浙江理工大学学报,第 49 卷,第 1 期,2023 年 1 月
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2023.01.011

基于按键动作识别的帕金森病早期预测方法

许　昊[1],童基均[1],齐鹏嘉[1],周思薇[2]
(1.浙江理工大学信息学院,杭州 310018;2.浙江康复医疗中心,杭州 310014)

摘　要:针对目前帕金森病早期预测方法普遍存在误诊率高、步骤繁多等问题,设计了基于 AdaBoost 算法的按键动作识别方法,实现对帕金森病早期的精准预测。该方法首先删除数据集的缺失值,并选取按键次数过万的数据;然后针对不同按键手,根据按键的时间间隔对预处理后的结果进行分类,以平均值、标准差、方差、偏度和峰度 5 个指标为特征,对每一位病人的数据进行分块,扩充数据集,并加入高斯噪声平衡数据集;最后应用 AdaBoost 算法进行分类预测。在公开的数据集上进行实验,结果表明:在按键数据集分类上,该方法的准确率、灵敏度和特异性分别为 95%、98%和 97%。该方法具有较高的准确率、灵敏度和特异性,为帕金森病早期的精准预测提供了一种有效的解决方案。

关键词:AdaBoost 算法;帕金森病;预测;分块;非平衡数据;按键
中图分类号:TP391.4　　**文献标志码**:A　　**文章编号**:1673-3851 (2023) 01-0083-06
引文格式:许昊,童基均,齐鹏嘉,等. 基于按键动作识别的帕金森病早期预测方法[J]. 浙江理工大学学报(自然科学),2023,49(1):83-88.
Reference Format: XU Hao, TONG Jijun, QI Pengjia, et al. Early prediction method of Parkinson′s disease based on keystroke recognition[J]. Journal of Zhejiang Sci-Tech University,2023,49(1):83-88.

Early prediction method of Parkinson′s disease based on keystroke recognition

XU Hao[1], *TONG Jijun*[1], *QI Pengjia*[1], *ZHOU Siwei*[2]
(1. School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2. Zhejiang Rehabilitation Medical Center, Hangzhou 310014, China)

Abstract: Aiming at the problems of high misdiagnosis rate and numerous steps in the early prediction method of Parkinson′s disease, a keystroke recognition method based on AdaBoost algorithm was designed to achieve accurate prediction of early Parkinson′s disease. By adopting this method, the missing values of the dataset were deleted, and the data with more than 10 000 keystrokes were selected. Then, for different keystroke hands, the results after pre-treatment were classified according to the time interval of the keys, the five indicators of calculated mean, standard deviation, variance, skewness and kurtosis were used as new features, and the data of each patient was blocked, so as to expand the dataset and Gaussian noise was added to balance the dataset. Finally, the predictive classification was carried out by AdaBoost algorithm to achieve early prediction of Parkinson′s disease. Experiments were conducted on the published dataset, and the experimental results showed that the accuracy, sensitivity and specificity of the algorithm were 95%, 98% and 97%, respectively, in the classification of the key dataset. Compared with other algorithms, this method has high accuracy, sensitivity and specificity, providing an effective solution for early prediction of Parkinson′s disease.

Key words: AdaBoost algorithm; Parkinson′s disease; prediction; block; unbalanced data; keystroke

收稿日期:2022-04-08　　网络出版日期:2022-09-07
基金项目:浙江省医院协会管理软科学项目(2021ZHA-KEB206);浙江省自然科学基金项目(LQ22F010006)
作者简介:许　昊(1997-　),男,浙江金华人,硕士研究生,主要从事信号处理方面的研究。
通信作者:童基均,E-mail:jijuntong@zstu.edu.cn

0 引 言

帕金森病(Parkinson's disease, PD)是一种老年神经系统退行性疾病,具有不可逆、难发现和严重影响生活质量等特点,该病自发现以来就广受人们关注,已成为众多学者研究的焦点。在我国年龄超过55岁的人群中,帕金森病患者约有170万人[1]。然而,目前帕金森病的早期诊断仍面临着巨大的挑战,其主要原因在于:一方面,帕金森病的发病机制复杂,涉及α-突触核蛋白积聚、神经炎症、氧化应激、线粒体功能障碍、神经黑色素过度累积等多个因素[2];另一方面,由于早期帕金森病隐蔽性较高,患者较难发现患病的真实情况。因此,开发一种精确、客观的早期预测手段,对诊断和治疗帕金森病具有重要的意义。

目前,帕金森病的早期预测主要有两种方式,即专业医学仪器预测和结合信息技术预测。通过专业医学仪器进行预测是现阶段临床预测的主要方式。例如,杨丽娟等[3]使用实时经颅彩色多普勒超声(Transcranial sonography, TCS)观察黑质形态、回声,通过判断黑质功能状态检测帕金森病,其优点在于操作方便、价格低廉,但也存在假阳性率高的问题。蔡增林等[4]给患者注射示踪剂^{18}F-氟脱氧葡萄糖(^{18}F-FDG),并通过正电子发射计算机断层成像仪进行正电子发射计算机脑断层显像,然后采用目测法对脑部各部位的葡萄糖代谢进行判断,但由于辐射以及部分患者对显像剂过敏等原因,该方式难以成为帕金森病早期检测的主流方式。此外,还有很多学者通过结合信息技术对早期帕金森病患者进行预测。例如,Prashanth等[5]通过问卷方式得到由国际帕金森和运动障碍病协会发起修订的统一帕金森病评定量表(Movement Disorder Society-Unified Parkinson's Disease Rating Scale, MDS-UPDRS),并建立了逻辑回归、随机森林、增强树和支持向量机模型进行帕金森病早期预测,但由于部分早期患者的症状并不明显,导致问卷结果无法反映真实情况,因此,该方法不能作为早期预测的主要手段。Cho等[6]通过识别帕金森病患者特定步态模式,采用主成分分析与线性判别分析对帕金森病进行预测,但因数据集过小,结果的可信度较低。综上可知,现有研究虽然取得了一定的成果,但是仍然存在误诊率高、步骤繁多等问题。

医学研究表明,帕金森病患者的主要症状有静止震颤、运动迟缓、姿势不稳、肌肉强直[7]以及其他非运动性症状。其中,70%~75%的患者存在静止性震颤,这是最常见且最易识别的症状;震颤以4~6 Hz的频率发生,并且在手等肢体的远端最为明显[7];运动迟缓通常表现为执行日常生活动作时动作迟缓,其中需要精细运动控制的任务尤为明显[8],对运动迟缓的评估通常包括让患者进行快速、重复、交替的手部运动[7](如手指敲击、手握、手内旋等)。此外,帕金森病的单侧性非常明显,可以作为临床参数对该疾病与其他神经退行性帕金森病综合征进行区分[9]。以上症状都可以通过人机交互的方式进行预测。例如,Noyce等[10]的研究表明,帕金森病患者和对照组在30 s内的按键次数、每个按键的平均停留时间、按键之间移动时间等指标上存在着显著差异。Giancardo等[11]将受试者按下和释放按键的时间间隔转换为nQi指数(neuroQWERTY index),再利用Bagging算法进行预测,但由于作者只选择了单一特征(按键保持时间)进行预测,所以预测结果并不理想。Madanchi等[12]使用多重分形去趋势分析(Multifractal detrended fluctuation analyses, MFDFA)分析了通过nQi采集到的受试者按键数据,并充分提取了早期帕金森病患者病情随时间逐渐加重的趋势特征。Iakovakis等[13]采用卷积神经网络通过触摸屏打字对早期帕金森病进行了预测与分析,但因采集到的按键次数过少,所以准确率较低。Adams[14]集合了8种不同机器学习模型,选用了更多的特征预测帕金森病,并对所有的预测结果进行了平均化处理,得到了95%的准确率。

以上研究结果表明,通过受试者的按键操作对帕金森病进行预测是可行的,但是若只选择按键中的某一特征进行预测,最终得到的准确率较低。因此,本研究提出了一种基于按键动作识别的帕金森病早期预测方法,选择按键手、按键保持时间和相隔按键之间的停顿时间为特征,对数据进行分块以扩充数据集,考虑到分块后数据集可能存在不平衡等问题,还引入了高斯噪声以进一步提升预测准确率,最后通过AdaBoost算法进行预测分类,以实现对帕金森病患者的早期预测。该方法由于选取了更多的按键特征,并对数据集进行了平衡化的处理,因此最终的预测结果将会有所提升。

1 数据采集与处理

1.1 数据集

本研究采用的实验数据是Goldberger等[15]在2017年发布的公开数据集。该数据集由来自美国、

加拿大、英国和澳大利亚的 200 多名参与者提供,收集到的每个数据文件都包括参与者使用各种 Windows 应用程序(如电子邮件、Web 搜索等)时按键操作的计时信息,按键采集软件“Tappy”提供按键按下和释放的时间戳的计时精度,可在几毫秒之内完成。数据文件包含两个部分:一部分是 Archived users,包括出生年份、性别、是否患有帕金森病、是否有震颤、是否有服用抑制帕金森病的药物等;另一部分是 Archived data,包含日期、按键手(Hand)、按键保持时间(按下和释放当前键之间的时间间隔,Hold Time)、按键方向(按下上一个键和按当前键的手,Direction)、延迟时间(按下上一个键和按当前键之间的时间间隔,Latency Time)、飞行时间(释放上一个键和按下当前键的时间间隔,Flight Time)等。

1.2　数据清洗

本研究只选取了数据集中具有完整登记信息且进行按键输入超过一万次的 215 名受试者,其中 164 人患病,51 人未患病,受试者具体年龄分布与患病情况如图 1 所示。另外,由于 Archived users 中所显示数据集的文件名为 User+_+UserKey 格式,Archived data 中所显示数据集的文件名为 UserKey+_+Date 格式,所以本研究将两个数据集通过 UserKey 进行合并,另外添加了标签 Label:Parkinsons 栏中 True 的 Label 值为 1,反之则为 0。合并后的数据集已去除 Parkinsons 栏中标记为 None 的数据,如表 1 所示。

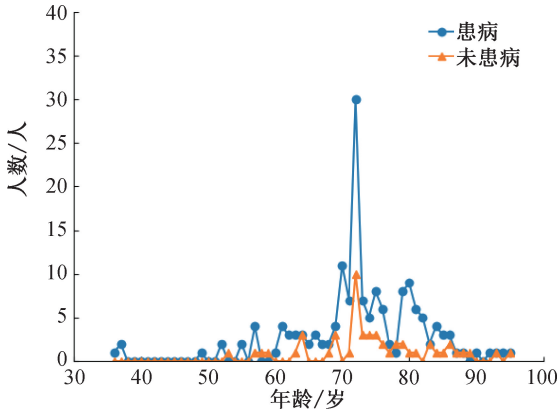

图 1　受试者具体年龄分布与患病情况

表 1　经过数据清洗、选择、合并文件后的数据

用户	指标						
	Hand	Hold Time/ms	Direction	Latency Time/ms	Flight Time/ms	Parkinsons	Label
0EA27ICBLF	L	0101.6	LL	0234.4	0156.3	True	1
…	…	…	…	…	…	…	…

2　研究方法

2.1　特征提取

在对受试者数据进行分块时,选择 Hand 列中标记为 L 的一类数据作为特征矩阵的行,并对 Hand 列中标记为 R 和 S 的数据用各自的平均值补齐。Hand 列中 L、R、S 分别表示在左手按键区域用左手按键、在右手按键区域用右手按键和用左手/右手按下空格键。Direction 列中 LL 表示第一次按键是在左手按键区域用左手按键,第二次按键也是在左手按键区域用左手按键;LR 表示第一次按键是在左手按键区域用左手按键,第二次按键在右手按键区域用右手按键;LS 表示第一次按键是在左手按键区域用左手按键,第二次按键是用左手/右手按下空格键;RL、RR、RS、SL、SR、SS 的含义与之类似。左手按键区域、右手按键区域和空格键如图 2 所示[14]。

本研究对数据的特征提取主要包括以下步骤:首先,对于所得到的每个数据列,计算平均值、标准差、方差、偏度和峰度 5 个指标,以此作为区分患病和未患病的特征。其中:平均值代表受试者长期以来按键的一般情况;方差和标准差在平均值的基础上进一步反映受试者长期以来按键的稳定程度,能够避免异常值对结果的影响;偏度能反映受试者按键异常值偏离的方向;峰度能反映受试者长期的按键离均值的分布情况,峰度越大,则按键数据在靠近均值部分的分布越多。经过组合后共得到 12×3×5=180 个特征,滤掉其中方差小于 5 的特征,最终得到 131 个特征。然后,对所得数据每一维度的特征进行标准差标准化,以避免数据不收敛,且规避在最终分类时大数据对小数据的影响。最后,将每一个受试者的数据进行分块,由于每个受试者均有较多数据,这样对于当前受试者最后一块数据可能出现未满 F_{flap} 情况的影响也较低。

2.2　机器学习模型

AdaBoost 算法[16]能够将预测精度低的弱学习器增强为预测精度高的强学习器。本研究采用 AdaBoost 算法对帕金森病进行预测,其中采用的弱

图 2 按键区域示意图

分类器为最大迭代深度为 15 的决策树算法,在每次迭代后会改变权值,最终将弱分类器进行整合成为一个强分类器。具体实现步骤如下:

a)初始化训练样本权重 $\boldsymbol{W}_j$:

$$\boldsymbol{W}_1=(w_{11},w_{12},\cdots,w_{1n}),w_{1i}=\frac{1}{n},i=1,2,\cdots,n \tag{1}$$

b)对 $j=1,2,\cdots,n$,使用权重 $\boldsymbol{W}_j$ 训练得到一个基学习器 $Q_j(x)$;

c)计算训练数据上的误差 $E_j(x)$:

$$E_j(x)=\sum_{i=1}^{n}P(Q_j(x_i)\neq y_i) \tag{2}$$

d)计算第 j 轮训练的系数 ε_j:

$$\varepsilon_j=\frac{1}{2}\ln\left(\frac{1-E_j(x)}{E_j(x)}\right) \tag{3}$$

e)得到第 $j+1$ 轮的训练结果:

$$Q_{j+1}(x)=\frac{Q_j(x)}{Z_j}\times\begin{cases}\exp(-\varepsilon_j), & Q_j(x_i)=y_i;\\ \exp(\varepsilon_j), & Q_j(x_i)\neq y_i\end{cases} \tag{4}$$

其中:Z_j 是规范化因子。

f)得到最终分类器 $Q(x)$:

$$Q(x)=\mathrm{sign}\left(\sum_{j=1}^{J}\varepsilon_jQ_j(x)\right) \tag{5}$$

2.3 评价指标

最终的预测结果通过 4 个指标进行评价,即准确率(Accuracy)、灵敏度(Sensitivity)、特异性(Specificity)和 AUC。其中,准确率通过十折交叉验证法得到,十折交叉验证法是测试算法准确率的常用方法,主要是将数据集分成 10 份,轮流将其中的 9 份作为训练数据,1 份作为测试数据进行试验。灵敏度 S_E 的计算公式为:

$$S_E=\frac{S_{ET}}{S_{ET}+S_{EF}} \tag{6}$$

其中:S_{ET} 表示被正确诊断为患病的病人,S_{EF} 表示被错误诊断为健康的病人。特异性 S_P 的计算公式为:

$$S_P=\frac{S_{PT}}{S_{PT}+S_{PF}} \tag{7}$$

其中:S_{PT} 表示被正确诊断为健康的健康人,S_{PF} 表示被错误诊断为患病的健康人。ROC 曲线被广泛用作度量一个二值分类器的优劣,横坐标为假阳性率,纵坐标为真阳性率。AUC 则表示 ROC 曲线下面的面积,其取值范围在 0.5 和 1 之间,越接近 1,则代表分类器性能越好。

3 结果分析与讨论

3.1 实验结果

预处理后的数据按照 0.8∶0.2 的比例分成训练集和数据集。为了验证不同 F_{flap} 值对结果是否存在影响,故让 F_{flap} 的值在(50,1000)之间以 10 为间隔进行变化,得到的结果如图 3 所示,其中:横坐标代表不同的分块数。

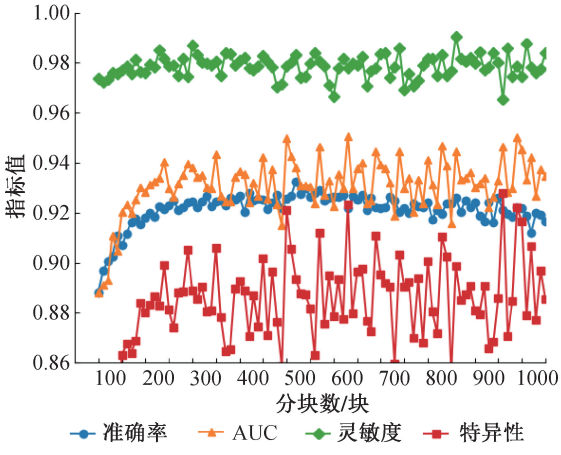

图 3 不同分块数下准确率、灵敏度、特异性和 AUC 的实验结果

此外,为了避免因数据集中出现非平衡数据(不得病的少于得病的)而导致的过拟合现象,原有数据处理基础上,对未患病病人数据(小样本)进行数据

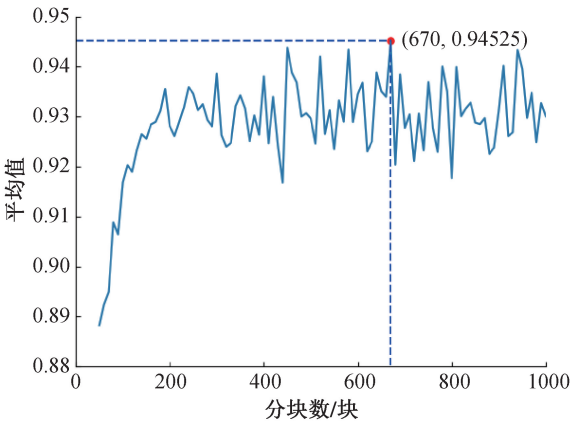

图 4　准确率、灵敏度、特异性和 AUC 四者的平均值随块数变化图

增强处理,根据图 4 选取准确率、灵敏度、特异性和 AUC 平均值最高的结果,即 $F_{flap}=670$ 进行数据平衡化处理,对其加入均值为 0、方差不同的高斯噪声,其得到的各指标值的结果如图 5 所示,其中横坐标代表方差。由图 5 可知,高斯噪声的方差对准确率、灵敏度、特异性和 AUC 的影响并不大;但由未加入高斯噪声前的混淆矩阵(表 2)和加入高斯噪声以后的混淆矩阵(表 3)相比可知,后者的准确率、灵敏度和特异性指标有明显的提升。

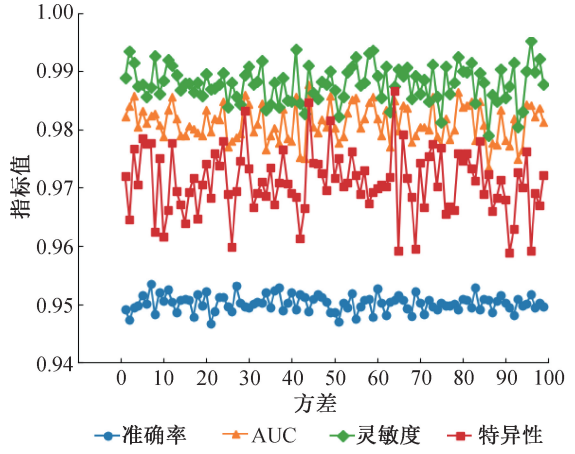

图 5　在 $F_{flap}=670$ 时加入均值为 0,方差不同的高斯噪声后准确率、灵敏度、特异性和 AUC 的实验结果

表 2　$F_{flap}=670$ 时的混淆矩阵

真实值	预测值	
	未患病	患病
未患病	386	51
患病	19	1032

表 3　$F_{flap}=670$,均值=0,方差=75 时的混淆矩阵

真实值	预测值	
	未患病	患病
未患病	1876	31
患病	26	1019

3.2　对比分析

为了验证本研究所提出方法的有效性,选取了准确率、灵敏度和特异性 3 种指标进行比较实验,结果如表 4 所示。该结果表明:本研究提出的基于按键动作识别的帕金森病早期预测方法,得到的结果相比专业医生的临床诊断有了明显提升;相比于新型的 TCS 技术,该方法在灵敏度和特异性方面均有所改善。与 Iakovakis 等[13]的研究相比,本研究所选取的受试者更多,因此在灵敏度和特异性等指标上都有明显提升;与 Adams[14]采用的多模型集合方法相比,本研究虽然只使用了一种模型,但通过加入高斯噪声的方式平衡数据集,使得最终预测结果的灵敏度得到提升。

表 4　不同诊断方法评价指标

诊断方法	灵敏度/%	特异性/%	准确率/%
按键[13]	79	79	—
按键[14]	96	97	95
非专家	74	79	—
专家	93	65	—
TCS[17]	81	68.6	—
按键(本研究)	98	97	95

4　结　论

本研究提出了一种基于按键动作识别的帕金森病早期预测方法。该方法以按键保持时间、按键手、相隔按键之间的停顿时间为特征,通过分块的方式扩充了数据集,并通过加高斯噪声的方式平衡了数据集,最后使用 AdaBoost 算法训练得到预测结果。与传统 MDS-UPDRS 评分表方式相比,本研究采用的方法避免了主观因素的影响,且准确率更高。该方法引入了计算机技术,具有较高的实际应用价值。

参考文献:

[1] Zhang Z X, Roman G C, Hong Z, et al. Parkinson's disease in China: Prevalence in Beijing, Xian, and Shanghai[J]. The Lancet, 2005, 365(9459): 595-597.

[2] 张森, 赵晓悦, 梁宇, 等. 帕金森病致病因素及发病机制研究进展[J]. 药学学报, 2020, 55(10): 2264-2272.

[3] 杨丽娟, 张京芬, 李月春, 等. 帕金森病的黑质超声表现[J]. 中华老年心脑血管病杂志, 2012,14(4):390-393.

[4] 蔡增林, 吴方萍, 周芯羽, 等. 早期帕金森病患者的临床与^{18}F-FDG PET 影像学特征研究[J]. 中国现代医药杂志, 2010, 12(2): 61-63.

[5] Prashanth R, Dutta Roy S. Early detection of Parkinson's disease through patient questionnaire and predictive

modelling [J]. International Journal of Medical Informatics, 2018, 119: 75-87.
[6] Cho C W, Chao W H, Lin S H, 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.
[7] Jankovic J. Parkinson's disease: Clinical features and diagnosis[J]. Journal of Neurology, Neurosurgery, and Psychiatry, 2008, 79(4): 368-376.
[8] Jahanshahi M, Jenkins I H, Brown R G, et al. Self-initiated versus externally triggered movements: I. An investigation using measurement of regional cerebral blood flow with PET and movement-related potentials in normal and Parkinson's disease subjects [J]. Brain, 1995, 118(4): 913-933.
[9] Cubo E, Martinez Martín P, Martin-Gonzalez J A, et al. Motor laterality asymmetry and nonmotor symptoms in Parkinson's disease[J]. Movement Disorders, 2010, 25(1): 70-75.
[10] Noyce A J, Nagy A, Acharya S, et al. Bradykinesia-akinesia incoordination test: Validating an online keyboard test of upper limb function[J]. PLoS One, 2014, 9(4): e96260.
[11] Giancardo L, Sánchez-Ferro A, Arroyo-Gallego T, et al. Computer keyboard interaction as an indicator of early Parkinson's disease[J]. Scientific Reports, 2016, 6: 34468.
[12] Madanchi A, Taghavi-Shahri F, Taghavi-Shahri S M, et al. Scaling behavior in measured keystroke time series from patients with Parkinson's disease[J]. The European Physical Journal B, 2020, 93(7): 126.
[13] Iakovakis D, Diniz J A, Trivedi D, et al. Early Parkinson's disease detection via touchscreen typing analysis using convolutional neural networks[C]//2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Berlin: IEEE, 2019: 3535-3538.
[14] Adams W R. High-accuracy detection of early Parkinson's disease using multiple characteristics of finger movement while typing[J]. PLoS One, 2017, 12(11): e0188226.
[15] Goldberger A L, Amaral L A, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals [J]. Circulation, 2000, 101(23): E215-E220.
[16] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.
[17] Wang L S, Yu T F, Chai B, et al. Transcranial sonography in differential diagnosis of Parkinson disease and other movement disorders [J]. Chinese Medical Journal, 2021, 134(14): 1726-1731.

(责任编辑:康 锋)