EpiTEAmDNA version 1.00

Authors: Fengfeng Zhou, Fei Li
Update: 2022-09-13
Email: FengfengZhou@gmail.com

Description:

EpiTEAmDNA is a cross-species predictor for multi epigenetic DNA modifications sites identification.

Develop environment:

System: WIN10
computer memory: 16G
GPU: 3060 12G
CUDA Version: 11.5

Installation:

Package	Version
Python	3.6.13
gensim	3.8.3
numpy	1.19.2
pandas	1.1.5
scikit-learn	0.24.2
scipy	1.5.4
tensorflow-gpu	2.4.1
xgboost	1.4.2

Software code structure

Folder of file name	description
config	Config.py contains parameters that control the training process

data	Original data files and tools which can read and format them.
dl	The DL-part of iDNA-TE
ensemble1	The main body of iDNA-TE
fs	Feature selection tool
prepare	Process original data
tools	Evaluation metrics
Main_1.py	This is the entrance to the program

Config:

You can change parameters in config/config.py to train models.

Parameter name	description	Default value
device	the device used to train model, 'cpu' or 'gpu'	'gpu'
is_feature_selection	whether to perform feature selection for ML-based part model	'True'
load_global_pretrain_model	whether to load a pre-trained model for the DL-based model	'False
global_model_save_path	the path of the pre-trained model	None
model_save_path	The path to save the model	None
batch_size	Number of samples send to DL-based model each batch	256
learning_rate	Learning rate of DL-based model	1e-2
num_epochs	epochs	500
patience	Early stopping	50 (epochs not_improvement)

Format of input data

The training set and test set are pandas.DataFrame with 2 columns (label, seq). The optional value of column 'label' is 1(Methylation) or 0(non- Methylation). and the column 'seq' is a 41bp sequence containing 4 bases 'ACGT'.

```
>N_1
CATCGTTGTATTGATGACAACTTATTGAGCGCTGCGCTTGC
>N_2
GCGGGTATTAGGTCGATATCCTGTAGTTACTCTTTTGTCGC
>N_3
AATCATTAAGGCCGGACGACCGTAAGGAGGGTGGTAATTAC
>N_4
ATAAAAGAAGTCCCCGTCTACAGGTAAGATTTAGGTGGAAT
>N_5
TCTGTGACAATCCCAAAATACGCTAATGCTGGCGAGCCACC
>N_6
TTTGTACGACGCTTTCCGGCCTACGGGGCGTCTCCCCACTT
>N_7
TCACTGTCTCAACTCTCTGTCACGGTGGTGCAACGCGCCCC
>N_8
```

Train and test model

Before running main_1.py, you may first set config.output_save_path.

If you want to use transfer learning, set config.load_global_pretrain_model=True. You can change the pre-trained model to yours. By running main.py, it will output the prediction metrics including ACC, SN, SP, MCC, AUC, and F1-score.