LDA 특징 추출과 LBG 클러스터링 기법을 통한 네트워크 패턴 기반 IoT 기기 내 봇넷 탐지 연구

이민경○, 홍충선*
경희대학교 컴퓨터공학과
{immigrationk, cshong}@khu.ac.kr

A Study on Botnet Detection in IoT Devices Using LDA Feature extraction and LBG Clustering

MinKyung Lee○, ChoongSeon Hong*
Department of Computer Science and Engineering, Kyung Hee University

요 약

최근 사물인터넷 환경 구축으로 인해 많은 기기들이 인터넷 기반으로 연결됨에 따라 보안 취약성이 대두되고 있고, 2016년 3분기 IoT 기기 기반 봇넷을 통한 공격이 발생하며 IoT 보안이슈가 더욱 중요시되고 있다. 봇넷을 탐지하는 방법들 중 네트워크 플로우 특징 선택 기반 기법이 있다. 그러나 플로우 기반 특징 선택 기법은 다양하게 변형되는 봇넷들의 형태를 전부 학습하여 추출해내는 일에 시간과 비용적 문제가 있다. 또한, 수많은 네트워크 플로우 중 치명적인 하나의 플로우를 놓칠 경우가 발생한다는 위험 부담이 존재한다. 따라서 본 논문에서는 네트워크 패턴 기반의 특징 추출을 위한 LDA와 클러스터링을 위한 LBG 알고리즘을 적용하여 학습을 한 후, IoT 기기 내 봇넷 침입 시도를 학습된 데이터를 기반으로 탐지하는 방법을 제안한다.

1. 서 론

최근 사물인터넷(IoT)기반의 환경이 발달함에 따라 수많은 기기들이 인터넷을 통해 연결되고 있다. 가트너에 따르면, IoT 기기가 2017년 84억대, 향후 2020년에는 204억대 육박할 것이라고 전망한다[1].

그러나 IoT가 발달함과 동시에 보안의 취약성은 점차 대두되고 있다. 실례로 2016년 10월 미라이(Mirai) 멀웨어가 보안이 취약한 IoT 기기에 침입하여 봇넷(botnet)을 형성한 후, 분산 서비스 거부 공격(DDoS)을 시행하였다. 그 결과 미국 동부 지역 인터넷 서버가 마비되었고, 이는 약 600Gbps로 DDoS 공격 관측 사상 최대 규모였다. 또한 보안이 취약한 IoT 기기들이 봇넷이 되어 DDoS 공격을 위한 거점으로 이용될 수도 있다는 우려가 현실화된 사건이었다. 이 사건을 기반으로 변형된 봇넷을 통한 사물인터넷 기기 대상 DDoS 공격이 늘어날 것이라고 전망되고 있다[2]. 따라서 IoT 기기 내 다양한 형태의 봇넷을 탐지하는 시스템의 필요성이 대두되고 있다[3].

한편, 기존의 봇넷 탐지 기법 중 플로우 기반 특징 선택 탐지 기법이 있다. PC로 유입되는 다량의 네트워크 트래픽의 플로우를 통해 봇넷을 탐지하는 방안이다. 그러나 플로우 기반 기법은 여러 특징을 포함한 봇넷의 경우 대두되는 특징을 판단하기가 어렵다는 한계가 존재한다[4]. 더불어 기존의 탐지 기법들을 IoT 환경 내에서가 아닌, PC 상 웹 기반의 봇넷 탐지 기법들로 진행된 연구라는 한계가 존재한다.

따라서 본 논문에서는 다양한 네트워크 패턴을 지닌 봇넷들의 특징을 LDA(Latent Dirichlet Allocation) 알고리즘을 통해 추출한 후, 이를 LBG(Linde-Buzo-Gray) 알고리즘 통해 클러스터링 하는 학습과정을 거쳐 새로운 봇넷이 침입을 시도할 때, 학습된 데이터 셋을 기반으로 이를 탐지하는 방법을 제안한다.

2. 관련 연구

2.1 플로우 기반 특징 선택을 통한 봇넷 탐지 기법

플로우 기반 특징 선택이란 네트워크 플로우 상에서 봇넷들의 특징을 추출하여 그룹화 하는 알고리즘이다. Elaheh Biglar Beigi 등은 [4]에서 세 단계로 구성된 플로우 기반 특징 선택을 통한 봇넷 탐지 방안 알고리즘을 제안하였다.

1) 결정트리를 기반으로 탐지의 정확도를 높인다.
2) 효과 평가를 위한 탐지율과 거짓 양성율 기반 분류의 정확도를 측정한다.
3) 특징 선택 알고리즘을 토대로, 그룹제거 및 특징 포함 단계를 수행한다. 그룹제거는 정확도를 저하시키는 특징들을 그룹화 시켜 배제하는 분류법이고 특징 포함은 정확도 향상을 위해 개체 특징들을 선택하여 그룹화 하는 분류법이다.

Exp	Byte based	Packet based	Time based	Behavior based	Detection Rate
Exp1		√	√	√	68.82%
Exp2	√		√	√	68.58%
Exp3	√	√		√	65.44%
Exp4	√	√	√		66.01%

표 1. 플로우 기반 특징 분류

본 논문은 산업통상자원부 산업핵심기술개발사업으로 지원된 연구결과입니다[10049079, 퍼스널 빅데이터를 활용한 마이닝 마인즈 핵심 기술 개발]
*Dr. CS Hong is the corresponding author

[4]는 표 1에서 제시된 바와 같이 봇넷들을 바이트 기반 특징, 패킷 기반 특징, 시간 기반 특징, 행동기반 특징 등으로 분류하였다. 그 후, 플로우 기반 특징 분류 값을 통해 특정트래픽이 유입될 시 해당 데이터 셋과 비교하여 봇넷 여부를 탐지하는 방안으로 제안된 기법이다[4].

2.2 LDA (Latent Dirichlet Allocation)

LDA(Latent Dirichlet Allocation)는 비지도학습 중 특징 추출 알고리즘으로, 문서 등의 데이터 집합에 대한 일반적인 확률 모델이다. 주어진 데이터인 문서를 분석하여 Topic 집단을 구성한 후, 이를 토대로 데이터들의 분포 지점을 구하여 특징 기반의 데이터들을 추출해내는 알고리즘으로 일반적 모델은 그림 1과 같다[5].

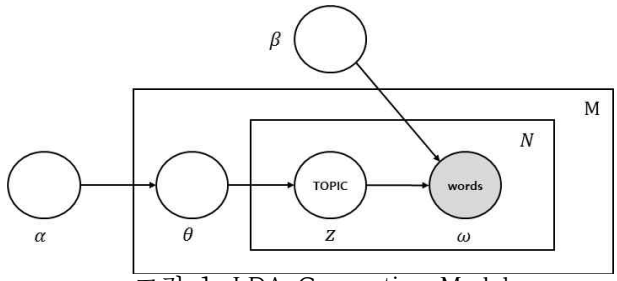

그림 1. LDA Generative Model

LDA Generative Model에서는 숨겨진 변수 또는 숨겨진 파라미터 전체를 고려하여 관측 값이 생성되는 과정을 나타낸다. 직접 구할 수 없는 사전·사후분포 및 가능성 또한 간접적으로 찾아주는 알고리즘이다[5].

2.3 LBG (Linde-Buzo-Gray)

LBG(Linde-Buzo-Gray)는 클러스터링 알고리즘으로, 초기값에 민감한 k-means 알고리즘을 개선한 것이다. 표준 k-means 알고리즘을 이용할 시보다, 속도 향상과 클러스터 분할 수행이 잘된다는 장점을 지니고 있다. k-means의 초기값을 이진 분할(binary split)로 구한 중심을 사용하여 k-means와 이진분할을 결합하여 최적화를 시도한 알고리즘이다[6].

3. 제안 사항

본 연구는 LDA알고리즘을 통해, 네트워크 패턴기반으로 봇넷 특성을 파악한 후, 얻은 결과값을 LBG알고리즘을 통해 클러스터링함으로 새로운 봇넷을 탐지하는 방안을 제안한다.

LDA를 통해 봇넷 침입 시 발생하는 네트워크 이상 현상 패턴을 분석하여 봇넷 특징을 추출하는 과정을 수행한다.

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1} \quad (1)$$

$$p(\theta, z, \omega|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(\omega_n|z_n, \beta) \quad (2)$$

$$p(\omega|\alpha, \beta) = \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_n}p(z_{dn}|\theta_d)p(\omega_{dn}|z_{dn,\beta})\right) \quad (3)$$

수식 (1)은 주어진 봇넷 데이터 셋 문서 집합의 전체 길이와 이를 기반으로 주어진 문서의 토픽 분포를 알 수 있다. 수식 (2)는 행렬로 구성된 문서 내 관측된 데이터 값들을 통해 알맞은 토픽 변수를 구하여준다. 수식 (3)은 해당 문서의 단어와 토픽들의 변수를 토대로 데이터들이 토픽에 따라 해당 단어를 생성할 확률을 정의해준다.

표 2. 변수 및 의미의 정의

변수	의 미
θ	주어진 문서i의 토픽 분포 랜덤 변수
α	$\alpha_i > 0$ 일 때, k 벡터 파라미터
k	디리클레 분포의 차원 수, 토픽의 개수(고정값)
z	문서i의 j번째 단어의 토픽 변수
ω	N의 단어 수, 관측된 단어 데이터
β	토픽 z에 따른 단어ω의 생성확률(k*V)
N	주어진 문서 집합의 전체 길이
α_k	문서 당 토픽k의 디리클레 사전 가중치
z_n	n번째 단어의 토픽 변수
θ_d	문서 당 샘플화 된 문서 단위 토픽의 비율
z_{dn}	문서d의 n번째 단어의 토픽 랜덤 변수
ω_{dn}	문서d의 n번째 단어 수준 변수

알고리즘 Process for each document ω

```
B ← flowed Botnet
BNP ← Botnet Network Pattern
TNW ← Total Number of Words in documents
ST ← Specific Topic (Index)
WM ← Words in Model
TP ← Topic Proportion in specific document
P ← Parameter
IF BNP from B is detected in IoT devices
   calculate TNW using Poisson
   IF TNW is calculated
    WM is from ST and TP is from P using Dirichlet
    FOR each of TNW words WM
        IF choose ST
        use Multinomial TP
        ELSE IF choose WM from p(WM|ST)
        use Multinomial probability conditioned on ST
```

알고리즘 과정을 통해, IP기반 혹은 Packet기반 등 다양한 네트워크 패턴을 지닌 봇넷의 토픽과 TCP/UDP로 유입되는 다량의 네트워크 패턴들이 특징이 추출된다. 마지막으로 LBG알고리즘을 통해, 네트워크 패턴 특징 별 클러스터링 과정을 수행한다.

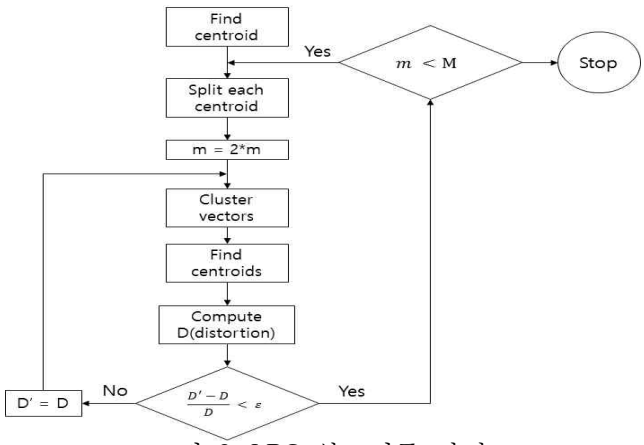

그림 2. LBG 알고리즘 과정

그림 2는 네트워크 패턴 특징 기반으로 추출된 데이터(라벨 값)들이 LBG 알고리즘을 거쳐 클러스터링 되는 과정을 보여준다. 이 때, 전달되는 패턴 특징 별 토픽 값들 중 클러스터의 중심을 이룰 수 있는 부분들을 찾는다. 토픽 값을 찾은 다음 패턴 별 클러스터 중심으로 토픽들이 모이게 된다. 이와 같은 과정을 반복하여 거치게 되면, 유입된 네트워크 패턴에 맞은 클러스터들이 형성되고, 이를 토대로 봇넷 유입 유무와 패턴 종류를 탐지하게 된다.

4. 성능 평가

4.1 시뮬레이션 환경

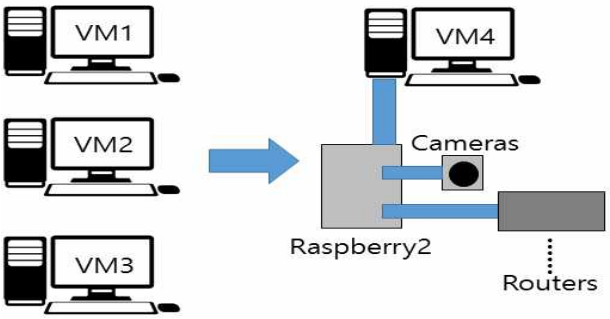

그림 3. 시뮬레이션 환경 구축

본 논문의 연구를 수행하기 위해 그림 3과 같은 시뮬레이션 환경을 구축하였다. 가상머신(VM) 1, 2, 3을 통해 봇넷, TCP, UDP 등 다량의 네트워크 트래픽이 라즈베리파이에 설치된 카메라 모듈과 라우터 모듈에 유입된다. 이 때, 파이와 연결된 가상머신(VM) 4를 통해 본 논문이 제안한 LDA와 LBG 알고리즘이 수행되며, 카메라 모듈과 라우터 모듈에 유입되는 다량의 네트워크 트래픽을 분석하여 봇넷 여부를 탐지한다.

4.2 성능 평가

본 장에서는 4.1에서 제시한 환경 구축을 바탕으로 Netflow 툴을 통해 구축한 데이터 셋의 네트워크 패턴 별 봇넷과 제시한 환경 내 제안한 머신러닝 알고리즘을 통하여 분류한 결과 값의 정확도를 측정하였다.

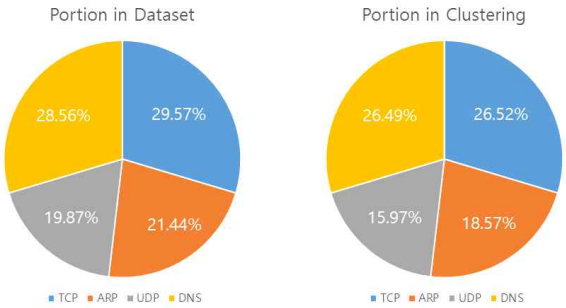

그림 4. 데이터 셋과 클러스터링을 통한 분류 비율 비교

그림 4는 Netflow를 통해 만든 데이터 셋의 네트워크 패턴 별 비율과 제안한 알고리즘을 통해 IoT 기기 단에 유입된 네트워크 패턴 별 분류 비율의 정도를 비교한 것을 보여준다. 네트워크 패턴의 경우, 이상행위를 발현하는 프로토콜을 기반으로 기준을 정하였다. Netflow를 이용하여 패턴 별로 문서화 한 데이터 셋의 비율과 알고리즘을 통해 IoT 기기 내에서 탐지된 봇넷의 비율 간의 차이 존재하나, 그 값이 크지 않음을 확인할 수 있다.

5. 결 론 및 향후 연구

본 논문은 보안이 취약한 IoT 기기 대상 봇넷 유입 시, Latent Dirichlet Allocation 알고리즘과 Linde-Buzo-Gray 알고리즘을 토대로, 네트워크 패턴을 특징 별로 추출한 후, 이를 클러스터링 하여 봇넷 여부를 탐지하는 방안을 제안하였다. 4장에서는 성능평가를 통해 IoT 기기 내 다량의 네트워크 트래픽이 유입될 경우, 이를 패턴 별로 클러스터링하여 탐지하는 방안을 정확도를 통해 검증하였다. 본 논문에서 제시한 네트워크 패턴 기반의 봇넷 탐지는 주어진 데이터 셋 기반 수행 시 정확도가 우수하나, 봇넷 소스 기반 데이터 셋 및 다양한 기기 모듈이 존재하는 IoT 환경 구축이 어렵다는 제약이 존재하였다.

따라서 향후 연구에는 카메라 및 라우터 모듈 외의 다양한 IoT 환경 구축 및 최근 이슈가 되고 있는 IoT 봇넷 관련 데이터 셋을 마련하여 IoT 기기 내 다량의 트래픽 유입 시 봇넷 여부를 탐지한 후 방어하는 방안을 진행하고자 한다.

참 고 문 헌

[1] Gartner, “Gartner Says 8.4 Billion Connected “Things“ Will Be in Use in 2017, Up 31 Percent From 2016”, availalble at : http://www.gartner.com/newsroom/id/3598917, Accessed on : February, 7, 2017

[2] James A.Jerkins, “Motivating a Market or Regulatory Solution to IoT Insecurity with the Mirai Botnet Code, 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), Pages: 1 - 5, DOI: 10.1109/CCWC.2017.7868464

[3] Sean Miller et al, “The Role of Machine Learning in Botnet Detection”, The 11th International Conference for Internet Technology and Secured Transactions, ICITST-2016, Pages: 359- 364, DOI: 10.1109/ICITST.2016.7856730

[4] Elaheh Biglar Beigi et al, “Towards Effective Feature Selection in Machine Learning-Based Botnet Detection Approaches”, 2014 IEEE Conference on Communications and Network Sercurity, Pages: 247 - 255, DOI: 10.1109/CNS.2014.6997492

[5] David M. Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3 (2003), Pages: 993-1022, Submitted 2/02, Published 1/03

[6] LBG Algorithm, available at : http://www.cs.nthu.edu.tw/~cchen/CS6631/2012/LBG.pdf Accessed on : 2012