Developing a generative adversarial network-based method for longitudinal microbiome data imputation

CS6824 project proposal

Presenter: Joung Min Choi

Background and Motivation

- **Microbiome studies**
  - The collection of microorganisms living in a certain environment
  - Key role in complex disease such as obesity, diabetes, cancer, and allergy outcomes
  - Potential as biomarkers for disease diagnosis or as therapeutic targets for treatment

- **Longitudinal studies of microbiome**
  - Microbiome can be altered over time by infections or medical interventions
  - Providing insights into the changes of microbiome composition over time and the association with disease outcomes

  - **Major challenges**

    : Limited sample size due to the uneven number of timepoints along the longitudinal timeline of different subjects

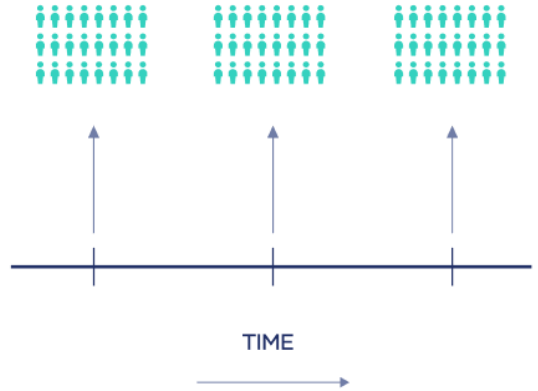

Figure reference: https://www.scribbr.com/methodology/longitudinal-study/

Background and Motivation

- **Generative adversarial network (GAN)**
  - Widely adopted in various fields to address the lack of data issue
  - Data augmentation framework to improve classification tasks by reducing overfitting
  - Recently, being utilized for imputation of missing values for a multivariate time-series with RNN

- **Applications of GAN in microbiome study**
  - MB-GAN (2020): a simulation framework for microbiome data based on GAN
  - DeepBioGen (2021): a data augmentation procedure that characterizes visual patterns of sequencing profiles

  ✓ **But still, the presented methods only simulate single time point microbiome data**

  ✓ **Data imputation for longitudinal microbiome data have not been addressed, yet.**

Proposed approach

- **A deep learning-based method for longitudinal microbiome data imputation based on GAN**

  - Help the researcher to utilize the incomplete longitudinal microbiome datasets
  - Promote the future microbiome analysis
  - Improve the accuracy for predicting the disease outcomes

Research Plan

1. **Data collection and preprocessing**
   - **Public longitudinal microbiome datasets**
     1) DIABIMMUNE three-country cohort dataset: 16S rRNA
     2) BONUSC-DF dataset: shotgun sequencing dataset

   - **Preprocessing**
     - Species-level relative abundance profiles

       : The percentages of the species in the total observed species

     - Centered log-ratio transformation

Research Plan

2. Developing the GAN-based model for longitudinal data imputation

1) Implementation of the biRNN-based GAN to generate samples by training the temporal relations between the observations

- Related papers for GAN-based time-series data imputation
  - "Multivariate Time Series Imputation with Generative Adversarial Networks" (*NeurIPS*, 2018)
  - "E2gan: End-to-end generative adversarial network for multivariate time series imputation" (*AAAI*, 2019)
  - "Time-series imputation and prediction with bi-directional generative adversarial networks" (*arXiv*, 2020)
  - "Missing value imputation in multivariate time series with end-to-end generative adversarial networks" (*Information Sciences*, 2021)

Research Plan

2) Incorporating the taxonomy relationship based on the phylogenetic tree

- Related papers for encoding the phylogenetic information by CNN
  - *"*PhyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data*"* (*Bioinformatics*, 2021)
  - "A novel deep learning method for predictive modeling of microbiome data" (*Briefings in Bioinformatics*, 2021)

3) Optimization of the hyperparameters for the model based on the grid search

- Number of hidden nodes, hidden layers, learning rate and learning epoch

Research Plan

3. Performance evaluation of our proposed model

- **Baseline methods**

1) Simple imputation methods

- Mean, Median

2) Traditional time-series imputation

- Linear curve fitting, Cubic curve fitting
- Moving-window-based imputation

3) Widely-used imputation method for longitudinal dataset

- Multiple imputation by chained equation (MICE)
- Last Observation Carried Forward (LOCF)

Research Plan

- **Experiments**
  - 10-fold cross-validation using test dataset as missing data
  - Evaluation of performance changes by increasing the missing rate
  - Validation of performance improvement for predicting the disease outcome when training the model by adding the incomplete dataset having missing samples for some time points with imputation

- **Performance evaluation metric**
  - Mean absolute error (MAE)
  - Classification accuracy and area under the ROC curve (AUC)

Project timeline

- Data collection and preprocessing (~ March 12th) (DONE)
- Implementation of the basic architecture of the biRNN-based GAN model (~ March 12th) (DONE)
- Add feature extraction module to GAN to incorporate the phylogenetic tree information (~ March 20th)
- Hyperparameter optimization (~ March 31th)
- Performance evaluation (~ April 13th)