UCLA

Missing Data Imputation In Time Series

Yuji Gao, Lei Wang, Shenghui Xu

Overall Project Goals

- Many unexpected accidents will cause missing values of data
  - software crash
  - communication outage
  - privacy

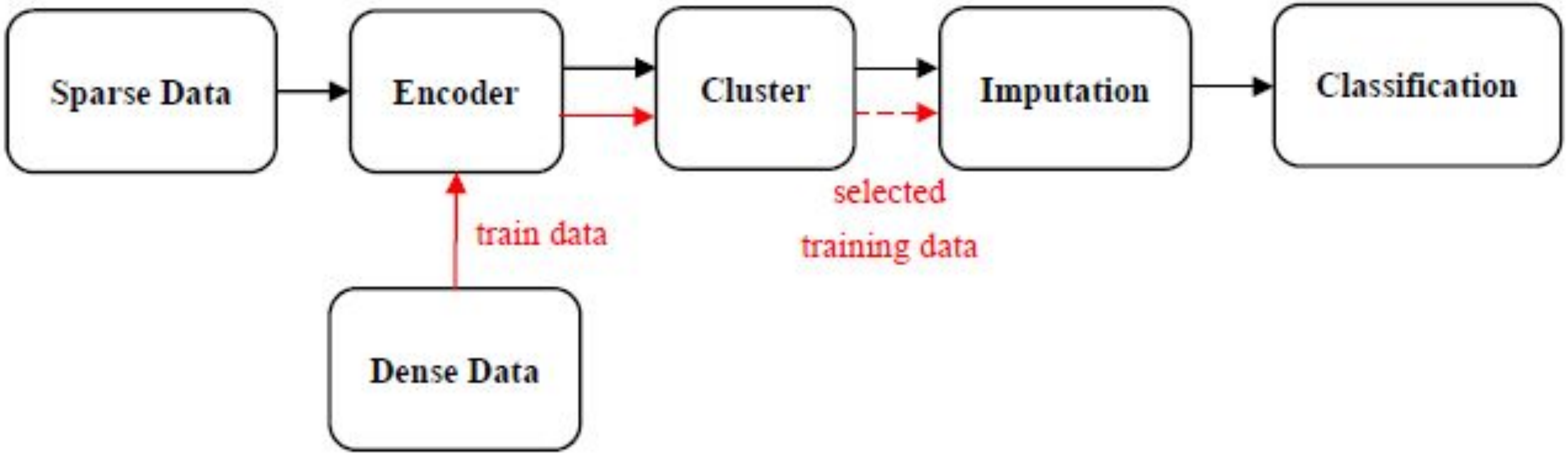
- Goal:
  - Improve the downstream classification accuracy when given sparse datasets with high natural missingness and few training data.
  - Explore a possible solution — a combination of encoder and clustering
  - Reduce computation time

Specific Aims

- **Principal Aim**: Realize the training and inference of an imputation model which works on serial data with missingness without background knowledge and history data.
- Train encoders with unrelated data and compress the dimensions of variables (from 48 to 4) to make it more clustering and classifying friendly
- Using K-means clustering to cluster and extract data that share the most similar pattern as training set
- Compare MRNN and BRITS models and finish training and imputation
- Evaluate the classification performance with the imputed data

UCLA

System Design



Current Methods
Traditional Methods

- Deletion methods
- Neighbor based methods
- Constraint based methods
- Statistical based methods

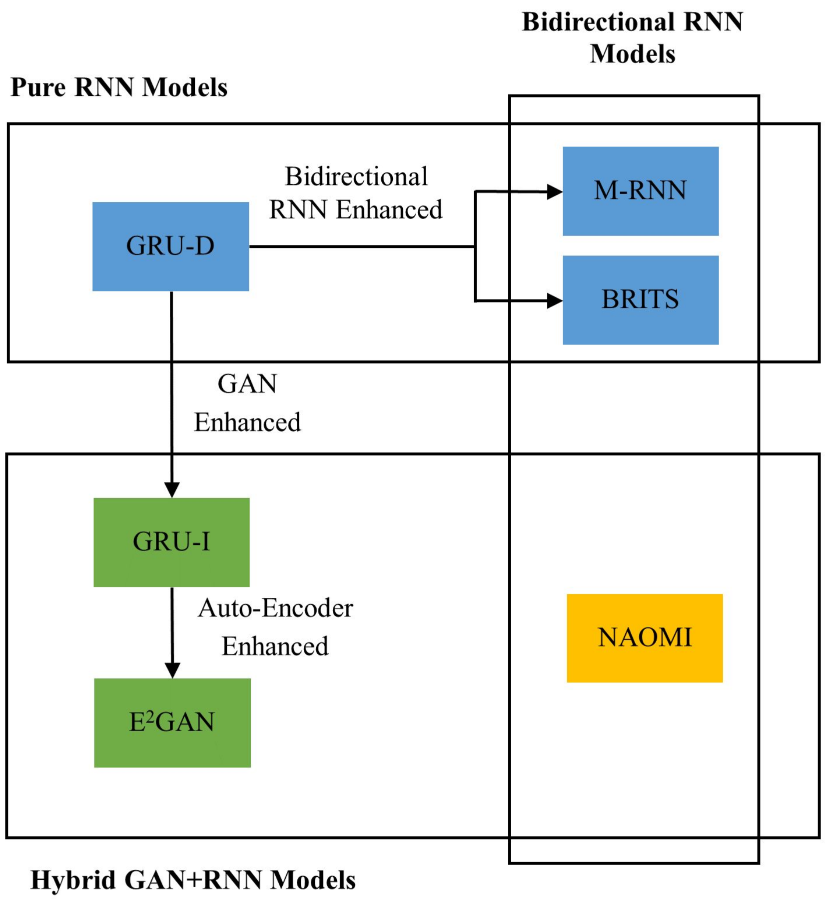
UCLA

Current Methods
Non-Neural Network Methods

- **Existing Methods**
  - ARIMA (Autoregressive Integrated Moving Average) based
    - **Smoothing by dth order difference**
    - **Auto Regression: pth order linear combination**
    - **Moving Average: qth order moving average of white Gaussian noise**
  - BN (Bayesian Netowrk) based
    - **Mixture (combination) of m Gaussian distribution**
    - **Estimate parameters with EM algo**
  - k-NN (k-Nearest Neighbour) based
    - **Find out kth nearest vectors, under a certain metric**
    - **Find the (weighted) average for imputation**
  - LSS (Local Least Squares) based
    - **Improved base on K-NN**
    - **Based on the k nearest vectors, apply linear regression to impute the missing value**
  - MCMC (Markov chain Monte Carlo) based
    - **Assum the sequence follows certain (parametric) distribution.**
    - **Learn the parameters with data of previous k states**
    - **Burn (abandom) the starting states, average the predictions correspoding to the rest parts**
  - PPCA (Probabilistic Principal Component Analysis) based
    - **Model the sequence with a latent Gaussian linear model, Gaussian noise and the mean value**

Current Methods
Neural Network Methods

- GRU-D
  - GRU (Gated Recurrent Unit)
- M-RNN & BRITS
  - Bidirectional RNN (Recurrent Neural Network)
- GRU-I & E^2GAN
  - GRU + GAN (Generative Adversarial Network)
- NAOMI
  - Bidirectional RNN + GAN

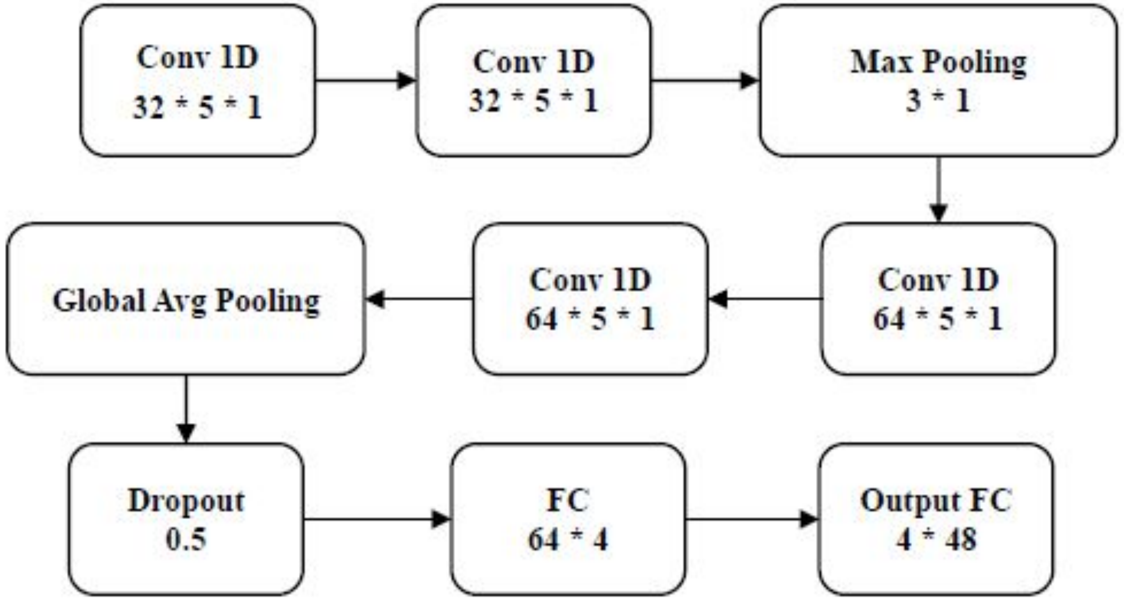


Novelty
Encoder & Cluster Model

- Propose a framework which extracts training data from different domains and scenarios to solve the shortage of training data
- Propose and evaluate several End-to-End encoder structures for feature selection from serial data
- Introduce clustering algorithm to accelerate the process of extraction of similar feature
- Explore different implementations of the imputation model and introduce downstream task to compare the performances

Technical Approach
Encoder Model

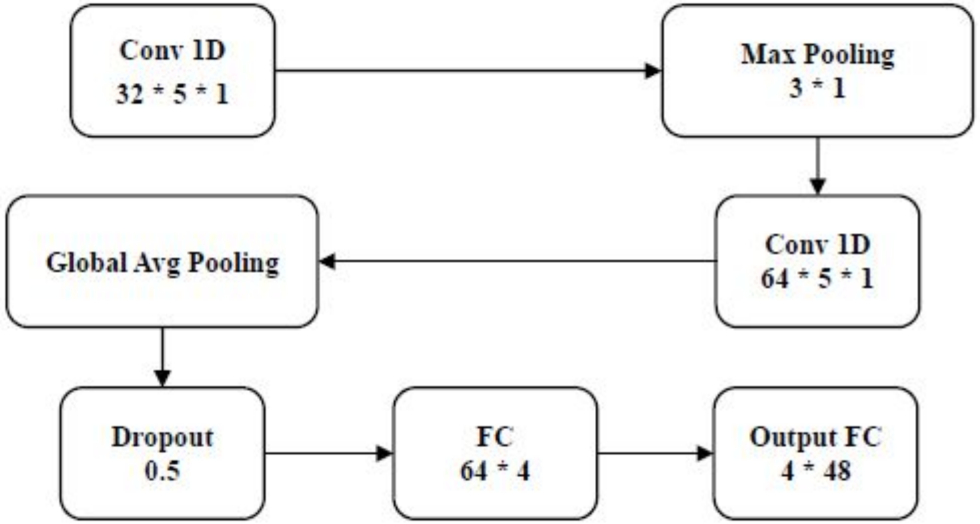
- Design a CNN1D–Based network
- Use the L2 loss to reconstruct the input univariate data from electricity and air quality datasets
- Save all the layers except the last output layer as an encoder model
- By inputting the PhysioNet data, we can get its 4-dimension representation

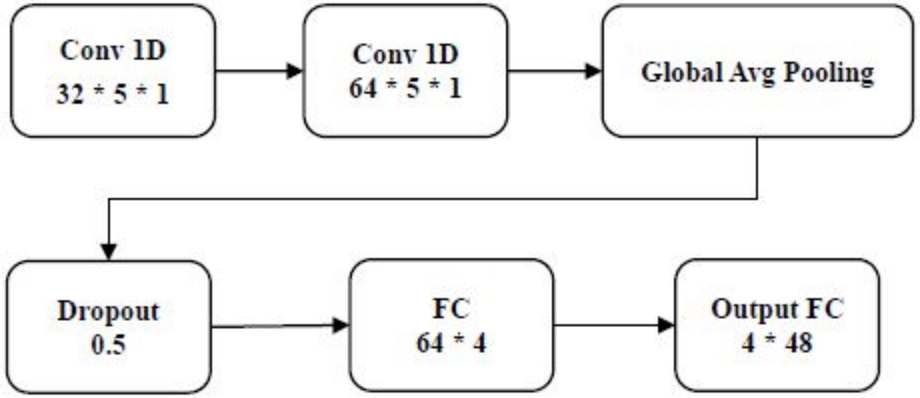


Technical Approach
Encoder Model

- We designed different encoders and compared their performances

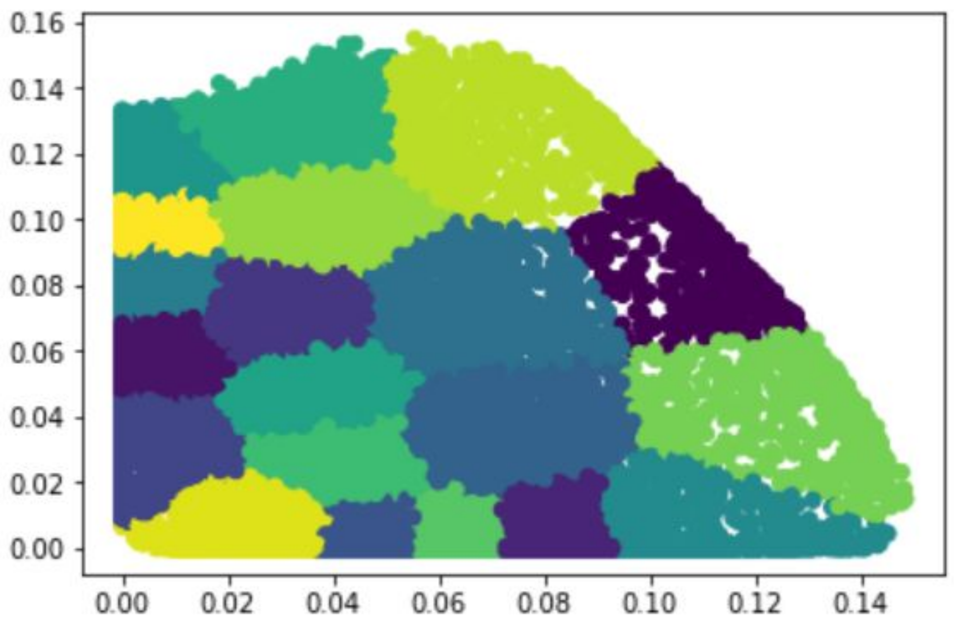
Encoder Models	Used	Left	Right
MSE	0.100	0.172	0.321

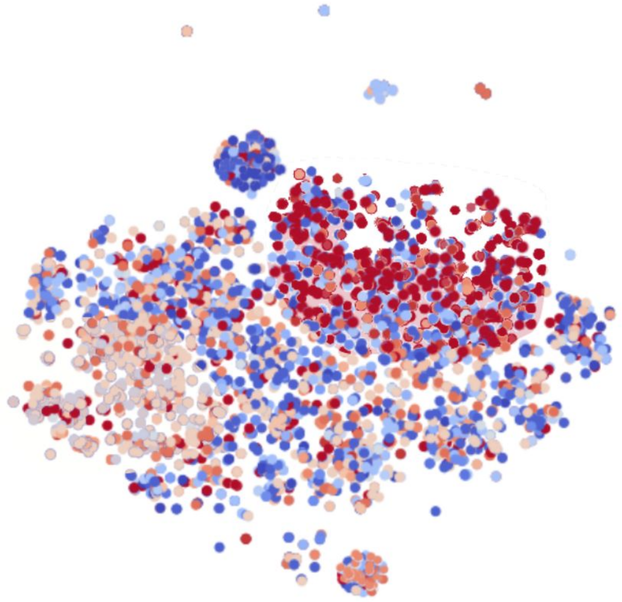




Technical Approach
K-means Clustering

- Cluster 20 classes using data from the encoded electricity and air quality datasets
- Find the top-3 nearest cluster centers of the encoded (low-dimensional) representation for each example of the PhysioNet dataset
- By voting we choose the most matched 3 cluster classes and use the original data of them as the final training data for M-RNN & BRITS





Technical Approach
Datasets & Metrics

- PhysioNet Challenge 2012
  - Medicine
  - Multivariate
  - 80% missing values
- Beijing Multi-Site Air-Quality
  - Weather
  - Multivariate
  - 1.6% missing values
- Electricity Load Diagrams
  - Electricity
  - Univariate
  - No missing values

- Imputation evaluation metrics
  - MAE(Mean Absolute Error)
  - RMSE(Root Mean Square Error)
  - MRE(Mean Relative Error)
- Downstream Classification Task
  - PhysioNet Challenge 2012 dataset (PhysioNet)
  - Predict mortality rate
- Other evaluation ideas
  - Early prediction capacity
  - Model scalability with growing data size
  - Time efficiency

https://www.physionet.org/content/challenge-2012
https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data
https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014

UCLA

Experiment Results

MAE/RMSE/MRE	Electricity	Air Quality
M-RNN	1.220 / 1.778 / 0.625	0.295 / 0.639 / 0.416
BRITS	0.843 / 1.284 / 0.430	0.156 / 0.535 / 0.216

- Baseline 1 (w/o): Directly split the PhysioNet data into the training and testing data of the imputation models
- Baseline 2 (random): Randomly select the same number of examples as our method from the other unrelated datasets
- Ours (encoder): Select the training data by our encoding and clustering algorithm
- Evaluation: classification accuracy, time consumption

Experiment Results

PhysioNet	ROC AUC	PR AUC	F1 Score	Time Consumption	Early Stop Setting
M-RNN (w/o)	0.822	0.463	0.303	-	-
M-RNN (random)	0.824	0.493	0.419	~1.5 h	5
M-RNN (encoder)	0.830	0.486	0.426	~1 h	30
BRITS (w/o)	0.819	0.461	0.359	-	-
BRITS (random)	0.840	0.514	0.405	~ 1.3 h	5
BRITS (encoder)	0.844	0.525	0.413	~ 1.2 h	30

Discussion

- Strengths
  - Solve the shortage of high quality data. Support Zero-Knowledge learning.
  - Better utilize training data with sliding window.
  - Filter unrelated data and select high quality data using K-means cluster
  - Reduce overhead of computation.
  - Improve the downstream classification accuracy with imputation techniques.

- Drawbacks
  - Require data preprocessing including truncating and encoding in advance
  - Rely heavily on the performance of encoder
  - Current cluster method is trivial so far (based on Euclidean Distance)

UCLA

Future work

- Generalization
  - Our data pool only contains two independent datasets and more datasets should be applied to test the generalization ability of this framework
  - Our encoder model needs fixed-length sequence and run-length encoding may be introduced
- Clustering Enhancement
  - The present model is to change the input sequence into a low-dimensional vector representative which is able to recover the original input sequence, and calculate distances using Euclidean Distance.
  - But Euclidean Distance may not represent the similarity between two sequences too well so we are looking at more advanced methods like metrics based learning.

Task Distribution

- Shenghui Xu: Data preprocessing with fixed windows, k-means clustering, voting algorithms for training data, GitHub repo
- Lei Wang: Encoder design and test, data processing for experiments, final slides, GitHub repo
- Yuji Gao: Different models building and experiments, imputation and classification implementation, model testing

Related Work

[1] **GRU-D:** Che, Zhengping, et al. "Recurrent neural networks for multivariate time series with missing values." Scientific reports 8.1 (2018): 1-12.

[2] **M-RNN:** Yoon, Jinsung, William R. Zame, and Mihaela van der Schaar. "Estimating missing data in temporal data streams using multi-directional recurrent neural networks." IEEE Transactions on Biomedical Engineering 66.5 (2018): 1477-1490.

[3] **BRITS:** Cao, Wei, et al. "Brits: Bidirectional recurrent imputation for time series." Advances in neural information processing systems 31 (2018).

[4] **GRU-I:** Luo, Yonghong, et al. "Multivariate time series imputation with generative adversarial networks." Advances in neural information processing systems 31 (2018).

[5] **E^2GAN:** Luo, Yonghong, et al. "E2gan: End-to-end generative adversarial network for multivariate time series imputation." Proceedings of the 28th international joint conference on artificial intelligence. AAAI Press, 2019.

[6] **NAOMI:** Liu, Yukai, et al. "NAOMI: Non-autoregressive multiresolution sequence imputation." Advances in neural information processing systems 32 (2019).

UCLA

Related Work

[7] **SAITS:** Du, Wenjie, David Côté, and Yan Liu. "SAITS: Self-Attention-based Imputation for Time Series." arXiv preprint arXiv:2202.08516 (2022).

[8] Zerveas, George, et al. "A transformer-based framework for multivariate time series representation learning." Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021.

[9] A. Goldberger, L. Amaral, L. Glass, Jeffrey M. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, and H. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation, 101 23:E215–20, 2000.

[10] [Li, Yuebiao, Zhiheng Li, and Li Li]”Missing traffic data: comparison of imputation methods.” IET Intelligent Transport Systems 8.1 (2014): 51-57.

[11] [Ahmed, Mohammed S., and Allen R. Cook]Analysis of freeway traffic time-series data by using Box-Jenkins techniques. No. 722. 1979.

[12] Lipton, Zachary C., David C. Kale, and Randall Wetzel. "Modeling missing data in clinical time series with rnns." Machine Learning for Healthcare 56 (2016).

Related Work

[13] [Troyanskaya, Olga, et al.] “Missing value estimation methods for DNA microarrays.” Bioinformatics 17.6 (2001): 520-525.

[14] [Kim, Hyunsoo, Gene H. Golub, and Haesun Park.] “Missing value estimation for DNA microarray gene expression data: local least squares imputation.” Bioinformatics 21.2 (2005): 187-198.

[15] [Ni, D., Leonard II, J.D.] “Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data”, Transp. Res. Rec., 2005, 1935, (1), pp. 57–67

[16] [Gilks, W.R., Richardson, S., Spiegelhalter, D.J.]”Markov chain Monte Carlo in practice” (Chapman & Hall, London, 1996)

[17] [Tipping, M.E., Bishop, C.M.]”Mixtures of probabilistic principal component analyzers”, Neural Comput., 1999, 11, (2), pp. 443–482

[18] [Lima, Juan-Fernando, Patricia Ortega-Chasi, and Marcos Orellana Cordero]”A novel approach to detect missing values patterns in time series data.” Conference on Information Technologies and Communication of Ecuador. Springer, Cham, 2019.

[19] [Dong, Y., Peng, C.Y.J.]Principled missing data methods for researchers. Springer-Plus 2(1), 222 (2013).

Thank you!