

基于机器学习的阿尔兹海默症分类预测

李彩¹, 范焯²

1. 山西医科大学基础医学院, 山西 太原 030001; 2. 山西医科大学转化医学研究中心, 山西 太原 030001

【摘要】目的:应用机器学习方法,将脑结构磁共振(sMRI)、年龄、性别、受教育年限和MMSE量表评分作为特征,对阿尔兹海默症进行分类预测。**方法:**特征选择后,用L1正则Logistic回归、L1正则支持向量机、梯度提升树分别对脑sMRI数据进行分类预测,选出最优模型后引入年龄、性别、受教育年限和MMSE量表评分特征优化模型,用10-折交叉验证评价模型性能。**结果:**L1正则Logistic回归分类效果最好,加入年龄、性别、受教育年限和MMSE评分后预测准确率提高0.89%~11.42%。**结论:**L1正则化Logistic回归模型的sMRI+年龄+性别+受教育年限+MMSE评分特征集对阿尔兹海默症有更好的分类效果,可作为辅助诊断阿尔兹海默症的依据。

【关键词】阿尔兹海默症;机器学习;L1正则Logistic回归;分类预测

【中图分类号】R318;R741

【文献标志码】A

【文章编号】1005-202X(2020)03-0379-06

Classification and prediction of Alzheimer's disease based on machine learning

LI Cai¹, FAN Zhao²

1. School of Basic Medicine, Shanxi Medical University, Taiyuan 030001, China; 2. Translational Medicine Research Center of Shanxi Medical University, Taiyuan 030001, China

Abstract: Objective To classify and predict Alzheimer's disease by machine learning method with brain structural magnetic resonance imaging (sMRI), age, gender, years of education and MMSE score as features. **Methods** After feature selection, L1-regularized Logistic regression, L1-regularized support vector machine and gradient boosting decision tree were used to classify and predict the brain sMRI data. After the optimal model was determined, the model was optimized by several features including age, gender, years of education and MMSE scores, and the performance of the optimized model was evaluated by 10-fold cross-validation. **Results** L1-regularized Logistic regression had the best classification performance. The prediction accuracy of the model optimized by the features of age, gender, years of education and MMSE scores was increased by 0.89% to 11.42%. **Conclusion** The feature set of sMRI, age, gender, years of education and MMSE scores in L1-regularized Logistic regression model has a better performance on the classification of Alzheimer's disease, which can be used as the basis for the auxiliary diagnosis of Alzheimer's disease.

Keywords: Alzheimer's disease; machine learning; L1-regularized Logistic regression; classification and prediction

前言

阿尔兹海默病(Alzheimer's Disease, AD)是一种以认知和智力损害、行为生活能力下降为主要特点的神经退行性脑疾病,发病原因尚不明确,且缺少有效彻底的治疗方案,是临床上导致痴呆的主要原

因^[1-2]。轻度认知障碍(Mild Cognitive Impairment, MCI)是介于正常衰老和痴呆之间的一种过渡状态,在6年后有80%转化为痴呆^[3]。如果在AD早期阶段对患者进行干预治疗,可延缓AD发病时间^[4]。所以,针对早期AD患者的高效诊断、识别AD的前驱阶段的研究十分有必要。目前广泛应用于AD分类预测的办法是应用各种机器学习算法,利用单模态影像学数据或多模态影像数据,或结合人口统计学特征和遗传学资料作为特征变量,以寻求最佳分类预测解决方案。逻辑回归已经用于多种临床应用^[5-7],但对AD分类还少有研究,逻辑回归的大量应用和迅速发展为其用于AD分类提供了良好借鉴。Challis等^[8]采用贝叶斯高斯过程逻辑回归对77例受试者分

【收稿日期】2019-11-20

【基金项目】留学回国人员科技活动项目(619017);山西省回国留学人员科研资助项目(2016-061)

【作者简介】李彩,硕士研究生,研究方向:阿尔兹海默症的分类预测, E-mail: licaizn@163.com

【通信作者】范焯,博士,副教授,研究方向:机器学习在阿尔兹海默症临床诊断的应用, E-mail: Fanzhao316@163.com

类,区分正常老年人(Normal Controls, NC)和早期轻度认知障碍(Early Mild Cognitive Impairment, EMCI)患者、EMCI患者和AD患者之间的差异,结果并不理想,且模型样本量相对特征数量较小,可能存在过拟合问题。此外,该模型没有考虑正规教育年限和性别对疾病的影响。Desikan等^[9]用逻辑回归建立自动核磁共振成像(Magnetic Resonance Imaging, MRI)测量内嗅皮层厚度、海马体积和超边缘回厚度,以识别MCI患者和AD患者的模型,该模型虽然在临床诊断的AD和MCI群体得到推广,但在存在一系列认知障碍和痴呆亚型的临床环境中,这些程序可能不太准确。已有的研究要么是使用单一影像学方法,要么样本含量不大,所得出的准确率都不是很高。由于使用多模态影像学数据检测会给患者带来经济压力,应用其他模型的分类器虽然获有不错的分类精确度,但在临床实际应用中仍存在困难。本研究基于结构性MRI(Structural MRI, sMRI)影像学资料、人口统计学特征(年龄、性别、受教育程度)和简易智力状态检查量表(Mini-Mental State Examination, MMSE),提出一种基于L1正则Logistic回归(L1-Regularized Logistic Regression, L1-LR)特征选择和自动识别NC、EMCI患者、晚期MCI(Late MCI, LMCI)患者和AD患者的辅助诊断工具,使病程分类更加细化,更清楚准确地判断病程阶段。

1 方法

1.1 预处理及特征分析

本研究把NC、EMCI患者、LMCI患者和AD患者这4组作为研究对象,进行sMRI,得到三维结构图像。用Freesurfer工具箱进行一系列算法,sMRI图像经空间标准化、图像平滑、分割、调制等预处理,计算得到各个大脑区域的面积、体积等形态学指标。

提取的形态学指标包括海马亚区体积(Hippocampal Subfield, HS)、皮层体积(Cortical Volume, CV)、皮层表面积(Surface Area, SA)、皮层下体积(Subcortical Volume, SV)、皮层厚度(Cortical Thicknesses, TA),作为特征选择的基础。

1.2 L1-LR方法

在特征选择和分类之前,首先要对数据进行归一化处理,把全部数据映射到0~1。转换的函数为 $(X - \text{Min}) / (\text{Max} - \text{Min})$ 。

1.2.1 L1-LR特征选择 本研究的每位研究对象均有272个形态学指标,涵盖全部大脑的皮层和海马亚区。文献^[10]指出AD病变脑区并非全脑,若使用全脑特征过于冗余,影响分类效果,所以必须进行特征选择。

L1-LR是一种组合的机器学习的特征选择方法,根据每个特征的最大数似然函数(Log-Likelihood, LL)值,选择对分类模型准确率的因变量的贡献率达到一定程度的指标才进入分类模型,剔除无贡献或贡献很小的特征变量。偏差值 $D = -2$ (当前模型的LL值-饱和模型的LL值)。饱和模型指各模型参数相同,似然值等于1。 D 值越小代表当前特征越重要。当加入一个新特征, D 值和似然比的减小等价。当样本量 N 足够大时,似然比呈分布,因此用似然比当作判断某一特征的依据。加入某一特征,如果前后模型偏差值的差大于或等于某一限定值,则认为此特征在类的判定中有比较大的贡献,否则认为此特征冗余。

本研究的特征选择分两部分完成特征空间维数压缩。首先,选择272项sMRI数据进行L1-LR方法筛选,并按照贡献率大小进行排序,并组成特征数据集;然后,在272项sMRI数据的基础上引入年龄、性别、受教育程度、MMSE量表评分,共276项指标,再次使用L1-LR方法筛选,组成新的特征集,得到用于训练分类器的特征集合。L1-LR特征选择的过程为:

(1) $i = 1$, 令 $S = \phi$, $R = (x_1, \dots, x_k)$;

(2) $i = i + 1$, $k^* = \arg \min_k D_k$, $k = 1, \dots, \|R\|$; R 为集合 R 的势;令 $S = S \cup \{x_{k^*}\}$, $R = R \setminus \{x_{k^*}\}$;

(3) 对任何非负 λ , 正则化形式为 $L(\lambda, \beta) = \arg \min \frac{1}{n} \sum_{i=1}^n (y_i - X_i' \beta)^2 + \lambda P(\beta)$; 其中 $P(\beta)$ 代表正则化项;

(4) 若 $D_{i-1} - D_i < [x_{DF}^2]^{-1} (0.01)$ 循环结束; 且 $S = S \setminus \{x_{k^*}\}$; 否则执行(2); 直至选出所有最优特征, 算法结束。

$A \setminus \{a\}$ 代表从集合 A 中剔除元素 a 。在此算法中, 两模型 D 值之差可近似看作自由度为 DF 的 χ^2 分布。 DF 大小取决于评价的特征数, 限值定为 0.01。当 $[x_{DF}^2]^{-1}(p)$ 为限值 p 时, 自由度为 DF 的 χ^2 分布的倒数。

1.2.2 L1-LR分类模型 L1-LR模型是一种稀疏的逻辑回归模型^[11], 模型是在广泛应用于二分类的逻辑回归模型^[12]的损失函数中加入惩罚项, 正则化技术^[13]能解决过拟合问题, 提高鲁棒性, 优化模型分类能力。在样本数相对于样本特征维数有点小的数据集中, L1-LR性能优于其他模型。

响应变量 $y \in \{0, 1\}$, 每次观测的 p 个预测变量值表示成向量, $x = [x_1, x_2, \dots, x_p]$ 对应的响应变量隶属于类别 1 的后验概率:

$$p(y = 1|x) = \phi(\beta^T x) = \frac{1}{1 + \exp(-\beta^T x)} \quad (1)$$

其中, $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ 为回归模型参数。

假定存在 n 个训练样本 $D_n = \{(x^i, y^i)\}_{i=1}^n$ 依据样本数据及隶属类别 D_n 去除 β , 优化损失函数为对数似然函数:

$$\arg \max_{\beta} \{ \log p(\beta | D_n) \} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \log (y^i \varphi(\beta^T x_i) + (1 - y^i)(1 - \varphi(\beta^T x_i))) \right\} \quad (2)$$

依据最小绝对收缩和选择算法的思想, 在逻辑回归的损失函数中加入对模型系数的 L1 范数惩罚项, 得到 L1-LR 模型函数:

$$\arg \max_{\beta} \{ \log p(\beta | D_n) - \lambda \|\beta\|_1 \} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \log (y^i \varphi(\beta^T x_i) + (1 - y^i)(1 - \varphi(\beta^T x_i))) \right\} - \lambda \|\beta\|_1 \quad (3)$$

本研究比较了另外两种常用的二分类模型, 一种是 L1 正则支持向量机 (L1-Support Vector Machine, L1-SVM), 是在 AD 二元分类中最常使用的支持向量机的决定函数中加入 L1 惩罚项, 试图找到性能更好的 SVM 模型^[14]。另一种是梯度提升树 (Gradient Boosting Decision Tree, GBDT), 是一种迭代的决策树算法, 可以灵活处理各种类型的数据, 鲁棒性强^[15]。有研究指出该机器学习算法的预测准确率高于 SVM^[16]。最后选取分类效果最好的模型作为分类预测模型。

本研究选择 10-折交叉验证 (10-fold Cross-Validation) 评价模型性能, 可以确保进行小样本训练的实验结果无偏差估计, 确保测试精确度。具体方法是随机将样本分为 10 份, 随机抽取 9 份作为训练样本, 剩余 1 份为测试样本, 每次试验 L1-LR 模型后得到训练和测试正确率 (或差错率), 重复 10 次, 将 10

次训练准确率和测试准确率的平均值作为 L1-LR 模型最终的分​​类准确率的估计。

2 实验数据

数据由 ANDI 数据库 (Alzheimer's Disease Neuroimaging Initiative) 提供, 受试者年龄选择 55~90 岁, 能够提供独立的功能评估, 排除特定的精神活性药物。数据包括 543 例的 sMRI 脑图像、MMSE 量表评分、年龄、性别、受教育程度这 5 项资料。获得 NC 组共 139 例 (男 65, 女 74); EMCI 组共 220 例 (男 119, 女 101); LMCI 组共 108 例 (男 58, 女 50); AD 组共 76 例 (男 44, 女 32)。

sMRI 数据统一选取场强 3.0T 飞利浦 MRI 扫描仪的数据, BOLD 序列: 射频重复时间/回波时间 (TR/TE) 6.8 ms/3.1 ms, 翻转角 (FA) 9°, 视野大小 (FOV) RL 204 mm、AP 240 mm、FH 256 mm; 分辨率 (1.0×1.0×1.2) mm³, 层厚 1.2 mm, 共 170 层。

3 实验结果

3.1 L1-LR 特征选择结果

3.1.1 272 项 sMRI 特征组结果 首先对实验数据集预处理后, 获得 272 项 sMRI 形态学指标, 使用 L1-LR 特征选择模型进行特征提取。在分类 NC-EMCI 组中, 272 项 sMRI 特征中有 65 个特征被挑选进入下一步骤的 L1-LR 分类预测模型; 同样的, 在对 NC-LMCI、NC-AD、EMCI-LMCI、EMCI-AD、LMCI-AD 分类中, 分别有 37、22、52、38、41 个特征进入 L1-LR 分类模型。特征选择结果按对分类的组别贡献由大到小依次给出, 显示贡献最大的前 10 项特征, 见表 1。

表 1 272 项 sMRI 特征中最重要的前 10 项特征
Tab.1 Top 10 important features of 272 sMRI features

分组	贡献最大的前 10 项特征
NC-EMCI	TA-左颞横回、SV-左侧脑室、SA-右颞横回、SA-左颞下回、SA-左岛叶、SA-左额上回、TA-右额中回下部、TA-左中央后回、TA-右扣带回后部、CV-右眶回等
NC-LMCI	SA-左颞下回、TA-左颞中回、HS-右海马前下托、CV-左梭状回、SA-右楔前叶、CV-左岛盖部、SA-右额极、CV-左缘上回、CV-左前扣带回后部、SA-左距状旁回等
NC-AD	TA-左颞横回、SV-左海马、TA-左颞中回、HS-右海马下托、SA-左颞极、HS-左海马前下托、CV-左额中回下部、TA-右前扣带回后部、SA-左距状旁回、HS-左海马下托等
EMCI-LMCI	CV-左扣带回后部、SV-左侧脑室、SV-左侧脑室、CV-左额上回、CV-右额上回、SA-右梭状回、CV-左前扣带回后部、SA-右顶上小叶、SV-左侧小脑白质、SV-左尾状核等
EMCI-AD	TA-左颞横回、SV-左侧脑室、CV-右额上回、HS-右海马前下托、SA-左颞极、SA-左颞极、SA-右楔前叶、SA-右梭状回、CV-左颞极、CV-左额中回下部等
LMCI-AD	TA-左颞横回、SA-右颞横回、SA-左额眶回内侧、SA-右扣带回后部、HS-右海马前下托、SA-左颞极、SA-右楔前叶、SA-右楔前叶、SA-右梭状回、SA-左眶回等

3.1.2 276项特征结果 为进一步增强分类准确率,本研究在272项形态学指标的基础上,再引入不同模态的3项人口学指标和1项MMSE量表评分。在NC-EMCI、NC-LMCI、NC-AD、EMCI-LMCI、EMCI-AD、LMCI-AD分类中,分别有67、42、11、56、21、20个特征进入L1-LR分类模型。这里同样显示前10贡献率的指标,见表2。

表2 276项特征中最重要的前10项特征
Tab.2 Top 10 important features of 276 features

分组	贡献最大的前10项特征
NC-EMCI	TA-左颞横回、SA-右颞横回、MMSE量表评分、SA-左岛叶、SA-左额上回、SV-左腹侧间脑、TA-右额中回下部、TA-右扣带回后部、年龄、CV-右眶回……性别、受教育年限在该组贡献率排第22、41位
NC-LMCI	TA-左颞横回、SA-左颞下回、MMSE量表评分、HS-右海马前下托、年龄、SA-右楔前叶、CV-右岛盖部、CV-左前扣带回后部、SA-左距状旁回、HS-左海马下托……受教育年限在该组贡献率排第23位
NC-AD	MMSE量表评分、HS-右海马前下托、HS-左海马前下托、性别、HS-左海马下托、SV-左杏仁核、受教育年限、TA-右前扣带回下部、SA-左扣带回峡部、SV-右侧脑室下角……
EMCI-LMCI	CV-左扣带回后部、SV-左侧脑室、SA-左颞下回、CV-左额上回、MMSE量表评分、CV-右额上回、HS-右海马前下托、年龄、SA-右梭状回、CV-左前扣带回后部……受教育年限在该组贡献率排第30位
EMCI-AD	TA-左杏仁核、SV-左侧脑室皮层下体积、MMSE量表评分、SA-左额眶回内侧、HS-右海马前下托、年龄、CV-左缘上回、SV-右脉络丛、CV-右岛盖、SV-左杏仁核……受教育年限在该组贡献率排第16位
LMCI-AD	MMSE量表评分、SA-左额眶回内侧、年龄、SA-左眶回、TA-左颞上回、TA-右三角部、TA-左颞上回、SV-右小脑白质SA-左额极、SA-左额极、SA-右内嗅区……受教育年限在该组贡献率排第12位

3.1.3 特征选择结果分析 272项结果中,只测试形态学指标sMRI时,在识别NC-EMCI、NC-AD、EMCI-AD、LMCI-AD分组时,TA-左颞横回对判断处于哪个病程阶段最为重要,颞横回为听觉皮质区,在AD疾病进程中,听觉的不断弱化是判断病程的重要依据。NC-LMCI组中,SA-左颞下回占分类决策的权重最大,颞下回负责学习和记忆,在进展为LMCI的时候,病人的学习记忆表现出更多差异;在EMCI-LMCI病程很接近难以区分的组中,CV-左扣带回后部起最重要的作用,后扣带回参与情感和自我评价功能,说明在EMCI进化为LMCI的过程中,病人的情感和自我评价功能有较明显的差别。值得注意的是,除了EMCI-LMCI组,其余组HS-左右海马前下托都可作为一个很重要特征去识别分类,尤其是HS-右海马前下托贡献更大,而海马与近期记忆有关,海马体积变化发生在疾病进展全程,说明记忆障碍体现在从发病开始持续到AD阶段。

276项结果中,在引入MMSE量表评分、年龄、性别、受教育程度后,特征选择结果改变,特征贡献率

也发生变化,其中MMSE量表评分作为非常重要的特征用于识别各阶段的疾病,临床上可将MMSE视为必不可少的依据,以提高诊断准确率。年龄和受教育程度是影响NC-EMCI、NC-LMCI、EMCI-LMCI、EMCI-AD、LMCI-AD疾病进展的重要因素。在NC-AD识别中,性别因素占很大比重,但年龄因素被剔除,说明性别差异导致男女患AD的可能性不同,在决定一个人患AD可能性大小的时候,性别比年龄因素更重要。

3.2 不同机器学习分类结果

比较L1-LR、L1-SVM、GBDT的分类准确率,分别是(85.93±2.53)%、(80.73±4.89)%、(71.77±6.79)%。结果显示L1-LR具有更好的分类效果。本研究选择L1-LR作为分类预测AD病程的模型。

为了进一步提高分类准确率,本研究在272项sMRI特征的基础上引入年龄、性别、受教育年限、MMSE评分;经特征选择后,构建另一个L1-LR分类预测模型。比较两种不同特征集合的分类效果,结果如表3、表4所示。

表3 基于两种特征数据集的分类准确率(%)
Tab.3 Classification accuracy based on two different feature datasets (%)

特征集	NC-EMCI	NC-LMCI	NC-AD	EMCI-LMCI	EMCI-AD	LMCI-AD
sMRI	84.02	88.73	86.24	82.04	88.21	86.33
sMRI+年龄+性别+受教育年限+MMSE评分	86.94	91.87	97.66	82.93	94.28	91.24

表 4 基于两种特征数据集的敏感度、特异性、AUC 值
Tab.4 Sensitivity, specificity and AUC values based on two different feature datasets

分组	sMRI			sMRI+年龄+性别+受教育年限+MMSE 评分		
	敏感度/%	特异性/%	AUC	敏感度/%	特异性/%	AUC
NC-EMCI	99.42	83.75	0.937 0	95.25	87.15	0.943 5
NC-LMCI	85.89	96.86	0.951 1	77.82	89.76	0.919 5
NC-AD	66.38	90.00	0.947 2	93.39	100.00	0.997 3
EMCI-LMCI	83.19	57.45	0.874 6	58.27	86.26	0.879 7
EMCI-AD	65.89	87.46	0.965 1	79.11	98.33	0.994 7
LMCI-AD	66.61	75.76	0.880 3	84.11	95.61	0.984 9

相比 272 项特征组模型,276 项特征组 L1-LR 分类模型在识别 NC-EMCI、NC-LMCI、NC-AD、EMCI-LMCI、EMCI-AD、LMCI-AD 组分类准确率依次提高 2.92%、3.14%、11.42%、0.89%、6.07%、4.91%。尤其在早期识别 EMCI、LMCI 和 AD 时,准确率高达 94.28%、91.24%,同时在区分难以鉴别的 EMCI 和 LMCI 时也可以达到 82.93% 的准确率。结合引入的 4 项特征后,特征选择中年龄、性别、受教育年限、MMSE 量表评分占很大贡献率,分类预测准确率的提高,说明在实际临床诊断时这几项指标可以看作

很重要的辅助诊断病程阶段因素,尤其是 MMSE 量表评分的评价。

用受试者工作特征 (Receiver Operating Characteristic, ROC) 曲线下面积 (AUC) 评估两种特征集的分类性能,AUC 值越大,分类性能越好。图 1 为两种特征集的 ROC 曲线图(0、1、2、3 分别代表 NC 组、EMCI 组、LMCI 组、AD 组)。图 1a 为基于 L1-LR 的 272 项特征集的 ROC 曲线下面积,为 $0.925\,9\pm0.386\,2$;图 1b 为基于 L1-LR 的 276 项特征集的 ROC 曲线下面积,为 $0.953\,2\pm0.475\,4$,从图中看出后者有更好的分类性能。

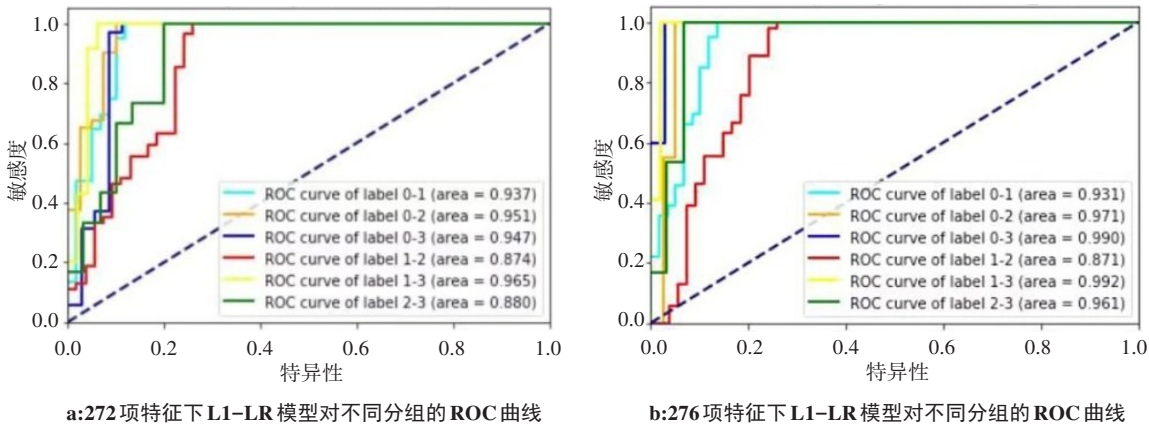


图 1 两种特征数据集的 ROC 曲线
Fig.1 Receiver operating characteristic curves of two different feature datasets

4 结 论

以往 AD 分类研究一般只指出大脑哪些区域发生病变,本研究不仅发现识别分组时的关键特征,还发现关键病变部位是体积或面积的变化,更有助于预测病程进展方向。本研究发现识别各组的特征集中在海马体、海马旁回、扣带回、大部分颞叶、岛叶、梭状回。其中 NC-MCI 变化集中发生在海马下托、海马旁回、海马前下托、上中下颞回、杏仁核、扣带回、梭状回、岛叶;MCI-

AD 变化主要集中在海马体、海马旁回、前扣带回,其中左右海马体积变化和海马亚区体积变化更为明显。已有研究发现在 NC 发展为 MCI 的过程,主要是负责认知功能的海马体和颞叶的萎缩,MCI 发展为 AD 的过程,认知进一步下降,同时负责行为障碍和生活能力的前扣带回萎缩更严重^[17-18]。本研究的研究重点是对 NC-EMCI、EMCI-LMCI 和 LMCI-AD 组的识别,除了与已有研究一致的脑部形态学变化之外,本研究还发现仅在某个分组中的特征变化有助于预测病程判断,在临床判断该

组分类时评价这些指标更有识别度和价值。SA-左岛叶、SA-左额上回、TA-右额中回下部、TA-左中央后回、TA-右扣带回后部、CV-右眶回、TA-右眶回、SA-右额眶回外侧、SA-左额中回后部、CV-左海马旁回、CV-左岛叶、SV-右脉络丛、SA-右额上回、CV-右枕叶外侧部、CV-左眶回、TA-左扣带回峡部、SA-左海马旁回部位的变化仅在判断NC-EMCI分类时占有较大贡献率;而CV-左扣带回后部、CV-左额上回、SV-左侧小脑白质、SA-左中央旁小叶、CV-右中央后回、SV-幕上、TA-左额中回后部、CV-右扣带回后部、CV-右三角部、SA-右额上回、TA-左额上回、CV-右额中回下部、CV-左额中回后部、TA-右楔前叶的变化只发生在EMCI与LMCI的分类中。CV-颞极、CV-右岛盖、TA-左右中央前回、CV-右楔前叶、SV-右丘脑、SA-右中央前回、TA-左额眶回外侧的变化在EMCI到AD的过程中发生得更明显。MMSE量表评分是很重要的影响诊断的因素,临床工作务必认真准确测量表;另外,年龄和受教育年限、性别也作为重要影响因素影响疾病进展。

本研究比较了3种常用机器学习方法,提出了一种基于sMRI图像的AD及前驱阶段的自动识别分类模式,结果证明L1-LR分类器可以作为临床辅助诊断AD病程的有效工具。实验采用sMRI数据+年龄+性别+受教育年限+MMSE量表评分特征集实现最优分类精度,能实现早期识别,起到阻碍病程进展的作用,提高辅助诊断系统的准确率。本研究在NC-AD组中达到97.66%的准确率,明显高于Bi等^[14]利用支持向量机得到的94.44%的准确率和杨晨晖^[19]应用随机森林得到的93%的准确率。在MCI-AD的识别中平均准确率为92.76%,高于Ardekani等^[20]基于多模态随机森林获得的82.3%的准确率,Bi等^[14]的研究也仅达88.73%。NC-MCI组中,本实验达到89.38%的平均准确率,Bi等^[14]的研究仅有81.45%。在与本文数据类型相似的研究中,齐雪丹^[21]比较了支持向量机、随机森林、决策树、K近邻对AD做分类预测,NC-EMCI、NC-LMCI、NC-AD、EMCI-LMCI、EMCI-AD、LMCI-AD组识别最高准确率为77.78%、88.00%、96.45%、81.82%、90.00%、84.21%,本研究使用的L1-LR模型准确率在各分组分别提高9.16%、3.87%、1.21%、1.11%、4.28%、7.06%,且与之不同的是,本研究还将人口学指标和MMSE量表评分引入模型。并且本研究提出的L1-LR分类器仅需检测sMRI,获取人口统计学指标和测评MMSE量表即可,具有经济实惠的社会效益。

sMRI是AD进行分类研究的基础,本研究的后续将通过增加认知评价、人口统计学资料、正电子发射型计算机断层显像、功能性磁共振成像、脑脊液检查等数据类型形成多模态数据,同时加大实验数据量,以获得更高精度、更稳定的分类器用于预测AD病程分类,以

期达到延缓疾病进展、提高生活质量、减轻国家和个人负担的目标。

【参考文献】

- [1] GAUGLER J, JAMES B, JOHNSON T, et al. 2016 Alzheimer's disease facts and figures[J]. *Alzheimers Dement*, 2016, 12(4): 459-509.
- [2] BARKER W W, LUIS C A, KASHUBA A, et al. Relative frequencies of Alzheimer disease, Lewy body, vascular and frontotemporal dementia, and hippocampal sclerosis in the State of Florida Brain Bank[J]. *Alzheimer Dis Assoc Dis*, 2015, 16(4): 203-212.
- [3] Alzheimer's Association. 2015 Alzheimer's disease facts and figures[J]. *Alzheimers Dement*, 2015, 11(3): 332-384.
- [4] Alzheimer's Association. 2014 Alzheimer's disease facts and figures[J]. *Alzheimers Dement*, 2014, 10(2): e47-e92.
- [5] PLITT M, BARNES K A, MARTIN A. Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards[J]. *Neuroimage Clin*, 2015, 7: 359-366.
- [6] TESHNIZI S H, AYATOLLAHI S M. A comparison of logistic regression model and artificial neural networks in predicting of student's academic failure[J]. *Acta Inform Med*, 2015, 23(5): 296-300.
- [7] CHAI H, LIANG Y, WANG S, et al. A novel logistic regression model combining semi-supervised learning and active learning for disease classification[J]. *Sci Rep*, 2018, 8(1): 13009.
- [8] CHALLIS E, HURLEY P, SERRA L, et al. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI[J]. *Neuroimage*, 2015, 112: 232-243.
- [9] DESIKAN R S, CABRAL H J, HESS C P, et al. Alzheimer's disease neuroimaging I. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease[J]. *Brain*, 2009, 132(Pt 8): 2048-2057.
- [10] LIU M, ZHANG D, SHEN D. View-centralized multi-atlas classification for Alzheimer's disease diagnosis[J]. *Hum Brain Mapp*, 2015, 36(5): 1847-1865.
- [11] STEIERT B, TIMMER J, KREUTZ C. L1 regularization facilitates detection of cell type-specific parameters in dynamical systems[J]. *Bioinformatics*, 2016, 32(17): i718-i726.
- [12] WU Y, JIANG X, WANG S, et al. Grid multi-category response logistic models[J]. *Bmc Med Inform Decis*, 2015, 15(1): 1-9.
- [13] SHI J, LIU F, PU H, et al. An adaptive support driven reweighted L1-regularization algorithm for fluorescence molecular tomography[J]. *Biomed Opt Express*, 2014, 5(11): 4039.
- [14] BI X A, SHU Q, SUN Q, et al. Random support vector machine cluster analysis of resting-state fMRI in Alzheimer's disease[J]. *PLoS One*, 2018, 13(3): e0194479.
- [15] DHAMI D S, SONI A, PAGE D, et al. Identifying Parkinson's patients: a functional gradient boosting approach[J]. *Artif Intell Med*, 2017, 10259: 332-337.
- [16] YUAN L, YE W, ZHAO X, et al. Classification of data from electronic nose using gradient tree boosting algorithm[J]. *Sensors*, 2017, 17(10): 2376.
- [17] RONDINA J M, FERREIRA L K, DE SOUZA DURAN F L, et al. Selecting the most relevant brain regions to discriminate Alzheimer's disease patients from healthy controls using multiple kernel learning: a comparison across functional and structural imaging modalities and atlases[J]. *Neuroimage Clin*, 2018, 17: 628-641.
- [18] ZU C, JIE B, LIU M, et al. Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment[J]. *Brain Imaging Behav*, 2016, 10(4): 1148-1159.
- [19] 杨晨晖. 基于随机森林的阿尔兹海默症计算机辅助诊断算法研究[D]. 厦门: 厦门大学, 2016.
- [20] ARDEKANI B A, BERMUDEZ E, MUBEEN A M, et al. Prediction of incipient Alzheimer's disease dementia in patients with mild cognitive impairment[J]. *J Alzheimers Dis*, 2017, 55(1): 1-13.
- [21] 齐雪丹. 基于脑MRI和机器学习的阿尔兹海默病病程分类研究[D]. 太原: 山西医科大学, 2018.
- [22] QI X D. Classification of Alzheimer's disease course based on brain MRI and machine learning[D]. Taiyuan: Shanxi Medical University, 2018.

(编辑:谭斯允)