머신러닝을 사용한 로그수집 시스템 설계 제안에 관한 연구

서덕원*, 윤호상**, 신동일*, 신동규*
*세종대학교 컴퓨터공학과
**국방과학연구소
e-mail:ejr3949@gec.sejong.ac.kr

A Study on the Logging System Design Suggestion Using Machine Learning

Deck-Won Seo*, Ho-sang Yooun**, Dong-Il Shin*, Dong-Kyoo Shin*
*Dept of Computer Engineering, Se-jong University
**Agency for Defense Development

요 약

현대사회에서는 사이버 해킹 공격이 많이 일어나고 있다. 공격이 증가함에 따라 이를 다양한 방법으로 방어하고 탐지하는 연구가 많이 이루어지고 있다. 본 논문은 OpenIOC, STIX, MMDEF 등과 같은 공격자의 방법론 또는 증거를 식별하는 기술 특성 설명을 수집해 놓은 표현들을 기반으로 머신러닝과 logstash 라는 로그 수집기를 결합하는 새로운 시스템을 제안한다. 시스템은 pc 에 공격이 가해졌을 때 로그 수집기를 사용하여 로그를 수집한 후에 로그의 속성 값들의 리스트를 가지고 머신러닝 알고리즘을 통해 학습시켜 분석을 진행한다. 향후에는 제안된 시스템을 실시간 처리 머신러닝 알고리즘을 사용하여 필요로그정보의 구성을 해주면 자동으로 로그정보를 수집하고 필터와 출력을 거쳐 학습을 시켜 자동 침입탐지시스템으로 발전할 수 있을 것이라 예상된다.

1. 서론

최근 현대사회에서는 사이버 해킹 공격이 많이 일어나고 있다. 공격이 증가함에 따라 이를 다양한 방법으로 방어하고 탐지하는 연구가 많이 이루어지고 있다. 보안을 위해서는 먼저 수집해야 할 많은 정보들이 있다. 그 정보들 중 해커들이 어떠한 정보를 거치며 어떠한 경로를 통해 들어 왔는지에 대한 정보를 수집하기 위해서는 로그 정보가 필요하다. 로그는 시스템의 모든 기록을 담고 있는 데이터이다. 이러한 로그 데이터는 IT 인프라 관리, 이상 징후 탐색 등 다방면으로 활용될 수 있다. 최근 이슈가 되고 있는 IT 보안 사고에 대비할 수 있는 방안으로 로그 데이터 관리에 대한 중요성이 대두되고 있다. 시스템의 모든 기록을 담고 있는 로그데이터를 수집하고 분석, 보관, 모니터링하여 내, 외부로부터의 공격을 사전에 대비하고 공격을 당하더라도 그 원인을 찾아 신속히 복구할 수 있기 때문이다[1]. 본 논문에서는 pc 에 공격이 가해졌을 때 logstash 라는 로그 수집기를 사용하여 수집 후 로그의 속성 값들의 리스트를 가지고 SVM (Support Vector Machine)머신러닝 알고리즘으로 학습을 시키는 정보 수집기를 제안한다. 향후에는 제안한 수집기를 실시간처리 및 자동화를 하여 자동 침입 탐지시스템을 연구 할 수 있을 것으로 예상한다.

본 논문에서는 2 장에서 관련연구에 대해 설명하고 3 장에서는 제안할 시스템의 구조를 기술하고 4 장에서 결론을 제시한다.

2. 관련 연구

2.1 OpenIOC, STIX, MMDEF

먼저 OpenIOC(Open Indicators of Compromise)는 알려진 위협, 공격자의 방법론 또는 다른 타협의 증거를 식별하는 기술 특성 설명을 위한 확장 가능한 XML 스키마이다[2,3]. STIX 는 사이버 위협 표현 규격은 개별 조직들이 보유하고 발전시켜 온 사이버 위협 정보의 개념을 표준화하고 구조화하여 사이버 위협에 대한 일관된 분석과 자동화된 해석이 가능하게 한 정보 표현 규격이다. 구성요소는 다양한 세부적인 속성을 가지고 있으며 XML 기반 언어이며 잠재적 인 사이버 위협 정보를 전달하는 데 사용된다[2,4]. MMDEF(Malware Metadata Exchange Format)는 맬웨어를 설명하고 해시, 파일 속성, 서명, 소프트웨어 패키지 및 레지스트리와 같은 맬웨어 파일 정보를 제공할 수 있는 XML 스키마이다[2]. 본 논문은 위에 설명한 기술을 기반으로 머신러닝을 이용하여 새로운 로

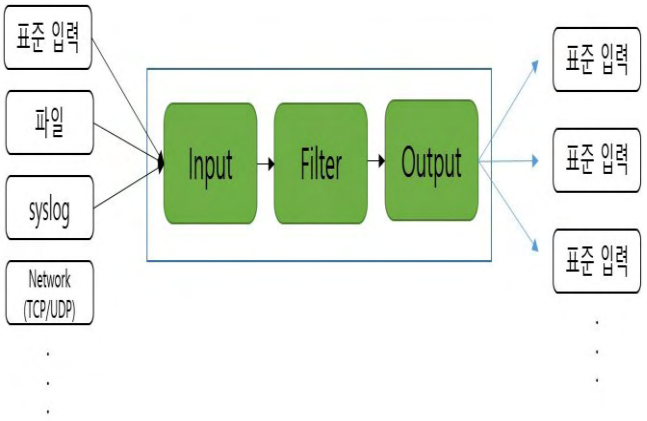
그 수집기에 대한 시스템을 제안한다.

2.2 학습을 사용할 알고리즘 + 학습 도구 설명

제안할 로그 수집기에서 분석을 위해 머신러닝 알고리즘으로는 SVM (Support Vector Machine)사용을 제안한다. SVM 은 1995 년에 개발되고 제안된 학습 알고리즘이다. 기계 학습의 분야 중 하나로 패턴 인식, 자료 분석을 위한 지도 학습 모델과 주로 분류와 회귀 분석을 위해 사용한다. 원래 이진 분류를 위하여 개발되었지만 현재에는 문자인식, 필기인식, 얼굴 및 물체 인식 등 다양한 분야에서 적용되고 있다. [5] 분석을 위한 알고리즘을 학습시키기 위한 학습 도구로는 WEKA 를 제안한다. WEKA 는 뉴질랜드에서 자바 언어로 개발된 기계학습 및 데이터 마이닝을 휘한 소프트웨어이다. 주요 기능으로는 명령어 라인 또는 GUI 환경으로 모두가 실행가능하고, 데이터 전처리 기능, 학습 알고리즘 및 평가 방법, 학습 알고리즘 비교 기능, 처리결과를 시각적으로 보여주는 가시화 기능 등이 제공된다[6].

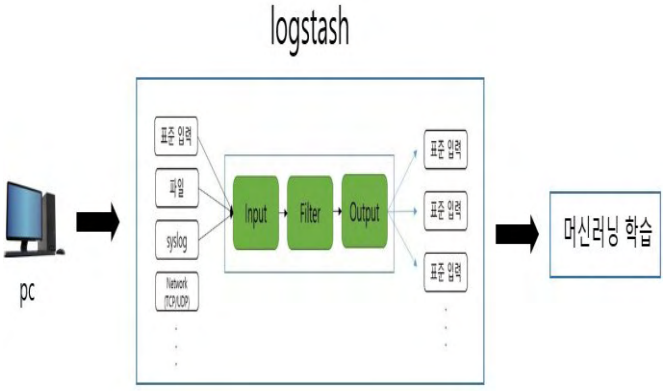
3. 제안하는 로그 수집 시스템 구조

logstash 는 데이터의 입력 변환 출력을 실시간 파이프라인으로 처리하는 오픈 소스 데이터 수집 엔진이다. 아래 (그림 1)은 logstash 의 기본 동작 구조를 보여준다.



(그림 1) logstash 동작

작동 과정은 입력(Input), 필터(Filter) 출력(Output) 3 단계 이벤트 파이프라인으로 처리한다. 먼저 입력(Input)을 통해 여러 시스템에 다양한 형태로 저장, 분산 되어있는 데이터들을 동시에 가져올 수 있다. 필터(Filter)를 통해 데이터가 소스 에서 저장소로 이동 중에 각 이벤트를 분석이 용이한 구조, 즉 비정형데이터를 정형데이터로 변환해주고 로그를 원하는 형태로 변환이 가능하다. 그리고 출력(Output) 과정에서는 자주 사용하는 보관소로 추출한 데이터를 전송하여 저장할 수 있다. logstash 의 장점은 다양한 플러그 인이 지원되어 다양하게 확장할 수 있다는 점이다[7,8].
시스템의 순서는 pc 에 attack 이 들어온 후 건드린 중요 로그의 정보를 수집해주는 일을 한다. 먼저 logstash 를 이용하여 입력에 들어갈 것은 수집하고자 하는 로그에 대해서 입력필터출력 구성을 먼저 한다. 코드를 통해 수집하고 싶은 정보를 입력해주면 된다. 필요 로그 정보를 수집한 후에 필터를 통해 입력데이터를 분해, 추가, 삭제, 변형 등의 과정을 거쳐 각 이벤트를 분석이 용이한 구조로 변환해준다. 그 후 다양한 저장소로 필터 된 데이터를 전송해준다.



(그림 2) 제안 시스템 동작

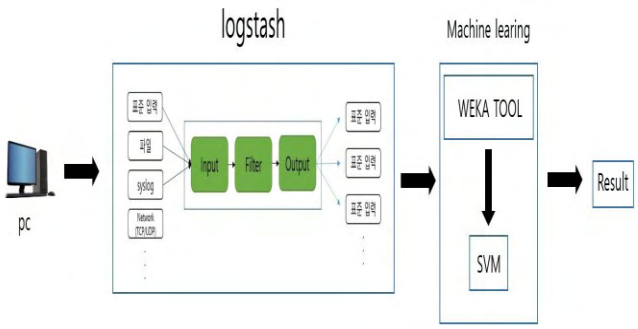
<표 1> 수집한 로그의 속성 값 정리

FileItem	FileItem/Md5sum
	FileItem/FileName
	FileItem/SizeInBytes
	FileItem/FileExtension
	FileItem/PEInfo/ImportedModules/Module/Name
	FileItem/PEInfo/ImportedModules/Module/ImportedFunctions/string
	FileItem/Created
	FileItem/FileExtension
	FileItem/FullPath
	FileItem/PEInfo/DetectedAnomalies/string
	FileItem/PEInfo/PETimeStamp
	FileItem/PEInfo/Exports/ExportsTimeStamp
	. . .
RegistryItem	RegistryItem/Path
	RegistryItem/Text
	RegistryItem/Value
	RegistryItem/ValueName
	RegistryItem/KeyPath
PortItem	PortItem/protocol
	PortItem/localPort

	PortItem/remotePort
	PortItem/remoteIP
Network	Network/DNS
	Network/String
HookItem	HookItem/HookDescription
	HookItem/HookingModule
UrlHistoryItem	UrlHistoryItem/URL
Snort	Snort/Snort
. . .	. . .

위의 <표 1> 처럼 수집한 로그의 속성 값을 정리한 데이터 셋을 토대로 SVM(Support Vector Machine)의 머신러닝 알고리즘을 사용하여 학습 및 분석을 시키는 구조이다.

아래 <그림 3>은 최종 시스템의 구조를 나타낸다.



(그림 3) 최종 시스템 동작

4. 결론

본 논문은 머신러닝과 logstash 라는 로그 수집기를 결합하는 새로운 시스템을 제안하였다. 시스템은 pc 에 공격이 가해졌을 때 로그 수집기를 사용하여 PC 들의 다양한 로그들을 수집한 후에 로그의 속성 값들의 리스트 하나의 파일로 정리를 한 데이터 셋을 만든다. 그 후 머신러닝 알고리즘을 통해 학습시켜 분석을 진행한다. 향후에는 제안된 시스템을 실시간처리 머신러닝 알고리즘을 사용하여 필요로그정보의 구성을 해주면 자동으로 로그정보를 수집하고 필터와 출력을 거쳐 학습을 시켜 자동 침입탐지시스템으로 발전할 수 있을 것이라 예상된다.

Acknowledgement

본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다(UD160066BD).

참고문헌

[1] http://www.dailysecu.com/?mod=news&act=articleView&idxno=22122
[2] Kampanakis, Panos. "Security automation and threat information-sharing options." *IEEE Security & Privacy* 12.5 (2014): 42-51.
[3] http://www.openioc.org/
[4] Rattan, A., Kaur, N., Chamotra, S., & Bhushan, S. Attack Data Usability and Challenges in its Capturing and Sharing.
[5] 김한성, 권영희, and 차성덕. "SVM 기반의 효율적인 신분위장기법 탐지." *정보보호학회논문지* 13.5 (2003): 91-104.
[6] 김종완. "WEKA 도구를 이용한 인공지능 수업 개선." *한국지능시스템학회 학술발표 논문집* 22.2 (2012): 170-171.
[7] 송중호, 김학민, and 윤진. "오픈소스를 이용한 윈도우 기반 PC 로그 수집 시스템." 정보과학회 컴퓨팅의 실제 논문지 22.7 (2016): 332-337.
[8] http://cs.sch.ac.kr/lecture/BigData/2017/10-Logstash.pdf