논문 21-46-11-15 The Journal of Korean Institute of Communications and Information Sciences '21-11 Vol.46 No.11
https://doi.org/10.7840/kics.2021.46.11.1896

Tomeklinks와 ClusBUS 기법을 사용한 UNSW-NB15 데이터셋 유해 트래픽 분류

윤 필 도◆, 황 경 호°

Malicious Traffic Classification in a UNSW-NB15 Dataset by Using Tomeklinks and ClusBUS

Pil-Do Yoon◆, Gyung-Ho Hwang°

요 약

UNSW-NB15 데이터셋은 9가지 공격 유형과 정상 유형에 대한 42개의 피처로 구성된 트래픽 정보를 포함하고 있다. 본 논문에서는 데이터셋 내에서 Exploits, Fuzzers, Generic, Normal의 데이터 유형에 대한 분류를 진행하였다. 4가지 유형 중 Fuzzers와 Normal에서 데이터 중복과 데이터 불균형을 확인하였고, Tomeklinks와 ClusBUS 기법을 통해 데이터 중복 문제와 불균형 문제를 해결하여 예측 성능을 높였다.

Key Words : Classification, Machine Learning, Data overlap, Data Imbalance, Malicious traffic

ABSTRACT

The UNSW-NB15 datasets have traffic information consisting of 42 features for 9 attack types and normal type. We classified 'Exploits', 'Fuzzers', 'Generic', and 'Normal' data types in the datasets. It was confirmed that there were data overlap and imbalance problems in Fuzzers and Normal types. We showed performance enhancement by using Tomeklinks and ClusBUS techniques to solve data overlap and imbalance problems.

Ⅰ. 서 론

네트워크를 통한 서비스 공격과 개인정보 침해 사고가 늘어나고 있어 침입 공격을 탐지하는 시스템 구축이 중요하다. 의도적으로 많은 패킷을 전송하여 서비스를 마비시키는 DoS(Denial of Service) 공격이나 운영체제나 시스템의 취약점을 찾기 위한 Fuzzzers 등의 공격 시도를 초기에 파악하여 대응해야 한다. 본 논문에서는 실제 네트워크에서 발생할 수 있는 9개의 공격 유형과 1개의 정상 유형을 가진 UNSW-NB15 데이터셋을 사용하여 유해 트래픽 여부를 판단하는 모델을 제안한다. 데이터셋은 42개의 피처로 구성되어 있고 본 논문에서는 Exploits, Fuzzers, Generic, Normal의 4가지 유형에 대한 분류를 진행한다. 이는 유형에 따라 데이터가 많은 상위 4개를 선택하였고, 향후 전체 유형에 대한 분류에 적용될 기법을 고안하기 위해 선택되었다. 이를 위해 정제 과정과 레이블 과정을 거쳐 제공된 학습용 데이터셋 147,577개와 테스트용 데이터셋 73,065개를 사용한다. 모델 학습 시에는 학습용 데이터셋 중 학습을 위해 80%, 검증을 위해 20%의 데이터셋을 사용하며 이 중, Fuzzers와 Normal 유형에서 데이터 중복과 불균형이 있음을 확인하고, 해결하는 방안을 제시한다.

Ⅱ. 관련 연구

2.1 앙상블 기법을 사용한 UNSW-NB15 데이터셋 에 대한 분류 모델

[1]에서 UNSW-NB15 데이터셋의 모든 유형에 대한 분류를 진행했다. 분류를 진행한 데이터 유형은 Analysis, Backdoor, DoS, Exploits, Fuzzers, Generic, Normal, Reconnaissance, Shellcode, Worms으로, 보팅 앙상블 기법을 사용하였다. 보팅 앙상블 모델을 구성하는 개별 모델로는 Logistic Regression, Random Forest Classifier를 이용하였고, 보팅 기법으로는 소프트 보팅을 사용하였다. 모델 학습을 위한 데이터 전처리 기법으로는 피처 중요도(Feature Importance)를 사용한 피처 선택과 LDA(Linear Discriminant Analysis)를 사용한 차원 축소를 사용하

※ 이 논문은 2021학년도 한밭대학교 교내학술연구비의 지원을 받았음
◆ First Author : (ORCID:0000-0003-4972-5937)Dept. Computer Engineering, Hanbat National University, yoonpd56@gmail.com, 학생(학사), 학생회원
° Corresponding Author : (ORCID:0000-0001-6795-8086)Dept. Computer Engineering, Hanbat National University, gabriel@hanbat.ac.kr, 정교수, 종신회원
논문번호 : 202107-155-B-LU, Received July 6, 2021; Revised August 6, 2021; Accepted August 13, 2021

였다. 이후, 모델의 하이퍼파라미터와 가중치 튜닝을 위해 Optuna를 사용하였다. 모델의 성능으로는 재현율 0.79, F1 스코어 0.80을 기록하였다.

2.2 데이터 중복과 불균형 문제를 해결하기 위한 ClusBUS 기법

ClusBUS는 데이터 중복 문제와 불균형 문제를 해결하는 기법이다.[2] ClusBUS는 다수-소수 클래스에서 다수 클래스를 삭제하는 언더 샘플링 기법을 기반으로, 클러스터링을 사용하여 다수 클래스의 데이터를 삭제하고 최종적으로 소수 클래스에 더 많은 중점을 둘 수 있도록 하는 기법이다.

ClusBUS 기법에서 사용되는 값들은 원본 데이터, 원본 데이터에서 특정 클러스터 기법을 통해 만들어진 클러스터 목록, 그리고 각 클러스터 내의 소수 클래스 데이터의 점유율이다. 소수 클래스 데이터의 점유율은 각 클러스터 내의 데이터 개수 중 소수 클래스 데이터의 비율을 의미한다. ClusBUS 기법은 임계치를 기반으로 클러스터 내의 다수 클래스의 데이터를 삭제하는 방식이다. 각 클러스터의 소수 클래스 데이터의 점유율이 정해진 임계치 이상일 경우 해당 클러스터 내의 다수 클래스의 데이터를 삭제한다. ClusBUS는 미리 몇 개의 클러스터로 나눌지 정해둔 파티셔닝 기반의 클러스터링 알고리즘을 적용해서는 안 되므로 본 논문에서는 밀도 기반의 클러스터링 기법인 Optics를 사용하였다.

2.3 UNSW-NB15 데이터셋 분석 및 시각화

[3]에서는 UNSW-NB15 데이터셋에 대한 시각화와 데이터 분석을 진행하였다. 시각화와 분석 시 stratified sampling 기법을 사용하여 전체 중 20%의 데이터셋을 사용하였다. 또한, t-SNE (t-distributed Stochastic Neighbor Embedding)를 비롯한 시각화 기법과 PCA, K-Means 알고리즘 등의 데이터 전처리 기법을 사용하여 데이터셋은 불균형 문제를 갖고 있으며, 특정 클래스에서는 데이터 중복 문제가 있음을 알 수 있었다.

Ⅲ. 제안하는 유해 트래픽 분류 방안

본 논문에서는 UNSW-NB15에서 Exploits, Fuzzers, Generic, Normal 데이터 유형을 분류하였고, 데이터 불균형과 중복 문제가 있는 Fuzzers와 Normal에 대한 재분류를 진행하였다. 데이터 불균형과 중복은 모델의 정확한 학습과 예측에 악영향을 끼쳐 과적합, 모델의 복잡성 증가 등의 문제를 일으키며 실제 침입 모니터링 시스템에 도입되었을 때 정상 트래픽을 차단하거나 유해 트래픽을 허용하여 시스템 장애를 불러올 수 있다. 그림 1은 데이터셋 내 Fuzzers와 Normal에서 데이터 중복 문제를 시각화한 것으로 t-SNE 기법을 활용하여 시각화 시 고차원의 정보를 저차원으로 축소하며 고차원에서의 거리 정보를 최대한 유지하였다. 제안된 방식은 데이터셋에 대한 전처리, Tomeklinks와 ClusBUS를 사용한 데이터 불균형과 중복 문제 해결, 재예측을 진행한다. 제안하는 유해 트래픽 분류 방안은 그림 2와 같다.

데이터 전처리를 위해 상수형 데이터에 대한 로그 변환을 적용했다. 로그 변환은 값의 분포가 비대칭성을 갖고 있을 때, 정규 분포 형태를 가질 수 있도록 변환하는 방식으로 각 값에 대해 로그를 취하는 방식으

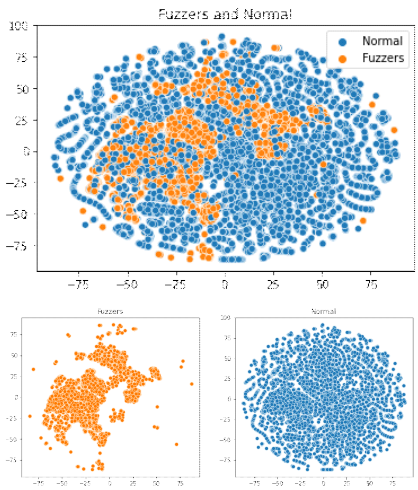

그림 1. Fuzzers와 Normal에서의 데이터 중복 시각화
Fig. 1. Visualization of data overlap problem in Fuzzers and Normal

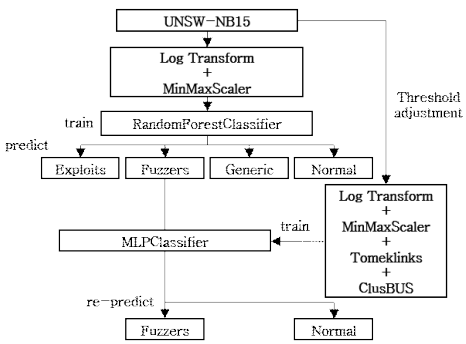

그림 2. 제안하는 유해 트래픽 분류 알고리즘
Fig. 2. Proposed malicious traffic classification

로 진행된다. 이후, 정규화를 위해 MinMaxScaler를 적용했다. 해당 데이터셋을 기반으로 4개의 공격 유형을 분류하기 위한 Random Forest Classifier 분류기를 학습시켰다. Random Forest Classifier는 Bootstrap 방식의 학습을 통해 예측에 대한 변동성과 과적합을 방지할 수 있다.

Fuzzers와 Normal을 재분류하기 위한 모델은 MLP Classifier를 사용하였다. 해당 분류기는 학습을 통해 고급 특징을 추출할 수 있고, 하이퍼 파라미터 튜닝을 통해 타 분류기에 비해 뛰어난 성능을 얻을 수 있지만, 학습에 오랜 시간이 소모될 수 있기에 전체 공격 유형이 아닌 Fuzzers와 Normal 분류에 사용하였다. 데이터셋 전처리를 위해 피처별 상관관계를 기반으로 한 피처 선택과 Tomeklinks를 사용한 전처리를 추가로 진행하였다. 타겟 값과 각 피처들의 상관관계를 계산하여 상관계수가 -0.1에서 0.1까지의 값을 갖는 피처를 삭제하여 최종적으로 dur, service, state, spkts 등의 20개 피처를 선택하였다. 선택된 피처들에 먼저 Tomeklinks를 적용하였다. Tomeklinks는 언더 샘플링 기법에 기반을 둔 방식으로 각 클래스의 경계선 부분에 다수 클래스와 소수 클래스의 쌍을 만들고 생성된 쌍에서 다수 클래스의 데이터를 삭제한다. 그다음으로 ClusBUS 기법을 적용하였다. ClusBUS 기법에서 임계치를 정하기 위해 다음과 같은 방법을 사용하였다. 먼저, 데이터셋을 두 가지로 구분하였다. 이는 Random Forest Classifier의 학습과 검증을 위해 로그 변환과 스케일을 적용한 데이터셋과 MLP Classifier의 학습과 검증을 위해 추가로 Tomeklinks와 ClusBUS를 적용한 데이터셋이다. 두 개의 데이터셋을 기반으로 Random Forest Classifier의 예측 결과 중 Fuzzers로 예측한 데이터에 대해서 MLP Classifier를 통해 재예측을 진행했다. MLP Classifier를 학습시키기 위해 ClusBUS를 적용할 때, 임계치 값을 0.2에서 0.75까지 0.05 단위로 증가시키며 위 과정을 진행하였고, 검증용 데이터셋에서 우수한 성능을 보인 0.4로 임계치 값을 정하였다. 이후 [1]에서 사용한 하이퍼파라미터 튜닝 프레임워크인 Optuna를 사용하여 Random Forest Classifier와 MLP Classifier에 대한 하이퍼파라미터 튜닝을 진행하였다. 표 1은 각 모델의 하이퍼파라미터 정보이다.

표 1. 각 모델의 하이퍼파라미터 정보
Table 1. Hyperparameter values for each model

Random Forest Classifier	
Hyperparameter	**Value**
n_estimators	100
max_depth	None
min_samples_split	2
MLP Classifier	
Hyperparameter	**Value**
hidden_layer_sizes	357
alpha	4.150012853868822e-5
activation	relu

Ⅳ. 성능 비교

Random Forest Classifier 모델과 MLP Classifier 모델을 학습한 뒤, 테스트 데이터셋을 통해 성능을 검사하였다. 그림 3은 Random Forest Classifier를 통한 4개 공격 유형 분류 방식과 MLP Classifier를 통해 재예측을 진행한 방식의 정밀도, 재현율, F1 점수를 비교한 것이다. 제안한 기법은 Random Forest Classifier를 통한 단일 예측과 비교했을 때, 정밀도는 유지하되 재현율과 F1 점수에서의 향상이 있었다. 재현율은 실제 레이블과 동일하게 예측을 성공한 비율로써, Fuzzers로 예측된 Normal을 재예측을 통하여 Normal로 정확히 예측한 비율이 높아짐을 알 수 있다. F1 점수는 정밀도와 재현율의 조화 평균으로, 재현율의 상승에 따라 상승되었음을 알 수 있다.

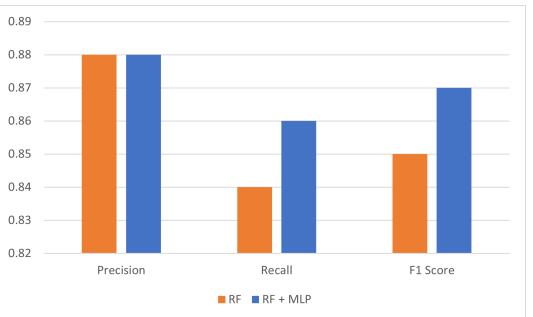

그림 3. 재예측 전후의 성능 비교
Fig. 3. Comparison of performance before and after re-prediction

Ⅴ. 결 론

본 논문에서는 UNSW-NB15 데이터셋 기반 Exploits, Fuzzers, Generic, Normal 데이터 유형에 대한 분류에서 Tomeklinks, ClusBUS 기법과 재예측을 통해 성능을 향상시켰다. 추후 제안한 기법을 다양한 공격 유형으로 확장할 예정이다.

References

[1] P. Yoon and G. Hwang, “Malicious traffic detection using ensemble learning based on UNSW-NB15 dataset,” in *Proc. Symp. KICS*, pp. 952-953, Feb. 2021.

[2] B. Das, N. C. Krishnan, and D. J. Cook, “Handling class overlap and imabalance to detect promt situations in smart homes,” *IEEE Int. Conf. Data Mining Wkshps.*, pp. 266-273, 2013.

[3] Z. Zoghi and G. Serpen, “UNSW-NB15 computer security dataset : Analysis through visualization,” in *Proc. Int. Conf. NCSET 2020*, 2020.