# 시계열 이상 탐지 성능 향상을 위한 코사인 유사도 기반 가중 보간 기법 제안

김파란하늘, 이연서, 이의수, 정태윤, 현장훈*
국립한밭대학교, 로보볼트㈜
jhhyeon@hanbat.ac.kr

## ABSTRACT

시계열 데이터의 결측값 보간은 이상 탐지와 같은 작업에서 필수적이나, 기존 보간 기법들은 급격한 변화가 있는 구간에서 정보 손실과 패턴 왜곡의 문제가 발생하는 경우가 많다. 이에 본 연구에서는 결측값 주변을 코사인 유사도로 정량화하고, 이를 시그모이드 함수로 변환한 후, KNN 기반 보간값과 외삽값을 가중 평균하여 결합하는 새로운 보간 기법을 제안한다. 제안된 기법은 PCA 기반 이상 탐지 알고리즘에 적용되었으며, 결측률 5~20%의 다양한 조건에서 성능을 평가하였다. 실험 결과, 제안 기법은 F1-score 0.651을 기록하였으며, 이는 선형 보간법 대비 약 34% 향상된 수치로, 결측 데이터 환경에서 이상 탐지 성능을 효과적으로 향상시킬 수 있음을 입증하였다.

## INTRODUCTION

- ESS의 이상 탐지에는 PCA 기반 방법이 널리 사용되지만, 결측 데이터가 발생할 경우 성능이 크게 저하된다.
- 기존 선형 보간법은 급변 구간에서 패턴을 평탄화하여 이상 탐지에 부정적 영향을 미친다.
- 본 연구는 시계열 패턴 유사도를 반영한 코사인 유사도 기반 가중 보간 기법을 통해 이상 탐지 성능 향상을 목표로 한다.

## METHODS

### 데이터 구성

- 1분 간격으로 측정된 전압 시계열 데이터
- 결측값 처리(센서 오류로 인한 측정값 3.3V 미만 ⇨ NaN)
- 학습/검증/테스트 데이터 분할

### 데이터 전처리

- 결측률: 5%, 10%, 15%, 20%
- 동일한 결측 마스크로 조건 통일
- K-fold 교차 검증(8:2)

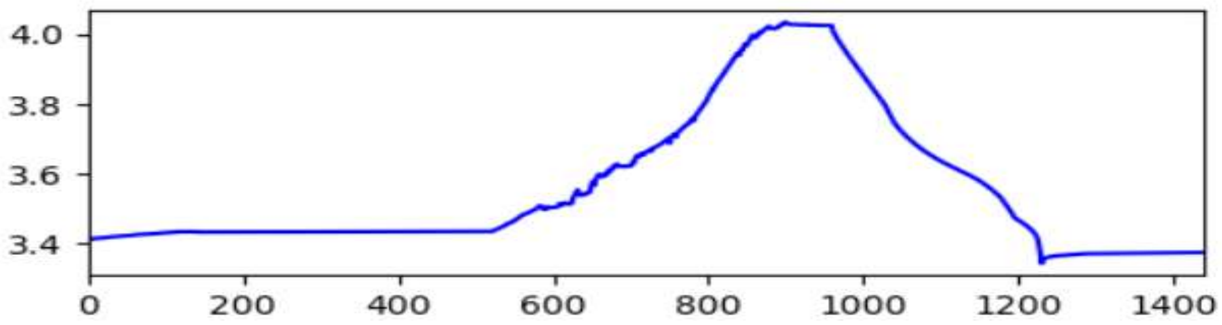


그림 1. 전처리된 전압 시계열 데이터

### 실험 방법

- PCA 기반 이상 탐지 알고리즘 설계.
- 학습 데이터를 사용하여 주성분 설정(분산 설명력: 95%).
- 검증 데이터를 활용하여 각 클래스별 MAE 분포 추출.
- Youden's J 통계량을 활용하여 normal / abnormal을 구분하는 임계값 설정

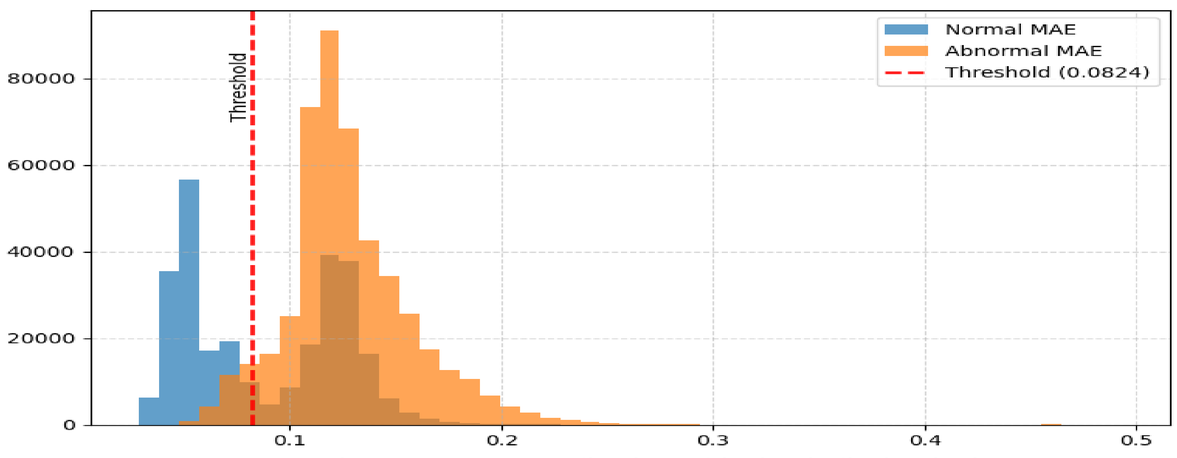


그림 2. MAE 분포에 따른 최적 임계값 설정

## PROPOSED METHOD

**초기 보간값 계산 (KNN Imputer)**

데이터 결측 구간에 KNN Imputer를 활용한 초기 보간값 생성.

**결측치 주변 패턴 유사도 계산**

1에 가까울수록 두 벡터의 방향성이 유사함.

$$\cos(\theta) = \frac{\mathbf{v}_{\text{left}} \cdot \mathbf{v}_{\text{right}}}{\|\mathbf{v}_{\text{left}}\| \cdot \|\mathbf{v}_{\text{right}}\|}$$

**가중치 변환 (Sigmoid)**

0 ~ 1 사이의 보간 가중치 w로 변환.

$$w = \frac{1}{1 + e^{-\alpha \cdot \cos(\theta)}}$$

**외삽값 계산**

결측치 기준 좌우 시계열 값의 기울기를 기반으로 산술평균하여 외삽값 계산.

$$x_{\text{ext}}^{\text{left}} = 2b - a, \quad x_{\text{ext}}^{\text{right}} = 2c - d \quad x_{\text{ext}} = \frac{1}{2}(x_{\text{ext}}^{\text{left}} + x_{\text{ext}}^{\text{right}})$$

**최종 보간값 산출**

유사도가 높을수록 기존 보간값에 가중치 부여하며 유사도가 낮을수록 외삽값에 가중치 부여.

$$\hat{x} = w \times x_{knn} + (1 - w) \times x_{ext}$$

## RESULTS

### 실험 결과

- 이상 탐지는 클래스 불균형으로 단일 평가 지표로 성능 평가 어려움.
- F1-score는 Precision과 Recall의 조화 평균으로 실질적인 이상 탐지 성능을 반영하는 핵심 지표로 간주됨.
- F1-score 기준 선형 보간 대비 약 34% 향상된 성능을 기록함.

| Missing Rate | Linear | | | Proposed | | |
|---|---|---|---|---|---|---|
| | F1 | PR | ROC | F1 | PR | ROC |
| 5% | 0.491 | 0.694 | 0.654 | 0.652 | 0.687 | 0.674 |
| 10% | 0.488 | 0.695 | 0.654 | 0.651 | 0.694 | 0.673 |
| 15% | 0.488 | 0.695 | 0.654 | 0.649 | 0.687 | 0.675 |
| 20% | 0.486 | 0.695 | 0.655 | 0.650 | 0.688 | 0.673 |
| Avg | 0.488 | **0.695** | 0.654 | **0.651** | 0.689 | **0.674** |

표 1. 결측 비율에 따른 이상 탐지 성능 평가 결과

### Reference

1. M. Crépey, A. Aouadi, and C. Rahal, "Anomaly Detection on Financial Time Series by Principal Component Analysis and Neural Networks," Algorithms, vol. 15, no. 3, pp. 1–21, 2022.
2. A. P. A. de Lima, G. F. Guedes, and R. M. S. Pereira, "Missing data in time series: A review of imputation methods and case study," Letters in the National Institute for Science and Technology in Machine Learning (L&NLM), vol. 20, no. 1, pp. 26–38, 2022.