Index Dispersion for Count 를 이용한 인터넷에서의 침입탐지예측

이경희*
*평택대학교 정보통신학과
e-mail : khlee@ptu.ac.kr

A Study on the Internet Intrusion Detection and Prediction by IDC

Kyunghee Lee*
*Dept. of Info. & Comm., PyeongTaek University

요 약

본 논문에서는 실시간 학습기능을 갖는 다층 신경회로망과 주어지는 인터넷 트래픽 데이터에서의 IDC(Index dispersion for count) 정보를 이용하여 인터넷에서의 공격(침입)을 탐지 예측 할 수 있는 모델을 제안하고 컴퓨터 모의실험을 통한 결과를 보인다

1. 서론

호스트 시스템에 대한 침입(공격)탐지방안에 대한 연구는 주로 일련의 시스템 호출명령(sequences of systems calls)을 대상으로 진행되었으며 일부 연구자들은 시스템 호출명령의 분포정보(frequency distribution of the system calls)를 사용하는 방안 등을 연구하였다[1][2][3][4]. 본 논문에서는 시스템 호출명령정보가 아닌 실제 네트워크에서 발생하는 인터넷 트래픽 데이터의 IDC(Index Dispersion for Count)정보를 이용하여 침입이 발생하기 이전에 미리 침입탐지가 가능한 다층 신경회로망 기반의 침입탐지 예측모델을 제안한다. 또한 컴퓨터 모의실험을 통하여 IDC(Index Dispersion for Count)정보의 효용성을 보인다.

2. 트래픽 데이터의 전처리과정

2.1 KDD Data Set

구현한 침입탐지 모델의 성능과 효용성을 시험하기 위하여 KDD(Knowledge Discovery and Data Mining)-99 의 인터넷 트래픽 데이터 세트를 사용하였다[5]. 기본적으로 KDD-99 의 TCP 패킷들은 근원지 주소(Source IP Address)에서 목적지 주소(Destination IP Address)로 시작 시각에서 적절한 시간 동안 이용 가능한 포트(Port)를 이용하여 사용한 후에 연결을 종료한 일련의 TCP 패킷들을 나타낸다. 이 데이터 세트에서는 정상 패킷과 공격(침입) 패킷으로 표현하고 있으며 공격 패킷의 유형은 크게 다음의 4 가지 유형으로 분류 된다.

- DoS : 서비스 거부 공격
- R2L : 원격지로부터 허가 받지 않은 접근 공격
- U2R : 지역에서의 관리자(root)권한 획득 공격
- Probing : 감시 및 다른 조사에 의한 공격시도

2.2 정보처리를 위한 전처리과정

제안한 신경회로망 예측 모델에서의 입력을 위하여 트래픽 데이터의 속성들에 대하여 아래와 같은 2 단계의 전처리 과정을 수행한다.

- 전처리-1 단계 : 각 속성의 값은 서로 다른 크기와 표현 형태를 갖게 되므로 효율적인 학습을 위하여 공통적인 표현과 표현 수치 스케일의 조정
- 전처리-2 단계 : 연속된 인터넷 패킷들의 선택된 각 속성에 대하여 적정구간(Lag)에서의 IDC 계산

3. IDC 기반의 신경회로망

3.1 IDC

일반적으로 n 개의 랜덤변수(random variable)의 분산은 다음과 같이 표현된다.

$$\mathrm{var}(X_{i+1}+\ldots+X_{i+n}) = n\,\mathrm{var}(X) + 2\sum_{j=1}^{n-1}\sum_{k=1}^{j}\mathrm{cov}(X_j, X_{j+k}) \quad \text{-- (1)}$$

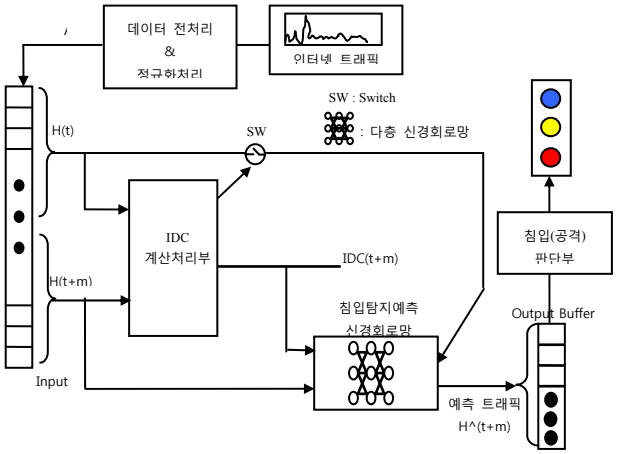
네트워크 트래픽 측면에서 식(1)은 트래픽의 특성이 자기공분산(autocovariance)에 의존적이며, 패킷도착 프로세스 표현에 각각의 도착 패킷에 대한 분산 정보가 유용함을 알 수 있다. 특히, 네트워크 트래픽의 성격을 파악하기 위하여 패킷 도착 프로세스(arrival process)의 변화성(variability)를 측정할 때에는 정규화한 Index of Dispersion 의 개념을 이용하여 트래픽의 성격을 분석할 수 있다. 본 논문에서는 침입 트래픽 데이터를 분석하기 위하여 Index of Dispersion 의 한 종류인 IDC (Index of Dispersion for Count)의 개념을 적용한다. IDC 는 주어진 간격 내에 도착한 트래픽 데이터 정보의 분산을 트래픽 데이터의 평균으로 나눈 값으로서 다음과 같이 정의 된다.

$$I_t = \frac{\mathrm{var}(N_t)}{E(N_t)} \quad \text{---------- (2)}$$

여기서, N_t 는 길이가 t 인 도착 간격에서의 패킷 정보의 갯수.

3.2 침입탐지 예측 모델

제안된 모델에서는 인터넷 트래픽을 입력받아 일정한 샘플링 간격에 따라 이동(shift)되는 입력버퍼(input buffer)에 저장한다. H^(t)트래픽과 H^(t+m) 트래픽은 IDC 계산처리부에 입력된다. 이때 H^(t) 트래픽은 침입탐지여부 신경회로망의 학습과정시 기대값으로도 입력된다. IDC 정보와 트래픽에 의하여 학습이 완료된 후에는 입력되는 인터넷 트래픽 데이터에 대한 침입여부 판단과정이 진행되어 IDC 정보와 H^(t+m) 트래픽이 침입탐지예측 신경회로망에 입력되고 학습된 침입탐지 예측 신경회로망의 출력결과에 따라 침입(공격)판단부에서 침입여부를 판단 할 수 있게 하는 모델이다.



(그림 1) 침입탐지예측을 위한 모델

4. 실험결과

4.1 네트워크 구조형성 및 학습

매트랩의 EBP(Error Back Propagation) 관련기능을 사용하여 제안모델(EBP-IDC)을 구현하여 실험하였다. 실험에 사용된 신경회로망은 3 계층으로 구성하였으며 입력층은 120 개, 중간층은 60 개, 출력층은 1 개의 노드를 가지도록 하였다. 각 계층의 전달함수(transfer function)은 Log-Sigmoid 함수를 사용하였고 EBP 의 학습 알고리즘으로는 공액경사도 역전파방법(conjugate gradient back-propagation)을 구현한 traincgi() 함수를 사용하였다. IDC 계산을 위한 구간(Lag)은 10 으로 설정하였다. 탐지예측기능의 비교를 위하여 IDC 를 사용하지 않는 EBP 신경회로망 모델(EBP-ORG) 역시 3 계층으로 입력층은 12 개 중간층은 6 개 출력층은 1 개의 노드를 갖는 구조이며 다른 파라메터는 제안모델의 경우와 동일하게 설정하였다.

4.2 침입 탐지예측 실험결과

KDD-99 데이터를 대상으로 침입탐지 예측을 실험하였다. KDD-99 데이터 중 침입 패킷이 일부 포함되어 있는 4000 개의 트래픽 데이타를 대상으로 하여 제안한 모델(EBP-IDC)에서 실험한 결과와, IDC 정보를 사용하지 않는 모델(EBP-ORG)에서 실험한 결과를 표 1 에 보인다. 실험은 IDC 계산에 사용된 구간(=10)을 넘어서지 않는 시간에서의 예측(t5 예측모델, t7 예측모델)와 예측과 넘어선 시간에서의 예측(t10,t15 예측모델)으로 구분하였다.

<표 1> 신경회로망 학습 결과 및 예측실험 결과

예측모델 비교	KDD-99 데이터 학습결과(MSE)			
	t5 예측모델	**t7 예측모델**	t10 예측모델	t15 예측모델
EBP-IDC	**0.0187044**	**0.0205032**	0.0220928	0.0242426
EBP-ORG	**0.0271990**	**0.0275904**	0.0276614	0.0300066

예측모델 비교	KDD-99 침입탐지 실험결과(예측비율)			
	t5 예측모델	**t7 예측모델**	t10 예측모델	t15 예측모델
EBP-IDC	**94.28%**	**93.65%**	91.93%	92.23%
EBP-ORG	**91.90%**	**91.83%**	91.70%	91.43%

표 1 에서 알 수 있듯이 EBP-IDC 에서 IDC 계산에 사용된 구간 내의 시간예측 모델들(t5,t7)은 침입탐지 예측비율이 비교적 우수하지만 구간 외의 시간예측 모델들(t10,t15)은 기본적인 EBP 의 경우와 크게 다르지 않음을 알 수 있으며 IDC 정보가 침입탐지예측 기능에 기여하고 있음을 볼 수 있다.

5. 종합토의

제안한 모델은 단순히 연속되는 인터넷 패킷들의 영역별 값만을 이용하여 침입을 예측하는 모델이 아니라 분포정보를 고려하여 예측할 수 있는 모델로서 모의실험을 통하여 인터넷 트래픽에서의 침입탐지 예측은 비교적 잘 동작하는 것을 확인할 수 있었다. 그러나 실제의 침입탐지 시스템에 적용하기 위해서는 IDC 계산을 위한 구간크기의 자동 조절 및 하드웨어화를 고려한 적절한 신경회로망 모델에 대한 추가 연구 등이 필요하다.

참고문헌

[1] Liu Z, Florea G, Bridges SM, “A comparison of input representations in neural networks: a case study in intrusion detection:, Proceedings of the 2002 International Joint Conference on Neural Networks, 2002.

[2] Wun-Hwa C, Sheng-Hsun H, Hwang-Pin S, “Application of SVM and ANN for intrusion detection”, Computer & Operations Research, 32, 2617-2634, 2005.

[3] Warrender C, Forrest S, Pearlmutter B. “Detecting intrusions using system calls: alternative date models”, Proceedings of 1999 IEEE Symposium on Security and Privacy, 1999.

[4] Liao Y, Vemuri VR, “Use of K-nearest neighbor classifier for intrusion detection”, Computer Security, 21, 439-448. 2002.

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

[6] Tarek S. “Wired and wireless intrusion detection system: Classifications, good characteristics and state-of-the-art”, Computer Standards & Interfaces, 28, 670-694, 2006.