특징 선택 방법에 따른 네트워크 공격 유형별 중요 특징 비교 연구

최다영, 이주홍, 박형곤

이화여자대학교 전자전기공학과 스마트팩토리융합전공

{dayoung.choi, joohong.rheey}@ewhain.net, hyunggon.park@ewha.ac.kr

Important Feature Comparison for Network Attack Types based on Feature Selection Methods

Dayoung Choi, Joohong Rheey and Hyunggon Park

Graduate Program in Smart Factory, Department of Electronic and Electrical Engineering,

Ewha Womans University

요 약

본 논문은 사물인터넷 기반의 실제 네트워크 트래픽 데이터셋인 NF-BoT-IoT-v2의 4가지 공격 유형에 대해 여섯 가지의 필터 방식 기반 특징 선택 방법을 이용하여 공격 유형별 특징 중요도를 측정하고, 중요도가 높은 상위 특징을 확인하였다. 중요도에 따른 특징 조합을 활용하여 공격 유형별 트래픽의 고유한 특성을 반영한다면, 다양한 네트워크 공격에 강건하고 효율적인 딥러닝 기반 네트워크 관리 시스템을 구축할 수 있을 것으로 기대된다.

Ⅰ. 서 론

사물인터넷(Inernet of Things, IoT)의 지속적인 발전으로 기기 간 연결의 복잡성이 증가함에 따라 방대한 사물인터넷 네트워크 트래픽이 생성되고 있는 추세이다. 이기종의 대규모 사물인터넷 기기들이 데이터를 교환하고 생성하는 과정에서, 네트워크를 통해 전달되는 데이터에 대해 악의적인 조작, 중단 또는 무단 접근 등을 기반으로 한 다양한 유형의 사이버 공격이 발생하고 있다[1]. 네트워크 시스템을 안전하고 효율적으로 운영하기 위해서는 점차 다양화되고 정교화되고 있는 사이버 공격을 정상 트래픽으로부터 구분하는 것이 중요하다. 따라서 네트워크 관리 시스템에 딥러닝을 적용하여 많은 수의 특징(feature)을 포함하는 고차원의 네트워크 공격 트래픽을 빠르고 정확하게 탐지하기 위한 연구가 계속되고 있다[1],[2]. 딥러닝 기반 네트워크 관리 시스템을 활용하는 과정에서 공격 유형별 특징 선택(feature selection) 방법을 적용하면, 중복성이 높은 특징을 제거함으로써 데이터의 차원을 축소하여 네트워크 관리 시스템의 복잡도를 낮추고 연산 처리 시간을 단축할 수 있다. 본 논문에서는 다양한 특징 선택 방법을 이용하여 서로 다른 공격 유형 별 특징 중요도를 측정하고, 중요도가 높은 상위 5개 특징에 대해 분석한다.

Ⅱ. 본론

A. 사물인터넷 데이터셋

본 논문에서 고려하는 데이터셋은 총 43개의 특징을 갖는 NF-BoT-IoT-v2 데이터셋[3]으로, 정상 트래픽과 총 4가지 유형의 공격(DoS, DDoS, Theft, Reconnaissance) 트래픽으로 구성되어 있다. NF-BoT-IoT-v2는 사물인터넷 기반의 실제 네트워크 트래픽인 BoT-IoT[4]로부터 공개적으로 제공된 pcap 파일을 이용하여 NetFlow 기반의 43개의 특징 집합으로 추출된 트래픽으로 구성된 데이터셋이다. 43개의 특징 집합은 흐름(flow) 및 패킷 수준의 측정값 기반 특징과 TCP(Transmission Control Protocol)에 관련된 특징으로 구성되어 있으며, 각 트래픽 데이터가 정상 트래픽인지 공격 트래픽인지를 나타내는 레이블(label)이 포함되어 있다.

B. 특징 선택 방법

본 논문에서는 NF-BoT-IoT-v2 데이터셋에 대해 공격 유형별로 특징 중요도를 측정하기 위해 필터 방식(filter method)의 여섯 가지 특징 선택 방법을 이용한다. 두 확률 변수 사이의 선형 상관 관계를 측정하는 PCC(Pearson Correlation Coefficient)[5], 평균값 차이를 이용하여 분산 분석을 수행하는 ANOVA(Analysis of Variance)[6], 평균 차이를 검정하는 T-test[7], 두 확률 변수 간의 독립성을 검정하는 Chi-squared[8], 두 확률 변수가 공유하고 있는 정보량을 측정하는 Mutual information[9] 그리고 확률 변수의 불순도(impurity)를 측정하는 Gini-index[10]를 이용하여 특징들과 레이블 간 관련성을 평가한다. 평가 결과, 레이블과 관련성이 높은 특징일수록 중요도가 높은 특징으로 간주한다. PCC, ANOVA, T-test, Chi-squared, Mutual information은 값이 클수록 중요도가 높은 특징인 반면, Gini-index는 값이 작을수록 중요도가 높은 특징이다. 레이블 외 42개의 특징 중 IP 주소와 포트 번호와 같은 소켓 정보를 제외하고 총 39개의 특징을 대상으로 특징 선택 방법을 적용하며, 중요도가 높은 특징일수록 공격 트래픽 판별에 영향이 큰 특징임을 의미한다.

C. 특징 중요도 순위

표 1은 특징 선택 방법에 따른 NF-BoT-IoT-v2 데이터셋의 4가지 공격 유형별 중요도 상위 5개 특징을 나타낸 표이다. PCC, ANOVA, T-test의 경우 Theft 공격을 제외한 모든 공격 유형에서 동일한 결과를 보이며, Theft 공격에서는 상위 3개의 결과가 동일하다. 또한, 모든 공격 유형에 대해 PCC, ANOVA, T-test와 Chi-squared, 그리고 Mutual information과 Gini-index에서 특징 중요도 순위의 경향성이 비슷하다. DoS 공격의 경우, PCC, ANOVA, T-test, Chi-squared 방법에서의 1순위 특징은 TCP와 관련된 것이며, 상위 5개 특징 조합은 모두 TCP 및 네트워크 흐름의 시간과 관련된 특징으로 구성되어 있다. 반면, DDoS 공격의 경우 모든 특징 선택 방법에서 *LT_PROTOCOL* 특징이 1순위를 차지하였으며, 상위 5개의 특징 조합에 대해서는 PCC, ANOVA, T-test, Chi-squared에서는 TCP를 비롯한 프로토콜과 관련된 특징들이 주로 포함되어 있으나,

	PCC	ANOVA	T-test	Chi-squared	Mutual Information	Gini-index
DoS						
1	TCP_FLAGS	TCP_FLAGS	TCP_FLAGS	TCP_WIN_MAX_IN	IN_BYTES	IN_BYTES
2	DURATION_IN	DURATION_IN	DURATION_IN	TCP_FLAGS	SRC_TO_DST_AVG_THROUGHPUT	NUM_PKTS_128_TO_256_BYTES
3	CLIENT_TCP_FLAGS	CLIENT_TCP_FLAGS	CLIENT_TCP_FLAGS	CLIENT_TCP_FLAGS	SHORTEST_FLOW_PKT	SHORTEST_FLOW_PKT
4	MIN_TTL	MIN_TTL	MIN_TTL	DURATION_IN	TCP_FLAGS	NUM_PKTS_UP_TO_128_BYTES
5	MAX_TTL	MAX_TTL	MAX_TTL	FLOW_DURATION_MILLISECONDS	NUM_PKTS_128_TO_256_BYTES	SRC_TO_DST_AVG_THROUGHPUT
DDoS						
1	L7_PROTO	L7_PROTO	L7_PROTO	L7_PROTO	L7_PROTO	L7_PROTO
2	PROTOCOL	PROTOCOL	PROTOCOL	TCP_FLAGS	IN_BYTES	LONGEST_FLOW_PKT
3	TCP_FLAGS	TCP_FLAGS	TCP_FLAGS	TCP_WIN_MAX_IN	SHORTEST_FLOW_PKT	SHORTEST_FLOW_PKT
4	CLIENT_TCP_FLAGS	CLIENT_TCP_FLAGS	CLIENT_TCP_FLAGS	CLIENT_TCP_FLAGS	SRC_TO_DST_AVG_THROUGHPUT	MAX_IP_PKT_LEN
5	MAX_TTL	MAX_TTL	MAX_TTL	FLOW_DURATION_MILLISECONDS	LONGEST_FLOW_PKT	SRC_TO_DST_AVG_THROUGHPUT
Theft						
1	DURATION_OUT	DURATION_OUT	DURATION_OUT	FLOW_DURATION_MILLISECONDS	IN_BYTES	IN_BYTES
2	PROTOCOL	PROTOCOL	PROTOCOL	DURATION_OUT	LONGEST_FLOW_PKT	SRC_TO_DST_AVG_THROUGHPUT
3	SERVER_TCP_FLAGS	SERVER_TCP_FLAGS	SERVER_TCP_FLAGS	SERVER_TCP_FLAGS	MAX_IP_PKT_LEN	LONGEST_FLOW_PKT
4	MAX_TTL	DST_TO_SRC_SECOND_BYTES	MAX_TTL	CLIENT_TCP_FLAGS	SRC_TO_DST_AVG_THROUGHPUT	MAX_IP_PKT_LEN
5	CLIENT_TCP_FLAGS	MAX_TTL	CLIENT_TCP_FLAGS	TCP_WIN_MAX_IN	L7_PROTO	L7_PROTO
Reconnaissance						
1	SERVER_TCP_FLAGS	SERVER_TCP_FLAGS	SERVER_TCP_FLAGS	DNS_TTL_ANSWER	CLIENT_TCP_FLAGS	SERVER_TCP_FLAGS
2	TCP_WIN_MAX_IN	TCP_WIN_MAX_IN	TCP_WIN_MAX_IN	TCP_WIN_MAX_IN	SERVER_TCP_FLAGS	CLIENT_TCP_FLAGS
3	DURATION_OUT	DURATION_OUT	DURATION_OUT	SERVER_TCP_FLAGS	TCP_WIN_MAX_IN	TCP_WIN_MAX_IN
4	CLIENT_TCP_FLAGS	CLIENT_TCP_FLAGS	CLIENT_TCP_FLAGS	TCP_WIN_MAX_OUT	IN_BYTES	OUT_BYTES
5	TCP_WIN_MAX_OUT	TCP_WIN_MAX_OUT	TCP_WIN_MAX_OUT	CLIENT_TCP_FLAGS	OUT_BYTES	DURATION_OUT

표 1. 특징 선택 방법에 따른 공격 유형별 중요도 상위 5개 특징

Mutual information, Gini-index에서는 패킷의 길이나 처리량과 관련된 통계 값을 나태내는 특징들이 높은 중요도를 보였다. Theft 공격의 경우에는 상위 5개 특징 조합에 대해서, ANOVA의 4순위 중요도를 갖는 *DST_TO_SRC_SECOND_ BYTES* 특징을 제외한다면 PCC, T-test와 ANOVA가 유사한 결과를 보였으며, PCC, ANOVA, T-test, Chi-squared에서는 네트워크 흐름의 지속 시간과 관련된 특징이 1순위를 차지하였다. 그리고 Mutual information, Gini-index에서는 *IN_BYTES* 특징이 1순위이며, 이 두 가지 방법의 중요도 상위 5개 특징 조합이 동일한 5개의 특징들로 구성되어 있다. Reconnaissance 공격의 경우에는 모든 특징 선택 방법의 중요도 상위 5개 특징 조합 내에 TCP와 관련된 특징인 *CLIENT_TCP_FLAGS*, *SERVER_TCP_FLAGS*, *TCP_WIN_MAX_IN*이 포함되어 있다. 특히, *SERVER_ TCP_FLAGS*는 PCC, ANOVA, T-test, Gini-index에서, *CLIENT_TCP_FLAGS*는 Mutual information에서 중요도 1순위 특징이다. 본 실험 결과를 통하여 공격 유형별, 특징 선택 방법별로 특징 중요도에 차이가 있음을 알 수 있다. 따라서 딥러닝 기반 네트워크 관리 시스템에 특징 선택 방법을 활용할 경우, 공격 유형별 특성을 고려하면 네트워크 공격 탐지 정확도를 높일 수 있을 것이라 기대된다.

Ⅲ. 결론

본 논문에서는 사물인터넷 기반의 실제 네트워크 트래픽 데이터셋인 NF-BoT-IoT-v2의 4가지의 공격 유형에 대해 여섯 가지의 특징 선택 방법을 적용하여 특징 중요도를 분석하였다. 분석 결과, 각 공격 유형마다 중요도가 높은 특징에 차이가 있음을 알 수 있다. 이를 활용하여 서로 다른 공격 유형에 따라 중요도가 높은 특징 조합을 선별하거나 공격 유형별 트래픽의 고유한 특성을 반영하는 특징 조합을 통해 강건한 네트워크 관리 시스템을 설계할 수 있다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2021-0-00739, 분산/협력AI 기반 5G+ 네트워크 데이터 분석 기능 및 제어 기술 개발)을 받아 수행된 연구임.

참 고 문 헌

[1] M. Saied, et al., “Review of artificial intelligence for enhancing intrusion detection in the internet of things,” *Engineering Applications of Artificial Intelligence,* vol. 127, pp. 107231-107252, 2024.

[2] J. Rheey, et al. “Low-complexity Anomaly Detection Method based on Feature Importance using Shapley Value,” *International Conference on Ubiquitous and Future Networks (ICUFN 2023)*, 2023.

[3] M. Sarhan, et al., “Towards a Standard Feature Set for Network Intrusion Detection System Datasets,” *Mobile Networks and Applications*, vol. 27, pp. 357 - 370, 2022.

[4] N. Koroniotis, et al., “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset,” *Future Generation Computer Systems*, vol. 100, pp. 779-796, 2019.

[5] K. Pearson, “Notes on regression and inheritance in the case of two parents,” *Proceedings of the royal society of london*, vol. 58, pp. 240-242, 1895.

[6] R. A. Fisher, “XV.—The correlation between relatives on the supposition of Mendelian inheritance,” *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399-433, 1919.

[7] B. L. Welch, “The generalization of "Student's" problem when several different population variances are involved,” *Biometrika,* vol. 34, no. 1-2, pp. 28-35., 1947.

[8] R. M. Fano, et al., “Transmission of information: A statistical theory of communications,” *American Journal of Physics*, vol. 29, no. 11, pp. 793-794, 1961.

[9] C. Gini, *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*, Tipogr. di P. Cuppini, 1992.