비지도 학습 기반 웹 로그 이상 탐지를 위한 공격 패턴 피처 엔지니어링 연구

양승원, 정주연, 강찬욱
건국대학교

yks8967@konkuk.ac.kr, jjjjyy0704@konkuk.ac.kr, kan0202@konkuk.ac.kr

Attack Pattern Feature Engineering Research for Unsupervised Learning-Based Web Log Anomaly Detection

Yang Seung Won, Jeong Ju Yeon, Kang Chan Wook
Konkuk Univ.

요 약

본 연구에서는 라벨이 존재하지 않는 대규모 웹 서버 환경에서 이상 로그를 탐지하기 위해, 웹 요청의 구조적 특성과 웹 공격에서 반복적으로 관찰되는 공격 패턴을 반영한 피처 엔지니어링 방법을 제안한다. URL 구조, HTTP 상태 코드, 헤더 정보 등과 함께 웹 공격에서 빈번히 사용되는 패턴을 출현 빈도로 반영하여 정량적 피처로 구성하고 이를 비지도학습 모델에 적용하였다. 실험 결과, 전체 로그 중 극히 일부에 해당하는 이상 로그를 효과적으로 고립시켰으며, 이상 점수 상위 100 개 로그에서는 실제 웹 공격 행위가 다수 관찰되었다. 이러한 결과는 공격 여부에 대한 라벨링 정보가 전혀 없는 실제 웹 서버 환경에서도 고위험 공격 가능성이 높은 로그를 식별할 수 있음을 보여준다.

Ⅰ. 서 론

웹 서버 환경에서의 사이버 공격이 증가함에 따라 효과적인 침입 탐지 시스템의 필요성이 점차 부각되고 있다. 기존의 시그니처 기반 탐지 기법은 사전에 정의된 공격 패턴에 의존하기 때문에, 이미 알려진 공격에 대해서만 탐지가 가능하다는 한계를 지닌다. 반면, 지도학습 기반 탐지 기법은 충분한 라벨 데이터가 확보될 경우 높은 탐지 성능을 기대할 수 있으나, 실제 운영 환경에서는 모든 로그에 대한 라벨링 비용이 매우 크다는 현실적인 제약이 존재한다. 또한 정상 로그 대비 공격 로그의 비율이 극히 낮은 클래스 불균형 문제로 인해 모델 학습이 어려우며, 특히 공격 유형이 변경될 경우 과거에 라벨링 되었던 데이터로 학습된 모델이 새로운 공격 패턴에 대해 효율적인 성능이 나온다고 보장할 수 없다. 이러한 현실적인 이유로 이상탐지 분야에서는 데이터 분포 특성만으로 이상 징후를 식별하는 비지도학습 기반 이상 탐지가 현실적인 대안으로 주목받고 있다.[1]

그러나 비지도학습 방식에서도 단순한 통계기반의 피처만으로는 공격 로그를 명확히 식별하기 어렵기 때문에, 웹 공격의 특성을 반영한 피처 엔지니어링이 중요하다. 따라서 본 연구는 대규모 웹 서버 환경에서 반복적으로 관찰되는 웹 공격 패턴을 기반으로, 로그 분포 상의 이질성을 반영하는 피처 엔지니어링 접근법을 제안한다.

Ⅱ. 문제 정의 및 데이터셋

본 연구에서는 사전 라벨 정보가 존재하지 않거나 매우 제한적인 실제 웹 서버 환경에서, 정상 로그 분포와 다른 특성을 가지는 로그를 이상 로그로 식별하는 것을 목적으로 한다. 이는 개별 로그를 정상 또는 공격으로 명확히 분류하는 것이 아니라, 라벨 정보가 없는 환경에서 정상 로그 분포로부터 구조적으로 이탈한 요청을 우선적으로 식별하기 위한 것이다. 실제 운영 환경에서는 공격 로그에 대한 명확한 라벨 데이터를 충분히 확보하기 어려운 경우가 많으며, 공격 발생 여부조차 명확히 정의하기 어려운 상황이다. 기존 연구들 또한 이러한 문제를 다루고 있으나, 공격 로그에 라벨을 부여한 지도학습 방식이거나 공격 패턴을 인위적으로 삽입하는 실험 설정을 사용하는 경우가 있다. 본 연구는 이러한 접근과 달리, 공격 로그를 가정하거나 생성하지 않은 실제 웹 서버 환경을 그대로 대상으로 삼고 라벨이 없는 환경에서의 공격행위를 찾는 비지도학습 기반 이상 탐지를 수행한다.

이러한 문제를 검증하기 위해 Kaggle 의 Web Server Access Logs 데이터셋을 사용하였다. 해당 데이터셋은 실제 웹 서버 환경에서 수집된 Access 로그로 구성되어 있으며, 별도의 라벨 정보가 제공되지 않은 약 900 만 건 이상의 대규모 로그 데이터를 포함한다. 본 논문에서 사용한 데이터셋의 일부 로그 정보는 다음과 같다.

```
31.56.96.51 - - [22/Jan/2019:03:56:32 +0330] "GET /static/images/amp/blog.png HTTP/1.1" 200 3863 "-" "Mozilla/5.0 (Linux; Android 6.0; ALE-L21 Build/HuaweiALE-L21) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/66.0.3359.158 Mobile Safari/537.36" "-"
```

Ⅲ. 피처 엔지니어링 및 모델 학습

라벨이 존재하지 않는 웹 서버 환경에서 이상 로그를 효과적으로 탐지하기 위해, 웹 요청의 구조적 특성과 공격 패턴을 반영한 피처를 설계하고 로그 한 건을 하나의 분석 단위로 취급하여, 구조 기반 피처와 공격 패턴 기반 피처를 결합한 다차원 피처 벡터를 구성하였다. HTTP Method 는 GET, POST 등과 같은 범주형 정보를 포함하므로, 각 Method 간의 순서적

의미를 배제하고 사용 여부를 독립적으로 표현하기 위해 One-Hot 인코딩을 적용하였다.

또한 HTTP 상태코드는 요청 결과의 의미를 직접적으로 반영하므로, 원본 상태코드를 그대로 사용하지 않고 의미 기반으로 그룹화하여 피처로 구성하였다. 2xx ~ 5xx 번 범위의 상태코드를 범주로 그룹화한 status_group 과 오류 응답 여부를 나타내는 is_error 피처로 구성하고 발생 빈도가 높은 상태코드에 대해서는 One-Hot 인코딩을 적용하여 비정상 요청에서 나타나는 응답 패턴을 반영하였다. URL 기반 피처로는 URL 길이, 경로 깊이, 쿼리 파라미터 존재 여부, 특수문자 빈도를 추출하였으며, User-Agent 와 Referrer 존재 여부를 함께 반영하여 자동화 요청 및 비정상 클라이언트의 특성을 포착하고자 하였다.

공격 패턴 기반 피처를 구성하기 위해, 웹 공격에서 반복적으로 관찰되는 문자열 흔적을 공격 패턴으로 정의하였다. 해당 패턴 집합은 SQL Injection, Command Injection, Directory Traversal, 웹 쉘 접근, 설정 파일 노출 등 다양한 웹 공격 유형과 관련된 문자열 패턴을 포함하며, 공개된 보안 분석 자료와 실제 공격 사례에서 빈번히 사용되는 키워드를 중심으로 구성하였다. 이러한 패턴들은 URL 및 User-Agent 필드에 포함 여부를 기준으로 집계되며, 개별 패턴의 공격 여부를 직접 판단하기 위한 규칙이 아니라, 모델이 로그 간의 분포 차이를 학습하고 이상 로그를 보다 효과적으로 고립시키기 위한 정량적 특성으로 활용된다.

Feature Category	Feature Name
HTTP/ Response	status_group, is_error
URL Structure	size, url_length, url_depth, has_guery_param, url_special_char_count
Header/Referrer	agent_length, ref_exists
Attack Pattern	uri_attack_count, ua_attack_count, attack_total_count
One-Hot(Method, Status)	method_*, status_*

표 1 최종 모델 학습 피처

이렇게 설계된 피처를 입력으로 하여 비지도학습을 적용하였으며, 웹 로그 데이터는 정상 트래픽이 대부분을 차지한다는 특성을 고려하여 정상 데이터의 분포로부터 상대적으로 고립되기 쉬운 로그를 이상치로 탐지할 수 있는 Isolation Forest 모델[2]을 사용하였다. 해당 모델은 무작위 분할을 통해 데이터 포인트를 고립시키는 방식으로 이상치를 탐지하며, 대규모 웹 서버 환경에서도 빠르고 효율적으로 적용 가능하다는 장점을 가진다.

Ⅳ. 이상 탐지 결과 및 분석

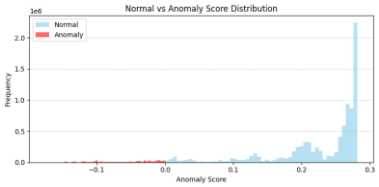

그림 1 이상점수 분포도

위에서 제안한 방식으로 Isolation Forest 모델을 적용하여 각 로그 요청에 대한 이상 점수를 산출하였다. 그림 1 은 정상 로그와 이상 로그의 이상 점수 분포를 나타내며, 정상 로그는 고밀도 영역에 분포하는 반면 이상 로그는 분포의 하위 영역에 집중되어 나타났다.

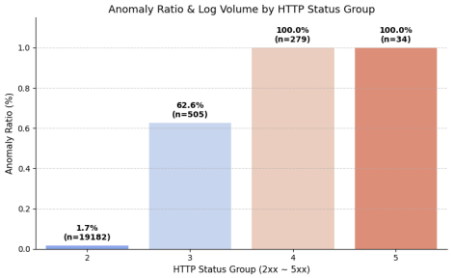

그림 2 HTTP 상태코드 그룹별 이상 탐지 비율

그림 2 는 모델이 HTTP 상태코드 그룹별 이상 로그를 탐지한 비율을 나타낸다. 정상응답에 비해 클리어언트 오류(4xx) 및 서버 오류(5xx) 응답에서 모델이 이상 로그를 보다 집중적으로 탐지하였으며, 탐지된 이상 로그가 의미적으로 공격 가능성이 높은 요청과 밀접한 관련이 있음을 확인할 수 있다.

그러나 상태코드별 분포 결과만으로는 실제로 탐지된 이상 로그의 구체적인 형태를 확인하기 어렵고 비지도 학습 환경의 특성상 정량적 성능 평가가 어려우므로, 본 연구에서는 이상 점수 기준 상위 로그에 대해 정성적 분석을 진행하였다. 상위 로그 100 개에 대한 분석 결과, 웹 공격에서 반복적으로 관찰되는 접근 패턴이 다수 확인되었다. 구체적으로는 엔드포인트에 대한 자동화된 접근(WordPress xmlrpc.php), 정상적인 서비스 이용 과정에서 생성되지 않는 비정상적인 API 엔드포인트 접근, 그리고 정상적인 입력 범위를 벗어난 파라미터 사용(count=null 등)이 관찰되었다. 또한 이들 중 일부 로그에서는 실제 취약점 악용 시도로 해석가능한 요청이 확인되었으며 이상 점수가 극단적으로 산출된 로그에 대한 추가 분석 결과 ThinkPHP 취약점을 악용한 RCE 공격이 확인되었으며 해당 로그의 구체적인 예시는 다음과 같다.

```
185.222.202.118 - - [22/Jan/2019:05:45:47 +0000] "GET /public/index.php?s=/index/₩think₩app/invokefunction&function=call_user_func_array&vars[0]=shell_exec&vars[1][]=cd /tmp;wget http://185.222.202.118/bins/rift.x86;cat rift.x86 > efjins;chmod 777 efjins;./efjins thinkphp HTTP/1.1" 403 134 "-" "python-requests/2.4.3 CPython/2.7.9 Linux/3.16.0-4-amd64"
```

Ⅴ. 결론

본 연구에서는 라벨링 되어있지 않은 실제 대규모 웹 서버 환경을 대상으로 비지도학습 기반 이상탐지를 통해 공격 가능성이 높은 로그를 식별하는 방법을 제안하였다. 웹 요청의 구조적 특성과 웹 공격에서 반복적으로 관찰되는 패턴을 반영한 피처를 설계하고 Isolation Forest 모델을 적용한 결과, 정상 로그 분포로부터 이탈한 이상 로그를 효과적으로 탐지되었다. 특히 추가적인 로그 분석을 통해 실제 웹 서버에서 WordPress, ThinkPHP 취약점 등과 같은 공격 시도를 탐지한 것을 확인할 수 있다. 이는 라벨 정보가 존재하지 않는 실제 웹 서버 환경에서도 비지도학습 기반 이상 탐지가 실제 공격 탐지에 효과적임을 확인한 결과이다.

참 고 문 헌

[1] R. Sommer and V. Paxson, Outside the Closed World: On Using Machine Learning for Network Intrusion Detection, IEEE Symposium on Security and Privacy, pp. 305-316

[2] Liu, F.T., Ting, K. M., & Zhou, Z.-H., Isolation Forest, Proc. IEEE Int. Conf. on Data Mining (ICDM), pp. 413-422