**ISUP标准下3D数据的联合模型**

**2019.9.30**

**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）。

**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，去除了log-sigma变换得到的特征。

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）、LASSO逻辑回归（LASSO-LR）。它们的简单说明可见附录。

**期态—特征类数模型训练结果：**

**（在进行训练时由于重新拆分了测试集，所以每个期态—特征类数模型的测试集各不相同，这使得模型的比较标准不一致，所以还需要重新训练。不过这对于最终的模型结果并不会造成很大的影响，因此我们还是将这个模型保存下来，并将其展示，从而对最终的分类效果有个大致的估计）**

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，在经过观察后，我们选出了每个期态—特征类数中的最佳模型，如表1所示。

其中‘Model’一栏中，表示的是最佳模型的模型类别和特征数（exp：Rf13表示使用的是随机森林（RF）模型，选取特征数为13）。

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** | *Rf13* | *1.000* | *0.902* | *0.858* | Rf7 | 1.000 | 0.925 | 0.761 | Rf5 | 1.000 | 0.890 | 0.728 |
| **动脉期** | Rf12 | 1.000 | 0.855 | 0.845 | *Rf8* | *1.000* | *0.820* | *0.920* | Rf12 | 1.000 | 0.876 | 0.860 |
| **静脉期** | Rf5 | 1.000 | 0.800 | 0.805 | *Rf11* | *1.000* | *0.847* | *0.820* | Rf9 | 1.000 | 0.852 | 0.702 |

可以看出所选的模型都是随机森林模型，这是因为我们主要是根据验证集的AUC来进行选取，而随机森林的模型特点就是会产生比较严重的过拟，所以在使用交叉验证的情况下，随机森林模型的验证集AUC值会偏高。

**联合模型的训练结果：**

**（在进行归一化时不应该用测试集的标准差与平均值进行计算，这也会造成最终AUC的误差，不过也不会造成巨大偏差）**

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Rf13模型，动脉期选择了3kind的Rf8模型，静脉期选择了3kind的Rf11模型。（已在表1中有斜体字标出）

由于随机森林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）

表 2. 联合模型结果汇总

|  |  |  |  |
| --- | --- | --- | --- |
| 期态 | Train | Test | 权重 |
| 平扫+动脉+静脉 | 0.982 | 0.973 | 2.525, 2.525, 0.252 |
| 平扫+动脉 | 0.982 | 0.970 | 2.556, 2.601 |
| 动脉+静脉 | 0.949 | 0.962 | 3.656, 0.548 |
| 平扫+静脉 | 0.952 | 0.968 | 3.473, 0.941 |

**结果讨论：**

1. 虽然在训练过程中存在两个问题，但是我认为并不会对最终结果造成过多的影响。总体而言结果相当不错，0.982和0.973的测试集和训练集AUC可以说已经没有很大的提高空间了。
2. 经过比较，对于平扫期而言，只选取形状特征和强度特征便已足够；而对于动脉期和静脉期而言，加入纹理特征是更好的策略。而变换特征则相对而言并不是十分必要。由于动脉期和静脉期中存在造影剂，所以纹理特征更为重要似乎是可以解释的。
3. 平扫期和动脉期对于最终的分类有着更大的贡献，静脉期图像的分类效果略差。这一点似乎也与预期的较为符合。平扫期的图像是基本；动脉期中加入造影剂后尚未经过很长时间，对于分类很有帮助；而静脉期则由于经过了较长时间，而对于不同病人，造影剂消退的程度不同，这也就导致了图像效果不太稳定。
4. 虽然静脉期的模型效果并不理想，但是加入静脉期模型依旧对联合模型有所帮助，（可以比较“平扫+动脉+静脉”和“平扫+动脉”的模型结果），只是提升效果十分有限。如果说静脉期的模型加入会使得模型更为复杂的话，出于这方面因素，可以考虑将静脉期模型舍去。

**工作展望：**

首先在保存下这个模型的同时，需要使用更加规范的方式来训练模型。如果结果差不多的话，应当使用后者作为最终模型，因为这更加便于解释。

关于如何选取最佳模型（包括如何选取期态—特征类数模型，以及如何选取期态最佳模型），还存在值得讨论的地方。接下来我会思考，是否有方法使得整个模型选取的过程更加系统规范，这样有利于解释，也便于重复。

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。

LASSO逻辑回归（LASSO-LR）：

即在逻辑回归的基础上，针对特征数量添加一个惩罚项，从而减少最终选取的特征数量，简化模型