A Data-Driven Intelligent Diagnostic Model for Alzheimer's Disease *

LIU Banteng[1], XIONG[2] Xinlong, LIANG Jingyao[1]*
(1.College of Information Science and Technology, Zhejiang Shuren University, Hangzhou Zhejiang 310015, China;
2.College of Information Engineering, Zhejiang University of Technology, Hangzhou Zhejiang 3110023, China)

Abstract: How to use sensors to collect signals to improve the accuracy of Alzheimer's disease diagnosis is a key issue in the field of computational psychiatry. Therefore, a data-driven intelligent diagnosis model is proposed for Alzheimer's disease. Firstly, feature extraction is performed on the signals obtained by sensors through ANOVA. Then, a decision tree model is constructed to diagnose diseases based on brain structural features and cognitive behavioral features. Finally, the effectiveness of the proposed algorithm is verified by using experimental data. The results show that in the three classification tasks, the proposed algorithm achieves an accuracy rate of 99.62%, a precision of 99.62%, a recall rate of 99.62%, an F1 score of 0.996 2, and an AUC indicator of 0.999 1. In the five classification task, the proposed algorithm has a cross validation accuracy of 99.40%, a precision of 99.40%, a recall rate of 99.40%, an F1 score of 0.993 9, and an AUC indicator of 0.999 7. Compared with other classic methods, the proposed algorithm has obvious advantages.

Key words: Alzheimer disease; ANOVA; feature extract; decision tree; intelligent diagnostic
EEACC: 6135 **doi**: 10.3969/j.issn.1004-1699.2025.10.014

基于数据驱动的阿尔茨海默病智能诊断模型研究 *

刘半藤[1], 熊鑫龙[2], 梁婧瑶[1]*
(1.浙江树人学院信息科技学院, 浙江 杭州 310015; 2.浙江工业大学信息工程学院, 浙江 杭州 310023)

摘 要: 如何利用传感器收集信号数据提高阿尔茨海默病诊断准确率是计算精神病学领域的关键问题。为此, 提出了一种基于数据驱动的阿尔茨海默病智能诊断模型。首先, 通过差异显著性分析对传感器获取的信号进行特征提取; 然后, 构造决策树模型分别对大脑结构特征以及认知行为特征进行疾病诊断; 最后, 利用实验数据验证所提出算法的有效性。结果显示, 在三分类任务中, 所提出的算法十折交叉检验的准确率为 99.62%, 精准率为 99.62%, 召回率为 99.62%, F1 得分为 0.996 2, AUC 指标为 0.999 1; 在五分类任务中, 所提出的算法十折交叉检验的准确率为 99.40%, 精准率为 99.40%, 召回率为 99.40%, F1 得分为 0.993 9, AUC 指标为 0.999 7。与其他经典方法对比, 所提算法具有较明显的优势。

关键词: 差异显著性分析; 特征提取; 决策树; 智能诊断
中图分类号: R749.16; TP391.4 **文献标识码**: A **文章编号**: 1004-1699(2025)10-1834-05

我国将步入中度老龄社会, 老年人口的医疗健康问题已经成为全体社会成员关注的焦点。阿尔茨海默病(Alzheimer's Disease, AD)是一种神经退行性疾病, 被称为威胁老人健康的"头号杀手"。近年来, 死于阿尔茨海默病所导致并发症的老人数量不断增加[1]。因此, 有效的筛查及诊断技术极为重要。但是阿尔茨海默病的病理复杂且临床表现多样, 患者往往因对其认知有限而错过最佳干预阶段。因此, 针对阿尔茨海默病诊断问题, 为患者及其家属制定科学高效的阿尔茨海默病识别方案, 根据患者大脑结构特征和认知行为特征完成阿尔茨海默病的智能诊断, 并进行早期干预已经成为缓解老人医疗健康问题以及推动健康中国建设的重要手段。

阿尔茨海默病是一种复杂且异质的精神疾病, 尽管基于症状学的诊断方法已经被广泛应用于各领域, 但这种方法并不利于病理机制的探讨。此外, 该诊断方法预测效率低下, 从而导致难以准确地评估和比较各种治疗方法的疗效。然而, 计算精神病学方法则能通过理论驱动和数据驱动两种互补的方法解决上述难题, 从而提高对阿尔茨海默病的认识。

近年来, 不少学者尝试基于计算精神病学的方法研究阿尔茨海默病。通过研究患者与正常对照者在 CT 影像数据以及功能磁共振成像数据等方面的区别, 运用机器学习方法进行疾病诊断。例如, 李长

项目来源: 浙江省人工器官与计算医学重点实验室开放/自主基金项目(SZD2025B016)
收稿日期: 2023-10-19 **修改日期**: 2023-12-14

胜等[2]基于功能磁共振成像数据利用集成学习的 Adaboost 算法在较小样本集实现阿尔茨海默病的诊断识别。该方法可以实现对阿尔茨海默病患者、轻度认知障碍患者的识别,综合识别准确率可以高达 96%。范炤等[3]基于功能磁共振成像数据利用 Adaboost 算法对阿尔茨海默病进行诊断,该方法可在十折交叉验证基础上达到 92.89% 的准确率。除 Adaboost 算法外,Kishore 等[4]进一步探讨不同机器学习算法对阿尔茨海默病的识别效果,并重点对支持向量机模型进行探讨。Balne 等[5]对识别阿尔茨海默病的传统机器学习算法以及深度学习算法进行综述,包括支持向量机、神经网络等。他们阐述各种方法的优势及劣势。Zhang 等[6]和 Lu 等[7]分别采用群优化的思想对传统模糊 KNN 算法进行改进,并建立改进的识别模型对阿尔茨海默病患者以及轻度认知障碍患者进行识别,结果显示综合识别准确率可以达到 97.64%。

由于功能磁共振成像获取的数据维度较高,若不加区分地进行疾病识别往往不能得到期望效果。因此,部分学者研究如何从数据提取关键特征信息从而实现高效的疾病识别效果。提取关键特征不仅可以提高患者识别准确率,也可以降低识别算法复杂度。例如,Muhammed 等[8]采用贪婪算法对阿尔茨海默病患病特征进行特征提取,并分别采用支持向量机模型以及 KNN 模型对患者进行识别,实验准确率可达到 90%以及 91%。Tu 等[9]采用卷积神经网络进行特征提取,并运用神经网络模型对疾病进行识别,结果显示阿尔茨海默病患者识别准确率为 96.2%,轻度认知障碍患者识别准确率为 87.4%。Sharma 等[10]采用深度学习的 VGG16 从功能磁共振成像数据中提取关键特征,并建立神经网络模型对阿尔茨海默病患者进行识别,准确率可以达到 90.4%。为能够获得更高的识别准确率,本文采用差异显著性分析的方法进行特征提取,并建立决策树模型对阿尔茨海默病患者、轻度认知障碍患者进行识别。

1　差异显著性分析的特征提取算法

1.1　数据预处理

构建高效识别模型的关键在于对功能磁共振成像获取的数据进行预处理。在缺失值填补中,往往使用均值、众数等方式进行填充,具备一定的随机性,容易造成较大的误差。因此,本文采用 Stekhoven 等学者提出的随机森林填补缺失值方法[11]。作为一种机器学习技术,研究也显示其具有超越 KNN 等填补方法的优势。

将完整的数据分为训练集和测试集进行训练。研究发现数据特征即使缺少 60%的数据量,其填补的准确度仍然能达到 90%。缺失数据填补算法的步骤如下:

Step1　将有缺失值的列根据缺失值的数量从小到大排列;

Step2　从缺失值较小列开始填充,被填充列记为缺失列,其余为自变量;

Step3　缺失列中没有缺失的视为训练数据的自变量,有缺失值的视为测试数据的因变量;

Step4　将训练后的模型应用于测试数据的因变量。

获得的功能磁共振成像数据可能存在由于意外因素导致的数据。如果这些数据没有得到适当修正,会影响模型对阿尔茨海默病定律的学习。因此,本文采用 3σ 准则判断数据中是否存在异常值,并对这些异常值进行修正,公式如下:

$$|x-\bar{x}|>3\sigma$$

对数据集进行标准化可以将不同维度的数据转换为统一范围内的数据。这有助于消除由于尺度差异引起的不同特征之间的偏差,数据标准化计算公式如下所示:

$$x^{*}=\frac{x_i-\bar{x}_i}{\sigma_i} \tag{1}$$

式中:x_i 为第 i 个指标的平均值,σ_i 为第 i 个指标的标准差。

1.2　单因素方差分析

经过数据预处理后,本文采用单因素方差分析探索各特征指标列与病症诊断的差异性,从而得到与阿尔茨海默症疾病诊断相关性较高的因素。进行单因素方差分析前,需满足其前提条件,即数据近似满足正态分布。根据中心极限定理:在一定条件下,独立随机变量的标准化的和随样本量变大会趋向正态分布,即它的累积分布函数收敛于标准正态分布。于是,可作如下假设:

$$\begin{cases}H_o:u_1=u_2=\cdots=u_r\\ H_1:u_1,u_2,\cdots,u_r\ \text{不全相等}\end{cases} \tag{2}$$

可将单因素方差分析模型表述为:

$$\begin{cases}X_{ij}=\mu_i+\alpha_i+\varepsilon_{ij}\\ \sum_{i=1}^{r}n_i\alpha_i=0\\ \varepsilon_{ij}\sim N(0,\delta^2)\end{cases} \tag{3}$$

将所有数据记录列与均值进行误差分析,可将误差分为两部分:

$$\begin{cases} S_A = \sum_{i=1}^{r} n_i(\bar{X}_i - X_i)^2 \\ S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \end{cases} \tag{4}$$

式中：S_A 是各组参数数学期望对总体数学期望的偏差平方和，其数值大小反映各个因素与因变量因素的差异，也称为组间误差；S_E 是各组参数对样本均值偏差的平方总和，其数值大小反映观测值的误差，通常被认为属于随机误差的一种，也称为组内误差。

因此，在单因素方差分析中，当原假设 H_o 正确时，存在如下关系：

$$\frac{S_A/(r-1)}{S_E/(n-r)} \approx 1 \tag{5}$$

若原假设不成立，则比值将远离 1。因此，可采取 F 检验方法，即在 α 分位数水平下，检验 $F_\alpha(r-1, n-r)$。结合单因素方差分析模型对大脑结构特征和认知行为特征展开分析，可以到影响疾病诊断的关键特征。

2 基于决策树的阿尔茨海默病识别模型

本文构建基于 C4.5 算法的决策树，利用增益比率(gain ratio)作为属性的选择标准。信息熵的计算方式如下：

$$\mathrm{Ent}(\boldsymbol{D}) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \tag{6}$$

式中：p_k 表示当前样本集 $\boldsymbol{D}$ 中第 k 类样本所占的比例，$k=1,2,\cdots,|y|$。

计算用属性 a 对样本集 $\boldsymbol{D}$ 进行划分所获得的“信息增益”：

$$\mathrm{Gain}(\boldsymbol{D},a) = -\mathrm{Ent}(\boldsymbol{D}) - \sum_{v=1}^{V} \frac{|\boldsymbol{D}^v|}{|\boldsymbol{D}|}\mathrm{Ent}(\boldsymbol{D}^v) \tag{7}$$

进而，信息增益率可以定义如下：

$$\text{Gain-ratio}(\boldsymbol{D},a) = \frac{\mathrm{Gain}(\boldsymbol{D},a)}{-\sum_{v=1}^{V}\frac{|\boldsymbol{D}^v|}{|\boldsymbol{D}|}\log_2\frac{|\boldsymbol{D}^v|}{|\boldsymbol{D}|}} \tag{8}$$

决策树采用启发式的方法对输入空间进行划分，选择第 j 个变量 $x^{(j)}$ 和它的取值 s，作为切分变量和切分点，并定义两个区域：$R_1(j,s)=\{x|x^{(j)}\leqslant s\}$ 和 $R_2(j,s)=\{x|x^{(j)}>s\}$；然后，寻找最优切分变量 j 和最优切分点 s[12]。具体可求解：

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2 + \min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{9}$$

对固定输入变量 j 可以找到最优切分点 s，将大于切分点的数据归为一类，小于数据切分点的数据归为一类。

数据集 $\boldsymbol{D}$ 的纯度可由基尼指数得出，具体公式如下：

$$\mathrm{Gini}(\boldsymbol{D}) = 1 - \sum_{k=1}^{|y|} p_k^2 \tag{10}$$

式中：p_k 表示按某个变量划分时，目标变量不同的概率。

最优划分属性的选择可基于各属性的基尼指数进行，其中某个属性 a 的基尼系数具体计算公式如下：

$$\mathrm{Gini\_index}(\boldsymbol{D},a) = \sum_{v=1}^{|y|}\frac{|\boldsymbol{D}^v|}{|\boldsymbol{D}|}\mathrm{Gini}(\boldsymbol{D}^v) \tag{11}$$

在各属性中，基尼指数越小则数据集 $\boldsymbol{D}$ 的纯度越高。因此，可选取划分后所得基尼指数最小的属性作为最优划分属性，并进行分类处理。

3 实验分析

为了验证本文所提方法的有效性，通过功能磁共振成像技术采集 4 182 名健康对照者、1 259 例阿尔茨海默病患者、1 195 例主观记忆抱怨患者(Subjective Memory Complains，SMC)、2 391 例早期轻度认知障碍(Early MCI，EMCI)、4 053 例晚期轻度认知障碍患者(Late MCI，LMCI)数据。

对磁共振成像采集的指标进行主成分建模，得到大脑结构特征的 7 项关键特征：Ventricles、Hippocampus、WholeBrain、Entorhinal、Fusiform、MidTemp、ICV，以及认知行为特征的 8 项关键特征：CDRSB、ADASII、ADASQ4、mmse、ADAS13、RAVLT_immediate、RAVLT_learning、RAVLT_forgetting。

选取将 70%的数据作训练集，即 8 826 条数据记录，剩余 30%数据作训练集。决策树模型训练参数如表 1 所示。

表 1 决策树关键参数

数据切分	叶子节点的最小样本数	节点分裂评价准则	叶子节点的最大数量	内部节点分裂的最小样本数	树的最大深度
0.7	1	Gini	50	2	10

运行 Python 程序，将阿尔茨海默病患者、轻度认知障碍患者以及健康对照者三类进行三分类，得到十折交叉检验的准确率为 99.62%，精准率为 99.62%，召回率为 99.62%，F1 得分为 0.996 2，AUC 指标为 0.999 1。三分类的 ROC 曲线如图 1 所示，混淆矩阵如图 2 所示。图 2 中，0 类表示阿尔茨海默病，1 类表示轻度认知障碍，2 类表示健康对照者。

对比传统支持向量机模型以及朴素贝叶斯判别模型,识别效果的参数对比状况如图 3 所示。从图 3 中不难发现,决策树的阿尔茨海默病疾病识别效果优于支持向量机模型以及朴素贝叶斯判别模型。

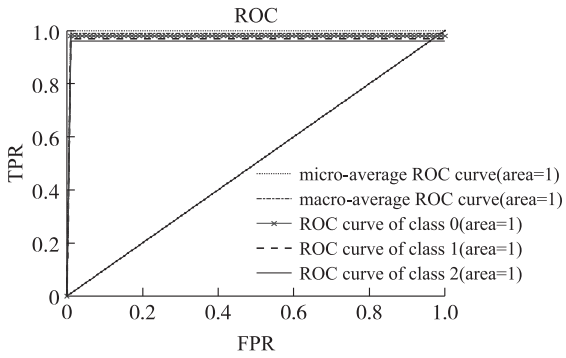

图 1　三分类的 ROC 曲线图

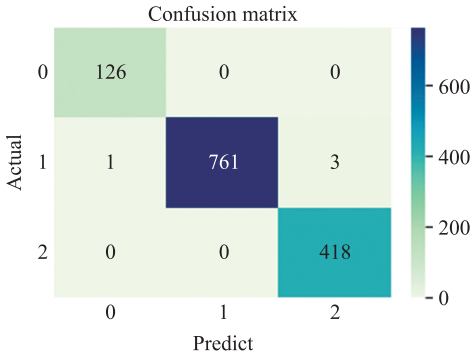

图 2　三分类混淆矩阵图

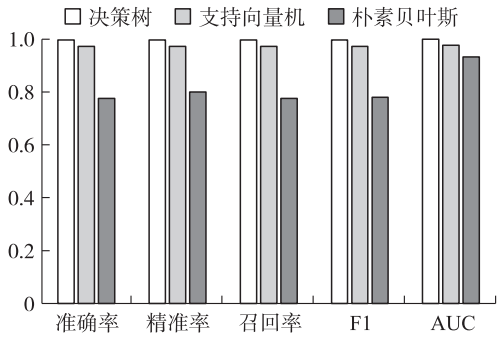

图 3　三分类各种判别模型的性能对比图

运行 Python 程序对阿尔茨海默病患者、早期轻度认知障碍、晚期轻度认知障碍、主观记忆抱怨以及健康对照者进行五分类,得到十折交叉检验的准确率为 99.40%,精准率为 99.40%,召回率为 99.40%,F1 得分为 0.993 9,AUC 指标为 0.999 7。五分类的 ROC 曲线如图 4 所示,混淆矩阵如图 5 所示。图 5 中,0 类表示阿尔茨海默病,1 类表示早期轻度认知障碍,2 类表示晚期轻度认知障碍,3 类表示主观记忆抱怨,4 类表示健康对照者。

对比传统支持向量机模型以及朴素贝叶斯判别模型,识别效果的参数对比状况如图 6 所示。从图 6 中不难发现,决策树的阿尔茨海默病疾病识别效果优于支持向量机模型以及朴素贝叶斯判别模型。

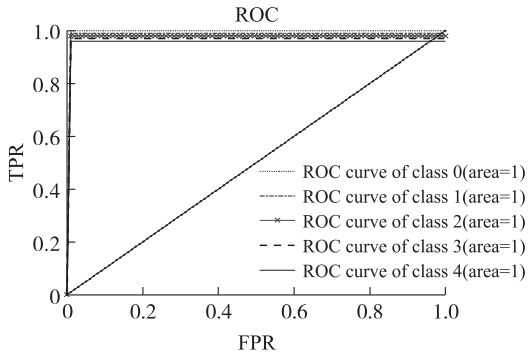

图 4　五分类的 ROC 曲线图

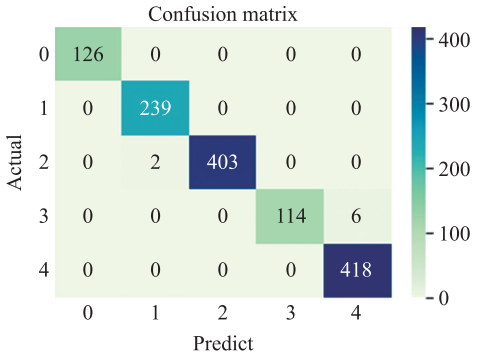

图 5　五分类混淆矩阵图

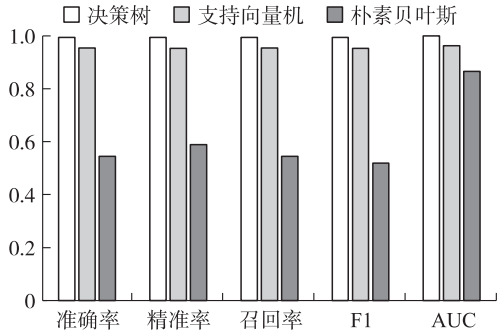

图 6　五分类各种判别模型的性能对比图

4　结论

本文提出了一种基于数据驱动的阿尔茨海默病智能诊断模型,对比其他学者的工作,本文将传统的二分类判别问题(是否患有阿尔茨海默病)延伸到疾病的多分类场景(是否患有阿尔茨海默病、认知行为障碍等),首先,通过差异显著性分析对传感器获取的信号进行特征提取;然后,构造决策树模型分别对大脑结构特征以及认知行为特征进行疾病诊断,并融合不同分类器的结果得到最终结论;最后,利用实验数据验证所提出算法的有效性。结果显

示,在三分类任务中,所提出的算法十折交叉检验的准确率为 99.62%,精准率为 99.62%,召回率为 99.62%,F1 得分为 0.996 2,AUC 指标为 0.999 1;在五分类任务中,所提出的算法十折交叉检验的准确率为 99.40%,精准率为 99.40%,召回率为 99.40%,F1 得分为 0.993 9,AUC 指标为 0.999 7。可见本文方法不仅可以判断患者患病与否,还可以进一步识别患病严重程度,更具应用价值,模型较高的识别准确率更具临床应用前景。在后续的工作中将进一步探讨采用深度学习的方法从磁共振影像数据中获得特征从而实现多分类的模式识别。

参考文献:

[1] 洪欣,黄铠沣,杨晨晖. 基于 Bi-ConvLSTM 时序特征提取的阿尔茨海默病预测 CTISS 模型[J]. 中国图像图形学报,2023,28(4):1146-1156.

[2] 李长胜,王瑜,肖洪兵,等. KPCA 和 Adaboost 算法在阿尔茨海默症功能磁共振影像分类中的应用[J]. 中国医学物理学杂志,2019,36(7):784-788.

[3] 范炤,姚丽丽. 基于功能磁共振成像的个体脑网络在阿尔茨海默病早期诊断中的应用[J]. 生物医学工程研究,2021,40(2):126-130,143.

[4] Kishore P,Usha K C,Kumar M N V S S,et al. Detection and Analysis of Alzheimer's Disease Using Various Machine Learning Algorithms[J]. Materials Today:Proceedings,2021,45:1502-1508.

[5] Balne S, Elumalai A. Machine Learning and Deep Learning Algorithms Used to Diagnosis of Alzheimer's[J]. Materials Today:Proceedings,2021,47:5151-5156.

[6] Zhang Q, Sheng J, Zhang Q, et al. Enhanced Harris Hawks Optimization-Based Fuzzy k-Nearest Neighbor Algorithm for Diagnosis of Alzheimer's Disease[J]. Computers in Biology and Medicine,2023,165:107392.

[7] Lu D, Yue Y, Hu Z, et al. Effective Detection of Alzheimer's Disease by Optimizing Fuzzy K-Nearest Neighbors Based on Salp Swarm Algorithm [J]. Computers in Biology and Medicine,2023,159:106930.

[8] Muhammed Niyas K P,Thiyagarajan P. Feature Selection Using Efficient Fusion of Fisher Score and Greedy Searching for Alzheimer's Classification[J]. Journal of King Saud University: Computer and Information Sciences, 2022,34(8):4993-5006.

[9] Tu Y,Lin S,Qiao J,et al. Alzheimer's Disease Diagnosis via Multimodal Feature Fusion[J]. Computers in Biology and Medicine,2022,148:105901.

[10] Sharma S, Guleria K, Tiwari S, et al. A Deep Learning Based Convolutional Neural Network Model with VGG16 Feature Extractor for the Detection of Alzheimer Disease Using MRI Scans [J]. Measurement: Sensors, 2022, 24:100506.

[11] Zhang F,Petersen M,Johnson L,et al. A Machine Learning-Based Multiple Imputation Method for the Health and Aging Brain Study-Health Disparities[C]//Informatics, 2023,10(4):77.

[12] 张国治,陈康,方荣行,等. 基于 DGA 与鲸鱼算法优化 LogitBoost-决策树的变压器故障诊断方法[J]. 电力系统保护与控制,2023,51(7):63-72.

刘半藤(1984—),男,汉族,浙江省宁波人,博士,教授,硕士生导师,研究方向:智能数据分析与挖掘,医学人工智能技术;

梁婧瑶(1993—),黑龙江哈尔滨人,博士,特聘副研究员,研究方向为阵列信号处理,机器学习,liangjingyao@ zjsru.edu.cn。