
Clustering for Incomplete Time Series Data

2023-01-06

Data Mining and Quality Analytics Lab

조억

발표자 소개

◆ 조억

- E-mail : ukjo@korea.ac.kr
- Data Mining & Quality Analytics Lab(김성범 교수님)
- 석박사 통합과정 학생(2019. 8월 입학)



종료 RL Fundamental
+
Cooperative Multi-Agent RL Framework
for Scalping Trading

발표자 : 조억

Cooperative Multi-Agent Reinforcement I

발표자: 조억

📅 2020년 1월 17일

종료 Reinforcement Learning for Generalization

2021-03-12
Data Mining and Quality Analytics Lab
조억

Reinforcement Learning for Generalization

발표자: 조억

📅 2021년 3월 12일

종료 Metric Studio
"Metric is all you need for Model Evaluation"

Data Mining & Quality Analytics Lab
2020년 8월 28일
발표자 : 조억

Metrics is all you need for model evaluation

발표자: 조억

📅 2020년 8월 28일

종료 Time Series Representation Learning

2022-01-14
Data Mining and Quality Analytics Lab
조억

Time Series Representation Learning

발표자: 조억

📅 2022년 1월 14일

❖ 관심분야

- ✓ Reinforcement Learning
- ✓ Time Series Data Analysis

목차

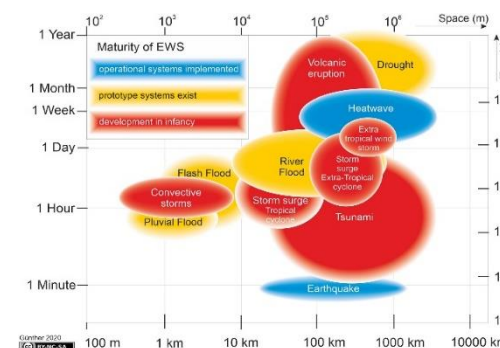
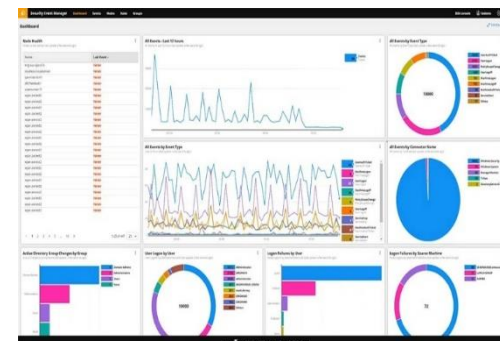
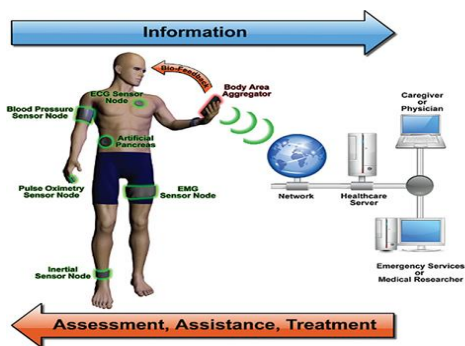
1. Time Series Clustering
2. Complete Time Series Clustering
3. Incomplete Time Series Clustering
4. Conclusions

Time Series Clustering

시계열 데이터

시계열 데이터 특성

- ❖ 헬스케어, 제조, 보안, 금융, 게임, 재난 예측 등 다방면에서 중요성 증대
- ❖ 데이터가 많으나 레이블링 얻기가 어려움
- ❖ 노이즈가 많고 고차원의 학습 양이 크고 정량적으로 판단이 쉽지 않음



시계열 클러스터링

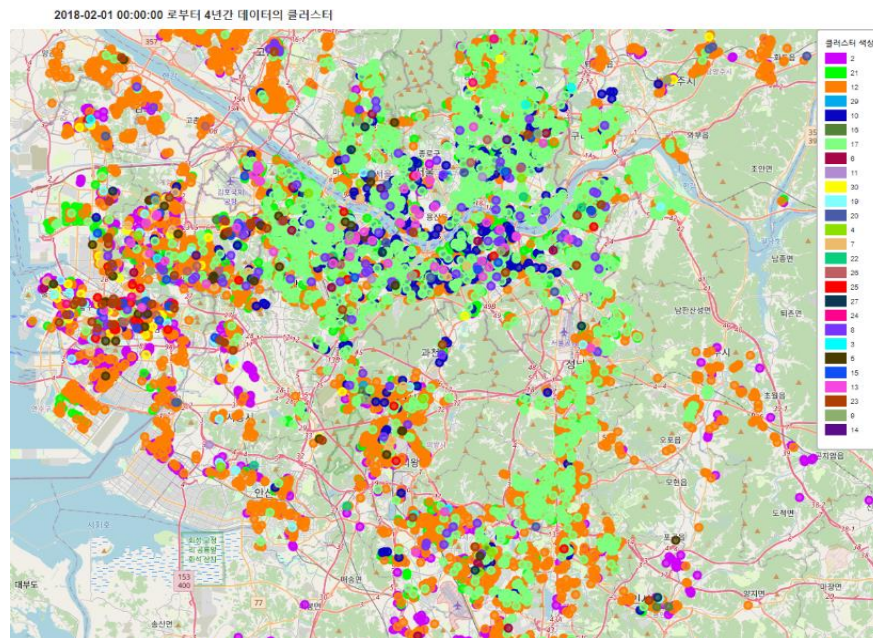
시계열 클러스터링 목적

- ❖ 복잡한 시계열 데이터에 대해 1차적으로 분석가로 하여금 인사이트 제공하며, 이상치 탐지, 트렌드 분석, 패턴 확인, 예측 등에 적용 가능

투자자산 가격 클러스터링을 통한 자산 배분 포트폴리오 구성



아파트 시세 클러스터링을 통한 유사 단지 실측치를 이용한 결측치 예측

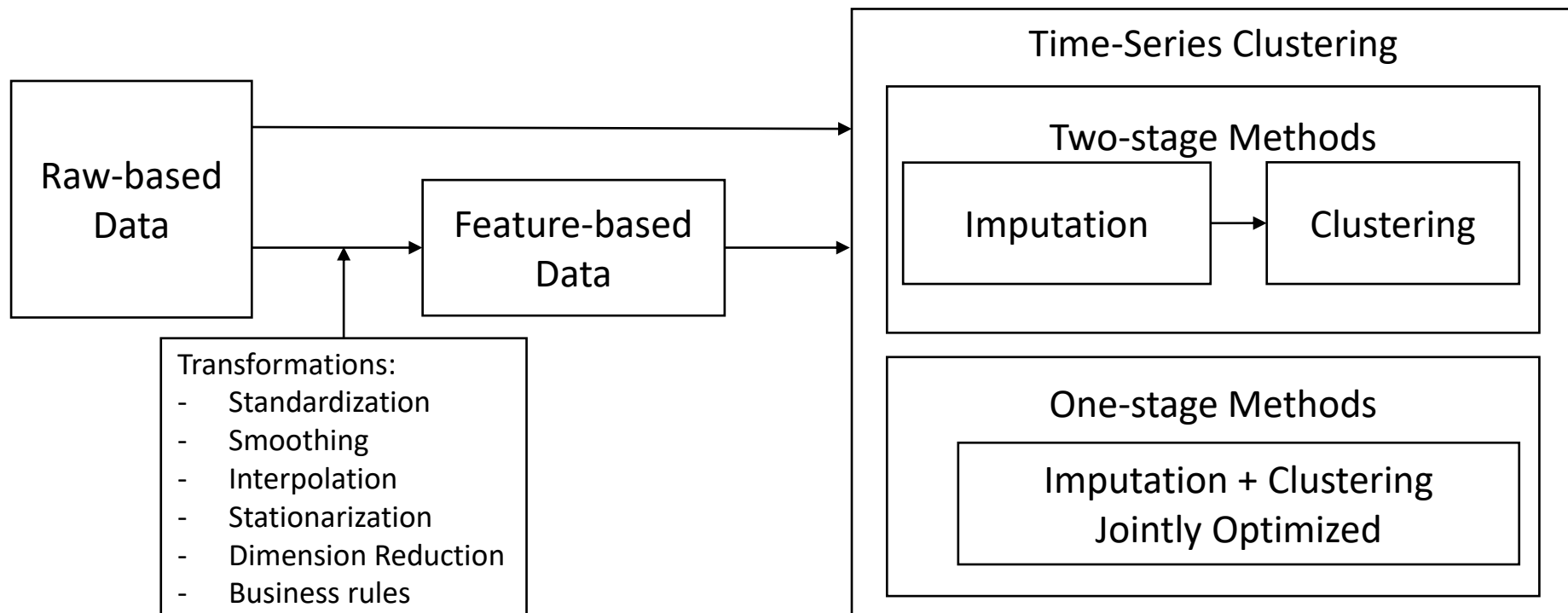


Tanca, L., Azzalini, D., Azzalini, F., & Mazuran, M. (2019, June). Tracking the Evolution of Financial Time Series Clusters. In *DSMM 2019-5th Workshop on Data Science for Macro-modeling with Financial and Economic Datasets*.

시계열 데이터 클러스터링

시계열 클러스터링 접근방법

- ❖ 기존 데이터를 사용하거나 도메인 전문가 판단에 의해 피쳐 추출 가능
- ❖ Two-stage에 의한 클러스터링 방법은 조합의 경우가 많아 최적 찾기 어려움



시계열 클러스터링

시계열 데이터를 클러스터링 적용시 한계점

- ❖ 널리 알려진 K-Means, K-Shape은 시계열 데이터가 아닌 일반 데이터로 많이 사용되어 시계열 데이터에 맞는 적합한 모델이 필요
- ❖ 고차원 데이터 존재하며, 길이가 길 경우 계산 복잡도가 크게 증가하며 특히 유사도를 측정하기 위해 DTW(Dynamic Time Warping)은 $O(n^2)$ 이므로 계산량 많음
- ❖ 노이즈가 심하고 복잡한 데이터이므로, 전문가를 통한 데이터 전처리가 필요
- ❖ 비선형 데이터 경우, 선형을 가정하는 일부 방법론은 사용 불가
- ❖ 데이터 수집 주기가 상이하며, 수집 에이전트 결함 문제로 현실 데이터는 결측치를 대부분 포함하므로 결측치가 클러스터링 성능을 떨어뜨릴 수 있음

Complete Time Series Clustering

Deep Temporal Clustering : Fully Unsupervised Learning of Time-Domain Features

❖ Deep Temporal Clustering : Fully Unsupervised Learning of Time-Domain Features

- ❖ 2018년 ICLR Reject되어 현재는 아카이브에 등록
- ❖ 2023년 1월 6일 기준, 87회 인용
- ❖ 시계열 데이터 군집 분석을 위한 초기 딥러닝 기반 모델이며 End-to-End로 학습하는 프레임워크 제시

Under review as a conference paper at ICLR 2018

DEEP TEMPORAL CLUSTERING: FULLY UNSUPER- VISED LEARNING OF TIME-DOMAIN FEATURES

Anonymous authors

Paper under double-blind review

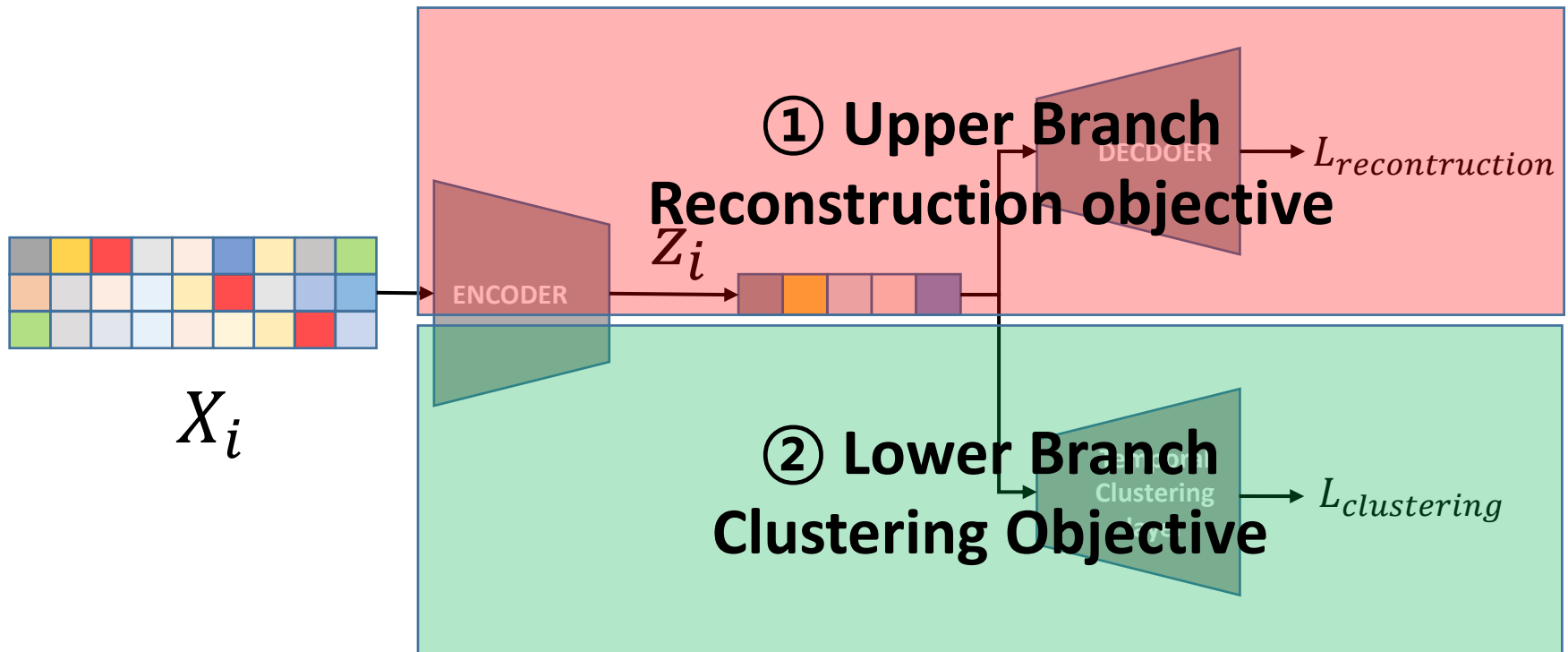
ABSTRACT

Unsupervised learning of time series data, also known as temporal clustering, is a challenging problem in machine learning. Here we propose a novel algorithm, Deep Temporal Clustering (DTC), to naturally integrate dimensionality reduction and temporal clustering into a single end-to-end learning framework, fully unsupervised. The algorithm utilizes an autoencoder for temporal dimensionality reduction and a novel temporal clustering layer for cluster assignment. Then it jointly optimizes the clustering objective and the dimensionality reduction objective. Based on requirement and application, the temporal clustering layer can be customized with any temporal similarity metric. Several similarity metrics and state-of-the-art algorithms are considered and compared. To gain insight into temporal features that the network has learned for its clustering, we apply a visualization method that generates a region of interest heatmap for the time series. The viability of the algorithm is demonstrated using time series data from diverse

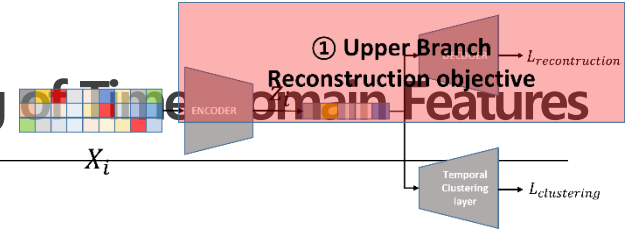
Deep Temporal Clustering : Fully Unsupervised Learning of Time-Domain Features

Overview

- ❖ 오토인코더 기반 표현 벡터를 얻기 위해 복원 목적 함수(dimensionality reduction)로는 클러스터링에 적합한 Representation을 얻기가 어려우므로, 딥러닝 모델에서 최초로 클러스터링에 적합한 표현을 얻기 위해 목적 함수(clustering objectives)를 추가
- ❖ 오버피팅과 낮은 성능이 문제인 Two-stage 대신 딥러닝 기반 최초 One-stage 방법 제안

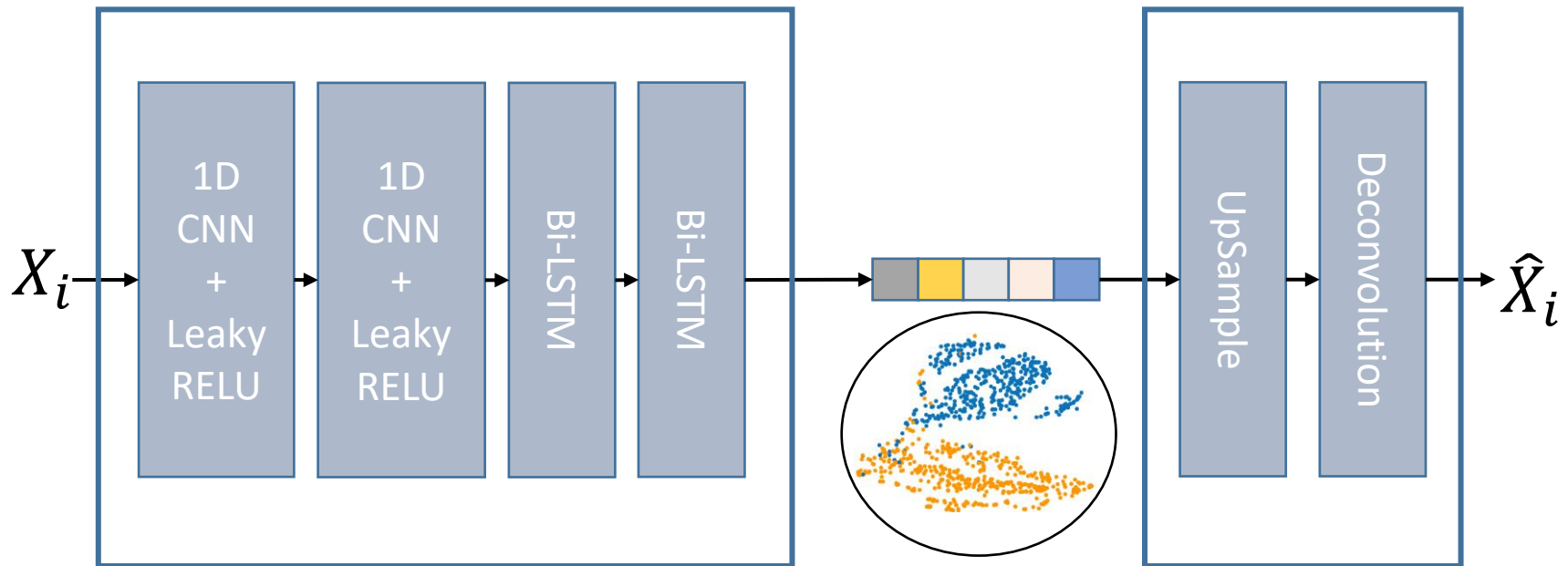


Deep Temporal Clustering : Fully Unsupervised Learning of Time

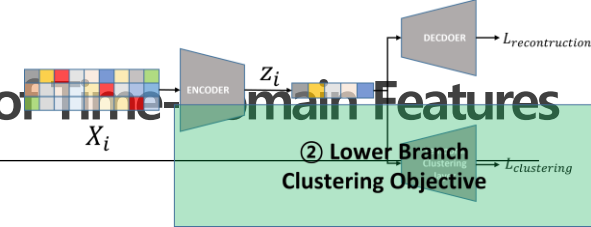


Reconstruction Objective

- ❖ 고차원 + 긴 시계열 데이터를 작은 차원을 가진 latent space로 변환하여 효율성 증대
- ❖ Noise에 강하고 관련이 없는 정보에 대해 필터링이 가능
- ❖ 일반 오토 인코더 구조와 다르게, 인코더와 디코더가 대칭을 이루고 있지 않음
- ❖ Short-Term의 시계열 패턴과 전체 관점의 Temporal 정보 얻기 위해 Convolution과 Bi-LSTM 사용



Deep Temporal Clustering : Fully Unsupervised Learning of Time-series main Features



Clustering Objective

- ❖ 인코더 출력이 잠재공간에서 k 개 클러스터 군집이 잘 형성되도록 Manifold를 가지게 학습
- ❖ 초기 학습된 인코더가 예측한 값들이 어느 정도 옳을 것이라라는 가정이 필요하며, 이를 기반으로 초기 각 클러스터의 중심을 구하고 이에 가장 가까운 클러스터를 높은 확률로 예측하도록 학습

- ① Reconstruction Objective으로만 인코더 초기 학습하여, 출력 벡터 z_i 생성
- ② Hierarchical Clustering with Complete Package를 통한 k 개 군집 진행
- ③ 각 군집에 대한 평균 centroid 인 $w_j, j \in k$ 생성
- ④ i 번째 샘플이 k 개 클러스터 각각의 중심과의 거리를 나타내는 벡터를 t분포 커널 ($\alpha=1$) 변환을 하여 확률로 변환

$$q_{i,j} = \frac{\left(1 + \frac{\text{similarity}(z_i, w_j)}{\alpha}\right)^{\frac{-(\alpha+1)}{2}}}{\sum_{j=1}^k \left(1 + \frac{\text{similarity}(z_i, w_j)}{\alpha}\right)^{\frac{-(\alpha+1)}{2}}}$$

Complexity Invariant Similarity(CID)

Deep Temporal Clustering : Fully Unsupervised Learning of Time-

Clustering Objective

- ❖ 인코더 출력이 잠재공간에서 k 개 클러스터 군집이 잘 형성되도록 Manifold를 가지게 학습
- ❖ 초기 학습된 인코더가 예측한 값들이 어느 정도 옳을 것이다라는 가정이 필요하며, 이를 기반으로 초기 각 클러스터의 중심을 구하고 이에 가장 가까운 클러스터를 높은 확률로 예측하도록 학습

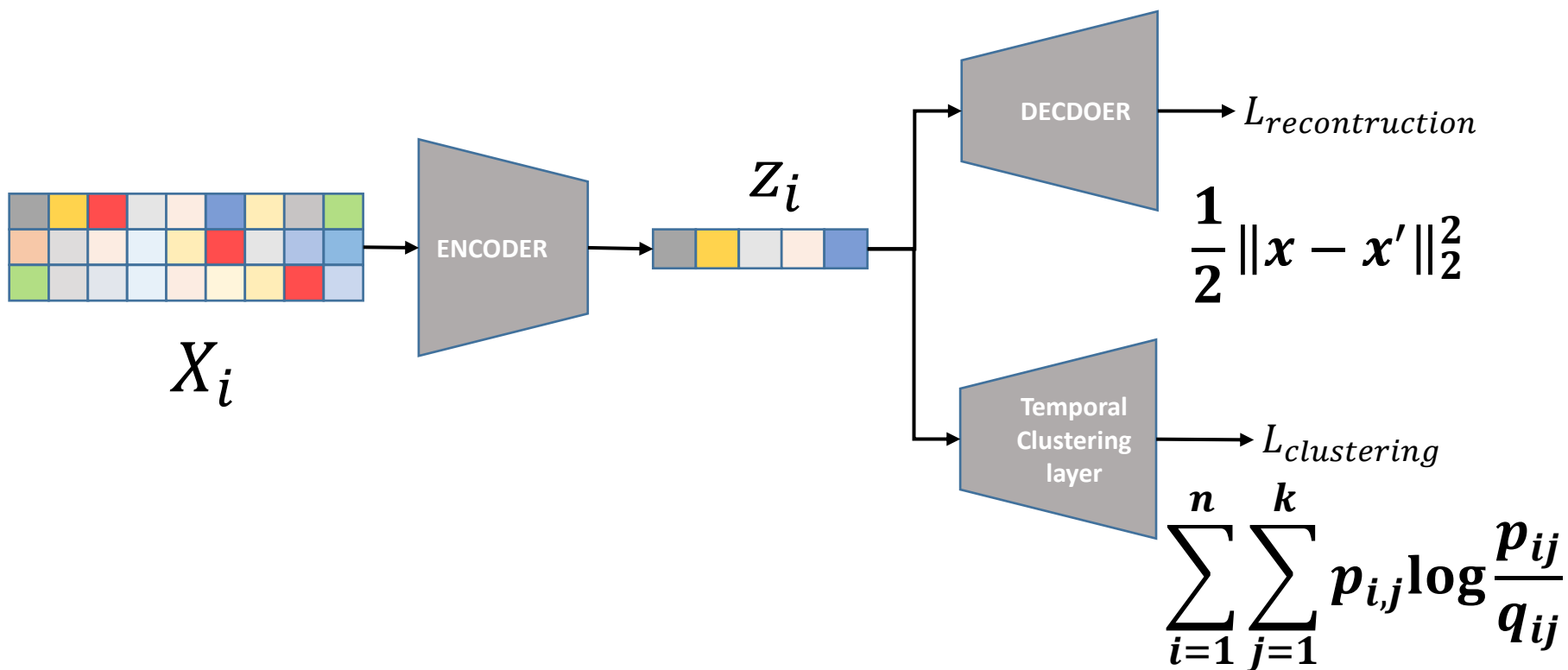
⑤ 비지도학습 특성상 p 의 값은 정답이어야 하는데, 비지도학습이라 centroid vector가 없으므로 아래와 같이 초반에 학습된 인코더의 출력의 q 값을 신뢰하고 q 값이 큰 것이 레이블이 크게 high confidence 예측이 되도록 제곱을 취하며 확률 값으로 변환

$$p_{i,j} = \frac{q_{i,j}^2 / f_j}{\sum_{j=1}^k q_{i,j}^2 / f_j}, \text{ where } f_j = \sum_{i=1}^n q_{i,j}$$

Deep Temporal Clustering : Fully Unsupervised Learning of Time-Domain Features

Deep Temporal Clustering : Fully Unsupervised Learning of Time-Domain Features

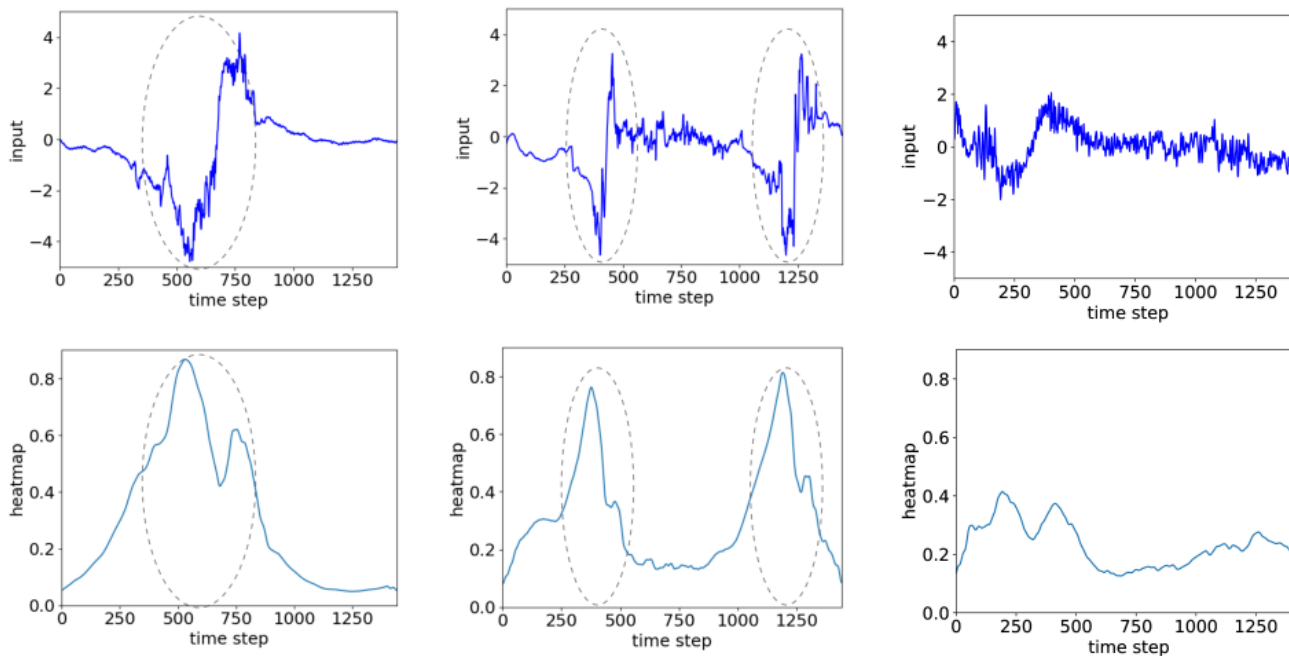
- ❖ 오토인코더의 모델 파라미터와 클러스터의 센터들을 L_c 와 L_{ae} 으로 학습
- ❖ 매번 SGD 업데이트마다 Target distribution p 를 업데이트를 하면서 학습



Deep Temporal Clustering : Fully Unsupervised Learning of Time-Domain Features

Experiments

- ❖ DTC 모델에서 나온 클러스터 정보를 라벨 정보로 활용하여 지도학습 진행
- ❖ 클러스터를 구분 지을 수 있는 특징에 대한 부분을 Heat Map Network으로 시각화



Learning Representations for Time Series Clustering

❖ Learning Representations for Time Series Clustering

- ❖ 2019 NIPS Accept paper
- ❖ 2023년 1월 6일 기준, 93회 인용
- ❖ DTC 방법론에 딥러닝에 적용 가능한 K-Means Loss와 Fake sample 감지하는 Auxiliary Task 통해 성능 향상시킴

Learning Representations for Time Series Clustering

Qianli Ma

South China University of Technology
Guangzhou, China
qianlima@scut.edu.cn

Jiawei Zheng*

South China University of Technology
Guangzhou, China
csjwzheng@foxmail.com

Sen Li *

South China University of Technology
Guangzhou, China
awslee@foxmail.com

Garrison W. Cottrell

University of California, San Diego
CA, USA
gary@ucsd.edu

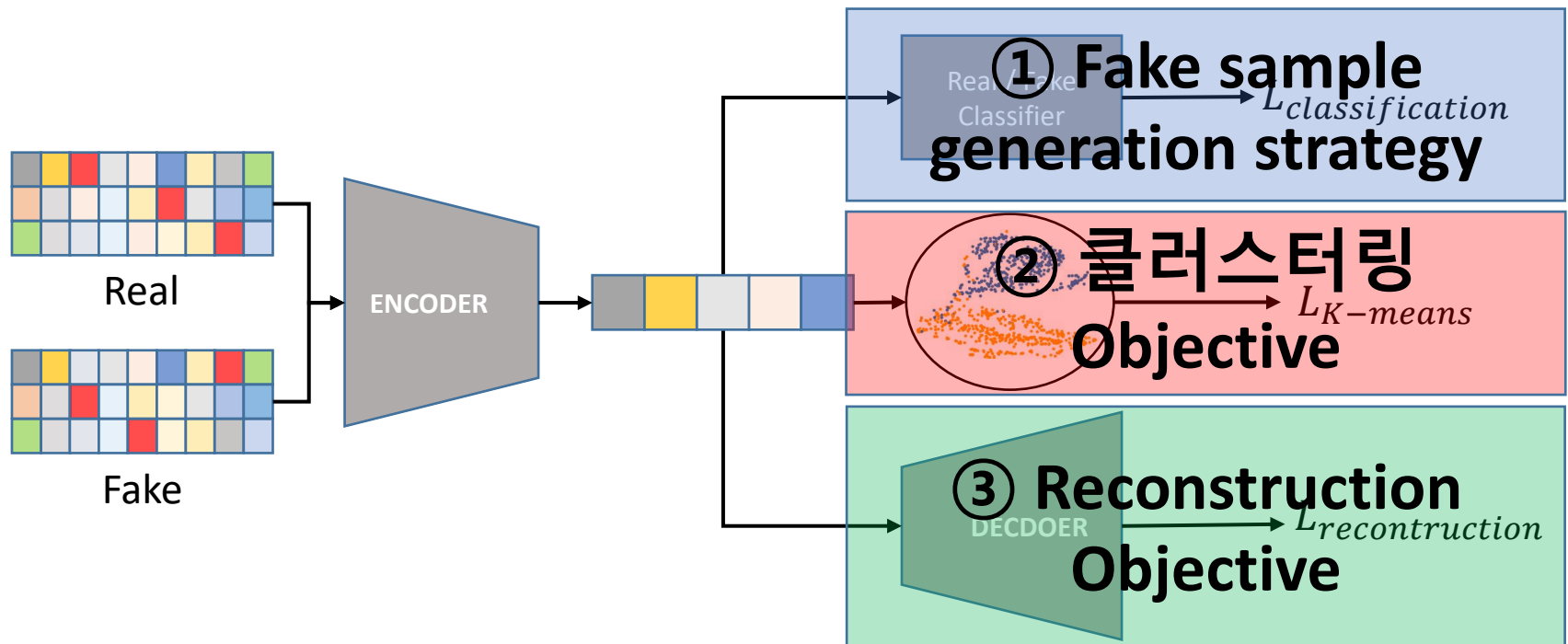
Abstract

Time series clustering is an essential unsupervised technique in cases when category information is not available. It has been widely applied to genome data, anomaly detection, and in general, in any domain where pattern detection is important. Although feature-based time series clustering methods are robust to noise and outliers, and can reduce the dimensionality of the data, they typically rely on domain knowledge to manually construct high-quality features. Sequence to sequence

Learning Representations for Time Series Clustering

Overview

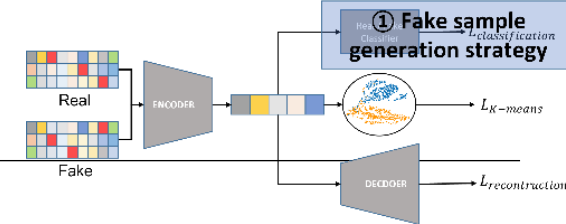
- ❖ Seq2seq 모델의 영감을 받아, 시계열 데이터를 적용한 논문
- ❖ DTC 과 유사하지만, 인코더(Bi-Directional Dilated Recurrent Neural Networks) 와 추가의 Objectives(K-Means Objective) 를 사용하여 성능 향상과 시각화로 학습이 잘됨을 보여줌



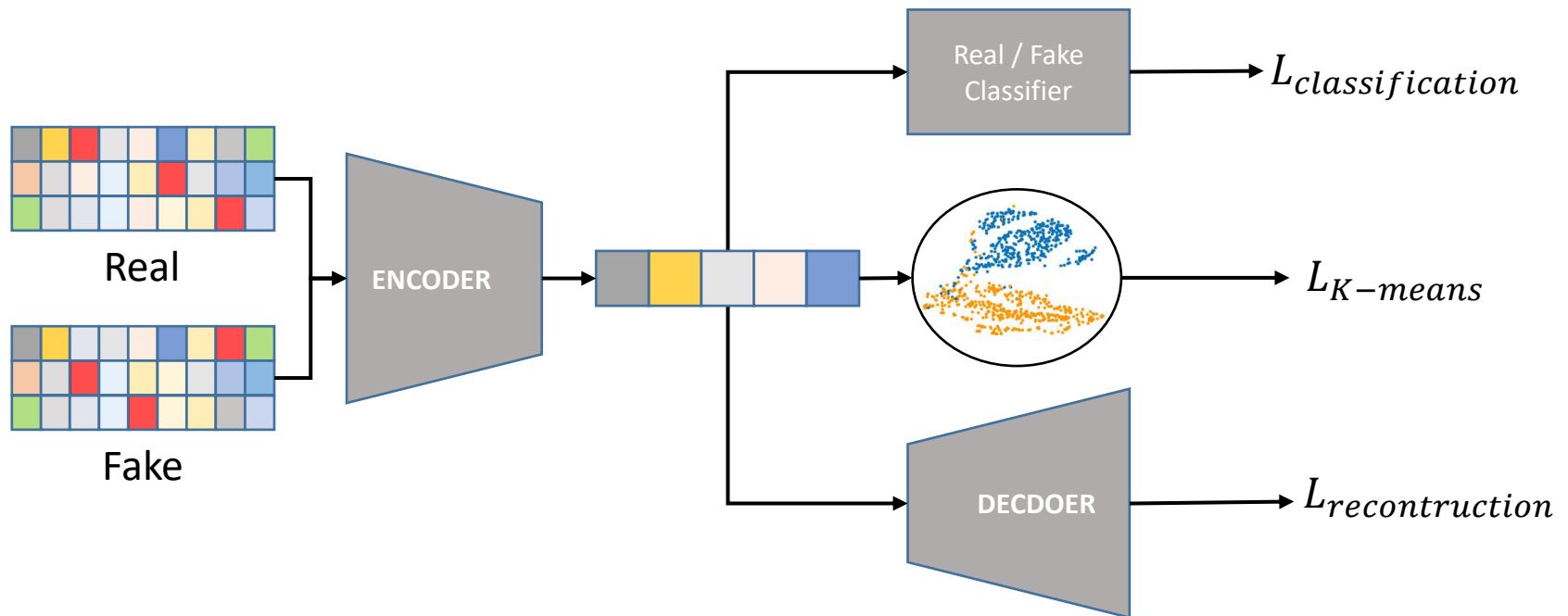
Learning Representations for Time Series Clustering

Fake sample generation strategy

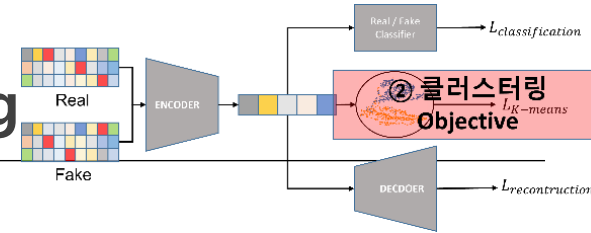
❖ 입력 시계열의 20%를 랜덤하게 Shuffling 하여 가짜 데이터를 생성



$$L_{classification} = -\frac{1}{2N} \sum_{i=1}^{2N} \sum_{j=1}^2 1\{y_{i,j} = 1\} \log \frac{\exp^{\hat{y}_{i,j}}}{\sum_{j=1}^2 \exp^{\hat{y}_{i,j}}}$$



Learning Representations for Time Series Clustering



클러스터링 Objective

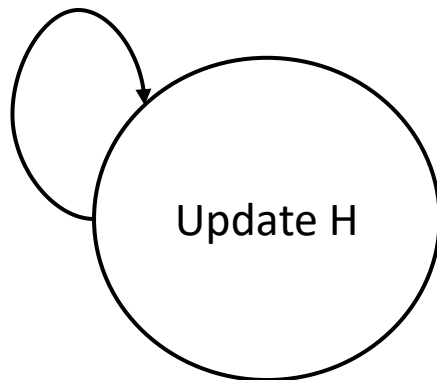
- ❖ K Means Clustering은 data vector와 cluster centroid 사이의 거리를 최소화하는 문제 푸는 문제로 정의하며, [1] 논문을 이용하여 trace maximization 문제로 재정의하여 딥러닝 학습이 가능한 목적식을 이용

$$\lambda L_{K-means} = \text{Tr}(H^T H) - \text{Tr}(F^T H^T H F)$$

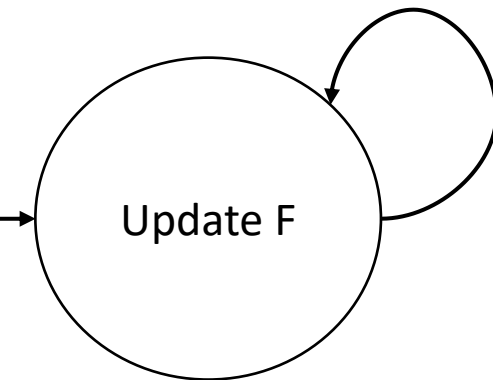
$$\text{where } H \in R^{h \times N}, F \in R^{N \times k}$$

$$s.t. F^T F = I \Rightarrow \text{relaxation} \Rightarrow F = k\text{-truncated SVD of } H \text{ (Ky fan theorem)}$$

by SGD



주기(하이퍼파라미터)마다

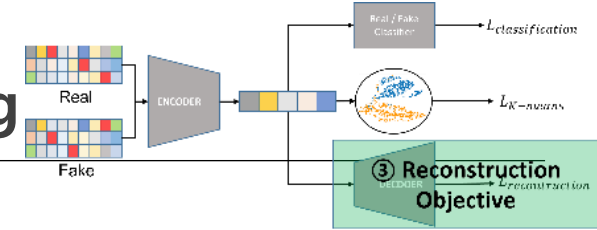


by Ky fan theorem

[1] Zha, H., He, X., Ding, C., Gu, M., & Simon, H. (2001). Spectral **relaxation** for k-means clustering. *Advances in neural information processing systems*, 14.

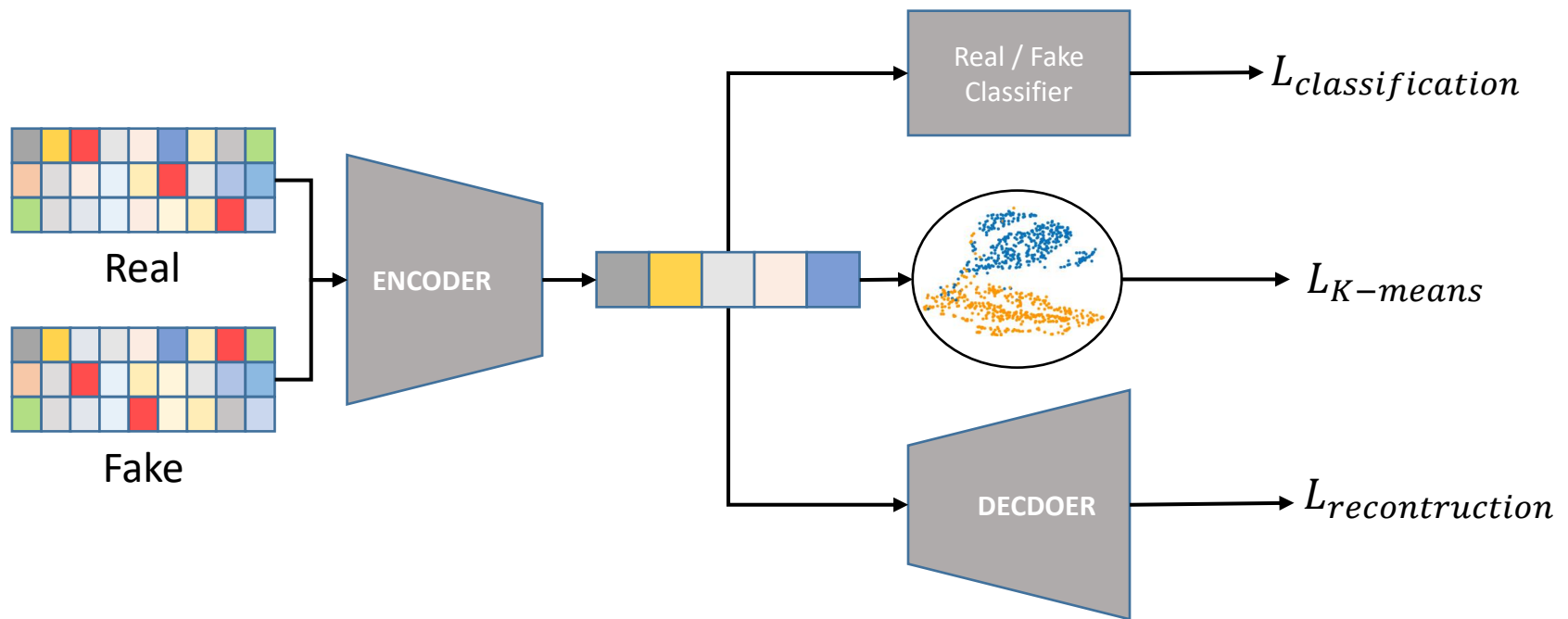
Learning Representations for Time Series Clustering

Reconstruction Objective



❖ 오토인코더의 목적함수와 동일하며 Mean Square Error를 통하여 원래 데이터로 복원

$$L_{reconstruction} = \frac{1}{n} \sum_{i=1}^n \|x_i - \hat{x}_i\|_2^2$$



Learning Representations for Time Series Clustering

Experiments

- ❖ 50 에폭 학습이 진행됨에 따라 클러스터내 샘플은 모이고 클러스터간 거리는 벌어짐을 시각화를 통해 확인

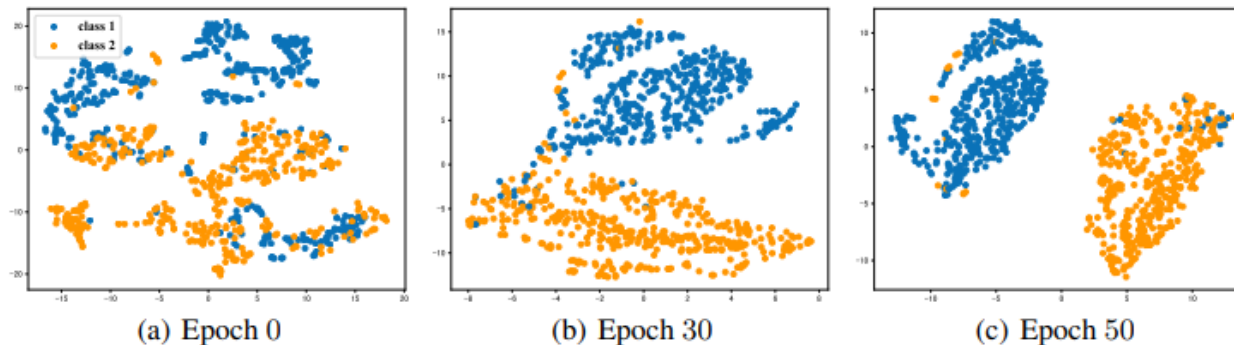


Figure 3: The learned representations on data set *ECGFiveDays* during the training process. From left to the right, the subfigure is obtained at Epoch 0, 30 and 50, respectively.

Learning Representations for Time Series Clustering

Experiments

- ❖ Ablation Study를 진행했을 경우, K-Means Loss와 Classification loss가 각각 도움이 되고 모두 적용이 되었을 경우 잘되는 것을 실험결과와 함께 t-SNE 시각화 결과까지보여줌

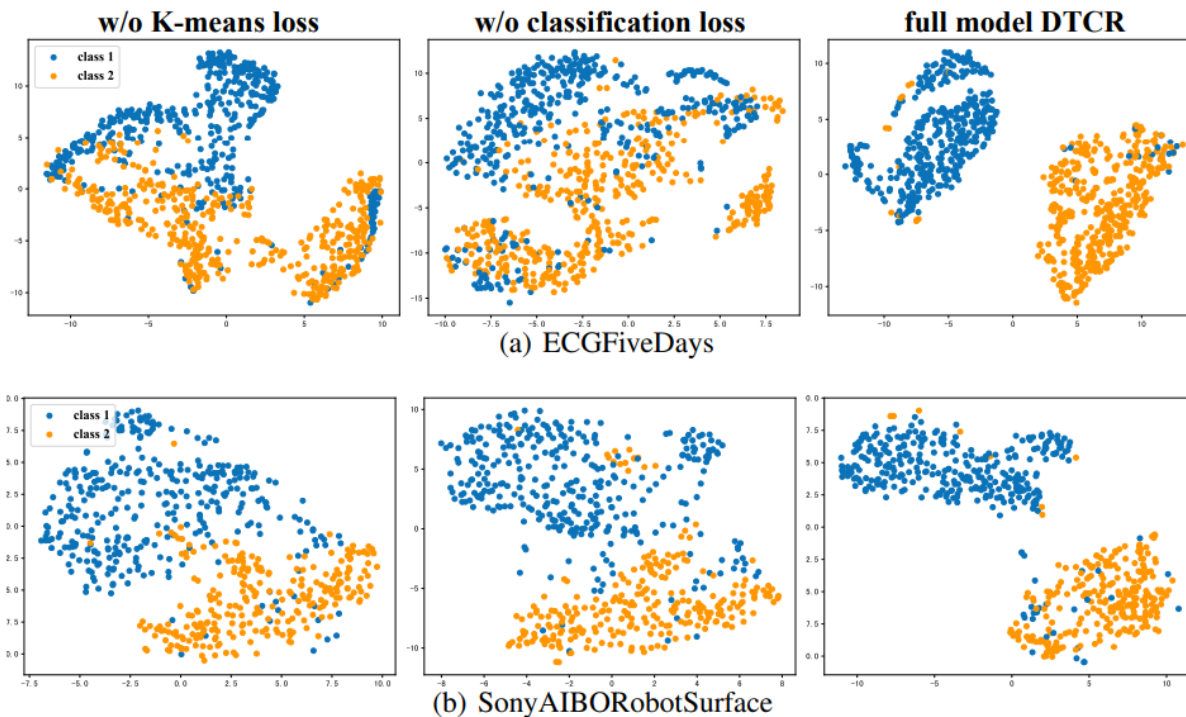


Figure 2: The visualizations with t-SNE on the datasets (a) *ECGFiveDays* and (b) *SonyAIBORobotSurface*. The colors of the points indicate the actual labels.

Learning Representations for Time Series Clustering

Experiments

- ❖ K-Means가 실수할 경우에도 학습을 잘 할 수 있는가에 대해 F 값을 랜덤하게 변형해서 실험했을 경우 $L_{reconstruction}$ 에 대한 학습이 도움이 되는 것을 볼 수 있다.

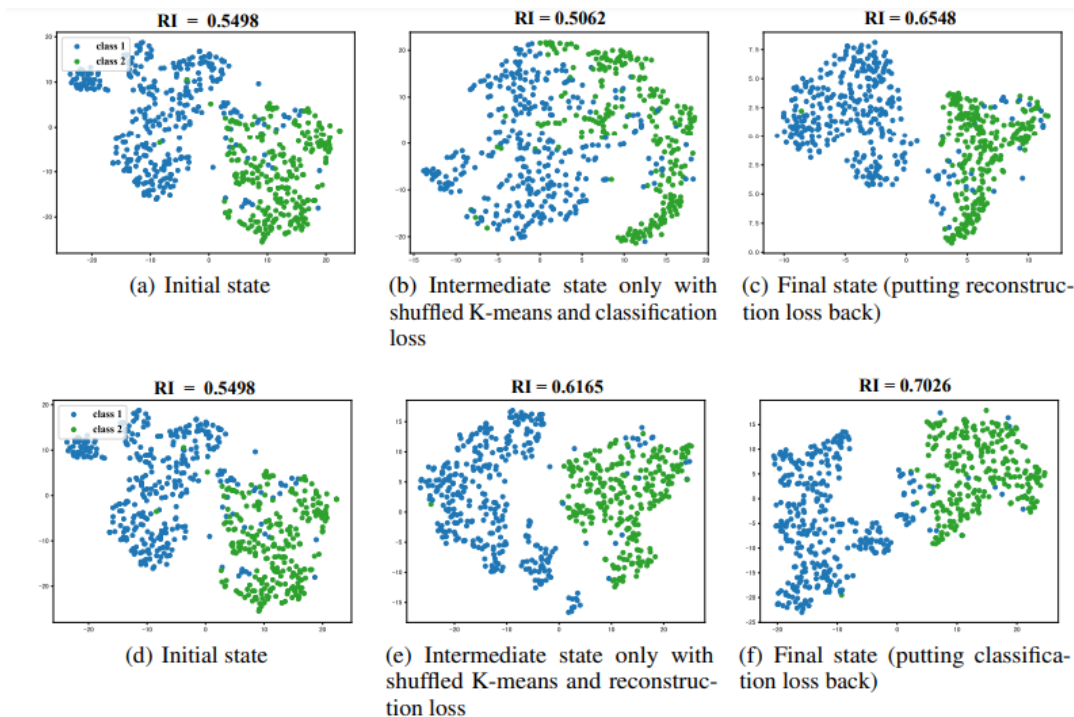


Figure 4: Robustness Analysis of DTCR on *SonyAIBORobotSurface*. Note that the (d) is the same as (a), replicated here for better illustration; hence the first and second rows start with the same state.

Incomplete Time Series Clustering

Learning Representations for Incomplete Time Series Clustering

❖ Learning Representations for Incomplete Time Series Clustering

- ❖ 2021 AAAI Accept paper
- ❖ 2023년 1월 6일 기준, 6회 인용
- ❖ DTCR 논문과 같은 저자가 쓴 논문으로, 결측치가 포함된 시계열 데이터에 대한 클러스터링 모델 제시

Learning Representations for Time Series Clustering

Qianli Ma

South China University of Technology
Guangzhou, China
qianlima@scut.edu.cn

Jiawei Zheng*

South China University of Technology
Guangzhou, China
csjwzheng@foxmail.com

Sen Li *

South China University of Technology
Guangzhou, China
awslee@foxmail.com

Garrison W. Cottrell

University of California, San Diego
CA, USA
gary@ucsd.edu

Abstract

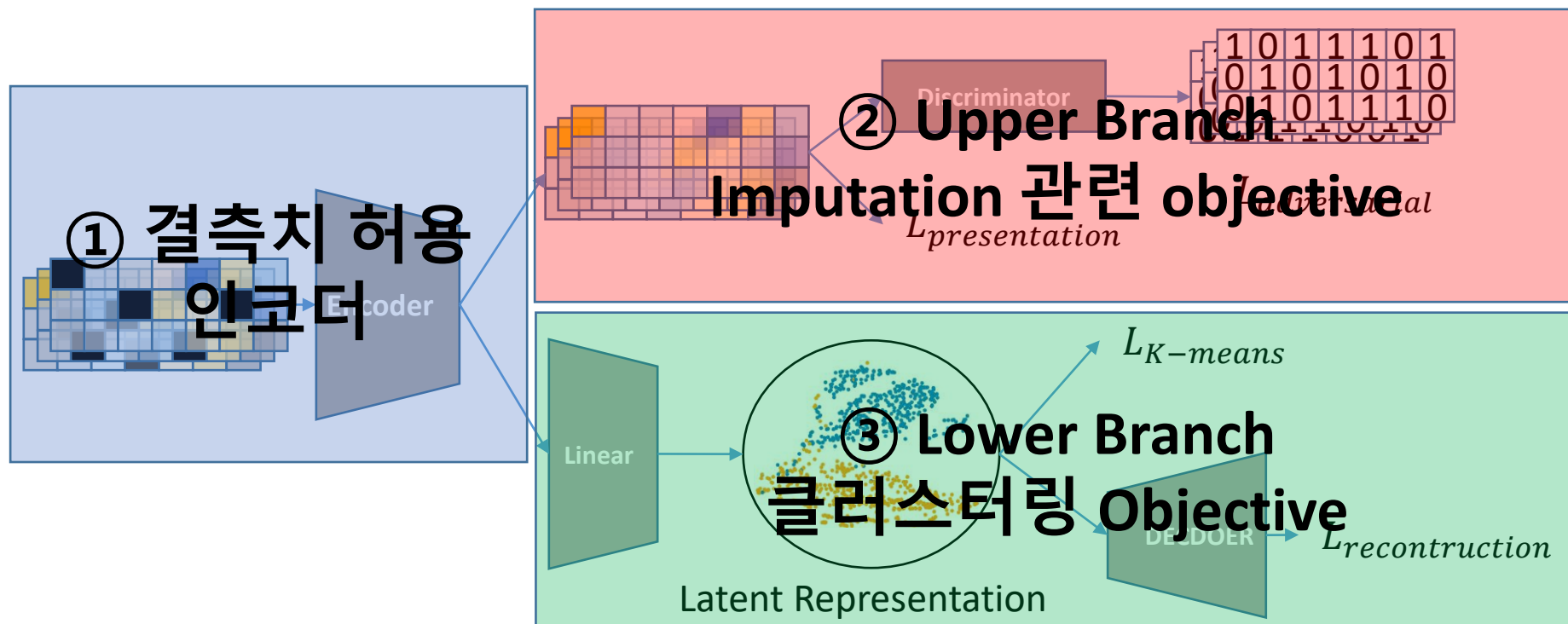
Time series clustering is an essential unsupervised technique in cases when category information is not available. It has been widely applied to genome data, anomaly detection, and in general, in any domain where pattern detection is important. Although feature-based time series clustering methods are robust to noise and outliers, and can reduce the dimensionality of the data, they typically rely on domain knowledge to manually construct high-quality features. Sequence to sequence

Learning Representations for Incomplete Time Series Clustering

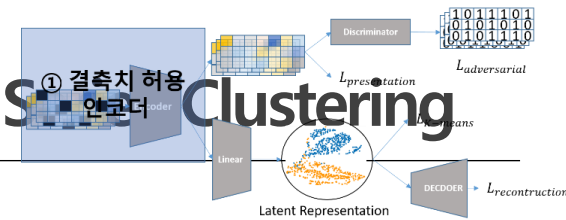
Overview

- ❖ 복원과 클러스터링을 위한 별도 학습 진행($L_{k-means}$) DTC, DTCR과 유사
- ❖ 기여점 : 결측치 포함한 데이터 학습이 가능한 구조로, 결측치 예측(L_{pre})과 Discriminator를 이용하여 결측치 여부를 판단하도록 추가 손실함수(L_{adv}) 정의함

$$L_{CRLI} = L_{prediction} + L_{reconstruction} + L_{adversarial} + \lambda L_{K-means}$$



Learning Representations for Incomplete Time Series Clustering



① 결측치 허용 인코더

- ❖ 인코더 부분은 결측치를 허용하기 위해 BRITS[1] 모델을 인코더 모델 구조 사용
- ❖ 결측치 부분이 regression layer를 통해 예측된 값으로 채워져 RNN에 사용

Original RNN

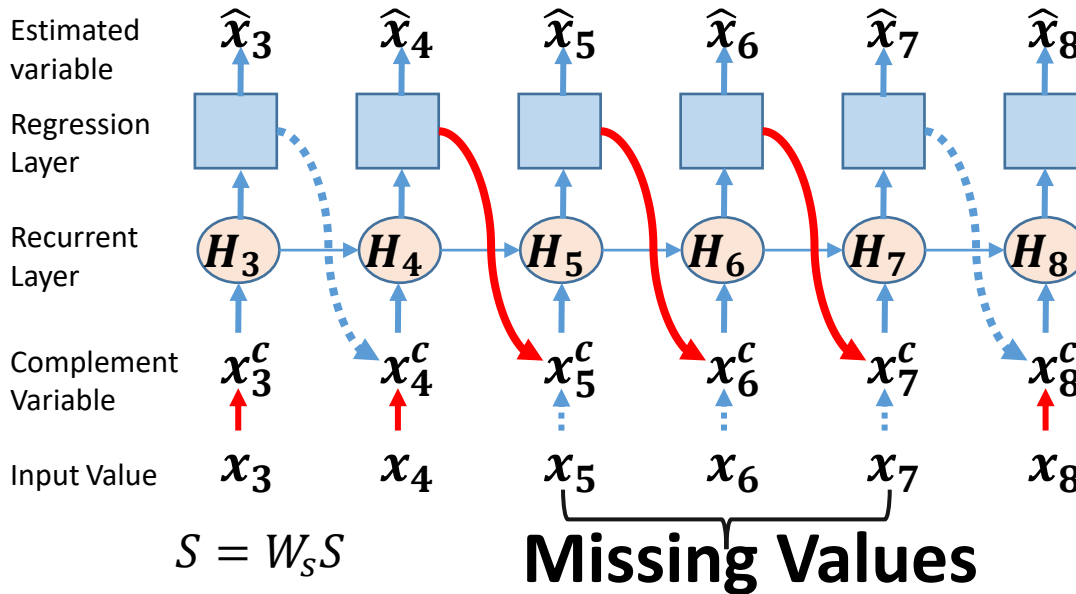
$$h_t = \tanh(W_h h_{t-1} + W_x x_t + b)$$



Revised RNN

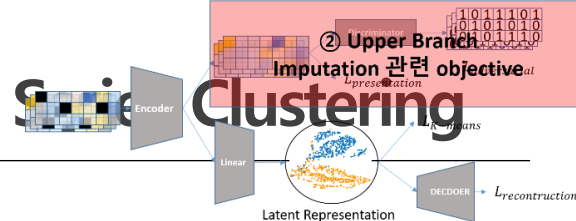
$$h_t = \tanh(W_h h_{t-1} + W_x \mathbf{u}_t + b)$$

$$\mathbf{u}_t = m_t \odot x_t + (1 - m_t) \odot \hat{x}_t$$



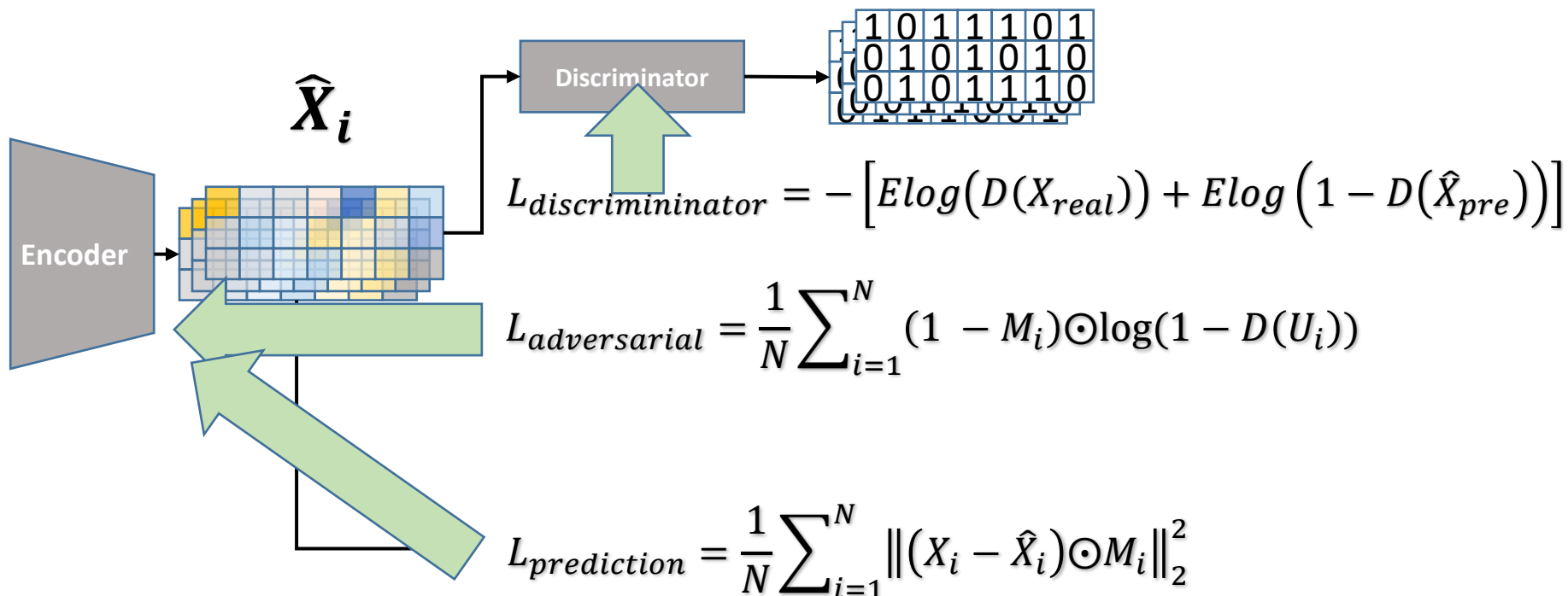
$$\hat{x}_t = W_{imp} h_{t-1} + b_{imp}$$

Learning Representations for Incomplete Time Series

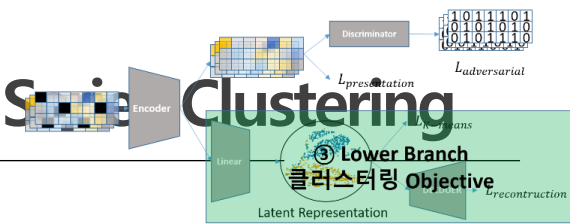


② Upper Branch - Imputation 관련 objective

- ❖ 인코더가 실측치를 잘 맞추도록 MSE Loss를 최소화하는 방향으로 학습함으로써 결측치 또한 잘 예측하도록 유도
- ❖ Encoder역할 외에도 Generator역할을 잘 하도록 Adversarial Loss를 최소화

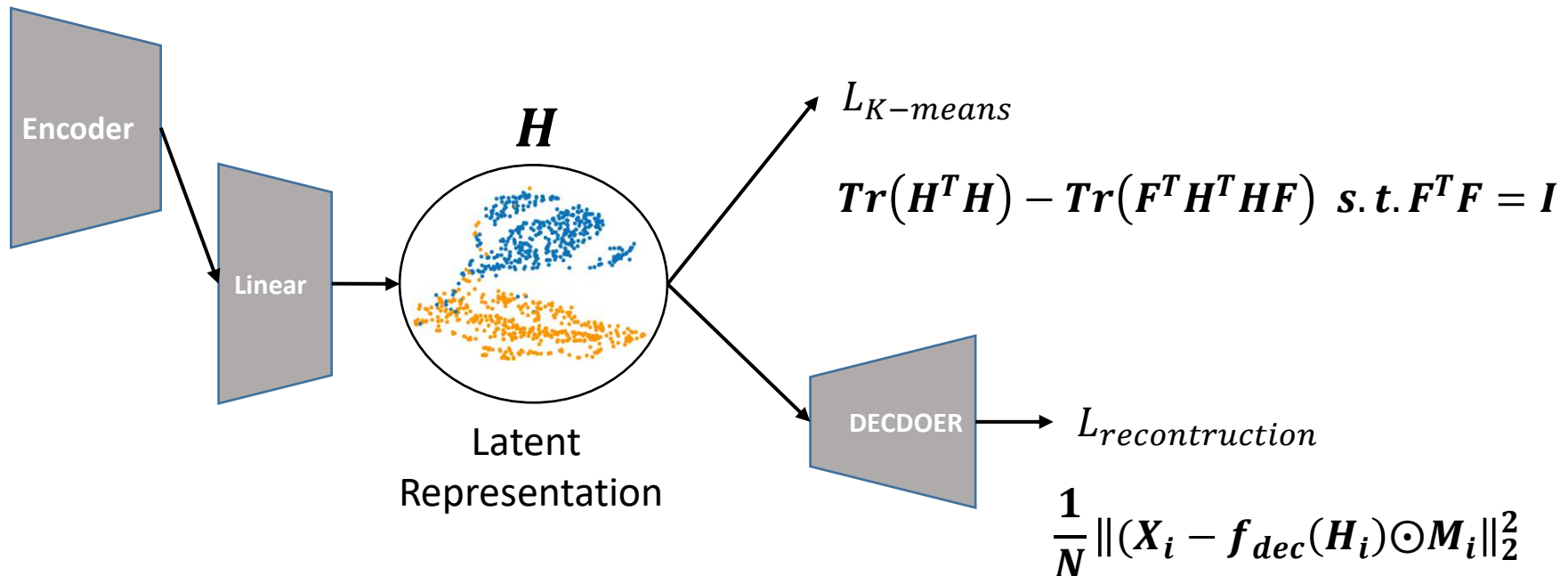


Learning Representations for Incomplete Time Series Clustering



③ Lower Branch - 클러스터링 Objective

- ❖ 시계열 Dynamics 를 이해하고 클러스터 형성을 위한 표현 벡터 생성 결측치 예측에서 클러스터링으로 에러가 전이 되는것을 방지하는 역할을 함
- ❖ DTCR 방법론과 마찬가지로 H가 변경됨에 따라 F가 변경이 되어야 하므로 iteratively 업데이트
- ❖ Temporal Reconstruction과 클러스터링이 잘되도록 K-Means Loss를 같이 목적함수로 활용



Learning Representations for Incomplete Time Series Clustering

Experiments

- ❖ 베이스라인에는 금일 주제인 DTC, DTCR 모델들이 포함되어 있으며, 총 8가지 데이터셋에 대해서 실험을 하며, Liang 데이터셋 외에는 더 좋은 성능을 내고 있음

Imputer	Cluster	Ali-v1	Ali-v2	Ali-v3	Blood	Chen	Vote	Liang	Pysionet	Best	avg Rank	avg RI	p-value
ZERO	KS	0.5	0.69	0.75	0.59	0.59	0.57	0.55	0.5	0	11.3	0.59	5.4e-4
	DEC	0.85	0.77	0.77	0.67	0.51	0.83	0.64	0.52	1	5.9	0.69	3.5e-3
	IDEC	0.72	0.79	0.77	0.67	0.5	0.84	0.7	0.52	0	5.6	0.69	2.9e-3
	DTC	0.47	0.74	0.73	0.72	0.56	0.57	0.64	0.5	0	9.4	0.62	2.7e-3
	DTCR	0.85	0.47	0.63	0.5	0.51	0.5	0.82	0.64	2	10.3	0.61	3.3e-2
GAIN	KS	0.63	0.7	0.75	0.62	0.58	0.69	0.64	0.5	0	9.6	0.64	6.8e-4
	DEC	0.46	0.51	0.51	0.67	0.77	0.81	0.42	0.51	1	10.6	0.58	1.0e-2
	IDEC	0.62	0.6	0.64	0.65	0.49	0.76	0.42	0.51	0	11.6	0.59	2.3e-5
	DTC	0.49	0.54	0.48	0.55	0.5	0.63	0.56	0.66	0	12.6	0.55	8.6e-4
	DTCR	0.56	0.78	0.72	0.51	0.51	0.5	0.75	0.73	0	8.8	0.63	1.5e-2
BRIT	KS	0.5	0.6	0.72	0.69	0.54	0.62	0.58	0.5	0	10.5	0.59	4.8e-4
	DEC	0.85	0.77	0.81	0.62	0.65	0.81	0.64	0.51	1	5.5	0.71	1.8e-2
	IDEC	0.82	0.83	0.74	0.64	0.61	0.87	0.7	0.5	0	5.9	0.71	2.8e-2
	DTC	0.47	0.7	0.7	0.75	0.54	0.69	0.64	0.5	0	9.8	0.62	1.8e-3
	DTCR	0.82	0.48	0.69	0.59	0.5	0.76	0.68	0.52	0	10.6	0.63	8.8e-3
VaDER		0.47	0.49	0.29	0.67	0.5	0.58	0.7	0.7	0	10.8	0.55	1.2e-2
CRLI		0.85	0.93	0.92	0.85	0.63	0.91	0.67	0.76	6	2	0.81	-

Table 1: Rand Index (RI) comparisons on 8 real incomplete time series datasets

$$RI = \frac{TP + TN}{n(n - 1)/2}$$

Learning Representations for Incomplete Time Series Clustering

Experiments

- ❖ 결측치만 평가했을 때는 클러스터링을 잘하기 위한 목적함수를 주었으므로 BRITS 대비 Imputation 성능인 RMSE가 떨어지지만, 클러스터링 성능에서는 우수한 성능을 보임

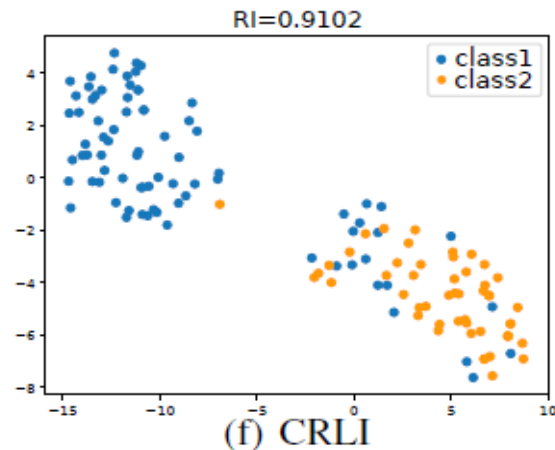
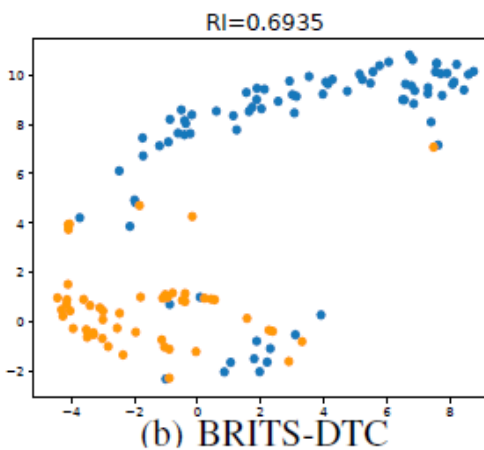
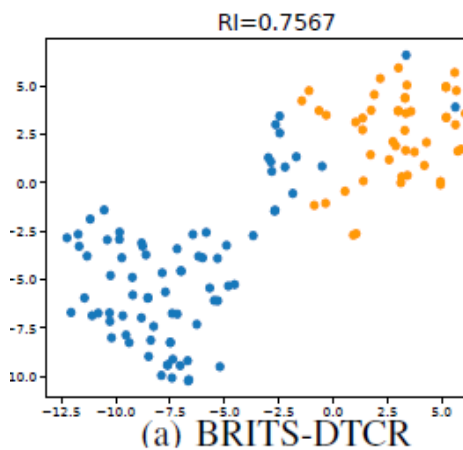
Metric	RMSE		RI		
Method	BRITS	CRLI	Raw_data	BRITS	CRLI
50words	0.67(0.19)	1.44(0.29)	0.941(0.01)	0.952(0.00)	0.952(0.00)
Adiac	0.93(0.47)	6.78(2.14)	0.921(0.02)	0.937(0.00)	0.934(0.00)
ArrowHead	1.36(0.50)	2.01(0.43)	0.556(0.00)	0.503(0.01)	0.575(0.01)
Beef	2.12(1.23)	2.90(0.99)	0.651(0.01)	0.648(0.01)	0.651(0.00)
BeetleFly	1.33(0.39)	3.03(0.16)	0.479(0.02)	0.516(0.01)	0.528(0.02)
BirdChicken	1.94(1.15)	3.88(0.85)	0.474(0.00)	0.474(0.00)	0.474(0.00)
Car	1.24(0.73)	3.04(1.12)	0.702(0.05)	0.678(0.02)	0.698(0.01)
CBF	3.37(0.20)	3.21(0.08)	0.704(0.01)	0.699(0.01)	0.710(0.01)
Chlorine	1.15(0.08)	3.10(0.17)	0.528(0.00)	0.528(0.00)	0.528(0.00)
CinC	2.19(0.80)	3.96(0.24)	0.693(0.01)	0.685(0.01)	0.695(0.00)
Coffee	1.11(0.36)	4.43(1.21)	0.484(0.00)	0.484(0.00)	0.506(0.01)
Computers	5.80(0.37)	6.40(1.03)	0.502(0.01)	0.510(0.00)	0.509(0.00)
Cricket_X	3.52(0.25)	4.07(0.60)	0.855(0.01)	0.855(0.00)	0.859(0.00)
Cricket_Y	2.89(0.34)	3.05(0.22)	0.858(0.01)	0.856(0.00)	0.861(0.00)
Cricket_Z	3.53(0.32)	4.15(0.89)	0.858(0.01)	0.855(0.00)	0.863(0.00)
Diatom	1.64(0.68)	2.00(0.38)	0.926(0.00)	0.925(0.00)	0.927(0.00)
DPO.AgeGroup	0.59(0.25)	1.04(0.04)	0.764(0.00)	0.763(0.00)	0.763(0.00)
DPO.Correct	0.58(0.28)	0.92(0.05)	0.505(0.00)	0.504(0.00)	0.504(0.00)
DPTW	0.85(0.50)	1.07(0.09)	0.774(0.02)	0.777(0.00)	0.775(0.00)
Earthquakes	13.02(0.13)	10.49(0.37)	0.499(0.03)	0.504(0.01)	0.511(0.01)
Average	2.492	3.548	0.684	0.683	0.691
AVG RANK	1.1	1.9	2.2	2.3	1.4
p-value	-	-	1.05E-02	3.19E-02	-

Table 3: RMSE & clustering RI comparison on the first 20 synthetic incomplete data sets of UCR archive. (the values in parentheses present standard deviations)

Learning Representations for Incomplete Time Series Clustering

Experiments

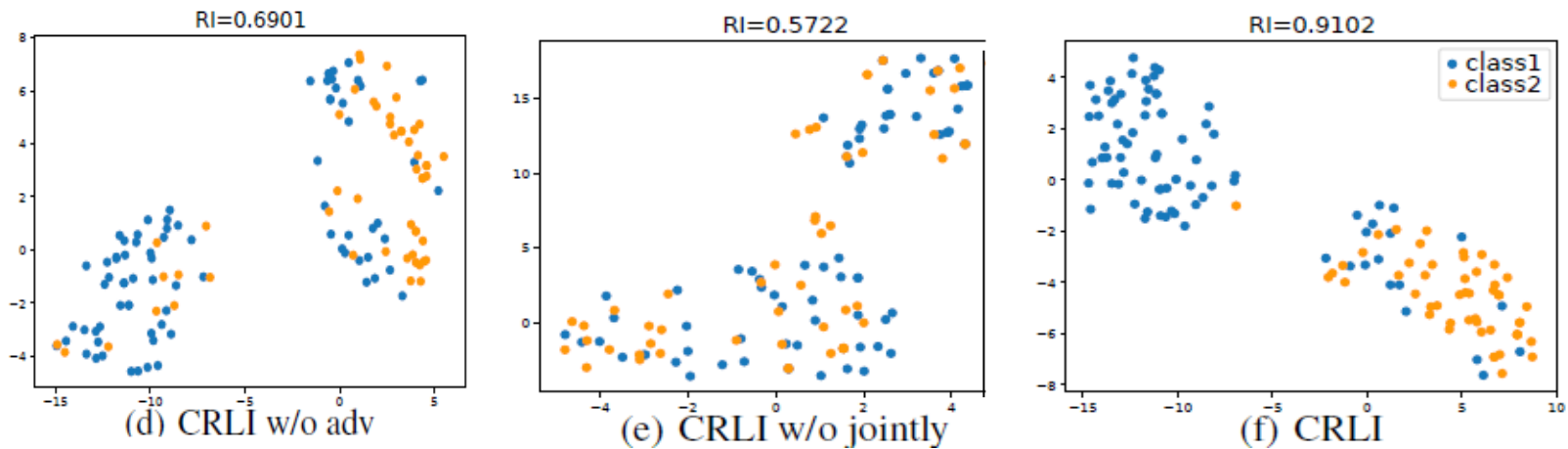
- ❖ CRLI 인코더에 사용한 BRITS 방법론을 통하여 결측치를 채운 후 DTCR, DTC에도 적용
- ❖ 베이스라인 대비 시각화를 해도 훨씬 더 잘 군집화가 된 것을 inter-class 들은 멀고 intra-class 들은 잘 모아져 있는 것으로 확인 가능



Learning Representations for Incomplete Time Series Clustering

Experiments

❖ Ablation 실험을 통해서도 adversarial과 jointly optimization, Lower branch 에 Linear Layer를 넣음으로써 성능이 올라간 것에 대한 성능 검증 함



-	w/o fd	w/o fc	w/o adv	w/o jointly	CRLI
Best	0	0	0	0	<u>8</u>
Rank	2.8	4.3	2.6	3.6	<u>1</u>
Ri	0.64	0.51	0.65	0.58	<u>0.81</u>
p-value	1.3e-2	6.6e-5	3.8e-3	3.0e-3	-

Table 2: Rand Index(RI) ablation study of CRLI

Conclusions

- ❖ 일반적인 클러스터링 알고리즘에 시계열 데이터 클러스터링 적용에 대한 어려움에 대해 알아보았으며, 이를 개선하기 위해 딥러닝 기반의 알고리즘들을 소개
- ❖ 어려운 점에 대해서 소개한 딥러닝 모델을 통하여 더 좋은 성능의 군집 분석을 확인
- ❖ One stage와 Two Stage 방법에 따른 딥러닝 모델에 대한 다양한 학습 방식에 대해 파악
- ❖ 결측치가 포함된 현실 데이터에 적용 가능한 딥러닝 모델에 대해 소개

Reference

- [1] Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). Brits: Bidirectional recurrent imputation for time series. Advances in neural information processing systems, 31
- [2] Madiraju, N. S. (2018). Deep temporal clustering: Fully unsupervised learning of time-domain features (Doctoral dissertation, Arizona State University).
- [3] Ma, Q., Zheng, J., Li, S., & Cottrell, G. W. (2019). Learning representations for time series clustering. Advances in neural information processing systems, 32
- [4] Ma, Q., Chen, C., Li, S., & Cottrell, G. W. (2021, May). Learning representations for incomplete time series clustering. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 10, pp. 8837-8846).
- [5] <https://www.heka.ai/en/our-publications/time-series-clustering>
- [6] <https://analyticsindiamag.com/all-you-need-to-know-about-time-series-clustering/>
- [7] <https://github.com/saeerur/dtc-tensorflow>
- [8] <https://github.com/HamzaG737/Deep-temporal-clustering>
- [9] <https://techy8855.tistory.com/22>
- [10] <https://stats.stackexchange.com/questions/395461/how-is-students-t-distribution-related-with-this-similarity-probability-equatio>