XGBoost 기반 침입탐지모델을 위한 데이터 스케일링 및 특성선택 기법 연구

김영원O, 이수진(교신저자)*
O국방대학교 국방과학학과 박사과정,
*국방대학교 국방과학학과 교수
e-mail: {kyoungwon21O, cyberkma*}@kndu.ac.kr

A study on data scaling and feature selection techniques for XGBoost-based intrusion detection model

Young-Won KimO, Soo-Jin Lee(Corresponding Author)*
ODept. of Defence Science, Korea National Defence University,
*Prof. Dept. of Defence Science, Korea National Defence University

● 요 약 ●

본 논문은 XGBoost 알고리즘 기반의 침입탐지모델의 성능을 향상하기 위한 스케일링(scaling) 및 특성선택(feature selection) 기법을 제안한다. 머신러닝 모델 개발 중 전처리 단계에서 스케일링 및 특성선택을 수행하면 데이터세트의 조건수가 감소하여 모델의 성능을 향상할 수 있다. 각 과정별로 다양한 기법이 있지만 기존의 연구에서는 이러한 기법들을 적용한 결과를 비교·분석하지 않고 특정 기법을 적용한 결과만을 나열하였고 스케일링 및 특성선택에 대해 최적의 조합은 제시하지 못하였다. 따라서 본 논문에서는 다양한 전처리 기법들의 적용결과를 비교하고 최적의 조합을 제안한다. 또한 기존의 연구들이 특정 데이터세트에만 적용 가능한 전처리 기법을 제안하는데 비해 본 논문은 다양한 데이터세트에 대해 공통적으로 적용 가능한 전처리 기법을 제안함으로써 제안 기법의 범용성과 실세계 적용 가능성을 증명한다.

키워드: 침입탐지(intrusion detection), 전처리(preprocessing), 스케일링(scaling), 특성선택(feature selection)

I. Introduction

머신러닝 모델 개발의 첫 번째 단계인 데이터 전처리는 원본데이터를 모델의 입력에 적합하도록 재가공하는 것이며 적용된 기법에 따라 모델의 성능에도 큰 영향을 미친다. 일반적으로 전처리는 결측치 처리 - 범주형 데이터 처리 - 이상치 제거 - 스케일링 - 특성선택 - 리샘플링의 순으로 진행되는데 이때 결측치 처리와 범주형 데이터 처리는 원본데이터를 모델에 입력할 수 있도록 형태를 재가공하는 것이고 이후 단계는 모델의 성능을 향상하기 위해 수행한다.

머신러닝 학습에서 조건수(condition number)는 입력값의 작은 변화에 대한 출력값의 변화량을 의미하며 조건수가 크면 입력값 사이의 차이가 작은 경우에도 오차(loss)값은 크게 달라지는 문제가 발생한다. 따라서 머신러닝에서는 데이터 전처리 단계에서 스케일링(scaling)을 통해 각 특성의 값을 특정 범위로 표준화 또는 정규화하고 특성선택(feature selection)을 이용하여 데이터세트의 차원을 감소시킨다.

본 논문은 XGBoost 기반의 침입탐지모델의 성능을 향상하기 위해 전처리 단계 중 조건수와 관련된 스케일링, 특성선택에 대해 다양한 기법들을 적용하여 결과를 비교함으로써 최적의 조합을 제안한다. 기존의 많은 연구들은 대부분 다양한 기법들간의 비교 없이 특정 기법의 적용 결과만을 단순 나열하고 있으며 전처리 과정 단계별 최적의 조합도 제시하지 못하고 있다. 또한 제안 방안을 특정 데이터세트에 한정하여 적용함으로써 해당 방법론이 다른 데이터세트 또는 실세계에서도 적용 가능한지는 알 수 없는 한계가 있다. 본 논문에서는 침입탐지와 관련된 연구에서 널리 사용되는 5개 데이터세트에 대해 공통적으로 성능향상을 기대할 수 있는 전처리 기법을 제안함으로써 이러한 한계를 극복한다.

본 논문의 특징은 다음과 같다.

① 침입탐지 데이터세트에 대하여 다양한 스케일링 및 특성선택 기법들을 적용한 결과를 비교한다.

② 최적의 스케일링 및 특성선택 조합을 제시한다.

③ 다양한 데이터세트를 대상으로 실험하여 제안기법의 범용성과 실세계 적용 가능성을 증명한다.

II. Preliminaries

1. 관련연구

1.1 스케일링(scaling)

데이터세트의 여러 특성들은 단위 차이로 인해 데이터의 분포가 서로 크게 다를 수 있다. 이 경우 모델은 데이터 분포가 좁은 특성값의 변화는 결과값을 예측하는데 영향이 적다고 판단하는 오류를 범한다. 따라서 전처리 단계에서 모든 특성의 분포를 동일하게 만들어야 한다. 다음은 전처리 단계에서 널리 쓰이는 스케일러(scaler)의 예이다.

Table 1. Comparison of Scalers

Scaler	Characteristic
StandardScaler	평균 0, 분산 1로 스케일링
RobustScaler	중앙값 0, IQR 1로 스케일링
MinMaxScaler	최소값 0, 최대값 1로 스케일링
MaxAbsScaler	절대값을 0~1로 스케일링
Quantile Transformer	0~1의 균등분포로 스케일링

스케일링은 우리말로 표준화, 정규화 등으로 번역되고 normalize 등의 용어와 혼용되기도 하지만 Normalization, Standardization, Regularization 등의 세부적인 전처리 기법과 혼동을 초래할 수 있어 본 논문에서는 스케일링의 원어를 그대로 사용한다.

침입탐지 관련 연구에서 스케일링은 전처리 과정의 일부로 자세한 설명 없이 특정 기법을 단순 적용하는 경우가 대다수이지만 ABDULAHEEM 등[1]이 QuatileScaler를 사용하여 희소데이터 분류성능을 향상한 사례가 있다.

1.2 특성선택(feature selection)

특성선택은 모델을 구성하기 위한 특성을 선택하는 과정이며 래퍼(wrapper)방식과 필터(filter)방식, 임베디드(Embedded)방식으로 구분된다[2].

Table 2. Comparison of feature selection Methods

Method	Characteristic
Wrapper	가능한 모든 특성 조합에 대해 특정 모델에 가장 적합한 조합을 선택
Filter	특성간의 상관관계를 확인하여 관계가 높은 특성중 하나를 제거
Embeded	모델의 정확도에 기여하는 특성을 선택

특성선택은 원본데이터의 전체 특성중 중요한 일부 특성만을 활용하므로 연산량을 줄일 수 있을 뿐 아니라 차원의 저주와 과적합을 방지하여 모델의 성능을 향상 할 수 있는 장점도 있다.

Kang 등[3]은 래퍼 방식인 다목적 유전자 알고리즘을 이용하여 선택된 최적의 특성 집합에 대해 필터 방식으로 피어슨 상관계수(pearson correlation coefficient)가 높은 특성 중 하나를 선택하는 방안을 제시한 바 있다.

2. 데이터세트(dataset)

침입탐지 데이터세트는 제안 방안의 성능을 검증하는 중요한 역할을 한다[4]. 상용제품에서 생산된 네트워크 트래픽 패킷은 개인정보보호 관련 문제로 접근이 어려우므로 대부분의 연구에서는 아래와 같이 공개적으로 사용할 수 있는 데이터세트를 사용한다.

Table 3. Comparison of Datasets

Dataset	Year
NSL-KDD	2009
ISCX 2012	2012
UNSW-NB15	2015
CIC-IDS2017	2017
CSE-CIC-IDS2018	2018

3. XGBoost

부스팅(boosting)은 여러개의 약한 결정트리를 조합해서 사용하는 앙상블(ensemble) 기법중의 하나이다. 그레디언트 부스트(gradient boost)가 대표적이며 이를 병렬학습이 지원되도록 구현한 라이브러리가 XGBoost이다. 표, csv 형식 등 정형화된 데이터세트에 대한 분류 및 회귀문제 해결 성능이 뛰어나 최근 침입탐지 연구분야에서도 널리 사용되므로[5-6] 본 연구에서도 XGBoost기반의 모델을 이용하여 실험을 진행한다.

III. The Proposed Scheme

1. 실험방법

본 연구에서는 앞 장에서 소개한 5가지 스케일링기법(Standard, Robust, MinMax, MaxAbs, Quantile)과 3가지 특성선택기법(Wrapper, Filter, Embeded)에 대해 가능한 모든 조합을 5개 데이터세트(NSL-KDD, ISCX 2012, UNSW-NB15, CIC-IDS2017, CSE-CIC-IDS2018)에 적용하고 XGBoost기반 모델을 이용한 분류결과를 비교한다. 데이터세트 중 원본 자체에 학습세트와 시험세트가 구분되어있지 않은 경우는 전체데이터를 8:2비율로 구분하여 학습세트와 시험세트로 사용하였고 특성 중 범주형데이터는 원-핫-인코딩(one-hot-encoding)으로 정수형데이터로 변환 후 실험하였다. 모든 실험은 Google Colab Pro+ 환경에서 GPU 가속을 사용하여 수행하였으며 pandas 및 sklearn 라이브러리를 사용하였다.

1.1 스케일링(scaling)

먼저 각 스케일러의 객체를 생성하고 sklearn preprocessing라이브러리의 fit()메서드를 각 데이터세트의 학습세트에 적용하면 스케일링을 위한 피팅값을 결정할 수 있다. 이어서 학습세트와 시험세트에 대하여 transform()메서드를 적용하여 스케일링을 완료한다. 이때 중요한 점은 학습세트에서 결정한 피팅값으로 학습세트와 시험세트 모두를 변환해야 한다는 것이다. fit()메서드를 시험세트에 적용하고 이 값을 이용할 경우 시험세트 데이터의 정보를 일부 이용하는 것이

되므로 주의하였다. 각 스케일링 기법은 다음과 같은 수식을 이용하여 데이터를 변환한다.

Table 4. Comparison of Scalers

Scaler	Equation
StandardScaler	$x' = \frac{x-\mu}{\sigma}$ μ : Mean, σ : Standard Devia
RobustScaler	$x' = \frac{x_i - x_{med}}{x_{75} - x_{25}}$
MinMaxScaler	$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$
MaxAbsScaler	$x' = \frac{x}{\max(\mid x \mid)}$
Quantile Transformer	1,000개의 quantile로 분포 후 0-1로 압축

1.2 특성선택(feature selection)

1.2.1 래퍼(wrapper)

래퍼 방식의 특성선택은 크게 전진선택, 후방제거, 단계별 선택의 세 가지로 분류된다. 본 연구에서는 이 중 후방제거 기법을 사용하였고 모든 특성을 가지고 시작하여 가장 덜 중요한 특성을 제거하면서 더 이상의 성능 향상이 없을 때까지 반복하였다.

1.2.2 필터(filter)

각 학습세트에서 특성들 간의 피어슨 상관계수를 계산하고 히트맵(heatmap)으로 시각화한다. 아래의 예시는 NSL-KDD의 학습세트에 대한 피어슨 상관관계 히트맵이며 흰색일수록 상관계수가 1에 가깝다.

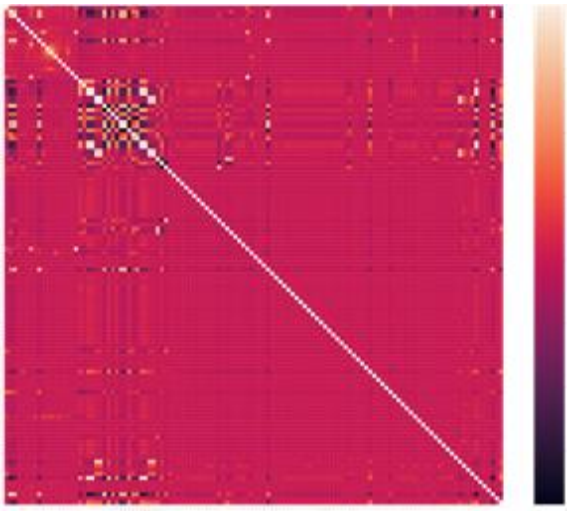

Fig. 1. Heatmap of feature corelation(NSL-KDD)

이어서 임계값(threshold)을 0.9로 설정하여 상관계수가 0.9 이상인 특성 쌍의 경우 둘 중 하나를 삭제한다.

1.2.3 임베디드(embedded)

sklearn 라이브러리의 SelectFromModel 모듈을 사용하여 모델의 정확도에 가장 큰 영향을 미치는 특성 집합을 선택하였다.

Table 5. Comparison of Experiment Results(f1-score)

Dataset	Feature Selection	Scaler					
		none	Standard	Robust	MinMax	MaxAbs	Quantile
NSL-KDD	none	0.730679	0.730679	0.730679	0.730679	0.730679	0.730500
	Wrapper	0.730679	0.730679	0.730679	0.730679	0.730679	0.738819
	Filter	0.739346	0.739346	0.739346	0.739346	0.739346	0.738819
	Embeded	0.736767	0.730679	0.730679	0.730679	0.730679	0.730679
ISCX 2012	none	0.946734	0.947512	0.947746	0.947746	0.947746	0.947050
	Wrapper	0.946222	0.947000	0.947230	0.947234	0.947231	0.940657
	Filter	0.953721	0.953884	0.953888	0.954564	0.954512	0.954380
	Embeded	0.946145	0.946678	0.946867	0.946678	0.946876	0.946567
UNSW-NB15	none	0.779430	0.779139	0.780504	0.780135	0.780135	0.780035
	Wrapper	0.778701	0.778640	0.778640	0.778438	0.778890	0.779425
	Filter	0.779636	0.780074	0.780227	0.781060	0.781060	0.780770
	Embeded	0.779704	0.779643	0.779643	0.779441	0.779893	0.779425
CIC-IDS2017	none	0.999058	0.999234	0.999423	0.999234	0.999423	0.999012
	Wrapper	0.969456	0.964554	0.962757	0.965644	0.967656	0.961423
	Filter	0.999678	0.999089	0.999798	0.999902	0.999809	0.999132
	Embeded	0.960354	0.962456	0.962712	0.965623	0.967622	0.961890
CSE-CIC-IDS 2018	none	0.983576	0.984787	0.984234	0.984534	0.984678	0.984564
	Wrapper	0.976786	0.977869	0.975674	0.975675	0.977896	0.970234
	Filter	0.990678	0.994576	0.994235	0.997242	0.994567	0.994890
	Embeded	0.974890	0.974513	0.974453	0.974811	0.974169	0.974197

1.3 분류모델(classification model)

전처리를 마친 각각의 학습세트로 sklearn wrapper를 이용하여 만든 XGBoost 분류모델을 학습한다. 이때 RandomizedSearchCV를 사용하여 n_estimators, colsamples_bytree, max_depth, min_child_weight, reg_alpha 등의 하리퍼파라미터를 튜닝하였고 GPU 가속을 위해 tree_method 하이퍼파라미터는 gpu_hist로 통일하였다.

2. 실험결과

침입탐지와 같이 레이블의 분포가 불균형한 경우에는 성능평가지표로 f1-score를 사용하면 모델의 성능을 정확하고 객관적으로 평가할 수 있다. 데이터세트별 스케일링과 특성선택 조합의 f1-score를 비교한 결과, 모든 데이터셋에서 필터기법으로 특성선택을 수행하였을 때 f1-score가 향상한 반면 래퍼기법과 임베디드기법은 f1-score가 소폭 하락하였다. 또한 모든데이터세트에 대해서 MinMaxScaler - 필터기법 특성선택 조합으로 전처리를 했을 때 모델의 분류성능이 가장 뛰어난 것을 확인하였다.

IV. Conclusions

본 논문은 XGBoost 기반 침입탐지 모델의 분류성능을 향상하기 위한 전처리기법에 대해 연구하였다. 데이터 전처리기법 중 조건수와 관련된 스케일링과 특성선택의 최적의 조합을 찾기 위해 실험을 진행하였고 제안기법의 실세계 적용 가능성을 확보하기 위하여 5개 데이터세트를 이용했다. 실험결과 MinMaxScaler로 스케일링후 필터기법으로 특성을 선택하는 경우 모든 데이터세트에 대하여 모델의 분류성능이 가장 뛰어난 것을 확인하였다. 향후에는 특성추출(feature extraction)등의 추가적인 전처리기법에 대해 연구할 예정이다.

REFERENCES

[1] Abdulraheem, Mohammed Hamid, and Najla Badie Ibraheem. "A detailed analysis of new intrusion detection dataset." Journal of Theoretical and Applied Information Technology 97.17 (2019): 4519-4537.

[2] Huang, Samuel H. "Supervised feature selection: A tutorial." Artif. Intell. Res. 4.2 (2015): 22-37.

[3] Kang Seung-Ho, Jeong In-Seon, Lim Hyeong-Seo. "A Feature Set Selection Approach Based on Pearson Correlation Coefficient for Real Time Attack Detection." Journal of convergence security 18.5 (2018): 59-66.

[4] Khraisat, Ansam, et al. "Survey of intrusion detection systems: techniques, datasets and challenges." Cybersecurity 2.1 (2019): 1-22.

[5] Bhati, Bhoopesh Singh, et al. "An improved ensemble based intrusion detection technique using XGBoost." Transactions on emerging telecommunications technologies 32.6 (2021): e4076.

[6] Dhaliwal, Sukhpreet Singh, Abdullah-Al Nahid, and Robert Abbas. "Effective intrusion detection system using XGBoost." Information 9.7 (2018): 149.