DNNGP 模型使用手册

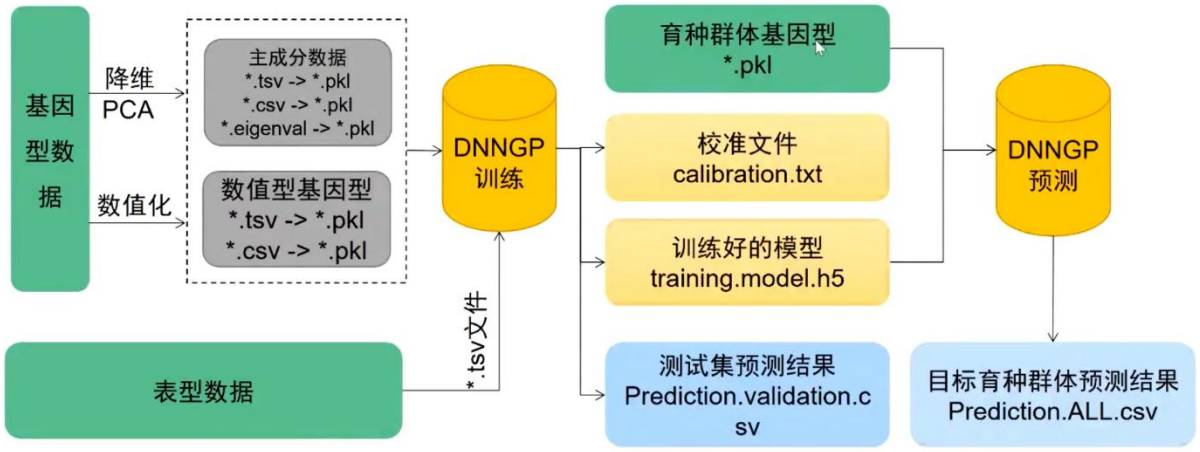
一、DNNGP 概况与框架结构

- **DNNGP:**

一种可用于全基因组预测，接受多组学数据并预测植物和动物的表型的深度学习模型。代码基于深度学习理念，使用Python 3.9及TensorFlow 2.6编写。支持GPU加速计算。

- 运算平台：

Windows、Linux、Mac



二、搭建运行环境

1.Windows 系统下的环境

(1) 下载项目地址：https://github.com/AIBreeding/DNNGP

(2) 运行 DNNGP 首先需要搭建运行环境:

首先安装：Miniconda（https://docs.conda.io/en/latest/miniconda.html），并将其添加进系统环境。（安装 miniconda 时可以勾选，使用 GUI 所必须的。）

Windows 平台下可以直接双击 **Double_click_me_first.bat** 文件实现一键搭建环境。若一键

搭建环境失败，则使用以下命令搭建运行环境：

- conda create -n DNNGP3 python=3.9.16
- conda activate DNNGP3
- cd dnngp
- conda install --yes --file requirements.txt
- conda install -c nvidia cuda-nvcc
- pip install framework-reproducibility==0.4.0

2.Linux 系统下的环境搭建

（1）下载项目地址：https://github.com/AIBreeding/DNNGP

（2）运行 DNNGP 首先需要搭建运行环境:

首先安装：Miniconda（https://docs.conda.io/en/latest/miniconda.html），并将其添加进系统环境。（安装 miniconda 时可以勾选，使用 GUI 所必须的。）

Linux 平台下可以直接通过 bash bash_click_me_first.sh 命令实现一键搭建环境。若一键搭建环境失败，则使用以下命令搭建运行环境：

- conda create -n DNNGP3 python=3.9.16
- conda activate DNNGP3
- cd dnngp
- conda install --yes --file requirements.txt
- conda install -c nvidia cuda-nvcc
- pip install framework-reproducibility==0.4.0

3.环境搭建后的激活

搭建好后，每次使用前还需要激活 DNNGP3 环境才能进行使用:

```
source activate DNNGP3
```

使用以上代码激活后，命令行的前缀由（bash）变为（DNNGP）说明激活成功

三、数据准备与处理：从 vcf 文件获取 PCA 矩阵，并转为 pkl 文件格式

（1）使用 plink2 把 vcf 文件转为 PCA 的主成分,获取 tsv 文件。**注意：一定要把训练的和要预测的基因型合并到一个 vcf 里面，然后在获得 PCA 的 tsv 文件后，再分为训练数据的 tsv 和要预测的 tsv。**

```
plink2 --threads 30 --vcf test.vcf --pca 283 --out pca283 --allow-extra-chr
cat pca283.eigenvec|sed 's/#IID/ID/' >pca283.tsv
```

#这里模型训练时所用的样本数为 283，假设合并后的 vcf 文件中，训练用的 283 个样本在前，你所需要预测的样本（假设为 100 个，实际替换为你的样本数）在后。将合并后的 vcf 文件降维，获得 PCA 的 tsv 文件后，再在 tsv 文件中，将中你的训练数据拆出。

```
tail -100 pca283.tsv >pca283.predic.tsv          ##提取出需要预测的 100 个群体
```

（2）使用脚本 tsv2pkl.py 把 PCA 的 TSV 文件转为 pkl 文件

```
tsv2pkl.py pca283.predict.tsv pca283.predict.pkl
```

至此，前期的数据处理已结束，可正式使用模型进行预测。

四、使用处理好的数据进行预测

1.使用模型预测

在得到处理好的pkl文件后，我们要对其表型性状进行预测。该部分需要两个输入文件，一个是模型文件，即对应性状的**training.model**，（如**sorbitol. training.model.h5**），第二个是需要预测群体的pkl格式文件。

预测表型性状命令示例：

```
python Pre_runner.py --Model "/Your_path/training.model.h5" --SNP "/Your_path/ pca283.predict.pkl " --output /Your_path/
```

DNNGP 预测参数说明：

--Model：训练模型时生成的.h5 模型文件路径

--SNP：待预测数据集的基因数据文件路径

--output：预测结果文件的生成目录

Windows 平台下可以通过双击 DNNGP 目录下的 **Start_DNNGP.bat** 启动 GUI 界面，然后根据 GUI 提示进行操作。

Linux 平台下可以通过上一步启动的 GUI 界面（bash bash_Start_DNNGP.sh），然后根据 GUI 提示进行预测操作。

2.模型预测输出文件

DNNGP 模型完成预测后将在指定目录下生成结果文件 **Prediction.ALL.csv**，该文件即是对育种群体所有个体的表型性状预测结果。

五、特别说明

Script 目录下含有名为 Pre-Batch_run.py 的 Python 脚本，可以批量进行模型预测。

运行示例命令:

```
python Pre-Batch_run.py
```