CV 기반 악성 URL 탐지 앙상블 스태킹 모델

이종호[1], 신용태[2]
[1]숭실대학교 컴퓨터학과 석사과정
[2]숭실대학교 컴퓨터학부 교수
leejongho@soongsil.ac.kr, shin@ssu.ac.kr

CV-based malicious URL detection ensemble stacking model

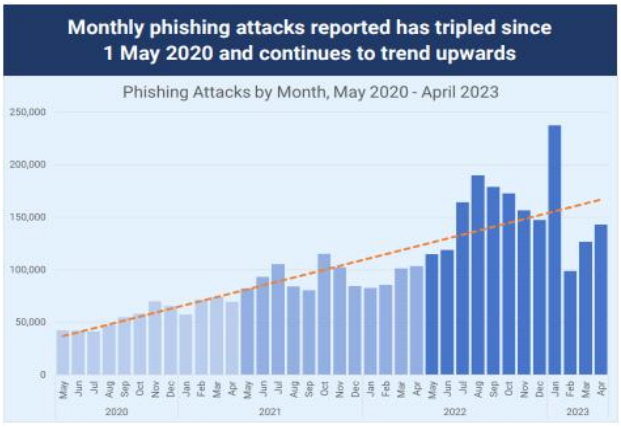
Jong-Ho Lee[1], Yong-Tae Shin[2]
[1]Dept. of Computer Science, Soong-Sil University
[2]Dept. of Computer Science, Soong-Sil University

요 약

다양한 분야에서 QR 코드가 급속도로 확산되면서, QR 코드를 악용하여 사용자를 악성 웹사이트로 리디렉션하는 '큐싱(Qshing)'이라는 새로운 형태의 사이버 범죄가 등장했다. 이에 본 연구에서는 일반화 성능을 향상시키기 위해 교차 검증(CV)을 활용하여 QR 코드 스캔과 관련된 악성 URL을 탐지하도록 설계된 스태킹 앙상블 모델을 제안한다. 이러한 통합은 실제 애플리케이션에서 높은 성능을 기대할 수 있도록 설계되었다. 본 연구는 이 모델이 기존의 연구보다 QR 코드 관련 사이버 위협에 대처하는 보다 효과적인 수단을 제공할 것으로 기대한다.

1. 서론

디지털 기술의 발전은 우리의 일상에 깊숙이 스며들었으며, 그 중심에는 QR(Quick Response) 코드가 있다. QR 코드는 속도와 효율성을 기반으로 다양한 분야에서 폭넓게 채택되고 코로나 19 사태가 시작된 이후 QR 코드의 사용량은 전 세계적으로 급격히 증가하였다.[1] 그러나 QR 코드의 보편화로 '큐싱(Qshing)'이라는 새로운 형태의 사이버 범죄가 생겨났다. 큐싱은 QR코드와 피싱(Fhishing)의 합성어로 악의적인 목적으로 조작된 QR 코드를 통해 사용자의 개인정보를 훔치거나, 악성 소프트웨어를 전파하는 행위를 말한다. 최근 국내에서 QR 코드를 이용한 큐싱 피해 사례가 증가하고 있다.[2] 범죄자들은 공유 자전거·킥보드의 공식 QR 코드 위에 가짜 QR 코드를 덧붙이거나, 주차장 위반 스티커와 과태료 고지서에도 가짜 QR 코드를 부착한다. 이러한 코드를 스캔하면 사용자는 악성 프로그램이 설치되거나 가짜 웹사이트로 유도되어 결제를 요구받는 피해를 입게 된다.[3] (그림 1)은 Interisle Consulting Group에서 발표한 악성 URL 공격을 포함한 피싱 공격 추이를 나타낸다.[4] 월별 피싱 공격 계속 상승하는 추세이며, 2020년 5월 이후 3배 이상 증가했다. QR 코드만으로 가짜 QR 코드를 식별하는 것은 큐싱 피해를 최소화하는 데 한계가 있다. 이에 따라, QR 코드를 스캔했을 때 접속되는 URL이 악성인지 여부를 판별하는 것이 필수적이다. 비록 머신러닝 및 딥러닝을 활용하여 악성 URL 탐지에 관한 많은 연구가 진행되었지만, 새롭게 등장하는 피싱 사이트에 대한 효과적 대응은 여전히 도전 과제로 남아 있다. 본 논문은 여러 머신러닝 모델을 결합한 앙상블 스태킹 방법을 통해 악성 URL 탐지의 정확도를 향상시키는 연구를 제안한다. 특히, 앙상블 스태킹 모델의 과적합 문제를 해결하고 일반화 능력을 강화하기 위해 교차 검증(Cross Validation)을 통합한다.[5]



(그림 1) 피싱 공격 추이

2. 관련 연구

2.1 앙상블 스태킹 모델

Stacking은 두 개 이상의 기본 모델로부터의 예측값을 결합하여 메타 모델에서 최종 예측을 도출하는 머신러닝 앙상블 기법이다. 이 기법은 여러 기본 모델의 예측을 상위 수준의 메타 모델로 공급한 후 이를 결합하여 최종 예측을 얻는다. Stacking은 개별 모델이 독립적이라고 가정하기 때문에, 이상치에 대한 대응력이 높아, 단일 모델의 오류율보다 낮은 값을 달성할 수 있다. Stacking은 Boosting이나 Begging과 같은 전통적인 앙상블 기법과는 달리, 서로 다른 특성을 가진 알고리즘들을 조합하여 사용합니다. 이러한 접근법은 각 데이터 모델의 독립적 가정을 활용, 각기 다른 강점을 결합함으로써 더욱 효과적인 결과를 도출할 수 있다. 따라서 Stacking에서는 앙상블의 각 구성 요소인 기본 모델과 메타 모델의 선택이 성공적인 분석 결과에 매우 중요하다.

2.4 k-fold CV(Cross Validation)

k-fold CV은 데이터 세트를 k의 fold로 나누고, 각 fold를 순차적으로 검증 데이터로 사용하는 방법이다. 나머지 k-1개의 fold는 학습에 활용된다. 이 과정을 k번 반복하여 모든 데이터가 한 번씩 검증에 사용된다.[6] 본 논문에서 제안한 앙상블 스태킹 모델의 과적합을 방지해주고. 일반화 능력을 향상시킨다.

2.3 kNN(k-Nearset-Neighbors) 알고리즘

kNN 알고리즘은 비모수적 지도 학습 분류기로, 근접성을 사용하여 개별 데이터 포인트의 그룹화에 대한 분류 또는 예측을 수행한다.[7] URL 길이, 특수문자와 같은 URL을 구성하는 문자열의 구조적 및 문법적 특징에서 우수한 성능을 보인다.[8]

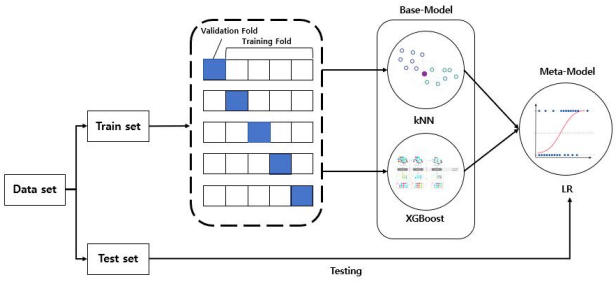
2.4 XGBoost(Extreme Gradient Boost)

XGBoost는 그래디언트 부스팅 알고리즘을 실행하기 위한 효율적이고, 유연하며, 이식 가능한 최적화된 분산 라이브러리다.[9] 이는 URL의 도메인 및 경로 기반 특성, 통신 프로토콜 특성 등 다양한 특성과 대규모 데이터셋에서 강력한 성능을 발휘할 수 있다.[10]

2.5 로지스틱 회귀 모델

로지스틱 회귀는 일반적으로 이진 분류 작업을 위한 예측 분석 모델로 활용되며, 여기서 로지스틱 함수를 사용하여 독립 변수의 선형 조합을 0과 1 사이의 확률 점수로 변환한다.[11] 다른 모델의 출력을 입력으로 사용하여 최종 예측 생성하고, 여러 모델의 예측을 효과적으로 결합하여 성능을 향상시킨다. 이는 최종 결정을 내리는 Meta-Model로 적합하다.

3. 제안



(그림 2) 악성 URL 탐지 모델 아키텍처.

그림은 제안하는 모델의 아키텍처를 나타내고 모델의 흐름은 다음과 같다.

3.1 데이터 수집 및 전처리

모델이 실제 상황을 반영하고 한쪽으로 편향되지 않도록 하기 위해, 정상 URL과 악성 URL의 비율을 60대 40으로 수집한다.[12] 이후 결측치, 이상치 데이터의 여부를 파악하여 수정 및 삭제하고 다음과 같은 주요 특징[13]을 추출한다.

1) HTTPS 여부: URL이 HTTPS 대신에 HTTP를 사용한다면 데이터를 가로채거나 조작하려는 시도가 있었다고 가정할 수 있다.

2) URL 길이: 긴 URL은 피싱 사이트라는 신호일 수도 있다.

3) WHOIS 정보: 도메인의 등록자 정보, 등록 날짜, 만료 날짜 등을 분석하여 악성 사이트를 식별하는 데 유용하다.

4) 특수문자 수: 특수문자가 지나치게 많은 URL은 신뢰할 수 사이트라고 보기 어렵다.

3.2 데이터 분할

추출된 특성을 포함하는 데이터셋 80%는 학습을 위한 학습 세트, 20%는 테스트를 위한 테스트 세트로 나눈다. 이는 다양한 특성과 패턴을 학습해 일반화 성능을 높이고, 모델의 과적합 방지 및 성능 평가를 위함이다.

3.3 앙상블 모델 학습

k-fold CV를 위해 학습 세트를 k개의 fold로 나눈다. 이때 각 fold는 학습과 검증에 순차적으로 사용된다. Base-model에서 각 fold의 k-1개를 학습 데이터로 사용하여 kNN 모델과 XGBoost 모델을 별도

로 학습한다. 그리고 남은 한 개의 폴드를 검증 데이터로 사용하여 두 모델의 예측값을 추출한다. k번의 fold마다 반복하여 모든 검증 데이터에 대한 예측값을 수집한다. 수집된 예측값들을 모아 새로운 학습 데이터셋을 구성한다. 새로운 학습 데이터셋을 사용하여 로지스틱 회귀 모델을 학습한다. 이 로지스틱 회귀 모델이 최종 Meta-Model이 된다.

3.4 모델 평가

최종 학습된 Meta-Model을 테스트 세트에 적용하여 모델의 성능을 평가한다.

본 연구에서 제안된 스태킹 앙상블 모델에는 kNN, XGBoost, 로지스틱 회귀 알고리즘이 포함되어 있다. 하이퍼파라미터 설정은 다음과 같다.

- kNN 알고리즘: 이웃의 수(k)는 초기에 5로 설정할 계획이다. 이 값은 향후 교차 검증을 통해 최적화 과정을 거쳐 최적의 k 값이 결정될 예정이다.
- XGBoost 알고리즘: 학습률(learning rate)은 0.01로, 트리의 최대 깊이(max_depth)는 6으로 설정하여 과적합을 방지하고 일반화 능력을 향상시키는 것을 목표로 한다. n_estimators는 100으로 설정하여 충분한 수의 트리가 생성되도록 한다.
- 로지스틱 회귀: 규제 파라미터(C)는 1.0으로 설정하여 학습 데이터에 대한 과적합을 최소화한다. 규제 유형으로는 L2 규제를 적용하여 가중치의 제곱합을 최소화할 계획이다.

하이퍼파라미터의 최종 선택과 설정은 교차 검증을 통해 이루어질 예정이며, 이 과정은 모델의 성능을 최적화하는 데 중요한 역할을 할 것이다. 이러한 접근을 통해 제안된 모델이 높은 정확도와 안정성을 달성할 수 있도록 할 것이다.

4. 결론

본 논문에서는 '큐싱'이라고 불리는 새로운 형태의 사이버 범죄에 대응하기 위해, 교차 검증을 통합한 스태킹 앙상블 모델을 제안했다. 이 모델은 다양한 머신러닝 알고리즘을 통합함으로써 각 알고리즘의 특성을 종합적으로 활용, 악성 URL을 보다 정확하게 탐지할 수 있는 솔루션을 제공한다. 특히, kNN과 XGBoost 알고리즘을 기본 모델로 채택하고, 로지스틱 회귀를 메타 모델로 사용하여 여러 기본 모델의 예측 결과를 효과적으로 결합하는 과정에서 교차 검증을 활용하였다. 이는 예측 성능을 신뢰성 있게 평가하고 일반화 성능을 향상시키는 데 중요한 역할을 할 것이다. 향후 연구에서는 이 모델을 다양한 데이터셋에 적용하여 실제 성능 평가를 수행하고, 이를 통해 모델의 강점과 한계를 더욱 명확히 파악할 계획이다. 본 연구는 악성 URL과 관련된 위협을 줄이는 데 크게 기여할 것으로 기대된다.

ACKNOWLEDGMENT

“이 기술은 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 개발한 결과물입니다.”

참고문헌

[1] “포스트코로나, 방역·결제·NFT 접속까지 넘나드는 ‘QR코드’”. Available: https://www.apple-economy.com/news/articleView.html?idxno=67307

[2] “"QR코드 찍었다가 천만원 날렸다"…한번 찍히면 끝 '경고' [이슈+]”. Available: https://news.nate.com/view/20240218n09878

[3] “[인천경찰24시]인천경찰청"주정차위반스티커와 공유자전거 '가짜 QR코드' 각별 조심"”. Available: https://news.ifm.kr/news/articleView.html?idxno=384599

[4] “Phishing Landscape 2023: An Annual Study of the Scope and Distribution of Phishing”. Available: https://interisle.net/insights/phishinglandscape2023

[5] Kisu Yang, Taesun Whang, Dongsuk Oh, Chanjun Park, Heuiseok Lim, “Cross-Validated Ensemble Methods in Natural Language Inference”, Journal of KIISE, Vol.48, No.2, pp.154-159, 2021. 2

[6] ”A Gentle Introduction to k-fold Cross-Validation“. Available: https://machinelearningmastery.com/k-fold-cross-validation/

[7] ”What is the k-nearest neighbors (KNN) algorithm?“. Available: https://www.ibm.com/topics/knn

[8] Saleem Raja A, Vinodini R, Kavitha A, ”Lexical features based malicious URL detection using machine learning techniques“, Materials Today: Proceedings, Volume 47, Part 1, Pages 163-166, 2021

[9] “XGBoost Documentation”. Available: https://xgboost.readthedocs.io/en/stable/

[10] Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey“, arXiv:1701.07179v3, 2017

[11] ”What is logistic regression?“. Available: https://www.ibm.com/topics/logistic-regressiond

[12] Apoorva Joshi, Levi Lloyd, Paul Westin, Srini Seethapathy, “Using Lexical Features for Malicious URL Detection - A Machine Learning Approach ”, arXiv:1910.06277v1, 2019

[13] Y Swathi, Pramod Hegde, P Sravani, Pragati Hegd ev, “Detection of Phishing Websites Using Machine Lea rning”, 2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplina ry Technology Management (IC-RVITM), Bangalore, In dia, 2023, pp. 1-6