**实验一 人体运动功能采集与智能分析（16学时）**

**数据库介绍**

该数据库包括45例脊髓型颈椎病（CSM）患者和45例正常人（HC）的步态数据，每个受试者均采集了两组数据。其中001-045代表受试者编号，-1和-2代表一个受试者的两组数据。

**实验目的**

1. 了解步态的基本知识；时域特征、时空参数的定义

步态基本知识：步行是人类生存的基础，是区别人类与其他动物的关键特征之一。步态，即人步行时所表现姿态，是人体通过髋、膝、踝、足趾乃至躯干与上肢的一系列连续活动，使身体向前移动的过程，其有赖于中枢神经系统、周围神经系统以及骨骼肌肉系统的协调工作，是人体运动和神经系统结构及功能以及心理活动在行走时的外在体现。当神经系统或骨骼肌肉系统因疾病或损伤而受到损害时，就有可能出现步态异常，步态特征对于每个人而言是独一无二的。

受疾病的影响，患者的步态发生改变，其关节运动学数据也会发生相应的变化。例如在临床上，CSM患者的脊髓受到压迫后会导致手足无力，步态不稳，下肢有踩棉花样感觉；帕金森患者则易出现慌张步态，即一旦起步便会不由自主地快速小步向前移动。因此，通常来说，CSM患者的下肢关节角度相较于HC更大，而帕金森患者的下肢关节角速度和加速度相较于正常人更大。根据这些特点，可对受试者下肢关节运动学数据时间序列进行时域特征提取，对比不同步态模式在关节运动学数据上的差异，如下表所示：

表1 本实验所使用时域特征名称及其定义

|  |  |
| --- | --- |
| 特征名称 | 特征定义 |
| 峰峰值（也可称关节活动范围，p2p） |  |
| 平均值（mean） |  |
| 方根幅值（rootAmplitude） |  |
| 标准差（std） |  |

时空参数：人的行走过程呈现周期性，在行走时，同一只脚从脚跟离地跨出，到该脚跟再次触地的行进过程称为步态周期。每一个步态周期主要分为两个阶段，即支撑相（也称站立相）和摆动相（也称迈步相）。支撑相，即步态周期中的承重期，从一侧的足跟触地开始，至同侧脚趾离地结束。摆动相则由脚趾离地开始，直至同侧脚跟触底结束。单个步态周期中人体不同阶段的行走姿势以及两侧脚的摆动与支撑状态如下图所示：

**步态周期示意图**

图1 单个步态周期各阶段示意图

受疾病的影响，正常人和患者之间、不同严重程度的患者之间以及不同疾病的患者之间的步态时空参数存在一定的差异。因此，根据单个步态周期不同阶段的人体状态，常用如下表所示的时空参数对步态进行量化计算：

表2 时空参数名称及其定义

|  |  |
| --- | --- |
| 参数名称 | 参数定义 |
| 步态周期（GC） | 从一侧脚跟离地跨出至同侧脚跟再次离地跨出的时间（s） |
| 步长（Step length） | 行走时同侧足跟着地至对侧足跟着地间的平均距离（m） |
| 步宽（Step width） | 两足跟中心点或重力点之间的水平距离（m） |
| 跨步长（Stride length） | 一侧足着地至同足再次着地间的距离，约为步长的2倍（m） |
| 步速（Velocity） | 单位时间内步行的平均速度（m/s） |
| 步频（Cadence） | 单位时间内步行的平均步数（步/s） |
| 周期变异系数（CV of GC） | std（GC）/mean（GC） |
| 左摆动相占周期比（LSP） | 左摆动相时间/步态周期时间\*100% |
| 右摆动相占周期比（RSP） | 右摆动相时间/步态周期时间\*100% |

如图所示分别为人体矢状面、冠状面和水平面（定义为数据的x、y、z三个方向）的设置以及光学运动捕捉实验过程中人体骨性标记物的位置（即反光标记物粘贴的对应位置）：

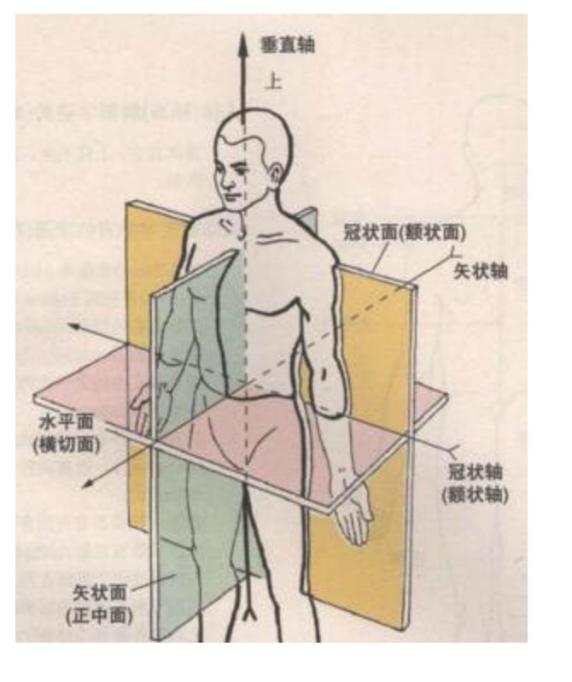
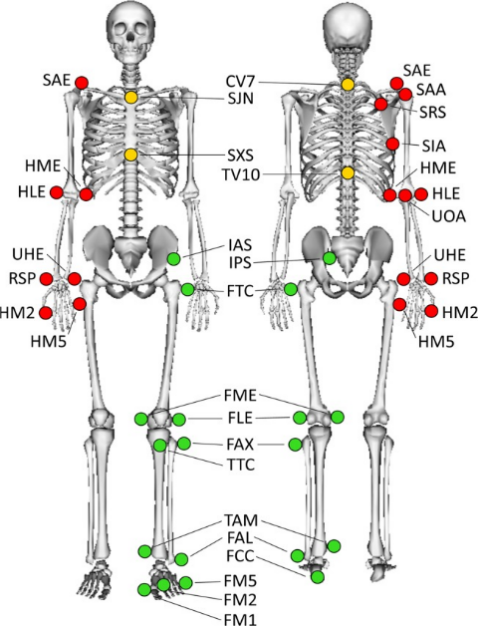
 

图2 人体三维方向设置 图3 反光标记物粘贴位点示意图

以步态周期、步长、步宽和左摆动相占周期比4个参数为例，其时间序列特征点提取结果如图4-图7所示。其中，步态周期参数通过提取粘贴于足踝部任一反光标记物的z轴时间序列波峰峰值计算，第n个周期的计算公式为



步长通过提取每个周期内分别粘贴于左右后跟骨（称为L\_FCC或R\_FCC）上的反光标记物在水平面（横切面）上的轨迹坐标最低点（在图中分别用红色“\*”和绿色“△”标记），使左右两侧标记物的x轴轨迹坐标数值相减，即可得每个周期内的步长数据，例如，第n个步长数值可表示为



步宽通过提取每个周期内分别粘贴于左右后跟骨（称为L\_FCC或R\_FCC）上的反光标记物在水平面（横切面）上的轨迹坐标最低点（在图中分别用红色“\*”和绿色“△”标记），使左右两侧标记物的y轴轨迹坐标数值相减，即可得每个周期内的步宽数据，例如，第n个步宽数值可表示为



左摆动相占周期比通过提取每个周期内分别粘贴于左脚后跟骨（称为L\_FCC）和左脚第一跖骨头（称为L\_FM1）上的反光标记物在水平面（横切面）上的z轴轨迹坐标最低点（在图中分别用红色“\*”和绿色“△”标记），使该两点轨迹坐标对应的拍摄帧数值相减并处以该步态周期所占拍摄帧数值，即可得该步态周期的左摆动相占周期比，例如，第n个左摆动相占周期比可表示为



****

图4 周期计算特征点提取 图5 步长计算特征点提取



图6 步宽计算特征点提取 图7 左摆动相占周期比计算特征点提取

1. 特征提取

本实验对受试者的步态特征进行了提取，其中时空参数根据受试者个体进行提取，每个受试者均获取了9个时空参数；时域特征分别根据每个受试者单侧下肢髋关节（L&R分左右hip）、膝关节Knee和踝关节ankle在矢状面、冠状面和水平面的三维角度时间序列数据进行提取，共可获36维特征数据。因此，本实验共获取45维特征。

1. 特征选择与降维

高维医疗数据往往具有大量不相关或冗余的特征。因此，需对这些数据进行处理，否则势必会影响算法的计算效率，也会干扰挖掘获得的信息，影响模型的分类精度。本实验利用SHAP值来计算和可视化特征的重要性贡献，并将其作为从原始数据集中选择特征的基础。这种方法具有一定的可解释性，为问题的分析带来了便利。首先对特征进行归一化处理。对于不同的分类器，基于 SHAP 值的特征选择结果是不同的。根据不同分类器的SHAP值对所有原始特征进行排序，然后按降序排列这些特征，逐个删除SHAP值最小的特征并计算每次的分类准确率，最终选择分类准确率最高的特征组合作为分类器的输入，并获取分类器的其他性能指标。

1. SVM、LR、RF、KNN、DT (基本知识)
2. 评价指标的含义

确定每个分类模型的最佳特征组合之后，需要将特征输入分类模型，获取最终的分类性能。本实验使用以下几个指标来评价分类模型的分类性能：

1、一级指标

（1）TP：true positive，即真实值是positive，模型认为是positive的数量；

（2）TN：true negative，即真实值是negative，模型认为是negative的数量；

（3）FN：false negative，即真实值是positive，模型认为是negative的数量，这是统计学上的第一类错误（Type I Error）；

（4）FP：false positive，即真实值是negative，模型认为是positive的数量，这是统计学上的第二类错误（Type II Error）。

将以上四个一级指标一起呈现在表格当中，就可以得到一个矩阵，即为混淆矩阵，如表3所示。

表3 混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| **混淆矩阵** | | 真实值 | |
| Positive | Negative |
| 预测值 | Positive | TP | FP (Type II Error) |
| Negative | FN (Type I Error) | TN |

2、二级指标

本实验所用二级指标由表4给出。二级指标可以将混淆矩阵中的样本数量结果转化为[0,1]之间的比率，便于进行标准化的衡量。

表4 二级指标说明表

|  |  |  |
| --- | --- | --- |
| **指标** | **公式** | **意义** |
| 准确度  ACC | Accuracy | 分类模型所有判断正确的结果占总观测值的比重 |
| 精确度  PPV | Precision | 在模型预测是positive的所有结果中，模型预测对的比重 |
| 灵敏度  TPR | Sensitivity=Recall | 在真实值是positive的所有结果中，模型预测对的比重 |
| 特异度  TNR | Specificity | 在真实值是negative的所有结果中，模型预测对的比重 |

3、三级指标

在上述二级指标的基础上再进行拓展，产生一个三级指标——F1分数（F1 Score）：



其中，P代表Precision，即精确度，R代表Recall，即灵敏度。F1-Score指标综合了Precision与Recall的产出的结果。F1分数是统计学中用来衡量二分类模型精确度的一种指标，它同时兼顾了分类模型的精确率和[召回率](https://baike.baidu.com/item/%E5%8F%AC%E5%9B%9E%E7%8E%87/560642" \t "https://baike.baidu.com/item/F1%E5%88%86%E6%95%B0/_blank)。F1分数可以看作是模型精确率和[召回率](https://baike.baidu.com/item/%E5%8F%AC%E5%9B%9E%E7%8E%87/560642" \t "https://baike.baidu.com/item/F1%E5%88%86%E6%95%B0/_blank)的一种调和平均，它的最小值为0，最大值为1，值越大代表模型输出结果越好。

4、其他指标

除了以上三类指标，本研究还使用了马修斯相关系数（MCC）、Kappa系数、ROC曲线和AUC值来评价模型的输出性能。

MCC是用于测量二分类模型分类性能的指标。该指标同时考虑了真阳性（TP）、真阴性（TN）、假阳性（FP）和假阴性（FN），通常被认为是一个比较均衡的指标，可应用于两类别的样本含量差别很大的情况。MCC本质上是一个描述实际分类与预测分类之间相关性的相关系数，它的取值范围为[-1,1]，取值为1时表示对受试对象实现了完美预测，取值为0时表示预测的结果还不如随机预测的结果，取值为-1时是指预测分类和实际分类完全不一致。MCC的计算公式如下所示：



Kappa系数是一种衡量分类精度的指标。它是通过把所有真实分类中的像元总数（N）乘以混淆矩阵对角线（Xkk）的和，再减去某一类真实像元总数与该类中被分类像元总数之积对所有类别求和的结果，再除以总像元数的平方减去某一类地表真实像元总数与该类中被分类像元总数之积对所有类别求和的结果所得到的。Kappa系数的计算公式如下：



其中，是每一类正确分类的样本数量之和除以总样本数，也就是总体分类精度 ACC，而在本研究的二分类问题之中，可用如下公式表示：



ROC曲线指受试者工作特征曲线，也称为接收器操作特性曲线(Receiver Operating Characteristic Curve)，是反映敏感性和特异性连续变量的综合指标。它使用构图法展现敏感性和特异性的相互关系，通过将连续变量设定出多个不同的阈值，从而计算出一系列FPR（False Positive Rate）和TPR（True Positive Rate），再以FPR为横坐标、TPR为纵坐标绘制成曲线，曲线下面积AUC值越大，表示分类准确性越高。