Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Siddharth Madhavan/Jhansi Badarvada

Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Problem Statement

With a rapid increase in the adoption of Electronic Health Records(EHRs), increasing amount of research is being devoted to applying machine learning/deep learning methods to EHRs for various clinical purposes and has exposed the challenges of the accessibility of EHRs.

The paper addresses some of the challenges faced by researchers in accessing the popular, public and free EHR dataset MIMIC IV

- The dataset is raw and needs domain specific knowledge to interpret the data
- Lack of standardized pre-processing steps lead to
  - Limited reproducibility
  - Limited ability to compare results of similar research work
  - Can lead to unreliable, biased or harmful designs

Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Original Paper - General Approach

- The paper presents a reproducible and customizable pipeline to extract, clean, and pre-process data available in the fourth version of the MIMIC dataset
  - Data extraction
    - Supports data extraction cohorts for MIMIC ICU/Non-ICU for prediction tasks in-hospital mortality, readmission, LOS, and phenotype prediction related to health conditions heart failure, chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), and coronary artery disease(CAD)
  - Data Preprocessing
    - Supports outlier removal, a grouping of medical features using standard coding systems, and filtering of data by time-series length based on user preferences to generate a personalized patient cohort
  - Predictive modeling
    - Supports machine learning and deep learning models for performing prediction tasks
  - Model evaluation
    - Supports standard evaluation using metrics (such as AUROC, AUPRC, accuracy, precision, recall, NPV), Fairness evaluation, and model calibration

Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Claims

Results claimed by the Original Paper:

- Pipeline is highly configurable and provides users with many options to define customized cohorts by allowing for feature selection options and other user-defined preprocessing steps
- Pipeline not only addresses the issue of reproducibility (by recording all design choices) but also promotes further research using different cohorts for prediction tasks.
- Pipeline processed data is compatible to use with several Machine Learning and Deep Learning Models

Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Scope of Reproducibility

The following claims will be verified

- Pipeline can reproduce the same output with the same inputs.
- Data from the pipeline can be used in multiple machine learning models
- Data from the pipeline can be used in multiple deep learning models

Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Testing Approach

Data Sourcing

- ➢ Secured access to MIMIC IV version 1.0 & 2.0 data from Physionet.org
- ➢ Downloaded a demo dataset and a full dataset from Physionet.org
- ➢ Extracted a subset of the MIMIC IV data for 500 patients (limited dataset)

Experiments were conducted with 3 datasets

- ➢ Demo dataset with 26 patients (✘ Incomplete Test)
- ➢ Subset of MIMIC IV data with 500 patients (✔ Successful Test)
- ➢ Complete MIMIC IV dataset with over 60K patients (✘ Incomplete Test)

Pipeline Testing

- ➢ Extracted cohorts with the same inputs multiple times
- ➢ Compared the results of each run using Beyond Compare to verify outputs was identical for identical inputs

Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Testing Approach - contd….

Machine Learning Model Testing

- Ran multiple machine models on the cohorts extracted during pipeline testing
- Compared the performance metrics across the model

Deep learning Model Testing

- Ran multiple deep learning models on the cohorts extracted during pipeline testing
- Compared the performance metrics across the model

Ablation Testing

- Ran Ablation testing by removing one of the fully connected layers from the LSTM model on the cohorts extracted during pipeline testing
- Compared the performance of the modified LSTM with the original model and observed little to no improvement

Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Test Results

	MIMIC IV Dataset		
Reproducibility task	**26 Patients**	**500 Patients**	**60K Patients**
Pipeline can reproduce the same output with the same inputs	✓	✓	✓
Data from the pipeline can be used in multiple machine learning models	✓	✓	✗
Data from the pipeline can be used in multiple deep learning models	✗	✓	✗
Ablation: Reduction in LSTM model’s depth	✗	✓	✗

Reproducibility of An Extensive Data Processing Pipeline for MIMIC-IV

Conclusions

What worked well

- ➢ Easy to follow instructions for Sourcing MIMIC IV data and running the Pipeline
- ➢ Pipeline performance was good for smaller datasets

What did not work well

- ➢ Pipeline performance was extremely slow for a complete MIMIC IV dataset with over 60K patients
- ➢ Few bugs were found which could have been prevented by having unit/integration testing

Thank You