基于决策树的阿尔茨海默病诊断分析

吴春模
兰州财经大学统计学院　甘肃　兰州　730020

摘　要：阿尔兹海默症作为主要的神经退行性疾病之一，已成为导致痴呆问题最常见的原因。识别该疾病需要医疗专家进行全面检查，消耗大量成本和时间。阿尔兹海默症患者提早确诊难度大，晚期治疗效果不明显。为此提出基于决策树算法的阿尔兹海默病诊断，对ANDI数据集进行数据预处理、诊断指标关联性分析、决策树识别构建，实现阿尔茨海默病诊断。实验结果表明，该模型识别准确率达89.4%，对阿尔兹海默症进行早期分类研究具有重要临床意义。

关键词：方差分析；决策树分类模型；阿尔兹海默症

引言

阿尔茨海默症（阿尔茨海默症）是一种神经系统退行性疾病，在老年人群中具有较高的发病率。阿尔茨海默症患者患病后常有记忆力衰减、冷漠抑郁、沟通障碍、神志不清等全面性痴呆症状，严重影响日常生活，甚至死亡，并且导致不可逆的大脑损伤。这种疾病通常开始于中年或者老年时期，有研究表明可能是由神经元及其周围蛋白质的累积而诱发的，与神经细胞突触功能障碍、脑神经元细胞死亡、脑补萎缩有关。相关研究人员称，因为阿尔茨海默症所导致的脑部结构相关变化可能比出现阿尔茨海默症临床症状还要提前20年。

全球范围内每三秒就新增一个阿尔茨海默症患者，据估计，2050年将有6.4亿人被诊断为阿尔茨海默症。然而人们对阿尔茨海默症的致病因素还不清楚，也没有有效的药物或治疗方法来阻止或逆转阿尔茨海默症的进展。如果能在患者发作早期诊断出阿尔茨海默症，现代医学的一些治疗方法可以有效减轻病情，延缓病情发展。阿尔茨海默症患者提早被确诊难度较大，并且对晚期患者的治疗效果不佳，因此，对阿尔茨海默症进行早期分类研究具有重要的临床意义。

1 阿尔茨海默病数据集分析

本文数据集来自ANDI数据集。阿尔茨海默病神经影像学计划 (ADNI) 是一项纵向多中心研究，旨在开发临床、成像、遗传和生化生物标志物，用于阿尔茨海默病 (AD) 的早期检测和跟踪。自十多年前推出以来，这一具有里程碑意义的公私合作伙伴关系为 AD 研究做出了重大贡献，实现了世界各地研究人员之间的数据共享。

研究表明，阿尔兹海默症表现为阶段性的患病症状，因此AD症存在不同的类型，这种病的患病程度由正常，到轻微患病，再到确诊为阿尔兹海默症的过程，可以划分为认知正常老年人（CN）、主观记忆障碍患者（SMC）、早期轻度认知障碍患者（EMCI）、晚期轻度认知障碍患者（LMCI）和阿尔茨海默病患者（AD），统计分布如图1所示：

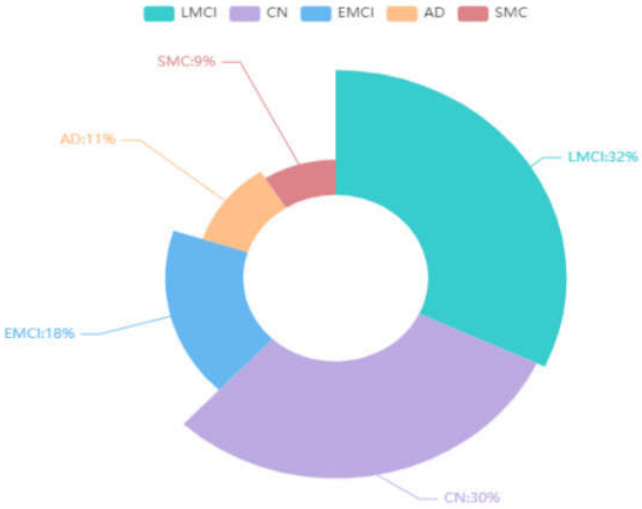

图1 ANDI数据集分析

2 阿尔茨海默病数据集预处理

该数据集中患者和正常老年人一共有16208例，但是附件所给到的数据中一共有16222例观测值，所以存在一定的差异。然后对特征指标进行详细细分后，发现题目中提到附件数据包含4850名认知正常老年人（CN），1416名主观记忆障碍患者（SMC），2968名早期轻度认知障碍患者（EMCI），5236名晚期轻度认知障碍患者（LMCI） 和1738名阿尔茨海默病患者（AD）患者收集在不同的时间点。

通过分析可知数据集中存在14个多余的空样本，对数据分析结果和模型建立误差和影响较小，综合考虑后直接对其进行剔除。

3 阿尔茨海默病数据集方差分析

对于决策树的阿尔茨海默病诊断识别而言，需要对

指标进行筛选，从而剔除相关性较小的指标，避免出现识别效果较差。在本文中，将采用方差分析进行判别[1]。

设X是一个特征指标，$E(X)$是特征指标X的数学期望，若$E(X^2)$存在，则称由式（1）定义的$V(X)$为X的总体方差，一般记为$V(X)$或$Var(X)$或者σ^2。

$$\sigma^2 = V(X) = Var(X) = E\left\{[X - E(X)]^2\right\} \quad (1)$$

方差$V(X)$的单位是特征指标的单位的平方，故在实际应用时，常取其算术平方根，令$\sigma = \sqrt{V(X)}$，称其为标准差或均方差。显然，标准差σ与特征指标X具有相同的量纲。

设$x_1,\cdots,x_n$是从阿尔兹海默症患病人群总体中随机抽取的样本含量为n的一个样本，并设$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$为该样本的样本均值，则样本方差由式（2）给出：

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right] = \frac{1}{n-1}SS \quad (2)$$

其中，$SS = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]$，为离均差平方和。离均差平方和是一个重要的统计量，在进行多因素试验设计资料的方差分析中，最核心的内容是对总离均差平方和的分解。设总体中个体的数目为N，观测指标为X，其总体平均值为μ，总体方差也可由式（3）给出：

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 \quad (3)$$

统计方差值为1的特征指标个数，最终选取了53个变量作为基础指标。随后利用筛选下来的数据进行相关性分析。RAVLT_perc_forgetting_bl、DX、CDRSB、FAQ、ADASQ4、ADAS13和ADAS11等7个指标与阿尔兹海默症病状细分程度DX_bl的相关较高，均大于0.4，其中RAVLT_perc_forgetting_bl与DX_bl的相关系数达到0.4917。而这7个指标都是跟AD症患者的认知行为特征相关，表明阿尔兹海默症的早期诊断，可以通过观察老年人的某些行为习惯是否存在异常来进行判断。

此外，全脑（WholeBrain_bl）、脑室心室（Ventricles_bl）、纺锤体（Fusiform_bl）等28个AD症患者的基本信息和大脑结构特征指标，都与AD症的患病程度呈现出负相关性，说明AD症患者的大脑结构保持良好状态，患者记忆障碍的症状会相对减轻。

4 基于决策树的阿尔茨海默症诊断

本文中所构建的决策树算法，是将阿尔兹海默症患者的样本属性作为节点，属性的取值作为分支，采用信息论的原理对大量患者特征信息的属性进行分析和归纳。决策树的根节点是所有样本数据中信息量最大的属性，即AD症病人的病状程度；中间节点是以该节点为根的子树所包含的样本子集中信息量最大的属性，叶节点是样本的类别值[2]。

决策树生成算法构造的结果是一棵二叉或多叉树：二叉树的内部结点（非叶子结点）一般表示为一个逻辑判断，树的边是逻辑判断的分支结果；多叉树的内部结点是属性，边是该属性的所有取值，有几个属性值就有几条边．构造决策树的方法通常采用自上而下的递归构造，分析步骤如下[3-5]：

1.通过训练集数据来建立决策数分类模型，得到决策树结构;

2.通过建立的决策树来计算特征重要性;

3.将建立的决策树分类模型应用到训练、测试数据，得到模型的分类评估结果;

4.由于决策树具有随机性，每次运算的结果不一样，若保存本次训练模型，后续可以直接上传数据代入到本次训练模型进行计算分类;

本文中采用了16208条AD症患者的认知行为特征和大脑结构等信息，一共53个特征指标，将其作为训练样本来构造决策分类树，提取相应的分类规则。

考虑到本题所涉及到的关于阿尔兹海默症病状程度的特征指标过多，一共有53个特征，需要将这53个特征进行分类，得到5个大类，即认知正常老年人（CN）、主观记忆障碍患者（SMC）、早期轻度认知障碍患者（EMCI）、晚期轻度认知障碍患者（LMCI）以及阿尔茨海默病患者（AD）患者这5大类，一共有53*5 = 265张叶子，绘制出来的多叉树图会比较密集，不能直观查看分类结果。因此，综合考虑后，构建决策树分类模型，得到各特征指标影响阿尔兹海默症诊断结果的重要性程度，同样可以识别出阿尔兹海默症的病状程度，从而提醒患者进行下一步的治疗。模型检验参数如表1所示：

表1 决策树模型参数

参数名	参数值
数据切分	0.7
节点分裂评价准则	gini
特征划分点选择标准	best
划分时考虑的最大特征比例	None
内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
叶子节点的最大数量	50
树的最大深度	10
节点划分不纯度的阀值	0

从表1中可以看出，构建的多叉树的最大深度为10，叶子节点的最大数量为50，内部节点分裂的最小样本数为2，叶子节点的最小样本数为1，与预期的结果基本吻合，这个多叉树图相对比较密集。

为更加直观的展示各特征指标对阿尔兹海默症诊断的影响程度，将所有特征的重要性从高到低进行排序，再次绘制条形图。DX的重要性比例最大，达到了0.41，这是因为DX_bl是DX的细分类别，DX的取值在一定程度上影响了AD症的诊断。DIGITSCOR_bl、DIGITSCOR_bl-CDRSB等12个指标的重要性比例较小，说明这12个大脑结构特征和认知认知行为特征帮助诊断AD症的效果有限。基于上述分析，给出决策树识别阿尔兹海默症病的混淆矩阵，如下所示：

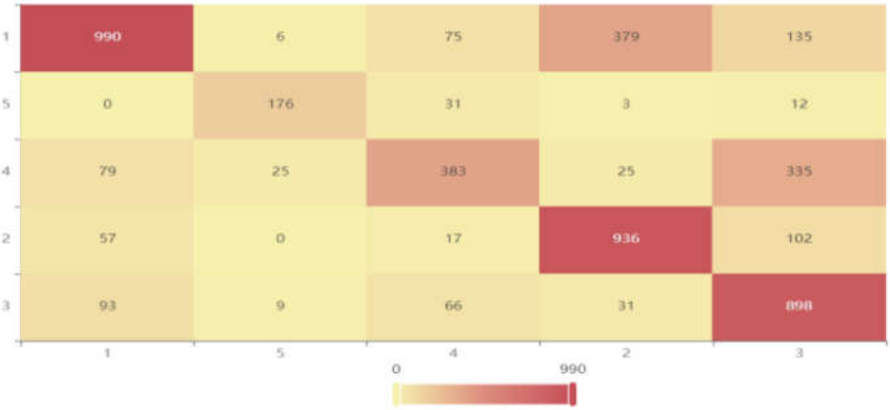

图2 混淆矩阵热力图

从图2中可以看出，横轴的5个标签表示AD症由正常到轻微认知障碍，再到AD症确诊的5个分类阶段；纵轴的标签表示分类的预测结果；矩阵中的取值X_{ij}表示将第i类预测程第j类的样本数目。矩阵主对角线上的样本数目较多，表明决策树分类模型在一定程度上能够准确诊断出阿尔兹海默症。但是将第2类SMC人群诊断为正常老年人的数据有379个，将第3类EMCI人群诊断为LMCI的样本数目有335个，说明决策树分类模型对AD症的诊断存在一定的误差，但整体上构建该模型来识别AD症还是有效的[6]。

表2 模型评估结果

集合	准确率	召回率	精确率	F1
训练集	0.849	0.849	0.852	0.849
测试集	0.696	0.696	0.711	0.69

表2展示了训练集和测试集的分类评价指标，准确率是预测正确样本占总样本的比例，两个数据集的准确率分别为0.849和0.696，说明模型在训练集和测试集上识别效果准确，能够用于阿尔茨海默病的诊断。

5 结论

阿尔茨海默症是一种神经性疾病，其早期症状不明显，因此往往在病情严重时才被诊断出来。本文利用ANDI数据集进行了数据预处理、诊断指标关联性分析和决策树构建，实现了阿尔茨海默病的诊断，训练集准确率达到了84.9%。本文的结果表明，该模型可以在一定程度上准确地诊断阿尔茨海默症，这不仅对于临床诊断具有重要意义，也为后续的研究提供了有价值的参考。

参考文献

[1]胡纯严,胡良平.如何正确运用方差分析——方差分析概述[J].四川精神卫生,2022，35(01):6-10.

[2]尹鹏飞,欧云.基于决策树算法的银行客户分类模型[J].吉首大学学报(自然科学版)，2014，35(05):29-32.

[3]程凤.探讨阿尔茨海默病患者采取综合护理对其临床依从性的意义研究[J].中国医药指南,2022,20(24):55-58.DOI:10.15912/j.cnki.gocm.2022.24.012.