론

RandomOverSampler와 SMOTE는 리콜을 극대화하는 데 효과적이었으나, 정밀도가 다소 하락하는 경향을 보였다. 반면, ADASYN과 Borderline-SMOTE는 정밀도와 리콜값 간의 균형을 좀 더 잘 유지하며, 다양한 클래스 탐지에 유리한 결과를 나타냈다. 본 연구의 핵심 목표는 리콜값을 최대한 높인 후, 정확도나 정밀도와 같은 다른 성능 지표를 개선하는 것이었으며, RandomOverSampler와 SMOTE를 적



(그림 2) 오버샘플링 기법 별 성능지표 추이

용했을 때 리콜값이 100%에 가깝게 도달하는 것을 확인할 수 있었다. 이러한 이유로, 이 두 기법이 본 연구의 목적에 적합한 방법임을 확인할 수 있었고, 리콜 지표를 최대한 향상시킨 후 정확도나 정밀도 같은 성능 지표를 개선하는 새로운 성능 평가 기준을 수립하는 것이 필요하다는 점을 확인할 수 있었다.

참고문헌

[1] Song, J.J., Lee, W, Relevance maximization for high-recall retrieval problem: finding all needles in a haystack, J Supercomput, vol. 76, no. 7734 - 7757, 2020.

[2] A. Gupta, A. Anand and Y. Hasija, Recall-based Machine Learning approach for early detection of Cervical Cancer, 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2021, pp. 1-5

[3] Chung J, Lee K, Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression, Sensors, 23, 18, 7788, 2023.

[4] R. Mohammed, J. Rawashdeh and M. Abdullah, Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results, 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 243-248

[5] Bagui, S.S., Mink, D., Bagui, S.C., Ghosh, T., Plenkers, R., McElroy, T., Dulaney, S., & Shabanali, S, Introducing UWF-ZeekData22: A Comprehensive Network Traffic Dataset Based on the MITRE ATT&CK Framework, Data, 8, 1, 18, 2023.