

# 基于机器学习的面部运动神经传导 检查数据的研究及应用

唐雨奇<sup>1</sup>, 李则辰<sup>2</sup>, 杨东东<sup>1</sup>, 金硕果<sup>1</sup>, 刘美君<sup>1</sup>, 李永红<sup>2</sup>

(1. 成都中医药大学附属医院神经内科, 四川 成都 610072; 2. 成都信息工程大学电子工程学院, 四川 成都 610225)

**摘要:**为初步研究面部运动神经传导检查数据, 提出运用机器学习方法进行深度数据挖掘、分析, 找出相关性最高的特征值, 以研究其主要的影响因素及探讨临床诊断预测的可能性。收集成都中医药大学附属医院 10 个月的肌电检查报告共 2352 份数据, 筛选符合标准的 575 份报告, 制作数据集, 利用编程的方式对其检查数据和报告结论进行量化分析, 分别建立 KNN、逻辑回归、随机森林、stacking 算法模型, 经过调参选取正确率最高的模型进行特征提取以研究其主要影响因素及研究临床判读预测的可能性。实验结果表明, 一方面在肌电图临床判读中随机森林算法正确率达到 92.69%, 精度为 92.78%, 召回率为 100%, 与逻辑回归相比较  $P$  值为 0.04271, 与 KNN 相比较  $P$  值为 0.00745, 均具有显著统计学意义, 即随机森林模型最适合于面部运动传导神经检查数据分析。另一方面, 运用随机森林方法提取特征值, 能够更加清晰迅速地找出影响面部运动神经病变的最主要因素。通过机器学习挖掘数据, 得出影响面部运动神经传导检查的主要因素为颞支右侧波幅数据和颊支右侧波幅等 8 个特征点, 并提出可使用临床获得的数据集进行判读预测并通过随机森林选取主要的特征点, 具体以减少临床操作时检查点位的形式达到缩短单人检查时间的目的。

**关键词:**机器学习; 面部神经; 肌电图; 特征值; 随机森林

**中图分类号:** TP181

**文献标志码:** A

**doi:** 10.16836/j.cnki.jcuit.2020.05.007

## 0 引言

21 世纪是分子生物学 (molecular biology, MB) 突破基础上精准医学的成熟及人工智能 (artificial intelligence, AI) 渗透到医学各个领域的时代<sup>[1]</sup>。近年来, 大数据分析及人工智能的快速发展, 为临床诊疗技术的发展提供了一个全新的思路及方向。目前已有众多研究者将数据挖掘及机器学习广泛运用于医疗领域<sup>[2-5]</sup>。在疾病的诊断、检出、评估、图像分割、配准与量化均取得了进步<sup>[6-7]</sup>, 而在电生理学相关领域, 尤其在临床肌电图检查中, 虽然已有研究者认为使用肌电图 (EMG) 信号诊断包括面神经在内的多种神经肌肉疾病时, 机器学习算法可被用作诊断神经肌肉疾病的决策支持系统<sup>[8]</sup>, 但相关研究仍较少。

随着对电生理学的研究日渐深入, 肌电图 (EMG) 相关检查技术日臻成熟<sup>[9]</sup>, 尤其在面部运动神经传导检查中运用极广泛。而在实际操作中, 临床医生对于繁多检查结果数据的分析判断仍需耗费大量时间, 因此对此项检查数据的快速诊断和预测对于节约医疗资源尤为重要。但传统的计算机医学数据分析十分繁琐, 主要是根据一些人工预先定义好的特征, 通过计

算得到其特征值。但这些特征对于准确描述医学数据包含的信息并不全面且较僵化, 无法提供个体水平的诊断和预测, 临床应用价值非常有限<sup>[10]</sup>。而使用机器学习的相关方法处理医学数据可弥补传统计算机医学数据分析的不足。旨在收集大量面部肌电检查报告信息和数据的基础上, 使用机器学习的逻辑回归 (LR)、KNN、随机森林、stacking 模型分别进行建模训练并对比正确率, 选择最优的模型, 找出其相关性最高的特征值。初步探讨面部运动神经传导检查数据的影响因素及探讨临床快速诊断和预测的可能性, 以达到节约医疗资源目的。

## 1 数据准备

### 1.1 肌电图检查参考标准

依据文献<sup>[11-12]</sup>及近几年针对面部神经研究的相关文献和研究资料进行分析对比<sup>[13-15]</sup>, 结合临床检查医生的诊断, 得到较为合理的检查项目正常与否的结果。

### 1.2 数据选用标准

数据纳入标准需满足以下几点: (1) 有完整的受检查人员信息: 包括病人的年龄、性别、是否住院、报告

时间。(2)有完整的面部运动神经传导研究(MNCS)数据:包括颊支左右两侧的潜伏期(latency)、波幅(amplitude)、面积(area)、颞支左右两侧的潜伏期、波幅、范围的数据且数据保持完整。(3)患者在检查过程中积极配合医生检查。

在满足纳入标准的情况下,还需要排除以下情况的数据:(1)受检查人员信息缺失。(2)仅有颊支和颞支部分数据,潜伏期、波幅、范围数据不完整。(3)仅有一侧颊支或颞支数据。

2 数据预处理方法

2.1 数据的收集

采用成都中医药大学附属医院肌电图室实测报告数据作为依据,时间为 2018 年 1 月 31 日-10 月 26 日,一共收集 2352 份报告数据。根据上述的排除标准和纳入标准,最终确定使用的报告有 575 份。

2.2 数据集的制作

对原始数据中的变量都进行量化,各项检查也使用英文表示,在数据集中,各项检查的 title 均为英文。为提升计算机执行效率,检查项的其他数据也尽量使用数值来表示,具体对照如表 1 所示。

表 1 数据集标题中英文对照表

中文	英文	中文	英文
是否住院	In/Out_Patient	左侧颊支潜伏期	BB_L_Latency
年龄	Age	左侧颊支波幅	BB_L_Amplitude
性别	Sex	左侧颊支面积	BB_L_Area
检查时间	Date	右侧颊支潜伏期	BB_R_Latency
左侧颞支潜伏期	RT_L_Latency	右侧颊支波幅	BB_R_Amplitude
左侧颞支波幅	RT_L_Amplitude	右侧颊支面积	BB_R_Area
左侧颞支面积	RT_L_Area	是否做了 ABR	Do_ABR
右侧颞支潜伏期	RT_R_Latency	是否做了瞬目反射	Do_Blink
右侧颞支波幅	RT_R_Amplitude	报告结论是否异常	abnormal
右侧颞支面积	RT_R_Area		

其中,是否住院使用 In, Out 表示;年龄用数值表示;性别用数值表示:1 表示男性,0 表示女性;检查时间使用 2018/\*\*/\*\* 的形式表示;潜伏期统一单位为 ms;波幅统一单位为 mV;面积统一单位为 mVms;在检查了面电波图之后是否还做了脑干诱发电位(ABR)使用 Do\_ABR 表示:1 表示做了,0 表示没做;在做了面电波图后是否还做了瞬目反射使用 Do\_Blink 表示:1 表示做了,0 表示没做;报告的结论是否异常使用 abnormal 表示:1 表示检查结果为异常,0 表示检查结果为正常。收集的延迟数据分布情况如图 1 所示。

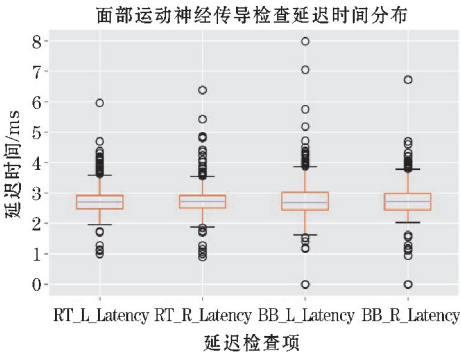


图 1 收集的延迟数据分布情况

如图 1 所示,收集的 RT\_L\_Latency 检查项的均值为 2.749 ms,均方差为 0.474,RT\_R\_Latency 检查项的均值为 2.760 ms,均方差为 0.501,BB\_L\_Latency 检查项的均值为 2.551 ms,均方差为 1.040,BB\_R\_Latency 检查项的均值为 2.534 ms,均方差为 1.008。

收集的波幅数据分布情况如图 2 所示。

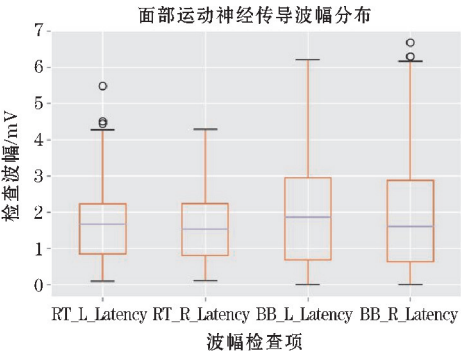


图 2 收集的波幅数据分布情况

如图 2 所示,收集的 RT\_L\_Amplitude 检查项的均值为 1.635 mV,均方差为 1.138,RT\_R\_Amplitude 检查项的均值为 1.569 mV,均方差为 0.889,BB\_L\_Amplitude 检查项的均值为 1.944 mV,均方差为 1.443,BB\_R\_Amplitude 检查项的均值为 1.869 mV,均方差为 1.461。

由于左、右侧颊、颞支面积可以由波幅与延迟时间推算得出,故未在文中以箱型图对面积数据进行描述。

收集到的病人年龄最小为 8 岁,最大为 86 岁,平均年龄为 45 岁。

2.3 数据预处理

常规数据预处理的步骤应该是:数据清洗,数据集成,数据变换,规范化。如此操作是因为获取的数据有很多不完整和无效值,但是经过之前的步骤筛选排除以及数据集制作之后,此时数据仅需要规范化,鉴于面部肌电图数据特征,采用“零-均值规范化”将数据进行规范化处理,其参考公式为

$$x^* = (x - \bar{x}) / \sigma \tag{1}$$

其中, $\bar{x}$  为原始数据的均值, $\sigma$  为原始数据标准差,该方法是目前数据规范化使用的最多的方法<sup>[16]</sup>。

3 数学模型的建立和算法比较

3.1 整体模型构建

由于处理数据较多,方案较为繁琐,由此以整体项目流程图的方式对方案进行说明,总体流程如图 3 所示。

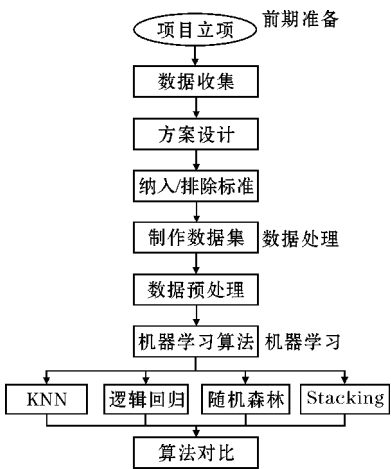


图 3 整体项目流程图

3.2 KNN

KNN(K nearest neighbor)是一种相对简单、较常用的监督学习分类方法<sup>[17]</sup>。算法的基本理念是通过训练集样本 D 中的相似度大小对 D 中样本进行排序,取前 K 个样本的标签众数作为测试样本 T 的标签。相似度由距离表达,计算公式如下:

dist(x,y)=(∑\_{i=1}^n|x\_i-y\_i|^p)^{1/p} (2)

式中 p=1 时,为曼哈顿距离,p=2 时即为最常用的欧式距离。使用 KNN 算法需要寻找最合适的 K 值,通过 [0,30] 的 K 值带入代码段进行错误测试,测试结果如图 4 所示。

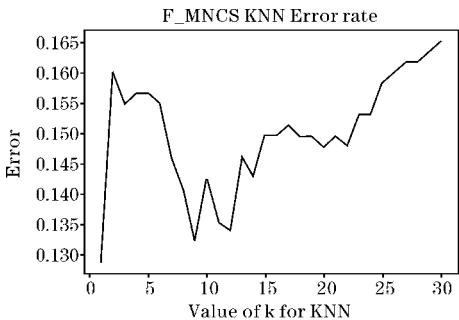


图 4 KNN 错误率

由图 4 可知,错误率最低的点为 k=1 时,因此将 k=1 作为选用 K 值,得到的准确率为0.85547。

3.3 逻辑回归

逻辑回归是利用 Logistic 函数将因变量的取值控

制在[0,1]的区间上,表示取值为 1 的概率的特殊的线性回归<sup>[18]</sup>。令 Logistic 函数为 P,则

P=1/(1+e^x) (3)

逻辑回归模型公式为

ln(p/(1-p))=β\_0+β\_1x\_1+β\_2x\_2+. . . +β\_px\_p+ε (4)

其中,自变量 x\_1,x\_2,⋯,x\_p 可以在任意范围内取值。通过使用逻辑回归算法得到的正确率为0.85729。

3.4 随机森林

随机森林是最经典的 bagging 模型之一,其特点是数据采样随机,特征选择随机,所谓森林,就是很多的决策树并行放在一起。

随机森林的原理是由多棵决策树构成,而构成决策树的算法较多,常见的如 ID3,C4.5,CART 等,决策树在医学中已有较多的实际应用<sup>[19]</sup>,文中采用基于二分法的 CART 算法作为单棵决策树的基础算法构建随机森林。

利用随机森林建模得到的算法正确率为0.9269。

3.5 Stacking 算法模型

Stacking 算法模型的原理是聚合多个分类或者回归模型,以达到更好的机器学习效果。Stacking 模型算法原理图如图 5 所示。

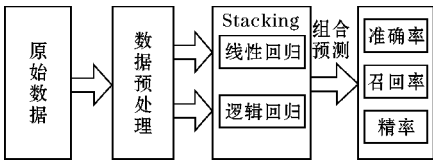


图 5 Stacking 算法原理图

论证使用 Stacking 模型聚合逻辑回归和线性回归两种算法,其正确率为0.8157。

各类算法正确率对比如表 2 所示。

表 2 各项算法正确率

算法名	正确率/%
Logistic 回归	85.72
KNN	85.55
随机森林	92.69
Stacking 算法	81.56

4 面部神经传导检查主要影响因素

由上述各类算法正确性比较可得,随机森林算法的验证准确率最高,因此将随机森林算法作为最优模型,通过随机森林提取各项数据的特征值。由于随机森林得到的特征值的数值较小,使用对数变化将其数值扩大以方便观察,经过变换之后的各项特征值数值



如图6所示。

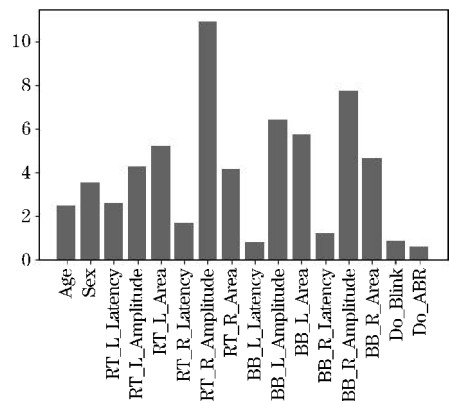


图6 随机森林提取特征值详情

由图6可以看出,在利用随机森林特征提取后得出最为重要检查点位参数是分别是 RT\_R\_Amplitude、BB\_R\_Amplitude、BB\_L\_Amplitud、BB\_L\_Area、RT\_L\_Area、BB\_R\_Area、RT\_L\_Amplitude、RT\_R\_Area。使用以上8个点位构建算法模型,检查判断准确率为91.02%,与报告所使用的检查方法相比较,结果如表3所示。

表3 精简点位前后效果对比

	精简点位前	精简点位后
准确率/%	92.69	91.02
检查电位数	12	8
预计节约时间/%	25	

由于 Stacking 模型算法的准确率较低,因此最终模型参数对比以逻辑回归、随机森林和 KNN 为主,模型的各项指标如表4所示。

表4 模型各项指标

	逻辑回归	随机森林	KNN
准确率/%	85.72	92.69	85.55
精度/%	45.36	92.78	90.72
召回率/%	94.62	100	70.96
P	0.04271	0.00745	

由表4可知,随机森林在准确率、精度、召回率上相较于逻辑回归和KNN都有较大优势,表中P值分别为逻辑回归与随机森林,KNN与随机森林对比的P值,此P值是上述数据结合模型多次交叉验证的准确率得到。两个P值均小于0.05,具有显著统计学差异,因此认定随机森林模型为最优模型。

由随机森林模型输出可知,当检查医生判断为正常时,机器学习模型输出也为正常的检查报告为90例;检查医生判断为异常时,机器学习模型判断为异常的报告为478例;医生判断为正常而机器学习误判为异常的报告有7例。随机森林模型的效果基本令人满意。

5 讨论

在面部运动神经传导检查大数据挖掘分析上创新出了一条新的思路,提出一种临床实测数据与机器学习相结合的预测方法,尤其在颞支和颊支的诊断预测上,随机森林算法判读正确率达到92.69%。在面部运动神经传导检查中,下一步将进一步验证随机森林以及其他机器学习算法在其他类别的肌电图检查数据中判读和诊断评估的临床潜力,如从肌电图检查的连续时间信号入手,结合文中方法,检查医生仅需要安装电极就能实现自动化神经传导检查。

6 结论

实验结果表明,运用机器学习的方法进行深度的数据挖掘、分析,为临床面神经相关诊疗技术的发展创新提供了一个全新的思路及方向。随机森林算法对面部运动神经传导检查数据的判读和特征提取尤为合适,为临床实测数据的诊断提供了重要的参考意见。在实际应用中,通过特征提取的方法能够减少临床检查医生的点位选取而基本不影响检查结果,从而节省单个病人的检查时间,提高医疗效率。

参考文献:

[1] 严律南. 肝脏外科发展三阶段论及中国医生的贡献[J]. 中国普外基础与临床杂志,2018,25(4):388-390.

[2] De Silva K, Jönsson D, Demmer RT. A combined strategy of feature selection and machine learning to identify predictors of prediabetes[J]. J Am Med Inform Assoc,2020,27(3):396-406.

[3] Dinh A, Miertschin S, Young A, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning[J]. BMC Med Inform Decis Mak,2019,19(1):211.

[4] van Os HJA, Ramos LA, Hilbert A, et al. Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke: Potential Value of Machine Learning Algorithms[J]. Front Neurol,2018(9):784.

[5] 何易晨,黄杰,熊仕奇,等. 慢性阻塞性肺疾病的健康管理系统设计[J]. 成都信息工程大学学报,2018,33(3):250-257.

[6] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective[J]. Artif Intell Med,2001(23):89-109.

- [7] Powell S, Magnotta VA, Johnson H, et al. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures[J]. Neuroimage, 2008(39):238-247.
- [8] Yaman E, Subasi A. Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification[J]. Biomed Res Int, 2019;13.
- [9] 李玲, 凡子莲, 戴东璟, 等. 听性脑干反应 V 波反应阈应用于听力测定的意义[C]. 国际数字医学会、Digital Chinese Medicine. 湖南中医药大学学报 2016/专集: 国际数字医学会数字中医药分会成立大会暨首届数字中医药学术交流会论文集. 国际数字医学会、Digital Chinese Medicine: 湖南中医药大学期刊杂志社, 2016:666-667.
- [10] 文宏伟, 陆菁菁, 何晖光. 机器学习在神经精神疾病诊断及预测中的应用[J]. 协和医学杂志, 2018, 9(1):19-24.
- [11] 党静霞. 肌电图诊断与临床应用[M]. 北京: 人民卫生出版社, 2005(12):17-484.
- [12] 王心刚. 神经电生理诊断技术规范[J]. 现代电生理学杂志, 2017, 24(2):117-126.
- [13] 王心刚. 神经电生理诊断技术规范(续)[J]. 现代电生理学杂志, 2018, 25(1):55-62.
- [14] 杨丽艳, 秦少福, 罗苑媚. 肌电图对面神经麻痹患者病情评估的研究进展[J]. 神经病学与神经康复学杂志, 2013, 10(1):31-32.
- [15] 叶爱萍. 面神经麻痹肌电图及面神经传导速度的观察[J]. 中国临床康复, 2003(1):106.
- [16] 林恬恬. 数据挖掘定量与定性相结合对李平教授冠心病诊疗规律的研究[D]. 北京: 北京中医药大学, 2018.
- [17] Song Y, Huang J, Zhou D, et al. IKNN: Informative K-Nearest Neighbor Pattern Classification[C]. Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings. DBLP, 2007.
- [18] 戴维. 逻辑回归解决文本分类问题[J]. 通讯世界, 2018(8):266-267.
- [19] 瞿海斌, 毛利锋, 王阶. 基于决策树的血瘀证诊断规则自动归纳方法[J]. 中国生物医学工程学报, 2005(6):709-711.

## Research and Application of Facial Motor Nerve Conduction Examination Data based on Machine Learning

TANG Yuqi<sup>1</sup>, LI Zechen<sup>2</sup>, YANG Dongdong<sup>1</sup>, JIN Shuoguo<sup>1</sup>, LIU Meijun<sup>1</sup>, LI Yonghong<sup>2</sup>

(1. Department of neurology, Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu 610072, China; 2. College of Electronic Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

**Abstract:** In order to conduct a preliminary study on facial motor nerve conduction examination data, the method of machine learning is proposed to conduct in-depth data mining and analysis to find the characteristic values with the highest correlation, so as to study the main influence factors and explore the possibility of clinical diagnosis prediction. We collected 10 months electrical inspection report totally 2352 data from Chengdu University of Traditional Chinese Medicine medical and selected 575 reports accord with standard to make datasets, we use programming to quantitatively analyze the inspection data and report conclusion, and building KNN, logistic regression, random forests, stacking algorithm model respectively. The model with highest accuracy is selected to extract features after adjustment of arguments, to research the main influence factors and the clinical interpretation prediction possibilities. Experimental results show that, on the one hand, random forest algorithm in emg clinical interpretation accuracy reached 92.69%, the precision is 92.78%, the recall rate was 100%,  $P$  value is 0.04271 compared with logistical regression and 0.00745 compared with KNN, mentioned above all make great sense. , namely the random forest model is best suited for facial motor conduction nerve examination data analysis, on the other hand, the use of random forests method to extract the characteristic value, can more clear quickly find out the main factors affecting facial movement neuropathy. This research mines data through machine learning, it is concluded that the main factors affecting facial motor nerve conduction check are temporal right amplitude data and buccal branch on the right side of amplitude and so on eight feature points. It is proposed that the clinical data sets can be used for interpretation prediction. By randomly selecting main feature points, specifically reduce the inspection points in clinical operations, to achieve the goal of reducing single person inspection time.

**Keywords:** machine learning; facial nerve; electromyography; characteristic; random forests