ECG-Classfier

This is an ECG classfier based on PTB-XL database, which is the homework for my master course **pattern recognization**.

这个仓库的代码是基于PTB-XL数据库建立的心电分类器, 用于本人研究生课程**模式识别**的课程设计

PTB-XL database

简介

在1989年10月至1996年6月的近七年中，使用Schiller AG的设备收集了PTB-XL ECG数据集的基础波形数据。通过从Schiller AG收购原始数据库，将全部使用权转让给了Schiller AG。 PTB。在Physikalisch-Technische Bundesanstalt（PTB）的一个长期项目中，整理了这些记录并将其转换为结构化数据库。该数据库已在许多出版物中使用，但是直到现在，访问仍然受到限制。机构伦理委员会批准了匿名数据在开放访问数据库（PTB-2020-1）中的发布。在2019年的公开发布过程中，对现有数据库进行了简化，特别是针对机器学习社区的可用性和可访问性。

数据采集

1. 原始信号数据已记录并以专有压缩格式存储。对于所有信号，我们在右臂上提供12组标准导线（I，II，III，AVL，AVR，AVF，V1，...，V6），并带有参考电极。
2. 一般对应的元数据（如age，sex，weight和height）被收集在数据库中。
3. 每条记录都用报告字符串（由心脏病专家生成或由ECG设备自动解释）进行注释，该报告字符串将转换为一组标准化的SCP-ECG声明（scp_codes）。对于大多数记录，还提取了心脏的轴（heart_axis）和梗塞区（infarction_stadium1和infarction_stadium2，如果存在的话）。
4. 大部分记录已由第二位心脏病专家确认。
5. 所有记录均由主要关注信号特性的技术专家验证。

数据预处理

心电图和患者由唯一的标识符（ecg_id和patient_id）标识。元数据中的个人信息（例如，验证心脏病专家，护士的姓名和记录的记录地点（医院等））是假名。出生日期仅作为ECG记录时的年龄，其中符合HIPAA标准的89岁以上的年龄在300的范围内。此外，每位患者的所有ECG记录日期均偏移了随机偏移量。用于注释记录的ECG声明遵循SCP-ECG标准。

数据集说明

数据集组织结构如下所示

```
ptbxl
├── ptbxl_database.csv
├── scp_statements.csv
├── records100
│   ├── 00000
│   │   ├── 00001_lr.dat
│   │   ├── 00001_lr.hea
│   │   ├── ...
│   │   ├── 00999_lr.dat
│   │   └── 00999_lr.hea
│   ├── ...
│   └── 21000
│       ├── 21001_lr.dat
│       ├── 21001_lr.hea
│       ├── ...
│       ├── 21837_lr.dat
│       └── 21837_lr.hea
└── records500
    ├── 00000
    │   ├── 00001_hr.dat
    │   ├── 00001_hr.hea
    │   ├── ...
    │   ├── 00999_hr.dat
    │   └── 00999_hr.hea
    ├── ...
    └── 21000
        ├── 21001_hr.dat
        ├── 21001_hr.hea
        ├── ...
        ├── 21837_hr.dat
        └── 21837_hr.hea
```

将下载后数据集解压后拷贝至`data`文件夹下：

```
PTB-XL-CLASSFIER
├── code
├── dataPreprocess.ipynb
├── data
    ├── ptbxl_database.csv
    ├── scp_statements.csv
    ├── records100
    │   ├── 00000
```

```
    │   │   ├── 00001_lr.dat
    │   │   ├── 00001_lr.hea
    │   │   ├── ...
    │   │   ├── 00999_lr.dat
    │   │   └── 00999_lr.hea
    │   ├── ...
    │   └── 21000
    │       ├── 21001_lr.dat
    │       ├── 21001_lr.hea
    │       ├── ...
    │       ├── 21837_lr.dat
    │       └── 21837_lr.hea
    └── records500
    ├── 00000
    │   ├── 00001_hr.dat
    │   ├── 00001_hr.hea
    │   ├── ...
    │   ├── 00999_numhr.dat
    │   └── 00999_hr.hea
    ├── ...
    └── 21000
        ├── 21001_hr.dat
        ├── 21001_hr.hea
        ├── ...
        ├── 21837_hr.dat
        └── 21837_hr.hea
...
```

该数据集包含18885名患者的10837秒的21837条临床12导联ECG记录，其中52%是男性，48%是女性，年龄范围从0到95岁（中位数为62，分位数范围为22）。数据集的价值来自对许多不同的同时发生病理的全面收集，也来自于大部分健康对照样品。诊断的分布如下，为了简化起见，我们将诊断语句限制为聚合到超类中的诊断语句（注意：由于每个记录可能有多个标签，因此语句的总和超过了记录数）：

Records	Superclass	Description
9528	NORM	Normal ECG
5486	MI	Myocardial Infarction
5250	STTC	ST/T Change
4907	CD	Conduction Disturbance
2655	HYP	Hypertrophy

波形文件以16位精度以WaveForm数据库（WFDB）格式存储，分辨率为1μV/LSB，采样频率为500Hz（records500/）。为了方便用户，我们还以100Hz（records100/）的采样频率发布了下采样版本的波形数据。

所有相关的元数据ptbxl_database.csv均以标识，每条记录存储一行。它包含28个列，可以分为：ecg_id

1. 标识符：每个记录都由唯一的来标识ecg_id。对应的患者通过编码patient_id。原始记录（500 Hz）的路径和记录的降采样版本（100 Hz）存储在filename_hr和中filename_lr。
2. 常规元数据：人口统计和记录元数据，例如年龄，性别，身高，体重，护士，部位，设备和recording_date。
3. ECG语句：核心组件是（SCP-ECG语句作为字典，带有以下形式的条目）scp_codesstatement: likelihood，其中可能性（如果未知，则设置为0）并进行报告（报告字符串）。附加字段，，，，，和。
4. 信号元数据：信号质量，例如噪声（和），基线漂移（）和其他伪像，例如。我们还提供了计数额外的心脏收缩和起搏器的信号模式，以指示起搏器处于活动状态。
5. 交叉验证折叠：建议进行10倍火车测试拆分（heart_axisinfarction_stadium1infarction_stadium2validated_bysecond_opinioninitial_autogenerated_reportvalidated_by_human static_noiseburst_noisebaseline_driftelectrodes_problemsextra_beats strat_fold）是在尊重患者分配的同时通过分层抽样获得的，即特定患者的所有记录都分配了相同的首折。第9和10折中的记录至少经过了一次人工评估，因此具有特别高的标签质量。因此，我们建议将1-8倍用作训练集，将9倍用作验证集，将10倍用作测试集。

与使用的注释方案有关的所有信息都存储在专用文件中，该文件中充斥着与其他注释标准（例如AHA，aECGREFID，CDISC和DICOM）的映射。我们提供了其他附带信息，例如可以将每个语句分配给的类别（诊断，形式和/或节奏）。对于诊断语句，我们还提供了一个建议的层次结构到和中。
scp_statements.csvdiagnostic_classdiagnostic_subclass

获取数据

- 该数据集的下载页面在这里：PTB-XL
- 压缩的数据集下载：下载ZIP文件
- 在终端下载可以直接通过 `wget` 命令行实现

```
wget -r -N -c -np https://physionet.org/files/ptb-xl/1.0.1/
```

数据预处理

数据预处理通过 `code/dataPreprocess.ipynb` 处理实现

主要的流程包括：平滑滤波、50Hz陷波、R波分段提取、基线漂移去除、数据集划分

平滑滤波

通过窗口滑动平均滤波实现，这里采用的是五点平滑滤波

50Hz陷波

通过 `scipy.signal.iirnotch` 实现

R波分段提取

R波的检测是通过固定差分阈值实现，基本原理是设定一个固定的峰值阈值作为限定条件，一般通过获取某一段时间内心电信号最大值和最小值，来设置捕获条件阈值，计算式如下：

$$Threshhold = (max(x) - min(x)) \times 0.7 + min(x)$$

这里主要是通过II导联的心电信号进行差分阈值得到R波的位置

基线漂移去除

基线漂移的去除是采用 `PR` 段作为基线，先取每个导联 `PR` 段上10个数据点的均值作为基线的近似值，然后用所有的数据减去该近似值，即可得到去基线的心电图数据

> 关于要不要去除基线这个问题持保留情况，主要是心电信号在低频部分也有部分很重要的有用信号，例如ST段，考虑到基线对心电前期预处理没有太大的干扰，这里就没有去除基线漂移

机器学习模型分类

数据超采样

这里的心电数据样本存在样本不均衡的问题，所有的分类中正样本比例均显著低于负样本比例，如果采用欠采样的话会使训练集丢失部分数据，而过采样会导致一个数据点在高维空间出现多次，增加过拟合风险，很多研究通过在过采样中加入少量随机噪声来减少这类风险

这里需要安装 `imblearn` 包，使用以下的命令行安装

```
pip install imbalanced-learn
```

分类模型

尝试了下svm，运行的时间太慢了，实在等不下去了，测试了下随机森林分类，效果在预期之内，就利用随机森林跑了一下分类，之后有时间再修改一下

利用随机森林对特定分类的病例二分类的分类器的准确率如下：

class	accuracy	precision	recall	f1-score	roc score
MI	0.904	0.993	0.535	0.695	0.950
STTC	0.899	0.987	0.516	0.678	0.936
CD	0.905	0.911	0.587	0.714	0.914
HYP	0.940	0.990	0.720	0.833	0.963