Statistics and Application 统计学与应用, 2023, 12(2), 267-273
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122028

Hans汉斯

基于传统方法和机器学习对帕金森病语音信号早期预警模型的对比研究

罗成敏

云南师范大学数学学院，云南 昆明

收稿日期：2023年3月7日；录用日期：2023年3月27日；发布日期：2023年4月14日

摘　要

近年来，研究发现大约90%以上的帕金森病(PD)患者在疾病的早期阶段表现出某种形式的声音障碍。本文从UCI机器学习存储库中选取经最先进的语音信号处理技术提取252名受试者的语音信号特征数据，运用Logistic回归和决策树、Bagging、随机森林建立PD语音信号早期预警模型。研究发现，基线特征、梅尔频率倒谱系数和小波变换与PD有明显的显著关系。经对比均方误差发现，无论是长期预测还是短期预测决策树预测模型的误判率约为0.10，能高精度分类出受试者是否患有帕金森病。Logistic模型的预测效果都显著不如机器学习预测，主要原因与数据的特殊性和复杂性相关，而非只在于模型本身。利用统计的方法在受试者监测早期进行预警分析，可提前做好预防准备，减轻频繁就医的麻烦。

关键词

PD语音信号，Logistic回归，决策树，Bagging，随机森林

Based on Traditional Methods and Machine Learning for Parkinson's Disease Speech Signals: A Comparative Study of Early Warning Models

Chengmin Luo

School of Mathematics, Yunnan Normal University, Kunming Yunnan

Received: Mar. 7th, 2023; accepted: Mar. 27th, 2023; published: Apr. 14th, 2023

Abstract

In recent years, studies have found that about 90% of people with Parkinson's Disease (PD) exhibit

文章引用: 罗成敏. 基于传统方法和机器学习对帕金森病语音信号早期预警模型的对比研究[J]. 统计学与应用, 2023, 12(2): 267-273. DOI: 10.12677/sa.2023.122028

some forms of sound disorder in the early stages of the disease. In this paper, the speech signal feature data of 252 subjects were extracted from the most advanced speech signal processing technology from the UCI machine learning repository, and the PD voice signal early warning model was established by using Logistic Regression, Decision Tree, Bagging, and Random Forest. It was found that the Baseline characteristics, FMCC and wavelet transform had obvious significant relationships with PD. After comparing the mean squared error, it is found that the false positive rate of the Decision tree prediction model is about 0.10, whether it is a long-term prediction or a short-term prediction, which can classify whether the subject has Parkinson's Disease with high accuracy. The prediction effect of Logistic models is significantly inferior to that of machine learning, mainly because of the particularity and complexity of the data, not just the model itself. The use of statistical methods to carry out early warning analysis in the early stage of subject monitoring can prepare for prevention in advance and reduce the trouble of frequent medical treatment.

Keywords

PD Voice Signal, Logistic Regression, Decision Tree, Bagging, Random Forest

1. 引言

帕金森病[1] [2] [3] (Parkinson's Disease, PD)是一种常见的神经系统变性疾病，老年人多见，平均发病年龄为 60 岁左右，40 岁以下起病的青年帕金森病较少见。据研究统计，我国 65 岁以上人群 PD 的患病率大约是 1.7%。大部分帕金森病患者为散发病例，仅有不到 10%的患者有家族史。PD 患者的持续增加引起了国内外相关研究者和医疗机构的高度重视，若是能在早期发现受试者有患帕金森病的趋势，就能及时进行早期干预，减小患病率。

近年来，研究发现大约 90%以上的帕金森病(PD)患者在疾病的早期阶段表现出某种形式的声音障碍，最近的 PD 远程诊断研究集中在检测受试者持续的元音发音或连续语音造成的声音障碍。故本文运用传统统计方法(Logistic 回归)和机器学习分类方法(决策树、Bagging、随机森林)分别对语音信号建立早期预警模型，分析各变量之间的关系，利用模型误判率和均方误差对比传统统计方法和机器学习分类方法对语音信号数据的适应性和预测准确性，分析出各模型之间差异性的原因，找出较佳的语音信号早期预警模型。

2. 模型简介

2.1. Logistic 回归

Logistic 回归[4]主要应用于二分类问题，适用于本文研究的语音信号数据。Logistic 回归模型如下所示：

$$\begin{aligned} &y = X\beta + \varepsilon \\ &y = N\left(\mu, \sigma^2 I\right) \text{ 其中 } \mu = E(y) = X\beta \\ &g(\mu) = X\beta \end{aligned} \tag{1}$$

其中 $y = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$ 表示因变量， $X = \{x_{ij}\}, (i = 1,2,\cdots,n; j = 1,2,\cdots,m)$ 表示自变量矩阵，模型误差项为 ε 服从独立正态分布， $g(\cdot)$ 称为连接函数。

对于两水平因变量 y，若它满足 Bernoulli 分布 Bernoulli(p)，则分布的均值记为 $\mu = p$。其中，把值域[0, 1]转换成实轴函数作为连接函数，能满足这种变换的连接函数有以下两种：一种是 logit 函数，即 $g(p) = \log\left(\frac{p}{1-p}\right)$，该种形式的广义线性模型称为 Logistic 模型：$\log\left(\frac{p}{1-p}\right) = X\beta$。另一种是正态积累分布函数的逆函数，即 $g(p) = \Phi^{-1}(p)$，该种形式的模型称为 Probit 模型：$\Phi^{-1} = X\beta$。

由于 Logistic 模型和 Probit 模型的预测精度差别极小，一般采用一种进行预测即可。故此处介绍 Logistic 模型的两种等式形式，其中观测值 $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})^{\mathrm{T}}, i = 1, 2, \cdots, n$，可使用最大似然估计来估计未知参数 $\beta = (\beta_1, \beta_2, \cdots, \beta_m)^{\mathrm{T}}$。

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^{\mathrm{T}}\beta, i = 1, 2, \cdots, n; \tag{2}$$

$$p_i = \frac{\mathrm{e}^{x_i^{\mathrm{T}}\beta}}{1+\mathrm{e}^{x_i^{\mathrm{T}}\beta}}, i = 1, 2, \cdots, n. \tag{3}$$

2.2. 决策树

决策树[4] (Decision Tree) (分类树)是经常被使用的分类方法之一。它是一种监督学习，所谓监督学习就是在给定固定样本，每个样本都有其属性和类别，且这些类别则是事先确定的，那么通过反复学习、训练得到一个成熟的分类器，该分类器可以对新的观测值进行预测并给出正确的分类。决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率。它是一种简单的模型，但它又是一些预测精度很高的机器学习方法的基本模块。

2.3. Bagging

Bagging [5]是一个基于自助法抽样的组合方法，可用于解决回归问题和分类问题。由于本文所研究的语音信号数据中的因变量为定性数据，此处利用 Bagging 分类来处理该问题。Bagging 分类是一个简单的基于分类树的组合方法。它从训练样本中做多次放回抽样，每次建立一个分类树，假设一共建立 A 棵树。对于每一个新的观测值，通过 A 棵树得到 A 个预测结果，然后，按照简单少数服从多数的原则来投票确定该观测值属于哪一类。

2.4. 随机森林

随机森林分类[5] (Random Tree)与 Bagging 分类非常相似，也是从原始数据中抽取一定数量数据的自助法样本。但与 Bagging 分类的区别在于，每个节点在所有竞争的自变量中，随机是选择几个(而不是所有变量)来竞争拆分变量。而选择个数是由选项 mtry 决定的。随机森林的这种随机选择少数自变量来竞争节点拆分变量的做法使得一些弱势变量有机会参加建模，可能会揭示仅仅靠一些强势变量无法发现的数据规律，可利用各种方法从不同角度展示自变量的重要性，并且计算出 OOB 交叉验证误差。

2.5. 均方误差

本文选用均方误差(RMSE)，定义为：

$$\mathrm{MSE} = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{4}$$

其中 y_i 表示测试集中原因变量，$\hat{y}_i$ 指的是依据训练集训练得到的相应模型对测试集第 i 个因变量的预测。如果 RMSE > 1，则说明模型精度极差，不可取。

3. 数据描述

3.1. 数据来源与可靠性

本文数据来自于 UCI 机器学习存储库[6] [7]，选取经最先进的语音信号处理技术提取 252 名受试者的语音信号特征数据，数据真实可靠，可以用于本文研究并赋予本文数据价值。本研究中使用的数据来自伊斯坦布尔大学医学院神经内科的 188 名 PD 患者(107 名男性和 81 名女性)，年龄在 33 至 87 岁(65.1 ± 10.9)之间。对照组由 64 名健康个体(23 名男性和 41 名女性)组成，年龄在 41 至 82 岁之间(61.1 至 8 ± 9)。在数据收集过程中，麦克风设置为 44.1 KHz，在医生检查后，从每个受试者中收集元音/a/的持续发声，重复三次。经各种语音信号处理算法得到 8 个一级指标和 754 个二级指标变量，该数据变量解释如下表 1 所示。

Table 1. Table of explanations for secondary indicator data variables
表 1. 二级指标数据变量解释表

一级指标		二级指标
变量名	描述	个数
Baseline Features	基线特征	23
Intensity Parameters	强度参数	3
Formant Frequencies	重要特征	4
Bandwidth Parameters	带宽参数	4
Vocal Fold	声带振动	22
MFCC	梅尔频率倒谱系数	84
Wavelet Features	小波变换的特征	182
TQWT Features	TQWT 特征	432

3.2. 变量重要性

由于本文所选用的数据量庞大，且语音信号数据具有一定的复杂性和特殊性，故本文根据语音信号数据提供者和数据本身的性质，将其分为 8 个模块，并对这 8 个模块的数据进行探索，得出变量的重要性以及特征性，为医护人员和医疗机构提供重要的帕金森病相关信息。

此处通过 Logistic 回归、决策树、Lagging 和随机森林模型，分别对 8 个模块的数据进行变量重要性观测和局部变量重要性作对比。通过对比分析出四种方法的变量重要性相差不大，在总体中重要的变量在局部中也重要。反之，总体不重要的变量，在局部也不重要。针对本文中的语音信号数据分析得出基线特征、梅尔频率倒谱系数和小波变换三种数据特征，在总体上对早期预警帕金森病有着特殊的意义，值得相关研究人员和医疗机构重视。

4. 模型的建立与分析

4.1. Logistic 模型的建立与分析

因为本文所使用数据极具复杂性，有 754 个变量，756 个观测值。对于所要建立的 Logistic 模型来说，

当自变量有较多的定性变量或者定性变量的水平比较多时，Logistic 回归时基本无法进行。本文语音数据除因变量为定性变量之外，其余自变量都为定量变量，故可建立 Logistic 模型。在 Python 建模过程中，由于起初假设值域 pt = 0.5 得到的预测结果不太理想，训练时的误差为 0.16，测试时更是高达高达 0.31。故对值域 pt 在区间[0.01, 0.99]内做了一个训练调整。最终寻找到当值域 pt = 0.2744 时，使 logistic 模型预测效果达到最佳，其误差由原来的 0.31 降为 0.28，但 Logistic 模型预测精度任然较差。

4.2. 决策树分类模型的建立与分析

决策树是机器学习分类里面的基本分类器，能够通过竞争原则去选择拆分变量或结束生长。此处选用 Gini 不纯度来作为分类树竞争原则，进行拆分变量及判别准则。

$$\sum_{k=1}^{756} p_k\left(1-p_k\right)=\sum_{k=1}^{756} p_k-\sum_{k=1}^{756} p_k^2=1-\sum_{k=1}^{756} p_k^2 \tag{5}$$

其中 $\sum_{k=1}^{756} p_k=1$， p_k 表示在一个节点的观测值中属于第 i 类的比例。若所有观测值为同一类，那么 Gini 不纯度的度量等于 0。故我们要做的拆分工作即：所选择的拆分变量要使得父节点的 Gini 不纯度与子节点 Gini 不纯度相差最大。二子节点的 Gini 不纯度应该为各个子节点观测值数目比上各个子节点 Gini 不纯度的加权平均来计算。拆分过程如下图 1 所示：

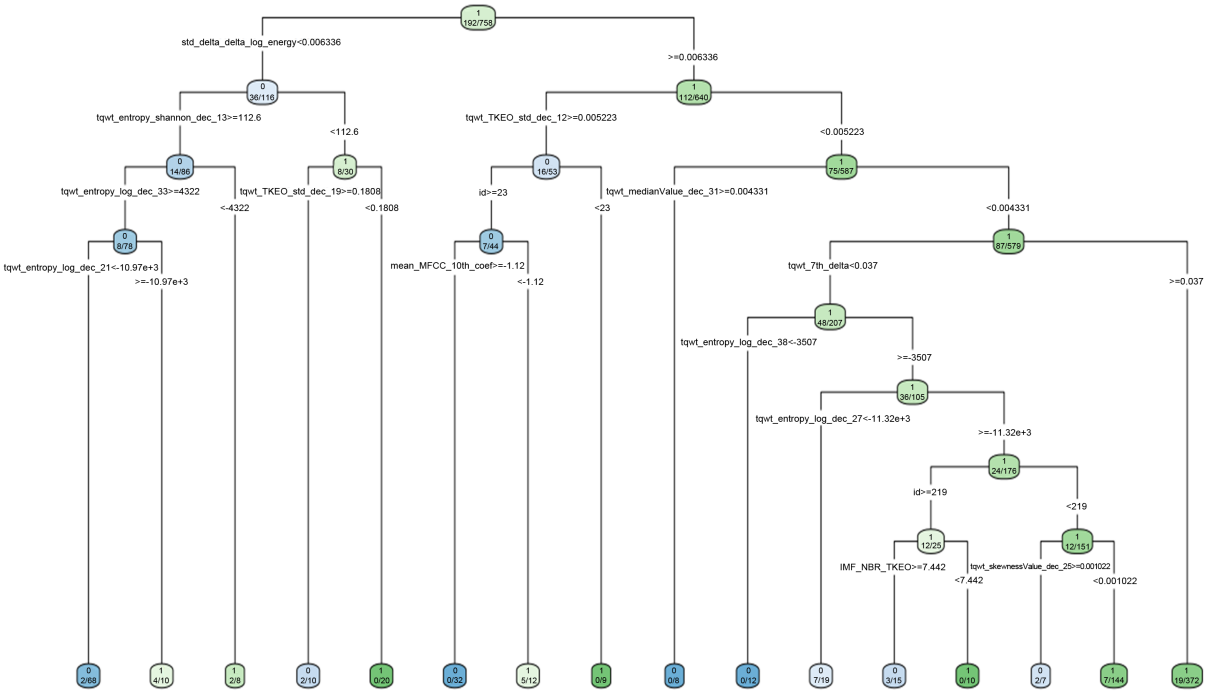

Figure 1. Decision trees for data classification
图 1. 数据分类的决策树

运用 R 建模过程中，通过不断拆分变量，得到较好的分类结果，训练时误差为 0，测试时误差为 0.10，相比 Logistic 模型精度高出许多。预测时，我们把因变量 class 标为缺失值，利用训练好的决策树来进行预测，得到预测值(下面相同)。

4.3. Bagging 分类模型的建立与分析

Bagging 分类是基于决策树的基础上进行的，按照简单少数服从多数的原则来投票确定待测观测值属

于哪一类。基于决策树分类的基础之上，在运用 R 建模的过程中，能取得较好的分类结果，训练的误差为 0，但测试时误差为 0.21，相比决策树分类精度稍逊一筹。

4.4. 随机森林分类模型的建立与分析

随机森林与 Bagging 一样，都是基于决策树的一种组合方法。随机森林的每棵树都不枝剪，让其充分生长。最终的预测结果时根据所有决策树按照各自的分类结果做建档投票，取得最多票的类。以下给出拟合过程中的节点直方图图 2。终节点个数又叫树的大小，可以通过函数 treesize()得到。图 2 中左图表示所有节点数，右图表示终节点个数。随机森林能取得较好的分类结果，训练时的误差为 0，测试时误差为 0.12，相比 Bagging 分类精度更高一筹。

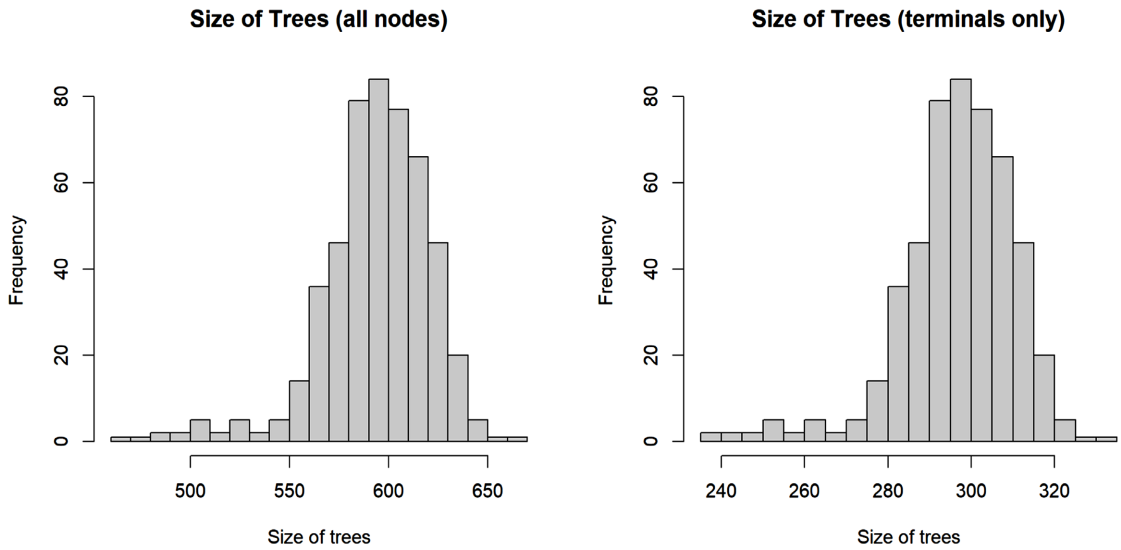

Figure 2. Histogram of all nodes (left) and number of terminal nodes (right) of a random forest
图 2. 随机森林所有节点数(左)和终节点数(右)的直方图

4.5. 对比与分析

为了对 Logistic 回归、决策树、Bagging 和随机森林预测模型进行分析比较，本文对语音信号数据进行分类的交叉验证，使用函数 Fold()来平衡各测试集变量之间的水平。

Figure 3. Four ways to cross-validate a bar chart with ten folds
图 3. 四种方法十折交叉验证条形图

从上图 3 分析得出，Logistic 回归、决策树、Bagging 和随机森林训练集训练过程中训练时的误差分别为：0.16，0，0，0。预测时的误差分别为：0.28，0.10，0.21，0.12。由此可以判断出，基于帕金森病语音信号数据，Logistic 回归、决策树、Bagging 和随机森林在预测时预测精度最高的是决策树，其次是随机森林，和决策树的预测精度相差了 0.02，两个模型在应用与探测早期帕金森病时都有不错的预测效果。而 Logistic 回归的二分类模型在此处的预测精度却显得格格不入，误差将近是决策树的 3 倍。

5. 结论

通过对比分析得出，Logistic 回归中的二分类在建模过程中需要满足很多条件，且预测的效果远远不如机器学习里面的决策树预测、Bagging 预测和随机森林预测，特别是在训练集较小时预测的效果极其不佳，经不断尝试初步得出若训练集越大，则训练出来的模型就会越成熟，预测值就会越接近真实值。而且当 Logistic 模型自变量有较多的定性变量或者定性变量的水平比较多时，Logistic 回归时完全无法再进行接下来的步骤。

相较而言，本文所选取的决策树、Bagging 和随机森林在做分类时，对自变量没有过多要求，以决策树为分类基础，Gini 不纯度为分类原则进行分类。取得了相对较好的预测效果其中决策树和随机森林的预测效果明显优于其它几种方法。经分析初步探知得出，机器学习方法在大部分数据中具有较佳的效果，能拟合出恰当的数据特征，但并不代表所有的机器学习方法就一定优于传统方法，因为也有一部分原因是数据本身的特殊性和复杂性所导致。在机器学习领域中，虽然决策树是组合模型 Bagging 和随机森林的基本分类基础，但是此处的预测效果却比 Bagging 和随机森林都还要好，足以证明，没有什么方法是绝对优于其它方法的，预测的效果不仅和模型相关和数据本身也有很大的关系。

参考文献

[1] 谭丽. 中晚期帕金森病脑深部电刺激治疗前后中医证候特征研究[D]: [硕士学位论文]. 北京: 北京中医药大学, 2021. https://doi.org/10.26973/d.cnki.gbjzu.2021.000437

[2] Heida, T. and Modolo, J. (2017) Models of Deep Brain Stimulation. http://www.scholarpedia.org/article/Models_of_deep_brain_stimulation

[3] Redgrave, P. (2007) Basal Ganglia. http://www.scholarpedia.org/article/Basal_ganglia

[4] 吴喜之, 张敏. 应用回归及分类——基于 R 与 Python 的实现[M]. 北京: 中国人民大学出版社, 2020.

[5] 吴喜之, 张敏. Python——数据科学的手段[M]. 北京: 中国人民大学出版社, 2020.

[6] Sakar, C.O., Serbes, G., Gunduz, A., *et al*. (2019) A Comparative Analysis of Speech Signal Processing Algorithms for Parkinson's Disease Classification and the Use of the Tunable Q-Factor Wavelet Transform. Applied Soft Computing, 74, 255-263. https://doi.org/10.1016/j.asoc.2018.10.022

[7] Rothlind, J.C., York, M.K., Luo, P., *et al*. (2021) Predictors of Multi-Domain Cognitive Decline Following DBS for Treatment of Parkinson's Disease. *Parkinsonism & Related Disorders*, **95**, 23-27. https://doi.org/10.1016/j.parkreldis.2021.12.011