**ISUP标准下3D数据的联合模型**

**2019.10.14**

**训练数据：**

本次实验所用的是ISUP标准下的3D数据，其中共156个案例（97个案例label为0，59个label为1），经过新的ICC筛选后，各提取了3158个特征（平扫期 1034个、动脉期 1090 个、静脉期 1034个）。

我们对这些数据进行了拆分，训练集和测试机集比例为7:3，各包含109例与47例。

**方法描述：**

**特征选取：**

我们将这些特征分成了四类，分别为形状特征（13个）、灰度特征（19个）、纹理特征（56个）、变换特征（594个）。其中变换特征我们只选取了小波变换。

随后，我们分别选取前两类、前三类、前四类特征作为训练数据进行训练，比较训练所得模型的结果，从而比较特征类型对结果的影响。在下文中我们将这三种特征集合成为2kind、3kind、4kind。

**数据平衡：**

另外，由于数据中的label值分布并不平衡，label为1的数据量明显较少，所以我们对这批数据进行了smote（Synthetic Minority Oversampling Technique）算法处理。它可以对少数类样本进行分析，并根据少数类样本人工合成新样本添加到数据集中。

**降维方式：**

在训练过程中我们采用PCC的降维方式，即通过计算特征之间的皮尔逊相关系数，从而筛去一些相关性较高的特征，从而达到降维的目的，减小模型的大小和参数量。

**分析方式：**

我们用方差分析的方式（ANOVA）来检验特征与label的相关性，从而使得显著相关的特征更加容易被提取到。

**归一化：**

我们对数据进行了z-score标准化，公式如下。



**训练模型：**

本次实验一共才用了四种训练模型：支持向量机（SVM）、随机森林（RF）、逻辑回归（LR）、。它们的简单说明可见附录。

**模型的挑选方式：一个标准误差法（One-standard Error）**

简称OSE，即当我们获得了模型的训练结果（AUC）与所选取的特征数的关系曲线后，选取最高的验证集AUC值作为基准，在它的一个标准差（训练数据）范围内选取特征数最少的模型，这样可以防止模型复杂与过拟的情况发生。

**期态—特征类数模型训练结果：**

对于平扫期、动脉期、静脉期，我们分别对2kind、3kind、4kind特征集进行训练，在使用OSE方法经过筛选后，我们对每种模型类别都选出了每个期态—特征类数中的最佳模型，在比较最终的训练结果后，我们决定使用分类效果较好，也便于解释的LR模型，它的结果如表1所示。

表 1. 最佳模型结果汇总

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **特征类数** | **2kind** | | | | **3kind** | | | | **4kind** | | | |
| **模型和AUC值** | Model | train | Val | Test | Model | train | Val | Test | Model | train | Val | Test |
| **平扫期** | **Lr12** | **0.894** | **0.863** | **0.828** | Lr8 | 0.874 | 0.843 | 0.749 | Lr5 | 0.882 | 0.853 | 0.753 |
| **动脉期** | Lr1 | 0.815 | 0.802 | 0.824 | Lr1 | 0.815 | 0.802 | 0.824 | **Lr6** | **0.888** | **0.861** | **0.810** |
| **静脉期** | **Lr1** | **0.820** | **0.816** | **0.791** | Lr1 | 0.820 | 0.816 | 0.791 | Lr1 | 0.821 | 0.816 | 0.791 |

其中‘Model’一栏中，表示的是最佳模型的模型类别和特征数（exp：LR2表示使用的是逻辑回归（LR）模型，选取特征数为2）。另外，我们是根据val的AUC来选出的模型，test数据不应在此时公布，此处列出仅作为展示。

为了获得最终的联合模型，我们必须在同一期态的三个模型中选出一个模型作为代表。经过比较，平扫期选择了2kind的LR12模型，动脉期选择了4kind的LR6模型，静脉期选择了2kind的LR1模型。（已在表1中有粗体字标出）

它们提取到的特征分别为：

平扫期：90%灰阶强度

能量（即像素值的平方和）

ROI区域最小轴长度

ROI区域近似椭球体主轴长

冠状面最大二维直径长

矢状面最大二维直径长

ROI区域的二维最大直径

ROI区域的三维最大直径

ROI区域近似椭球体次轴长

ROI区域与球体的相似度

ROI区域表面积

ROI区域近似球体紧凑程度

动脉期：ROI区域的二维最大直径

ROI区域的三维最大直径

ROI区域与球体的相似度

小波变换后的图像依赖熵数值

小波变换后图像中依赖关系的相似性

纹理图案的强度值均匀程度

静脉期：ROI区域的二维最大直径

**联合模型的训练结果：**

由于LR模型可以输出预测概率值，所以我们将每个期态的预测概率值输入逻辑回归模型之中，从而得到联合模型。经过两两组合，共获得了4组联合模型，结果如表2所示。（其中“权重”与“期态”相对应；敏感性与特异性为测试集的数据）

表 2. 独立模型与联合模型结果汇总

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 期态 | Train AUC | Test AUC | 敏感性 | 特异性 | 权重 |
| 平扫 | 0.876 | 0.828 | 0.778 | 0.759 | / |
| 动脉 | 0.860 | 0.810 | 0.833 | 0.759 | / |
| 静脉 | 0.807 | 0.791 | 0.833 | 0.759 | / |
| 平扫+动脉+静脉 | 0.885 | 0.839 | 0.833 | 0.793 | 1.79，1.88，0.25 |
| **平扫+动脉** | **0.888** | **0.841** | **0.833** | **0.828** | **1.86，1.96** |
| 动脉+静脉 | 0.860 | 0.822 | 0.778 | 0.793 | 2.62，0.80 |
| 平扫+静脉 | 0.868 | 0.833 | 0.778 | 0.793 | 2.49，0.91 |

**结果讨论：**

1. 最佳模型：

综合考虑之下，虽“平扫+动脉”的模型效果最好，AUC值可以达到0.841，同时它0.833的敏感性与0.828的特异性也是所有模型中最高的，不知道临床上这两个数值是否可以接受。

1. 独立模型之间的比较：

独立模型中，平扫期效果好于动脉期好于静脉期，这是可以解释的；但是经过wilcoxon双边检验以后，平扫期与静脉期模型间的p值为0.120，说明二者并不存在显著性差异，也就是说三期的分类效果并没有明显差异。

另外，虽然平扫期模型的auc值更高，但是敏感性更低，请问这是否有临床上的解释。

1. 联合模型间的比较：

联合模型之间的差距并不明显，但“动脉+平扫”的分类效果相对突出，也是一个可以解释的结果。

1. 所提取到的特征

综合而言，这些分类模型表明，形状特征是主要的分类依据，其中ROI区域的二维个三维最大直径被提及3次，与球体的近似程度被提及2次（静脉期中该特征被ICC筛除，因此没能提取到）。

**工作展望：**

综合而言模型的分类效果尚可，并且具有一定的解释性，可以被用作最终模型，不知道该模型表现在临床上或者科研探讨上，是否具有一定的积极意义?

接下来我将最后尝试着提高模型的分类效果，并正式开始论文的一些方法部分的撰写工作。

附录：

支持向量机（SVM）：

SVM是一种高效稳定的分类器，其思想是建立一个最优决策超平面，使得该平面两侧距离该平面最近的两类样本之间的距离最大化，从而对分类问题提供良好的泛化能力。

随机森林（RF）：

随机森林是一种有监督学习算法。就像你所看到的它的名字一样，它创建了一个森林，并使它拥有某种方式随机性。所构建的“森林”是决策树的集成，大部分时候都是用“bagging”方法训练的。

逻辑回归（LR）：

这是一种线性回归模型，即将问题构建为 y=wx+b 的形式，然后将其输入sigmoid函数从而得到分类结果。