

비군집화 임베딩의 군집화 가능성 탐구: 다중 도메인 다변량 시계열 데이터 기반 임베딩 모델 비교 연구*

서형철^{O1} 유지훈¹ 이상민¹ 진동현¹ 최진혁¹ 이지수¹ 김천구¹ Vaishnavi Vats¹ 서영균¹

¹경북대학교 컴퓨터학부

{wjdqh6544, knu12370, lsmin3388, loonaticvibe, daniel040607, jisulee74, kcg1009, vatsvaishnavi28, yksuh}@knu.ac.kr

Exploring the Clusterability of Unclustered Embeddings: A Comparative Study of Embedding Models Based on Multi-Domain Multivariate Time Series Data

HyeongCheol Seo^{O1} JiHun Yu¹ SangMin Lee¹ DongHyeon Jin¹ JinHyuk Choi¹
JiSu Lee¹ CheonGu Kim¹ Vaishnavi Vats¹ YoungKyoong Suh¹

¹School of Computer Science and Engineering, Kyungpook National University

요약

최근 시계열 데이터의 범용 임베딩 학습에 관한 연구가 증가하고 있으나, 군집화와 같은 비지도학습 테스크에 대한 평가는 부족하다. 본 연구는 최신 시계열 임베딩 방법론 5종(TS2Vec, PatchTST, MERIT, TimeCMA, TimeKD)의 군집화 성능을 체계적으로 분석하고, 학습 패러다임에 따른 특성을 비교한다. UEA 아카이브의 데이터셋 4종에 대하여, TS2Vec은 평균 RI 0.768 및 평균 NMI 0.429를 기록하여 선형 및 비선형 군집화 구조 모두에서 가장 우수한 성능을 보였다. 반면, 지식 종류와 교차 모달리티 정렬 기반을 기반으로 한 TimeKD와 TimeCMA는 비교적 높은 RI 수치를 보였으나, NMI는 0에 근접하여 불균형한 군집화 결과를 보였다. PatchTST와 MERIT은 Spectral 군집화에서 비선형 다양체 구조 포착 능력을 입증하였으며, 대조학습 및 자기지도학습 기반 방법론이 레이블 없는 시계열의 내재적 구조 학습에 더 효과적임을 확인하였다. 본 연구는 시계열 임베딩의 군집화 성능을 체계적으로 측정하고, 데이터 특성에 따른 방법론 선택 지침을 제시하여 실무 응용의 기반을 마련하였다.

1. 서 론

의료 모니터링, 산업 공정관리 등의 다양한 도메인에서 다변량 시계열(Multivariate Time Series, MTS) 데이터가 폭발적으로 생성되고 있으며, 각 도메인은 고유한 데이터 구조와 스케일 특성을 가진다. 전통적으로는 도메인별로 독립적인 임베딩 모델을 학습하고 관리하여야 하며, 이는 모델 개발 및 유지보수 비용의 기하급수적인 증가를 초래한다. 하지만 최근 빠르게 발전하고 있는 대규모 언어 모델(Large Language Model, LLM)과 자기지도학습 기반 방법론을 활용하면, 여러 도메인에 적용할 수 있는 공통 임베딩을 얻을 수 있다. 이를 통해 분류, 예측, 이상 탐지, 군집화를 포함하는 다양한 다운스트림 테스크를 동시에 지원하여 효율성을 크게 개선할 수 있다.

그러나 LLM 및 자기지도학습을 기반으로 군집화와 같은 완전 비지도학습 영역에 공통 임베딩을 적용한 연구는 드물다. 이는 레이블이 없는 데이터에 대한 표현의 유용성을 평가하기 어렵고, 공통 임베딩 접근법이 이러한 비지도학습에서 보이는 특성에 대한 실증적 근거가 부족하기 때문이다. 특히 군집화는 완전 비지도학습을 대표하는 핵심 테스크로서, 레이블이 없는 데이터 내의 패턴을 발견하고 구조를 해석하기 위한 핵심 도구이다. 따라서 공통 임베딩의 비지도학습 성능을 이해하려면 군집화 테스크에서의 특성을 면밀히 파악할 필요가 있다.

이에 본 연구에서는 기존 임베딩 접근법을 LLM 기반 및 Non-LLM 기반 방법론으로 구분하고, 각 접근법이 군집화 테스크에서 보이는 특성을 체계적으로 분석하고자 한다. TS2Vec[1], TimeKD[2], PatchTST[3], TimeCMA[4], MERIT[5]의 다섯 가지 모델을 사용하여, UEA 시계열

아카이브[6]의 4개의 데이터셋(BasicMotions, Epilepsy, HandMovement Direction, Libras)에 대한 군집화 성능을 측정한다. 통일된 실험 프레임워크에서 K-Means[7] 및 Spectral 군집화[8] 알고리즘을 적용하고, RI(Rand Index)[9] 및 NMI(Normalized Mutual Information)[10] 지표를 통해 군집화 성능을 측정하여 각 모델이 생성한 임베딩의 특성을 비교한다. 이를 근거로 LLM 및 Non-LLM 기반 접근법의 군집화 특성을 파악하고, 특정 응용 시나리오에 적합한 접근법을 선택하기 위한 실증적인 기반을 제공하고자 한다.

본 논문의 공헌은 다음 세 가지로 요약된다.

- 본 논문은 최신 시계열 공통 임베딩 기법 5종의 비지도 군집화 성능을 체계적으로 측정하고, K-Means 및 Spectral 군집화 알고리즘의 성능 차이를 실험적으로 규명하였다.
- LLM 기반 방법론에서의 RI-NMI 불균형 현상을 발견하고, Non-LLM 기반 방법론과의 군집화 성능 특성 차이를 정량적으로 비교 및 분석하였다.
- 데이터 특성(주기성, 비선형성, 궤적 유사성)과 군집화 알고리즘 선택에 따른 임베딩 방법론별 적합성을 실증적으로 제시하여, 실무 응용을 위한 방법론 선택 가이드라인을 도출하였다.

2. 관련 연구

본 장에서는 다섯 가지의 시계열 표현학습 방법론을 비교 및 분석한다. 이들은 크게 LLM 기반 방법론과 Non-LLM 기반 방법론으로 구분된다.

2.1 LLM 기반 방법론

LLM 기반 방법론은 사전 학습된 LLM의 지식을 시계열 표현 학습에 접목한다. MERIT은 LLM을 의사결정 도구로 활용하여 고품질 증강

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (2021-0-01082)

표 1. 비교 분석 대상 시계열 표현 학습 모델의 특성

Methods	Base Frameworks	Learning Paradigms	LLM Roles	Key Techniques	Pre-trained LLM Required
PatchTST	Transformer	Self-supervised	N/A	Patching + Channel Independence	×
TS2Vec	Dilated CNN	Contrastive	N/A	Hierarchical Contrast (Instance + Temporal)	×
MERIT	LLM	Self-supervised	Augmentation Designer	Multi-agent System for View Generation	✓
TimeCMA	LLM	Cross-modal	Alignment Guide	Time Series ↔ Text Prompt Alignment	✓
TimeKD	LLM → Transformer	Knowledge Distillation	Teacher Model	LLM Teacher → Lightweight Student	✓

표 2. 군집화 성능 평가에 사용된 데이터셋

Datasets	# of Samples	Time Series Length	# of Channels	# of Classes	Domains
BasicMotions	80	100	6	4	Motion Recognition
Epilepsy	552	178	3	4	Healthcare
Libras	360	45	2	15	Gesture Recognition
HandMovement Direction	234	400	10	4	Motion Tracking

뷰를 생성한다. TimeCMA는 교차 모드 정렬을 통해 시계열 임베딩과 프롬프트 임베딩의 뒤틀림 문제를 해결한다. TimeKD는 LLM을 교사 모델로 활용하며, 경량 학생 모델로 지식을 중류한다.

2.2 Non-LLM 기반 방법론

Non-LLM 기반 방법론은 시계열 데이터 특화 아키텍처와 자기지도 학습 기법을 활용하여 임베딩 표현을 학습한다. PatchTST는 패칭 및 채널 독립성을 통해 계산 효율성과 예측 성능을 개선하였으며, TS2Vec은 팀스탬프 수준 표현과 계층적 대조학습을 통해 다양한 다운스트리밍 테스크에 범용적으로 활용할 수 있는 표현을 학습한다.

3. 실험 설계 및 결과 분석

3.1. 실험 설정

본 연구의 실험은 Intel Core i7-9700K 3.6GHz CPU, 64GB DDR4 메모리, NVIDIA GeForce RTX3090 24GB GPU, Ubuntu 20.04, CUDA 11.8을 사용한 환경에서 진행하였다. 각 모델의 학습은 공식 구현체를 기반으로 하며, 하이퍼파라미터 및 데이터 전처리를 위한 파라미터는 기존 논문의 권장 설정을 따랐다. 또한, 무작위 시드를 사용하여 재현성을 보장하였으며, 공정한 비교를 위하여 모든 모델의 최종 임베딩 벡터 차원은 320으로 고정하였다.

임베딩의 범용성을 다양으로 평가하기 위하여, UEA 시계열 아카이브에서 서로 다른 특성을 가진 데이터셋 4종을 선정하였다(표 2). BasicMotions는 주기적 동작 패턴, Epilepsy는 복잡한 비선형 의료 신호(EEG), Libras는 클래스별 궤적 유사성이 높은 2D 수화 동작, HandMovementDirection은 상, 하, 좌, 우의 네 방향 움직임에 대한 3D 좌표 데이터이다. 데이터셋을 모델의 입력 구조에 맞추기 위해, 길이가 M인 시계열 샘플 N개로 구성된 4가지 데이터셋을 시간 축을 기준으로 연결하여 ((NxM), 1) 형태의 연속적인 다변량 시퀀스로 변환하였다. 이후 평균 0, 표준편차 1로 정규화하여 스케일 차이를 보정하였고, 변환된 시퀀스를 고정된 윈도우 길이(Length=96)로 분할하였다. 또한, 차원 축소는 별도로 진행하지 않았다.

5가지 모델을 사용하여, 각 데이터셋에 대한 임베딩을 추출하였다. 이후 추출된 각 임베딩에 대해 K-Means 군집화 및 Spectral 군집화를 적용하여, 임베딩 공간의 구조를 평가하였다. K-Means 군집화는 거리

기반으로 볼록한 군집을 형성하는 알고리즘으로, 선형적으로 분리할 수 있는 구조를 평가한다. 반면 Spectral 군집화는 그래프를 기반으로 비선형 및 비볼록 군집을 탐지하여, 상호 보완적인 두 알고리즘으로 군집화를 진행하였다. 군집 수는 데이터셋이 가진 클래스 수로 설정하였으며, 군집화 알고리즘은 scikit-learn[11]으로 구현하였다. 또한, Spectral 군집화를 위한 유사도 구조는 k-최근접 이웃 그래프 방식을 채택하였으며, 하이퍼파라미터(이웃의 수) k는 15로 설정하였다.

군집화 성능 평가 지표는 RI 및 NMI를 사용하였다. RI는 샘플 쌍의 군집 할당이 일치하는 정도를 의미하며, NMI는 군집이 실제 클래스 정보를 반영하는 정도를 의미한다. 두 지표 모두 0에서 1 사이의 값을 가지며, 1에 가까울수록 성능이 우수하다. 또한, 정답 레이블은 최종 평가 단계에서만 사용하여, 학습 과정이 정답에 의해 영향받지 않도록 하였다.

3.2. 실험 결과 및 분석

표 3은 각 데이터셋에 대한 5가지 모델의 군집화 성능을 나타낸다. TS2Vec이 K-Means 군집화에서 평균 RI 0.768 및 평균 NMI 0.429, Spectral 군집화에서 평균 RI 0.808 및 평균 NMI 0.568을 기록하여 두 알고리즘 모두에서 가장 우수한 성능을 달성하였다. 이는 계층적 대조학습이 시간 일관성을 효과적으로 포착하여, 선형 및 비선형 군집화 구조 모두에 적합한 임베딩을 생성하였기 때문인 것으로 해석할 수 있다. PatchTST는 TS2Vec에 근접한 성능을 보였으며, Spectral 군집화에서 K-Means 군집화 대비 평균 5.52%p 향상된 RI 점수를 기록하여 비선형 구조 포착 능력을 입증하였다.

TimeKD와 TimeCMA는 비교적 높은 RI 수치를 보였으나, NMI가 0.05 미만으로 나타나 두 지표가 불균형한 모습을 보였다. 이는 군집이 형성되었으나 실제 클래스 정보를 반영하지 못함을 의미하며, 지식 종류나 교차 모달리티 정렬 방식이 시계열의 시간적 패턴과 국소 구조를 충분히 보존하지 못한 데 기인한다. 반면, MERIT은 Spectral 군집화에서 K-Means 군집화 대비 평균 25.34%p 향상된 RI 성능을 보였으며, 이는 다중 에이전트 기반 증강이 국소적 비선형 다양체 구조 보존에 기여한 결과로 판단된다. 이러한 결과는 비지도 군집화에서 대조학습과 자기지도학습이, 지식 종류나 모달리티 정렬보다 시계열의 내재적 구조를 더 효과적으로 학습함을 시사한다.

데이터셋별 성능 분석 결과, 데이터 특성에 따른 차이가 명확하게 나타났다. 주기적 패턴을 가진 BasicMotions은 모든 방법론이 0.70 이상의 RI를 기록한 반면, 네 방향 패턴 데이터인 HandMovement Direction은 모든 모델에서 0.64 미만의 평균 RI로, 저조한 성능을 보였다. 이는 현재의 임베딩 방법론으로 동작 방향성과 같은 추상적 패턴을 포착하기 어려움을 시사한다. 또한 Libras와 Epilepsy에서는 TS2Vec과 PatchTST가 상대적으로 높은 성능을 보였으며, 이는 대조학습과 패칭 전략이 복잡한 시공간 패턴 학습에 효과적임을 시사한다.

표 3. K-Means 군집화 및 Spectral 군집화를 활용한 성능 평가 결과

Datasets	Metrics	Methods									
		TS2Vec		TimeKD		PatchTST		TimeCMA		MERIT	
		K-Means	Spectral	K-Means	Spectral	K-Means	Spectral	K-Means	Spectral	K-Means	Spectral
BasicMotions	RI	0.8540	1.0000	0.5130	0.4380	0.8397	1.0000	0.5497	0.5058	0.7506	0.7361
	NMI	0.8200	1.0000	0.0045	0.0099	0.8000	1.0000	0.1598	0.0554	0.5025	0.4454
Epilepsy	RI	0.7060	0.7334	0.5804	0.5962	0.7004	0.6897	0.4985	0.4990	0.1064	0.6611
	NMI	0.3120	0.5658	0.0015	0.0013	0.2742	0.2377	0.0033	0.0015	0.1411	0.1447
HandMovement Direction	RI	0.6090	0.6024	0.2665	0.3864	0.5987	0.5983	0.6350	0.6350	0.5877	0.6175
	NMI	0.0440	0.0300	0.0000	0.0000	0.0334	0.0276	0.0002	0.0002	0.0064	0.0056
Libras	RI	0.9040	0.8960	0.4998	0.4994	0.8701	0.8867	0.6114	0.6118	0.8673	0.8831
	NMI	0.5420	0.6770	0.0001	0.0080	0.3060	0.3153	0.0026	0.0038	0.1974	0.2169
Average	RI	0.7683	0.8080	0.4649	0.4800	0.7522	0.7937	0.5737	0.5629	0.5780	0.7245
	NMI	0.4295	0.5682	0.0015	0.0048	0.3534	0.3952	0.0415	0.0152	0.2119	0.2032

4. 결론 및 향후 연구

본 연구는 다섯 가지 최신 시계열 임베딩 방법론의 비지도 군집화 성능을 체계적으로 분석하였다. TS2Vec은 선형 및 비선형 군집화 구조 모두에서 가장 우수한 성능을 보였으며, PatchTST와 MERIT은 Spectral 군집화를 통해 비선형 다양체 구조를 효과적으로 포착함을 확인하였다. 반면 TimeKD 및 TimeCMA는 높은 RI에도 불구하고 0에 가까운 NMI를 기록하였으며, 이는 지식 종류 및 교차 모달리티 정렬 기반 학습이 비지도 시계열의 내재적 구조를 충분히 포착하지 못함을 보여준다. 이러한 결과를 바탕으로, 주기적 패턴 시계열에는 TS2Vec을, 복잡한 비선형 구조의 의료 및 센서 데이터에는 Spectral 군집화와 결합한 PatchTST 또는 MERIT를 적용하는 실무 지침을 제시한다. 또한, 순수 비지도 환경에서는 대조학습 및 자기지도학습 기반 방법이 적합한 것으로 판단된다.

본 연구는 기존 시계열 임베딩 방법론의 군집화 성능 특성을 분석하고, 각 방법론이 단일 도메인에서 학습된 후 다른 도메인에 적용될 때의 특성을 평가하는 데 초점을 맞추었다. 하지만, 다양한 임베딩 벡터 차원(128, 256 등)에 따른 성능 변화를 면밀하게 분석하지 못하였다는 한계가 있다. 또한, RI와 NMI 뿐 아니라 ARI, AMI, Silhouette Score 등과 같은 다양한 지표를 통하여 모델의 강건성을 입증할 필요가 있다. 향후 연구에서는 다양한 임베딩 차원에 따른 성능 변화와 확장된 평가지표를 활용한 모델의 일반화 성능을 심도 있게 분석하고자 한다. 나아가, 군집화뿐만 아니라 이상 탐지 및 분류 등 다양한 태스크로 평가 범위를 확장하여, 범용 임베딩이 도메인별 개별 모델 개발 비용을 실질적으로 절감할 수 있음을 실증할 예정이다.

참고문헌

- [1] Yue, Zhihan, et al. "Ts2vec: Towards universal representation of time series." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 8. 2022.
- [2] Liu, Chenxi, et al. "Efficient multivariate time series forecasting via calibrated language models with privileged knowledge distillation." arXiv preprint arXiv:2505.02138 (2025).
- [3] Huang, Xinyu, Jun Tang, and Yongming Shen. "Long time series of ocean wave prediction based on PatchTST model." Ocean Engineering 301 (2024): 117572.
- [4] Liu, Chenxi, et al. "Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 18. 2025.
- [5] Zhou, Shu, et al. "MERIT: Multi-Agent Collaboration for Unsupervised Time Series Representation Learning." Findings of the Association for Computational Linguistics: ACL 2025. 2025.
- [6] Bagnall, Anthony, et al. "The UEA multivariate time series classification archive, 2018." arXiv preprint arXiv:1811.00075 (2018).
- [7] McQueen, James B. "Some methods of classification and analysis of multivariate observations." Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.. 1967.
- [8] Ng, Andrew, Michael Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." Advances in neural information processing systems 14 (2001).
- [9] Rand, William M. "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical association 66.336 (1971): 846–850.
- [10] McDaid, Aaron F., Derek Greene, and Neil Hurley. "Normalized mutual information to evaluate overlapping community finding algorithms." arXiv preprint arXiv:1110.2515 (2011).
- [11] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825–2830.