README

1. Preparation(only need to be take once)

Step1:

Get the **home path** of your program. For example, my file is unzipped at `C:/Users/NORTH/Dropbox/SemiSupervised_RiskPrediction/`, then the corresponding home file directory is `C:/Users/NORTH/Dropbox` **(with no '/' at last)**. Write it down, and we'll need it later.

Step2:

Read important file: `Rscript_run.sh` . In this step, check every file mentioned in `Rscript_run.sh` and replace the `mdir` with the **home path** you get in Step:1

```
* Utilities/get_fold_new.R
* Utilities/convert_data_NonLongitudinal.R
* Method/run_DeepHit.R
* Evaluation/eval_DeepHit.R
* Method/Method_AE/train_AE2.py
* Real/plot_main_benchmark_tune.R
```

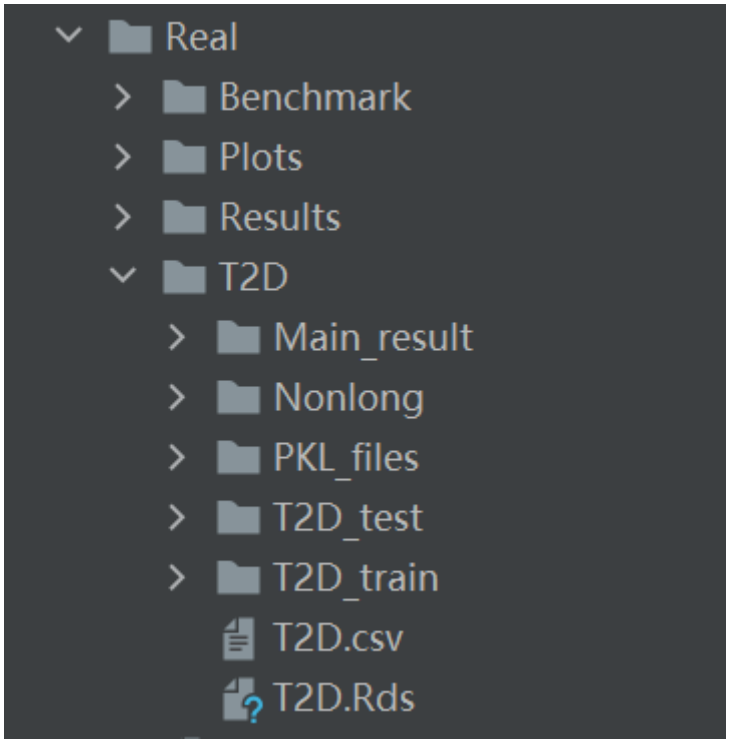
Input Structure

Original Data File:

2 files at path `Real/T2D/`, respectively named `{phe.nm}.csv` and `{phe.nm}.Rds`, {phe.nm} is the argument input by you.

E.G.

phe.nm = T2D



I've left the 2 files(T2D.csv and T2D.Rds) in the zipped document, study them to get more detailed information of input structure before utilizing the model

Data Structure

1. csv file:

```
"","ID","Y","T","X.1","X.2","X.3","X.4","X.5","X.6","X.7","X.8","X.9","X.10","S","MAP"
"1",1,0,0,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"2",1,0,1,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"3",1,0,2,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"4",1,0,3,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"5",1,0,4,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"6",1,0,5,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"7",1,0,6,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"8",1,0,7,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"9",1,0,8,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0.
"10",1,0,9,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-0
"11",1,0,10,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-
"12",1,0,11,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-
"13",1,0,12,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-
"14",1,0,13,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-
"15",1,0,14,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-
"16",1,0,15,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-
"17",1,0,16,-0.380166666666667,-0.382633333333333,0.622166666666667,-0.374966666666667,0.497366666666667,-0.5009,-0.5016,-
```

v1	ID	Y	T	X.i	S	MAP
ordinal number	patients' ID	current status	current time	i_th dimension of feature	Surrogate	MAP result

```
single_patient = data.frame(ID = ID, Y = Y, T = t, X = X_feature, S = S_t, MAP = MAP_t)
single_patient = as.matrix(single_patient)
return(single_patient)
```

2. Rds file

X, T(event time), C(censoring), S(surrogate), I(final status)

```
saveRDS(list(X=X,T=T,C=C,S=S,I=I,
             #P=P,
             MAP=MAP),file = dataname)
```

Arguments denote:

```
# Switch to target directory
cd C:\Users\NORTH\Dropbox\REAL\SemiSupervised_RiskPrediction

#-------------------------Split data to train set and test set

RScript Utilities/get_fold_new.R  ${total_number} ${labeled_num} ${num_train} ${num_test} ${phe.nm} ${observation_time}
# 1.total_number 2.labeled_num 3.num_train 4. num_test 5.phe.nm 6. observation time
# e.g. RScript Utilities/get_fold_new.R 30000 1000 300 170 T2D 5

#-------------------------Train and Evaluate DeepHit
RScript Utilities/convert_data_NonLongitudinal.R ${phe.nm} ${observation_time} ${maximum_time}
# 1.phe.nm 2. observation time(baseline time)(5 for simulation) 3. maximum time
# e.g. RScript Utilities/convert_data_NonLongitudinal.R T2D 5 30

RScript Method/run_DeepHit.R ${num_train} ${phe.nm}
# 1.num_train   2. phe.nm
# e.g. RScript Method/run_DeepHit.R 300 T2D

RScript Evaluation/eval_DeepHit.R ${num_train} ${phe.nm} ${observation_time} ${maximum_time}
#  1.num_labels  2. phe.nm 3. observation_time 4. maximum time
# e.g. RScript Evaluation/eval_DeepHit.R 300 T2D 5 30

#-------------------------Train and Evaluate method
python Method/Method_AE/train_AE2.py --phe_nm='' --epochs=
# e.g python Method/Method_AE/train_AE2.py --phe_nm=T2D --epochs=4

#-------------------------Plot the result
RScript Real/plot_main_benchmark_tune.R ${num_train} ${phe.nm} ${observation_time} ${maximum_time}
#  1.num_labels 2. phe.nm 3. observation_time 4. maximum time
# e.g. RScript Real/plot_main_benchmark_tune.R 300 T2D 5 30
```

- `total_number`: total number of patients(ID)
- `labeled_number`: the number of labeled samples
- `num_train:` number of samples for training
- `num_test:` number of samples for testing
- `phe.nm`: the name of the phenotypes(**must be correspond to the file name**)
- `obervation_time` : the observation time **(baseline time)** *e.g. 5 for SimData*
- `maximum_time`: the maximum time in data *e.g. 30 for SimData*
- `epochs:` training epochs for Main_method

You can run the *example* codes provided under each command line to run the full process on provided T2D data

Run the Program

Run `Rscript_run.sh` to run the whole process! Remember to `cd` to project workspace at first.

```
# Switch to target directory
cd C:\Users\NORTH\Dropbox\REAL\SemiSupervised_RiskPrediction

#-------------------------Split data to train set and test set

RScript Utilities/get_fold_new.R  ${total_number} ${labeled_num} ${num_train} ${num_test} ${phe.nm} ${observation_time}
# 1.total_number 2.labeled_num 3.num_train 4. num_test 5.phe.nm 6. observation time
# e.g. RScript Utilities/get_fold_new.R 30000 1000 300 170 T2D 5

#-------------------------Train and Evaluate DeepHit
RScript Utilities/convert_data_NonLongitudinal.R ${phe.nm} ${observation_time} ${maximum_time}
# 1.phe.nm 2. observation time(baseline time)(5 for simulation) 3. maximum time
# e.g. RScript Utilities/convert_data_NonLongitudinal.R T2D 5 30

RScript Method/run_DeepHit.R ${num_train} ${phe.nm}
# 1.num_train   2. phe.nm
# e.g. RScript Method/run_DeepHit.R 300 T2D

RScript Evaluation/eval_DeepHit.R ${num_train} ${phe.nm} ${observation_time} ${maximum_time}
#  1.num_labels  2. phe.nm 3. observation_time 4. maximum time
# e.g. RScript Evaluation/eval_DeepHit.R 300 T2D 5 30

#-------------------------Train and Evaluate method
python Method/Method_AE/train_AE2.py --phe_nm='' --epochs=
# e.g python Method/Method_AE/train_AE2.py --phe_nm=T2D --epochs=4

#-------------------------Plot the result
RScript Real/plot_main_benchmark_tune.R ${num_train} ${phe.nm} ${observation_time} ${maximum_time}
#  1.num_labels 2. phe.nm 3. observation_time 4. maximum time
# e.g. RScript Real/plot_main_benchmark_tune.R 300 T2D 5 30
```

You can also run the `Rscript_example.sh` to run the process on example data

Update V_2 10/31/22

Promote of home path setting

- now you only need to update the `mdir` in 2 files:
  - `set_mdir.R`
  - `train_AE2.py`