基于迭代决策树的帕金森 UPDRS 预测模型研究

林 钢,季 薇
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:迭代决策树(GBDT)属于机器学习算法的一种,该算法具有较好的真实分布拟合能力,可用于解决大部分回归问题。根据帕金森病对不同年龄的男性和女性患者语音的影响不同这一现实依据,提出将性别和年龄这一先验知识融入到 GBDT,实现对统一帕金森评定量表(UPDRS)的预测。将性别和年龄作为先验知识,对 UPDRS 预测模型进行模型分解;根据迭代决策树的原理,对分解后的各模型运用决策树进行模型重构,并在各自残差减少的梯度方向上迭代训练新的决策树;将得到的以叶子节点作为增益的决策树作为最终的 UPDRS 预测模型。在远程帕金森数据集的仿真实验中,得到的 total-updrs 和 motor-updrs 平均绝对误差值分别为 4.498 0 和 3.531 8,与最小二乘法相比,分别提高了 52.19% 和 53.36%,与决策树相比,分别提高了 52.66% 和 52.89%。实验结果表明,根据先验知识,使用性别和年龄的组合进行预测模型分解,并对分解各模型分别进行模型重构,能够有效提高 UPDRS 预测的准确率。

关键词:帕金森疾病;语音;统一帕金森评定量表;性别划分;年龄划分;迭代决策树
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2019)01-0216-05
doi:10.3969/j.issn.1673-629X.2019.01.045

Research on Parkinson UPDRS Prediction Model Based on GBDT

LIN Gang, JI Wei
(School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: As a kind of machine learning algorithm, gradient boosting decision tree (GBDT) can be used to solve most of the regression problem due to fine fitting ability of the true distribution. Based on the fact that the effect of Parkinson's disease on the speech of male and female patients of different ages is different, we use the prior knowledge of gender and age into GBDT to predict unified Parkinson's disease rating scale (UPDRS). Use sex and age as a prior knowledge to decompose the prediction model of UPDRS. Applying decision tree to reconstruct each new model and new decision tree is iteratively trained in the direction of the gradient of the respective residuals. Decision tree with leaf node as the gain is the final prediction model of UPDRS. In the simulation experiments of remote Parkinson data set, the mean absolute error (MAE) of total-updrs is 4.498 0 and the motor-updrs is 3.531 8, which are 52.19% and 53.36% higher than that of least squares method (LS), and 52.66% and 52.89% higher than that of the classification and regression tree (CART). The experiment show that GBDT based on sex and gender partition can improve the accuracy of UPDRS prediction.

Key words: Parkinson's disease; speech; unified Parkinson's disease rating scale; gender partition; age partition; gradient boosting decision tree

0 引 言

帕金森病(Parkinson's disease,PD)是常见的神经退行性疾病之一,主要是由于中脑黑质致密部多巴胺能神经元变异以及残存神经元细胞多巴胺生物合成能力下降导致纹状体区多巴胺缺乏引起的。据流行病学调查,国内帕金森病患者自 2001 年就已经达到 200 万人,约占全球的 65%[1]。据国际数据统计,这一数字正以每年 10 万人的速度持续增长,预计到 2030 年国内患帕金森病的人数将达到 500 万[2]。然而,由于帕金森病是神经系统疾病,对其诊断具有极大的难度,主要依赖于权威专家多年的临床经验[3]。同时,帕金森的诊断和治疗,需要患者按时前往医院进行详细的检查。这不仅给患者,尤其是老年患者带来了不便,同时也给医师带来了巨大的工作量。随着国内老年化进程

收稿日期:2018-01-21 **修回日期**:2018-05-24 **网络出版时间**:2018-09-21
基金项目:国家自然科学基金(61603197,61772284);南京邮电大学科研基金(NY215104)
作者简介:林 钢(1991-),男,硕士研究生,研究方向为机器学习;季 薇,博士,副教授,研究方向为无线通信、机器学习。

网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20180920.1535.014.html

的加深,现急需通过计算机技术开发一款费用低廉,可以应用于基层医疗服务的自助诊疗系统[4]。

研究表明,大约 90% 的帕金森症患者都会患有某种程度的语音障碍,因此,直接通过非接触式的方法采集语音进行研究,如阵列式麦克风等,相比较其他诊断方法更加方便有效[5-6]。这使得基于语音的帕金森诊疗方案的研究得到了极大关注。Titze 等分析了不同性别的人在 20 ~ 90 岁时基频的变换情况,认为可以使用患者的语音特征,结合机器学习算法来解决帕金森病的诊疗方案探索的问题[7]。随着计算机技术的快速发展,近年来,人们应用了许多基于机器学习算法研究帕金森病的诊疗方案,以求能彻底取代临床决策。这些诊疗方案主要归为两类:(1)判断用户是否患有帕金森病,即实现帕金森病的诊断[8-10],常见的是基于数据线性可分与非线性可分,应用线性 SVM 和非线性 SVM[11]进行研究;(2)预测帕金森病患者病情严重等级,即通过预测 UPDRS(unified Parkinson's disease rating scale)实现帕金森病进展跟踪[12-14]。

对于预测帕金森病患者病情的严重等级这一问题,主要是采用回归预测算法将提取的语音特征映射为 UPDRS。目前国际上普遍采用 UPDRS 衡量帕金森症病情的严重性,主要涉及三方面内容:(1)精神、行为和情绪(1 ~ 4),用于测量患者的生理或心理状况;(2)日常活动(5 ~ 17),用于测量帕金森症患者能否在无协助状态下完成日常生活;(3)运动症状(18 ~ 44),用于测量身体肌肉状况。总的来说,UPDRS 总共含有 44 项测试内容,每项评测结果分为 0 ~ 4 个等级,其中 0 表示健康,4 表示严重,采用这三项测量的总评分,即 total-updrs,作为病情的总体衡量标准。其中,第三项包含了帕金森症的大部分症状[13],采用此测量评分,即 motor-updrs,作为病人运动能力的衡量标准。

Tsanas 等使用了各种语音信号处理算法提取了相关病理特征,并利用 LS(least squares)和 CART(classification and regression tree)来预测帕金森症患者的 motor-UPDRS 和 total-UPDRS 值[12,15]。然而,在开发安卓应用预测 UPDRS 的过程中发现预测结果的正确率较低,原因是传统的机器学习方法需要假设数据是同分布的[16],该假设忽略了不同年龄的男性和女性的语音样本之间存在差异性的事实。文献[13]中证明了帕金森病对男性患者和女性患者语音的影响有着明显的不同。对此,文中首先分析了样本之间的差异性是如何影响预测模型建立的,然后对 UPDRS 进行回归预测的问题做了进一步的研究分析,提出了根据性别和年龄的先验知识进行 UPDRS 预测模型的分解,并对分解的模型分别进行模型重构。其次,考虑到远程帕金森数据集共包含 5 875 个语音样本,数据规模较大,因此模型重构采用集成学习算法 GBDT。最后,通过与传统的 LS 与 CART 算法进行仿真实验比较。

1　基于 GBDT 的帕金森病 UPDRS 的预测

1.1　数据集描述

文中使用了 UCI 中的远程帕金森数据集,该数据集包含 42 位患有帕金森病的语音样本,样本源于每周对患者采集 1 次持续元音/a/的语音,每次采集 6 条,持续时间 6 个月,共计 5 875 条样本。通过语音信号处理算法,从这些语音样本中提取的特征数总共为 16 个,其中衡量基频变化的特征有:跳动 Jitter(%)、跳动绝对值 Jitter(Abs)、相对幅度摄动 Jitter:RAP、5 点周期摄动熵 Jitter:PPQ5、周期绝对差与平均周期比 Jitter:DDP;衡量振幅变化的特征有:局部闪烁 Shimmer、局部闪烁(dB)Shimmer(dB)、3 点幅度摄动熵 Shimmer:APQ3、5 点幅度摄动熵 Shimmer:APQ5、11 点幅度摄动熵 Shimmer:APQ11、相邻周期幅度差的平均绝对差 Shimmer:DDA;噪声谐波比 NHR;谐波噪声比 HNR;循环周期密度熵 RPDE;趋势波动分析 DFA;基音周期熵 PPE,最终得到 5 875 ×16 的样本集。

1.2　数据集划分

计算机可以通过机器学习构建各种学习模型,来解决相同或相近的问题。然而,良好的学习模型的建立,除了依靠合适的算法,计算机的计算能力,还需要针对数据集做数据的预处理。文献[15]中利用 UCI 中的远程帕金森数据集,使用 LS 进行预测模型的训练,取得了一定的预测效果。文中以两个存在差异性的数据集为例,使用普通最小二乘法(ordinary least squares)训练出一个回归预测模型 model 3,如图 1 所示。考虑到样本数据分布情况,很显然 model 3 为了拟合两个不同域的对象样本做了折中,该模型的真实预测效果并不是很理想。

回归模型的学习,是通过计算对所有数据的预测值和真实值之间的平方误差,并对误差进行累加使得正差值和负差值相互抵消。这样会导致在分布不同,即域不同的样本数据集上做回归预测,却能得到相同的模型。数据集分布的概念比较复杂,总的来说,每个种类的样本数量,应该有利于模型的建立和评价。根据安卡姆剃刀原理[17]:在模型的选择时,能够很好地解释已知数据并且十分简单的模型才是最合适的模型。分析远程帕金森数据集,在这两个对象上建立的 model 3 显然是欠拟合的,但如果将两个对象按照一定的先验知识划分成两个域的数据,然后在两个域上对原模型进行分解,那么在两个域上分别建立起的 model 1 和 model 2 则能很好地拟合各自域中的数据,如图 1 所示。因此,按照对象的性别和年龄来划分数据集,

并在不同的数据集上分别建立起预测模型以达到有效的模型分解。在实际应用中,采集到用户的语音,并根据用户的年龄和性别,通过推荐系统给出合适的模型,实现对用户的 UPDRS 的预测,其原理如图 2 所示。

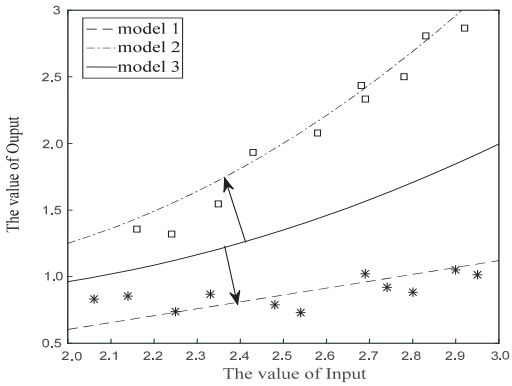


图 1 回归训练模型

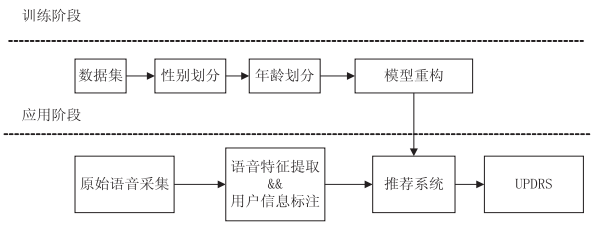


图 2 实现流程

1.3 迭代决策树

集成学习能够显著提高学习系统的泛化能力,受到了机器学习界的广泛关注[18]。GBDT 是由 Jerome Friedman 于 1997 年提出的,并 1999 年重新改进后应用于回归预测的问题中[19-20]。随后,GBDT 回归算法因性能突出吸引了大批学者的关注,陈天奇博士针对 GBDT 的众多改进,对 GBDT 算法进行了总结和优化,得到了 XGboost 算法[21]。Apache Spark 已将 GBDT 回归算法封装到 Mllib 库中,使得在 Spark 平台上使用 GBDT 做回归预测更加方便。

GBDT 属于集成学习算法,由多个弱学习器组合产生一个强学习器。该算法由多棵决策树持续迭代而成,通过迭代使得所有树的预测结论和残差等于或趋近于 0 为止,最终得到一个高准确度的预测模型。

假设初始得到的学习器为:

$$F_0(x) = \arg\min_{\alpha} \sum_{i=1}^{n} L(y_i, \alpha) \tag{1}$$

其中, $L(y_i,\alpha)$ 是模型 $f(x)$ 的损失函数。$((x_i,y_i))_{i=1}^{n}$ 是样本训练集, x_i 是特征,与之相对应的是结果 y_i。

首先根据当前数据,使得损失函数最小,得到初始化损失函数模型。迭代次数设定为 M,每次迭代产生一个模型,为了让每次迭代生成的模型对训练集的损失函数最小,根据式 1,每次迭代时通过向损失函数的负梯度方向移动来使得损失函数越来越小,就可以得到越来越精确的模型。算法迭代的主要步骤如下:

第一步,计算残差 r_{i1}。

用初始模型 $F_0(x)$ 计算出负梯度,如式 2。损失函数的负梯度对于当前模型 $F_0(x)$ 的值作为残差。对于平方损失函数该值是残差,对一般损失函数,该值是残差的估计值。

$$r_{i1} = -\left[\frac{\partial L(y_i, F_0(x_i))}{\partial F_0(x_i)}\right], i = 1,2,\cdots,n \tag{2}$$

第二步,训练出一个基学习器。

使用 $((x_i,r_i))_{i=1}^{n}$ 拟合出一棵 CART 回归树,得到由叶子节点组成的决策树 $h_1(x)$。

第三步,寻找合适的步长。

在 GBDT 算法中用到的梯度下降,步长是通过计算得到的。计算规则是使得到的新学习器的 $F_1(x)$ 损失函数 β_1 值最小。

$$\beta_1 = \arg\min_{\alpha} \sum_{i=1}^{n} L(y_i, F_0(x_i) + \alpha h_1(x_i)) \tag{3}$$

第四步,根据梯度和步长,迭代得到回归树模型 $F_1(x)$,如式 4 所示。

$$F_1(x) = F_0(x) + \beta_1 h_1(x) \tag{4}$$

通过上面四个步骤,就可以从初始模型 $F_0(x)$ 优化得到第二个模型 $F_1(x)$,迭代 $M-1$ 次这四个步骤,就可以得到最终的 GBDT 模型。

2 实 验

2.1 数据处理

2.1.1 数据预处理

数据预处理一般包括特征选择、数据清洗等操作。

特征选择是指从原始的特征集中选出部分重要特征组成的子集。特征选择能够除去冗余的或不相关的特征,以达到提高模型泛化能力的目的[22]。文中使用的特征选择算法是 ReliefF[23]。最早提出的 Relief 算法用来解决二分类的问题,该算法设计了一个相关统计向量来评估每个特征的重要程度,向量的每个分量是对其中一个初始特征的评价值,特征子集的重要性为子集中所有特征的相关统计量之和,这个相关统计量被视为是每个特征的权值,即 ReliefF 属于一种特征权重算法。公式如下:

$$W(f_j) = \frac{1}{q}\sum_{i=1}^{q}\left\{-\frac{1}{|\mathrm{NH}(X_i)|}\cdot\sum_{X_n\in\mathrm{NH}(X_i)}\|x_{i,j}-x_{n,j}\| + \sum_{y_i\neq y_l}\frac{1}{|\mathrm{NM}(X_i)|}\cdot\frac{p(y=y_l)}{1-p(y=y_i)}\cdot\sum_{X_n\in\mathrm{NM}(X_i)}\|x_{i,j}-x_{n,j}\|\right\} \tag{5}$$

其中, $W(f_j)$ 表示第 j 个特征拥有的权重; q 表示随机选择的样本数量; X_i 表示一个数据样本,

$|NH(X_i)|$ 表示离样本 X_i 最近的同类样本数量,$|NM(X_i)|$ 表示离该样本 X_i 最近的不同类样本数量;$\|\cdot\|$ 表示距离测度,常用欧氏距离或曼哈顿距离。

文中用 ReliefF 算法得到特征权重,并根据权重对特征进行重要性排序,选取排在前面的 13 个重要特征。

在完成了特征选择之后,并不能将数据直接用于计算,很多情况下需要对数据进行一些基本的处理。这些基本的处理包括缺失值的处理、非数字形式的特征值处理、异常值的处理等等,远程帕金森数据集已经进行过数据的清洗。

2.1.2　基于性别信息的数据划分

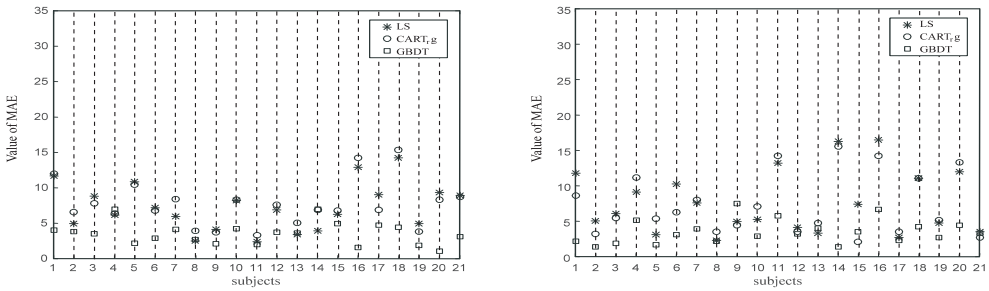
帕金森疾病对不同年龄的男性患者和女性患者语音的影响有着明显不同,将性别和年龄信息作为先验知识,对 UPDRS 回归预测问题进行进一步分解。首先,按照性别将数据集分成两个子集,其中男性 28 个对象共计 4 008 条样本,女性 14 个对象共计 1 867 条样本。然后,根据年龄再次划分数据集,最终分为 6 组:女性<60 岁 756 条样本,女性 60 ~ 70 岁 573 条样本,女性>70 岁 538 条样本;男性<60 岁 1 121 条样本,男性 60 ~ 70 岁 1 554 条样本,男性>70 岁 1 333 条样本。

2.2　模型训练

通过对数据集的划分,得到了 6 组数据集,使用 GBDT 算法在这 6 个数据集上分别进行模型训练,最终会得到 6 个来自不同域的模型。在实际系统的应用阶段,当用户自助采集语音时,会将用户的性别和年龄信息作为标签进行存储。然后对用户采集的语音提取 13 维特征作为输入,根据标签信息为用户推荐一个域匹配的模型,最终输出预测的 UPDRS 值。设 A、B、C、D、E、F 分别代表男性 60 岁以下、男性 60 岁至 70 岁、男性 70 岁以上、女性 60 岁以下、女性 60 岁至 70 岁、女性 70 岁以上。使用 100 棵决策树用于模型的训练,并设置每棵树默认的叶子节点数是 64 个,为了避免出现过拟合的现象,规定每个节点样本数不得少于 30 个。以 A 组为例,选取该组中任意一个对象的前 1/3 的数据作为测试集,剩余数据和该组的其他对象的数据作为训练集训练模型[24],最后通过训练所得的模型得出预测值,使用预测值和真实值的 MAE 作为衡量模型训练效果的评价准则。

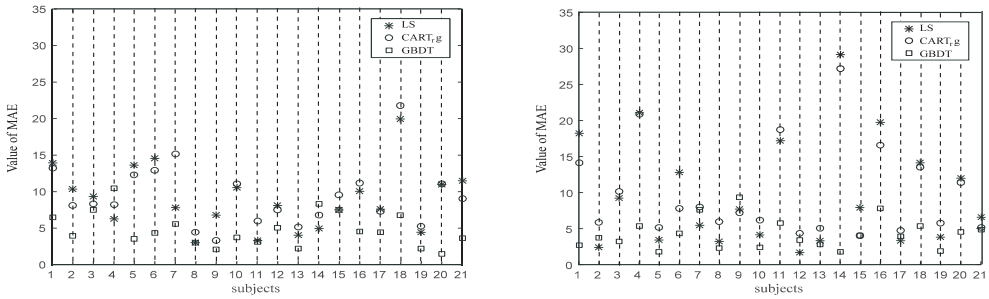
2.3　实验结果分析

将 GBDT 应用到经过预处理得到的远程帕金森数据集以及更进一步划分得到的 6 组远程帕金森数据集上。考虑到 Tsanas 等未进行分组实验,因此在对比实验中,对分组实验的仿真结果进行了汇总,并将对 42 个对象的仿真分为两组显示,如图 3 和图 4 所示。



(a)motor 误差（前 21 个对象）　　(b)motor 误差（后 21 个对象）

图 3　远程帕金森数据集上对 motor-UPDRS 预测的 MAE



(a)total 误差（前 21 个对象）　　(b)total 误差（后 21 个对象）

图 4　远程帕金森数据集上对 total-UPDRS 预测的 MAE

观察建立的模型在不同对象上进行 motor-updrs 和 total-updrs 的预测结果,可以看出,对绝大部分对象预测的表现上,GBDT 总是比 LS 和 CART 的误差更小。使用 GBDT 在 6 个小组上预测 motor-updrs 的 MAE 是:3.589 2、2.954 4、3.567 0、4.345 4、4.000 3、3.236 9;预测 total-updrs 的 MAE 是:4.308 0、4.471 5、4.612 3、5.461 9、4.080 5、3.665 5。

使用 LS 预测 motor-updrs 的 MAE 是 7.461,预测 total-updrs 的 MAE 是 9.409;使用 CART 预测 motor-updrs 的 MAE 是 7.497,预测 total-updrs 的 MAE 是 9.645;使用 GBDT 预测 motor-updrs 的 MAE 是 3.532,预测 total-updrs 的 MAE 是 4.500。分析可知,

将性别和年龄作为先验知识融入到 GBDT,实现对统一帕金森评定量表 UPDRS 的预测相比传统的 LS 和 CART,最低提高 motor-updrs 的预测准确率 52.89%,最低提高 total-updrs 的预测准确率 52.19%。实验结果验证了基于性别和年龄的组合进行预测模型的分解,并在各自域上进行模型重构,能够有效提高 UPDRS 预测的准确率。

3 结束语

将集成学习方法迭代决策树 GBDT 应用到远程帕金森数据集中,实现了对 UPDRS 的预测。考虑到性别和年龄对帕金森病的影响,把数据集按照性别和年龄进一步做了划分,并在各集合上采用 GBDT 算法进行预测模型的训练,最后采用模型推荐进行最佳预测模型匹配。实验结果表明,最终得到的预测效果较普通决策树效果提高了一半有余,同时也进一步验证了帕金森病受年龄和性别的影响较大。性别和年龄属于先验知识,使用这些先验知识来分解预测模型,不但简单高效,而且具有一定的现实指导意义。当然,除性别外,生活中还存在许多其他的先验知识,如患者的健康状况等,对于进一步提高帕金森病 UPDRS 预测的准确率,发现更多有用的先验知识是重要的研究工作。

参考文献:

[1] ZHANG Z,ROMAN G C,HONG Z,et al. Parkinson's disease in China:prevalence in Beijing,Xi'an,and Shanghai[J]. Lancet,2005,365(9459):595-597.

[2] DORSEY E R,GEORGE B P,LEFF B,et al. The coming crisis:obtaining care for the growing burden of neurodegenerative conditions[J]. Neurology,2013,80(21):1989-1996.

[3] CHO C,CHAO W,LIN S,et al. A vision-based analysis system for gait recognition in patients with Parkinson's disease[J]. Expert Systems with Applications,2009,36(3):7033-7039.

[4] 韩 艳,张晓红,陈 彤,等. 帕金森病诊治现状调查[J]. 中华保健医学杂志,2008,10(1):18-20.

[5] HO A K,IANSEK R,MARIGLIANI C,et al. Speech impairment in a large sample of patients with Parkinson's disease[J]. Behavioural Neurology,1998,11(3):131-137.

[6] LOGEMANN J A,FISHER H B,BOSHES B,et al. Frequency and co-occurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients[J]. Journal of Speech & Hearing Disorders,1978,43(1):47-57.

[7] TITZE I R. Principles of voice production[M]. 2nd ed. Iowa City:National Center for Voice and Speech,2000.

[8] SAKAR C O,KURSUN O. Telediagnosis of Parkinson's disease using measurements of dysphonia[J]. Journal of Medical Systems,2010,34(4):591-599.

[9] LITTLE M A,MCSHARRY P E,HUNTER E J,et al. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease[J]. IEEE Transactions on Biomedical Engineering,2009,56(4):1015.

[10] TSANAS A,LITTLE M A,MCSHARRY P E,et al. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease[J]. IEEE Transactions on Biomedical Engineering,2012,59(5):1264-1271.

[11] 李 航. 统计学习方法[M]. 北京:清华大学出版社,2012.

[12] ATHANASIOS T,LITTLE M A,MCSHARRY P E,et al. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests[J]. IEEE Transactions on Biomedical Engineering,2010,57(4):884-893.

[13] TSANAS A,LITTLE M A,MCSHARRY P E,et al. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity[J]. Journal of the Royal Society Interface,2011,8(59):842-855.

[14] SAKAR B E,KURSUN O. Telemonitoring of changes of unified Parkinson's disease rating scale using severity of voice symptoms[C]//Proceedings of the 2nd international conference on e-health and telemedicine. Istanbul:[s. n.],2014:114-119.

[15] TSANAS A. Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning[D]. UK:University of Oxford,2012.

[16] 戴文渊. 基于实例和特征的迁移学习算法研究[D]. 上海:上海交通大学,2008.

[17] 王 珏,周志华,周傲英. 机器学习及其应用[M]. 北京:清华大学出版社,2006:7-8.

[18] 张春霞,张讲社. 选择性集成学习算法综述[J]. 计算机学报,2011,34(8):1399-1410.

[19] FRIEDMAN J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis,2002,38(4):367-378.

[20] FRIEDMAN J H. Greedy function approximation:a gradient boosting machine[J]. Annals of Statistics,2001,29(5):1189-1232.

[21] CHEN Tangqi,GUESTRIN C. XGBoost:a scalable tree boosting system[C]//ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco,California,USA:ACM,2016:785-794.

[22] 李 敏,卡米力·木依丁. 特征选择方法与算法的研究[J]. 计算机技术与发展,2013,23(12):16-21.

[23] MARKO R S,IGOR K. Theoretical and empirical analysis of ReliefF and ReliefF[J]. Machine Learning,2003,53(1-2):23-69.

[24] WU D,CHUANG C H,LIN C T. Online driver's drowsiness estimation using domain adaptation with model fusion[C]//International conference on affective computing and intelligent interaction. Xi'an,China:IEEE,2015:904-910.