**ISUP标准下3D数据的联合模型**

**2019.10.9**

**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，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）、决策树（DT）。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SVM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** | Svm2 | 0.830 | 0.812 | 0.837 | **Svm3** | **0.848** | **0.837** | **0.862** | Svm1 | 0.824 | 0.814 | 0.799 |
| **动脉期** | **Svm1** | **0.847** | **0.833** | **0.844** | Svm1 | 0.842 | 0.829 | 0.844 | Svm1 | 0.842 | 0.829 | 0.844 |
| **静脉期** | **Svm1** | **0.793** | **0.788** | **0.876** | Svm1 | 0.793 | 0.788 | 0.876 | Svm1 | 0.792 | 0.787 | 0.826 |

其中‘Model’一栏中，表示的是最佳模型的模型类别和特征数（exp：SVM2表示使用的是支持向量机（SVM）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。

**联合模型的训练结果：**

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3kind的SVM3模型，动脉期选择了4kind的SVM1模型，静脉期选择了2kind的SVM1模型。（已在表1中有粗体字标出）

由于SVM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 | 0.837 | 0.862 | 0.769 | 0.765 | / |
| 动脉 | 0.810 | 0.844 | 0.769 | 0.794 | / |
| 静脉 | 0.778 | 0.876 | 0.846 | 0.765 | / |
| **平扫+动脉+静脉** | **0.834** | **0.871** | **0.769** | **0.794** | **1.619，1.599，0.271** |
| 平扫+动脉 | 0.838 | 0.867 | 0.769 | 0.794 | 1.709，1.670 |
| 动脉+静脉 | 0.815 | 0.855 | 0.692 | 0.853 | 2.220，0.844 |
| 平扫+静脉 | 0.829 | 0.869 | 0.769 | 0.824 | 2.293，0.759 |

**结果讨论：**

1. 最佳模型：

根据AUC值选出的最佳模型是三期联合模型，它的AUC值为0.871。与之前所训练出来的联合模型的0.973的AUC值虽然相去甚远，但这也说明了之前模型训练时，测试集泄露带来的影响十分巨大，这是十分不科学的，我们应当采用这次的训练模型。

另外，它的敏感性和特异性分别为0.769和0.794，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：

相比之下，静脉期的模型表现最好，动脉期的模型表现最差多少有些出乎意料。另外动脉期的特异性、静脉期的敏感性最高，虽然存在一定的随机因素，但不知是否有医学上的解释。

1. 联合模型间的比较：

从AUC值而言，最终结果是三期联合模型的结果最为理想，但是它的敏感性没有达到静脉期那么好，特异性也没达到另外两个联合模型这么高。

**工作展望：**

在较为严格的训练过程控制下，训练出来的模型没有预期的这么好，但是还可以接受。只是其中静脉期的测试集结果比训练集要好不便于解释，这可以理解为随机事件。另外在smote的过程中略有不严格之处，导致了训练结果产生了一些波动，虽然不会造成显著差异，之后会尽快将模型重新训练一次，不过不用担心结果会有很大变动。

下一步的工作便是开始论文的撰写，同时也会探索一下模型选取的特征，并尝试着比较模型之间是否存在显著差异。

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。

决策树（DT）：

决策树(Decision Tree）是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。

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。