XGBoost, LightGBM, CatBoost 알고리즘을 활용한 침입탐지시스템 성능 비교 분석

진혜정, 이현우*

명지대학교, KENTECH*

okh1686@mju.ac.kr, hwlee@kentech.ac.kr*

Comparative Analysis of XGBoost, LightGBM, and CatBoost on Intrusion Detection Systems

Hyejeong Jin, Hyunwoo Lee*

Myungji Univ., KENTECH*.

요 약

본 연구는 XGBoost [1], LightGBM [2], CatBoost [3] 알고리즘을 사용하여 포트 스캐닝, 텔넷 사전 공격, UDP 플러딩 공격에 대한 침입탐지 성능을 비교 분석한다. 세 알고리즘은 모든 공격 패턴에 대해 높은 Precision을 보였으나, 공격 패턴에 따라 Recall이 차이가 많이 나는 것을 확인할 수 있었다. 실수형 특징이 분류에 중요한 역할을 하는 포트 스캐닝과 텔넷 사전 공격에서는 세 알고리즘이 유사한 성능을 보였으나, 카테고리형 특징이 분류에 중요한 역할을 하는 UDP 플러딩의 경우, LightGBM에서 낮은 성능을 보였다. 이는 결정트리 기반의 알고리즘들을 침입탐지에 사용할 때, Recall을 높이기 위한 보조 장치의 필요성과 공격 패턴의 특징에 따른 모델 선택의 중요성을 보여준다.

Ⅰ. 서 론

사이버 공격이 날로 증가함에 따라 침입탐지시스템(IDS)은 그 중요성이 더욱 부각되고 있다. 인공지능(Artificial Intelligence, AI)과 머신러닝(Machine Learning, ML) 기술이 다양한 산업 분야에서 혁신을 이끌어내는 흐름에 따라, IDS 영역에서도 AI/ML 기술 적용이 논의되고 있다. AI/ML을 활용한 침입탐지는 전통적인 규칙기반 탐지와 달리 모르는 공격도 탐지할 수 있다는 장점을 가진다 [4]. 이에 따라, 최신 연구들은 AI/ML 알고리즘들을 활용하여 침입 탐지 정확도를 크게 향상시키고, 보다 빠르고 정교한 보안을 제공하고자 하였다 [5].

본 연구는 XGBoost [1], LightGBM [2], CatBoost [3] 알고리즘을 미라이 봇넷 데이터셋 [6]에 적용하여 성능을 비교하고, 침입탐지에 가장 적합한 모델을 선정하고자 한다. 미라이봇넷 데이터셋은 다양한 IoT 장치에서 발생할 수 있는 공격을 모사한 데이터로, 침입탐지시스템의 성능을 평가하는 데 유용한 정형데이터를 제공한다. 이 데이터셋은 세 가지 유형의 사이버 공격을 포함하고 있어, 알고리즘들의 성능을 비교 분석하는 데 적합한 벤치마크 역할을 한다. 침입탐지시스템은 일반적으로 정형데이터를 처리하며, 이러한 정형데이터는 결정트리 기반 알고리즘에서 뛰어난 성능을 발휘한다고 알려져 있다 [7]. XGBoost, LightGBM, CatBoost와 같은 알고리즘은 중요한 특징을 빠르게 학습하여 효과적인 침입탐지를 가능하게 한다. 본 연구는 AI 기반 사이버 보안 기술의 발전에 기여하며, 보안 시스템 개발자와 IDS 연구자들에게 중요한 참고자료가 되고자 한다.

Ⅱ. 데이터셋과 알고리즘

본 연구에서 다루는 데이터셋에는 다음 세 가지 공격을 담고 있다.

- **포트 스캐닝 공격:** 포트 스캐닝은 네트워크나 시스템을 스캔하여 열려 있는 포트와 서비스를 찾는 공격이다. 이는 공격자가 시스템의 취약점을 파악하기 위한 단초로 활용된다.
- **텔넷 사전 공격:** 텔넷 사전 공격은 텔넷 프로토콜을 이용하여 대상 시스템에 침투하기 위해 사전에 정의한 여러 개의 ID와 비밀번호를 활용하여 접속을 시도하는 공격 방식이다.
- **UDP 플러딩 공격:** UDP 플러딩은 시스템에 대량의 UDP 패킷을 보내 시스템의 네트워크 자원을 고갈시키는 공격이다. 이 공격은 시스템의 성능 저하를 일으키며, 서비스 거부 공격(DoS)으로 분류된다.

본 연구는 세 가지 주요 공격 패턴에 대해 개선된 결정트리 알고리즘들을 적용하여 성능을 비교 분석한다. 이들의 근원이 되는 Gradient Boosting은 여러 결정트리를 결합하여 성능을 개선하는 앙상블 기법이다. XGBoost, LightGBM, CatBoost는 Gradient Boosting를 확장하여 빠르고 효율적인 학습, 과적합 방지, 범주형 데이터 처리 등에서 개선된 성능을 제공한다. 각각의 알고리즘은 아래와 같다.

- **XGBoost:** XGBoost (Extreme Gradient Boosting)는 병렬화와 분산 학습을 지원함으로써 Gradient Boosting 대비 학습 속도를 향상시키고 과적합 방지 기법(예: L1/L2 정규화, 조기 종료)을 내장하여 안정적인 성능을 제공한다. 특히 대규모 데이터셋에서 뛰어난 성능을 보인다.
- **LightGBM:** LightGBM (Light Gradient Boosting Machine)은 XGBoost와 유사하지만, 히스토그램 기반 학습으로 속도와 메모리의 효율성을 개선한 알고리즘이다. 데이터를 구간화하여 빠르고 효율적인 학습을 가능하게 하며, 리프 기반 학습을 통해 정교한 예측을 한다. 대규모 데이터셋에서 우수한 성능을 보인다.
- **CatBoost:** CatBoost (Categorical Boosting)는 범주형 데이터를 자동으로 처리할 수 있는 Gradient Boosting 알고리즘이며, XGBoost와 LightGBM보다 더 작은 수의 하이퍼파리미터를 가져서 보다 안정적인 성능을 보여준다. 범주형 데이터 처리에 강점을 가지며, 정확도와 속도가 뛰어나고 과적합 방지 기능도 내장되어 있다.

Ⅲ. 결정트리 알고리즘들의 비교 분석

결정트리 알고리즘들의 성능을 비교 분석하기 위한 실험은 미라이봇넷 훈련데이터로부터 41개의 플로우 기반 특징(예, 특정 시간당 평균 패킷 개수)을 추출하여 벡터화하고 [6], 이 벡터들을 바탕으로 공격 패턴 별로 XGBoost, LightGBM, CatBoost 모델을 생성한 뒤, 마찬가지로 테스트 데이터로부터 41개의 플로우 기반 특징을 추출한 벡터들에 대해 모델이 공

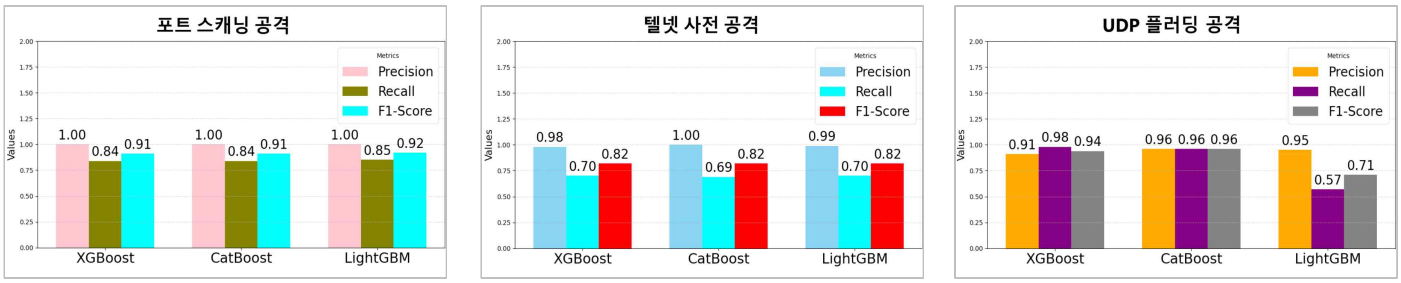

그림 1 공격 패턴에 따른 알고리즘의 성능 결과

알고리즘	상위 중요 특징
XGBoost	1) 패킷의 평균 길이, **2) 프로토콜 번호**, 3) 패킷의 최대 길이, 4) 패킷간 거리의 표준편차, 5) 패킷들의 전체 길이
CatBoost	**1) 프로토콜 번호**, 2) ACK 패킷 개수, 3) 패킷간 거리의 최소값, 4) 일방향 패킷간 거리의 최소값, 5) 패킷들의 전체 길이
LightGBM	1) 패킷 길이의 평균값, 2) 패킷간 거리의 최대값, 3) 패킷 길이의 최소값, 4) 전체 패킷 개수, 5) 패킷간 거리의 최소값

표 1 UDP 플러딩 탐지를 위해 생성된 모델 별 생성에 영향을 끼친 상위 5개의 중요 특징들 (위 특징들은 특정 기간 내에서 값 추출)

격 여부를 추론하게 하면서 진행하였다. 각 공격 패턴에 대해 알고리즘의 성능을 평가하기 위해, Precision, Recall, F1-score의 3 가지 성능 지표를 사용하였다. 모델의 성능의 원인을 분석하기 위해 각 모델 생성에 영향을 끼친 중요도 순서대로 특징들(이하 중요 특징)을 추출해 보았다. 그림 1은 공격 패턴에 따른 알고리즘의 성능 결과이며, 표 1은 UDP 플러딩 탐지에 대한 모델별 상위 5개의 중요 특징을 담은 것이다.

우리는 실험을 통해 다음 세 가지에 주목하였다.

첫째, 알고리즘들은 모든 공격에 대해 0.9 이상의 높은 Precision을 보였으며, Recall은 공격 패턴 별로 차이가 있었다. 이는 양성을 잘 분류해내기 위해 중요 특징들을 찾아내는 결정트리 알고리즘들의 특성 상 자연스런 결과이다. 텔넷 사전 공격의 Recall이 낮다는 것은 해당 공격이 다른 공격들에 비해 공격 여부를 잘 가르는 특징이 없다는 것으로 해석된다. 세 모델 모두 텔넷 사전 공격을 분류해내기 위해 패킷 길이의 최대 혹은 평균값 등 패킷 길이에 맞춰 분석하였는데, 텔넷 사전 공격용 패킷과 일반 텔넷 패킷의 길이가 매우 유사하였다는 것을 의미하며 약 30%의 양성 패킷들은 모델의 중요 특징에 의해 탐지가 배제된 것으로 보인다. 이는 결정트리 알고리즘들이 정탐을 하는데 유리한 반면에 Recall을 높이기 위해서는 다른 보조 장치가 필요하다는 것을 보여준다.

둘째, **포트 스캐닝**과 **텔넷 사전 공격**에서는 세 알고리즘 모두 뛰어난 성능을 보였으며, 특히나, 포트 스캐닝에 대한 F1-score는 약 0.91-0.92로 유사하게 높은 결과를 기록했다. 이는 세 알고리즘이 이 두 공격 패턴에서 비슷한 성능을 나타내며, 공격 탐지에서 모두 우수한 능력을 발휘했음을 시사한다. 포트 스캐닝의 경우, XGBoost와 CatBoost가 모든 지표에서 동일한 결과를 보였는데, 그 이유는 두 모델의 중요 특징들이 3개였고, 완전히 동일(SYN패킷 개수, 패킷 길이 최대값, 패킷 길이 최소값)하였기 때문이었다. 실수형 특징들이 중요한 역할을 하는 공격 유형에서는 XGBoost와 CatBoost가 유사한 동작을 하는 것으로 보인다.

셋째, **UDP 플러딩 공격**에서 알고리즘 별로 성능 차이가 뚜렷하게 나타났다. XGBoost와 CatBoost는 F1-score가 각각 0.94와 0.96으로 높은 성능을 보였으나, LightGBM은 0.71로 상대적으로 낮은 성능을 보였다. 이 원인을 분석하기 위해 세 가지 모델들의 상위 5개의 중요 특징들을 비교해 보았다. 프로토콜 번호를 첫 번째 혹은 두 번째 중요 특징으로 잡은 XGBoost와 CatBoost와 달리, LightGBM은 상위 5개의 특징에 프로토콜 번호가 들어가지 않았다 (표 1 참조). 본 데이터셋의 분포를 분석해본 결과, 데이터셋에 들어있는 UDP 패킷은 99% 이상이 UDP 플러딩 공격에 사용되어 있었으며, 따라서 프로토콜 번호는 UDP 플러딩 탐지에 매우 유용한 특징이었다. 프로토콜 번호는 대표적인 카테고리형 특징으로, 이러한 특징에 유용한 CatBoost는 이 특징을 잘 활용하였다. 반면에, 리프 기반 트리 방식인 LightGBM에서는 각 리프에서 가장 큰 성능 향상을 이끄는 특징을 우선적으로 선택하게 되며, '프로토콜 번호'와 같은 카테고리형 특징은 상대적으로 적은 정보 이득(information gain)을 제공하는 경우가 있어, 분할에서 우선 고려되지 않은 것으로 보인다 [6]. 패킷 데이터들을 다루는 특징에는 실수형과 카테고리형 특징들이 섞여 있기에 이 부분에 유의해야할 필요가 있어 보인다.

Ⅳ. 결론

본 연구에서 정형 데이터 기반의 미라이봇넷 데이터셋에 대한 XGBoost, CatBoost, LightGBM 모델을 활용한 침입탐지시스템 성능 비교가 이루어졌다. 실험을 통해 우리는 공격 패턴 별로 유용한 모델이 다를 수 있다는 것을 확인하였으며, 공격 패턴 별 서로 다른 모델을 적용하여 상호 보완적인 침입탐지시스템의 구축이 필요하다는 것을 보았다.

ACKNOWLEDGMENT

This work was supported by the KENTECH Research Grant (202200048A)

참 고 문 헌

[1] Tianqi Chen, Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System," arXiv preprint, arXiv:1603.02754, 2016.

[2] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 3146 - 3154.

[3] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin. "CatBoost: unbiased boosting with categorical features," arXiv preprint, arXiv:1706.09516, 2019.

[4] Hyunwoo Lee, Anand Mudgerikar, Ashish Kundu, Ninghui Li, Elisa Bertino. "An Infection-Identifying and Self-Evolving System for IoT Early Defense from Multi-Step Attacks," Computer Security - ESORICS 2022, Lecture Notes in Computer Science, vol. 13555, Springer, 2022, pp. 549 - 568.

[5] E. Bertino, S. Bhardwaj, F. Cicala, S. Gong, I. Karim. Machine Learning Techniques for Cybersecurity, Springer, 2023.

[6] "IoTEDef Project," GitHub Repository, (https://github.com/iotedef)

[7] Ammar Mohammed, Rania Kora. "A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 4, Feb. 2023, DOI:10.1016/j.jksuci.2023.01.014.