비지도 학습기반 IoT 데이터의 이상탐지 기술 연구

문애경, 송윤정
한국전자통신연구원 지역산업IT융합연구실
akmoon@etri.re.kr, yjsong@etri.re.kr

Study of Unsupervised Learning based Anomaly Detection for IoT Data

Aekyung Moon, Yunjeong Song
ETRI, Regional Industry IT Convergence Research Section

요 약

본 논문은 신호 처리 알고리즘을 이용하여 데이터를 압축하면서 소수의 데이터 프로파일을 사용하여 데이터의 정상과 비정상을 구별하는 비지도 학습기반의 데이터 이상 탐지 기술을 제안한다. 제안된 방법은 기존의 손실압축 방법에 데이터 이상 감지 기능을 확장한 것으로, 데이터 압축 효과와 이상 탐지 기능을 동시에 구현하였고 데이터 오염률에 따라 이상 데이터를 삽입한 합성데이터를 사용하여 기존의 이상 탐지 기법과 비교하였다.

Ⅰ. 서 론

센서 기술의 발전으로 IoT 기술을 활용하는 애플리케이션의 역활이 증대하고 있고, 수집된 IoT 데이터의 양이 증가함에 따라 효율적 데이타 관리 및 분석에 대한 요구사항이 증가하고 있다[1,2]. 또한, 센서 및 네트워크의 오류와 장애로 인하여 이상 데이터가 발생할 수 있고 센서 노드에서 사용할 수 있는 리소스의 제약 및 전송 기능에 대한 해결도 필요하다. 데이터 이상 현상은 다양한 분야에서 발생할 수 있는 만큼, 특정 비즈니스 도메인에 특화된 이상 탐지를 위한 연구가 진행되고 있다. 정상(normal) 상황과 이상(anomaly)인 상황을 명확히 규정할 수 있다면, 특정 이벤트가 발생했을 때 명확하게 이상을 탐지할 수 있을 것이다. 하지만 이전에 발생했던 이상 상황에 대해서는 정의할 수 있겠지만, 예상치 못한 이상 상황은 항상 존재하고 모든 이상 관련 규칙을 만들 수 없어서 데이터 이상 상황을 명확하게 판단하는 것은 어려운 문제이다.

본 논문에서는 IoT 데이터 스트림을 변환한 후 변환된 정보에서 소수의 샘플링 데이터를 추출하고, 샘플링된 데이터로 얻은 정보에서 정상과 비정상간 데이터를 구별할 수 있는 이상 탐지 기법을 제안한다. 즉, 원본 데이터에 변환을 적용하고 상위 k 개의 주요 구성 요소를 샘플링하며 k 값의 차이를 기반으로 데이터 이상을 탐지한다. 제안된 알고리즘의 성능을 검증하기 위하여 실제 무선 센서 노드에서 수집 한 IoT 데이터 세트(온도, 습도 및 CO_2)를 사용하여 압축 성능을 검증하고, 기존의 기법과 데이터 오염률에 따른 이상 탐지 정확성을 검증한다.

2. IoT 이상데이터 감지 기술

수집 된 IoT 데이터를 효과적으로 관리하기 위해서는 다음과 같은 이슈가 있다. 첫째, 기존 IoT 환경과 마찬가지로 센서 노드는 처리 능력, 대역폭, 에너지 및 스토리지 측면에서 제한된 리소스와 기능을 가지고 있다[3], 지속적으로 데이터를 수집하는 센서 노드는 저장 및 전송 비용을 줄일 수 있도록 데이터의 효율적 관리 방식이 필요하다. 둘째, 센서 노드에서 수집 된 데이터는 센서 및 네트워크 오류로 인해 이상이 발생할 수 있고 이것은 제어노드 또는 게이트웨이에서 부정확 판단이나 필요치 않은 제어 작업을 수행할 수 있다. 이러한 예상치 못한 이상 현상은 수집된 IoT 데이터 세트에서도 자주 발생한다. 그림 1은 실제 이상 데이터를 포함하는 수집된 온도 데이터 샘플을 보여준다. 따라서 제어 작업을 수행하기 전에 이러한 이상 징후가 있는지 감지하고 애플리케이션에서 확인할 수 있도록 이상 데이터 상황을 시스템에 전달해야 한다[4 5]. 효과적인 이상 탐지 기법은 IoT 시스템을 활용한 의사결정 지원시스템에서 발생 이상 상황에 따라 적절한 대응과 조치를 할 수 있게 한다[6].

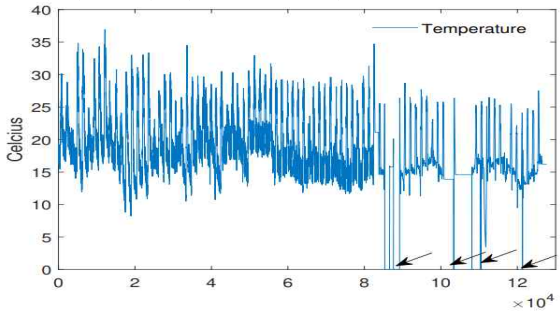

그림 1 이상 데이타를 포함하는 수집된 온도 데이터

이상 탐지는 예상한 정상 데이터 패턴과 다른 패턴을 찾는 과정이다[5], 최근의 이상 탐지 기술은 학습 단계 후 정상 상태와 이상 상태를 판단하는 머신러닝 (ML) 방법을 적용하고, 지도(supervised), 비지도(unsupervised), 및 준지도(semi-supervised)의 세 가지 형태로 분류 할 수 있다[7,8.]. 지도학습 기반 이상 탐지 모델은 레이블이 지정된 학습 데이터에서 정상 및 이상 동작을 분류하는 방법을 학습한다. 그러나 적절한 레이블이 있는 대규모 데이터 세트를 얻는 것은 어려운 문제이고, 레이블을 어노테이션하기 위해서는 전문가의 도메인 지식이 필요하다. 따라서, 실제 시나리오에서 대표적인 데이터 패턴에 레이블을 지정하는 것과 정확한 레이블링은 애플리케이션에 따라 달라지기 때문에 해결에 어려움이 있다. 반면에 비지도 학습방법은 학습시 레이블을 가진 데이터가 필요 없다는 장점이 있는 반면에 때문에 데이터 자체의 특성을 파악이 필요하다.

Ⅲ. 이상감지 알고리즘

제안하는 이상 탐지 기법은 간단하면서도 효율적인 데이터 손실 압축

알고리즘을 사용하여 낮은 샘플링 데이타로 원본 데이터로 재구성가능한 센서 노드의 기능을 확장한다. 그림 2는 센서 노드가 데이터 포인트를 수집하여 손실 압축을 통하여 추출한 프로파일을 사용하여 데이터 이상을 감지하는 비지도 학급기반 이상 탐지 기법을 보여준다. 제안된 기법은 전체 원본 데이터 포인트를 검사하지 않고 샘플링 된 데이터 포인트의 k 값을 조사하여 이상을 감지하기 때문에 리소스가 제한된 센서 노드에서 스토리지 및 계산 사용에 효율적이다[9]. 그림 2에서와 같이 원본 데이터 X를 DCT 기반 구성 요소로 변환하여 데이터를 희소하게 만든다. 이러한 방식으로 t 번째 샘플링 기간 ($1 \le i \le N$)의 원래 데이터 포인트 $X_i(t)$는 변환 된 데이터 포인트 $\hat{X}_i(t)$로 변환된다. 단, 여기서 i는 블록을 의미하는 것으로 압축 및 이상 탐지 적용 단위가 된다.

IV 평가

이상 탐지 기법의 성능을 평가하기 위해 다음의 성능평가지수를 사용하고 데이터 오염 비율을 변경하면서 평가한다. 제안된 알고리즘은 압축의 기능도 포함하고 있어서 압축비율도 추가적으로 측정한다.

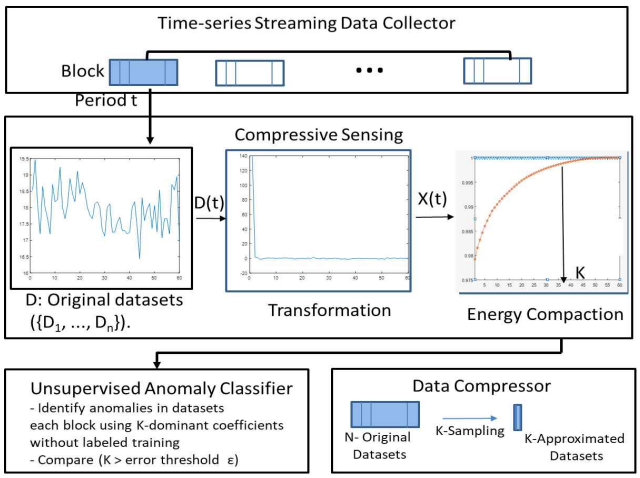

그림 2 이상탐지 알고리즘 개요

- 압축 비율 : 달성 가능한 압축 비율:

$$R = \frac{|D| - |D'|}{|D|} \times 100\%$$

- 이상감지의 정확성 (Accuracy): $\frac{TP + TN}{TP + TN + FP + FN}$
- PSNR (Peak Signal-to-Noise Ratio) : 복원 오류 측정

$$PSNR_M = 20\log10(Max(x) - Min(x)) - 10\log10(MSE)$$

압축 비율과 압축에 따른 오류률은 (온도, 습도 및 CO_2) 데이터를 평가하였고 온도데이타의 경우 압축률과 PSNR은 98.3%, 32,41로 각각 나타났다. 오염률 0.1은 비정상 데이터의 10 %가 각각 학습 및 테스트 데이터 세트에 포함되어 있음을 의미한다. 그림 3에서 볼 수 있듯이 제안된 기법은 92 % -100 %의 이상을 감지 할 수 있을 뿐만 아니라 오염 비율이 증가함에 따라 감지 정확도도 향상되는 것으로 나타났다. 더 높은 오염 비율과 함께 제안된 이상 탐지의 더 높은 검출 정확도의 이유는 더 많은 이상 데이터가 존재할 때 K가 이에 따라 증가하여 결과적으로 증가하는 오염 비율에 잘 적응하는 것으로 보인다.

V. 결론

본 논문에서 IoT 데이터 세트의 신호변환을 통한 추출 프로파일을 사용하여 이상을 감지하는 비지도 기반의 이상 탐지 기법을 제안했다. 센서 노드에서 수집 된 IoT 데이터가 신호변환을 통한 데이터 추출 프로파일과 데이터 이상과 상관관계를 갖는다. 제안된 이상탐지 기법을 합성데이타에 신호변환을 적용하여 변환후 데이터 프로파일 추출하고 이를 이상탐지 기법에 활용하고 기존의 기법들과 성능을 비교하였다. 제안된 이상 탐지 기법은 원본 데이터의 98.3%,(온도), 98.1.9%(습도), 98.3%(CO_2)로 압축이 가능하고 92%-100 %의 이상탐지 성능을 나타내었다.

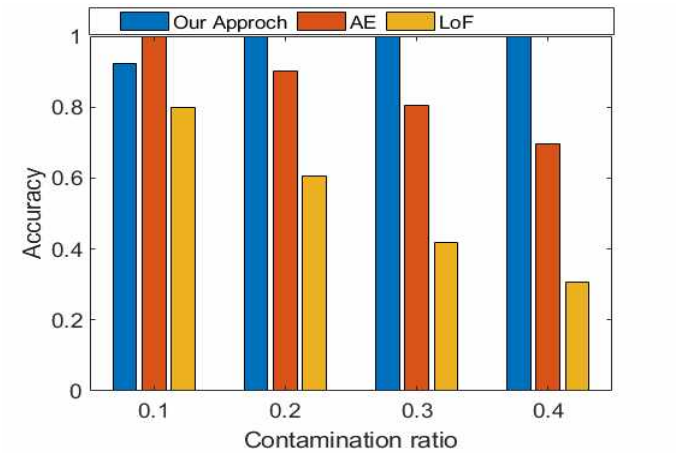

그림 3 이상 탐지 기법의 성능 비교

ACKNOWLEDGMENT

본 연구는 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음.[21ZD1100, , 대경권 지역산업 기반 ICT융합기술 고도화 지원사업

참 고 문 헌

[1] A. Ukil, S. Bandyopadhyay, and A. Pal, “IoT Data Compression: Sensor-Agnostic Approach,” in2015 Data Compression Conference, April, 2015, pp. 303 - 312

[2] O. Tuncer, E. Ates, Y. Zhang, A. Turk, J. Brandt, V. J. Leung, M. Egele,and A. K. Coskun, Online Diagnosis of Performance Variation in HPC Systems Us1ing Machine Learning,” in IEEE transactions on parallel and distributedsystems, 2019.

[3] S. Kartakis and J. A. McCann, “Real-time Edge Analytics for Cyber Physical Systems using Compression Rates,” in 11th International Conference on Autonomic Computing (ICAC ’14), 2014,

[4] H. Ren, B. Xu, Y. Yujing Wang, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, and Q. Zhang, “Time-Series Anomaly Detection Service at Microsoft ,” in KDD, 2019.

[5] S. A. Haque, M. Rahman, and S. M. Aziz, “Sensor Anomaly Detection in Wireless Sensor Networks for Healthcare,” Sensors, vol. 15, no. 4, pp. 8764 - 8786, 2015.

[6] L. Mart´ı, N. Sanchez-Pi, J. M. Molina, and A. C. B. Garcia, “Anomaly Detection Based on Sensor Data in Petroleum Industry Applications,” Sensors, 2015.

[7] A. Borghesi, A. Libri, L. Benini, and A. Bartolini, “Online Anomaly Detection in HPC Systems ,” in IEEE International Conf

[8] A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, and L. Benini,“Anomaly Detection using Autoencoders in High Performance Com-puting Systems ,” inProceedings of the AAAI Conference on ArtificialIntelligence, 2018

[9] A. Moon, X. Zhuo, J. Zhang, and S. W. Son, “AD2: Improving Qualityof IoT Data through Compressive Anomaly Detection,”

IEEE Big Data, 2019