Data Science Project: Epigenetic Classifier for Twin Zygosity

Zexi Liu, Haoze Xia, Atefeh Zamani, Ni Zhang, Tianyu Zhou

Supervised by:

Dr. Jiadong Mao and Prof. Michael Kirley

Group Members

Zexi Liu, Atefeh Zamani, Haoze Xia, Ni Zhang, Tianyu Zhou

Duties:

Literature review, Data preprocessing, Model development, Documentation, …

Host Organisation

Twins Research Centre of Epidemiology and Biostatistics UoM

- Dr. Shuai Li and Prof. John Hopper

What is the Data Science Problem?

- Data linkage of medical data gives more opportunity for research in recent years

- The organisation wants to explore machine learning techniques to this sudden available data, and this project is to explore options.

Overview

What is zygosity of Twins and why are they important?

- Monozygotic twin (MZ) or identical twin
- Dizygotic twin (DZ) or fraternal twin

Importance:

- Cheaper classification method compared to traditional questionnaires from hospitals
- It helps better understand disorders that’s correlated with twins, e.g. spine disorders.

Problem Domain and Research Gap

- Domain - Methylation Data
  - Measured on DNA Sequence
  - The scale can easily goes up to 400 K

- Identify Research Gap - Semester 1
  - State of arts models van Dongen et a. [2021]
    - yields low performance on Datasets they used but not report in their writing
    - yields low performance on Datasets our client provided

Aim

1. Problem domain: Binary classifier predict zygosity using DNA Methylation Data.

2. We proposed a machine learning based algorithm that outperformed state-of-arts models van Dongen et a. [2021] in the same problem domain.

Key Innovations of Proposed Algorithms

Innovation 1 - Use more training data => Model more robust

Innovation 2 - Use feature selection => More cost efficient for real biologist

Innovation 3 - Use more recent machine learning algorithms => More accurate

Key Challenges

- Biological terms and techniques
- Data preprocessing
  - Different data format
  - Missing and complex metadata
- Scalability
  - Big datasets (450K columns)
  - Variable selection
  - HPC - Spartan
- Machine learning algorithms research and test

Project Pipeline

- Data preprocessing
  - Missing values; preparing training & testing data
- Data engineering
  - Feature selection using random forest / logistic regression
- Machine learning
  - SVM, Random Forests, Logistic Regression etc.
- Ensemble learning
  - Stacking & Voting
- Results

Datasets

	N Total	N MZ Twins	N DZ Twins	N XZ Twins	N Family Members	Tissue
E-Risk	1658	860	608	190	-	Blood
AMDTSS	479	132	132	-	215	Blood
BSGS	614	135	223	-	256	Blood
Denmark	180	94	86	-	-	Blood
E-MTAB	648	240	408	-	-	Fat and Skin

Table 2. Data sets provided with client to train machine learning classifiers for twin zygosity

Data Preprocessing

1. Remove samples of family members
2. Remove invalid data from the Denmark dataset
3. Replace missing values by their column means
4. Concatenate four datasets as training and the left one as testing - Innovation 1
5. Further split training data into training (75%) & validation (25%)

Repeat	Test Data	Train Data
1	E-MTAB	E-Risk, BSGS, Denmark, AMDTSS
2	AMDTSS	E-Risk, BSGS, Denmark, E-MTAB
3	Denmark	E-Risk, BSGS, AMDTSS, E-MTAB
4	BSGS	E-Risk, AMDTSS, E-MTAB, Denmark
5	E-Risk	BSGS, AMDTSS, E-MTAB, Denmark

Table 3. Concatenating four datasets as training and the left one as testing

Data Engineering

Variable Selection - Innovation 2

- Methods
  - Random Forest
  - Logistic regression
- Original number of variables: 833
- Current number of variables: ~300

Classical Machine Learning Models

- Logistic Regression (LR)
- Multinomial Naive Bayes (MNB)
- Support Vector Machine (SVM)
- Random Forest (RF)
- Gradient Boosting (GB)

Ensemble Learning

- Stacking: LR + MNB + SVM + RF + GB
- Voting: LR + SVM + RF + GB

Results - New Model

Innovation 3

AUC	NC - Classifier 13	Stacking - rf	Stacking - lr	Voting - rf	Voting - lr
E-Risk	0.728	0.7207	0.7189	0.7338	0.7235
BSGS	0.784	0.8068	0.8201	0.8076	0.8219
Denmark	0.544	0.6546	0.6379	0.7227	0.6734
E-MTAB	0.522	0.6967	0.6842	0.7202	0.6976
AMDTSS	0.648	0.7013	0.723	0.6889	0.7167
Agerage	0.6451	0.7160	0.7168	0.7346	0.7266

Table 4. Comparison between new trained classifier with original classifier

Conclusion

- Innovation 1: Using 4 datasets concatenated as training and left 1 dataset as testing => Better generalizability
- Innovation 2: Number of variables is reduced using RF (from 833 to ~300)
- Innovation 3: Ensemble learning - voting classifier
- Best model found: variable selection using random forest + voting classifier
- AUC increased from 0.65 to 0.73, by ~13%
- Future work - Spartan: select variables from the original raw dataset with 450k columns (CPGs), then test the result

Tried Models

The performance of the following methods had been checked but with no further investigation:

- Adaboosting - Low AUC
- Deep Learning - Low AUC
- Partial Least Square (PLS) - Using variable selection instead

Thanks for watching!