악성 사이트 탐지를 위한 설명가능한 인공지능(XAI) 기반 기계학습 특징 선별에 관한 연구

김홍비, 신삼신
한국인터넷진흥원
khb@kisa.or.kr, sss@kisa.or.kr

A Study on Explanable Artificial Intelligence (XAI)-based Machine Learning Feature Screening for Malicious Web Site Detection

HongBi Kim, SamShin Shin
Korea Internet & Security Agency(KISA)

요 약

악성 사이트를 통한 사이버 공격이 급증함에 따라 정확하고 신속한 대응을 위해 기계학습 기반 악성 사이트 탐지 기술 연구가 활발히 진행되고 있다. 그러나, 기계학습에 사용되는 악성 사이트 특징들이 수백개에 이르는 상황이며, 모든 특징을 이용하여 기계학습을 진행하는 것에는 한계가 존재한다. 이에 따라 본 논문에서는 악성 사이트 탐지를 위한 기계학습에서 사용되는 각 특징의 가중치를 측정하여 해석을 제공하는 설명 가능한 인공지능(XAI) 방법론에 기반하여 악성 사이트 탐지에서 중요한 특징을 선별한다. XAI 방법론을 통해 측정되는 가중치를 이용하여 각 특징의 기여도를 확인할 수 있으며, 기여도 점수를 활용하여 주요 특징을 선별하고, 비교 실험을 통해 XAI 기반의 기계학습 특징 선별의 유효성을 검증한다.

Ⅰ. 서 론

인터넷이 보편적으로 보급되기 시작함과 동시에 사용률이 급격하게 증가하였다. 빠른 정보 습득, 편리한 정보 공유, 다양한 IoT 기기들의 연결 등의 이점으로 인해 인터넷 사용률은 높아졌으며, 공격자들은 이러한 점을 악용하여 악성 사이트를 통한 공격을 수행하고 있다. 대표적으로 악성 URL을 이용해서 공격자가 의도한 특정 사이트로 연결시켜 바이러스나 악성 공격을 수행하도록 하는 공격이 발생하고 있으며[1], 한국인터넷진흥원의 악성코드 은닉사이트 탐지 동향 보고서에 따르면, 2021년 하반기 대비 2022년 상반기 악성 URL(유포지)이 38% 증가하였음을 확인할 수 있다[2]. 이에 따라 악성 URL 즉, 악성 사이트에 대응하기 위한 기술이 요구되고 있으며, 악성 사이트의 특징 정보를 분석하고 기계학습, 심층학습 등 인공지능 모델을 통한 탐지 기술 연구가 활발히 진행되고 있다. 그러나, 악성 사이트 탐지를 위해 공개된 특징 정보들만 수백개에 이르며, 모든 특징 정보를 이용한 기계학습 기반의 악성 사이트 탐지는 낮은 효율성, 저속 탐지 시간 등의 한계가 존재한다.

이에 따라, 본 논문에서는 악성 사이트 탐지에서 직면한 문제 상황을 해결하고자 설명 가능한 인공지능(XAI)를 이용하여 기계학습의 주요 특징을 선별한다. XAI를 통해 각 특징의 가중치를 측정함으로써 기계학습에 기여한 정도를 점수로 도출 가능하며, 도출된 점수에 기반하여 주요 특징을 선별한 후 비교 실험을 진행한다. 실험 결과, XAI를 통해 측정한 가중치 기반의 기여도가 주요 특징 선별에서 유의미하며, 선별된 특징만으로도 우수한 성능 달성이 가능함을 확인하였다.

Ⅱ. XAI(eXplainable Artificial Intelligence)

설명 가능한 인공지능 즉, XAI[3]는 사용자가 인공지능 시스템의 전반적인 강점 및 약점을 이해하도록 도와준다. 복잡성으로 인해 예측에 대한 판단 근거를 알 수 없었던 인공지능의 한계 극복을 위해 사용된 각 특징의 가중치를 측정하고 이에 기반한 해석을 제공한다. XAI 방법론에는 LRP(Layer-wise Relevance BackPropagation), LIME(Local Interpretable Model-Agnostic Explanations), SHAP(SHapley Additive exPlanations) 등이 있으며, 본 논문에서는 SHAP를 적용하여 악성 사이트 탐지를 위한 기계학습 특징을 선별한다.

SHAP[4]는 게임 이론을 바탕으로 하나의 특징에 대한 중요도를 알기 위해 여러 특징들의 조합을 구성하고 해당 특징의 유무에 따른 평적인 변화를 통해 얻어낸 값인 SHAP value를 통해 인공지능의 해석을 제공한다. SHAP는 다른 방법론과 달리 견고한 이론적 기반을 가지고 가중치를 측정하기 때문에 도출되는 값에 대한 높은 신뢰도를 가지고 있으며, SHAP value에 따라 각 특징의 영향력을 해석할 수 있다. 본 논문에서는 예측 해석을 위해 측정되는 SHAP value를 도출함으로써, 각 특징의 기여도를 확인하고, 영향력이 높은 특징을 선별하여 악성 사이트 탐지를 진행한다.

Ⅲ. XAI 기반 기계학습 특징 선별

본 논문에서는 기계학습 기반의 악성 사이트 탐지에서 사용되는 수많은 특징 중 주요 특징을 선별하기 위해 SHAP를 적용하고, 선별 전/후 및 선별 특징을 제외한 경우의 비교 실험을 통해 유효성을 검증한다.

3.1 데이터 세트 및 특징

악성 사이트 탐지 실험을 위해 오픈피시, 피시탱크, OSINT 사이트를 통해 약 1년동안('21년) 중복을 제거하여 확보된 악성/피싱 URL 30만개 데이터 세트를 활용한다[5]. 정상 및 악성 URL이 5:5의 비율로 구성되어 있으며, 전체 데이터 세트 중 80%는 학습으로, 20%는 테스트로 사용한다. 수집된 데이터 세트를 이용해 특징 추출을 진행하며, [5] 및 github 등에서 공개된 특징 조사를 통해 lexical/HTML/JS/Domain 기반의 특징을 추출하고, XAI 기반의 특징 선별의 유효성을 직관적으로 확인하기 위해 총

26개 특징을 사용하였다.

3.2 Tree 알고리즘 기반 악성 사이트 탐지

악성 사이트 탐지 분야에서는 다양한 기계학습 알고리즘이 적용되어왔으며, 우수한 성능이 증명되었다. 본 논문에서는 높은 성능을 나타내는 기계학습 알고리즘 중 tree 알고리즘을 기반으로 실험을 진행하며, tree 알고리즘에서도 가장 우수한 성능으로 악성 사이트를 탐지하는 XGBoost를 통해 최종적으로 XAI 기반 특징 선별 결과 비교 실험을 진행한다. 표 1은 세 가지의 tree 알고리즘 기반 악성 사이트 탐지 결과를 나타내며, XGBoost의 성능이 가장 우수함을 확인할 수 있다.

표 1 tree 알고리즘 기반 악성 사이트 탐지 결과

	XGBoost	LightGBM	RandomForest
Accuracy	96.95%	96.28%	90.28%
Precision	96.95%	96.33%	90.32%
Recall	96.95%	96.28%	90.28%
F1-score	96.95%	96.28%	90.28%

3.3 XAI 기반 특징 선별 및 결과

특징 선별을 위해 XGBoost 학습 모델과 특징이 추출된 데이터 세트를 입력으로 SHAP value를 도출한다. 표 2는 학습 데이터 세트에서 각 특징들의 가중치인 SHAP value를 평균내어 계산한 악성 사이트 예측 기여 점수를 나타내며, 값이 클수록 기여도가 높은 특징으로 해석이 가능하다. 해당 표를 기반으로 상위에 위치한 13개의 특징을 주요 특징으로 선별하여 XGBoost 학습 후 선별 전과의 비교를 진행한다.

표 2 특징별 SHAP value

Num	feature	SHAP value mean
1	alexa_top	1.033919
2	external_link_ratio	0.979367
3	ssl_remain_day	0.740422
4	html_lenth	0.715884
5	sharp_tag_link_ratio	0.707500
6	form_tag_link	0.664450
7	meta_script_link_tag_ratio	0.608668
8	whois_domain_expiration	0.462986
9	url_length	0.454397
10	having_https_protocol	0.418627
11	external_script_link_ratio	0.404173
12	open_port	0.374217
13	url_number_cnt_ratio	0.362291
…	…	…
…	…	…
…	…	…
26	slash_cnt	0.002110

선별된 특징은 alexa_top, external_link_ratio, ssl_remain_day 등 상위의 13개 특징이며, 표 3은 선별된 상위의 13개 특징만을 이용한 경우와 상위의 13개 특징을 제외한 나머지 하위 13개의 특징을 이용하여 기계학습 기반 악성 사이트를 탐지한 경우의 결과를 나타낸다. 선별된 상위의 13개 특징을 이용하였을 경우 악성 사이트 탐지 정확도가 96.59%로, 전체 특징을 이용하여 악성 사이트를 탐지한 결과인 96.95%와 유사하게 우수한 성능을 나타내는 것을 확인할 수 있다. 반대로, SHAP을 통해 영향력이 높다고 해석되는 상위의 13개 특징을 제외하여 하위 특징만을 사용하게 되면 악성 사이트 탐지 정확도가 71.80%로 성능이 현저히 떨어지는 것을 확인할 수 있다. 이를 통해 SHAP를 통해 산출되는 SHAP value를 사용하여 특징을 선별할 경우 악성 사이트 탐지에서 우수한 성능을 도출할 수 있으며, 수많은 특징 중 실제 크게 기여한 특징을 확인함으로써 기계학습에서 중요한 특징을 선별할 수 있음을 확인할 수 있다.

표 3 SHAP 기반 주요 특징 및 주요 특징 제외 악성 사이트 탐지 결과

	top rank feature	low rank feature
Accuracy	96.59%	71.80%
Precision	96.59%	73.10%
Recall	96.59%	71.80%
F1-score	96.59%	71.42%

Ⅳ. 결론

본 논문에서는 광범위적인 악성 사이트 특징 중 실제 기계학습에서 중요한 특징을 선별하기 위해 XAI 방법론 중 하나인 SHAP을 사용하여 기계학습 특징을 선별하였다. 선별된 특징을 이용하여 기계학습을 수행하고 특징 선별 전/후 및 선별 특징을 제외한 경우 비교를 통해 XAI 기반으로 선별된 특징만으로도 높은 정확도로 악성 사이트 탐지가 가능함을 확인하였으며, XAI를 통한 특징의 기여도 산출의 유의미성을 검증하였다. 기계학습 성능의 저해 요소 중 하나인 대량의 특징 정보를 본 논문과 같은 방식으로 선별함으로써 악성 사이트 탐지에 기여할 수 있을 것으로 예상되며, 향후 SHAP 뿐만 아니라 다른 XAI 기술들의 추가 연구를 통해 악성 사이트 탐지에서 가장 우수한 성능을 도출할 수 있는 특징 선별 연구를 진행할 예정이다.

ACKNOWLEDGMENT

본 연구는 대한민국 정부(산업통상자원부 및 방위사업청) 재원으로 민군협력진흥원에서 수행하는 민군기술협력사업의 연구 결과로 수행되었음 (협약번호 UM19204RD2)

참 고 문 헌

[1] H. Kwon, S.J. Park and Y.C. Kim. “Design of detection method for malicious URL based on Deep Neural Network,” Journal of Convergence for Information Technology, vol. 11, no. 5, pp. 30-37, 2021

[2] KISA, [Online], Available: https://www.krcert.or.kr/filedownload.do?attach_file_seq=3667&attach_file_id=EpF3667.pdf (downloaded 2022)

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (XAI): Concepts, tax- onomies, opportunities and challenges toward responsible AI,” Inf. Fusion, vol. 58, pp. 82 - 115, Jun. 2020

[4] LS. Shapley. “17. A value for n-person games.” Princeton University Press, 2016.

[5] S.S. Shin and S.G. Ji. “A Study on the Machine Learning Feature Selection for Malicious Site Detection”, 2022