·康复工程·

随机森林算法应用于步态评估的研究*

贺 刚[1] 何汉武[1] 韦宇炜[1,2]

摘要

目的:从采集的步态数据中寻找对步态类别识别具有可分性判据最优特征子集,评估各个步态特征重要性及其对异常步态的分类识别效果。

方法:选取2017年6月—2018年8月广东省广州市第二中医院采集到的异常步态类别的数据60例和200名健康人群的信息,且均为自愿参加。本实验采用三维步态捕捉系统获取被测者的步态参数,进行统计分析和数据处理,然后采用随机森林算法获得各个步态特征的权值,并在此基础上进行特征选择,最后再用不同算法所得的分类的准确性对特征选择的效果加以评定。

结果:应用随机森林算法给出各个步态特征的权重评分,并依照其评分大小进行特征重要性排名,去掉对步态分类识别相对而言几乎不起作用的特征,发现经过特征选择后的步态特征子集,相比于特征选择之前,其步态分类识别的效果要更好,各个分类算法的多项评价指标均有所提升。

结论:应用随机森林算法能够对步态特征在步态类别的识别中进行权重打分,从而进行步态特征选择,提高对步态类别识别的准确率。

关键词 随机森林;分类;步态特征;权重;识别

中图分类号:R743.3,R496 **文献标识码:**B **文章编号:**1001-1242(2020)-05-0585-05

跌倒行为是造成我国居民伤害死亡的第四位"元凶",在65岁以上老年人位居首位,除了导致老年人死亡,还导致大量残疾。据报道,每年大概有30%左右65岁以上的老年人会发生跌倒行为,并且发生跌倒的次数和概率会随着其年龄的增长而进一步提高。跌倒行为会损害老年人的身心健康,并给其生命质量带来极大负面影响。相关研究表明,异常步态会增加跌倒风险,大约有40%的老年人都存在步态问题,而其中更是有一半存在严重的步态问题。而要较为准确地识别异常步态,则需要选择最优分类能力的步态特征并进行分析,过滤冗余步态特征,提高对异常步态识别的准确率。从而指导老年人的步行活动,避免跌倒行为的发生。

1 系统与方法

1.1 步态设备采集系统

该步态采集设备基于加速度传感器研发而成,由8块尺寸5cm×3cm、采样频率为100Hz的步态关节传感器及1个蓝牙接收器组成;该步态捕捉系统主要功能由受试者信息管理、数据实时采集、数据分析三大功能模块组成。

病理信息管理模块主要用来登记管理被测人员的基本信息(身高、体重、姓名、年龄、病历号、测试距离、测试日期、临床诊断等基本信息),以及对测试记录进行增加、删除、修改、查看等基本操作并将数据存入数据库。数据采集模块由数据采集和数据动态显示、动画模拟仿真等主要功能组成。首先,需要将步态传感器进行正确佩戴,分别正确匹配到受试者髋、膝、踝等关节上,并且左右两侧不能混淆佩戴,否则会出现报错。正确佩戴好之后,躯体挺直,双腿并拢进行校零。然后录入受试者的基本物理信息,选择菜单,步态传感器与软件系统检测到连接成功之后,受试者沿直线方向走一段距离,然后步态传感器实时采集信息,并将数据录入数据库。

数据分析模块主要包括数据回放和数据预处理,其中数据回放模块主要根据速度和角加速度等变化等数据,采用数字图像处理的方法识别下肢的时间和空间位置的信息,以此来提取受试者的步态特征,主要包括运动学和动力学参数,如髋、膝、踝关节角度伸展、屈曲最大值、最小值,以及步频、步幅、支撑相、步态周期、步长偏差等,然后生成测试报告,报告中含有被测者的性别、年龄、身高体重等基本信息、提取到的运动学、动力学及关节活动角度图等参数。

DOI:10.3969/j.issn.1001-1242.2020.05.014

*基金项目:广东省科技计划项目(2017B020210009);广州市科技计划项目(201704020110)

1 广东工业大学机电工程学院,广州,510000; 2 通讯作者

第一作者简介:贺刚,男,硕士研究生; 收稿日期:2018-08-29

1.2 分析方法

1.2.1 数据采集：为了实现异常患者步态模式识别，而考虑患有运动疾病的异常步态受试者偏少，造成数据收集上的困难，因此，本文基于步态传感器收集60例患有运动功能障碍的患者和200例健康人群的步态信息分别命名为P组和N组，并利用三维步态捕捉系统的数据采集模块获取被测者的步态参数，本文中患有运动功能障碍的患者由广东省第二人民医院的志愿者自愿参加，而作为对照的健康组成员则由下肢健全、未有损伤或患有神经系统方面疾病的人员组成。所有人员均是在知情且同意的情况下取得。在测试过程中，被测试人员在放松自然的转态下沿直线正常行走，所有测试环节均在相关医护人员的指导下完成，见表1。

表1 受试者基线信息 （$\bar{x}\pm s$）

组别	例数	年龄（岁）	体重（kg）	身高（cm）
P组	60	50.83±9.29	63.64±12.83	163.45±6.9
N组	200	52.88±7.66	62.43±8.68	162.38±8.63

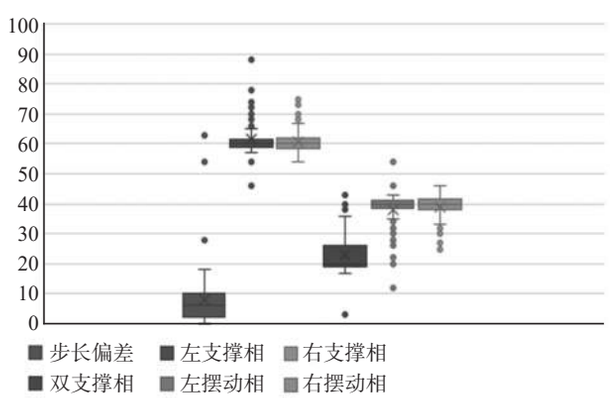
注：P组：运动功能障碍患者组；N组：健康人组

1.2.2 数据预处理：本实验中主要通过填补异常值、分析异常值及数据规范化的方法来完成数据预处理。

缺失值分析：在测量的步态数据中编号为36的那一栏，步幅出现了空值，为填补该空缺值，本文采取的方法是：查找在已有的数据集中，查找与该编号各步态参数最接近的样本，然后将该样本的步幅参数填入到该空缺值中。

异常值处理及分析：采用箱型图对实验数据进行分析，根据采集到的实验数据绘制出，见图1。

图1 异常值检测的箱型图



通过对异常样本进行回溯，发现摆动相异常和支撑相异常出现较高的一致性。左支撑相和左摆动相异常样本均为19,23,36,38,40,42,54，而右支撑相和右摆动相异常的样本则为12,28,37,46,55,72，结合箱型图进一步得知，19,54为异常样本中左支撑相偏低，左摆动相偏高的样本，而12则为右支撑相偏低，右摆动相偏高的样本，经过调取病例资料发现，12,19,54为单侧患肢较严重的偏瘫患者，因此，在行走过程中，会有一定程度的拖腿、划步等姿态，造成患肢支撑相和摆动相异常。

数据规范化处理：由于身高和步速、步频以及跨步周期和站立相时间有明显的正相关性[1]，因此，需要通过相应的数据转换方式，来消除身高因素对步态参数造成的影响，为此引入身高和重力加速度g，并作如下处理：

$$\hat{l}=l/l \tag{1}$$

$$\hat{V}=\frac{v}{\sqrt{gl_0}} \tag{2}$$

$$\hat{f}=\frac{f}{\sqrt{gl_0}} \tag{3}$$

$$\hat{t}=\frac{t}{\sqrt{l_0/g}} \tag{4}$$

$$\hat{X}=\frac{x-u}{\sigma} \tag{5}$$

其中，l为步长参数，t为时间参数，包括双支撑相、步态周期、左支撑相以及右支撑相等多种步态时间特征参数，f为步频，v则表示步速，g为重力加速度常量，其值为9.8m/s^2，用以消除身高对步速以及时间等步态特征参数的影响，式（5）为归一化处理，以消除各个步态特征因量纲和取值范围等差异的影响。

1.3 特征选择

1.3.1 相关性分析：区分变量之间的线性相关程度的强弱，需要进行相关性分析，判断两个变量之间是否具有线性相关关系的最直观的方法是绘制散点图，本文中需要同时考察多个变量之间的关系。因此，本文采用散点图矩阵同时绘制各变量间的散点图，从而快速发现多个变量间的主要相关性，图2。

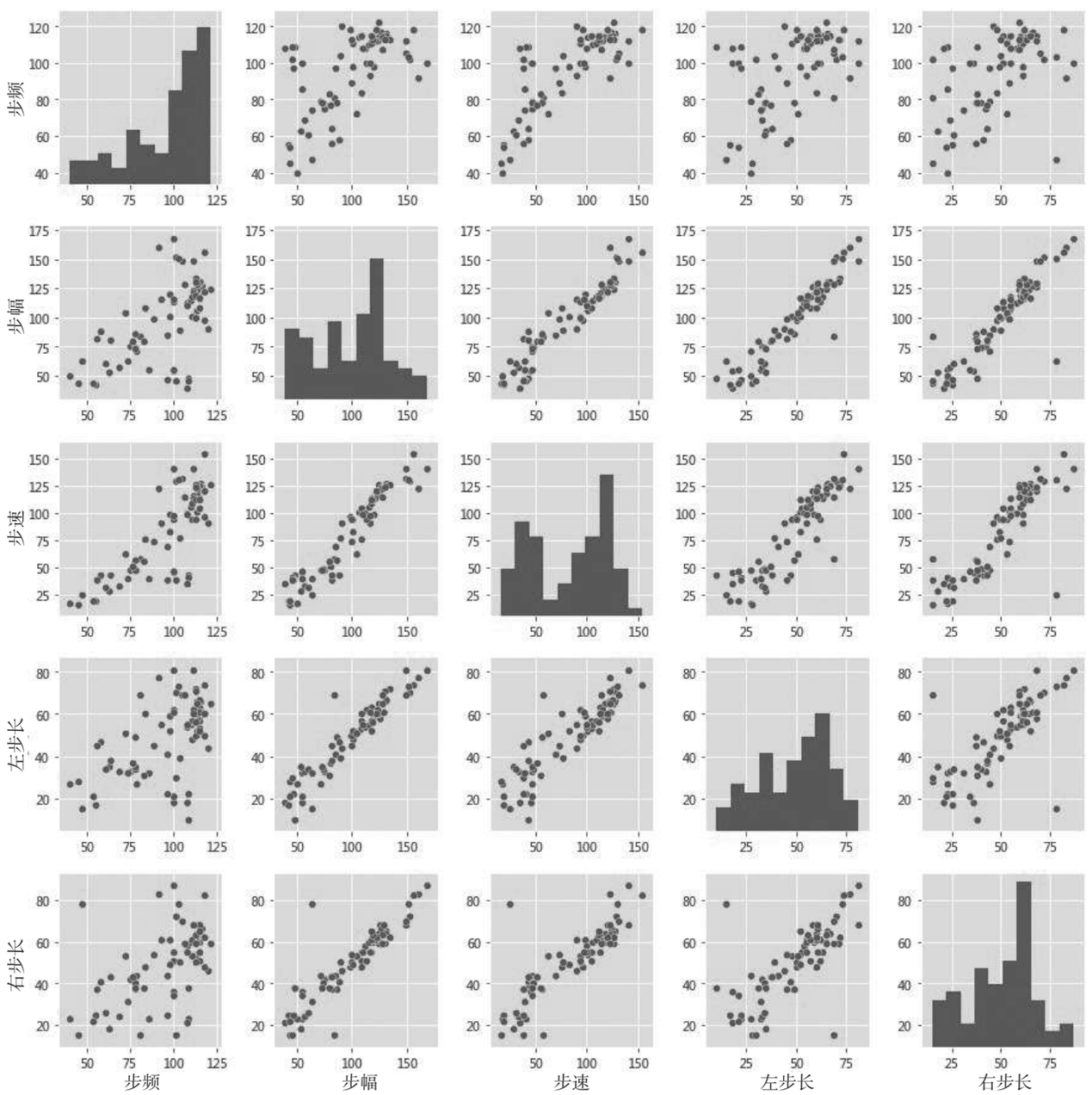
从散点图矩阵可以发现，步频与步速、左步长及右步长呈现出较高程度的线性比例关系。

1.3.2 计算相关系数：为了更加准确的描述变量之间的线性相关程度，可以通过计算相关系数进行相关分析，相关系数的计算公式如下。

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{6}$$

相关系数r的取值范围：$-1\leqslant r\leqslant 1$，根据其绝对值的大小可以判断变量之间的相关程度：r的绝对值≤0.3为不存在线性相关；r绝对值在0.3—0.5之间为低度线性相关；r绝对值在0.5—0.8之间为显著线性相关；r绝对值>0.8则为高度线性相关；而当绝对值等于1则表示两者完全线性相关。通过计算相关系数，可以找出各个步态特征之间的关联程度。通

图2 步态特征散点矩阵图



过计算发现步幅与步速的相关系数为1,成完全线性相关,与左步长和右步长的相关系数为0.9,成高度线性相关。因此,为去掉冗余数据特征,在数据中保留步幅特征,去掉步速、左步长和右步长等数据项。

1.4 随机森林模型构建

1.4.1 随机森林算法简介:随机森林算法是由美国学者Breiman于2001年提出的一种集成学习算法,常见的集成学习算法包括有提升算法(如AdaBoost)以及装袋算法等。而作为一种机器学习模式,目前正成为该领域研究的热点,因为集成学习能够弥补单一学习方法的不足。由于其具有良好的性能表现而在诸如生物信息[2]、医学研究[3]、文本分类[4]、商业管理[5]、经济金融[6]等相关领域得到了广泛应用,并且都取得了不错的效果。随机森林是以决策树为基本分类器的一个集成学习模型,由包含多个Bagging集成学习技术训练得到的决策树,当输入待分类的样本时,最终的分类结果由单棵决策树的输出结果投票决定。其主要由Bagging[7]和特征

子空间[8]的两大随机化思想决定的：

Bagging思想：从原来样本集中有放回地随机抽取与原样本集同样大小的训练集，然后每次通过抽取的训练样本集构造与之相对应的决策树。

特征子空间思想：在对决策树每个节点进行分裂时候，从样本集的全部属性中等概率地随机抽取一个属性子集，然后再从这个子集中选择一个最优属性来对节点进行分裂。

建立随机森林的基本方法步骤是，通过bootstrap重采样原理，然后反复生成所需要的训练集和测试集，在得到训练集的基础上，进一步生成一组决策树分类器，且自变量为独立同分布的随机向量，在给定自变量的情况下，每个决策树分类器通过投票来决定最优的分类结果。随机森林有两个重要参数，一是树节点个数，二是随机森林中所选择树的棵数。

1.4.2 随机森林变量的重要性评分的步骤：设有N个待分类的样本，表示各个步态参数的变量依次为$X_1,X_2,\cdots,X_m$。应用自助法重采样技术有放回地抽取K个新的自助样本集，在此过程中得到k个分类回归树，每次未被抽中的样本则组成了k个袋外数据(Out-of-bag,OOB),该部分数据样本作为测试样本用于评估各个步态参数变量在分类中的重要性。具体步骤如下：

①用自助样本可以得到若干个树形分类器，同时对相应的OOB进行分类，得到k个OOB样本中的每一个样本的投票数，记为$v_1,v_2,\cdots,v_k$。

②将变量X_i的数值在k个OOB中的顺序进行随机改变，形成新的OOB测试集，根据样本判别的正确数，所得到的结果可以表示为：

$$\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ v_{31} & v_{32} & \cdots & v_{mk} \end{bmatrix}$$

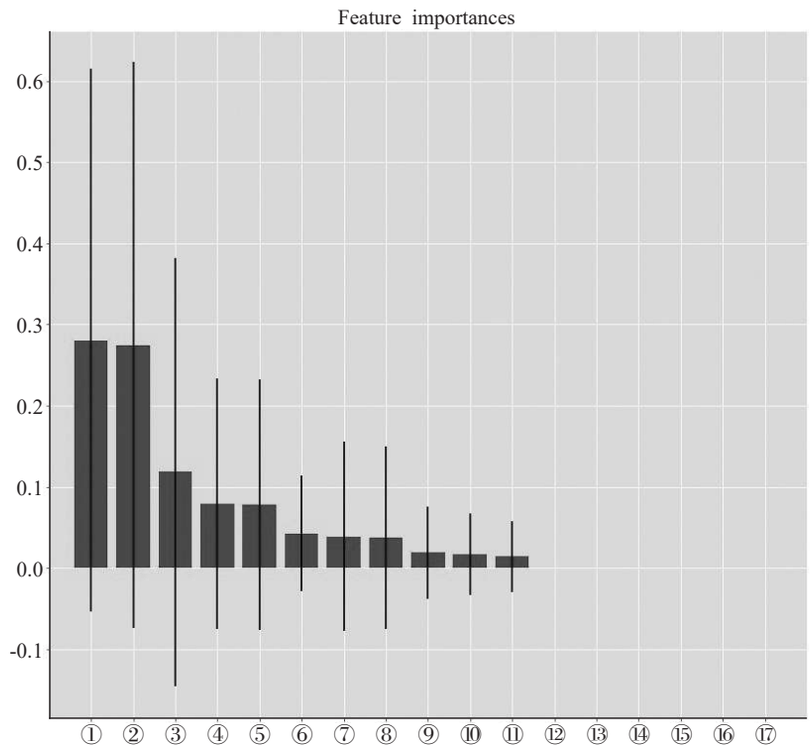
③$v_1,v_2,\cdots,v_b$与矩阵(7)对应的第i行向量相减，求和平均后得到各个变量X_i的重要性评分即：

$$\text{score}_i = \sum_{j=1}^{b}\left(vote_j - vote_{ij}\right)\Big/b \qquad (7)$$

各个步态特征的评分排名如图3所示。

从图中可以看出，有的步态特征如左摆动相、双支撑相等对权重评分很低，在分类中几乎不起作用。因此，可以考虑去除掉这些特征。

图3 各个步态特征的评分排名



①步速；②步态周期；③步频；④右摆动相；⑤右膝屈值；⑥右髋屈值；⑦左髋伸值；⑧右支撑相；⑨左支撑相；⑩右髋伸值；⑪步长偏差；⑫左髋屈值；⑬左膝伸值；⑭左摆动相；⑮双支撑相；⑯左膝屈值；⑰右膝伸值

1.5 模型训练

在将数据清洗的预处理准备工作完成之后，对于经过筛选好的步态数据集，本文在采用随机森林算法的同时，添加KNN(k-nearest neighbor)、NB(Naive Bayesian)和SVM(Support Vector Machine)等分类算法对正常人群和异常步态人群的步态模式进行训练并分类识别。本文采用5折交叉验证方法，其具体做法是：将该步态样本数据集随机切割成5个互不相交的容量相等的样本子集，然后将4个样本子集训练和优化模型，用留下的剩余子集测试、评估模型。将这一步骤重复进行5次，从中选择出测试误差最小的模型。

1.6 模型的评价指标

在二分类问题中，评价分类器性能的指标一般用分类准确率(accuracy)、召回率(recall)、精确度(precision)、F-measure来进行衡量。通常情况下，我们把关注的类记正类，而将其他类记作为负类，各个评价指标的计算公式如下所示[9—10]：

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \times 100\% \qquad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (9)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (10)$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad (11)$$

2 结果

未用随机森林算进行特征选择时,各算法的分类预测效果见表2。用随机森林算进行特征选择后,各算法的分类预测效果见表3。

表2 未进行特征选择时,各算法的分类识别效果

算法	准确率(%)	召回率(%)	精确度(%)	F-measure(%)
KNN	79.32	60.83	83.42	68.34
NB	77.64	57.62	76.58	61.32
SVM	85.68	63.35	87.67	72.44

表3 用随机森林进行特征选择后,各算法的分类识别效果

算法	准确率(%)	召回率(%)	精确度(%)	F-measure(%)
KNN	80.85	62.34	85.21	69.82
NB	79.74	59.37	79.33	63.55
SVM	88.92	66.75	90.21	75.88

3 讨论

跌倒行为严重影响老年人的身心健康,除了会导致残疾之外,严重者甚至会造成死亡。而引起跌倒行为很大一部分比例是由于步态异常引起的,并且随着年龄的增长,其跌倒次数和跌倒的概率会进一步增加。而对异常步态的及时识别则能尽早采取干预措施,指导其步行活动及其相关注意事项,则能避免跌倒行为的发生。因此,本研究的主要目的是提高对异常步态的识别率,从而减少跌倒行为。对比表2和表3可以发现,经过随机森林算法对步态特征进行选择后,训练得到的异常步态的识别模型,从准确率、召回率、精确度等指标来看,要比未经随机森林算进行特征之前的各项指标要高。由此可证,随机森林算法进行特征选择的处理后,能够提高分类模型对异常步态的识别率。因为随机森林算能够对不同的步态特征的分类能力进行权重打分,根据分数大小将步态特征进行排序,由此可以去除掉在分类中几乎不起作用的特征,保留分类能力较强的、包含丰富信息特征,从而达到了除去冗余步态信息,提高模型分类识别准确率的目的。由于随机森林的评分效果,能够对各变量的重要性进行打分,可以直观看出各个步态特征对步态的影响程度的大小,因此,在一定程度上能够反映步态特征和运动功能障碍的疾病的关系。造成异常步态原因有多种,如神经性或非神经性,另外关节、骨骼、肌肉等也会造成步态异常。若能找到步态特征与其病因的映射关系,则能在实现识别与分类的基础上,分析其原因、找到其病理,还能为患者的康复方案提供参考依据,这是接下来需要进一步研究探讨的内容。

综上所述,随机森林特征选择得到的分类模型,能够一定程度上提高对异常步态识别的准确率。本文分别用三种分类算法在特征选择前后的各项指标进行对比,完成相关验证。相比于以往研究,本研究能够更加充分地利用步态特征信息,从而提高对异常步态识别率,为相关研究提供了另一种思路。

参考文献

[1] 毕素清, 瓮长水, 毕胜,等. 偏瘫患者步态空间—时间参数对自由和最大步行速度的影响[J]. 中国康复理论与实践, 2004,10(12):736—737.

[2] Pang H, Data D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes[J]. Bioniformatics, 2010,26(2):250—258.

[3] Xie Y, Li X, Ngai E, et al. Customer churn prediction using improved balanced random forests[J]. Expert Systems with Applications, 2009, 36(3): 5445—5449.

[4] 张华伟, 王明文, 甘丽新. 基于随机森林的文本分类模型研究[J]. 山东大学学报(理学版), 2006,41(3):5—9.

[5] Kim S,Lee J, Ko B, et al. X-ray image classification using random forests with local binary patterns [C] // In proceedings of the 9th International Conference on Machine Learning and Cybernetics. Qingdao, China: IEEE Computer Society, 2010:3190—3194.

[6] 方匡南,朱建平. 基于随机森林方法的基金超额收益方向与交易策略研究[J]. 经济经纬, 2010,(2):61—65.

[7] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123—140.

[8] Ho T. The random subspace method for constructing decisionforests [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20 (8): 832—844.

[9] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves [C]. Proceedings of the 23rd international conference on Machine learning, ACM, 2010:233—240.

[10] Wang AG, An N, Chen G, et al. Predicting hypertension without measurement: A non- invasive, questionnaire- based approach [J]. Expert Systems with Applications, 2015,42(21): 7601—7609.