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

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

요 약

시계열 데이터의 범용 임베딩 학습에 관한 연구가 증가하고 있다. 하지만 대부분의 연구는 분류나 예측 등 지도 학습 태스크에 집중하고 있으며 군집화와 같은 비지도 학습 태스크에 대한 평가는 부족하다. 본 연구는 최신 시계열 임베딩 방법론의 군집화 성능을 체계적으로 분석하여, LLM 기반(MERIT, TimeCMA, TimeKD) 및 Non-LLM 기반(PatchTST, TS2Vec) 접근법의 특성을 비교한다. UEA 아카이브의 4개 벤치마크 데이터셋에 K-Means 군집화 및 Spectral 군집화를 적용한 결과, Non-LLM 방법론인 TS2Vec은 0.768의 평균 RI 및 0.430의 평균 NMI를 기록하여 가장 우수한 성능을 보였다. 반면 LLM 기반 방법론 중 TimeKD 및 TimeCMA의 경우, RI 수치는 일정하였으나 NMI 수치가 0에 가까워 불균형을 나타냈으며, 오직 MERIT 모델만 균형 잡힌 성능을 유지하였다. PatchTST와 MERIT는 Spectral 군집화에서 K-Means 군집화 대비 높은 성능치를 기록함으로써 비선형 구조 포착 능력을 입증하였으며, TS2Vec의 경우 선형 분리 구조에 최적화되어 있음을 확인하였다. 본 연구는 시계열 임베딩의 군집화 성능을 최초로 측정하여 연구 공백을 해소하고, LLM 기반 방법론이 비지도 학습 태스크에 근본적인 한계가 있음을 실증적으로 규명하였다.

Abstract

Research on universal embedding learning for time-series data is increasing. However, most studies focus on supervised learning tasks such as classification or prediction, and evaluations for unsupervised learning tasks like clustering are lacking. This study systematically analyzes the clustering performance of state-of-the-art time series embedding methodologies, comparing the characteristics of LLM-based (MERIT, TimeCMA, TimeKD) and non-LLM-based (PatchTST, TS2Vec) approaches. Applying K-Means clustering and Spectral clustering to four benchmark datasets from the UEA Archive, the non-LLM methodology TS2Vec achieved the best performance with an average RI of 0.768 and an average NMI of 0.430. Conversely, among LLM-based methodologies, TimeKD and TimeCMA exhibited consistent RI values but imbalanced NMI values close to zero, while only the MERIT model maintained balanced performance. PatchTST and MERIT demonstrated superior performance compared to K-Means clustering in spectral clustering, proving their ability to capture nonlinear structures, while TS2Vec was confirmed to be optimized for linearly separable structures. This study fills a research gap by being the first to measure the clustering performance of time-series embeddings and empirically demonstrates that LLM-based methodologies have fundamental limitations in unsupervised learning tasks.

Keywords : Multivariate Time Series, Time Series Embedding, Clustering, Unsupervised Learning, Large Language Models

Ⅰ. 서 론

HVAC 시스템, 의료 모니터링, 산업 공정 관리 등 다양한 도메인에서 다변량 시계열(Multivariate Time Series, MTS) 데이터가 폭발적으로 증가하고 있으며, 각 도메인은 고유한 데이터 구조와 스케일 특성을 가진다. 전통적으로는 도메인별로 독립적인 임베딩 모델을 학습하고 관리해야 하였으나, 이는 모델 개발 및 유지보수 비용의 기하급수적인 증가를 초래하였다. 하지만 최근 빠르게 발전하고 있는 대규모 언어 모델(Large Language Model, LLM)과 자기 지도 학습 기반 방법론을 활용하면, 여러 도메인에 적용할 수 있는 공통 임베딩을 얻을 수 있다. 이를 통해 분류, 예측, 이상 탐지, 군집화를 포함하는 다양한 다운스트림 태스크를 동시에 지원함으로써 효율성을 크게 개선할 수 있다.

그러나 LLM 및 자기지도학습을 기반으로, 군집화와 같은 완전 비지도 학습 영역에 공통 임베딩을 적용한 연구는 현저히 부족하다. 이는 레이블이 없는 데이터에서 표현의 유용성을 평가하는 방법에 대한 어려움 때문이며, 최신 공통 임베딩 접근법들이 군집화에서 어떠한 특성을

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 1. 비교 분석 대상 시계열 표현 학습 모델의 특성

보이는지에 대한 실증적 근거가 부족한 상황이다. 군집화는 대규모 레이블이 없는 데이터 환경에서 데이터의 패턴을 발견하고 구조를 해석하기 위한 핵심 도구이므로, 공통 임베딩이 진정한 범용성을 달성하기 위해서는 군집화 태스크에서의 특성을 명확히 이해할 필요가 있다.

이에 본 연구에서는 기존의 임베딩 접근법을 Non-LLM 및 LLM 기반 방법론으로 구분하고, 각 접근법이 군집화 태스크에서 보이는 성능 특성을 체계적으로 분석하고자 한다. 우리는 TS2Vec[1], TimeKD[2], PatchTST[3], TimeCMA[4], MERIT[5]의 다섯 가지 최신 모델을 대상으로, UEA 시계열 아카이브에서 선정한 4개의 데이터셋(BasicMotions, Epilepsy, HandMovementDirection, Libras)에 대한 군집화 성능을 측정하고자 한다. 이를 위해 통일된 실험 프레임워크에서 K-Means 군집화[6] 및 Spectral 군집화[7] 알고리즘을 적용한다. 이후 RI(Rand Index)[8]와 NMI(Normalized Mutual Information)[9] 지표를 통해 군집화 성능을 측정하고, 각 모델이 생성하는 임베딩의 특성을 비교한다. 이를 통해 LLM 기반 접근법과 Non-LLM 기반 접근법 각각의 군집화 성능 경향성을 파악하고, 특정 응용 시나리오에 적합한 접근법을 선택하기 위한 실증적 기반을 제공하고자 한다.

본 논문의 공헌은 다음 세 가지로 요약된다.

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

II. 관련 연구

본 연구에서는 표 1에 제시된 다섯 가지 최신 시계열 표현 학습 방법론을 비교 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. MERIT는 다중 LLM 에이전트 시스템을 통해, LLM을 의사결정 도구로 활용하여 고품질 증강 뷰를 생성한다. TimeCMA는 교차 모드 정렬을 통해 시계열 임베딩과 프롬프트 임베딩의 뒤얽힘 문제를 해결한다. TimeKD는 LLM을 교사 모델로 활용하며, 경량의 학생 모델로 지식을 증류하는 방법을 제안한다.

2.2 Non-LLM 기반 방법론

Non-LLM 기반 방법론은 시계열 데이터에 특화된 아키텍처와 자기지도학습 기법을 활용하여 임베딩 표현을 학습한다. PatchTST는 패칭과 채널 독립성을 통해 계산 효율성과 예측 성능을 개선하였으며, TS2Vec은 타임스탬프 수준 표현과 계층적 대조 학습을 통해 다양한 다운스트림 과업에 범용적으로 활용할 수 있는 표현을 학습한다.

III. 실험 및 결과 분석

3.1 실험 환경

실험을 진행한 하드웨어 환경은 표 3과 같다. 각 모델의 학습은 공식 구현체를 기반으로 진행하였으며, 하이퍼파라미터는 기존 논문의 권장 설정을 따랐다. 또한, 무작위 시드는 고정하여 재현 가능성을 보장하였다.

3.2 데이터셋

임베딩의 범용성을 다각도로 평가하기 위해, UEA 시계열 아카이브[10]에서 서로 다른 특성을 가진 4개의 벤치마크

Datasets	# of Samples	Time Series Length	# of Channels	# of Classes	Domains	Characteristics
BasicMotions	80	100	6	4	Motion Recognition	Periodic and Clear Patterns
Epilepsy	552	178	3	4	Healthcare	EGG-Based Complex, Non-Linear Pattern
Libras	360	45	2	15	Gesture Recognition	2D Coordinate Data High Class-Specific Trajectory Similarity
HandMovement Direction	234	400	10	4	Motion Tracking	3D Coordinate Data Four-Way Motion Pattern

표 2. 군집화 성능 평가에 사용된 데이터셋

Specifications	
GPU	2 × NVIDIA GeForce RTX 3090 24GB
CPU	Intel Core i7-9700K @ 3.60GHz (8 cores)
RAM	64GB
OS	Ubuntu 20.04.6 LTS
CUDA	11.8

표 3. 실험 환경

데이터셋을 선정하였다. BasicMotions는 주기적이고 명확한 패턴을 갖는 4가지 기본 동작 데이터이다. Epilepsy는 복잡하고 비선형적인 의료 신호(EEG) 데이터이다. Libras는 브라질 수화 동작을 2차원 좌표로 기록한 데이터로, 클래스별 궤적 패턴의 유사성이 높다는 특징이 있다. 마지막으로 HandMovementDirection은 피실험자가 손을 상, 하, 좌, 우의 네 가지 방향으로 움직일 때의 3D 공간 좌표를 기록한 데이터이다. 모든 실험 과정에서 데이터의 레이블 정보는 최종 성능 평가 단계에서만 사용되었으며, 표현 학습 과정은 완전히 비지도 방식으로 진행되었다. 선정된 데이터셋의 상세 명세는 표 2와 같다.

3.3 실험 절차 및 구현 세부 사항

본 연구의 실험은 다음 3단계에 따라 수행되었다.

1. 표현 학습: 2장에서 소개한 5가지 모델로, 각 데이터셋에 대한 임베딩을 추출한다. 공정한 비교를 위해 모든 모델의 최종 임베딩 벡터 차원은 320으로 통일하였다.

2. 비지도 군집화: 학습된 인코더를 통해 추출된 임베딩 벡터 전체를 대상으로, K-Means 군집화 및 Spectral 군집화를 사용하여 임베딩 공간의 구조를 다각도로 평가한다. K-Means 군집화는 가장 보편적인 알고리즘으로, 임베딩이 선형적으로 분리할 수 있는 볼록 구조를 학습하였는지 평가하는 베이스라인이다. 반면, Spectral 군집화는 K-Means 알고리즘이 탐지하기 어려운 비선형 및 비볼록 구조의 클러스터를 효과적으로 탐지하여, 딥러닝 임베딩의 복잡한 공간을 평가한다. 두 알고리즘이 선형과 비선형이라는 상호 보완적인 클러스터 가정을 대표하기 때문에, K-Means 및 Spectral 군집화 알고리즘으로 실험을 진행하였다. 두 알고리즘 모두 군집의 수(k)는 각 데이터셋의 실제 클래스 수로 설정하였다. 또한, 모든 군집화 알고리즘은 scikit-learn[11]을 사용하여 구현하였다.

3. RI 및 NMI 기반 성능 평가: RI와 NMI의 두 가지 지표를 사용하여, 군집화 알고리즘의 할당 결과를 실제 정답과 정량적으로 비교한다. RI는 모든 샘플 쌍에 대해 "같은 군집에 속하는지"의 여부를 측정하여 전체적인 군집 할당의 정확도를 평가하며, NMI는 정보 이론 관점에서 군집 할당이 실제 클래스 정보를 얼마나 잘 반영하는지를 측정한다. 두 지표 모두 0에서 1 사이의 값을 가지며, 1에 가까울수록 우수한 성능을 의미한다. 우연에 의한 일치를 보정한 이 두 지표를 함께 사용함으로써 군집화 성능을 다각도로 평가할 수 있다.

3.4 실험 결과 및 분석

3.4.1 K-Means 군집화 성능 평가

TS2Vec은 평균 RI 0.768, NMI 0.429로 가장 우수한 성능을 보였다. 이는 TS2Vec의 계층적 대조학습이 시계열의 전역적 시간 일관성을 효과적으로 포착하여 K-Means가 선호하는 선형 분리 가능한 임베딩을 생성하기 때문으로 해석된다. PatchTST는 평균 RI 0.752, NMI 0.353으로 TS2Vec에 근접한 성능을 보이며 두 번째로 우수한 결과를 기록하였다.

TimeKD(RI 0.465, NMI 0.002)와 TimeCMA(RI 0.574, NMI 0.042)는 RI-NMI 불균형을 보였다. 높은 RI에도 불구하고 0에 가까운 NMI 수치는 할당된 군집이 실제 클래스 정보를 전혀 반영하지 못한다는 것을 의미한다. 이는 지식 증류나 교차 모달리티 정렬 방식의 LLM 활용이, 레이블 없는 시계열의 내재적 구조 학습에 실패했음을 시사한다.

데이터셋별로 살펴보면, BasicMotions에서는 TS2Vec(RI 0.854, NMI 0.820)의 성능이 가장 우수하였으며, MERIT(RI 0.751, NMI 0.503)가 그 뒤를 따랐다. Epilepsy에서도 TS2Vec(RI 0.706)이 가장 우수하였던 반면,

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

표 4. K-Means 군집화 및 Spectral 군집화를 활용한 성능 평가

MERIT는 비교적 저조한 성능(RI 0.106)을 보여 데이터 특성에 따른 안정성 문제를 드러냈다. Libras에서는 TS2Vec(RI 0.904), PatchTST(RI 0.870), MERIT(RI 0.867) 순서로 높은 RI 수치를 기록하였다.

HandMovementDirection에서는 모든 방법론이 낮은 NMI(<0.05)를 보였는데, 이는 현재의 임베딩 방법론으로는 네 방향 움직임 패턴이 의미 있게 분리되지 않음을 시사한다. 전반적으로, K-Means 군집화 알고리즘에서 선호하는 선형 분리 가능한 임베딩 구조는, 계층적 대조학습 방식인 TS2Vec 방법론이 가장 효과적으로 생성함을 확인하였다.

3.4.2 Spectral 군집화 성능

표 4의 Spectral 군집화 결과에서, 일부 방법론의 경우 K-Means와는 다른 양상을 보이는 현상을 발견하였다. TS2Vec의 경우 평균 RI가 0.768에서 0.808로 상승하였으며(+0.040), 두 알고리즘 모두에서 높은 성능을 유지하였다. 또한, PatchTST(RI +0.042)는 Spectral 군집화에서 RI 향상을 보였다.

TimeKD(RI 0.480, NMI 0.005)와 TimeCMA(RI 0.563, NMI 0.015)는 Spectral 군집화에서도 의미 있는 클러스터 형성에 저조한 성능을 보였다. 반면, 같은 LLM 기반 방법론인 MERIT의 경우 평균 RI 0.725, NMI 0.203을 기록하여 상대적으로 균형 잡힌 성능을 유지함을 확인할 수 있었다. 특히 MERIT의 RI 수치가 큰 폭으로 상승(+0.147)하였는데, MERIT의 임베딩이 국소적 비선형 다양체 구조를 보존하기 때문으로 해석된다.

데이터셋별로 살펴보면, BasicMotions에서는 TS2Vec(RI 1.000, NMI 1.000) 및 PatchTST(RI 1.000, NMI 1.000) 모델이 완벽한 군집화를 달성하였으며, Libras에서는 MERIT(RI 0.883)가 높은 RI를 기록하였다. 그러나 HandMovementDirection의 경우 모든 방법론이 0.03 미만의 낮은 NMI를 기록하였다. 이는 군집화 알고리즘의 종류와 무관하게 네 방향 움직임의 패턴은 구분될 수 없음을 시사한다. 전반적으로 Non-LLM 방법론인 TS2Vec이 두 군집화 알고리즘 모두에서 가장 안정적인 성능을 보였으며, PatchTST와 MERIT는 K-Means 대비 Spectral 군집화에서 성능 향상을 보여 비선형 구조 포착 능력을 입증하였다.

IV. 결론 및 향후 연구

본 연구는 Non-LLM 기반 모델(TS2Vec, PatchTST)의 군집화 성능이, LLM 기반 모델보다 우수함을 확인하였다.

TimeKD와 TimeCMA는 높은 RI에도 불구하고 0에 가까운 NMI 수치를 기록하여, 현재의 LLM 지식 전이 방식이 레이블이 없는 시계열 구조에 일반화되지 못함을 입증하였다. TS2Vec는 선형 분리성에, PatchTST와 MERIT는 비선형 다양체 구조 포착에 강점을 보였다.

향후 연구에서는 다양한 도메인과 데이터 특성을 포괄하는, 광범위한 범위에서의 성능 평가를 진행할 필요가 있다. 특히, HandMovementDirection 데이터셋에서 모든 방법론이 실패한 원인을 규명하고, 군집화 난이도에 직접적인 영향을 미치는 데이터 특성을 찾아내기 위한 체계적인 분석이 필요하다. 또한, LLM 기반 방법론에서 관찰된 RI-NMI 수치 불균형 현상에 대한 심층 분석이 필요하다. 불균형이 발생하는 임베딩의 기하학적 특성을 규명하고, 이러한 불균형을 해소할 수 있는 방법론의 개발이 요구된다. 이후 LLM 기반 방법론의 군집화 성능에 대한 근본적인 한계를 극복하기 위하여, 군집화 친화적 손실함수 통합, 대조학습과 LLM의 하이브리드 접근법, 비지도 학습에 특화된 새로운 LLM 활용 전략의 모색 등이 필요하다. 나아가, 스트리밍 환경에서의 온라인 군집화 성능과 임베딩 갱신 전략 연구를 통하여 범용 임베딩에 대한 실용적 가치를 높일 수 있을 것으로 기대한다.

REFERENCES

[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] McQueen, James B. "Some methods of classification and analysis of multivariate observations." Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.. 1967.

[7] Ng, Andrew, Michael Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." Advances in neural information processing systems 14 (2001).

[8] Rand, William M. "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical association 66.336 (1971): 846-850.

[9] McDaid, Aaron F., Derek Greene, and Neil Hurley. "Normalized mutual information to evaluate overlapping community finding algorithms." arXiv preprint arXiv:1110.2515 (2011).

[10] Bagnall, Anthony, et al. "The UEA multivariate time series classification archive, 2018." arXiv preprint arXiv:1811.00075 (2018).

[11] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.