스마트홈 환경에서 활동 데이터를 활용한 랜덤포레스트 기반 침입탐지 기법

이필원*, 신용태**
*숭실대학교 컴퓨터학과
**숭실대학교 컴퓨터학부
pwlee@soongsil.ac.kr

Random Forest Based Intrusion Detection Method using Activity Data in Smart Home Environment

Pil-Won Lee, Yong-Tae Shin
Department of Computer Science, Soongsil University

요 약

최근 IoT 기술의 발전을 통해 스마트홈 서비스가 사용자에게 활발하게 보급이 되고 있다. 스마트홈 서비스에서 발생하는 데이터는 개인정보를 내포하고 있으므로 보안이 매우 중요한 요소이다. 그러나 매해 스마트홈 해킹 신고가 증가하고 있으며 기존 네트워크 침입탐지 시스템은 관리자 계정을 탈취당했을 경우 대응할 방법이 미비하다. 본 논문에서는 스마트홈 환경에서 발생하는 활동 데이터를 인공지능 알고리즘의 종류 중 하나인 랜덤포레스트를 통해 학습하고 분류모델을 구현했다. 구현한 모델은 87%이상의 높은 정확도로 측정되었다. 따라서 활동 데이터를 통해 분류를 시행하므로 네트워크에 이미 침입한 사용자를 탐지하여 대응할 수 있다.

1. 서론

최근 IoT(Internet of Things) 기술의 발전으로 개인 주거공간의 가전기기 및 디지털 디바이스를 네트워크에 연결하여 원격제어가 가능한 스마트홈 서비스가 통신사를 중심으로 활발하게 제공되고 있다. 스마트홈 서비스는 일반적으로 ISP(Internet Service Provider)의 퍼블릭 네트워크를 통해 서비스가 제공되기 때문에 개인정보의 탈취 위험에 항시 노출되어 있다. KISA(한국인터넷진흥원)에 따르면 스마트홈 관련 해킹 우려 신고 건수가 2018년 기준 387건으로 매해 증가했다. 스마트홈 서비스 제공업체는 암호화 등 보안 솔루션을 통해 해킹 피해를 최소화하려는 노력을 하고 있다. 그러나 관리 계정을 탈취하여 네트워크에 침입했을 경우 사용자가 해킹을 인지하고 계정을 변경하지 않는 이상 대응할 수 있는 장치가 미비하다. 따라서 본 논문에서는 스마트홈 환경에서 사용자의 활동 로그를 기계학습 알고리즘 중 하나인 랜덤포레스트를 통해 분류하여 네트워크 침입을 판단하고 대응할 수 있는 기법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 기존에 활용되는 IoT 침입 탐지 방법에 대해 알아보고 요구사항을 도출한다. 3장에서는 본 논문에서 제안하는 랜덤포레스트 기반 네트워크 침입 탐지 기법을 설계하고 평가한다. 마지막 4장에서는 결론을 제시한다.

2. 관련연구

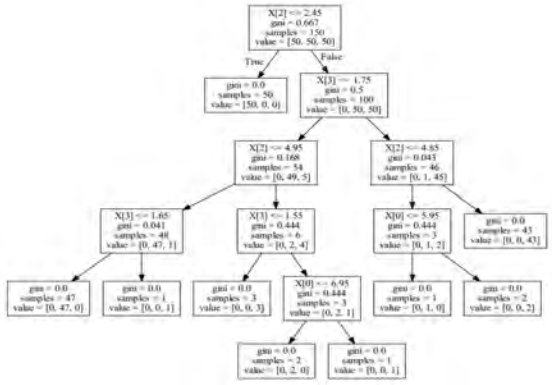
2.1 침입탐지 시스템

침입탐지 시스템은 IoT 네트워크가 아닌 기존 네트워크 환경에서의 많은 연구가 이루어졌다. 따라서 대부분의 연구에서 기존 네트워크 환경의 Dataset을 활용하여 침입탐지 모델을 구성한다[1]. 대표적인 Dataset은 KDDCUP99로써 DoS(Denial of Service Attack), U2R(User to Root Attack), R2L(Remote to Local Attack), Probing attack 총 4가지의 공격 종류가 포함되어있다[2]. KDDCUP99는 시스템에 직접적으로 개입하여 공격하는 시나리오를 중점으로 구성되어 있으며 전통적인 네트워크 공격 탐지 모델을 효과적으로 구성하고 평가할 수 있다. 그러나 위 Dataset으로 침입탐지 모델을 구성할 경우 관리자 계정을 탈취당하여 시스템에 이미 침입을 했다면 더 이상 대응할 수 있는 수단이 없다는 단점

이 있다.

2.2 랜덤포레스트 분류기

랜덤포레스트는 다수의 의사결정나무를 활용하여 앙상블(ensemble)로 구성한 분류기이다[3]. 앙상블은 여러 개의 학습 모델을 합쳐서 하나의 결과로 만드는 것이다. 랜덤포레스트를 구성 할 때 각각의 의사결정트리는 데이터를 무작위로 부여하여 학습하므로써 독립적인 모델로 구성이 가능하다. 따라서 과적합(overfitting)에 강인하여 정확한 분류를 수행할 수 있다.



(그림 1) 의사결정나무 도식화 예제

3. 제안하는 기법

제안하는 침입탐지 모델의 구현은 랜덤포레스트 알고리즘의 학습이 선행되어야 하므로 데이터가 필수적이다. 스마트홈 서비스 환경에서 수집이 가능한 데이터를 <표 1>과 같이 정의한다.

<표 1> 스마트홈 환경의 데이터 스키마

Type	Description
time_stamp	접근 상세 시간
access_home_device	접근한 스마트홈 기기명
action	요청한 기기 조작 명령
access_device	접근 요청한 디바이스명
ip	ip 주소
protocol	프로토콜

위 정의된 데이터를 통해 사용자 개개인에 적용되는 맞춤형 침입탐지 모델을 구성한다. 학습에 필요한 데이터는 사용자 유형을 두 가지로 나누어 각각 학습데이터와 평가데이터로 활용한다. 예를 들어 유형1의 데이터는 주로 오후 3시 부터 오후 7시 까지 스마트홈 디바이스에 접근하는 데이터이다. 유형2의 데이터는 오전 2시부터 3시까지 집중적으로 스마트홈에 접근하는 데이터로 구성한다. 이처럼 서로 접근 시간 뿐 아니라 접근 디바이스, 조작 명령, 접근 요청 디바이스 데이터가 모두 상이하지 않지만 접근 패턴이 다른 두 가지 유형의 Dataset으로 학습과 평가를 진행한다. (그림 2)는 생성한 데이터를 기반으로 학습한 의사결정나무를 도식화 한 것이다.

(그림 2) 구현된 의사결정나무 도식화

(그림 2)의 의사결정나무는 최대 깊이를 4로 설정하였을 때 0.89로 가장 높은 정확도를 나타내었다. 또한, 데이터를 무작위로 분류하여 다수의 의사결정나무를 다시 구성하고 평균 수치를 종합하면 랜덤포레스트가 구현된다. 다수의 의사결정나무를 구현하고 평균치를 종합하여 정확도를 평가한 결과 정확도는 0.87로 하나의 의사결정나무보다 낮아졌다.

4. 결론

본 논문에서는 스마트홈 환경에서 기존 네트워크 침입탐지에 한계점에 대해 알아보고 관리자 계정을 탈취 당했을 경우 대응할 수 있는 방법이 미비하다는 것을 확인했다. 따라서 네트워크에 침입을 하여도 행동 로그를 분석하여 허가된 사용자인지 침입한 사용자인지 구분하는 방법을 제안하였다. 인공지능 알고리즘의 종류 중 하나인 랜덤포레스트를 기반으로 스마트홈 환경에서 침입을 탐지하는 모델을 구현한 결과 정확도가 0.87로 측정되어 높은 정확도로 침입을 탐지할 수 있음을 확인했다. 향후 스마트홈에서 발생하는 실제 데이터를 통해 학습을 진행하고 정확도를 측정하여 기존 스마트홈 서비스의 적용 가능 여부에 대해 연구가 필요하다.

ACKNOWLEDGMENT

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음” (IITP-2020-2020-0-01602)

참고문헌

[1] Imtiaz Ullah and Qusay H. Mahmoud, “A Two-Level Hybrid Model for Anomalous Activity Detection in IoT Networks,” 2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp.1-6, 2019

[2] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp.1-6, 2009

[3] Breiman, Leo. "Random forests." Machine learning 45.1, pp.5-32, 2001