중요도를 고려한 의사 결정 트리 기반 DDoS 공격 분석

염성관 · 박상윤 · 신광성*

원광대학교

DDoS attack analysis based on decision tree considering importance

Sungkwan Youm · Park Sangyoon · Kwang-Seong Shin*

Wonkwang University

E-mail : skyoum@gmail.com / tkddbs7070@naver.com / waver0920@wku.ac.kr

요 약

침입 탐지 시스템에 의해서 DDoS와 같은 공격을 탐지되며 조기에 차단할 수 있다. 의사 결정 트리를 이용하여 DDoS 공격 트래픽을 분석하였다. 중요도가 높은 결정적인 속성(Feature)을 찾아서 해당 속성에 대해서만 의사 결정 트리를 진행하여 정확도를 확인하였다. 그리고 위양성 및 위음성 트래픽의 내용을 분석하였다. 그 결과 하나의 속성은 98%, 두 가지 속성은 99.8%의 정확도를 각각 나타냈다.

ABSTRACT

Attacks such as DDoS are detected by the intrusion detection system and can be prevented early. DDoS attack traffic was analyzed using the decision tree. Deterministic features with high importance were found, and the accuracy was verified by proceeding the decision tree for only those properties. And the contents of false positive and false negative traffic were analyzed. As a result, the accuracy of one attribute was 98% and the two attributes were 99.8%, respectively.

키워드

의사결정트리, 기계학습, 침입 공격, DDoS

I. 서 론

패턴 기반 침입탐지는 비정상 트래픽을 감지하는 알고리즘을 적용하여 외부의 위협으로부터 사전에 차단하고자 한다. 많은 논문에서 효율적인 알고리즘을 검증하기 위해 실제 망에서 수집된 데이터에 적용해서 검증하고 있다. 가장 많이 사용하고 있는 데이터인 KDDCUP 99는 20년이 지났음에서 많은 연구에서 활용되고 있다. 네트워크 환경을 구성하여 시뮬레이션으로 얻어진 네트워크 트래픽을 tcpdump로 만들어진 공격 데이터이다. DARPA에서 네트워크 이상 탐지 또는 공격 판별 테스트를 위해 만든 데이터 셋으로써, 지속시간, 프로토콜 종류 등 41개의 속성과 공격여부를 라벨링한 정보까지 총 42개의 속성을 가지고 있다[1]. 그리고 CIC-IDS-2017는 캐나다 사이버보안연구소에서 전수공격(Brute force), Heartbleed, Botnet, DDoS 등 6가지 공격 시나리오를 생성하여 수집한 데이터 세트로써 80개의 속성을 가지고 있다[2].

공격 탐지 성능 향상을 위해 머신러닝을 이용한 탐지 기법이 많이 도입되고 잇다. 머신 러닝의 학습을 위해서 트래픽의 특징과 공격패턴을 사용하며 학습한 네트워크을 바탕으로 실제 네트워크에 적용하여 시험한다. 이때 주로 사용되는 데이터 세트가 KDDCUP99, CICIDS2017이다[3]. 본 논문은 가장 최근의 데이터 세트인 CICIDS2017를 이용하여 결정 트리 침입 탐지 알고리즘을 검증하고 데이터의 유효성에 관해서 확인하고자 한다.

* corresponding author

II. 의사 결정 트리 알고리즘

의사 결정 트리 학습법은 기계 학습으로써 입력 및 목표 변수를 연결해주는 의사 결정 트리를 사용한다. 일반적으로 의사 결정 트리 방법으로 분류(Classification) 및 회귀(Regression)가 가능하다. 분류는 목표 값이 이산적일 경우 즉, 목표 값이 가질 수 있는 값이 유한한 경우에 적용하며 회귀는 목표 값이 실수인 경우에 적용한다.

지니 중요도(Gini Importance)를 측정하여 속성 중 가장 중요한 값을 찾아서 해당 속성에 대해서만 결정트리를 학습시켜 학습 시간을 최소화 할 수 있다. 지니 중요도로 중요한 속성을 파악하여 해당 속성을 분류의 기준으로 삼는다. 속성 중요도는 모든 속성 평균 가 사용되는 모든 트리의 노드에 대해서 차감을 더한 후 평균을 구한다[4-5].

$$Imp(X_j) = \frac{1}{M}\sum_{m=1}^{M}\sum_{t\in\varphi_m} 1(j_t = j)\left[p(t)\triangle i(s_t,t)\right], \quad (1)$$

지니 지수를 불순도 함수로 사용하는 경우에 측정값을 지니 중요도라고 한다.

III. 결정 트리를 이용한 DDoS 분석

다음은 'Total Length of Fwd Packets'와 'Fwd Packet Length Max'을 모두 이용하여 의사 결정트리를 수행하면 트리는 그림 1와 같으며 성능은 표 4와 같이 얻을 수 있다. 그림에서 x1은 'Total Length of Fwd Packets'를 x2는 'Fwd Packet Length Max'를 나타낸다. 트리는 단순화 되어 있지만 표 1에서 보는 바와 같이 정확도가 99.8에 달한다. 속성이 많아지면 학습하는데 소비되는 자원이 많아지기 때문에 최소의 속성으로 최대의 성능을 달성해야 한다. 이처럼 2개의 속성만으로 전체 속성을 학습시킨 결과와 유사한 정확도를 얻을 수 있었다.

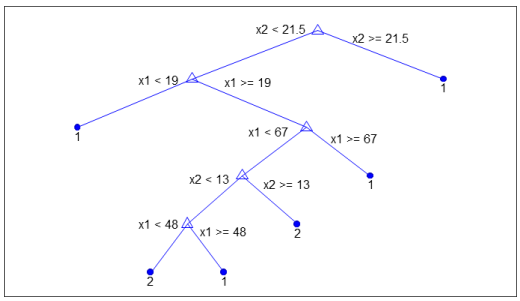

그림 1 'Total Length of Fwd Packets and Fwd Packet Length Max'기반 의사 결정 트리

표. 1 정확도

Accuracy(%)	False positive	False negative
99.8405	5	31

IV. 결 론

본 연구에서는 의사 결정 트리를 이용하여 CICIDS2017 데이터의 DDoS 공격를 분석하였다. 분석 방법으로 먼저 전체 트래픽 속성을 적용해서 분석한 후 예측 중요도로 다시 중요한 트래픽 속성을 파악한다. 해당 트래픽 속성으로 다시 의사 결정 트리를 학습하여 정확도, 위양성 및 위음성를 확인하였다. 2가지 트래픽 속성으로 분석하여도 99.8%의 정확도를 얻을 수 있었다. 하지만 위음성의 개수는 늘어나 탐지를 피하는 공격 트래픽이 늘어 날 가능성이 있다.

Acknowledgement

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 한국연구재단에서 부여한 과제번호 : NRF-2018R1D1A1B07050277)

References

[1] E. M. Yang and C. H. Seo, " A Study on Intrusion Detection in Network Intrusion Detection System using SVM," Journal of Digital Convergence, vol. 16, no. 5, pp. 399-406, May. 2018.

[2] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, " Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," In Proceeding of the 4th International Conference on Information Systems Security and Privacy, Funchal: FNC, pp. 108-116, Jan. 2018.

[3] S. H. Choi, M. H. Jang, and M. S. Kim, "A Study on AI algorithms to Improve Precision Rate in a Managed Security Service," The transactions of The Korean Institute of Electrical Engineers, vol. 69, no. 7, pp. 1046-1052, Jul. 2020.

[4] B. H. Menze, B. M. Kelm, R. Masuch, R, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its Gini importance with standard chemometric

methods for the feature selection and classification of spectral data," BMC Bioinformat, vol. 10, no. 213, pp. 1-16, Jul. 2009.

[5] G. Louppe, "Understanding random forests," Ph. D. dissertation, University of Liège, liège, Be, Jul. 2014.