메타 특성을 이용한 네트워크 데이터 이상탐지기법 성능 개선

남연하, 정다은, 박형곤
이화여자대학교
{ttr9904, daeun.jung}@ewhain.net, hyunggon.park@ewha.ac.kr

On Improving Network Data Anomaly Detection Performance based on Meta Characteristics

Yeonha Nam, Daeun Jung and Hyunggon Park
Ewha Womans University

요 약

비균등하며 예측하기 어려운 네트워크 데이터의 특성으로 인해 네트워크 데이터 이상 탐지가 점점 더 어려워짐에 따라 네트워크 데이터 이상 탐지에 대한 연구의 중요성이 증가하고 있다. 이에 본 논문에서는 대용량 네트워크 데이터에서 메타 특성을 추출하고 이를 이용하여 이상 탐지 기법의 성능을 향상시키는 방법을 제안하였다. 실제 네트워크 데이터를 적용한 실험을 통하여 제안한 메타 특성을 이용한 방식이 기존의 방식에 비해 더 정확한 예측을 짧은 시간에 내 가능함을 확인하였다.

Ⅰ. 서 론

네트워크 내 다양한 공격으로 인한 침입은 인터넷 컴퓨팅 환경의 변화와 새로운 서비스 출현으로 가속화 되었다. 지능화 되어가는 공격으로 인한 규칙-기반 침입탐지시스템의 취약점을 극복하기 위해 머신러닝 기술을 활용한 네트워크 이상 검출에 대한 관심이 집중되고 있다. 5G 및 사물 인터넷의 등장으로 네트워크 동작 과정이 점점 더 복잡해짐에 따라 네트워크 데이터는 무수한 네트워크 소스로부터 넓은 범위에서 생산되고 있다. 빠른 네트워크 입력과 출력을 통해 생성된 네트워크 데이터는 계산량이 많고, 불균형한 특성을 가지며, 네트워크 데이터 분포의 지속적인 변화로 인해 데이터 분석 및 비정상 동작 탐지가 더욱 어려워지고 있다 [1]. 매일 새로운 유형의 공격이 발견되고 있지만, 현재 수신되는 데이터의 방대한 양, 속도, 다양성 및 정확성으로 인해 이러한 공격을 조기에 탐지하게 어렵기 때문이다. 이처럼, 빅 데이터 시대는 의료, 제조, 통신, 은행, 교육 및 운송과 서비스를 제공하기 위해 네트워크 아키텍처가 복잡해짐에 따라 데이터 처리 및 분석 이상 탐지 감지의 필요성이 증가되고 있다. 따라서 효과적인 네트워크 데이터 이상 탐지를 위해서 방대한 네트워크 데이터를 효율적으로 처리할 수 있는 방법이 요구된다 [2].

본 논문에서는 이상 탐지를 하고자 하는 네트워크 데이터를 메타 특성을 이용하여 낮은 계산량으로 이상 탐지 성능이 향상됨을 보였다. 메타데이터란 구조화 된 형식에 포함된 데이터에 대한 데이터, 정보에 대한 정보로, 데이터베이스에 수록된 데이터 집합을 기술하는 정보를 담고 있는 데이터를 말한다 [3]. 비가공 데이터에 비해 메타 특성은 요약된 정보로 데이터의 크기가 작다. 따라서 적절한 메타 데이터를 정의하여 비가공 데이터 대신 활용해 계산량과 정확도에서 더 높은 성능을 기대할 수 있다.

실험을 통하여 기존의 비가공 데이터를 이용한 이상 탐지 기법의 성능보다 더 높아짐을 확인하였다.

Ⅱ. 본론

A. 문제 정의

본 논문에서는 네트워크로부터 수집한 데이터 중 네트워크 트래픽의 TCP dump 데이터 X 를 메타 특성으로 변환하여 이상 탐지의 성능을 높이고자 한다.

비가공 데이터 X 로부터 매 시간 t 초 내 데이터 $X_t = [X_t^d : X_t^p]$ 은 패킷 도착 시간 X_t^d 과 패킷 사이즈 X_t^p 으로 구성되어 있으며, 각각은 $X_t^d = [x_1^d, \cdots, x_n^d]^T$, $X_t^p = [x_1^p, \cdots, x_n^p]^T$ 으로 정의하였다. 매 t 마다 들어오는 패킷의 수는 상이하므로 n 은 시간에 독립적이다. X_t 의 레이블은 네트워크 상태로, 정상(y_0)과 비정상(y_1)으로 표기하였다. 즉, $\{X_t, y_i\}_{i \in \{0,1\}}$ 은 하나의 데이터 포인트를 의미하며, 본 논문에서는 X_t 로부터 M개의 메타특성 $F_t = [f_1, f_2, \cdots, f_M]$ 으로 확장시켜 새로운 데이터 포인트로 삼았다.

B. 제안 방법

본 논문은 비가공 네트워크 데이터를 메타 특성으로 변형하여 이상 탐지에 이용하는 것을 제안하였으며, 각 제안하는 메타 특성 추출 및 확장 구조는 그림 1과 같다.

먼저, X_t^d 와 X_t^p 의 통계적 특성 i 개씩을 구한 X_t 의 특성을 $F_t^S = [f_1, f_2, \cdots, f_{2i}]$ 라고 한다. 다음으로는, X_t^d 와 X_t^p 사이의 특성을 구하여 확장한 데이터 X_t 의 특성을 $F_t^I = [f_{2i+1}, \cdots, f_M]$ 로 표기하였다. 앞서 구한 특성 F_t^S 와 F_t^I 을 합하여 최종적으로 데이터 X_t 의 특성 F_t 을 메타 특성으로 정의하였다.

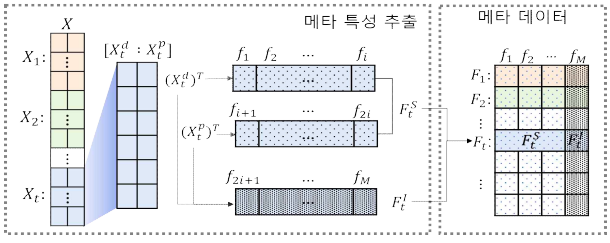

그림 1. 메타 특성 추출 및 확장 구조

Ⅲ. 실험 및 실험 결과

A. 실험 데이터

비정상 네트워크 탐지를 위해 다음과 같은 네트워크 데이터를 사용하였다.

1. CIDDS 데이터 셋: 이상 징후 기반 네트워크 침입 탐지 시스템에 대한 평가에 이용되는 데이터 셋으로 흐름 기반 데이터 집합으로 레이블이 지정되며 단방향 NetFlow 데이터를 포함한다. NetFlow의 10가지 속성과 레이블이 지정된 클래스 및 공격 유형 등이 포함되어있다 [7].

2. Network attack 데이터 셋: 네트워크 공격 분류를 위해 네트워크 패킷 캡처를 수행하여 수집된 데이터 셋으로 85개의 특성과 함께 정상과 비정상(DDoS)으로 레이블링 되어있다. 네트워크 장치에 생성된 IP 흐름의 정보, 즉 소스 및 대상 IP 주소, 포트 번호, 도착 간격, 해당 흐름에서 클래스로 사용되는 프로토콜 등을 포함하고 있다 [8].

3. IP Network traffic 데이터 셋: 86개의 특성과 소스 및 대상 IP 주소, 포트 번호, 도착 간격, 사용된 프로토콜을 보유하고 있으며 트래픽 흐름과 트래픽 데이터의 기능을 나타낸다 [9].

네트워크 데이터 특성 중에서 패킷 도착 시간과 패킷 사이즈 정보만을 활용해 메타 특성으로 확장하였다.

B. 실험 시스템

본 논문에서는 메타 특성을 활용한 이상탐지기법의 성능이 비가공 데이터를 활용 할 때보다 높아짐을 보이기 위해 최근에 개발된 세 가지의 이상탐지기법([4]-[6])에 서로 다른 이상 네트워크 데이터 셋([7]-[9])을 적용하였다. 본 논문에서의 실험에서는 비가공 데이터를 이용하여 메타 특성을 만드는 과정에서 F_t^S은 X_t^d와 X_t^p의 평균, 분산, 표준편차 여섯 개의 특성을 이용하였고, F_t^I로는 X_t^d와 X_t^p의 간의 내적값, 상호 정보, 두 확률 분포의 차이 계산하여 총 6개의 특성으로 F_t를 구성하였다.

비가공 데이터 X_t와 메타 특성으로 이루어진 데이터 F_t의 학습 성능을 평가하기 위해 네 가지의 성능지표 : 시간 복잡도, 재현율, 정밀도, AUC가 활용되었다. 재현율과 정밀도는 결함을 탐지하지 못하는 상황과 오검출의 발생에 대해 평가할 수 있는 지표로 이상 탐지 분야에서는 널리 사용되는 지표이다. 또한 임계값의 변화에 따른 재현율의 변화를 그린 ROC 곡선 아래의 너비인 AUC으로 이상 탐지 모델의 성능을 평가하였다.

C. 실험 결과

세 가지의 이상 탐지 모델에 비가공 데이터와 메타 특성을 활용하여 각각의 경우에 대하여 이상 탐지를 수행한 뒤 성능 지표를 통해 제안방법의 성능을 평가하였다.

결과는 표1과 같이 세 가지의 데이터 셋에서 모두 제안한 방법이 기존의 방법보다 시간 복잡도는 감소하고, 4가지의 성능지표에서는 값이 증가하였다. 최대 0.46%까지 계산 복잡도가 감소하였으며 재현율과 정밀도에서 최대 360%, 350% 증가를 보였다. 이는 메타 특성이 비균등한 이상탐지데이터를 효율적으로 처리하는 것을 알 수 있다.

Ⅳ. 결론

본 논문에서는 이상 탐지기법의 성능을 높이기 위하여 메타특성을 이용하는 이상 탐지기법을 제안하였다. 비가공 데이터를 이용하는 이상 탐지기법은 계산량이 많아 데이터의 특성을 잘 살리되, 계산량을 줄일 수 있는 방법을 제안하였다. 두 종류의 실험 데이터 셋에 대한 실험 결과를 통해 제안된 방식이 기존 방식보다 높은 성능을 가짐을 알 수 있다.

CIDDS		시간복잡도[초]	재현율	정밀도	AUC
비가공데이터	[4]	8994.048	0.956	0.956	0.957
	[5]	20742.663	0.453	0.601	0.401
	[6]	55.639	0.233	0.434	0.267
제안방법	[4]	1082.773 (12.04%)	0.992 (104%)	0.992 (104%)	0.993 (104%)
	[5]	331.457 **(1.60%)**	0.998 (220%)	0.986 (164%)	0.937 **(192%)**
	[6]	1.784 (3.21%)	0.519 **(223%)**	0.931 **(215%)**	0.456 (154%)
Network attack		시간복잡도[초]	재현율	정밀도	AUC
비가공데이터	[4]	7136.457	0.395	0.395	0.306
	[5]	19657.476	0.218	0.214	0.185
	[6]	52.696	0.519	0.627	0.273
제안방법	[4]	684.034 (9.59%)	0.963 **(244%)**	0.963 (244%)	0.971 (244%)
	[5]	277.781 **(1.41%)**	0.491 (225%)	0.749 **(350%)**	0.611 **(277%)**
	[6]	0.757 (1.44%)	0.649 (125%)	0.942 (150%)	0.824 (139%)
IP Network traffic		시간복잡도[초]	재현율	정밀도	AUC
비가공데이터	[4]	35346.387	0.146	0.149	0.197
	[5]	1475.354	0.131	0.568	0.267
	[6]	61.829	0.491	0.839	0.646
제안방법	[4]	164.198 **(0.46%)**	0.429 (294%)	0.429 **(288%)**	0.361 **(288%)**
	[5]	216.469 (14.67%)	0.471 **(360%)**	0.613 (108%)	0.428 (241%)
	[6]	0.702 (1.14%)	0.667 (136%)	0.997 (119%)	0.833 (129%)

표 1. 비가공 네트워크 데이터와 메타 특성 데이터의 이상 탐지 성능표

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(, No.2021-0-00739분산/협력 AI 기반 5G+ 네트워크 데이터 분석 기능 및 제어 기술 개발, No. 2019-0-00024,네트워크 자동화를 위한 개방형 네트워크 데이터 분석 기반 지도형 애자일 머신러닝 기술 개발)과 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2020R1A2B5B01002528).

참 고 문 헌

[1] Terzi. " Big data analytics for network anomaly detection from netflow data, "Computer Science and Engineering (UBMK), pp. 592-597, Oct. 2017

[2] Habeeb RA, Nasaruddin F, Gani A, Hashem IA, Ahmed E, Imran M, "Real-time big data processing for anomaly detection: A Survey," International Journal of Information Management, vol. 45, pp 289-307, 2019

[3] Nam, Taewoo and Oh, Dong-Geun, "메타데이터의 의미론적 확장에 관한 연구," 한국문헌정보학회지, vol. 44, no. 4, pp. 373 - 393, Nov. 2010.

[4] Liron Bergman and Yedid Hoshen, "Classification-Based Anomaly Detection For General Data," ICLR, May, 2020

[5] Yue Zhao, "LSCP: Locally Selective Combination in Parallel Outlier Ensembles," SDM (SIAM International Conference on Data Mining), 2019

[6] Hongzuo Xu, "MIX: A Joint Learning Framework for Detecting Both Clustered and Scattered Outliers in Mixed-Type Data," IEEE International Conference on Data Mining(ICDM), 2019

[7] https://www.kaggle.com/kartikjaspal/server-logs-suspicious

[8] https://www.kaggle.com/akshat4112/attackdataset

[9] https://www.kaggle.com/akshat4112/networkanamolydetection