

DOI:10.13232/j.cnki.jnju.2021.01.007

基于随机森林的民俗体育对身体指标影响评估方法

李佳佳¹, 丁伟^{2*}, 王伯伟¹, 聂秀山³, 崔超然¹

(1. 山东财经大学计算机科学与技术学院, 济南, 250014; 2. 山东财经大学体育学院, 济南, 250014;

3. 山东建筑大学计算机科学与技术学院, 济南, 250101)

摘要:随着体育事业的发展,传统体育项目越来越受重视。作为我国传统体育运动的重要组成部分,民俗体育运动的发展显得尤为重要。为研究民俗体育项目对健身效果的影响,组织多个对象分别进行一段时间内的角力类、竞足类和技巧类三类民俗体育运动训练,并观察对象训练后在身体形态、身体机能、身体素质三方面共计 32 个代表性身体指标的变化情况。进一步,以身体指标变化情况作为特征表示观察对象,采用随机森林算法预测不同对象在训练阶段进行民俗体育运动的种类,在过程中基于信息增益进行特征选择,从而度量不同类别民俗体育运动对各项身体指标的影响程度。将获得的运动对各项身体指标的影响程度与真实影响情况进行评估分析,揭示民俗体育运动与人体机能的关系。此外,实验结果证明,和基准线算法相比,提出的算法有更高的预测准确性。

关键词:民俗体育,随机森林,身体指标,数据挖掘,特征选择

中图分类号:TP391

文献标识码:A

An evaluation method for the influence of folk sports on body indicators based on random forest

Li Jiajia¹, Ding Wei^{2*}, Wang Bowei¹, Nie Xiushan³, Cui Chaoran¹

(1. School of Computer Science and Technology, Shandong University of Finance and Economics, Ji'nan, 250014, China;

2. Physical Education College, Shandong University of Finance and Economics, Ji'nan, 250014, China;

3. School of Computer Science and Technology, Shandong Jianzhu University, Ji'nan, 250101, China)

Abstract: With the development of sports, more and more attention has been paid to traditional sports. As an important part of Chinese traditional sports, the development of folk sports is particularly important. In order to study the influence of folk sports on the effect of physical fitness, this article organizes a number of subjects to conduct three types of folk sports training including wrestling, foot races and skills sports for a period of time, and observes a total of 32 representative body indicators in three aspects: body shape, physical function and physical fitness after training. As for a further comment, based on the changes of indicators, this article uses changes in physical indicators as features to represent individuals. Random forest algorithm is used to predict what kind of folk sports different objects will perform in the training phase. In the process, feature selection is based on information gain to measure impact. Thus, it measures the influence degree of different types of folk sports on body indicators. In this paper, the influence degree of different folk sports on body indicators and actual impact are evaluated and studied to reveal the correlation between folk sports and human body functions. In addition, the experimental results show that the proposed algorithm has higher prediction accuracy than the baseline algorithm.

Key words: folk sports, random forest, body indicators, data mining, feature selection

基金项目:国家社会科学基金(17BTY056),山东省科技发展计划(2011GGH21816)

收稿日期:2020-10-09

* 通讯联系人, E-mail: scdw2013@163.com

随着社会的发展,体育形式的发展趋于多样化,我国处于不断积累的快速发展阶段,正由体育大国向体育强国发展.民俗体育是中国体育发展的重要组成部分,民俗体育可以定义为“活动形态下的一种以身体运动为基本手段,以促进身心健康发展为根本目的的文化活动”^[1].民俗体育项目囊括走、跑、跳、投、悬垂、支撑、爬越、平衡等身体运动形式,具有动员身体各个部位参与运动的特征.本着项目分类原则,依照运动性质与形式特点,民俗体育项目可以划分为:角力类、竞足类、技巧类、投射类、表演类和棋牌类.本文主要研究角力类、竞足类和技巧类三类运动.角力类运动主要包含掰手腕、斗拐、推手、拉钩、推车、拔河、拉棍和顶杠等徒手力量型体育项目.竞足类运动主要包括网鱼、滚铁环、老鹰捉小鸡、编花篮、跳绳和开火车等跑跳项目.技巧类运动主要包括抓石子、砸沙包、跳房子、跳皮筋、侧手翻、丢手绢、跳山羊和踢毽子等高灵活性项目.

民俗体育运动有一定健身功效,对人们的身体形态、身体机能、身体素质有不同的影响,主要体现为进行民俗体育运动后相关身体指标的变化,所以民俗体育运动对身体指标影响的评估即为评估研究民俗体育运动对身体指标的影响程度.部分学者提出利用计算机技术和数学工具来挖掘体育运动数据的内蕴价值,其中最主流的研究^[2-9]是依靠统计学方法来实现这一目标.现有的诸多研究多用简单描述分析^[3-6]方法,如统计不同群体不同体质指标的平均水平、标准差等;体质指标之间、体质指标与其他影响变量之间相互影响分析时,主要应用简单相关系数法^[6].此类研究只能研究指标间一对一的影响关系,不具有综合代表性,不能确定不同因素对总体体质水平的影响;同时,体质数据多为截面数据,相关性不显著,影响作用不能被反映出来.综上,基于机器学习的数据挖掘方法很少应用在体育数据中,尤其是身体指标数据影响方面.

为研究民俗体育对身体指标的具体影响,本文提出基于随机森林算法的民俗体育对身体指标影响评估方法,具体是利用随机森林算法根据对象的身体指标变化情况来分类预测个体进行的民俗体育运动种类,在这个过程中,进一步基于信息

增益的思想进行特征选择,从而获得不同运动项目对各种身体指标的影响程度,进一步对比各类运动的实际真实影响程度,利用实验评价指标分析研究民俗体育运动对身体指标的影响程度.图1给出我们评估方法的一个框架性描述.

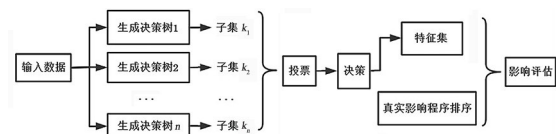


图1 基于随机森林的民俗体育对身体指标影响评估法模型示意图

Fig.1 The illustration of the random forest based evaluation method for the influence of folk sports on body indicators

本文的主要贡献如下:

(1)构建并发布了民俗体育运动对身体指标影响评估数据库Folk-CS(Folk Custom Sports),该数据库来自200名中小学学生真实训练数据,对数据进行了整理、标注,可为今后体育运动数据的挖掘提供数据研究基础.

(2)首次将机器学习算法运用到民俗体育对身体健身功效的影响当中,利用随机森林算法揭示了不同的民俗体育运动对人体不同指标的影响程度.

(3)利用评价指标对本文提出的评价方法进行了实验对比,实验结果表明,该方法取得了较好的评估性能,有更高的准确性.

1 相关工作介绍

本文基于随机森林算法,研究民俗体育运动对身体指标的影响.本节首先介绍体育数据挖掘方面的工作,随后介绍随机森林算法方面的工作.

1.1 体育数据挖掘 随着体育数据量的迅速增长,体育数据挖掘逐渐引起相关学者的关注.研究民俗体育运动对身体指标数据的影响即体育数据挖掘问题.数据挖掘(Data Mining)是指从大量的、不完全的、有噪声的、模糊的、随机的实际应用中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程.体育数据挖掘则是将数据分析应用在体育运动中.在国

内,体育数据挖掘技术主要应用在体育教学、学生体育调研、运动训练监控、临场比赛优化和体育信息管理等,其中涉及离散数据挖掘、图像数据挖掘及视频数据挖掘。本文研究的是基于学生体育调研的离散体质数据挖掘方面。

对身体指标数据的挖掘是指对体质的研究,国外对该研究开始较早,研究内容和成果较为丰富。早在19世纪60年代,美国医学博士爱德华·希契科克就以学生为研究对象,进行身体能力的测试^[2]。许浩和姜文凯^[3]利用江苏省2000年国民体质监测数据,运用数理统计分析方法,着重分析该省成年人身体质量指数(Body Mass Index, BMI)特征。张勇等^[4]选取2015年体质监测数据,以2014年全国国民体质监测公报为依据,利用统计学方法对浙江师范大学男性教职工进行分组并比较分析。周皎等^[5]采用问卷调查方式,搜集了山西大学3894名大学生的体质健康水平和生活方式因素,研究不同生活方式对体质健康等级的影响。乔克满和孙卫^[6]以2005年上海市国民体质监测数据库中11407条20~39岁成年人数据为对象,运用关联规则数据挖掘技术,对市民身体形态、机能和素质中的21项重要体质指标进行处理和分析。陶弥锋^[7]分析江西省高校学生2013—2014年的体测数据,并结合问卷调查结果研究其体质健康存在的问题。杜志峰和周艳华^[8]简要论述大数据技术在中小学校学生体质健康监测中应用价值,系统分析其利用方式,并提出具体的保障措施,旨在为当前中小学校实施学生体质健康工作提供必要的指导。张京舒等^[9]利用1985—2014年全国学生体质与健康调研数据,分析不同年份学生体质健康达标优良率的变化趋势以及不同特征学生优良率的差异,并采用log-binomial回归模型分析学生体质健康的相关因素。

近些年来,学者更关注对图像及视频数据的挖掘。Aoki et al^[10]用一个系数来量化预测体育赛事的结果,这个系数测量的是观察到的体育联盟的最终结果与在技能方面理想化的完美平衡比赛之间的距离,还提出一个概率图形模型来学习团队的技能,并分解运气和技能在每一场比赛中的相对权重。Decroos^[11]介绍了一种新的先进的足球度量方法,用于评估球员在球场上的任何类型

的动作,无论有无球。我们的度量标准基于行动对游戏结果的影响来评估每个玩家的行动,同时考虑行动发生的环境。宋兆铭等^[12]认为C5.0决策树算法可以用来判断影响警察院校学生体质测试成绩的关键因素,为深层挖掘相关警务数据内涵与监测提供了实证依据。冯敏等^[13]利用同济大学568名学生的问卷调查基于随机森林模型对大学生体质健康影响因素降维,利用相关性分析获得重要影响因素及重要影响因素之间的相互作用。刘玉琪^[14]提出人工蜂群优化的随机森林算法,作为分类器应用于人体运动模式识别,可以得到较高的分类精度。

传统的研究方法主要基于定期的国民体质监测报告或调查报告数据,虽然对体育运动功效的研究很重要,但没有合适的公开数据集,而本研究则是利用真实的数据。另外,其他研究者多使用统计学方法。因此,基于机器学习的体育数据挖掘的地位就非常重要。

1.2 随机森林 本文研究的问题属于分类问题,随机森林算法就是解决分类问题常用的一种算法。随机森林是利用多棵树对样本进行训练并预测的一种分类器。随机森林算法比神经网络更容易被人接受,精确度更高,对有噪声和缺失的数据鲁棒性更强、运算更快,因此数据挖掘中随机森林算法比较常用。

随机森林^[15]是2001年Breiman提出的一种集成机器学习方法,是一个由一组决策树分类器构成的集成分类器,每个决策树分类器通过投票来决定最优的分类结果,其输出的类别是由个别树输出的类别的众数而定。

随机森林算法是许多决策树集成在一起的分类型器,如果把决策树看成分类任务中的一个专家,随机森林就是许多专家在一起对某种任务进行分类。随机森林算法具有较高的准确性,而且对离群点不敏感,其泛化误差随着树的棵数的增多而收敛,因此不易过拟合。另外,每棵树划分时需要考虑的候选特征较少,计算速度快,最重要的是能给出变量的重要性的估计。自Breiman提出随机森林算法后,由于其良好的性能,已被广泛应用于各个领域^[12-14],如生物领域信息分类和回归、经济和金融领域中的客户信用和欺诈的分析、计算

机视觉领域中对人体的监视和跟踪、目标识别和行为与事件识别、语音领域中的语音识别和语音合成和数据挖掘领域中的异常检测。

深度森林算法是 Zhou and Feng^[16]提出的一种集成的森林模型,是传统的森林模型在广度和深度上的一种集成。深度森林算法利用多粒度扫描使输入数据的差异性有足够的体现,利用级联森林提高分类准确度,其作为一种基于树的方法,更容易做理论分析。基于深度森林算法的模型被应用于分类问题及预测问题^[17-20],如文本情感分析、网络行为分类、用户购买行为预测、火焰预测等。

本文方法是基于随机森林算法提出的评估方法,应用于数据挖掘领域及体育领域。本文提出一种基于随机森林的民俗体育对身体指标影响评估方法,主要利用随机森林特征选择的特性,评估民俗体育运动对中小学学生身体素质各项指标的影响程度。

2 基于随机森林的民俗体育对身体指标影响评估方法

首先对民俗体育运动进行数据建模,介绍本文构建的民俗体育数据库 Folk-CS,随后介绍利用随机森林算法实现民俗体育运动对身体指标影响的评估方法。

2.1 Folk-CS 数据库 为研究民俗体育对身体指标的影响,首先基于真实数据,构建了一个民俗体育对身体指标影响的数据库(Folk-CS)。

Folk-CS 数据库来自对 200 名中小学学生的实际测量。研究团队首先依照民俗体育运动类别,把民俗体育分为角力、竞足、技巧三类。另外,为体现不同运动对身体指标影响,本文增加了现代体育运动和无体育运动两类。因此,本文所用到的数据共分五类:角力类运动,主要包括掰手腕、斗拐、倒拉牛、拔河和拉棍项目;竞足类运动,主要包括网鱼、滚铁环、老鹰捉小鸡、编花篮和跳绳项目;技巧类运动,主要包括抓石子、砸沙包、跳房子、跳皮筋和踢毽子项目;现代体育类运动,主要包括篮球、足球、排球和跑步项目;无体育类运动则不做任何项目。

Folk-CS 数据库中身体指标共有 32 项,其中,身高、坐高、上臂围松紧差、胸围、呼吸差、腰围、臀围、肩宽、骨盆宽、体重和体脂率属于身体形态指标;基础心率、心功指数、脉压差、肺活量和最大摄氧量属于身体机能指标;握力、背肌力、一分钟仰卧起坐、立定跳远、俯卧撑、选择反应时、50 m 短跑、十字变向跑、往返跑、反复横跨、纵横叉、坐位体前屈、转肩、立位转体、闭目单足立和一分钟抛网球属于运动素质指标。

测试学生共分五组,每组对应一种类别,研究团队首先测试每组训练前的身体指标,身体指标数据记为 P_0 ,然后专人指导每组进行相应类别运动的训练,每周锻炼三次,每次锻炼 40 min,其中包括 7 min 的准备部分和 3 min 的结束部分,基本部分为 30 min,实验时间为三个月。三个月后的身体指标数据记为 P_1 。表 1 给出了数据库中角力类运动的一部分测量数据示例。

表 1 Folk-CS 数据库的测量数据举例

Table 1 An example of measured data of Folk-CS database

	Person	身高(cm)	坐高(cm)	转肩(cm)	50 m 短跑(s)	平均握力(kg)	背肌力(kg)	纵横差(cm)
P_0	1	151.5	81.3	68.4	9.5	20.65	60.2	19.4
P_0	2	154.9	83.2	63.5	9	25	48.9	18.8
P_0	3	148.6	79.4	77.5	9.2	19.75	51.2	28.85
P_0	4	155.3	84.3	64.5	8.9	22.6	58.5	20.9
P_1	1	153.2	81.9	70.4	9.2	25.15	65.7	18.8
P_1	2	156.4	84	65.2	8.8	28.6	55.2	18.2
P_1	3	150.3	80.2	78.7	8.9	24.15	57.1	28.25
P_1	4	156.8	85.1	66.2	8.7	26.7	64.4	20.3

对收集的数据经过数据清洗及数据预处理等步骤,共获得200人的身体指标变化数据并进行分类标注.进一步,研究团队和体育专家合作,获得体育数据对身体指标的真实影响程度排序标注.式(1)表示对于给定一个类别的运动, z 位投票人根据经验基于此类运动对身体指标的影响排序投票求和后,获得身体指标的重要程度票数($Impact$):

$$Impact_C(a_p) = \sum_{m=1}^z vote(m, a_p) \quad (1)$$

其中, $Impact_C(a_p)$ 表示第 C 类运动对第 p 项身体指标的重要程度, m 代表第 m 位投票人, z 代表投

票人的总数.本实验共邀请五位体育专家作为投票人,即 $z=5$.

具体的,给定每个专家角力类、竞足类和技巧类三类运动,请专家根据经验针对各类运动对已测量的32个指标的各自影响程度进行排序投票,然后对五位专家投票结果进行相加,按照最终票数的由高到低排序,获得三类运动影响显著的有序的前10位身体指标序列,构建体育运动对身体指标影响的真实影响程度(Ground-truth),最终形成民俗体育运动对身体指标影响数据库.表2给出了三个类别运动对各项身体指标影响程度的真实排序.

表2 各类运动对身体指标影响程度的真实排序

Table 2 The real order of the impact of various sports on body indicators

	1	2	3	4	5	6	7	8	9	10
角力类	背肌力	平均握力	上臂围 松紧差	立位转体	心功指数	呼吸差	一分钟 仰卧起坐	肺活量	俯卧撑	体脂率
竞足类	往返跑	反复横踢	立位转体	十字 变向跑	50 m 短跑	立定跳远	心功指数	闭目 单足立	俯卧撑	纵横叉
技巧类	选择 反应时	反复横踢	心功指数	闭目 单足立	50 m 短跑	1 min 抛网球	十字 变向跑	立位转体	往返跑	纵横叉

2.2 基于随机森林的评估方法 本文利用随机森林的特征选择方法研究数据特征的重要性,推断对应体育运动对相应身体指标的影响.通过测量获得分组训练前后身体指标数据,计算得到民俗体育训练前后身体指标变化数据;利用随机森林的特征选择方法,获得民俗体育对身体指标影响较大的指标属性集.构建随机森林模型的数据集输入时,分别将每一类民俗体育数据与无体育运动数据作为输入数据集进行训练来做二分类;另外,也将每一类民俗体育数据与现代体育运动数据放在一起作为输入数据集进行训练,形成对照实验,以便能更好地体现各类民俗体育对身体指标的影响程度.具体方法如下:

受测者在训练前后的身体指标数据分别记为矩阵 $P_0 = \{x_{ij}\}^{N \times 32}$ 和 $P_1 = \{y_{ij}\}^{N \times 32}$,其中 N 为受测人数, P_0 为民俗体育训练前的身体指标数据, P_1 为训练三个月后的身体指标数据.如式(2),训练前后身体指标变化矩阵记为:

$$P = \left| (P_1 - P_0) / P_0 \right| \quad (2)$$

数据库的类别向量记为 $C = \{C_1, C_2, C_3, C_4, C_5\}$,其中, C_1 至 C_5 分别表示民俗体育角力运动类、竞足运动类、技巧运动类、现代体育运动类和无体育运动类.表3给出了一个最终数据库的部分示例.

数据库中反映民俗体育运动对身体影响的身体指标,如“平均握力”“背肌力”,称“属性”或“特征”,则数据库的属性集记为 $A = \{a_1, a_2, \dots, a_p\}$.其中数据库 P 中含有32维属性,每维属性有多个可能的取值.图2展示了算法的流程,且本部分以竞足类民俗体育运动为例,来介绍一下算法的流程.

本文首先把无体育运动类(标记为 C_5)和民俗体育竞足运动类(标记为 C_2)作为样本集 D , D 中有32个指标,每个指标下有多个样本点,对于指标属性 a , a 有 V 个可能的取值 $\{a^1, a^2, \dots, a^V\}$,若用指标 a 对训练集 D 进行划分,则会产生 V 个分支节点,其中第 v 个分支节点包含了 D 中所有在属性 a 上取值为 a^v 的样本,记为 D^v . $Ent(D)$ 是

表 3 Folk-CS 数据库的示例

Table 3 An example of Folk-CS database

	Person	身高	坐高	转肩	50 m 短跑	平均握力	背肌力	纵横差	
	C_1	1	0.011221122	0.00738	0.02924	0.031578947	0.218729427	0.091362126	0.03145
	C_1	2	0.009683667	0.009615	0.026772	0.022222222	0.144020739	0.128834356	0.032537
	C_1	3	0.011440108	0.010076	0.015484	0.032608696	0.22341629	0.115234375	0.021081
	C_1	4	0.009658725	0.00949	0.026357	0.02247191	0.181575643	0.100854701	0.029159

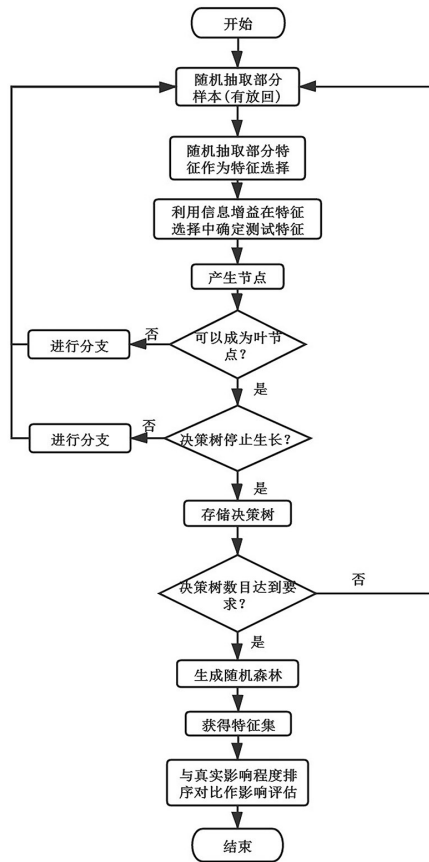


图 2 基于随机森林的民俗体育对身体指标影响评估算法流程示意图

Fig. 2 The flow diagram of the random forest based evaluation method for the influence of folk sports on body indicators

D 的“信息熵”^[15], 其计算式为:

$$Ent(D) = - \sum_{q=1}^{|y|} p_q \log_2 p_q \quad (3)$$

其中, $p_q (q = 1, 2, \dots, |y|)$ 表示当前样本集合 D 中第 q 类样本所占的比例。

考虑不同分支节点所包含的样本数不同, 给分支节点赋予权重 $|D^v|/|D|$, 则可计算用属性 a

对样本集 D 进行划分所获得的“信息增益”^[15]:

$$Gain(D, a) = Ent(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} Ent(D^v) \quad (4)$$

每次剔除一定比例的特征, 利用信息增益进行属性选择, 得到新的属性集 a_* .

$$a_* = \arg \max_{a \in A} Gain(D, a) \quad (5)$$

将现代体育运动类(标记为 C_4)和民俗体育竞足运动类(标记为 C_2)作为样本集进行以上算法, 亦得到另一个新的属性集; 进行属性集重要性排序, 对比得到的两个属性集; 研究对竞足运动影响更明显的指标, 最终得到针对竞足这一类民俗体育对身体指标的影响重要的属性。

通过将民俗体育竞足运动类和有无现代体育运动类分别作为样本集, 构建指标影响评估算法, 不仅可以获得民俗体育对身体指标的影响程度, 而且可以分析出民俗体育竞足类对比现代体育运动类对身体指标影响更重要的指标。

在研究其他类别的民俗体育时, 同样通过以上算法, 最终获得各类民俗体育对身体指标影响较大的指标属性集。通过研究获得的影响较大的指标属性集, 分析推断出对应体育运动对身体指标的影响情况。本文通过算法得出的民俗体育对身体指标影响程度与 Ground-truth 进行对比评估, 利用评价方法验证算法的准确性。

3 实 验

本节使用构建的数据集测试了提出的基于随机森林对身体指标影响评估模型, 并且利用两种数理统计方法(平均值法与方差法)和另外两种机器学习算法(支撑向量机^[21]与递归特征消除算法^[22])作为基准线算法(baseline)来进行对比实验分析。

3.1 实验设置 本文方法在 Folk-CS 数据库上

进行了实验,数据库设置为训练集和测试集,比例为4:1. 本文主要研究民俗体育运动及现代体育运动对身体的指标变化影响,民俗体育运动实验获得数据作为正类,其他数据作负类,对两种数据进行对比实验. 另外,用四种基准线算法进行了对比实验,这五种方法与数据集中 Ground-truth 数据进行对比得到评估结果.

3.2 实验评价方法 本文利用随机森林方法进行特征选择,选择出的特征对分类问题来说是最重要的特征,即对民俗体育影响最大的指标.

本文的实验评估准则是 $\text{top}@k$ 的准确率,定义为基于算法获得的身体指标影响与真实情况 Ground-truth 相匹配的比率. 准确率越高,代表算法越有效. 准确率(Precision)的计算式为:

$$\text{Precision} = \frac{k}{n} \quad (6)$$

其中, k 代表与 Ground-truth 相符的影响指标个数, n 代表在 Ground-truth 中选取的指标总数.

3.3 实验对比分析

3.3.1 不同运动对身体指标的影响情况 本文利用随机森林方法,基于特征增益对三种民俗体育运动对身体指标的影响进行特征重要性排序,发现和现代学校体育运动相比,民俗体育运动对某些特定指标的影响更大.

图3给出角力类运动对身体指标的影响程度排序,可以看出影响最大的是平均握力、心功指数、一分钟仰卧起坐、腰围、立位转体、50 m短跑和背肌力等. 角力类运动主要锻炼的就是力气,实验结果符合我们的认知.

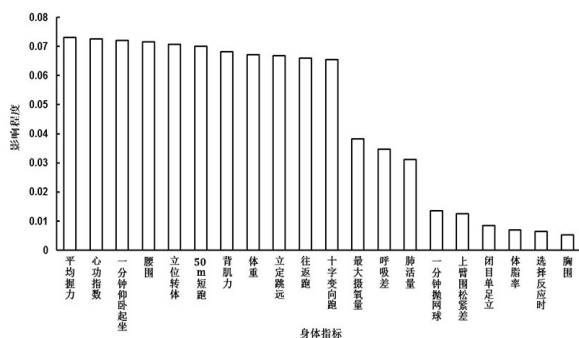


图3 角力类运动影响最大的20个身体指标

Fig. 3 The 20 body indicators with the greatest influence of wrestling

图4给出竞足类运动对身体指标的影响程度排序,可以看出影响最大的是腰围、心功指数、立位转体、50 m短跑、反复横踢、体重和十字变向跑等. 竞足类运动主要锻炼身体机能及反应力,尤其是下半身的身体机能,实验结果符合我们的认知.

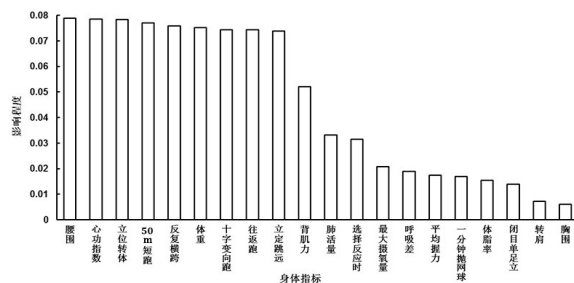


图4 竞足类运动影响最大的20个身体指标

Fig. 4 The 20 body indicators with the greatest influence of foot race

图5给出技巧类运动对身体指标的影响程度排序,可以看出影响最大的是腰围、心功指数、立位转体、闭目单足立、50 m短跑、体重、反复横踢和十字变向跑等. 影响程度较高的这八种指标,其中三种指标反映身体机能的变化,五种指标反映反应力的变化. 技巧类运动主要锻炼反应力,实验结果符合我们的认知.

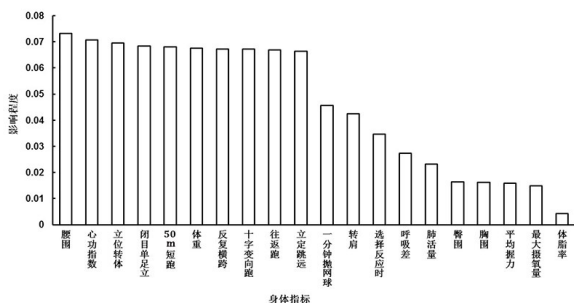


图5 技巧类运动影响最大的20个身体指标

Fig. 5 The 20 body indicators with the greatest influence of skill sports

3.3.2 对比实验结果分析 本文实验和采用均值、方差、支持向量机算法(SVM)和递归特征消除算法(RFE)的方法进行了对比,其中训练集和测试集设置方式相同,实验结果如表4所示.

可以看出,本文方法对三类民俗体育数据得到的对身体影响指标评估的准确率,在 $\text{top}@5$ 上都是最高的. 针对不同的准确率参数设置进一步

表 4 五种方法在 Folk-CS 数据集上的分类性能比较

Table 4 The classification performance of five methods on Folk-CS dataset

方法	角力类			竞足类			技巧类		
	top@3	top@5	top@10	top@3	top@5	top@10	top@3	top@5	top@10
均值	0.33	0.4	0.7	0	0.2	0.4	0.33	0.4	0.3
方差	0.33	0.2	0.7	0	0	0.4	0	0.2	0.3
SVM	0	0.4	0.6	0.33	0.6	0.7	0	0	0.7
RFE	0	0.4	0.5	0	0.2	0.4	0	0.2	0.5
Our-RF	0.33	0.6	0.5	0.33	0.6	0.7	0.33	0.8	0.7

实验,可以看到 top@3 和 top@10 的结果,本文的方法得到的准确率都明显高于其他方法,证明本文的方法对于民俗体育运动对身体指标的影响评估更有效。

从实验结果还可以看出,不同类别的民俗体育运动,其锻炼的功效不同,影响的身体指标也不同。民俗体育运动不仅能帮助学生改善身体状况,还能锻炼他们的反应力和力量。

4 结 论

本文利用随机森林算法研究民俗体育运动对学生身体健康各项指标的影响,同时构建了影响评估数据集,为下一步的研究打下基础。利用本文方法得到的实验结果与人们的主观认知和体育专家的判断基本相符。同时,本文还提供了一种简单易用的指标影响评估方式,对于民俗体育运动发展具有重要的参考价值。

参考文献

- [1] 林顺治. 体育强国建设中民俗体育文化发展研究. 军事体育进修学院学报, 2010, 29(1): 26—29. (Lin S Z. Researching on the development of folk-custom sports culture under a world sports power construction. Journal of PLA Institute of Physical Education, 2010, 29(1): 26—29.)
- [2] Powers S K, Dodd S L. Total fitness: exercise, nutrition, and wellness. Boston: Allyn and Bacon, 1999, 398.
- [3] 许浩, 姜文凯. 对江苏省成年人 BMI 指数的研究. 体育与科学, 2001, 22(6): 39—43. (Xu H, Jiang K W. A study of the body mass index of adults in Jiangsu Province. Journal of Sports and Science, 2001, 22(6): 39—43.)
- [4] 张勇, 蒋月玲, 钱永东. 2015 年浙江师范大学男性教职工体质状况研究. 当代体育科技, 2015, 5(35): 129—130. (Zhang Yo, Jiang Y L, Qian Y D. Study on physical fitness of male faculty in Zhejiang Normal University in 2015. Contemporary Sports Technology, 2015, 5(35): 129—130.)
- [5] 周皎, 乔京转, 马慧敏. 生活方式对大学生体质健康的影响研究. 辽宁体育科技, 2015, 37(6): 67—69. (Zhou J, Qiao J Z, Ma H M. A Study on the life style influencing undergraduates' physical health. Liaoning Sport Science and Technology, 2015, 37(6): 67—69.)
- [6] 乔克满, 孙卫. 身体形态、机能、素质指标的关联分析. 周口师范学院学报, 2010, 27(5): 144—146. (Qiao K M, Sun W. Association analysis on indices of body pattern, function and content. Journal of Zhoukou Normal University, 2010, 27(5): 144—146.)
- [7] 陶弥锋. 江西高校学生体质健康现状及主要问题分析. 当代体育科技, 2015, 5(35): 101—102. (Tao M F. Analysis of the current situation and main problems of college students' physical health in Jiangxi Province. Contemporary Sports Technology, 2015, 5(35): 101—102.)
- [8] 杜志峰, 周艳华. 中小学校学生体质健康大数据技术应用研究. 青少年体育, 2020(4): 37—38. (Du Z F, Zhou Y H. Application research on big data technology of physical health of students in primary and middle schools. Youth Sports, 2020(4): 37—38.)
- [9] 张京舒, 闫晓晋, 胡佩瑾等. 1985—2014 年中国汉族 13~18 岁学生体质健康达标优良率变化趋势及相关因素分析. 中华预防医学杂志, 2020, 54(9): 981—987. (Zhang J S, Yan X J, Hu P J, et al. Trends of prevalence of excellent health status and physical fitness among Chinese Han students aged 13 to 18 years from 1985 to 2014. Chinese Journal of Preventive Medicine, 2020, 54(9): 981—987.)

- [10] Aoki R Y S, Assuncao R M, De Melo P O S, et al. Luck is hard to beat: the difficulty of sports prediction//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2017:1367—1376.
- [11] Decroos T, Bransen L, Van Haaren J, et al. Actions speak louder than goals: Valuing player actions in soccer//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: ACM, 2018: 1851—1861.
- [12] 宋兆铭,叶菁,董如军.数据挖掘:C5.0决策树算法在警察院校学生体质分析中的应用.四川体育科学,2020,39(1):52—55,74.(Song Z K, Ye J, Dong R J. Ta mining: application of C5.0 decision tree algorithm in physique analysis of police college. Sichuan Sports Science, 2020, 39(1): 52—55, 74.)
- [13] 冯敏,冯辉,张一雨等.随机森林模型分析大学生体质健康影响因素:来源于同济大学568名学生的问卷调查.中国组织工程研究,2019,23(23):3722—3728.(Feng M, Feng H, Zhang Y Y, et al. Factors influencing physical health of college students analyzed by random forest model: questionnaires of 568 students from Tongji University. Chinese Journal of Tissue Engineering Research, 2019, 23(23): 3722—3728.)
- [14] 刘玉琪.基于随机森林算法的人体运动模式识别研究.硕士学位论文.北京:北京邮电大学,2018.(Liu Y Q. Human motion pattern recognition research based on Random Forest algorithm. Master Dissertation. Beijing: Beijing University of Posts and Telecommunications, 2018.)
- [15] 周志华.机器学习.清华大学出版社,2016,74—76.
- [16] Zhou Z H, Feng J. Deep forest: towards an alternative to deep neural networks. 2017, arXiv: 1702.08835.
- [17] 葛绍林,叶剑,何明祥.基于深度森林的用户购买行为预测模型.计算机科学,2019,46(9):190—194.(Ge S L, Ye J, He M X. Prediction model of user purchase behavior based on deep forest. Computer Science, 2019, 46(9): 190—194.)
- [18] 蒋鹏飞,魏松杰.基于深度森林与CWGAN-GP的移动应用网络行为分类与评估.计算机科学,2020,47(1):287—292.(Jiang P F, Wei S J. Classification and evaluation of mobile application network behavior based on Deep Forest and CWGAN-GP. Computer Science, 2020, 47(1): 287—292.)
- [19] 朱晓好,严云洋,刘以安等.基于深度森林模型的火焰检测.计算机工程,2018,44(7):264—270.(Zhu X Y, Yan Y Y, Liu Y A, et al. Flame detection based on deep forest model. Computer Engineering, 2018, 44(7): 264—270.)
- [20] 韩慧,王黎明,柴玉梅等.基于强化表征学习深度森林的文本情感分类.计算机科学,2019,46(7):172—179.(Han H, Wang L M, Chai Y M, et al. Text sentiment classification based on deep forests with enhanced features. Computer Science, 2019, 46(7): 172—179.)
- [21] 于剑.机器学习:从公理到算法.北京:清华大学出版社,2017,48—51,156—162,170—173.
- [22] 吕世聘,王秀坤,孙岩等.改进的支持向量机特征选择算法.计算机工程,2009,35(1):171—172.(Lü S P, Wang X K, Sun Y, et al. Improved feature selection algorithm for SVM. Computer Engineering, 2009, 35(1): 171—172.)

(责任编辑 杨可盛)