MMATAD: 멀티모달 정렬 기반 다변량 시계열 이상탐지 모델

정현욱[1], 조영완[2], 홍인표[2], 김은지[2], 이효정[3], 김정은[2], 박상현[2†]
홍익대학교 소프트웨어융합학과[1], 연세대학교 컴퓨터과학과[2], 연세대학교 인공지능학과[3]
ppid0930@naver.com, {jyy1551, hip9863, kejh66, hyojoy, wjddms2216, sanghyun}@yonsei.ac.kr

MMATAD: A Multimodal Alignment Framework for Multivariate Time Series Anomaly Detection

Hyunwook Jeong, Youngwan Jo, Inpyo Hong, Eunji Kim, Hyojeong Lee, Jeongeun Kim, Sanghyun Park†
Dept. of Software & Communications, Hongik University
Dept. of Computer Science, Yonsei University
Dept. of Artificial Intelligence, Yonsei University

요 약

대규모 언어 모델 등장 이후, 시계열 예측 과업에 이를 적용하려는 연구가 활발히 진행중이다. 하지만, 대부분 제안된 방법은 단변량 시계열에 초점을 맞추고 있어 변수 간 상호작용을 충분히 반영하지 못하는 한계가 있다. 이에 본 논문에서는 단변량 시계열 예측 모델의 표현 학습 능력을 확장하여, 변수 간 관계를 통합적으로 고려한 다변량 시계열 이상탐지 모델인 A Multimodal Alignment Framework for Multivariate Times Series Anomaly Detection(MMATAD)을 제안한다. 제안한 모델은 Positive-Masked Matrix와 Grouped-Query Attention을 결합하여 변수 간 관계 강도와 의미를 동시에 반영하며 기존 방법 대비 향상된 성능을 보였다.

1. 서 론

최근 대규모 언어 모델(LLM)을 시계열 분석에 적용하려는 시도가 활발히 이루어지고 있다. 그러나 대부분의 LLM 기반 시계열 모델은 단변량 시계열 처리에 초점을 맞추고 있어서 다변량 시계열 데이터에서 변수 간 상호작용을 충분히 반영하지 못한다는 한계가 존재한다[1]. 실제로 센서, 금융, 의료 데이터와 같은 복잡한 시계열 데이터는 변수 간 강한 상관관계를 가지며 이를 적절히 고려하지 못할 경우 중요한 패턴을 인식하지 못하게 된다.

이러한 문제를 해결하기 위해 최근 연구에서는 변수 간 관계를 확률적 행렬 형태로 모델링하는 Masked Matrix 개념이 제안되었다[2]. Masked Matrix는 변수 간 연결의 존재 여부를 확률적으로 표현하고 이를 이진화 하여 관계가 있는 변수들 간의 상호작용을 학습하는 방식이다.

본 연구에서는 이러한 Masked Matrix 개념을 확장하여, 관계의 존재 여부 뿐만 아니라 강도(intensity)까지 반영할 수 있는 Positive-Masked Matrix(PMM)를 제안한다. PMM은 변수 간 관계의 존재는 Masked Matrix로 확립하되 각 관계의 강도를 확률적 가중치로 부여함으로써 보다 세밀한 관계 표현이 가능하도록 설계되었다. 이를 통해 변수 간 영향력 차이를 세밀하게 반영하며 다변량 시계열 내 복합적 상호작용을 내재적으로 표현할 수 있다. 또한 변수별 표현력을 강화하기 위해 Grouped-Query Attention(GQA)을 도입하였다[3]. GQA는 변수별 패치 표현을 언어적 표현 공간으로 정렬(alignment)함으로써, LLM이 시계열 내 변수 간 의미적 상호작용을 효과적으로 학습할 수 있다. 따라서 본 논문에서는 Positive-Masked Matrix와 GQA를 결합한 멀티모달 정렬 기반 다변량 시계열 이상탐지 모델(MMATAD)을 제안한다. 본 연구의 기여는 다음과 같다:

1. 변수 관계 강도를 반영한 **Positive Masked Matrix 기반의 다변량 시계열 표현 학습 구조**를 제안한다.
2. LLM의 언어 표현 정렬 특성을 활용한 **GQA 기반 다변량 시계열 이상탐지 모델**을 제시한다.
3. 다양한 공개 시계열 데이터셋에서 정량적 실험을 통해 제안 모델이 우수한 성능을 보였다.

2. 본 론

2.1 Overall Architecture

모델의 전체 구조는 그림 1과 같다. 입력은 $B \times L \times N$ 차원의 multivariate time series $X \in \mathbb{R}^{B \times L \times N}$ (배치 크기 B, 시계열 길이 L, 변수 개수 N)이며, RevIN을 통해 정규화된다[4]. 정규화된 입력은 패치 길이 p, stride s를 갖는 patching을 거쳐 $P = \left\lfloor \frac{L-p}{s} \right\rfloor + 1$개의 패치로 변환된다. 각 패치는 변수 축을

† 교신저자 : sanghyun@yonsei.ac.kr
*이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. RS-2025-02312833)과 국토교통부의 스마트시티혁신인재육성사업으로 지원을 받아 수행된 연구임.

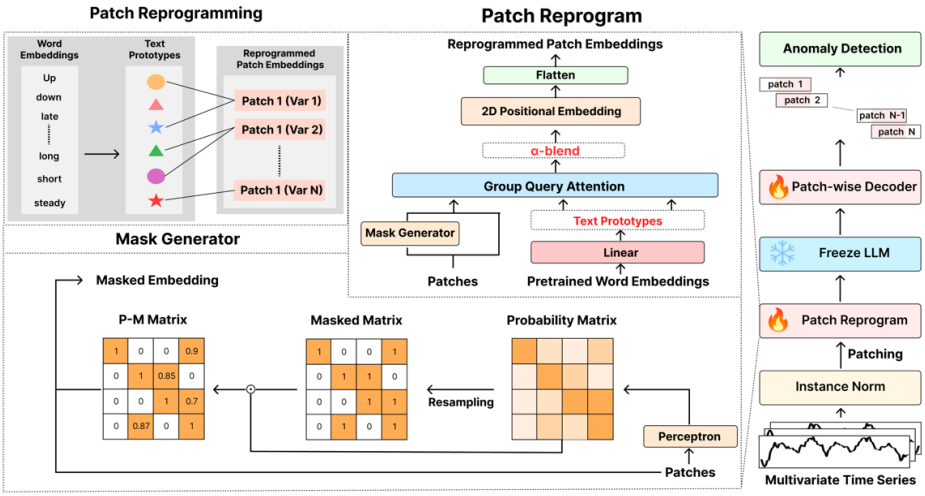

그림 1. MMATAD 모델 프레임워크

유지한 채 $X_p \in \mathbb{R}^{B\times P\times N\times p}$ 로 임베딩된다.

2.2 Mask Generator and Positive-Masked Matrix

각 패치 임베딩은 Perceptron 기반 Mask Generator를 통과하여 변수 간 연결의 활성 확률을 나타내는 확률 행렬 $P_{prob} \in \mathbb{R}^{N\times N}$ 을 산출한다. 이 행렬은 Gumbel-Softmax를 통해 이진화된 Masked Matrix $M \in \{0,1\}^{N\times N}$ 로 표현된다[5].

이후 확률 행렬과 마스크 행렬의 요소 곱을 통해 PMM를 생성한다:

$$\tilde{M} = M \odot P_{prob} \quad (1)$$

이를 통해 변수 간 연결의 존재 여부와 그 강도를 동시에 반영한다. 생성된 $\tilde{M}$를 이용하여 패치 표현 X_p에 마스킹을 적용하여 변수마다 마스킹된 패치 표현 $\tilde{X}_p \in R^p$을 얻는다. 각 패치 표현은 패치별 N개의 변수들에 대해 변수별 패치표현 $p_{patch} \in \mathbb{R}^p$ 과 마스킹된 표현 $p_{masked} \in \mathbb{R}^p$ 형태로 분리되고 GQA를 통해 언어 표현으로 정렬된다.

2.3 Group Query Attention with Soft Gating

변수별 패치표현 p_{patch}과 마스킹된 표현 p_{masked}은 Group Query Attention을 통해 LLM의 언어 표현으로 정렬된다. 이를 위해 LLM의 언어 임베딩에 선형 변환을 수행하여 압축된 표현 벡터 집합인 Text Prototypes를 구성한다. 이는 시계열-언어 alignment에 사용되는 GQA에 Key와 Value로 사용된다. GQA는 그룹별 동일한 Key/Value를 공유하여 메모리를 절감하면서 성능도 챙긴 연산방법이다[3]. 본 연구에서는 그룹을 변수 기준으로 구성하였으며 각 그룹은 이전에 구성한 두 표현의 Query를 사용하며 변수 간 상호작용을 학습한다. 각 그룹의 Attention 결과 Z는 마스킹 표현의 강도를 조절하는 학습 가능한 파라미터 α로 결합되어 패치별, 변수별로 표현된다:

$$Z = (1-\alpha)Z_{patch} + \alpha Z_{masked} \quad (2)$$

이렇게 얻은 결과표현 Z는 그림 1의 Patch Reprogramming에 묘사한 것처럼 Text Prototypes의 결합으로 표현된다. 이 과정으로 시계열 데이터를 LLM이 알고 있는 언어 표현으로 나타낸다.

2.4 Positional Encoding and LLM Projection

결과 표현 Z에 패치 위치와 변수 위치를 반영하고자 2D positional embedding을 추가하고 Flatten하여 Reprogrammed Patch Embeddings를 얻는다. 이렇게 얻은 Reprogrammed Patch Embeddings는 사전 학습된 LLM (LLAMA[6])을 통과하며, 이때 LLM 파라미터는 freeze하여 추론을 수행한다.

2.5 Patch-wise Decoder

LLM 출력은 patch-wise decoder를 통해 원래 시계열 패치 단위로 복원된다. 겹치는 구간은 overlap-average 방식으로 병합되어 최종 재구성 시계열 $\hat{X}$를 얻는다.

2.6 Training Objective

학습은 MSE 기반 재구성 손실과 Mask Generator 학습을 위한 InfoNCE Loss[7] 기반의 Mask 손실로 학습된다. 수식 3, 4는 Loss이다:

$$L_{rec} = \|X - \hat{X}\|_2^2 \quad (3)$$

$$L_{Mask} = -\frac{1}{N}\sum_{k=1}^{N} \log\left(\frac{\sum_{m=1}^{N} M_{k,m}\cdot \exp\left(\frac{cos\left(p_{masked}^{k,i}, p_{patch}^{m,i}\right)}{\tau}\right)}{\sum_{l=1}^{N}\exp\left(\frac{\cos\left(p_{masked}^{k,i}, p_{patch}^{l,i}\right)}{\tau}\right)}\right),$$

$$p_{masked}^{k,j}, p_{patch}^{m,j} \in \mathbb{R}^d, j = \{1, \dots N\} \quad (4)$$

최종 학습 손실은 두 손실의 가중합으로 가중치 λ를 사용하여 다음과 같이 구성한다:

$$L_{total} = L_{rec} + \lambda L_{Mask} \quad (5)$$

2.7 Anomaly Scoring

학습된 모델의 이상 탐지는 입력 시계열과 복원된 시계열 간의 평균 제곱오차(MSE)를 이상 점수로 사용하며 오차 값이 클수록 해당 시점이 이상일 가능성이 높다고 판단한다. 이상 판단을 위한 임계값은 학습 데이터의 재구성 오차 분포를 기반으로 하는 백분위수(percentile) 탐색 방법을 통해 결정하였다.

3. 실험 및 결과

3.1 실험 환경

본 논문에서는 모델의 성능을 평가하기 위해 다변량 이상탐지에서 널리 사용되는 데이터셋인 MSL과 GECCO를 사용하였다. 평가 지표는 구간 단위 평가지표인 Affiliated-F1 score (Affi-F1 score), 시점 단위 평가지표인 ROC-AUC와 F1 score를 사용하였다[8]. 성능 비교 모델은 단변량 시계열 이상탐지에서 우수한 성능을 보인 CATCH[2], TimesNet (TsNet)[9]을 사용하였다. CATCH는 주파수 기반 패치 재구성과 채널 마스크 어텐션을 결합하여 다변량 시계열 이상탐지를 수행하는 모델이다. TsNet은 시계열을 여러 주기의 2차원 패치로 변환해 주기적 패턴을 학습하는 트랜스포머 기반 모델이다. 모델 학습은 배치 크기 16으로 설정하고 총 10 epoch 동안 수행하였다. 입력 시퀀스 길이는 96으로 설정하였으며 시퀀스는 길이 16의 패치 단위로 분할하고 stride는 8로 설정하였다. 마스크 손실에 대한 가중치는 0.005로 고정하여 실험하였다.

3.2 실험 결과 및 분석

표 1. 실험 결과. 1등은 굵은 글씨, 2등은 밑줄로 표시

Datasets	Metrics/ Models	MMATAD (Ours)	CATCH	TsNet
MSL	Affi-F1 score	**0.7217**	<u>0.7148</u>	0.6218
	ROC-AUC	**0.6288**	<u>0.6269</u>	0.5922
	F1 Score	**0.1652**	<u>0.1220</u>	0.1105
GECCO	Affi-F1 score	<u>0.8116</u>	**0.8157**	0.8070
	ROC-AUC	0.7456	<u>0.9073</u>	**0.9082**
	F1 Score	0.1410	<u>0.2221</u>	**0.2253**

표 1은 제안한 모델과 비교 모델(CATCH, TsNet)의 실험 결과를 보여준다. 제안 모델은 MSL 데이터셋에서 모든 평가지표에서 비교 모델보다 우수한 성능을 보였다. 반면 GECCO 데이터셋에서는 시점 단위 평가지표(ROC-AUC, F1 Score)가 다소 낮았으나, 구간 단위 평가지표(Affinity-F1)에서는 안정적인 성능을 유지하였다.

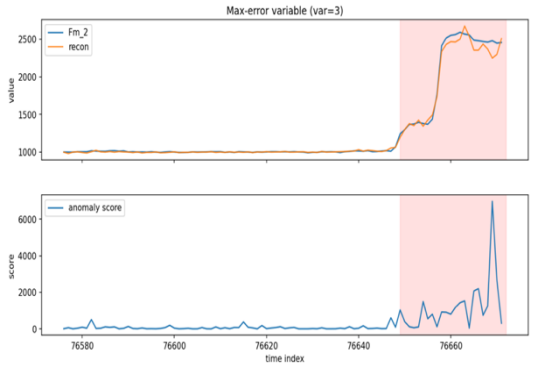

그림 2. GECCO 샘플 데이터 시각화

그림 2는 GECCO 평가 데이터셋의 일부 시점에 대한 이상탐지 결과를 시각화한 것이다. 제안한 모델은 다변량 시계열 전체에 대해 이상 점수를 계산하지만, 시각적 직관성을 위해 이상 구간 내에서 재구성 오차가 가장 크게 나타난 단일 변수(Fm_2)를 대표로 시각화 하였다. 붉은 음영 영역은 실제 이상 구간(ground truth)으로 해당 구간에서 Anomaly Score의 상승을 확인할 수 있다.

4. 결 론

본 연구는 PMM와 GQA을 결합한 다변량 시계열 이상탐지 모델 MMATAD를 제안하였다. 제안한 모델은 변수 간 관계의 강도를 반영하여 MSL 데이터셋에서 우수한 성능을 보였다. 반면 GECCO 데이터셋에서는 시점 단위 평가지표(ROC-AUC, F1 Score)에서 다소 낮은 성능을 보였는데 이는 시계열-언어 정렬 과정에서 이상 패턴까지 정상적으로 복원하려는 경향이 나타났기 때문이다.

향후 연구에서는 변수 간의 학습된 정상 패턴 차이를 유지하면서 시계열-언어 정렬을 수행하는 학습 구조를 통해 이러한 한계를 개선하고자 한다.

참고문헌

[1] Jin, Ming, et al. "Time-llm: Time series forecasting by reprogramming large language models." International Conference on Learning Representations, 2024.

[2] Wu, Xingjian, et al. "Catch: Channel-aware multivariate time series anomaly detection via frequency patching." International Conference on Learning Representations, *2025.*

[3] Ainslie, Joshua, et al. "Gqa: Training generalized multi-query transformer models from multi-head checkpoints." Empirical Methods in Natural Language Processing, 2023.

[4] Kim, Taesung, et al. "Reversible instance normalization for accurate time-series forecasting against distribution shift." *International conference on learning representations*, 2021.

[5] Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax." *arXiv preprint arXiv:1611.01144*, 2016.

[6] Dubey, Abhimanyu, et al. "The llama 3 herd of models." *arXiv e-prints* (2024): arXiv-2407, 2024.

[7] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748*, 2018.

[8] Alexis Huet, Jose Manuel Navarro, and Dario Rossi. Local evaluation of time series anomaly detection algorithms. In Proceedings of the ACM SIGKDD international conference on knowledge discovery & data mining, pp. 635–645, 2022.

[9] Wu, Haixu, et al. "Timesnet: Temporal 2d-variation modeling for general time series analysis." *arXiv preprint arXiv:2210.02186*, 2022.