Statistical hazard-harm control in health institutions

Mid term evaluation seminar

Chi Zhang

2019/05/02

Outline

Progress overview

Paper 1

Interveneable predictions of hospital acquired infection via a hierarchical lasso procedure using Electronic Health Records

Paper 2

Feature learning on heterogeneous temporal EHR data

Future works

Progress overview

Paper 1

- Simulation, (most part of) manuscript done
- Need a real data example to be complete

Paper 2

- Started: Feb 2019 (3 months in)
- Open data, preparation work done
- Concept: formed
- Analysis: started on small sample

Time remaining: 1 year 3 months

Paper 1

Interveneable predictions of hospital acquired infection via a hierarchical lasso procedure using Electronic Health Records

Motivation:

- a framework based on an interpretable model to predict an outcome, tradeoff between interpretabilitty and predictivity
- Outcome: HAI on a certain day in the future
  - patients with pneumonia, urinary tract infection, etc
- Data type: time series predictors and response
  - Lab tests: positive results, high white blood cell count, …
  - Patient Characteristics: BMI, fever, ...
  - Procedures and medication: antibiotics, vasopressor drugs, ...
  - Staff: specialist nurses, working overtime, ...

2 steps approach

step 1: variable selection

select important time series covariates via hierarchical lasso penalty

$$min_\beta \sum_{t=1}^{T} ||y^{(t)} - \sum_{j=1}^{p}\sum_{l=0}^{L} \beta_j^{(l)} x_j^{(t-l)}||_2^2 + \lambda \sum_{j=1}^{p}\sum_{l=0}^{L} ||\beta_j^{(l)}||_1$$

The fitted model prediction from the hierarchical variable selection is

$$\hat{y}^{(t)} = \sum_{j=1}^{p}\sum_{l=0}^{L} \hat{\beta}_j^{(l)} x_j^{(t-l)}$$

step 2: prediction improvement

refit on residuals and historical lags of the response to improve prediction

Denote the residuals as $r^{(t)} = y^{(t)} - \hat{y}^{(t)}$

$$\tilde{y}^{(t)} = \sum_{k=1}^{K} \hat{\phi}^{(k)} y^{(t-k)} + \sum_{j=1}^{p} \sum_{l=0}^{L} \hat{\beta}_j^{(l)} x_j^{(t-l)}$$

Intervention

Change the value of certain variables at time t.

Paper 2

Feature learning on heterogeneous temporal EHR data

Background

Data: MIMIC-III Critical Care Database (Medical Information Mart for Intensive Care III)

Challenges of EHR data

- heterogeneity: multiple sources; unstructured text/numeric measurements
- unequal length
- unevenly sampled

Representatiton learning via

- SAX: symbolic aggregate approximation
- Dynamic time warping: time series similarity
- Tensor decomposition: extract patient latent feature for classification task

Plan for future works

Paper 2 (priority)

- finish implementing our method on small cohort
- compare with other method (i.e. LSTM Autoencoder by Suresh2018 paper)

Paper 1

Paper 3

(some of my interests:)

- open dataset: eICU (50 centers)
- continue feature learning, patient phenotyping and tensor
- privacy preserving ML for EHR data
- software development from paper 2