시계열 이상 탐지를 위한 패턴 유사도 기반 가중 보간 기법

A Pattern Similarity-Based Weighted Interpolation Method for Time Series Anomaly Detection

Abstract

Imputation of missing values in time series data is essential for tasks such as anomaly detection, yet conventional methods often suffer from information loss and pattern distortion in regions of abrupt change. This study proposes a novel interpolation method that quantifies the pattern similarity between neighboring segments using cosine similarity, transforms it via a sigmoid function, and combines KNN-based imputation and extrapolation values through weighted averaging. The proposed method is applied to a PCA-based anomaly detection algorithm and evaluated under various missing rates (5 - 20%) in comparison with existing techniques, including linear, spline, KNN, and polynomial interpolation. Experimental results show that the proposed method achieves the highest average F1-score (0.651) across all conditions, demonstrating its effectiveness and robustness in anomaly detection.

Ⅰ. 서론

최근 친환경 에너지 확대와 전력망 효율화에 대한 요구가 증가함에 따라 에너지 저장장치(Energy Storage System, ESS)의 중요성이 부각되고 있다. 그러나 ESS의 급속한 보급과 함께 배터리 셀의 제조 결함이나 과충전으로 인한 화재 발생 사례가 빈번히 보고되며 안전성 확보에 대한 우려가 커지고 있다. 이에 따라 ESS 전압 데이터를 기반으로 한 이상 탐지 기술의 필요성이 대두되고 있으며, 현재는 오토인코더(Auto encoder)나 주성분 분석(Principal Component Anal ysis, PCA) 기반의 재구성 오차를 활용한 접근이 주로 사용되고 있다[1,2]. 하지만 이러한 방법들은 센서 노이즈나 통신 지연 등으로 인해 결측이 포함될 경우 탐지 성능이 급격히 저하되는 한계를 지닌다. 실제 ESS 운용 환경에서는 전체 데이터의 15% 이상 결측이 발생할 수 있으며[3], 이처럼 데이터가 불완전한 상황에서는 모델의 신뢰성을 충분히 확보하기 어려운 실정이다. 이에 기존 연구에서는 결측 구간을 선형, K-최근접 이웃(KNN) 기반 보간법 등을 활용해 결측 문제를 완화하고자 하였으나, 이들 기법은 데이터의 급변 구간에서 정보를 과도하게 평탄화하여 이상 패턴을 왜곡하는 문제점이 존재한다[4]. 이에 본 논문에서는 코사인 유사도 기반 가중치를 활용하여 KNN 보간값

과 외삽값을 가중 혼합하는 새로운 보간 기법을 제안한다. 제안 기법은 PCA 이상 탐지 알고리즘을 통해 기존 기법들과 성능을 비교하였으며, 이를 통해 결측이 존재하는 환경에서도 패턴 정보를 효과적으로 보존하고, 모델의 이상 탐지 성능이 향상됨을 확인하였다.

Ⅱ. 데이터 구성

본 연구에서는 1분 간격으로 측정된 ESS 전압 시계열 데이터를 사용하였다. 전체 데이터는 정상 데이터셋, 화재가 발생했던 비정상 데이터셋으로 구성되며, 각각 모델 학습에 사용할 학습 데이터, 벨리데이션(Validation) 데이터, 테스트 데이터로 분류하였다.
공정한 성능 비교를 위해 [그림 1]과 같이 데이터를 1,440분(1일) 단위로 재구성하였으며, 전압값이 3.3V 미만으로 측정된 지점은 이상값으로 간주하여 해당 지점을 결측값(NAN)으로 처리하였다. 또한, 하루 기준으로 결측 개수가 일정 임계치를 초과하는 경우, 해당 일자를 고장 발생일로 판단하고 분석 대상에서 제외하였다.

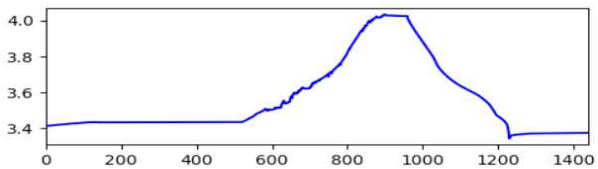

그림 1. 전처리된 전압 시계열 데이터

Ⅲ. 실험 설계

3.1 결측 데이터 삽입

각 보간 기법의 성능을 비교하기 위하여 전처리된 데이터에 인위적으로 결측치를 무작위 삽입하였다. 결측 삽입 비율은 5%, 10%, 15%, 20%로 설정하였으며, 동일한 결측 마스크를 원본 시계열 데이터에 일관되게 적용함으로써 보간 기법의 성능을 동일한 조건에서 비교할 수 있도록 하였다.

3.2 비교 보간 기법 설정

시계열 데이터는 센서의 물리적 특성, 결측 분포, 시간적 패턴 등에 따라 데이터 특성이 달라지며, 이에 따라 보간 기법의 성능 또한 상이하게 나타날 수 있다[5]. 따라서 선형성, 비선형성, 국소 유사성(Local Similarity) 등의 서로 다른 데이터 특성을 반영하는 보간 기법을 적용하는 것이 중요하며, 이를 통해 결측률 변화에 따른 이상 탐지 성능을 정량적으로 평가할 수 있다. 이에 본 연구에서는 실제 ESS 운용 환경에서 널리 활용되는 대표적인 기법인 선형 보간(Linear), 3차 스플라인 보간(Cubic Spline), K-최근접 이웃 기반 대체(KNN Imputer), 다항식 보간(Polynomial)을 비교 대상으로 선정하였다.

3.3 이상 탐지 실험 방법

각 보간 기법의 이상 탐지 성능을 비교하기 위해 [그림 3]의 주성분 분석(Principal Component Analysis, PCA) 기반 이상 탐지 알고리즘을 설계하여 실험을 수행하였다. 모델 학습에는 정상 데이터만을 사용하였으며, 데이터 분산의 95%를 설명하도록 주성분의 수를 설정하였다. 이후, 학습된 모델을 벨리데이션 데이터에 적용하여 각 클래스의 평균 절대 오차(Mean Absolute Error, MAE) 분포를 추출한 뒤, Youden's J 통계량을 활용하여 [그림 2]와 같이 정상, 비정상 클래스를 구분하는 최적 임계값(Threshold)을 설정하였다[6]. 설정된 임계값은 테스트 데이터의 MAE 분포에 적용되며, MAE 값이 임계값을 초과하는 구간을 이상(Anomaly)으로 판별하였다. 최종 이상 탐지 성능은 F1-score, PR-AUC, ROC-AUC의 세 가지 지표를 기준으로 평가하였으며, 실험의 통계적 신뢰도를 높이고 데이터 분할에 따른 편향을 최소화하기 위해 k-fold 교차 검증을 반복 수행하였다.

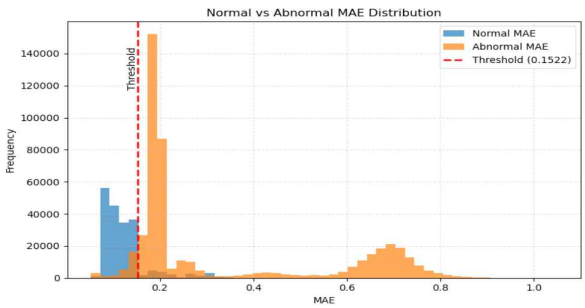

그림 2. MAE 분포에 따른 최적 임계값 설정

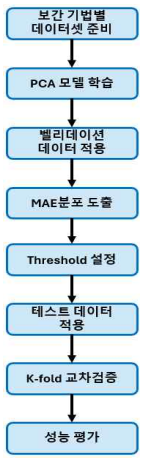

그림 3. PCA 기반 이상 탐지 알고리즘

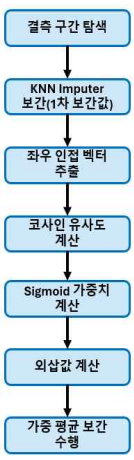

그림 4. 패턴 유사도 기반 가중 보간 기법 알고리즘

Ⅳ. 제안 기법

4.1 제안 기법 개요 및 설계 원리

이상 탐지 분야에서 널리 활용되는 전통적인 보간 기법들은 국소적 연속성에 기반하여 결측값을 추정하므로, 전역적인 패턴 구조를 고려하지 못하고 비정상 구간에서의 정확도가 저하되는 문제가 존재한다[7]. 이러한 문제점을 해결하기 위해 본 연구에서는 [그림 4]를 이용한 패턴 유사도 기반 가중 보간 기법을 새롭게 제안한다. 제안 기법의 핵심 설계 원리는 결측 지점을 기준으로 좌우 인접 구간의 시계열 패턴 유사도를 계산하고, 이를 기반으로 기존 보간값과 외삽값의 가중 평균을 통해 최종 보간값을 결정하는 것이다. 이러한 방식은 패턴이 안정적인 구간에서는 기존 보간 기법의 단순성과 효율성을 유지할 수 있으며, 급변 구간에서는 외삽 기반의 보정을 통해 정보 손실을 최소화할 수 있다는 장점을 가진다.

4.2 알고리즘 구성 및 수식 정의

본 절에서는 제안한 패턴 유사도 기반 가중 보간 기법의 세부 처리 과정과 수식을 구체적으로 서술한다.

① 초기 보간값 생성(1차 보간)

시계열 데이터 $X = \{x_1, x_2, ..., x_T\}$에 대해 결측값이 존재하는 구간에 KNN Imputer를 활용하여 각 결측 위치 t에 대한 초기 보간값 x_{knn}을 산출한다.

② 좌우 벡터 구성 및 패턴 추출

시계열 데이터의 결측 위치 t를 기준으로, 좌측에서 가장 가까운 유효값 a, b와 우측에서 가장 가까운 유효값 c, d를 선택하여 식 (1)과 같은 벡터를 구성한다.

$$\vec{v}_{\text{left}} = [a, b], \quad \vec{v}_{\text{right}} = [c, d] \tag{1}$$

③ 좌우 패턴 유사도 계산

결측 구간의 좌우 벡터 간 구조적 유사성을 정량화하기 위해 식 (2)를 통해 코사인 유사도(Cosine Similarity)를 계산한다. 유사도는 -1에서 1사이의 값을 가지며, 값이 1에 가까울수록 두 벡터의 방향성이 유사함을 의미하고, -1에 가까울수록 두 벡터의 방향성이 상반됨을 의미한다.

$$\cos(\theta) = \frac{\mathbf{v}_{\text{left}} \cdot \mathbf{v}_{\text{right}}}{\|\mathbf{v}_{\text{left}}\| \cdot \|\mathbf{v}_{\text{right}}\|} \tag{2}$$

④ 가중치 변환

계산된 코사인 유사도 값은 식 (3)의 시그모이드 함수를 통해 0에서 1사이의 보간 가중치 w로 변환된다. 이때 α는 유사도 값에 대한 민감도를 조절하는 계수로 작용한다.

$$w = \frac{1}{1 + e^{-\alpha \cdot \cos(\theta)}} \tag{3}$$

⑤ 외삽값 계산

결측치 기준 좌우 인접한 시계열 값의 기울기를 기반으로 식 (4)를 계산하고, 최종 외삽값은 식 (5)를 이용하여 산술 평균으로 계산한다.

$$x_{\text{ext}}^{\text{left}} = 2b - a, \quad x_{\text{ext}}^{\text{right}} = 2c - d \tag{4}$$

$$x_{\text{ext}} = \frac{1}{2}(x_{\text{ext}}^{\text{left}} + x_{\text{ext}}^{\text{right}}) \tag{5}$$

⑥ 최종 보간값 산출

가중치 w를 활용하여 식 (6)을 통해 최종 보간값을 계산한다. 식 (6)은 패턴 유사도가 높을수록 기존 보간값 x_{knn}에 더 많은 가중치를 부여하고, 유사도가 낮을수록 외삽값 x_{ext}에 더 많은 가중치를 부여하는 구조를 가진다.

$$\hat{x} = w \times x_{knn} + (1 - w) \times x_{ext} \tag{6}$$

Ⅴ. 성능 평가 및 분석

일반적으로 이상 탐지 문제는 극심한 클래스 불균형(Class Imbalance)을 동반하며[8], 이로 인해 단일 평가 지표만으로는 모델의 성능을 평가하기 어렵다. 특히 Recall이 낮을 경우, 실제 이상(Anomaly)을 놓치는 FN(False Negative)이 증가하게 되며, 이는 안전성이나 보안이 중요한 응용 분야에서는 심각한 위험 요소가 될 수 있다. 따라서 본 연구에서는 F1-score, PR-AUC, ROC-AUC의 세 가지 지표를 활용하여 모델의 성능을 종합적으로 평가하였다. 이 중 F1-score는 Precision과 Recall의 조화 평균으로, 클래스 불균형을 갖는 상황에서 실질적인 이상 탐지 성능을 가장 적절히 반영하는 핵심 지표로 간주된다. 실험 결과, 본 논문에서 제안한 보간 기법은 F1-score 기준 평균 0.651로, 비교 기법 중 가장 우수한 성능을 기록하였다[표 1]. 또한, AUC 지표에서도 높은 성능을 보여 결측률이 증가하더라도 이상 탐지 성능을 안정적으로 유지함을 확인할 수 있었다. 한편, KNN Imputer와 Cubic Spline 기법은 PR-AUC 및 ROC-AUC에서 상대적으

표 1. 결측 비율과 보간 기법에 따른 이상 탐지 성능 평가 결과

Missing Rate	Linear			Cubic Spline			Polynomial			KNN Imputer			Proposed		
	F1	PR	ROC	F1	PR	ROC	F1	PR	ROC	F1	PR	ROC	F1	PR	ROC
5%	0.491	0.694	0.654	0.333	0.711	0.700	0.531	0.631	0.570	0.647	0.672	0.672	0.652	0.687	0.674
10%	0.488	0.695	0.654	0.262	0.701	0.685	0.538	0.630	0.568	0.647	0.672	0.672	0.651	0.694	0.673
15%	0.488	0.695	0.654	0.262	0.698	0.680	0.525	0.632	0.573	0.647	0.671	0.672	0.649	0.687	0.675
20%	0.486	0.695	0.655	0.262	0.695	0.677	0.510	0.631	0.572	0.646	0.694	0.676	0.650	0.688	0.673
Avg	0.488	**0.695**	0.654	0.280	**0.701**	**0.686**	**0.526**	0.631	0.571	**0.647**	0.677	**0.673**	**0.651**	**0.689**	**0.674**

로 높은 성능을 보였으나, F1-score가 낮아 실제 이상을 자주 놓치는 경향이 나타났다. 이는 안정성 확보가 중요한 이상 탐지 문제에서 치명적인 한계로 적용할 수 있음을 시사한다.

Ⅵ. 결론 및 향후 연구 방향

본 논문에서는 기존 보간 기법의 한계를 극복하고 시계열 데이터의 결측 구간을 효과적으로 보완하기 위한 방법으로 패턴 유사도 기반 가중 보간 기법을 제안하였다. 제안 기법의 성능을 검증하기 위해 다양한 결측률에서 이상 탐지 실험을 진행하였으며, 실험 결과 제안 기법은 선형, 3차 스플라인, KNN Imputer, 다항 보간 등 기존 보간 기법들과 비교하여 F1-score, PR-AUC, ROC-AUC 지표 전반에서 우수한 성능을 기록하였다. 향후 연구에서는 제안된 보간 기법을 실시간 이상 탐지 시스템에 적용하여 실제 환경에서의 활용 가능성을 분석하고 다양한 조건에서의 일반화 성능을 검증할 예정이다.

참고문헌

[1] 박노진. (2023). 시계열 이상치 탐지: 오토인코딩을 활용한 사례 분석. 한국데이터정보과학회지, 34(4), 649-657. 10.7465/jkdi.2023.34.4.649

[2] 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.

[3] N. U. M. Khanum, H. Dahrouj, R. C. Bansal, and H. M. Tawfik, “An Overview of the Prospects and Challenges of Using Artificial Intelligence for Energy Management Systems in Microgrids,” arXiv preprint arXiv:2505.05498, May 2025. [Online].Available: https://arxiv.org/abs/2505.05498

[4] 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.

[5] M. Lepot, J.-B. Aubin, and F. H. L. R. Clemens, “Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment,” Water, vol. 9, no. 10, p. 796, Oct. 2017.

[6] Youden, William J. "Index for rating diagnostic tests." Cancer 3.1 (1950): 32-35.

[7] X. Tang, H. Yao, Y. Sun, C. Aggarwal, P. Mitra, and S. Wang, “Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values,” Proc. of the AAAI Conf. on Artificial Intelligence (AAAI), vol. 34, no. 4, pp. 5938 - 5945, Feb. 2020.

[8] T. Lee, L. K. Lee, and C. Kim, “Performance of Machine Learning Algorithms for Class-Imbalanced Process Fault Detection Problems,” in IEEE Transactions on Semiconductor Manufacturing, Vol. 29, No. 4, pp. 436 - 445, 2016.