Unsupervised Anomaly Detection in Time-series: An Extensive Evaluation and Analysis of State-of-the-art Methods

https://arxiv.org/abs/2212.03637·

0. Introduction

- 라벨 없이 정상 데이터만으로 이상을 탐지하는 비지도 시계열 이상 탐지는 난이도가 높음
- 기존 연구는 단순 지표(F1 등)에 집중, 실용적 요소는 미흡
- 본 논문은 여러 최신 모델을 공통 조건에서 비교하고 성능·안정성·복잡도·이상 유형별 성능을 함께 평가함

1. Overview

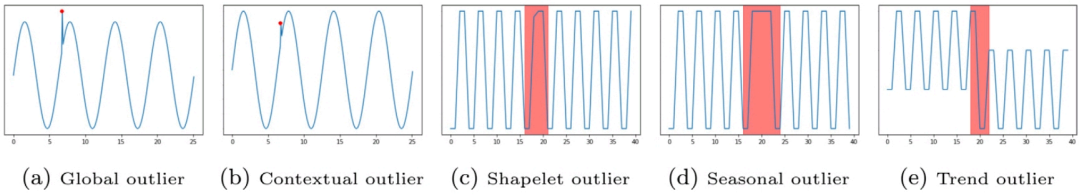
- 전통, 머신러닝, 딥러닝 기반 모델을 동일 프로토콜로 실험
- 평가 항목: 표준 지표 + 시계열 특화 지표 + 모델 크기 + 안정성
- 다양한 공개 데이터셋 사용
- 목표: 공정 비교를 통해 실제 적용 가능성 평가 및 연구 방향 제시

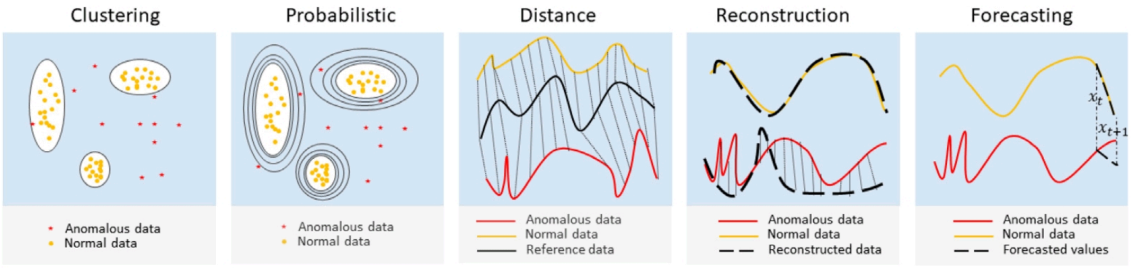
2. Challenges

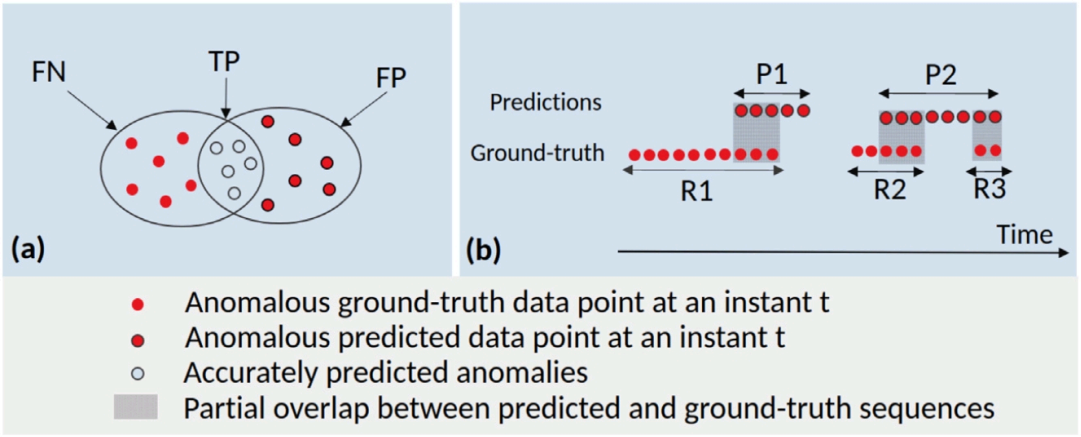
- 비지도 학습: 이상 라벨 부재로 경계 설정 어려움
- 데이터 불균형: 이상 비율 낮고, 구간 이상까지 고려 필요

- 평가 지표 한계: 시계열 특성 반영 부족
- 모델 복잡도: 성능과 효율성의 균형 문제
- 안정성: 초기화나 데이터 변화에 따른 성능 편차 큼
- 재현성: 연구별 설정 차이로 공정 비교 어려움

3. Method







- 비교 프레임워크: 동일 전처리·데이터 분할·평가 기준 적용
- 평가 대상: ARIMA, Isolation Forest, Autoencoder, VAE, Transformer 등
- 지표 구성:
  - 기본: Precision, Recall, F1

    - 특화: 구간 단위 정확도, 탐지 지연, 안정성, 복잡도
- 이상 유형별 분석: point / collective / contextual
- Ablation 실험**:** 노이즈·비율·길이 변화에 따른 민감도 평가

4. Experiments

- 데이터셋: NASA bearing, ECG, KPI 등 벤치마크 사용
- 전처리: 동일한 정규화·분할 기준 적용
- 비교 방식:
    - 동일 설정하 반복 실험(5회 평균)
    - 이상 비율·노이즈·샘플링 변화 실험 포함
- 평가 항목: 성능, 복잡도, 안정성, 이상 유형별 결과

5. Results

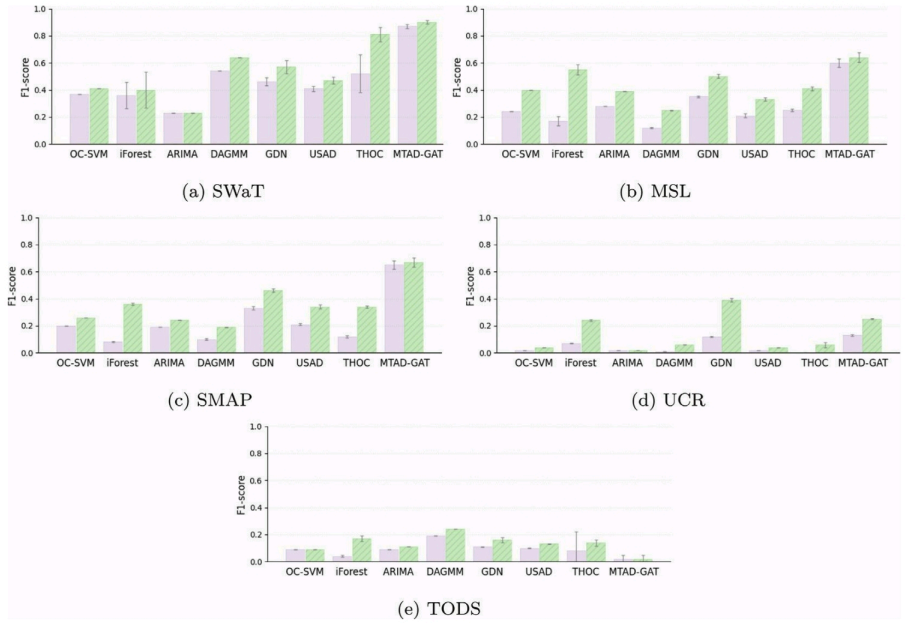


Fig. 6. Mean F1-Score on the five datasets. The non-hatched and hatched bars correspond to the mean F1-Score with and without Point Adjustment (PA), respectively. The vertic black line represents the standard deviation over five runs.

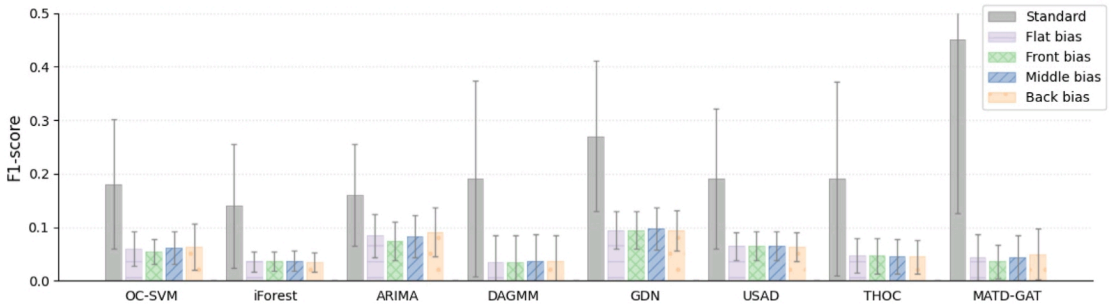


Fig. 7. The mean performance per method on all datasets using the range-based metrics of Tatbul et al. (2018), with different location biases.

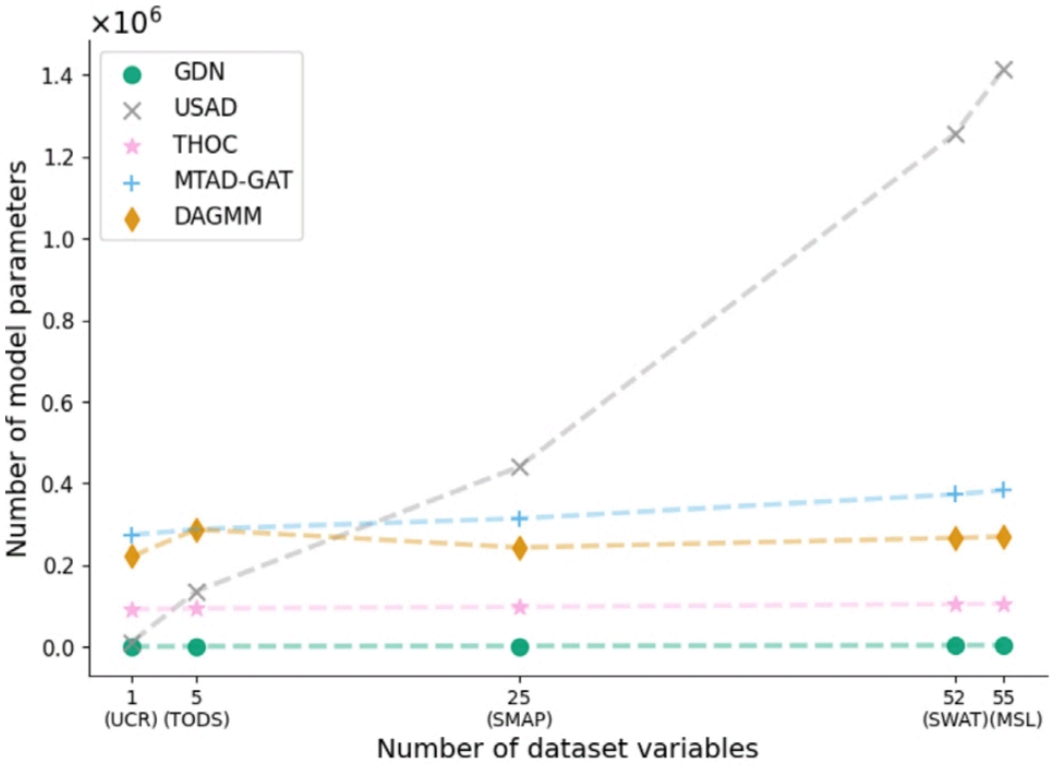


Fig. 8. Relation between the number of the parameters of the model and the number of features in the considered dataset.

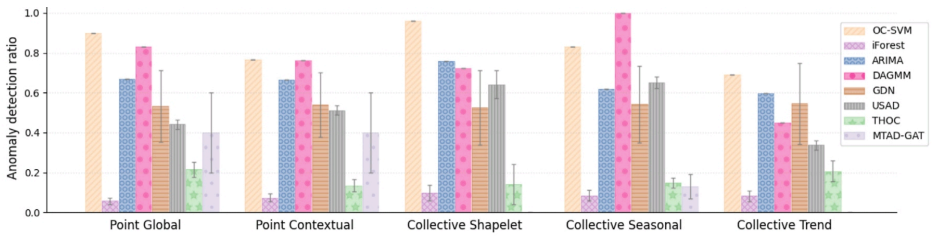


Fig. 9. The ratio of true anomalies detected for each tested method when varying the anomaly types. All methods succeeded in partially detecting each anomaly type, except MTAD-GAT which was unable to detect any collective trend anomaly.

- 성능: 딥러닝 모델이 전반적으로 우수하나 유형별 편차 존재

- 시계열 지표: 긴 구간 이상엔 일부 모델 취약
- 효율성: Transformer는 느리고 무겁고, AE/IF는 빠름
- 안정성: 통계 모델은 일관성 높고, 딥러닝은 변동 큼
- 유형별 성능:
  - Point: 대부분 우수
  - Collective: LSTM/VAE 강세
  - Contextual: Transformer 유리

6. Insight

- 단순 지표보다 복잡도·안정성·유형별 성능을 함께 고려해야 함
- 높은 성능보다 안정적이고 효율적인 모델이 실무에 적합
- 이상 유형별로 최적 모델이 다름 → 사전 분석 필요
- 향후 방향:
  - 반지도/약지도 접근 결합
  - 경량화·실시간 탐지 연구
  - 모델 선택 자동화 및 앙상블 활용