유해 네트워크 트래픽 탐지를 위한 트래픽 분류 기법 성능 비교

최진태, 뉘엔 신 응억, 김경백
전남대학교 전자컴퓨터공학부
e-mail : jefron1100@gmail.com, sinhngoc.nguyen@gmail.com, kyungbaekkim@gmail.com

Performance Comparison of Traffic Classification Methods for Detecting Malicious Network Traffic

Jin-Tae Choi, Sinh-Ngoc Nguyen, Kyung-baek Kim
Dept. Electronics and Computer Engineering, Chonnam National University

요 약

유해 네트워크 트래픽이란 네트워크 환경에 유입되어 서비스의 질을 떨어뜨리고 특정 서버, 호스트 등에 피해를 입히는 트래픽을 뜻한다. 최근 이러한 유해 네트워크 트래픽을 탐지하기 위한 많은 연구들이 제안되고 있다. 본 논문에서는 이러한 연구에서 사용된 다양한 트래픽 분류 기법들을 대규모 네트워크 트래픽 데이터 셋에 적용하여 실험하고, 이에 따른 탐지율, 실용성, 가용성 등의 결과를 분석하였다. 또한, 분석 결과에 따른 문제점에 대한 개선방안을 제시하였다.

1. 서 론

네트워크가 활성화된 이래로 DoS, Worm 등의 다양한 악성코드가 끊임없이 등장하고 있다. 이러한 악성코드들은 네트워크상에 다양한 유해 트래픽을 증가시켜 네트워크 서비스의 질을 저하시키며 특정 서버, 호스트 등에 피해를 입힌다. 이러한 문제점을 개선하기 위해 최근에는 다양한 트래픽 분류 기법을 이용한 악성 트래픽 탐지 연구가 계속되고 있다. 본 논문에서는 이러한 다양한 트래픽 분류 기법의 성능을 비교하고 해당 기법들의 문제점에 대한 개선방법을 제안한다.

2. 관련 연구

신동혁 등은 DoS와 Worm악성 트래픽을 K-평균 클러스터링을 이용하여 탐지하였다. 이를 위해서 사전에 악성 트래픽을 탐지하기위한 특징을 추출하였다.[1] 한명지 등은 클러스터의 개수를 스스로 선택할 수 있는 X-means 클러스터링을 이용하여 악성 트래픽을 탐지 하였다. 하지만, 이는 각 공격 트래픽을 탐지하기 전 단계로 악성 트래픽을 분석하고 각 공격에 맞는 feature를 지정해주어야 한다. 또한 각 악성 트래픽 마다 서로 다른 feature set을 가지고 있기 때문에, 각 공격 마다 각각의 클러스터링을 적용할 필요가 있다.[2] 최병하 등은 모바일 봇넷 탐지를 위해 HVM과 SVM기법의 비교를 하였다. 이 논문에서는 탐지율은 매우 높게 나타났지만, 오탐율 또한 적지 않게 나타났다. 이러한 오탐율은 평범한 트래픽의 통신에 장애를 일으켜 네트워크 서비스의 질을 떨어뜨릴 수 있다.[3]

이러한 최근 연구들에서 사용된 네트워트 트래픽 분류 기법으로는 SVM, Naïve Bayes, KNN기법 등이 있는 것으로 분석되었다. 본 논문에서는 KDD Cup 1999 Data set에 SVM, Naïve Bayes, KNN기법을 적용하여 네트워크 유해 트래픽 탐지율에 대한 성능 비교를 수행 후 분석하고 문제점과 개선방안을 제시하고자 한다.

3. 트래픽 분류 기법의 성능 분석 및 평가

3.1 KDD CUP 1999 Data Set

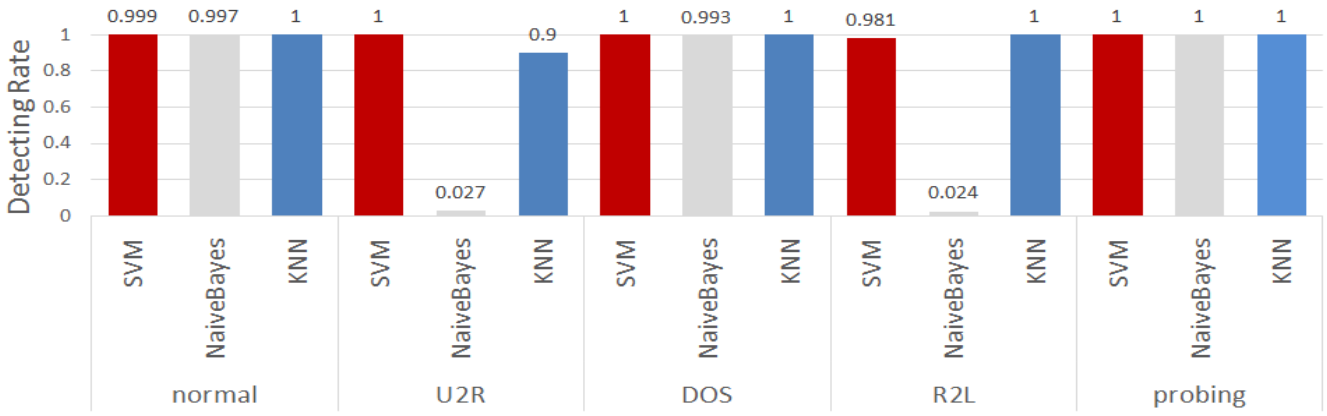
KDD Cup 1999 Data는 제 3 회 국제 지식 발견 및 데이터 마이닝 도구 공모전에 사용 된 데이터 세트이다. 이 데이터 세트는 링컨 연구소에서 미 공군 LAN을 시뮬레이팅하여 얻은 데이터이며 다수의 네트워크 유해 트래픽을 포함한다. 표 1은 해당 데이터 세트에 포함된 네트워크 유해 트래픽에 대한 공격의 범주 및 종류를 나타낸다.[4] 본 논문의 실험에서는 트레이닝 데이터로 896,064개의 레코드를 사용하였고 테스트 데이터로 494,021개의 레코드를 사용하였다. 각 레코드는 표1에 해당하는 공격의 범주 라벨이 설정되어 있고, 일반적인 트래픽 레코드는 Normal 라벨이 설정되어 있다. 각 레코드에는 총 41개의 feature가 있으며, 본 실험에서는 표2과 같은 9개의 대표적인 feature set을 사용하여 네트워크 분류 기법을 적용하였다.

3.2 비교 트래픽 분류 기법

본 논문에서는 네트워크 트래픽 분류 기법으로 자주 사용되는 SVM, KNN, Naïve Bayes 기법을 비교하였다.

(표 1) KDD Cup 1999 Dataset 레코드의 공격 범주 및 종류

Categories	Attack types
DoS	Back, land, Neptune, pod, smurf, teardrop
u2r	Buffer_overflow, loadmodule, perl, rootkit
r2l	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
Probe	Ipsweep, nmap, portsweep, satan



(그림 1) 트래픽 분류기법의 네트워크 유해 트래픽 탐지 정확도 비교 결과 (Precision)

(표 2) KDD Cup 1999 Dataset 레코드의 기본 Feature

Feature Name	Description	Type
duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol, e.g. tcp, udp, etc.	discrete
service	network service on the destination, e.g., http, telnet, etc.	discrete
src_bytes	number of data bytes from source to destination	continuous
dst_bytes	number of data bytes from destination to source	continuous
flag	normal or error status of the connection	discrete
land	1 if connection is from/to the same host/port; 0 otherwise	discrete
wrong_fragment	number of "wrong" fragments	continuous
urgent	number of urgent packets	continuous

SVM 기법은 각각의 벡터를 이용하여 최적의 decision boundary를 찾는 기법이다. KNN 기법은 새로운 data가 입력으로 들어왔을 때 이 데이터가 가진 feature와 가장 유사한 k개의 트레이닝 데이터들이 가진 클래스 중에 가장 많은 클래스를 할당해 주는 기법이다. Naïve Bayes란 모든 feature들이 서로 독립적이며 같은 분포를 갖는다는 전제로 확률을 계산하여 분류하는 Classification 기법이다. [5]

3.3 비교 실험 결과 및 분석

그림 1의 그래프는 각 네트워크 분류 기법인 SVM, Naïve Bayes, KNN을 해당 데이터세트에 적용할 때의 Precision 결과를 나타낸다. 이 결과에서는 SVM기법과 KNN기법이 매우 높은 탐지율을 보였다. 그 중에서도 SVM기법은 모든 경우에 1에 가까운 탐지율을 보였다. KNN기법의 경우, u2r공격 트래픽에 대해서는 0.9의 탐지율을 보였고 나머지 공격 트래픽에 대해서는 1에 가까운 탐지율을 보였다. Naïve Bays기법은 DOS공격 트래픽과 Proving 트래픽은 1에 가까운 탐지율을 보였으나, u2r 공격과 r2l 공격에 대해서는 매우 낮은 탐지 결과를 보였다.

이 실험결과를 통해 Precision 측면에서 SVM기법이 다른 기법들에 비해 매우 뛰어남을 확인할 수 있다. 또한 SVM의 True Positive Rate는 최소 0.978로 매우 높게 나타났으며 False Positive Rate는 최대 0.001로 매우 낮게 나타났다. 하지만, 탐지율이 뛰어난 SVM과 KNN기법은 실험결과를 얻기까지 시간이 각각 11시간, 12시간 정도 소요되었다.

4. 결론

본 논문에서는 네트워크 유해 트래픽을 탐지하기 위한 트래픽 분류 기법 간의 성능을 비교 및 분석하였다. 이를 위해 KDD Cup 1999 DataSet에 SVM, KNN, Naïve Bayes 기법을 적용하여 유해 트래픽 탐지율을 비교 하였다. 실험결과 SVM과 KNN은 매우 높은 탐지율을 보였지만 매우 느린 처리속도가 문제가 되는 것을 확인하였다. 향후, 이러한 문제점을 개선하기 위해 해당 분류 기법의 분산 처리 시스템 적용 연구를 수행할 계획이다. 또한, 네트워크 트래픽 레코드를 이미지화 하여 CNN을 적용하는 등의 딥러닝 기법을 활용한 트래픽 분류 기법연구를 수행할 계획이다.

Acknowledgement

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017R1A2B4012559)

참 고 문 헌

[1] 신동혁, 안광규, 최성춘, 최형기 (2016). K-평균 클러스터링을 이용한 네트워크 유해트래픽 탐지. 한국통신학회논문지, 41(2), 277-284.

[2] 최병하, 조경산 (2014). 모바일 봇넷 탐지를 위한 HMM과 SVM 기법의 비교. 한국컴퓨터정보학회논문지, 19(4), 81-90.

[3] 한명지, 임지혁, 최준용, 김현준, 서정주, 유철, 김성렬, 박근수 (2014). X-means 클러스터링을 이용한 악성 트래픽 탐지 방법. 정보과학회논문지, 41(9), 617-624.

[4] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[5] http://sanghyukchun.github.io/