

불확실성 정보를 활용한 오토인코더 기반 이상탐지 모델

류승형, 유용균, 서호건*
한국원자력연구원

ashryu@kaeri.re.kr, ygyu@kaeri.re.kr, *hogeony@kaeri.re.kr

Autoencoder based anomaly detection model utilizing uncertainty information

Seunghyoung Ryu, Yonggyun Yu, Hogeon Seo*
Korea Atomic Energy Research Institute

요약

이상 탐지는 사고와 같은 비정상적인 상태를 분류하는 문제로, 그 특성과 영향력으로 인해 다양한 산업분야에서 그 중요성이 증가하고 있다. 이에 따라 최근 빅데이터와 인공지능 기술의 발전에 힘입어 데이터 중심의 이상탐지 기술이 활발히 연구되고 있으며, 대표적인 딥러닝 기반 이상탐지 기법으로 오토인코더를 활용한 이상탐지 기술을 예로 들 수 있다. 오토인코더 기반 이상탐지 모델은 정상 데이터를 학습하여 비정상 데이터에 대한 복원오차를 기준으로 정상과 비정상을 분류한다. 본 연구에서는 이와 같은 오토인코더 기반의 이상탐지 모델의 성능향상을 위하여 분위수 오토인코더 (Quantile autoencoder, QAE)와 이상치 누적 (Abnormality accumulation, AA) 기법을 제안하였다. 제안된 QAE와 AA 기법을 통해 복원오차뿐만 아니라 데이터의 불확실성 정보와 재귀적 복원오차를 이상점수 계산에 활용할 수 있으며, 기존 오토인코더 기반 모델 대비 다수의 데이터셋에서 더 높은 이상탐지 성능을 얻을 수 있었다.

I. 서론

이상 탐지는 사고, 결함, 오류와 같이 정상적이지 않은 패턴 및 데이터를 분류하는 문제로, 발생 확률은 낮지만 부정적인 영향을 미치기 때문에 사기 탐지[1], 영상 감시[2], 예방 정비[3] 등을 포함한 다양한 산업 분야에서 그 중요성이 점차 증가하고 있다. 이에 따라, 최근 빅데이터와 인공지능 기술 발전에 힘입어 딥러닝 기반의 이상탐지 (Deep anomaly detection, DAD) 기술이 활발히 연구되고 있다. 일반적으로 이상 탐지는 이상 점수에 대한 이진 분류 문제로서, 이상 판별 기준값에 대비하여 이상점수가 높은 경우 비정상으로, 낮은 경우 정상으로 판단한다. 대표적으로 오토인코더를 활용한 이상 탐지 모델에서는 복원오차를 이상 점수로 활용한다. 이상 탐지를 수행하기 위해 오토인코더는 정상 데이터로 훈련되며, 정상 데이터를 잘 복원하기 위한 특징들을 학습하게 된다. 따라서 학습된 정상 데이터에 대해서는 낮은 복원 오차를 갖고, 학습되지 않은 비정상 데이터에 대해서는 높은 복원 오차를 갖으므로 이러한 값의 차이를 활용하여 정상과 비정상을 분류하게 된다. 오토인코더 구조를 바탕으로 이후 다양한 이상탐지 모델들이 개발되었으며, 이상점수로 잠재공간에서의 복원 오차를 활용하거나 [4] 또는 GAN 구조에서의 판별자의 손실함수를 이상점수로 함께 활용하는 방법 [5] 등이 제안되었다.

본 논문에서는 오토인코더 기반 이상탐지 모델의 성능 향상을 위한 방법으로 이상 점수 계산에 있어서 기존 복원 오차와 더불어 데이터의 불확실성 정보와 재귀적 복원오차를 활용하였다. 이를 수행하기 위하여 분위수 오토인코더 (Quantile autoencoder, QAE)와 이상치 누적 (Abnormality accumulation, AA) 기법을 제안하였고,

QAE-AA의 적용을 통해 얻은 오차 벡터로부터 마할라노비스 거리 (Mahalanobis distance)를 계산하여 이상 점수로 사용하였다.

II. 분위수 오토인코더와 이상치 누적 기법

기존 오토인코더 모델이 입력과 출력의 평균제곱오차 (Mean squared error, MSE)를 최소화하도록 훈련되는 것과 달리 QAE는 기존 입력값에 대한 복원뿐만 아니라 출력 분포에 대하여 특정 분위수 $\tau \in \{0,1\}$ 에 대한 예측을 수행한다. 즉, 누적 확률 분포의 값이 τ 가 되는 지점의 변수값을 얻을 수 있으며, QAE는 중앙값을 포함한 세가지 분위수에 대한 예측 ($\tau_l < 0.5, \tau_m = 0.5, \tau_u > 0.5$)을 수행한다. QAE의 학습을 위한 손실함수로서 오토인코더에서 사용되는 MSE 대신 pinball loss를 사용하며, QAE 학습이 완료되면 주어진 입력 샘플에 대하여 이상점수를 계산하기 위해 다음의 오차항 $\epsilon = [x - \hat{x}_{\tau_m}, \hat{x}_{\tau_u} - \hat{x}_{\tau_l}]$ 을 산출한다.

이상치 누적 기법은 오토인코더의 출력값을 다시 오토인코더의 입력으로 사용하여 재귀적으로 복원오차를 다시 얻는다. ϵ^1 이 최초로 QAE를 통과하여 얻은 오차항이라고 한다면, ϵ^2 는 $Q(\hat{x}_{\tau_m})$ 의 출력값으로부터 계산한 오차항에 해당한다.

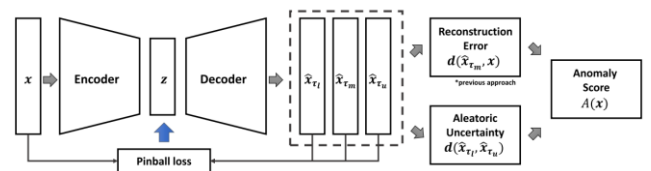


Figure 1. QAE 구조도

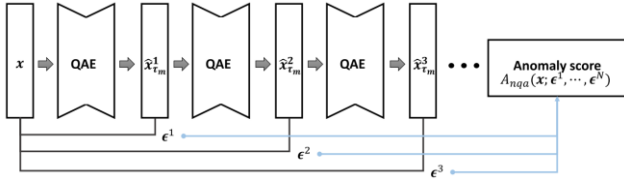


Figure 2. QAE-AA 기법 개념도

복원값을 재귀적으로 네트워크의 입력하는 과정을 N 회 수행하면, 해당 오차항들을 합쳐 하나의 벡터 $[\epsilon^i]_{i=1}^N$ 를 생성할 수 있다. 최종적으로 QAE-AA 의 이상점수는 마할라노비스 거리를 사용하며 다음의 수식을 통해 계산된다.

$$([\epsilon^i]_{i=1}^N - \mu) S^{-1} ([\epsilon^i]_{i=1}^N - \mu)^T$$

이때 S 와 μ 는 학습에 사용한 정상 데이터의 오차항에 대한 평균과 Covariance matrix 이다.

III. 실험 결과

제안된 모델의 이상 탐지 성능을 평가하기 위하여 6 개의 이상탐지 데이터셋을 구성하여 AUROC 를 계산하였으며, 성능 비교를 위해 오토인코더를 포함한 비지도학습 기반 이상탐지 모델 (AE, OCSVM, DSVDD)의 AUROC 를 표 1 에 정리하였다. 이상탐지 문제와 데이터셋의 구성은 [4]의 구성을 참고하였다. AUROC 가 100 에 가까울수록 정상과 비정상 데이터의 이진 분류가 잘 이뤄진다는 것을 의미하며, 결과를 살펴보았을 때 제안한 QAE-AA 모델이 평균적으로 가장 높은 AUROC 를 갖는 것을 확인할 수 있다.

Table 1. 모델별 이상탐지 성능 비교

Data	OCSVM	AE	DSVDD	QAE-AA
MI-F	77.7	60.7	73.4	80.7
MI-V	84	89.7	66	92.0
EOPT	59.7	61	50.8	62.7
RARM	74.9	68.7	75.6	80.7
SNSR	95.9	97.9	95.3	99.4
SNSR(M)	52.4	61.3	50.6	74.4
Avg	74.1	73.2	68.6	81.7

IV. 결론

본 논문에서는 오토인코더 기반 이상탐지 모델의 성능을 향상시키기 위하여 데이터의 불확실성과 재귀적 오차를 활용하는 QAE-AA 모델을 제안하였다. 다양한 데이터셋에 대한 실험 결과 데이터의 불확실성 정보와 재귀적 복원오차를 추가로 사용하였을 때 기존 오토인코더 모델 대비 이상탐지 성능이 증가하는 것을 확인할 수 있었다. 추후 연구로 시계열이나 이미지 데이터에 대한 이상탐지 연구로 확장할 계획이다.

ACKNOWLEDGMENT

This work was supported in part by the Korea Atomic Energy Research Institute (KAERI) Research and Development

Program under Grant KAERI-524450-22, and in part by the National Research Foundation of Korea (NRF) funded by the Korean Government Ministry of Science and ICT under Grant NRF-2021R1F1A1051290..

참고 문헌

- [1] U. Porwal and S. Mukund, "Credit card fraud detection in e-commerce: An outlier detection approach," 2018, arXiv:1811.02196.
- [2] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2017, pp. 1577-1581.
- [3] T. Kuzin and T. Borovicka, "Early failure detection for predictive maintenance of sensor parts," in Proc. ITAT. Princeton, NJ, USA: Citeseer, 2016, pp. 123-130.
- [4] K. H. Kim, S. Shim, Y. Lim, J. Jeon, J. Choi, B. Kim, and A. S. Yoon, "RaPP: Novelty detection with reconstruction along projection pathway," in Proc. Int. Conf. Learn. Represent., 2019, pp. 1-14.
- [5] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semisupervised anomaly detection via adversarial training," in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2018, pp. 622-637.