https://doi.org/10.33778/kcsa.2024.24.2.003

인공지능 기반 침해분석 도구 활용에 관한 연구

양 환 석*

요 약

최근 몇 년 동안 사이버 위협의 수와 복잡성이 증가하고 있다. 이러한 위협은 개인 소유 장치를 업무에 사용하는 것의 위험성을 증가시킨다. 이 연구는 인공지능을 활용한 침해분석 도구의 활용 방안에 대해 다루고 있다. 이를 위해 자동화된 분석 프로세스를 통해 분석자의 업무 부담을 줄이고 분석 효율을 향상시키는 인공지능 기반 침해분석 도구를 개발하고 활용 가능성을 제안하였다. 이를 통해, 분석자는 더욱 중요한 업무에 집중할 수 있다. 본 논문에서는 인공지능 기반 침해분석 도구의 개발과 활용 가능성을 제시하는 것이다. 이를 통해 침해분석 분야의 새로운 연구 방향을 제시하고, 자동화 도구의 성능, 적용 범위, 사용 편의성을 향상시켜 조직이 효과적으로 사이버 공격에 대응할 수 있도록 하는 것이 필요하다는 것을 제시하였다. 연구 방법으로는 인공지능 기술을 활용하여 침해분석 도구를 개발하고, 이를 통해 다양한 활용 사례를 연구하였다. 또한, 자동화 도구의 성능, 적용 범위, 사용 편의성을 평가하고, 침해 사고의 예측 및 예방, 자동 대응을 위한 연구도 진행하였다. 본 연구는 인공지능 기반 침해분석 도구의 개발과 활용을 위한 기초가 될 것으로 이를 통해 효과적으로 사이버 공격에 대응 방안을 실험을 통해 확인할 수 있었다.

A Study on the Utilization of Artificial Intelligence-Based Infringement Analysis Tools

Yang Hwan Seok*

ABSTRACT

Recently, in order to build a cyber threats have increased in number and complexity. These threats increase the risk of using personally owned devices for work. This research addresses how to utilize an AI-enabled breach analysis tool. To this end, we developed and proposed the feasibility of using an AI-based breach analysis tool that reduces the workload of analysts and improves analysis efficiency through automated analysis processes. This allows analysts to focus on more important tasks. The purpose of this research is to propose the development and utilization of an AI-based breach analysis tool. We propose a new research direction in the field of breach analysis and suggest that automated tools should be improved in performance, coverage, and ease of use to enable organizations to respond to cyberattacks more effectively. As a research method, we developed a breach analysis tool using A.I. technology and studied various use cases. We also evaluated the performance, coverage, and ease of use of automated tools, and conducted research on predicting and preventing breaches and automatically responding to them. As a result, this research will serve as a foundation for the development and utilization of AI-based breach analysis tools, which can be used to respond to cyberattacks more effectively through experiments.

Key words : Artificial Intelligence, Analytical Tools, Incident Analysis, Malware

접수일(2024년 06월 03일), 게재확정일(2024년 06월 28일)

* 중부대학교/정보보호학과

1. 서 론

최근 기업들은 업무 생산성을 높이기 위해 모바일 기기를 업무에 활용하기 시작했으며, 그 결과 폐쇄적이던 기업 환경이 점차 개방형 구조로 변화되고 있다. 이러한 기업의 업무 환경은 BYOD(Bring Your Owen Device)라는 개념으로 새로운 기업 환경으로 자리 잡고 있다[1]. 기업업무 환경을 구성에 있어서 이러한 BYOD 장치(PC, 노트북, 휴대폰, 각종 모바일 디바이스 등)들은 단말(Endpoint)에 해당하며 이곳에서 수많은 위협 요소와 취약점이 노출되고 있어 이에 대한 대응이 중요해지고 있다[2]. 그러므로 개인 소유 장치를 업무에 사용하는 것의 위험성을 제거하기 위해 우선 소유자 개인의 정보가 기업 보안 인프라의 일부로 관리되어야 할 것이며 이 과정에는 상호 수용 가능한 보안모델이 필요하다는 필요성이 제기되고 있다[3.]

본 논문에서는 사이버 공격으로 인해 감염된 단말(Endpoint)에서 시스템이나 애플리케이션이 자동으로 생성한 데이터 및 확보되는 데이터들을 분석하여 각 데이터의 특징들을 가지고 효과적으로 식별할 수 있는 인공지능 엔진을 제안한다. 또한 이러한 식별된 결과를 TTP(Tactic, Technique, Procedure) 형식으로 저장하여 이를 다양한 응용에 활용될 수 있도록 하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 침해분석 엔진에서 사용되는 대표적인 기법에 대하여 살펴보고 3장에서는 본 논문에서 제안한 인공지능을 활용한 침해사고 분석 엔진에 관해 기술하였다. 4장에서는 제안한 기법의 성능 평가를 위해 실험하고 마지막으로 5장에서 결론을 맺는다.

2. 인공지능

2.1 LLM

LLM(Large Language Models)은 번역, 생성, 요약, 분류 등 다양한 자연 처리가 가능하며, 대용량의 모델을 지원한다[4]. 또한, 맞춤형 모델 학습을 통해 유연한 텍스트의 생성이 가능한 장점이 있다. 특히, 질문에 답변하는 작업에 LLM은 더욱 세밀하게 사람의 의도를 이해하고, 적절한 답을 할 줄 알고 좀 더 활용하면 원하는 답변이 나올 때까지 특정 분야에 대한 필요한 지식을 학습시키며 자동화 도구를 만들 수 있다.

2.2 프롬프트 엔지니어링

프롬프트 엔지니어링은 LLM 모델을 활용하여 주어진 프롬프트(prompt)에 대한 생성형 인공지능 모델에 사용자의 의도를 효과적으로 전달하여 원하는 결과물을 얻기 위해 입력하는 프롬프트를 설계, 최적화하는 과정 즉, 자동완성을 수행하는 기술이다[5]. 프롬프트는 입력된 시작 문장을 의미하며, LLM 모델은 주어진 프롬프트를 기반으로 자연어 생성을 수행하지만 서로 다른 단어가 서로 뒤따를 가능성을 기반으로 응답을 생성한다. 이 가능성은 학습 데이터에서 유사한 문맥에서 발생한 빈도를 따른다.

2.3 랜덤 포레스트(RandomForest)

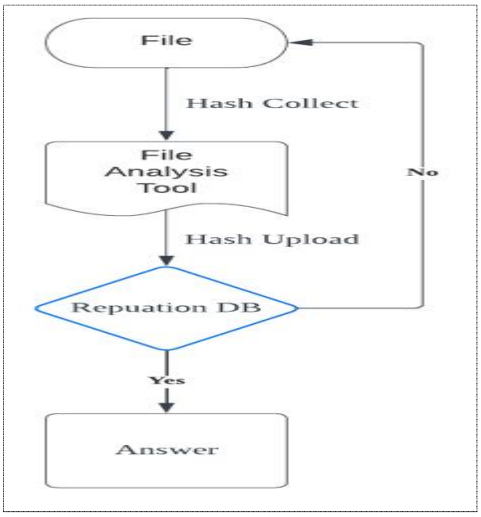
랜덤 포레스트는 여러 개의 결정트리(Decision Tree)를 활용한 배경 방식의 알고리즘이다[6]. 결정트르리의 쉽고 직관적인 장점을 사용한다. 앙상블 알고리즘 중 비교적 빠른 수행속도를 가지고 있어 빠른 학습이 가능하다. 다양한 로그를 비교 분석해야하는 상황에서 좋은 성능을 가지고 있다.

3. 시스템 개요

본 장에서는 인공지능을 기반으로 한 침해사고 분석 도구에 대해 각 엔진들에 대하여 설명하였다. 제안한 기법은 정확한 탐지를 위하여 LLM, 프롬프트 엔지니어링, 랜덤 포레스트 알고리즘 기술을 적용하였다.

3.1 악성 파일 여부 확인 엔진

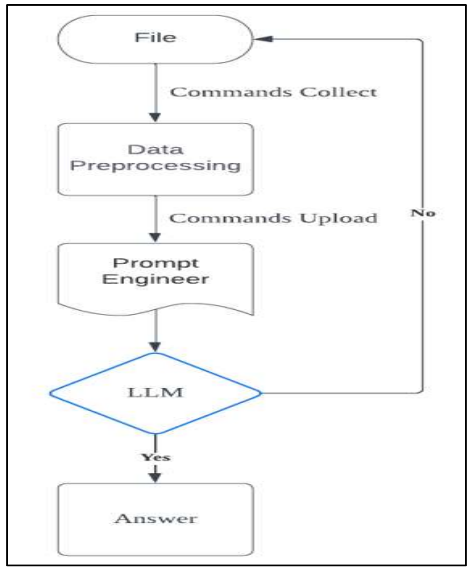
침해 사고 초동 분석은 사고의 원인과 피해 범위를 파악하는 중요한 단계이다. 악성 파일 여부 확인 엔진은 이 단계에서 악성 파일의 유무를 빠르게 확인하여 분석의 방향을 정하기 위해 사용된다. 악성 파일 여부 확인 엔진은 파일의 해시값을 추출하여 악성 파일 여부를 판단한다. 해시값은 파일의 내용을 압축하여 고유한 값으로 표현한 것이다. 파일이 동일하면 해시값도 동일하게 된다. 따라서 악성 파일의 해시값을 평판 조회 DB에 등록하여, 새로운 파일의 해시값과 비교하면 악성 파일 여부를 빠르게 확인할 수 있다. 해시값 기반 평판 조회 분석은 악성 파일 여부를 빠르게 확인할 수 있다는 장점이 있다. 하지만, 악성 파일의 해시값이 변경될 경우 악성 파일을 탐지하지 못할 수 있다는 단점이 있다. 따라서, 해시값 기반 평판 조회 분석과 함께 다른 분석 엔진을 함께 사용하면 악성 파일 탐지율을 높일 수 있다.



(그림 1) 악성 파일 여부 확인 엔진 수행도

3.2 파일 행위 비교 엔진

파일 행위 비교 엔진은 악성 파일의 행위 패턴을 파악하는 데 중점을 둔다. LLM 모델은 악성 파일의 행위에 대한 보다 구체적인 정보를 제공한다. 예를 들어, 악성 파일이 시스템에 어떤 영향을 미칠 수 있는지, 어떤 데이터를 수집할 수 있는지, 어떤 시스템에 영향을 미칠 수 있는지 등을 파악할 수 있다. LLM 모델은 방대한 데이터를 학습하기 때문에 악성 파일의 새로운 행위 패턴도 탐지할 수 있다. 따라서, 파일 행위 비교 엔진과 LLM 모델을 함께 사용하면 악성 코드 공격을 보다 효과적으로 탐지할 수 있다.

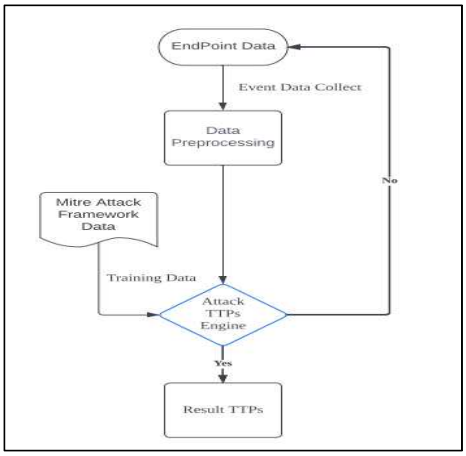


(그림 2) 파일 행위 비교 엔진 수행도

3.3 공격 행위 TTPs 엔진

공격 행위 TTPs 엔진은 악성코드가 시스템에 침입한 후 수행할 수 있는 공격을 TTPs로 예측하는 엔진이다. 이 논문에서는 MITRE ATT&CK Framework에 매핑된 Atomic Red Team 라이브러리를 학습 데이터로 사용한다. MITRE ATT&CK Framework는 다양한 공격 TTPs를 문서화한 프레임워크이다. Atomic Red Team 라이브러리는 MITRE ATT&CK Framework에 매핑된 공격 방식을 코드로 표현한 데이터베이스이다. RandomForest 알고리즘을 사용하여 공격 방식을 코드로 표현된 데이터 정보들과 매핑한다. RandomForest 알고리즘은 의사결정나무 모델을

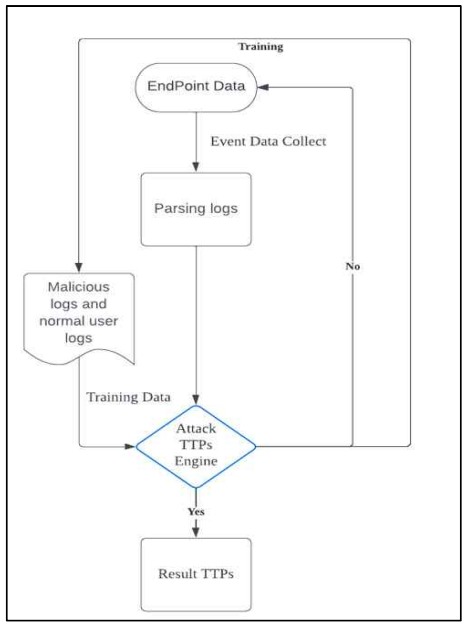
여러 개 생성하여 결과를 종합하여 예측하는 알고리즘이다.실행, 지속, 권한 상승, 방어 회피 등의 정보를 중점으로 학습한다. 이러한 정보는 공격자가 시스템에 침입한 후 수행할 수 있는 일반적인 공격 방식이다. 매핑된 데이터를 스토리라인으로 구상하여 머신러닝 모델을 학습시킨다. 스토리라인은 공격자의 행동을 연속적으로 연결한 것이다. 스토리라인으로 구상된 데이터를 학습시킴으로써 공격자의 행동을 보다 정확하게 예측할 수 있다.



(그림 3) 공격 행위 TTPs 엔진 수행도

3.4 자동화 비교 수행 엔진

자동화 비교 수행 엔진은 앞서 소개된 엔진들에서 발견되지 않았던 초기 공격 기법을 탐지하는데 사용된다. 이러한 공격 기법은 악성 파일이 존재하지 않거나, 악성 파일의 행위가 분석되지 않을 때 발생할 수 있다. 자동화 비교 수행 엔진은 엔드포인트 단에서 확인되는 로그를 이용하여 초기 공격 기법을 찾는다. 로그에는 악성 행위가 발생했을 때 발생하는 정보가 포함되어 있다. 자동화 비교 수행 엔진은 Atomic에서 제공되는 데이터를 학습시킨 머신러닝을 사용하여 로그를 분석한다. Atomic 데이터는 실제 공격에서 발생한 로그를 수집하여 만든 데이터베이스이다. 머신러닝은 로그에서 발견되는 패턴을 학습하여 초기 공격 기법을 탐지한다.



(그림 4) 자동화 비교 수행 엔진 수행도

4. 실험 및 결과

4.1 실험 결과 분석

파일의 해시값을 기반으로 악성 여부를 판단하기 때문에 파일의 내용을 압축하여 고유한 값으로 표현한 파일의 해시값을 이용하여 엔진을 구축하였다. 이러한 방식으로 악성 파일 여부 확인 엔진은 침해 사고 초동 분석 시 분석의 방향성을 정하고, 추가적인 분석을 위한 시간을 확보할 수 있다.

(그림 5) 악성 파일 여부 확인 엔진 결과

LLM을 사용하여 악성 파일의 행위에 대한 정보

를 얻는다. ChatGPT와 Google Bard LLM을 비교한 결과, Google Bard가 더 나은 성능을 보였다. Google Bard는 ChatGPT보다 더 많은 데이터를 학습하여 다양한 질문에 대한 답변을 제공할 수 있다. 성능 향상을 위해 프롬프트를 개선하였다. 하나의 질문보다는 여러 개의 질문을 통해 악의적인 행위의 과정을 파악하는 것이 효과적이다.

(그림 6) 파일 행위 비교 엔진 결과

파일 행위 비교 엔진에서 악성 코드 행위를 분석하여 나온 데이터를 TTPs 형식에 맞게 매핑하고, 이를 바탕으로 공격자의 행위를 도출하는 엔진이다. TTPs 학습 데이터는 MITRE ATT&CK Framework를 매핑한 Atomic Red Team 라이브러리 데이터를 사용한다. TTPs 학습은 RandomForest 알고리즘을 사용하여 공격 방식을 코드로 표현된 데이터 정보들을 의사결정나무 모델로 훈련시킨다. 의사결정나무 모델은 데이터의 특성에 따라 분류 또는 회귀를 수행하는 머신러닝 모델이다.

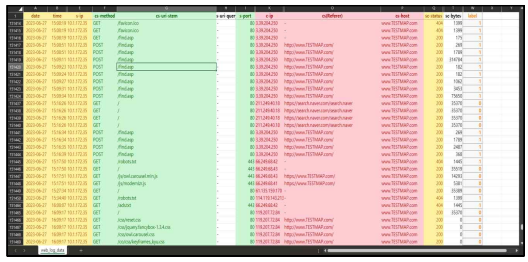
(그림 7) 공격 행위 TTPs 엔진 결과

본 논문에서는 실행, 지속, 권한 상승, 방어 회피 등의 정보를 중심으로 학습시켰다. 이러한 정보는 공격자가 시스템에 침입한 후 수행할 수 있는 일반적인 공격 방식이다. 학습된 모델은 악성 파일이 시스템에 침입한 후 수행한 가능성이 높은 공격을 TTPs로 예측할 수 있다. <표 1>에서는 본 논문에서 실험한 파일 행위 비교 엔진 실험의 평균 결과를 보여주고 있다.

<표 1> 평균 인식 결과

학습 데이터	Precision
100	0.664
1500	0.826
2000	0.985

자동화 비교 수행 엔진은 엔드포인트 단에서 확인되는 로그를 분석하여 초기 공격 기법을 탐지하는 엔진이다. 로그에는 악성 행위가 발생했을 때 발생하는 정보가 포함되어 있다. 자동화 비교 수행 엔진은 이러한 로그를 분석하여 초기 공격 기법을 탐지한다. 로그를 분석하기 위해 특정 필드를 정리하여 머신러닝에 학습시킨다. 정리된 필드는 date, time, s-ip, cs-method, cs-uri-stem, s-port, c-ip, cs(Referer), cs-host, sc-status, sc-bytes 등이다.



(그림 8) 각 필드로 나눈 학습 데이터

<표 2>에서는 본 논문에서 실험한 자동화 비교 수행 엔진 실험의 평균 결과를 보여주고 있다. 학습된 데이터의 숫자가 일정 수준 이상이 되면 정확도가 높아진다. 예를 들어, 학습된 데이터의 숫자가 1,000개 미만일 때 정확도는 76% 정도이지만,

15,000개 이상일 때 정확도는 99% 이상으로 높아진다. 탐지된 초기 공격 기법은 TTPs와 매핑하여 공격의 종류를 파악할 수 있다.

<표 2> 평균 인식 결과

학습 데이터	Precision
6000	0.762
105523	0.848
158985	0.989
179431	0.985

5. 결 론

인공지능 기반 침해 분석 도구는 침해 사고 대응의 효율성을 향상시킬 수 있는 잠재력을 가지고 있다. 인공지능 기반 침해 분석 도구를 통해, 조직은 침해 사고로 인한 피해를 최소화할 수 있다. 또한, 중소기업이나 기관의 사이버 보안 강화에 기여할 수 있다.

본 본문에서 인공지능 기반 침해 분석 도구의 개발과 활용 가능성을 제시하였다. 인공지능 기반 침해 분석 도구는 대량의 데이터를 효과적으로 처리하고 분석하여, 침해 사고를 보다 빠르고 정확하게 식별할 수 있다. 또한, 기존의 침해 분석 도구에서 발견하기 어려운 새로운 위협을 식별할 수 있다. 그리고 자동화된 분석 프로세스를 통해, 분석자의 업무 부담을 줄이고 분석 효율을 향상시킬 수 있다.

참고문헌

[1] EKE, Christopher Ifeanyi; NORMAN, Azah Anir; MULENGA, Mwenge. Machine learning approach for detecting and combating bring your own device (BYOD) security threats and attacks: a systematic mapping review. Artificial Intelligence Review, 2023, 56.8: pp. 8815-8858.

[2] AlShalaan, Manal Rajeh, and Suliman Mohamed Fati. "Enhancing Organizational Data Security on Employee-Connected Devices Using BYOD Policy." Information 14.5 (2023): 275.

[3] SHYKYRYNSKA, Oleksandra, et al. Formation of Scientific Research Competence of Master's Degree Students by Means of Byod Technology. In: SOCIETY. INTEGRATION. EDUCATION. Proceedings of the International Scientific Conference. 2023. pp. 329-337.

[4] Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and benefits of large language models for the environment. Environmental Science & Technology, 57(9), pp. 3464-3466.

[5] Velásquez-Henao, Juan David, Carlos Jaime Franco-Cardona, and Lorena Cadavid-Higuita. "Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering." Dyna 90.230 (2023): pp. 9-17.

[6] Wang, Beibei, et al. "A Dynamic Trust Model for Underwater Sensor Networks Fusing Deep Reinforcement Learning and Random Forest Algorithm." Applied Sciences 14.8 (2024): 3374.

〔저자소개〕

양 환 석 (Hwan-seok Yang)
1998년 2월 조선대학교 이학석사
2005년 2월 조선대학교 이학박사
2007년 3월 호원대학교 연구교수
2011년 9월 ~ 현재 중부대학교 정보보호학과 부교수
email : yanghs@joongbu.ac.kr