Journal of The Korea Society of Computer and Information
Vol. 19, No. 4, April 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.4.081

모바일 봇넷 탐지를 위한 HMM과 SVM 기법의 비교

최 병 하*, 조 경 산**

Comparison of HMM and SVM schemes in detecting mobile Botnet

Byungha Choi *, Kyungsan Cho **

요 약

스마트폰 같은 모바일 장치의 대중적 보급과 발전으로 인해 PC 기반의 악성코드가 모바일 기반으로 빠르게 이동하고 있다. 특히 봇넷은 PC에서의 강력한 악성행위와 피해를 모바일 장치에서 재생산하며 새로운 기법을 추가하고 있다. 기존 PC 기반의 봇넷과 달리 모바일 봇넷은 동시에 다양한 공격 경로의 탐지가 어려워 네트워크 기반보다는 호스트 기반의 탐지 기법이 주를 이루고 있다. 본 논문에서는 호스트 기반 기법의 한계를 극복하기 위하여 네트워크 기반으로 모바일 봇넷을 탐지하는 HMM과 SVM을 적용한 2 가지 기법을 비교한다. 기계학습에 많이 사용되는 시계열 데이터와 단위시간 데이터를 추출하여 두 기법에 적용하여, 실제 봇넷이 설치된 환경의 트래픽 검증 분석을 통해 이들 데이터에 따른 두 기법의 탐지율과 탐지 특성을 제시한다.

▸ Keywords : 모바일 봇넷, 탐지 시스템, HMM, SVM, 탐지율

Abstract

As mobile devices have become widely used and developed, PC based malwares can be moving towards mobile-based units. In particular, mobile Botnet reuses powerful malicious behavior of PC-based Botnet or add new malicious techniques. Different from existing PC-based Botnet detection schemes, mobile Botnet detection schemes are generally host-based. It is because mobile Botnet has various attack vectors and it is difficult to inspect all the attack vector at the same time. In this paper, to overcome limitations of host-based scheme, we compare two network-based schemes which detect mobile Botnet by applying HMM and SVM techniques. Through the verification analysis under real Botnet attacks, we present detection rates and detection properties

•제1저자 : 최병하 •교신저자 : 조경산
•투고일 : 2013. 12. 5, 심사일 : 2014. 1. 22, 게재확정일 : 2014. 3. 2
* 단국대학교 정보통신융합기술연구원 (Research Institute of Information and Communication Convergence Technology)
** 단국대학교 소프트웨어학과(Dept. of Software Science, Dankook University)

of two schemes.

▸ Keywords : Mobile Botnet, Detection System, HMM, SVM, Detection Rate

I. 서 론

"malicious code" 또는 "malware"로 불리는 악성코드는 운영체제 커널이나 보안에 민감한 애플리케이션의 동작을 변화시키는 코드로써, 사용자의 동의 없이 이루어지거나 운영체제 또는 응용프로그램의 문서화된 기능(예: API)을 사용하여 그러한 변화를 탐지할 수 없도록 행해지는 프로그램이다. 또한, 악성코드는 정보유출과 금전적 이익 등의 악의적 목적으로 사용되며 최근 그 피해가 모바일 장치로 이동하고 있다[1,2].

악성코드는 일반적으로 기능에 따라 다음과 같이 5가지로 분류할 수 있다[1,2]. 1) 유용한 프로그램으로 가장하여 컴퓨터 시스템에 침투하는 트로이목마(trojan horse), 2) 다른 파일에 삽입되어 자기복제 하는 바이러스(virus), 3) 독립적으로 자기복제 하는 웜(worm), 4) C&C(Command and Control) 서버에서 원격으로 명령을 제어하는 봇의 네트워크인 봇넷(Botnet), 5) 루트(Root) 권한을 탈취하는 루트킷(Rootkit). 이들 악성코드들은 여러 기능을 혼합하여 공격한다[3]. 예를 들면, 봇넷은 트로이목마처럼 침입하여 루트권한을 탈취하고 자기복제하여, 봇들을 통해 DDOS (Distributed Denial Of Service), 스팸 메일, 정보 유출 등 다양한 피해를 발생시킨다. 이러한 과정에서 봇넷은 C&C 서버의 명령으로 조직적인 공격을 수행하므로 다른 악성코드보다 더 위험하다. 이의 대응책으로 호스트 기반과 네트워크 기반의 탐지 기법이 가능하다.

스마트폰의 대중적인 이용 증가에 따라 모바일 봇넷도 증가하고 있다. 그러나 모바일 장치의 다양한 공격 경로(SMS, MMS, 3/4G, WiFi)를 동시에 탐지하는 것은 어려우므로 모바일 봇넷 탐지의 대부분은 호스트 기반이다. 그러나 C&C 서버의 명령에 의한 DDOS, 스팸 메일 등의 공격을 호스트 기반으로 탐지해도, 이들 봇들이 미확인 네트워크에 존재한다면 네트워크 전체에 트래픽 체증 등의 문제를 발생시킬 수 있다. 또한 호스트 기반 기법은 모바일 장치의 배터리와 컴퓨팅 자원을 소진할 수 있다. 따라서 모바일 네트워크와 장치의 원활한 서비스를 위하여 네트워크 기반의 봇넷 탐지 기법이 필요하다.

본 논문의 선행연구로 모바일 장치와 IDS(Intrusion Detection System)를 VPN으로 연결하여 봇넷의 C&C 트래픽의 탐지기반을 구축하였다[2]. 본 논문은 이를 기반으로 대표적인 기계학습 기법인 HMM(Hidden Markov Model)과 SVM(Support Vector Machine)의 두 탐지 기법을 제안하여 비교한다. 동일한 기계학습 기법이라도 사용되는 특징(features)에 따라 그 성능이 달라지므로, 본 연구는 봇넷 C&C 트래픽을 시계열(time series) 특징과 일반적으로 사용되는 단위 시간의 트래픽 특징의 2가지로 구성하여 HMM과 SVM의 탐지기법에 적용한다. 또한 적용한 결과를 기반으로 두 기법의 특성을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 봇넷과 기존 탐지 기법을 분석하고 제시한다. 3장에서는 VPN을 이용한 모바일 봇넷 탐지 환경을 구축하여 HMM과 SVM의 탐지 기법을 제안하여 비교한다. 4장에서는 제안 기법의 장단점을 비교 분석하고, 5장의 결론으로 본 논문을 마무리 짓는다.

II. 관련 연구

1. PC 기반의 봇넷 특성

봇넷은 악성코드의 한 유형이며 감염된 컴퓨터 또는 악성 프로그램인 봇(Bot)의 집합으로 이루어진다. 봇은 C&C 채널을 통해 C&C 서버의 통제 아래 원격 조정되고 있으며, C&C 서버들은 봇마스터(Bot master)에 의해 그림 1과 같이 운영된다.

봇은 연결 구조에 따라 중앙 집중적 봇넷(centralized Botnet)과 P2P 봇넷(P2P Botnet, decentralized Botnet)으로 분류할 수 있다. 중앙 집중적 봇넷은 봇마스터가 몇몇의 C&C 서버를 운영하고 그 C&C 서버가 다시 여러

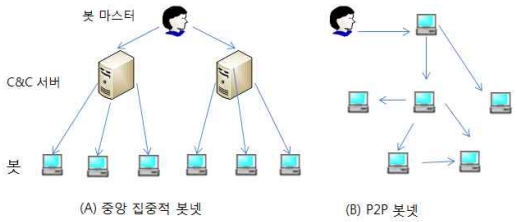

그림 1 봇넷의 구조
Fig. 1. Botnet Architecture

봇을 C&C 채널을 통해 관리를 하는 구조이다. 그러므로 명령 전달이 신속하고 집단적인 악성 행위를 하기에 적절한 구조이다. 프로토콜은 HTTP와 IRC 기반의 프로토콜이 대표적이며 Pull 방식으로도 불리는 HTTP 기반 봇은 주기적인 시간 간격으로 C&C 서버에게 연결을 요청하여 C&C 명령을 받아오고, Push 방식이라고도 불리는 IRC 기반 봇은 연결된 상태를 유지하며 C&C 서버가 명령을 봇에게 전달한다. 반면에 P2P 봇넷은 C&C 서버가 봇을 직접 관리하는 것이 아니라 봇들이 서로에게 명령을 전달하는 체계를 가지므로, 중앙 집중적 봇넷 보다는 유연성이 있다. 즉 한 봇이 제거되어도 전체 연결 구조는 붕괴되지 않으며, 다만 봇마스터가 명령을 전달하였을 때 언제 전체 봇으로 전달되고 언제 실행하는지 알 수 없으며, 봇넷 전체의 유지보수가 어렵다는 단점이 있다[2,4]. P2P 봇넷에서는 UDP 기반과 TCP 기반의 다양한 프로토콜이 사용된다[4].

봇넷의 구조 중에서 C&C 채널은 C&C 서버와 봇 그리고 봇마스터를 연결시키는 경로로, 봇넷의 가장 취약한 부분이다. 따라서, 중앙 집중적 봇넷에서는 이를 탐지하면 탐지된 C&C 서버를 통해 봇넷 전체를 탐지 및 추적할 수 있다[4].

봇넷의 탐지 기법은 이미 알려진 악의적인 행위에 대한 특정 패턴과 비교하는 시그너처 기반 탐지와 네트워크의 트래픽 흐름 및 응용 프로그램 정보 등에 대한 정상행위 모델과 비교하는 비정상 행위 기반 탐지가 제시되었다[5]. 시그너처 기반 탐지는 오탐이 낮고 정확한 탐지를 제공하지만 새로운 기법의 악성행위는 탐지하기 힘들다. 반면, 비정상 행위 기반 탐지는 새로운 기법의 악성 행위는 탐지가 가능하나 오탐이 많다. 또한 탐지 위치에 따라 호스트 기반과, 네트워크 기반의 탐지기법이 제안되었다. 호스트 기반으로 Anti-virus는 시스템 내부에서 악성코드를 찾는 전통적인 기법으로 여러 기업이 제공하고 있으며, 호스트의 비정상 행위 탐지로 SystemAPI-call의 로그(log) 정보를 이용한 탐지 기법이 제안되었다[4]. 네트워크 기반으로 Snort는 네트워크 기반에서 가장 널리 알려진 침입 탐지 시스템으로써 미리 정해진 시그너처로 탐지한다[4]. Botminer는 봇의 C&C 트래픽과 공격 트래픽의 로그 정보를 각각 클러스터링 기법에 적용하여 봇넷을 판단한다[6].

2 모바일 봇넷의 특성

모바일 봇넷은 심비안(Symbian)에서 처음 발견된 이후 다른 모바일 장치로 활성화되고 있다. 2009년 심비안의 SymbOS. Yxyes를 시작으로 같은해에 iOS에서 Ikee.B라는 봇넷과 2010년 Geinimi가 안드로이드에서 발견되었다. 모바일 장치는 사용자의 사생활과 사적인 정보를 가지고 있으며, 효과적인 인터넷과 컴퓨팅 자원을 사용할 수 있기 때문에 봇넷의 매력적인 목표가 되어 왔다[2]. 모바일 봇넷은 PC 기반의 봇넷과 몇가지 차이점이 있다. 첫째는 PC 기반 봇넷과 달리 모바일 봇넷은 SMS, WiFi, 3/4G 등의 다양한 공격 경로를 가지며 둘째는 PC에서는 IRC 기반의 봇넷이 많지만 모바일 봇넷은 HTTP 기반이 대부분이다. 실제 C&C 채널을 경로로 사용하는 대부분 모바일 악성코드 중에서 90%이상이 봇넷으로 전환가능한 패킷의 페이로드(payload)를 가지고 있으며, 이들 대부분이 HTTP 기반 패킷이라고 분석되었다[7]. 모바일 봇넷은 이동성과 모바일 네트워크 환경의 자주 변경되는 IP 주소로 인해 지속적인 연결이 되지 않으므로, HTTP 기반의 Pull 방식의 정기적인 연결 요청으로 봇넷을 유지하것으로 분석되었다[2].

모바일 봇넷의 대표적인 기법은 표 1과 같다. 이들 탐지 기법 중 대부분은 모바일 악성코드를 탐지하는 호스트 기반 기법이며, 순수하게 모바일 봇넷을 위한 기법은 SMS spam-detection 기법이다. 또한 PC 기반 기법과 유사하게 Anti-virus가 제시되었고, 탐지기 Crowdroid는 K-means 기법으로 훈련하여 탐지하는 비정상 행위 기반 탐지 기법이다. 이와 유사한 기법으로 호스트 기반의 HMM 기반과 SVM 기반의 비정상 행위 기반 기법이 제시되었다[8,9,10].

호스트 기반 탐지기법은 모바일 네트워크 악성트래픽에 의한 트래픽 체증 등을 해결할 수 없는 한계가 있으며, 클라우

표 1. PC 모바일 봇넷의 탐지 기법
Table 1. Mobile Botnet Detection Schemes

기법		시그너처 기반	비정상 행위 기반
호스트 기반		SMS spam-detection[13]	Crowdroid[8],클라우드기반[11], HMM 기반[9], SVM 기반[10]
네트워크 기반	3/4G	N/A	scan detection[12]
	WiFi, 블루투쓰	N/A	

드 기반의 기법은 비동기화 되었을 때 공격에 취약하다. 또한 대부분의 호스트 기반은 다른 OS와 다양한 모바일 장치에 적용하기가 어려우며 모바일 장치의 컴퓨팅 자원의 한계 등이 문제가 될 수 있다. 아울러 3/4G의 탐지 기법과 클라우드 기법의 경우 3/4G의 네트워크 트래픽을 사용한다면 이에 따른 과금이 발생할 수 있다. 모바일 봇넷 탐지에 기존 PC 기반에서 우수한 탐지율을 보인 HMM과 SVM의 기법을 사용할 수 있으나, 표 3의 기존 HMM 기반과 SVM 기반 기법은 호스트 기반의 기법으로 적용되었다.

이들 HMM과 SVM 중 어떤 기법이 우수한지는 단순 비교는 어렵고 표 2에서처럼 적용대상에 따라 다를 수 있다.

표 2. HMM과 SVM 기법의 비교
Table 2. Comparision of HMM and SVM

적용 환경 또는 대상	우수기법
필기체의 서명 인식[14]	SVM
음성을 이용한 감정인식[15]	HMM
기어박스 결함 탐지[16]	SVM

III. 비교를 위한 제안 탐지 기법

본 장에서는 HMM과 SVM의 모바일 탐지에 적용하고 이들중 어떤 것이 우수한지 비교하기 위하여 다음과 같은 시스템 을 제안한다. 제안 시스템의 주요한 구성요소인 VPN과 IDS의 구조를 제시하고 이를 이용하여 3/4G와 WiFi에서 모바일 봇넷 C&C 채널을 탐지할 수있는 HMM 기반의 HDS (HMM-based Detection Scheme)와 SVM을 기반으로 구현한 SDS (SVM-based Detection Scheme)를 제안한다.

1. 제안시스템의 구조

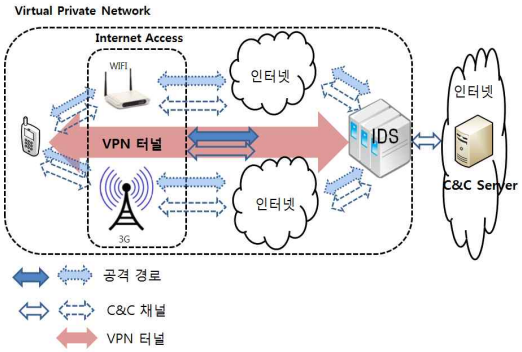

그림 2 제안 시스템 환경
Fig. 2. Environment of Proposed System

제안 시스템의 환경은 그림 2와 같이 PPTP 기반의 VPN을 설치하여, WiFi 또는 3/4G에서 모바일 장치가 IDS를 통해 인터넷에 연결되도록 한다. 봇넷 C&C 트래픽을 탐지하기 위해 IDS에서 트래픽의 플로우(flows)를 수집하여 MySQL에 저장할 수 있도록 그림 3과 같이 PMACCT (Promiscuous Mode IP Accounting Package)를 설치한다. PMACCT는 플로우의 바이트, 패킷수, 플로우 수 등의 다양한 특징 추출과 시그너처를 탐지 할 수 있는 기능을 가지며, 이들 특징들을 저장한다. 기본적으로 저장되는 정보는 플로우에 포함된 패킷수, 바이트수 등이며 SQL(Structured Query Language) 질의문을 이용하여 시간당 패킷수의 양과 평균, 바이트의 양, 평균 같은 시간 단위의 특징으로 생성이 가능하다.

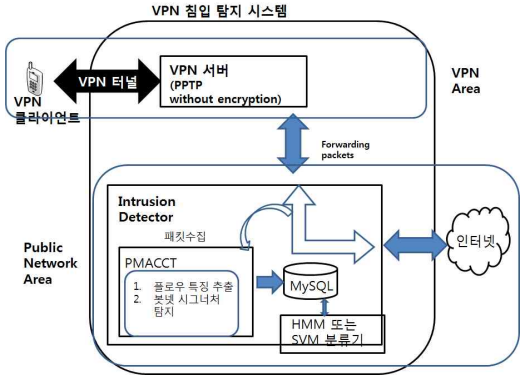

그림 3 IDS의 구조
Fig. 3. IDS Architecture

본 연구에서는 HMM과 SVM을 적용하기 위해 시계열 특징과 시간 단위의 트래픽의 특징을 사용하여 구현된 HMM 기반의 HDS와 SVM 기반의 SDS는 그림 4와 같이 탐지 과정을 수행한다. 그림 2의 환경에서 모바일 악성 코드의 데이터셋을 수집하고, 이를 Wireshark[17]로 분석하여 정상 또는 비정상트래픽(C&C 트래픽)으로 분류하고, 블랙리스트(blacklist)를 작성한다. 블랙리스트의 C&C 트래픽만을 훈련하여 HMM 또는 SVM의 각 기법별로 비정상 행위의 모델을 생성한다. HDS와 SDS 분류기에서 각 기법에서 제공하는 비정상 모델을 이용하여 서로 비교하거나 확률을 구하여 탐지할 수 있다.

2. HMM 기반과 SVM 기반의 탐지

비정상 행위 기반 탐지 기법을 위한 기계학습 기반으로 다양한 알고리즘이 존재하며 모든 상황에 우수한 성능의 알고리

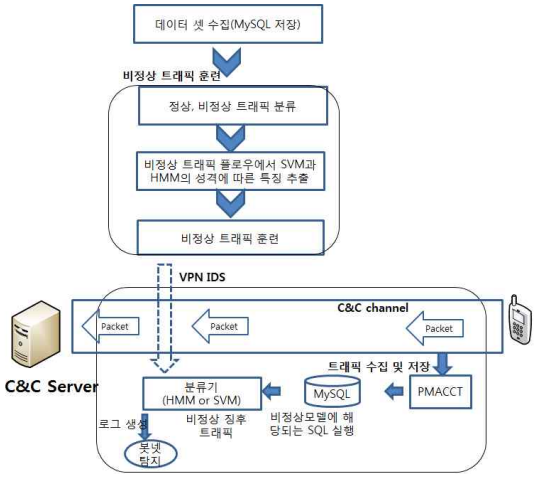

그림 4 HDS와 SDS의 탐지과정
Fig 4. Detection process of HDS and SDS

즘은 찾기 힘들다. 다만 탐지 대상의 속성, 특징벡터의 입력 순서, 탐지기법의 활성 함수 등을 어떻게 최적화시키느냐에 따라 알고리즘의 성능은 좌우된다. HTTP 기반의 봇넷의 특징은 규칙적인 주기로 C&C 서버에 명령을 요청하는데, 그 요청의 패킷 수와 바이트 수는 비슷하거나 동일한 경우가 많다. 이러한 속성에는 시간적인 규칙이 있으므로, 시계열 기반의 HMM을 적용할 수 있다. 또한 관련연구에서는 SVM 기법 또한 높은 탐지율을 보여주고 있다[10].

2.1 데이터의 분석

공학 분야 외에 다양한 분야에서 실태 및 현상 분석과 효과적인 예측 또는 정책 등을 수립하기 위해 데이터의 심층적인 분석이 요구되고 있다. 이를 위해 시간의 흐름에 따라 데이터가 어떻게 변화하는지 보여주는 시계열 데이터(Time series data)와 일정 시점에서 데이터가 어떤 값을 갖는가를 확인하는 횡단면 데이터(Cross section data), 마지막으로 동일한 표본으로부터 지속적으로 동일한 문제에 대해 일정기간마다 얻어지는 횡단면 데이터를 패널 데이터(Panel data)라고 한다[21]. 즉 패널 데이터는 시계열 데이터와 횡단면 데이터의 정보를 모두 내포하고 있다. 본 연구에서는 네트워크 트래픽을 시계열 데이터와 일정 시간별로 트래픽을 추출한 패널 데이터(단위시간 데이터)를 기계학습 특징(features)으로 추출하여 분석한다.

2.2 HMM 기반 탐지와 특징 추출

HMM은 숨겨진 모수를 결정하기 위해 관찰이 가능한 기호로 모델링하는 이중의 확률적 과정이다. 이는{S, V, π, A, B}로 표현되며 각 구성요소는 표3과 같은 의미를 가진다[9]. 이들 5가지 구성요소로 특정 모델 λ를 규정할 수 있는데, 이는 $\lambda = (\pi, A, B)$ 으로 간단하게 나타낸다.

표 3. HMM의 5 가지 구성요소
Table 3. 5 tuples of HMM

구성요소	의미
S	N개의 은닉상태 값의 집합= $\{S_1, S_2, \cdots, S_n\}$
V	M개의 관측 가능한 관찰기호의 집합=$\{v_1, v_2 \cdots v_m\}$
π	각 상태의 초기 확률 집합= $\{\pi_1, \pi_2, \cdots \pi_i\}$
A	상태 천이 확률의 집합 = $\{a_{ij}\}$ aij : 상태 S_i 에서 S_j 로 천이할 확률 $=a_{ij}=P(qt+1=S_j \mid qt=S_i)$, $1 \le i, j < N$
B	상태에 대한 출구 확률의 집합 = $\{b_i(v_k)\}$ b_i: 상태 (k) Si에서 관찰기호 v_k를 출력할 확률

HMM은 널리 알려진 학습 방법으로 두 가지로 Baum-Welch 기법과 Segmental K-means 기법이 있다. HMM의 전통적인 학습 방법인 Baum-Welch 기법은 최대 우도 평가(Maximum likelihood estimation)를 사용하는데, 인식률을 극대화하는 모델의 파라미터 값을 신속하게 생성하지 못하는 단점이 있다. 이를 해결하기 위해 Segmental K-means 학습 기법이 제안되었으며, 학습시에 모든 경로에 대한 우도를 이용하는 Baum-Welch 기법보다 최적의 분할 경로의 수렴으로 효율적인 계산과 신속함을 수학적으로 증명되었다[18].

HMM은 단위 시간 데이터와 시계열 데이터를 모두 적용 가능하다. 표 4의 단위 시간 특징을 이용하는 것으로, IP 주소와 포트 번호를 기준으로 플로우를 군집화하여, 1 시간 동안 전송된 패킷수, 바이트수, 패킷수의 평균량, 바이트 수의 평균량, 플로우 사이의 시간 간격으로 구성된다. 이들 단위 시간 특징은 PC 기반 봇넷을 탐지하는 다른 연구에서도 많이 사용되는 일반적인 특징들이다[6]. 이들 특징을 "포트" - "http_user_ agent" - "시간당 전체 패킷수" - "시간당 전체 바이트수" - "플로우 평균 시간 간격" - "시간당 평균 플로우 수" 와 같이 연결하여 HMM의 시퀀스로 적용 가능하다.

HMM은 시계열의 데이터를 기계학습의 특징(features)으로도 추출하여 인식할 수 있다. 즉 공격시에 어떤 특징이 시간에 따라 일정한 패턴의 순서가 존재할 경우에는 HMM의 시퀀스에 적용하여 특정한 공격을 탐지할 수 있다. 따라서 하나의 플로우에서 다음과 같이 TCP flags, 플로우의 패킷수, 플로우의 바이트양, HTTTP의 user-agent, 발생시간 간격 등을 배열 형태로 하나의 관찰 기호로 만들고, 1시간동안의

표 4. 각 기법별 특징
Table 4. features of each scheme

특징	특징 항목	내용
시계열 특징	http user-agent	1. 이들 특징을 배열형태의 하나의 관찰기호로 설정하고, 시퀀스 생성시 이들 기호를 연결한다. 2.시퀀스 생성시 1시간 동안의 동일한 포트와 아이피의 관찰기호를 연결하여 생성, 최소 1시간 동안 3 개 이상의 플로우가 존재해야 시퀀스 생성가능
	Tcp_flags	
	플로우의 패킷	
	플로우의 바이트수	
	동일 아이피와 포트의 이전 플로우와 시간간격	
단위 시간 특징	포트(port) 번호	1. 이들 특징을 1시간 단위로 SQL 질의문을 이용하여 생성 2. 12시간 이내에 동일 IP 주소와 port로 3번 이상의 플로우가 존재해야 탐지 대상이 됨
	http user-agent	
	전체 패킷수/시간	
	전체 바이트/시간	
	평균 패킷수/시간	
	평균바이트수/시간	
	1시간동안 동일 IP, port의 플로우 평균시간 간격	
	평균플로우수/시간	
	전체플로우수/시간	

관찰기호를 연결하여 다중 관찰 시퀀스(sequence of multivariate observations)로 생성한 후 Baum-Welch 또는 Segmental K-means 학습 기법으로 훈련하여 탐지할 수 있다. 관찰기호에서 사용된 특징들은 표 4의 시계열 특징을 이용하여 그림 5의 다중관찰 시퀀스 생성과정으로 나타낼 수 있다.

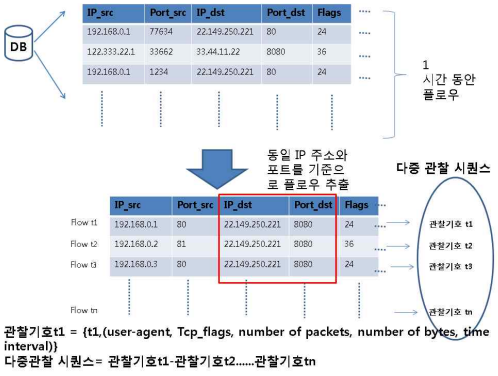

그림 5 시계열의 다중 관찰 시퀀스 생성과정
Fig. 5. process of creating a sequence of multivariate observations

본 연구에서 제안하는 HMM 기반의 HDS에서는 C&C 채널의 탐지를 위해 시계열 또는 단위 시간의 시퀀스를 생성하고, 훈련을 통해 비정상 트래픽의 모델인 λ를 생성한다. 실제 탐지시에도 트래픽을 시퀀스로 변환하여 모델 λ에서 해당 시퀀스가 발생할 확률을 구하는 확률 평가(Probability Evaluation)를 이용하여 C&C 트래픽을 탐지한다.

2.3 SDS 기반 탐지와 특징 추출

본 절에서는 그림6의 탐지 과정을 구현한 SVM 기반의 SDS를 제안한다. SVM은 어떤 그룹들을 분류할 때 서로의 간격(margin)을 가장 크게 하는 분류 초평면을 찾는 기법으로, 일반화 능력이 뛰어나고, 다양한 분야에서 우수한 성능을 나타낸다[16].

SVM은 선형분리가 불가능할 경우 커널함수를 이용하여 비선형으로 분리하여 문제를 해결할 수 있는데, 주로 많이 사용되는 커널함수는 RBF(Radial Basis Function), linear, polynomial, sigmoid 등이 있다.

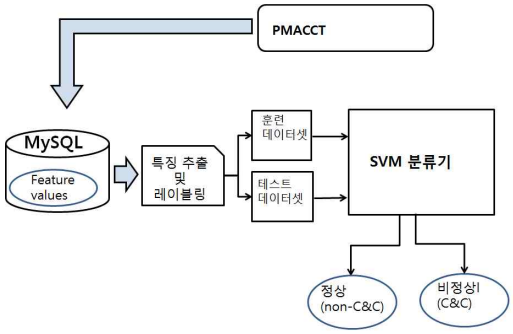

그림 6. SDS의 탐지과정
Fig. 6. Detection process of SDS

본 연구에서는 커널 파라미터와 SVM 수식의 파라미터에 복잡도 영향이 비교적 적은 RBF를 사용한다. 이 기법을 위한 응용프로그램은 LibSVM이라는 SVM을 구현한 라이브러리(Library)로 수행할 수 있다. SVM은 HMM과 달리 시계열 특징을 처리할 수 없으므로 단위 시간 특징을 사용한다. 탐지 과정은 그림 6과 같이 PMACCT에서 트래픽의 특징을 MySQL에 저장한 후 SVM에 사용할 단위 시간 데이터를 추출하고 데이터가 정상인지 비정상인지 레이블링(Labeling)한 후 탐지한다.

IV. 검증 및 분석

본 장에서는 HDS와 SDS 기법을 통해 앞장에서 제시한

기법들의 탐지율 및 적용 특성을 비교 및 제시한다.

표 5. 데이터셋 생성을 위한 악성코드
Table 5. Malwares for creating datasets

봇넷 악성코드	악성행위
droidkungfu변종 2개	루팅권한을 취득, 기기 정보 유출
carberpmobile변종 2개	러시아의 모바일 뱅킹 사용중 신임장을 C&C 서버에 유출
SMSzombie변종 8개	관리자권한 획득, 모바일 거래 정보 탈취. WIFI 불능시 SMS로 발송
androidexprespam	개인정보유출과 거짓 메시지를 보여줌
androidobada	사용자의 연락처 정보, 통화 기록, SMS 메시지 수신 내용, 설치되어 있는 앱에 관한 정보를 수집
ChuliA,	DDOS 공격, 다른 악성코드 설치
Hongtoutou,	기기정보 등을 C&C서버에 유출
Luckycat,	개인정보 유출
new zitmo, zitmo2013	SMS나 인터넷으로 모바일 뱅킹정보를 전송
SMSSend packageinstaller	C&C 서버로 정보유출, 프리미엄 서비스로 과금유도
SMSSendNopoc	사용자의 SMS 내용 외부유출
SpamSoldier	SMS spam 발송, C&C서버에서 명령 수신
SMSBotnet	Spamsoldier의 변종
Stels flashplayer	어도비 플래쉬 플레이어 업데이트로 위장하여 SMS내용을 원격서버로 전송

모바일 악성코드 출처 : http://contagiodump.blogspot.kr/

표 6. 데이터셋
Table 6. Datasets

구성요소	데이터셋1	데이터셋2	데이터셋3
플로우 수 (C&C플로우) (정상플로우)	17433 (16451) (982)	14664 (13833) (831)	14993 (13997) (996)
수집기간	7/16 20:22 - 7/17 13:57	7/17 14:40 - 7/18 10:40	7/18 14:14 - 7/19 13:17
분석된 채널 수(C&C 채널 /정상 채널)	46(13/33)	45(14/31)	32(7/25)

1. 데이터셋

검증에서는 표 5에서 제시된 23개의 봇넷을 그림 2와 같은 제안시스템 환경에 있는 안드로이드폰(HTC Desire -Android2.2, SKY IM-A690L-Android2.2.1)에 설치하여, 봇넷의 C&C 트래픽을 수집한다. 수집된 트래픽은 표 6과 같이 3개의 데이터셋으로 구성한다. 탐지된 플로우의 페이로드를 확인하여 검증할 수 있도록 Wireshark로도 패킷을 수집한다.

2. 검증

표 7.시계열 특징을 이용한 HMM의 탐지율
Table 7. Detection Rate for HMM using Time series features

구성요소	데이터셋2	데이터셋3
C&C플로우	13833	13997
탐지플로우(탐지율)	13797(99.73%)	13994(99.97%)
오탐 플로우(오탐율)	80(9.62%)	41(4.11%)

표 8.단위 시간 특징을 이용한 HMM의 탐지율
Table 8. Detection Rate for HMM using unit time features

구성요소	데이터셋2	데이터셋3
C&C플로우	13833	13997
탐지플로우(탐지율)	12772(92.32%)	13767(98.35%)
오탐 플로우(오탐율)	183(22.02%)	230(23.09%)

데이터셋 1은 훈련을 위해, 데이터셋 2와 데이터셋 3은 탐지율 검증을 위해 사용한다. JaHmm 라이브러리로 HDS를 구현한 HMM 기법의 검증을 위해서, Segmental K-means로 훈련한 시계열 특징과 Baum-Welch로 훈련한 단위 시간 특징으로 HMM 기반 기법의 탐지율을 비교 및 분석한다. 또한 SVM 기법은 단위 시간 특징을 적용했을 때의 탐지율을 분석하고 시계열 HMM 기법과 비교한다. 표 7, 표 8, 표 9에서 오탐은 "오탐 = (C&C로 판단한 정상트래픽 / 정상트래픽)"으로 계산하였다.

표 9.단위시간 특징을 이용한 SVM의 탐지율
Table 9. Detection Rate for SVM using unit time features

구성요소	데이터셋2	데이터셋3
C&C플로우	13833	13997
탐지 플로우(탐지율)	13769(99.53%)	13972(99.82%)
오탐 플로우(오탐율)	49(5.89%)	16(1.60%)

3. HMM과 SVM의 성능 비교

3.1 시계열 HMM과 단위시간 HMM의 비교

표 7과 표 8은 HMM 기반의 HDS를 이용해 분석한 두 가지 HMM 기법의 탐지율을 제시한다. 표 7은 시계열의 특징을 적용하였고 표 8은 단위 시간 특징을 적용하였다. 분석

결과는 시계열의 HMM 기법이 단위시간의 HMM 기법보다 최대 7.65%의 격차를 보이며 99.73 %의 우수한 탐지율을 보인다. 단위 시간 특징의 HMM 기법은 92.32 - 98.35%의 탐지율로 시계열 특징의 HMM 기법보다 비교적 낮은 탐지율을 보이는데, 23.09%의 높은 오탐율을 보이므로 신뢰성에 문제가 있다. 따라서, HMM에서는 시계열 특징을 적용한 기법이 탐지율과 오탐율에서 우수한 성능을 보인다.

표 7 과 표 8에서 제시된 오탐 원인의 대부분은 봇넷 트래픽과 비슷한 주기적인 요청과 비슷한 트래픽량을 가진 구글에 대한 응용프로그램의 트래픽과 daytime 프로토콜의 트래픽이다.

3.2 시계열 HMM과 단위시간 SVM의 비교

표 9는 표 8와 동일한 단위 시간 특징을 SVM 기반의 SDS를 이용해 분석한 SVM의 탐지율이다. 표 9의 SVM 기법의 탐지율과 표 8의 HMM 기법의 탐지율을 비교해 보면, 99.53%-99.82%의 탐지율과 1.60%-5.89%의 오탐율을 보인 SVM 기법이 더 우수하다. 그러나 시계열의 HMM의 표 7과 단위시간 특징을 사용한 SVM의 표 8를 비교해보면, 탐지율 면에서 시계열의 HMM 기법이 약간 더 우수하고, 오탐율은 SVM 기법이 우수하다.

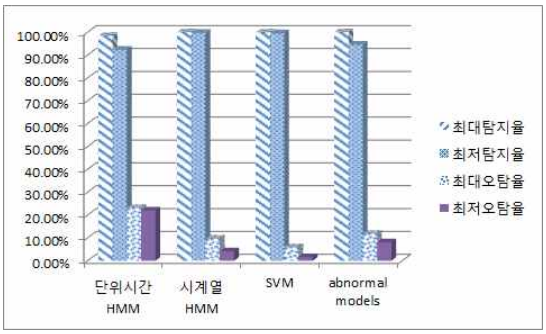

그림 7. HMM, SVM과 다른 연구의 비교
Fig. 7. Comparision of HMM, SVM and other work

따라서 시계열의 데이터를 추출할 수 있거나 변환할 수 있는 데이터의 경우에는 HMM을 적용하는 것이 탐지율과 오탐율의 측면에서 우수하고, 단위 시간 데이터 같은 패널 데이터가 존재할 경우, SVM 기법이 우수한 성능을 나타내는 특성이 나타난다.

3.3 타 기법과의 비교

또한 그림 7에서 보여주듯이 모바일 봇넷을 탐지하는 기존 연구인 Abnormal model 기법[2]은 탐지율은 비슷하나 오탐율에서 시계열의 HMM 기법과 SVM 기법보다 높아 취약한 편이다. 또한 HMM과 SVM을 필기체의 서명을 인식하는데에 적용한 기존 연구에서는 SVM이 우수한 것으로 나타났다[14]. 이는 HMM에서 시계열 특징을 사용 않았고, SVM에 우수한 패널 데이터를 사용한 것으로 분석된다.

V. 결 론

최근 모바일 봇넷은 PC에서 모바일 장치로 이동하고 있으며, 따라서 그 피해가 모바일 장치로 재생산되고 있다. 본 논문에서는 VPN을 이용하여 모바일 장치와 IDS를 연결하고 그 IDS에서 C&C 트래픽을 탐지하고 3장에서 제안한 HDS와 SDS의 두 탐지 시스템의 구현을 통해 SVM과 HMM의 특성을 분석하고 제시하였다. 실제 봇넷이 공격되는 환경에서 단위 시간 특징을 이용한 HMM 기법은 92.32 ~ 98.35 %의 탐지율과 22.02 ~ 23.09%의 오탐율을 보이고 시계열 특징을 사용한 HMM 기법에서는 99.73% - 99.97%의 탐지율과 4.11 ~ 9.62%의 오탐 결과를 보여 abnormal Models 기법보다 시계열의 HMM 기법의 탐지율이 우수한 것으로 분석되었다. SVM에서는 단위 시간 특징을 이용한 결과 99.53 ~ 99.82%의 우수한 탐지율과 1.60-5.89% 낮은 오탐율을 보였다. HMM, SVM 모두 탐지율과 오탐율면에서 우수한 모바일 봇넷의 탐지기법으로 분석되었으며, 이들은 다음과 같은 특성을 갖는다.

1. 시간의 흐름에 따른 시계열 데이터에 대해서는 시계열 특징을 사용한 HMM 기법의 탐지율이 우수하다.
2. 단위 시간에 측정되는 특성의 패널 데이터가 존재하는 경우에는 SVM 기법의 탐지율이 우수하다.

HMM과 SVM은 탐지율 측면에서 우수하지만 계산량과 계산 시간 측면에서 많은 자원을 사용하는 것으로 알려져 있다. 이를 극복하기 위하여 적은 자원과 시간을 사용하는 결정트리 같은 다른 기계 학습 기법을 HMM과 SVM과 비교분석하는 향후 연구를 제시한다.

참고문헌

[1] Byungha Choi, Kyungsan Cho, "An Improved Detecting Scheme of Malicious Codes using HTTP Outbound Traffic," Journal of the Korea society of computer and information vo.14 no.9 pp.47-54, SEP. 2009.

[2] ByungHa Choi, Sung-kyo Choi, Kyungsan Cho,

"Detection of Mobile Botnet Using VPN," Procs. of The Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2013), pages 142-148, 2013.

〔3〕 G. Delac, M. Silic and J. Krolo, "Emerging security threats for mobile platforms," Procs. of the 34th International Convention, MIPRO 2011, pp. 1468- 1473, 23-27 May. 2011.

〔4〕 AK. Tyagi, G. Aghila "A Wide Scale Survey on Botnet," Procs. of International Journal of Computer Applications, Vol. 34, No.9, pp. 10-23, Nov. 2011.

〔5〕 Byungha Choi, Kyungsan Cho, "Two-Step Hierarchical Scheme for Detecting Detoured Attacks to the Web Server," ComSIS, vol 10, no 2, 633-649, 2013.

〔6〕 Gu, Guofei, et al. "BotMiner: Clustering analysis of network traffic for protocol-andstructure-independent botnet detection," Procs. of the 17th conference on Security symposium. 2008.

〔7〕 NQ Mobile, NQ Mobile 2011 Mobile Security Report, 2012.

〔8〕 Iker Burguera, Urko Zurutuza, Simin Nadjm-Tehrani, "Crowdroid: behavior-based malware detection system for android," Procs. of the 1st ACM workshop on Security and privacy in smartphones and mobile devices. ACM, pp. 15-26, 2011.

〔9〕 L. Xie, X. Zhang, J. P. Seifert, S. Zhu, "pBMDS: a behavior-based malware detection system for cellphone devices," Procs. of the third ACM conference on Wireless network security. ACM. pp. 37-48, 2010.

〔10〕 A. Bose, X. Hu, K. G. Shin, T. Park, "Behavioral detection of malware on mobile handsets," In Procs. of the 6th international conference on Mobile systems, applications, and services. ACM. pp. 225-238, 2008.

〔11〕 Portokalidis, Georgios, et al. "Paranoid Android: versatile protection for smartphones," Procs. of the 26th Annual Computer Security Applications Conference. ACM, 2010.

〔12〕 Falletta, Vincenzo, and Fabio Ricciato. "Detecting scanners: empirical assessment on a 3G network," International Journal of Network Security vol. 9, no. 2, pp.143-155, 2009.

〔13〕 Vural, Ickin, and Hein S. Venter. "Combating Mobile Spam through Botnet Detection using Artificial Immune Systems," Journal of Universal Computer Science 18.6 pp. 750-774. 2012.

〔14〕 Edson J.R. Justino, Flávio Bortolozzi, Robert Sabourin, "A comparison of SVM and HMM classifiers in the off-line signature verification," Pattern Recognition Letters, vol 26, Issue 9, pp. 1377-1385, 2005.

〔15〕 Yi-Lin Lin, Gang Wei, "Speech emotion recognition based on HMM and SVM," Machine Learning and Cybernetics, 2005. Procs of 2005 International Conference on, vol. 8, pp. 18-21, 2005.

〔16〕 Miao, Qiang, Hong-Zhong Huang, and Xianfeng Fan. "A comparison study of support vector machines and hidden Markov models in machinery condition monitoring," Journal of Mechanical Science and Technology, pp. 607-615, 2007

〔17〕 Wireshark, http://wireshark.com/

〔18〕 B-H. Juang, Lawrence R. Rabiner. "The segmental K-means algorithm for estimating parameters of hidden Markov models," Procs. of Acoustics, Speech and Signal Processing, IEEE Transactions on 38.9, pp. 1639-1641, 1990.

저 자 소 개

최 병 하

2007: 독학학위제 컴퓨터과학(이학사)
2009: 단국대학교 정보통신대학원 정보통신학과(공학석사)
2014: 단국대학교 컴퓨터학과 (공학박사)
2014~현재: 정보통신융합기술연구원
관심분야: 네트워크 보안
Email : notanything@hanmail.net

조 경 산

1979: 서울대학교 전자공학과(학사)
1981: 한국과학원 전기전자공학과 (공학석사)
1988: 텍사스 대학교(오스틴) 전기전산공학과(Ph.D.)
1988~1990: 삼성전자 컴퓨터부문 책임 연구원, 실장
1990~현재: 단국대학교 소프트웨어학과 교수
관심분야: 네트워크 보안 및 성능분석, 이동통신보안, 컴퓨터시스템
Email : kscho@dankook.ac.kr