

DOI: 10.13382/j.jemi.B1902715

机器学习在听性脑干诱发电位数据 分析中的应用研究*

李则辰¹ 唐雨奇² 刘涛¹ 杨东东² 金硕果² 陈超²

(1. 成都信息工程大学 电子工程学院 成都 610225; 2. 成都中医药大学附属医院 成都 610075)

摘要:近年来,许多学者将机器学习算法应用到肌电信号(EMG)数据分析中,取得了良好的效果,但是主要针对手势识别等应用研究,较少有学者对辅助临床诊断进行研究。针对模型训练所需数据较大和机器学习在听性脑干诱发电位(ABR)数据分析中的应用较少两种问题。研究了机器学习方法在基于小型 ABR 数据集数据的计算机辅助诊断中的应用。收集了四川省中医医院的 2 352 份肌电图检查报告,通过设计纳入标准并进行数据清洗,构建了包含 233 份 ABR 报告的数据集。之后,使用数据标准化方法对数据进行数据预处理,再使用随机森林、线性回归、Logistic 回归和人工神经网络 4 种机器学习算法对数据集进行分析处理。4 种算法的性能对比表明随机森林算法性能最优,其准确率、召回率、精确率分别达到了 0.995 7、0.989 7、0.950 0。此外还对每种算法在数据标准化前后的性能进行了比较,表明标准化处理对准确率的提高有一定的提升效果。随机森林算法输出的特征重要性表明,ABR 检查中最重要的指标是 L_latency_5、L_latency_A 和 L_Interval_35,其次是 L_latency_b 和 L_latency_4。这些指标重要性融入上位机软件有助于提高临床诊断效率,在临床应用中具有较高的临床判读潜力。

关键词:机器学习;肌电图;特征提取;随机森林;ABR

中图分类号: TP181;R318.5 **文献标识码:** A **国家标准学科分类代码:** 520.20

Application of machine learning in auditory brainstem response data analysis

Li Zechen¹ Tang Yuqi² Liu Tao¹ Yang Dongdong² Jin Shuoguo² Chen Chao²

(1. College of Electronic Engineering, Chengdu University of Information Technology, Chengdu 610225, China;

2. Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China)

Abstract: In recent years, many scholars have applied machine learning algorithm to electromyogram (EMG) data analysis and achieved good results, but the main direction is gesture recognition, few scholars have applied machine learning to EMG clinical diagnosis. There are two problems: The amount of data needed is large; Machine learning is rarely used in auditory brainstem response (ABR) data analysis. Aiming at these two problems, this paper studies the application of machine learning method in computer-aided diagnosis of ABR data based on small data set. In this paper, 2 352 EMG examination reports of Sichuan Traditional Chinese Medicine Hospital were collected. A data set containing 233 ABR reports data was created by inclusion criteria and data cleaning. Then, four machine learning algorithms, linear regression, logistic regression, random forest and Artificial neural network, are used to analyze and process this data set. According to the performance comparison, the random forest is considered to be the best one, the accuracy, recall and precision of this algorithm is 0.995 7, 0.989 7 and 0.950 0 respectively. In addition, this paper also compares the effect of each algorithm with and without data standardization, this experiment shows that data standardization can improve the accuracy to some extent. The random forest model outputs the importance of each indicator, the most important indicator in ABR are L_latency_5, L_latency_A and L_Interval_35, followed by L_latency_b and L_latency_4. The integration of the importance of these indicators into the upper computer software helps to improve the efficiency of clinical diagnosis and has certain diagnostic evaluation potential in clinical application.

Keywords: machine learning; electromyogram; feature extraction; random forest; auditory brainstem response

0 引言

神经电生理学旨在研究神经和肌肉细胞的生物电活动原理和应用,已被临床应用近一个世纪。以神经电生理学为基础的综合性临床电生理检测技术包括脑电图、肌电图和诱发电位等方面。这些检查技术在神经源性疾病和肌源性疾病的鉴别诊断以及周围神经病变的定性定位、病理范围和预后方面具有重要的临床意义^[1]。

从 21 世纪初开始,有学者开始对肌电图数据进行定量分析,极大地促进了肌电图应用的发展。2003 年, Doherty 等^[2]设计了肌电信号(electromyogram, EMG) 定量分析的方法和初步方案。从此,肌电图定量分析方法成为一个热门研究课题。2005 年, Boe 等^[3]对上肢远端和近端肌肉运动单位的数量进行了定量分析,对基于肌电信号的手势识别具有重要的参考价值。2006 年, Stashuk 等^[4]发表了对肌电图的定量解释。

机器学习算法作为人工智能的核心部分,如传统的机器学习算法、深度学习算法、强化学习算法等,已经在医学领域得到了广泛的应用,在疾病的辅助诊断和治疗中发挥着越来越重要的作用,在临床诊断中的应用已经获得了初步应用成果。近十年来,随着机器学习和深度学习的蓬勃发展和应用,许多学者开始将机器学习算法应用到肌电信号数据分析中。2013 年, Abdulhamit^[5]应用支持向量机(support vector machine, SVM) 方法对加齐安泰普大学神经内科 27 例患者的二头肌电数据进行分类,在计算机辅助诊断中取得了良好的效果。在研究中,他提出了将粒子群优化算法(particle swarm optimization, PSO) 和 SVM 相结合的方法来提高分类效果。2014 年, Yousefi 等^[6]使用了传统的机器学习方法,包括自组织特征映射网络(self-organizing feature map, SOFM)、决策树、贝叶斯技术、人工神经网络(artificial neural network, ANN)、神经模糊系统(neural-fuzzy system, NFS) 等,对 57 名参与者进行了分类,其中 17 名参与者有非特异性手臂疼痛症状,40 名健康。通过比较各种算法的优缺点,得出了 NFS 方法对肌电图数据有较好的分类效果的结论。2018 年, Phinyomark 等^[7]发表了一篇关于肌电图数据在辅助临床诊断中的意义和大数据时代未来发展方向的综述性论文。从数据收集到数据分析, Phinyomark 的论文提出了一个完整的方案和设想。

这些研究虽然在生物医学工程领域取得了很大的成功,但仍存在一些不足。

首先,这些最新的研究成果主要是基于深度学习和大量样本的大规模数据集^[8]。而实际情况却是,即使在较为著名的医院中,其一年内的肌电图检查量在进行分类后也仅有几百份。因此,很难获得数以万计的海量样

本,无法满足深度学习对数据数量的要求。这一实际情况阻碍了深度学习方法的辅助诊断和临床应用研究。相反,传统的机器学习算法通过人工选择合适的特征,即使小数据集也能获很好的预测结果。因此,本文选择传统机器学习的方法对肌电图辅助诊断和治疗进行研究和应用^[9]。

其次,临床人机智能协同医疗是目前的热门研究方向^[10-12]。然而时下关于肌电数据的人工智能(artificial intelligence, AI) 辅助诊断的研究罕有报道,尤其是听性脑干诱发电位(auditory brainstem response, ABR) 数据。因此,本文创新性的将传统机器学习方法应用于肌电检查的临床辅助诊断,具有极大的研究潜力。

本文将传统的机器学习算法应用于 ABR 肌电数据,以小型数据集对其临床应用进行了相关研究。首先,从四川省中医院收集了 10 个月内的 2 352 份肌电图检查报告。设计并使用纳入标准和数据清洗之后得到了 233 份 ABR 报告。然后,将线性回归、Logistic 回归、人工神经网络(ANN) 和随机森林 4 种算法分别应用于数据集,并对 4 种算法的处理结果进行了详细的比较和讨论,并对数据标准化前后的性能效果进行了比较分析,结论说明随机森林性能最优,有很好的临床应用价值和辅助诊断价值。随机森林算法给出的按重要性高低的特征排序,具有很大的临床应用价值。

1 数据

1.1 数据收集

从四川省中医院采集了大约 10 个月左右的 2 352 份肌电图报告。检查报告所用的测试设备是日本光电株式会社生产的肌电诱发电位检测装置 MEB-9 200 K,设备如图 1 所示。该装置具有较强的可扩展性,在肌电信号制造领域处于领先地位,已在成都中医药大学附属医院(四川省中医院)投入使用多年。

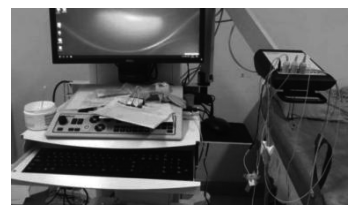


图 1 肌电图检查设备 MEB-9 200 K 外观
Fig. 1 EMG inspection equipment meb-9 200k appearance diagram

本文所收集到的 2 352 份肌电数据包括许多类别,将获取到的数据依照由 2012 年中华医学会神经病学分会肌电图和临床神经电生理学组发布的《肌电图规范化

检测和临床应用共识》^[13-14]进行分类,分类结果如表1所示。

表1 肌电图检查项目的类型

Table 1 Types of electromyography examination items	
神经传导	运动神经传导速度(MNCS)
	感觉神经传导速度(SNCS)
	重复电刺激(Repetitive stimulation)
	F-波(Fwave)
	H-反射(Hreflex)
	瞬目反射(Blink reflex)
体感诱发电位	体感诱发电位(SEP)
	短潜伏期体感诱发电位(SSEP)
	心电图触发的体感诱发电位(ECG-SSEP)
听觉诱发电位	脑干诱发电位(ABR)
	中潜伏期反应(MLR)
	缓慢颅顶反应(SVR)
视觉诱发电位	翻转视网膜诱发电位(PR-VEP)
	眼罩视诱发电位(G-VEP)
	闪光视诱发电位(F-VEP)

由于肌电图检查的子项目数量比较多,本研究仅研究其中比较有代表性的数据,即 ABR 数据,同时也考虑到该组数据有较低的数据维度且较大的样本数量。

科学界通常使用 MCV (motor nerve conduction velocity) 来描述运动神经传导速度检查,而本实验设备 MEB-9 200 K 的输出为 MNCS (motor nerve conduction study)。因为本文的所有研究均基于 MEB-9 200 K 的输出,经过研究决定遵照设备输出进行英文缩写。

ABR 的原理是通过一定的声强刺激听觉器官,并记录听觉系统在皮层中产生的一系列电活动^[15]。ABR 广泛应用于新生儿听力筛查、器质性耳聋和功能性耳聋的测定、听神经瘤的术中监测、耳毒性药物对听力的监测等。ABR 也可作为面瘫、头晕、头痛、耳鸣或听力丧失等临床表现的辅助检查。

常见的 ABR 检查报告包含的数据字段较多,表 2 为报告中关于 ABR 检查的每个数据字段及其相应的含义。表 2 中,延迟时间和声激励强度的单位分别为 ms 和 dB。间隔时间和振幅的单位分别为 ms 和 μV 。字段 DO_MNCSF 表示是否同时进行面部运动神经传导速度检查。

如图 2 所示 MEB-9 200 K 输出的典型的 ABR 数据曲线,原始报告中该图没有标注 x 轴和 y 轴。图 2 中, x 轴和 y 轴分别表示时间和电压,每个水平轴单元代表 1 ms,图中有 8 个独立波形,“+”符号为检查医生通过判断选取的特征点位。

1.2 数据纳入标准

在医学数据处理领域,纳入标准是一种常见的数据预处理方法。这种方法可以看作是一种特殊的数据清洗,目的是纳入感兴趣的数据或者排除一些有问题的数

表2 ABR 数据集各字段及描述

Table 2 Description of fields in ABR data set

数据字段	含义描述
In/Out_Patient	住院状态
Age	患者年龄
Sex	患者性别
Do_MNCSF	是否同时进行 MNCS 面部检查
Abnormal	检查结果
L_Latency_1	左侧延迟 1
L_Latency_2	左侧延迟 2
L_Latency_3	左侧延迟 3
L_Latency_4	左侧延迟 4
L_Latency_5	左侧延迟 5
L_Latency_A	左侧延迟 A
L_Latency_B	左侧延迟 B
L_Smit	左侧声刺激频率
L_Interval_13	左侧 I-III 间隔时间
L_Interval_35	左侧 III-V 间隔时间
L_Interval_15	左侧 I-V 间隔时间
L_Amp_5 A	左侧 V-A 振幅
L_Amp_1B	左侧 I-B 振幅
R_Latency_1	右侧延迟 1
R_Latency_2	右侧延迟 2
R_Latency_3	右侧延迟 3
R_Latency_4	右侧延迟 4
R_Latency_5	右侧延迟 5
R_Latency_A	右侧延迟 A
R_Latency_B	右侧延迟 B
R_Smit	右侧声刺激频率
R_Interval_13	右侧 I-III 间隔时间
R_Interval_35	右侧 III-V 间隔时间
R_Interval_15	右侧 I-V 间隔时间
R_Amp_5 A	右侧 V-A 振幅
R_Amp_1B	右侧 I-B 振幅

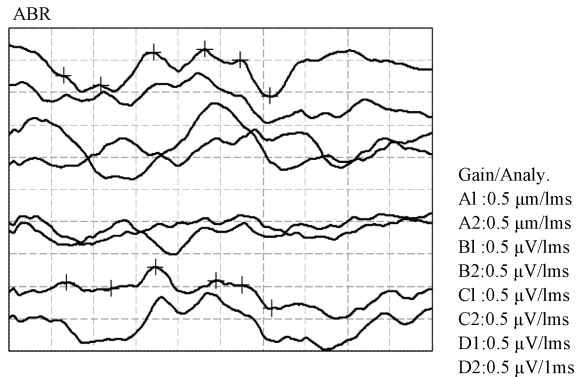


图2 一份典型的 ABR 肌电图检查报告波形

Fig. 2 A typical waveform of ABR inspection report

据。根据纳入标准,数据行将作为一个整体纳入或排除。考虑到这些数据都是从临床环境中收集的,一些数据会或多或少的受到不确定和不可预测的因素的影响。例如,一些病人只进行了部分而非全部的检查。这些数据

不能直接应用于机器学习的训练中,需要事先筛选。为了实现这一目标,必须按照医学数据处理的标准来设计一些纳入标准。根据这些准则,2 352 份原始报告中只有一部分保留下来。本次研究没有设计排除准则。纳入准则如下:

- 1) 报告中的数据应包含被检查人员的完整信息,如年龄、性别、住院时间和报告时间;
- 2) 检查项目、数据和结论的完整性;
- 3) 在检查过程中患者与检查医生积极配合。

1.3 数据清洗

数据清洗是机器学习的关键环节,其方法和结果将直接影响机器学习模型的性能和最终的结论。数据清理将取决于具体任务。常用的方法包括不同类型的任务相关预处理、数据丢弃或修复^[16]。

在本文的 ABR 数据中,有些患者只进行左侧或右侧面部检查,而没有做全部项目的检查。这将导致一些检查结果值为 NaN,NaN 是数据处理中常用的空值表示形式,全称为 Not a Number。在这种情况下,需要进行 NaN 值清洗,常见方法包括删除、均值填充和零填充。考虑到 ABR 数据集本来数据量就不大,删除方法显然不适用;均值填充可能会影响后续机器学习算法的特征选择。因此,本文最终采用了零填充方法。

其次,数据的重复主要是由于在数据采集期间报表的重复打印。重复的报告不会连续出现在数据集表中,因此很难手动删除。本文采用的识别/删除策略是:通过程序进行两两比对,如果两行或多行数据的性别、年龄、检查时间和前两个数据字段相同,则只保留其中一行。

基于上述纳入标准和数据清洗,从 2 352 份原始报告中筛选出了 233 份 ABR 检查报告,构造了数据集。这些数据集以 csv 格式的数据表的形式存储,以供后续程序代码调用。表 3 为数据集大小,其中正例 39 个,反例样 194 个。

表 3 ABR 正常及异常报告统计

Table 3 ABR normal and abnormal report statistics

	样本数
0/正常/反例	194
1/异常/正例	39
总计	233

1.4 数据标签

本文肌电检查医生给出的诊断结论被视为标签(Label)。一般来说,诊断结论大致可分为正常、轻度神经损害、中度神经损害和严重神经损伤共 4 类。在实际的治疗过程中,需要考虑病人是否正常,因此,本文将医生的诊断结论只分为正常和异常两种,对应表 2 的 Abnormal 字段的取值。特别需要注意的是,Abnormal 的

值 0 表示反例,即正常或无疾病;1 表示正例,即异常或有疾病,这是典型的二分类问题。

2 方法

将 ABR 数据集分别应用于四种机器学习算法,即随机森林算法、线性回归算法和 Logistic 回归算法。在每个算法之前需要进行数据标准化,这是机器学习的一个关键步骤。然后在有标准化和无标准化的情况下,对 4 种算法进行性能比较。详细的比较将基于 4 种算法的准确性和召回率。

2.1 数据标准化

数据标准化是机器学习的重要步骤,其核心思想是将原始数据通过变换函数映射到特定的数据区间^[17],这种映射可以是线性的或者非线性的;映射的区域可能是 [0,1] 或者其他区间。其中一个原因是不同特征可能有不同的量纲或者不同的数量级,数量级较大的特征会严重影响其他特征。如果没有标准化预处理,会受到严重影响预测结果,或者造成学习过程中收敛困难甚至无法收敛^[18]。常用的标准化方法包括小数定标标准化、标准差标准化和离差标准化。考虑到 ABR 数据的多维度特点和数据间的相关性,本文采用离差标准化。

离差标准化是将原始数据线性映射到 [0,1],0 对应原始数据的最小值,1 对应原始数据中的最大值,以此消除量纲和数量级的影响,同时保持原始数据之间的线性关系。转换公式可表示如下:

$$x^* = (x - \min) / (\max - \min) \quad (3)$$

式中: \max 表示样本数据的最大值, \min 表示样本数据的最小值。

2.2 线性回归

线性回归方法能表示自变量与因变量之间的线性关系^[19],如式(4)所示。

$$y_{\theta}(X) = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n = \theta^T X \quad (4)$$

其中,自变量是 $X = [x_1, \dots, x_n]$, 因变量是 y_{θ} , 待拟合系数矢量是 $\theta = [\theta_0, \theta_1, \dots, \theta_n]^T$ 。

不同的系数矢量 θ 对应不同的线性回归模型,且具有不同的数据拟合性能。为了找到一个更准确的线性模型来描述输入-输出数据之间的线性关系,需要引入一个函数来描述线性回归模型与实际数据之间的差异,称之为代价函数,常用的代价函数定义如下:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m [y_{\theta}(X^{(i)}) - y^{(i)}]^2 \quad (5)$$

上述代价函数的作用与 L_2 距离或欧氏距离相似。上标(i)代表第 i 个样本,m 表示总训练样本数。由于较小的代价函数值表示拟合模型与实际数据之间更接近,性能更好,因此优化目标是找到所有代价函数 $J(\theta)$ 中具

有最小代价值 $J(\hat{\theta})$ 的最优解 $\hat{\theta}$:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} J(\theta) \quad (6)$$

2.3 Logistic 回归模型

Logistic 回归模型是一种概率非线性回归模型^[20], 其中因变量只有 0 和 1 两个值。假设存在 k 个独立变量 $\mathbf{X} = [x_1, x_2, \dots, x_k]$, 则 $y = 1$ 的后验概率表示为 $p = P(y = 1 | \mathbf{X})$ 。 $y = 1$ 和 $y = 0$ 之间的概率比等于 $p/(1-p)$, 称为事件优势比。采用自然对数的优势比, 可以得到 Logistic 回归变换公式:

$$\operatorname{Logit}(p) = \ln\left(\frac{p}{1-p}\right) \quad (7)$$

将 $\operatorname{Logit}(p)$ 表示为 z , Logistic 回归函数可表示为:

$$p = \frac{1}{1 + e^{-z}} \quad (8)$$

Logistic 回归模型可以用 Logistic 函数来构造, 定义为:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \quad (9)$$

2.4 随机森林模型

随机森林是一种基于套袋法 (Bagging) 思想的集成学习模型。许多简单且判决性能较低的决策树被集成到一个复杂的森林中进行最终预测。任何一个决策树都只能给出较差的预测, 而组合树或森林则能提高预测性能。决策树有多种算法, 目前主流的有 C4.5^[21]、ID3^[22] 和分类回归树算法 (CART)^[23], 每种算法各有优劣, 而随机森林的效果因不同的决策树算法具有较大的差异。根据不同算法的不同特性, 本文采用 CART 作为单个决策树的算法创建随机森林。

基于 CART 算法的决策树选择了基于 Gini 指数的最优切分特征, 且每个分支都是二分支的。Gini 指数在理论上与熵相似。对于随机变量 \mathbf{X} , 其具有 k 个状态和对应概率 p_1, p_2, \dots, p_k , 其 Gini 指数定义为:

$$\operatorname{Gini}(\mathbf{X}) = \sum_k p_k (1 - p_k) = 1 - \sum_k p_k^2 \quad (10)$$

根据式 (10), 伯努利分布的 Gini 指数可表示为:

$$\operatorname{Gini}_B(\mathbf{X}) = \sum_k p_k (1 - p_k) = 2p(1 - p) \quad (11)$$

$X \sim \operatorname{Bernoulli}(p)$

对于训练集 D , 假设有 K 类, 而 C_k 表示对应于第 k 类的数据子集, 则训练集 D 的 Gini 指数为:

$$\operatorname{Gini}(D) = \sum_k \frac{|C_k|}{D} \left(1 - \frac{|C_k|}{D}\right) = 1 - \sum_k \left(\frac{|C_k|}{D}\right)^2 \quad (12)$$

其中, $|\cdot|$ 表示子集的大小。

CART 算法使用特征 A 来分割数据集。如果特征 A 是离散的, 则根据 A 的值 a 可将 D 分为子集 D_1 和 D_2 :

$$D_1 = \{D | A = a\}; D_2 = \{D | A \neq a\} \quad (13)$$

结合上述公式, 可以得到已知特征 A 条件下集合 D 的 Gini 指数为:

$$\operatorname{Gini}(D, A) = \frac{|D_1|}{|D|} \operatorname{Gini}(D_1) + \frac{|D_2|}{|D|} \operatorname{Gini}(D_2) \quad (14)$$

Gini 指数的含义类似于条件熵, Gini (D, A) 值越大, 样品的不确定度越大, 反之越小。因此, 当选择特征 A 时, Gini (D, A) 值越小越好。

在随机森林机器学习训练过程中, 随着决策树的构建, 进一步引入属性的随机选择。具体而言, 假设在当前分割点有 d 个特征, 则从 d 特征集中随机选择一个包含 k 个特征的子集。推荐 k 取 $k = \log_2 d$ 。

2.5 人工神经网络

人工神经网络是 20 世纪 80 年代兴起的神经网络模型, 以信息处理作为切入点对人体神经网络进行抽象建立^[24]。

人工神经网络开创性的意义在于将人类的神经元以数学模型和网络结构的方法进行描述, 证明了单个神经元能够执行逻辑功能, 从而开创了 NN (neural network) 时代。ANN 相较于其他网络主要的有点在于具有较强的自学习功能。

2.6 系统设计

整体处理过程分为如下 4 个步骤: 1) 数据采集和清理; 2) 数据标准化; 3) 4 种机器学习算法的实现; 4) 性能比较与分析。

整个项目使用 Python 编写, 以 Sklearn 包为核心, 辅以 Numpy、Pandas 和 Matplotlib 等第三方开源包进行数值计算和模型实现。

3 成果

ABR 数据集大小为 233, 其中 193 个反例样本和 40 个正例样本, 如表 3 所示。数据集划分策略是: ABR 数据集的 233 个样本被平均分为 2 份, 训练集有 116 个样本, 测试集有 117 个样本。对 ABR 数据集使用上述四种机器学习算法, 分别构建预测模型。

将训练集的数据用于线性回归模型的训练和参数拟合, 对 ABR 测试集进行无放回测试。使用 abnormal 作为因变量, 除了 abnormal 之外的其他数据作为自变量进行模型验证。在进行离差标准化之前的准确率为 0.133 05, 对数据进行离差标准化之后准确率为 0.141 63。标准化后略有提高, 但整体的准确率很低。

与上述线性回归模型的训练略有不同, Logistic 回归模型的训练和参数拟合中, 对 ABR 训练集和测试集进行多次随机划分、参数拟合和交叉验证测试, 最后对准准确率

取平均。在进行标准化前准确率为 0.854 13,在标准化后的准确率为 0.991 56。

随机森林模型训练中,设置 200 棵决策树,每棵树有 3 层深度。使用上次交叉验证,最终的准确率取交叉验证结果的平均值。在进行标准化前准确率为 0.780 28,在标准化后的准确率为 0.995 73。

表 4 为 ABR 数据集的 4 种预测模型在有/无离差标准化的情况下的准确率对比。在每一行中,有离差标准化处理时的准确率总是高于无离差标准化的准确率。其中,线性回归仅有不到 1%的提升;Logistic 回归有大约 14%的提升;随机森林约有 22%的提升,性能改善非常明显,验证了标准化是提高预测性能的有效方法之一。在每一列的纵向比较中,线性回归准确率非常差,不到 15%,无法用于实际预测;每一列中,逻辑回归和随机森林的性能都很好。总体比较之下,带标准化的随机森林算法性能最好,已经达到了 99.6%的较高水平,完全可以在临床医疗中使用。

表 4 各算法判断 ABR 检查结果准确率对比

Table 4 Comparison of accuracy of ABR results by different algorithms

算法名	数据标准化前准确率	数据标准化后准确率
线性回归	0.133 0	0.141 6
Logistic 回归	0.854 1	0.991 5
ANN	0.603 4	0.695 2
随机森林	0.780 2	0.995 7

由表 4 可知,由于线性回归在不论是否数据标准化,准确率均大幅低于 Logistic 回归与随机森林的准确率,因此在后续的研究中先排除了线性回归的研究。在有数据标准化的情况下,随机森林的准确率略高于 Logistic 回归,约高 0.4%。但在没有数据标准化的情况下,随机森林的准确率低于 Logistic 回归。为了找出产生这一问题的原因,本文研究了两种算法在几个交叉验证试验中的准确率。表 5 为随机森林和逻辑回归的详细交叉验证结果。

表 5 随机森林与逻辑回归的交叉验证结果对比

Table 5 Comparison of cross-validation results between random forest and logistic regression

交叉验证次数	随机森林		Logistic 回归	
	离差标准化前准确率	离差标准化后准确率	离差标准化前准确率	离差标准化后准确率
1	0.787 2*	1	0.791 7*	0.957 4
2	0.914 9	0.987 2	0.914 9	1
3	0.893 6	1	0.893 6	1
4	0.847 8	1	0.869 6	1
5	0.913 0	1	0.911 1	1
平均值	0.871 3	0.995 7	0.876 1	0.991 5

从表 5 可以看出,在进行离差标准化前,后 4 次交叉验证中随机森林的准确率与 Logistic 回归没有显著性差异,而第一次交叉验证的准确率显著低于前几次。通过研究小组讨论和相关文献查找,认为这种减少有两个原因。1) 计算机中的随机值不是真实的或理想的随机,而是一种近似的伪随机,这可能导致数据分割中某些特征的过度集中,从而降低了分割的准确率,由这种原因造成的模型准确率骤降,可以通过切分数据集时分别按比例切分正负样本再组合成训练集来解决。2) ABR 检验数据数量少,具体而言,只有 233 个数据,这就可能造成随机森林模型训练不足。

对 ABR 检测数据进行了特征提取,以验证离差标准化是否影响了数据的相关性。从随机森林中提取的 ABR 数据集的特征相关性权重如图 3 所示。

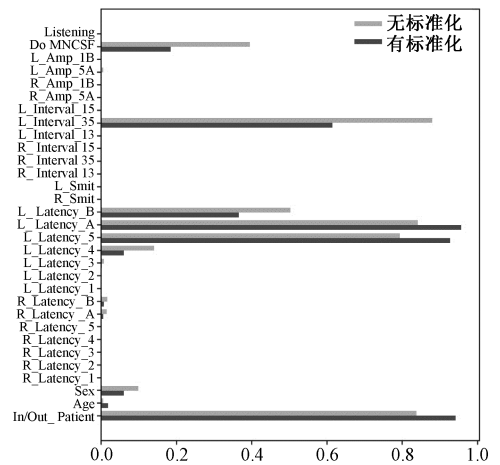


图 3 基于随机森林的 ABR 数据特征提取

Fig. 3 ABR data feature extraction based on random forest

如图 3 所示,在加入离差标准化前后,随机森林提取的特征值权重有明显差异。在实际临床诊断中,年龄特征和性别特征对 ABR 检查的结果没有显著影响,故可以不用考虑。在标准化前后特征重要性虽有明显变化,但 L_Latency_5、L_Latency_A、L_Interval_35、L_Latency_B 和 L_Latency_4 等依旧是最重要的几类指标。如果不考虑两项外部因素:1) 是否入院/出院字段;2) do_MNCSF 字段(与 ABR 一起进行面部运动神经传导速度检查),那么 ABR 检查中最重要的字段依次为 L_Latency_5、L_Latency_A、L_Interval_35、L_Latency_B 和 L_Latency_4 对结果也有一定的影响。

上述特征重要性及排序对于临床诊断具有很好的指导作用。肌电医生在获取测试数据后,在下诊断结论前,可以重点先关注 L_Latency_5、L_Latency_A、L_Interval_35 三个指标,然后再辅以 L_Latency_B 和 L_Latency_4 两个指标,最后再参考其他指标。这样可以极大的提高诊断的效率,也可以为非医学专业人士进行医疗辅助诊断

系统或者决策系统的软件开发提供一些指导和参考。各个模型实际决策结果如表6所示。

表6 模型实际效果

Table 6 Models effect

随机森林实际效果		逻辑回归实际效果		ANN 实际效果	
0	1	0	1	0	1
0	192	2	193	5	156
1	1	38	0	35	34

表6中0和1表示实际检查报告中,由检查医生所判断的病人情况为正常或异常的数量,各个模型下方的0和1表示在模型预测的情况下,由各项机器学习模型判断的对应的病人正常或异常的数量。从表6可以看出,Logistic回归和随机森林的效果远优于人工神经网络。

判断模型的性能往往不能通过单一的指标,除了使用准确率(Accuracy)外,常用的还有召回率(Recall)、精确率(Precision)等^[25]。其中,192名病例被标记为正常,2名病例被随机森林误判为异常。在38名被标记为异常的病人中,只有1名被随机森林误判为正常。因此可以求得随机森林、ANN和逻辑回归的准确率、召回率、精确率等指标如表7所示。

表7 随机森林、ANN和逻辑回归的效果对比

Table7 Comparison of random forest, ANN and logical regression

	随机森林	ANN	Logistic
准确率	0.995 7	0.695 2	0.991 5
召回率	0.989 7	0.808 3	0.974 7
精确率	0.950 0	0.821 1	0.875 0

由表7可以看出,ANN不论是准确率、召回率还是精确率都与随机森林、逻辑回归差距较大。随机森林得到的结果,不管是准确率、召回率还是精确率,均略高于逻辑回归。虽然Logistic回归的准确率、召回率都超过了97%,但是Logistic回归的精确率略低,不足90%。因此本文认为在小样本的ABR数据辅助临床诊断的实际应用中,随机森林的效果优于Logistic回归,前者在临床辅助诊断中应为首选算法。

在使用原始设备配套软件中,对于单一的ABR检查项目,检查需要选取至少12个波形的点,才能判断给出报告以及判断出病人的病理情况是否为异常,而如果再配套软件中嵌入本文所设计的模型,检查医生仅需要选取5个点位以及能直接从刷就诊卡就能获取的病人住院情况以及是否在做了ABR检查的同时做了面部运动神经传导检查,就能判断出病人的状态是否异常。具体如表8所示。

表8 选取点位数变化和预计节省时间

Table 8 Changes in number of selected points and estimated time savings

	嵌入本模型前	嵌入本模型后
选取波形点位数	12	5
预计节省检查时间	30%	

4 讨论

4.1 数据

一方面,由于数据数量少,导致模型训练不足。因此,从ABR数据训练中得到的随机森林模型的性能还有很大的提升空间。可以预见,随着数据集规模的增加,无标准化的随机森林预测模型的性能会进一步提高,离差标准化对算法准确率和特征值的影响也会逐渐减小。另一方面,在数据集较小的情况下,维数对模型有很大的影响。

由于即使在国内级别最高的三甲医院,一年内单项肌电检查数据也难以过千,而国内又缺乏完备的医疗数据共享平台,因此难以收集数以万计甚至百万计的数据量;同时,由于肌电检查数据相较于图像分类或者目标检测,数据维度较小,难以满足深度学习对数据复杂度的要求。因此,综合上述情况,目前传统的机器学习算法在肌电图检测数据处理方面暂时优于深度学习方法。

4.2 未来研究方向

本课题未来的研究方向应该是根据ABR肌电图数据的波形识别,并根据波形的特点确定各种测试的值。目前,这一过程是由检查医生根据波形的特点手动选择,然后填写到打印给病人的检查报告中。结合本文的研究方法和下一步研究的内容,可以形成一个完整的系统。最终的目的是缩短单个病人的肌电图检查的时间,提高医生的检查效率,增加医院病人的吞吐量,充分利用医疗资源。

5 结论

本文通过实地收集ABR肌电图数据,并进行数据纳入,构建数据集等前期工作,然后使用4种机器学习算法构建预测模型。通过对4种算法的准确率性能比较,表明随机森林算法在ABR肌电图数据的诊断中优于线性回归算法和Logistic算法。而且离差标准化是提高准确率性能的有效方法之一。其次,还发现影响ABR临床诊断的结论的特征或指标中,最重要的指标是L_latency_5、L_latency_A和L_Interval_35,其次是L_latency_b和L_latency_4。因此,利用本文得出的特征值重要性结论,对

生物学工程的工作者而言通过对上位机软件的更新,纳入本文所设计的模型,能有效降低检查医生需要手动选取特征点的数量,从而提高检查效率;对检查医生而言,从检查众多的检查参数中快速找到侧重关注点,从而达到减轻工作量和提高工作效率的目的。

本文的研究结果表明,机器学习在数据挖掘和分析中的应用,为临床相关诊疗技术的发展提供了一种创新的思路和方向。此外,该结果也为临床计量资料的诊断和提高医疗效率提供了有意义的参考。

参考文献

- [1] MIOTTO R, WANG F, WANG S, et al. Deep learning for healthcare: Review, opportunities and challenges [J]. *Briefings in Bioinformatics*, 2017, 19 (6), DOI: 10.1093/bib/bbx044.
- [2] DOHERTY T J, STASHUK D W. Decomposition-based quantitative electromyography: Methods and initial normative data in five muscles [J]. *Muscle & Nerve*, 2003, 28(2):204-211.
- [3] BOE S G, STASHUK D W, DOHERTY T J. Within-subject reliability of motor unit number estimates and quantitative motor unit analysis in a distal and proximal upper limb muscle [J]. *Clinical Neurophysiology*, 2006, 117(3):596-603.
- [4] STASHUK D W, PINO L, HAMILTON-WRIGHT A, DOHERTY T, et al. Interpretation of QEMG data [C]. *Proceedings of the 2007 General Meeting of the American Association of Neuromuscular and Electrodiagnostic Medicine*, 2007.
- [5] ABDULHAMIT S. Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders [J]. *Computers in Biology and Medicine*, 2013, 43(5):576-586.
- [6] YOUSEFI J, HAMILTON-WRIGHT A. Characterizing EMG data using machine-learning tools [J]. *Computers in Biology and Medicine*, 2014, 51:1-13.
- [7] PHINYOMARK A, SCHEME E. EMG pattern recognition in the era of big data and deep learning [J]. *Big Data and Cognitive Computing*, 2018, DOI: 10.3390/bdcc2030021.
- [8] 张学红,王光磊,王洪瑞,等.基于中智学理论的冠脉 OCT 图像的钙化斑块分割 [J]. *电子测量与仪器学报*, 2019, 33(10):112-117.
- ZHANG X H, WANG G L, WANG H R, et al. Calcified plaque segmentation of coronary OCT images based on the theory of neutrosophic [J]. *Journal of Electronic Measurement and Instrumentation*, 2019, 33(10):112-117.
- [9] ZIEMNIAK T. Use of machine learning classification techniques to detect atypical behavior in medical applications [C]. *Sixth International Conference on It Security Incident Management & It Forensics*, IEEE Computer Society, 2011.
- [10] 郑悦,景晓蓓,李光林.人机智能协同在医疗康复机器人领域的应用 [J]. *仪器仪表学报*, 2017, 38(10):2373-2380.
- ZHENG Y, JING X B, LI G L. Application of human-machine intelligent collaboration in the field of medical rehabilitation robot [J]. *Journal of Instrumentation*, 2017, 38(10):2373-2380.
- [11] 史庆瑞,马泽亮,周智,等.基于电子舌和模式识别的中成药品种识别方法研究 [J]. *电子测量与仪器学报*, 2017, 31(7):1081-1090.
- SHI Q R, MA Z L, ZHOU ZH, et al. Study on identification methods of Zhong cheng drugs based on electronic tongue and pattern recognition [J]. *Journal of Electronic Measurement and Instrumentation*, 2017, 31(7):1081-1090.
- [12] 胡金通,杨森.智能蓝牙心率检测贴片 [J]. *国外电子测量技术*, 2017, 36(9):114-117.
- HU J T, YANG M. Intelligent Bluetooth heart rate detection patch [J]. *Foreign Electronic Measurement Technology*, 2017, 36(9):114-117.
- [13] 中华医学会神经病学分会肌电图和临床神经电生理学组.肌电图规范化检测和临床应用共识(一) [A]. *中国康复医学会电诊断专业委员会、中国康复医学会康复医学教育专业委员会、南京医科大学第一附属医院.中国康复医学会肌电图与临床电生理学习班资料汇编* [C]. *中国康复医学会电诊断专业委员会、中国康复医学会康复医学教育专业委员会、南京医科大学第一附属医院:中国康复医学会*, 2012.
- EMG and clinical electrophysiology group of Neurology branch of Chinese Medical Association. Consensus on standardized detection and clinical application of EMG (I) [A]. *Electronic diagnosis Professional Committee of Chinese Rehabilitation Medical Association, rehabilitation medical education professional committee of Chinese Rehabilitation Medical Association, the First Affiliated Hospital of Nanjing Medical University. Materials collection of EMG and clinical electrophysiology class of Chinese Rehabilitation Medical Association* [C]. *The First Affiliated Hospital of Nanjing Medical University: China Rehabilitation Medical Association*, 2012.
- [14] 中华医学会神经病学分会肌电图和临床神经电生理学组.肌电图规范化检测和临床应用共识(二) [A].

中国康复医学会电诊断专业委员会、中国康复医学会康复医学教育专业委员会、南京医科大学第一附属医院. 中国康复医学会肌电图与临床电生理学学习班资料汇编 [C]. 中国康复医学会电诊断专业委员会、中国康复医学会康复医学教育专业委员会、南京医科大学第一附属医院: 中国康复医学会, 2012.

EMG and clinical electrophysiology group of Neurology branch of Chinese Medical Association. Consensus on standardized detection and clinical application of EMG (II) [A]. Electric diagnosis Committee of Chinese Rehabilitation Medical Association, rehabilitation medical education committee of Chinese Rehabilitation Medical Association, the First Affiliated Hospital of Nanjing Medical University. Materials collection of EMG and clinical electrophysiology class of Chinese Rehabilitation Medical Association [C]. The First Affiliated Hospital of Nanjing Medical University: China Rehabilitation Medical Association, 2012.

- [15] 黄召辉, 林霖, 王涛. 最大长度序列诱发听性脑干反应的线性与非线性成分引出率和稳定性分析 [J]. 中国生物医学工程学报, 2016, 35(2): 148-154.
HUANG ZH H, LIN L, WANG T. Linear and nonlinear component extraction rate and stability analysis of auditory brainstem response induced by maximum length sequence [J]. Chinese Journal of Biomedical Engineering, 2016, 35(2): 148-154.
- [16] 孟维良, 王胜男. 人体脉搏信号的希尔伯特-黄特征提取 [J]. 电子测量技术, 2017, 40(9): 271-274.
MENG W L, WANG SH N, HHT feature extraction of human pulse signal [J]. Journal of Electronic Measurement and Instrumentation, 2017, 40(9): 271-274.
- [17] KIM M, SHIN S Y, KANG M, et al. Developing a standardization algorithm for categorical laboratory tests for clinical big data research: Retrospective study [J]. JMIR Medical Informatics, 2019, 7(3): e14083.
- [18] ADAM T J, CHI C L. Big data cohort extraction for personalized statin treatment and machine learning [J]. Methods in Molecular Biology, 2019, 1939: 255-272.
- [19] BUCKLEY J, JAMES I. Linear regression with censored data [J]. Biometrika, 1979, 66(3): 429-436.
- [20] KLEINBAUM D G, KLEIN M. Logistic regression [J]. Statistics for Biology & Health, 1994, 68(3): 497-507.
- [21] ZHU X, YAN H, JIAN W, et al. Research and application of the improved algorithm C4.5 on Decision tree [C]. International Conference on Test & Measurement, 2010.
- [22] JIN C, DE-LIN L, FEN-XIANG M. An improved ID3

decision tree algorithm [C]. International Conference on Computer Science & Education, 2009.

- [23] TEIXEIRA A. Discriminant analysis with binary decision tree (CART: Classification and regression tree) [J]. Oral Diseases, 2004, 21(1): 1174-6.
- [24] MESQUITA D P P, GOMES J P P, RODRIGUES L R. Artificial neural networks with random weights for incomplete datasets [J]. Neural Processing Letters, 2019, 50(1): 2345-2372.
- [25] KAMIYA N. Deep learning technique for musculoskeletal analysis [J]. Advances in Experimental Medicine and Biology, 2020(1213): 165-176.

作者简介



李则辰, 2015 年于成都信息工程大学获得学士学位, 现为成都信息工程大学硕士研究生, 主要研究方向为生物医学工程, 机器学习。

E-mail: 307841726@qq.com

Li Zechen received his B. Sc. degree from Chengdu University of Information Technology in 2015. Now he is a M. Sc. candidate at Chengdu University of Information Technology. His main research interests include biomedical engineering and machine learning.



唐雨奇, 2017 年于成都中医药大学获得学士学位, 现为成都中医药大学硕士研究生, 主要研究方向为神经内科。

E-mail: 654020899@qq.com

Tang Yuqi received his B. Sc. degree from Chengdu University of TCM in 2015. Now he is a M. Sc. candidate at Chengdu University of Traditional Chinese Medicine. His main research interest includes internal medicine-neurology.



刘涛 (通信作者), 2000 年于重庆大学获得学士学位, 2003 年于重庆大学获得硕士学位, 2009 年于电子科技大学获得博士学位, 现为成都信息工程大学教师, 主要研究方向为气象卫星遥感信号处理、雷达信号处理、图像处理、人工智能、软件设计等。

E-mail: opentony@cuit.edu.cn

Liu Tao (Corresponding author) received B. Sc. and M. Sc. from Chongqing University in 2003 and 2000, Ph. D. from the University of Electronic Science and Technology of China in 2009, respectively. Now he is a teacher at Chengdu University of Information Technology His main research interests include meteorological satellite remote sensing signal processing, radar signal processing, image processing, artificial intelligence, software design.