

· 康复工程 ·

基于决策树和逻辑回归算法的辅具适配模型研究与实现*

张小雪^{1,2,3} 张祖婷^{1,5} 刘 莉⁴ 罗椅民^{1,2,3}

摘要

目的:为实现辅具适配过程中用户自主自助及辅具适配的精准预测,研究建立辅具适配模型。

方法:通过调查问卷的方式收集到102份辅具配置信息作为数据集,基于决策树和逻辑回归算法建立辅具适配模型。使用python语言实现模型并绘制主要参数对模型性能的影响,优化后,计算准确率、召回率和F1分数评价模型的预测能力。

结果:基于决策树算法建立模型准确率达到0.91,且召回率和F1分数性能指标分别为0.91和0.90。基于逻辑回归算法建立的模型性能较低。

结论:本研究以患者使用辅具目标等信息为特征建模。基于决策树算法建立的模型性能优于逻辑回归模型,具有实用价值。

关键词 辅具适配技术;决策树模型;逻辑回归模型;python语言

中图分类号:R496 **文献标识码:**B **文章编号:**1001-1242(2023)-08-1108-006

康复辅具(以下简称辅具)研究旨在维持或改善个人身体功能和独立性,或预防损害和继发的健康问题,它不仅包含促进个体福祉的辅助产品(包括任何外部的装置、器材、器具或软件),还包括相关技术^[1]。世界卫生组织与此相关定义为:辅助技术是指应用与辅助产品相关的有组织的知识和技能,包括系统和服务^[1-2]。因辅具使用者包括了残疾人、老年人和伤病人等各类功能障碍者群体,每名使用者在年龄、身体功能障碍、使用环境、支付能力以及个人活动目标等方面存在明显个体差异,所以对辅具的需求是个性化的。同时,辅具产品种类丰富,根据国家标准《康复辅助器具分类和术语》(GB/T16432—2016),康复辅助器具产品划分为12个主类、130个次类、794个支类^[3]。纷繁复杂的辅具种类和产品与使用者个性化的需求如何精准匹配,并确保使用效果与预期目标一致,所应用的技术可称为辅具适配技术。

辅具配置的传统模式是用户直接购买辅具或政府直接配发辅具,涉及较少的专业评估。随着近年来行业的发展,目前现有模式是先全面评估,再配置辅具,最后再进行效果评估的模式^[4]。但是现有模式依然面临着配置的辅具遭到弃用且满意度不高、专业服务力量不足的难题,且辅具配置经验几乎没有被充分利用^[5]。随着新一代信息技术的发展,人工智能、大数据在各领域得到了很好的应用。国内学者在辅具适配行业也进行了初步探索,主要有以下研究:董理权等

基于《国际功能、残疾和健康分类》,从全民健康覆盖健康服务体系要素出发,利用新一代信息技术,提出辅助器具评估与适配服务体系架构,并分析实现“互联网+”、智能化的路径^[6]。范秀琴等进行了基于案例推理—规则推理混合推理脊髓损伤智能辅具适配系统研究,应用临床的125例SCI案例建立案例库,通过K邻近和相似度检测算法并利用规则推理进行修正后推出适配的辅具组合^[7]。

现有模式是从辅具配置机构的角度进行的,其中患者常是被动接受的角色。实际上,由于专业机构和人员队伍有限等原因,对于需要使用辅具的用户,不是均有条件进入专业辅具配置机构。然而,在配置辅具过程中,患者本身往往已经具有一定的目标,也就是对于自己的疾病和身体功能障碍和使用辅具要解决什么问题是比较清楚的^[8]。为了解决辅具配置时一律依赖专业人员主导,而不能很好发挥用户的主观能动性的问题,本研究提出以患者使用辅具目标等信息为特征建立辅具适配模型,模型基于决策树和逻辑回归两个经典算法分别建立,性能评估后进行了优化,旨在实现辅具适配过程中用户自主自助及辅具适配的精准预测。本文将初步研究结果报告如下。

1 资料与方法

有监督学习指经过现有的部分输入信息以及输出信息

DOI:10.3969/j.issn.1001-1242.2023.08.014

*基金项目:中央级公益性科研院所基本科研业务费专项

1 国家康复辅具研究中心,北京市,100176; 2 北京市老年功能障碍康复辅助技术重点实验室; 3 民政部智能控制与康复技术重点实验室;

4 北京大学第三医院; 5 通讯作者

第一作者简介:张小雪,女,助理研究员; 收稿日期:2022-11-02

之间的相关关系,生成相应的函数关系表达式,最后将相应的输入信息映射到相应的输出信息上^[9]。监督学习中预测值为离散值的为分类学习任务^[10],本文研究的辅具适配问题符合此分类学习任务应用场景。决策树和逻辑回归是两个经典的分类器,应用广泛、运算代价低。其中,逻辑回归算法是非常简单、常见且具有代表性的广义线性模型^[11],是将回归算法用于分类问题的典型,该算法的思想是用 Sigmoid 函数将线性预

测界限于区间[0,1],并将其作为标签概率的预测。决策树算法基于特征对数据实例按照条件不断进行划分,最终达到分类或者回归的目的,当样本特征取值不具有数量意义时,决策树算法是处理类别特征的最佳选择^[12]。故本研究选择决策树和逻辑回归两种算法建立康复辅具适配模型,并对模型性能进行评估和优化。决策树与逻辑回归算法特点对比如表1。

模型的建立和优化主要包括收集辅具配置信息、数据预

表1 决策树与逻辑回归算法特点对比

序号	特点	决策树	逻辑回归
1	对整体结构的分析	对全局把握差	针对整个数据的拟合,对全局把握好
2	对局部结构的分析	采用分割的方法,能够深入数据局部	缺乏探查局部数据结构的内在机制
3	对极值是否敏感	对极值不敏感,表现好	对极值敏感,易受到影响
4	分析线性关系	较差	擅长
5	分析非线性关系	能力强	需要进行转换
6	输出结果	直观的决策规则,结果可解释,只能得到观察对象所在组别。	每个观察对象的概率

处理、提取标签及特征矩阵、切分训练集及测试集、实例化模型、提升模型性能六个步骤。其中实例化模型时使用决策树或逻辑回归算法。模型的建立和优化流程如图1所示。

1.1 获取初始数据

本研究通过在线随机调查问卷的方式收集辅具配置信息,调查问卷包括5个问题,分别是:“①辅具使用者年龄”。“②目前正在使用的辅具,曾经使用过的辅具以及需要的辅

具有什么”。“③使用辅具解决了什么问题”。“④上述辅具使用环境是否进行了专门改造?如有,请简单描述”。“⑤使用辅具的相关疾病是什么”。同时记录答题IP所在省市。上述信息简称:①患者年龄;②辅具名称;③辅具使用目的;④辅具环境;⑤患者伤病残情况;⑥患者省市。

收集到102份问卷,该102份问卷的患者均使用了辅具且解决了一定问题。患者年龄最小的2岁,最大的超过90岁。因一部分患者使用了不少一件辅具,102例患者共配置了133件辅具,涉及17个省份22个城市。

1.2 数据预处理

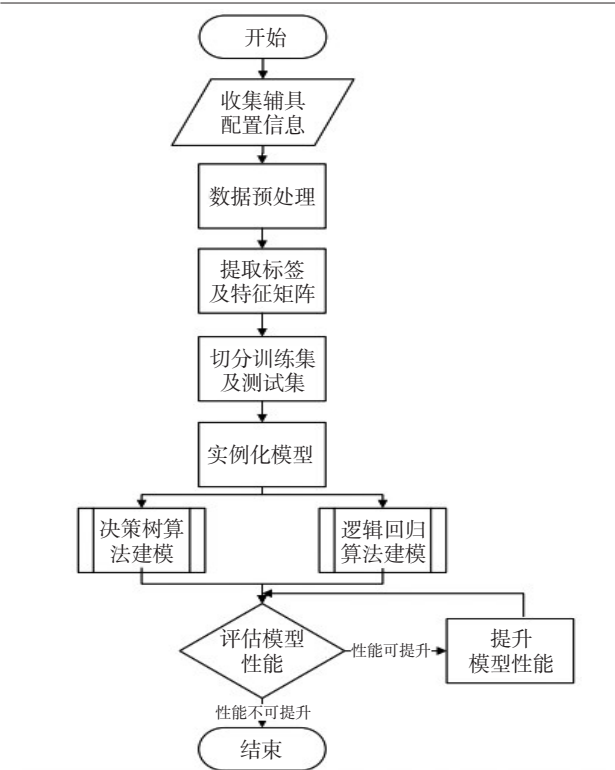
在数据预处理阶段,首先,检查数据是否存在缺失,对缺失值进行丢弃或补全。其次,对各项数据进行适当的数据类型转换和规范化。

数据类型转换:为后续建模做好准备提高数据质量。如:在问卷设计阶段,为方便调查者规范填写年龄段信息,问题1设计为患者年龄,而在数据处理时,用下列年龄段信息进行一个转换。年龄段划分按照1—6岁为童年、7—17岁为少年、18—40岁为青年、41—65岁为中年、66岁以上为老年。

数据规范化:主要是两个方面,一是在问卷设计阶段,为简化流程和避免过于强调术语,没有设辅具类别、产品及适应证的选项,因问卷面向正在使用辅具、需要辅具的人群及其家人和朋友,调查对象对辅具专用名词及相关知识背景的了解差别较大,所以收集上来的辅具、适应证等名称需要进行归类处理。另外,一是初始的一行数据对应着一名患者,一名患者对应着一款或多款辅具,需要对其拆分,拆分后补全信息,以一名患者及其配置的一款辅具作为一行数据。二是考虑省份城市对结果影响较小,不作为特征值。

该数据集数据存储于 excel 表格,规范化处理后共有133行5列,包含3个用户标签特征、1个改造环境特征和1个辅

图1 模型的建立和优化流程图



具产品。

特征变量解释、数据分布及缺失情况如表2所示。

1.3 实现语言

Python 语法简洁、清晰、语言通俗易懂且其中含有丰富和功能强大的类库,非常适于深入的数据处理和机器学习等领域。本研究运用python语言,以辅具配置信息为数据基础,基于决策树和逻辑回归算法预测目标变量概率。程序包括人机交互、数据读写、数据管理和性能评估四个子程序,在这里充分利用了python丰富的工具库和扩展性。主要调用了numpy、openpyxl、matplotlib和scikit-learn四个模块。其中调用的NumPy是Python的一种开源的科学技术库,主要用于处理维度数组和矩阵及运算^[13]。调用的openpyxl库用于读取和处理excel内的数据集数据。调用的matplotlib用于绘制散点图及决策边界使数据可视化。调用的scikit-learn用于对不同算法进行实验^[14]。

2 模型实现

2.1 建模准备

导入数据集,共133行,应用python字典使文本型变量编码转换为数值型变量。提取标签和特征矩阵,将年龄(age)、病种(diseases)、目的(purpose)、环境(environment)四列信息作为特征,产品(product)列作为标签。使用train_test_split切分训练集和测试集,随机数种子选555,测试集占比30%。

变量编码赋值如下:

age列变量编码赋值:老年=1,青年=2,中年=3,少年=4,童年=5;

diseases列变量编码赋值:截肢=1,老年人腰腿痛和不便=2,脑卒中=3,术后=4,儿麻后遗症=5,股骨头和髋关节置换

=6,骨折=7,膝损伤=8,腰腿疼痛=9,脊髓性肌萎缩症=10,听力障碍=11,关节炎=12,脊髓损伤=13,扁平足=14,截瘫=15,快乐木偶症=16,扭伤=17,视力障碍=18,腰椎间盘突出=19,孕晚期=20,骨瘤=21,脊柱侧凸=22,肩痛=23,颈椎不适=24,脑瘫=25,瘫痪=26,糖尿病=27,先天性下肢畸形=28,长期卧床=29,足跟外翻=30,腰肌劳损=31;

purpose列变量编码赋值:出行=1,助行=2,缓解疼痛=3,稳定关节=4,术后恢复=5,助听=6,防护=7,如厕=8,肢体矫形=9,洗澡=10,助视=11,代偿缺失上肢=12,翻身=13,固定肢体=14,进食=15,美观=16,平足=17,双腿不等长畸形=18,卧床照料=19;

environment列变量编码赋值:台阶=1,卫生间=2,连接下水=3,无障碍=4,无=5;

product列变量编码赋值:矫形器=1,轮椅=2,假肢=3,拐杖=4,助行器=5,助听器=6,坐垫=7,扶手=8,护理床=9,洗澡椅=10,眼镜=11,坐便椅=12,呼吸训练器=13,马桶增高器=14,膝部按摩仪=15,自助具=16。

2.2 建模并进行优化

2.2.1 决策树建模并进行模型优化:决策树算法的重点在于选择最优划分属性,使节点的“纯度”越来越高。这通过计算信息增益得到。

$$\text{信息熵} \text{Ent}(D) = - \sum_{k=1}^{|D|} p_k \log_2 p_k$$

$$\text{信息增益} \text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{D_v}{D} \text{Ent}(D_v)^{[10]}$$

基于决策树算法建模的计算逻辑^[15-16],见图2。

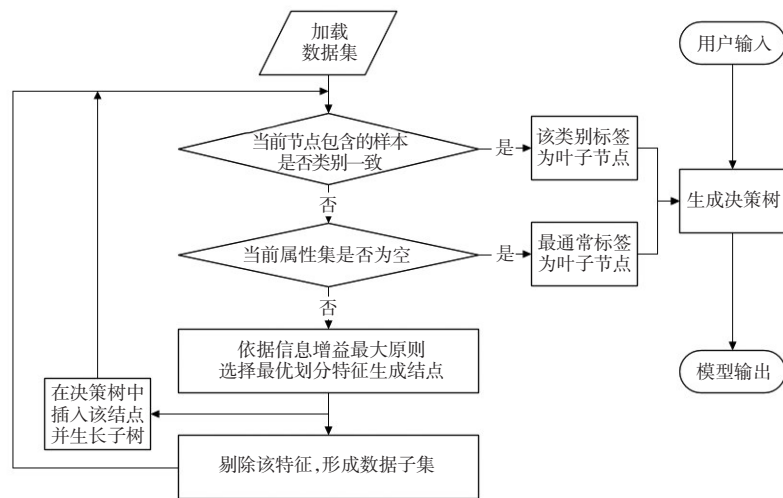
实例化模型即使用数据集进行训练,初始树深11。通过将预测结果和真实结果进行对比评估模型性能,得到决策树模型在训练集上准确率为92.4%,然而此时测试数据上的预测结果初始准确率为62.5%,模型过拟合严重。观察发现在

表2 数据集一览

变量	年龄	病种	目的	应用环境	产品
解释 选项及数据 分布	年龄段 老年:56 青年:37 中年:31 少年:5 童年:4	与使用辅具相关的伤病残情况 截肢:20 老年人腰腿(痛)不便:16 脑卒中:9 术后:8 儿麻后遗症、股骨头/髋关节置换、 骨折、膝损伤、腰腿疼痛:各7 脊髓性肌萎缩症:6 听力障碍:5 关节炎、脊髓损伤:各4 扁平足、截瘫:各3 快乐木偶症、扭伤、视力障碍、腰椎 间盘突出、孕晚期:各2 脊柱侧凸、肩痛、颈椎不适、脑瘫、瘫 痪、糖尿病、先天性下肢畸形、长期 卧床、足跟外翻、腰肌劳损:各1	辅具使用目的、拟解决的问题 出行:45 助行:24 缓解疼痛:22 稳定关节:9 术后恢复:6 助听:5 防护:4 如厕、肢体矫形:各3 洗澡、助视:各2 代偿缺失上肢、翻身、固定肢 体、进食、美观、平足、双腿不等 长畸形、卧床照料:各1	改造辅具应用环境 台阶:3 卫生间:3 连接下水:1 无障碍:1	辅具产品 矫形器:44 轮椅:30 假肢:16 拐杖:12 助行器:9 助听器:5 坐垫:3 扶手、护理床、洗澡 椅、眼镜、坐便椅:各2 呼吸训练器、马桶增 高器、膝部按摩仪、 自助具:各1
个数缺失	133 否	133 否	133 否	133 否	133 否

数据集中5个变量,缺失值检验结果如表2所示,全部为否。

图2 基于决策树算法建模的计算逻辑图



133行辅具配置信息中,因数据不平衡,当标签值为少数案例时,由于长尾效应存在,导致准确率降低。

为提升模型性能,一是复制少数案例,使相同案例数最少为6。二是编写for循环并可视化,调整主要参数。其中最大树深max_depth、叶子节点最小样本数min_sample_leaf、子树继续划分最少样本数min_samples_split,最大特征数

max_features和信息增益min_impurity_decrease 5个参数对模型的影响如图3所示。

2.2.2 逻辑回归建模并进行模型优化:逻辑回归算法的激活函数为 $g = \frac{1}{1 + e^{-x}}$, 模型假设是 $h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$, 代价函数为 $J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))]$, m为样本数量。

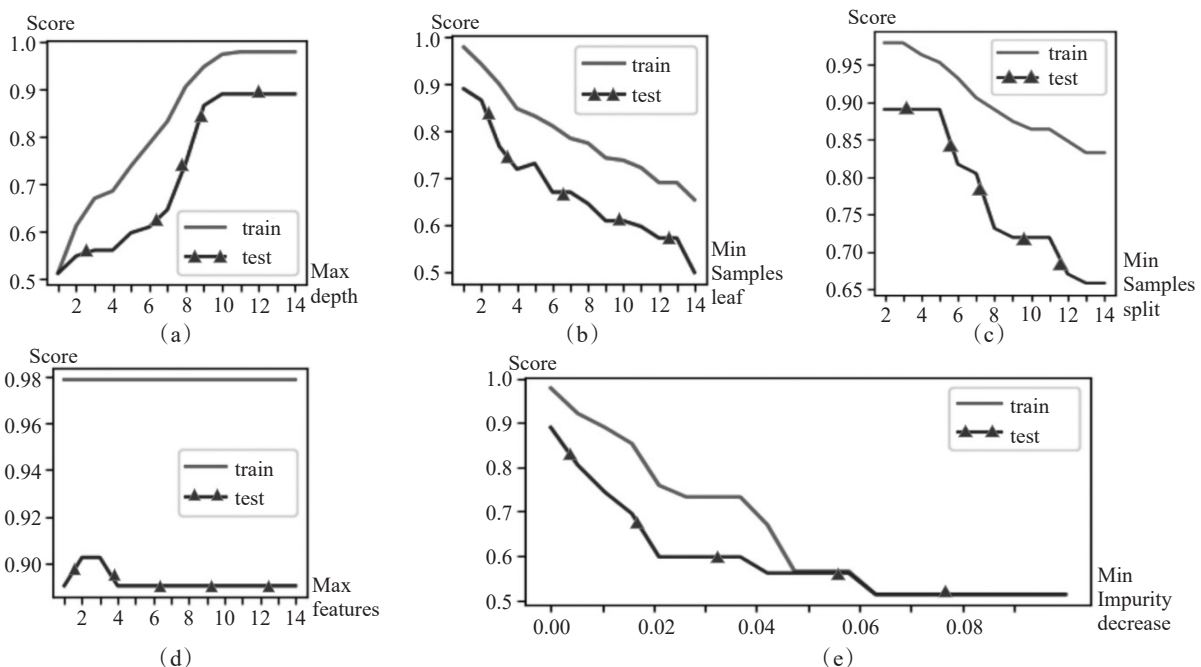
用梯度下降法计算使代价函数最小的参数,算法为每个参数减去代价函数对它的偏导数,同时更新所有参数。求导后得到:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

基于逻辑回归算法建模的计算逻辑,如图4所示。

实例化后的逻辑回归模型训练集准确率为55.4%,测试集准确率为47.5%,整体数据集准确率为53.0%,预测效果很差。为提升模型性能,一是平衡数据,减少长尾效应,二是实验性的对数据归一化处理,增加模型拟合能力,三是改变样本权重,四是调整模型训练过程的超参数,包括迭代次数max_iter、正则化参数C、学习率 α ,如图5所示。

图3 决策树主要参数对模型性能的影响



(a)最大树深max_depth;(b)叶子节点最小样本数min_samples_leaf;(c)子树继续划分最少样本数min_samples_split;(d)最大特征数max_features;(e)信息增益min_impurity_decrease

3 结果

在数据集增加至273行后,随机种子random_state=25,当决策树模型参数调至max_depth=10,Max_features=3,min_samples_leaf、min_samples_split和min_impurity_decrease分别默认1、2和0.0时,测试集准确率最高为91.5%,此时训练集上准确度96.9%,整体数据集上准确率95.2,准确度和过拟合情况得到改善,具有实用价值。决策树如图6所示,模型性能评估结果见表3。

基于逻辑回归算法的模型经过上述四个步骤的优化,模型性能提升效果并不理想,测试集最高准确率为63.4%,低于基于决策树算法建立的模型,不具有较高实用性^[17]。

辅具适配模型用于预测和分类辅具产品,而辅具产品品类丰富,这是预测结果不确定性的主要原因。并且一名患者可能配置多件辅具,这也降低了预测结果得分。此外,由于模型的训练数据来自调查问卷,相比临床实际数据,信息量不够丰富,随着该模型的应用,进而完善模型特征及加强数据集,模型性能还将提高。

辅具需求者和使用者对自身伤病残情况和辅具使用目的有着较为清晰认知,而对辅具产品了解甚少。在当今互联网购物盛行的大背景下,在此模型基础上继续开发具有图形界面的输入输出模块,以及连接企业的辅具产品专业配置服

图4 基于逻辑回归算法建模的计算逻辑图

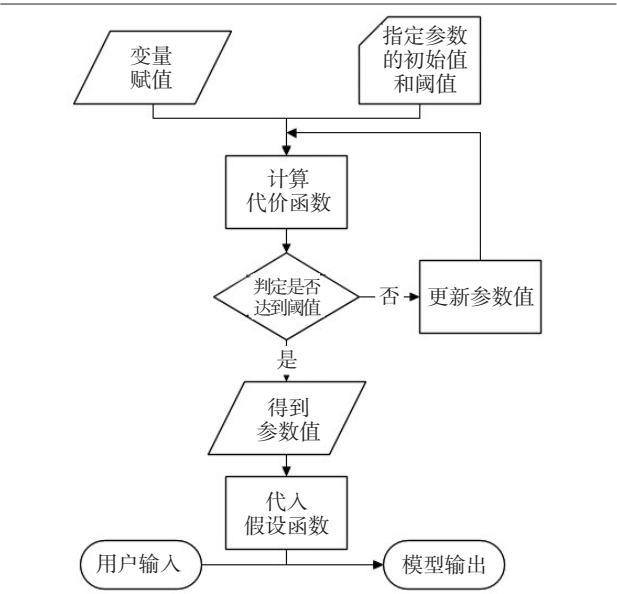
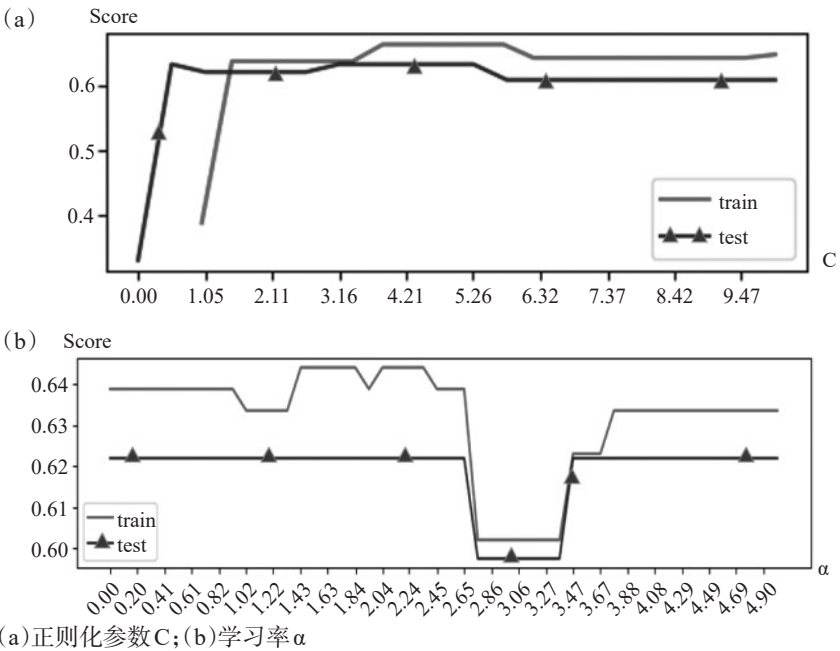


图5 逻辑回归主要参数对模型性能的影响



务沟通模块,即可为患者、机构和企业使用。通过模型获得辅具配置的有益信息可为患者自主自助决策提供参考,也可以为相关辅具配置机构提供参考。

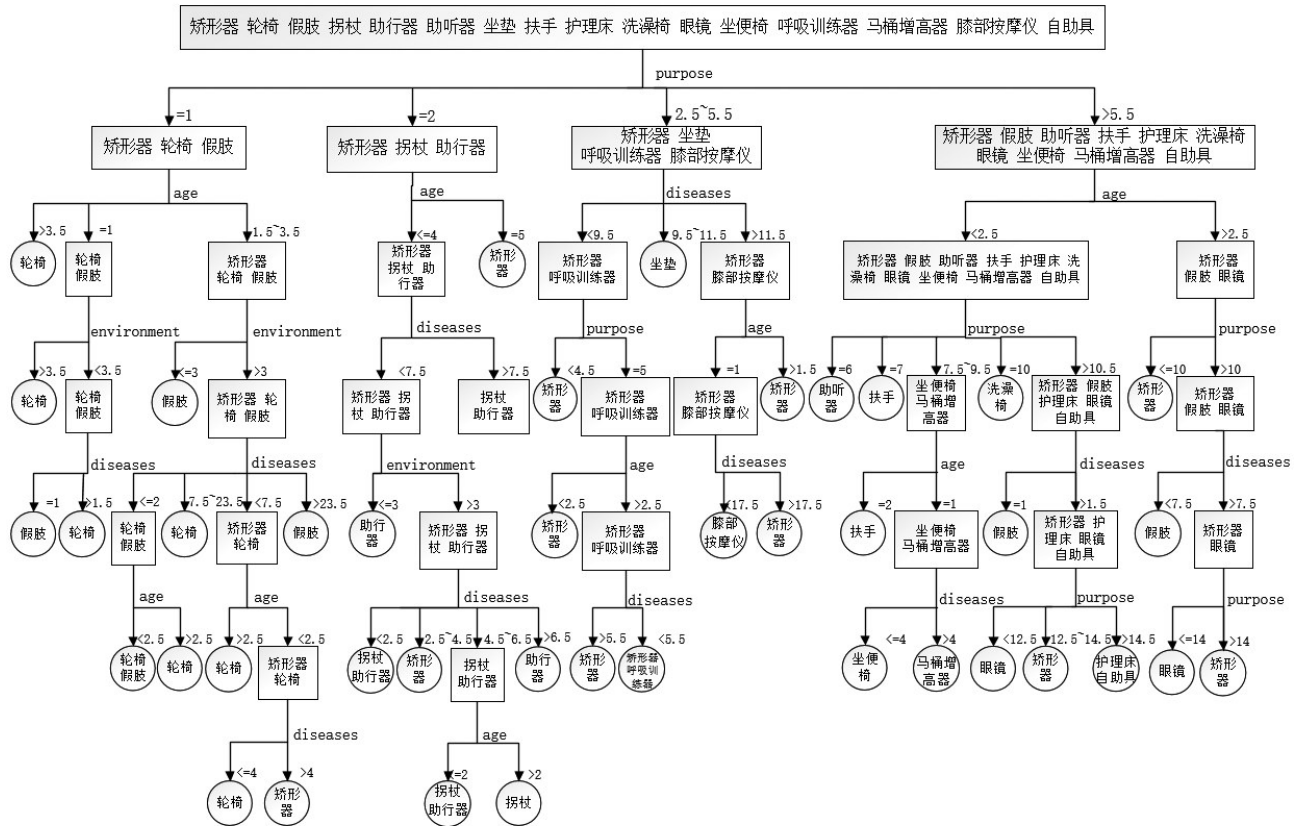
4 结论

本研究通过调查问卷的方式收集辅具配置信息,提取患者年龄段、与使用辅具相关的伤病残情况、辅具使用目的/拟解决的问题和改造辅具应用环境信息为特征,提取辅具产品信息为标签。首次基于决策树算法与逻辑回归算法分别建立了辅具适配模型,经过性能提升,决策树模型优于逻辑回

表3 决策树性能评估结果

	准确度	召回率	F1 分数	案例数
1	0.93	1.00	0.96	27
2	1.00	0.93	0.96	14
3	0.91	1.00	0.95	10
4	0.00	0.00	0.00	4
5	0.33	0.33	0.33	3
6	1.00	1.00	1.00	2
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	1
9	1.00	1.00	1.00	4
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	1.00	1.00	1.00	3
14	1.00	1.00	1.00	2
15	1.00	1.00	1.00	3
16	1.00	1.00	1.00	3
正确率			0.91	82
宏平均	0.88	0.88	0.88	82
微平均	0.89	0.91	0.90	82

图6 辅具适配决策树



归模型。

模型评估结果显示,基于决策树算法建立的辅具适配模型具有实用价值,可以达到实现辅具适配过程中用户自主自助及辅具适配的精准预测的效果,具有良好的应用前景。

参考文献

- [1] World Health Organization. Priority assistive products 2016 list [EB/OL]. [2022-10-31]. <https://www.who.int/publications/i/item/priority-assistive-products-list>.
- [2] World Health Organization & United Nations Children's Fund (UNICEF). Global report on assistive technology. World Health Organization [EB/OL]. [2022-10-31]. <https://apps.who.int/iris/handle/10665/354357>.
- [3] 全国残疾人康复和专用设备标准化技术委员会. 康复辅助器具分类和术语[S]. 北京: 中国标准出版社, 2016.
- [4] 郎倩倩, 郑俭. 国外常用辅助技术服务模式的比较[J]. 中国康复理论与实践, 2018, 24(2):224—228.
- [5] World Health Organization. Policy brief: access to assistive technology [EB/OL]. [2022-10-31] <https://www.who.int/publications/i/item/978-92-4-000504-4>.
- [6] 董理权, 吴小高. 构建辅助器具适配体系探讨[J]. 残疾人研究, 2014, 2:16—19.
- [7] 范秀琴, 喻洪流, 杨宇辉, 等. 基于案例推理—规则推理混

合推理的脊髓损伤智能辅具适配系统[J]. 中国康复医学杂志, 2022, 37(8):1084—1088.

- [8] Jesus TS, Bright F, Kayes N, et al. Person-centered rehabilitation: what exactly does it mean? Protocol for a scoping review with thematic analysis towards framing the concept and practice of personcentered rehabilitation[J]. BMJ Open, 2016, 6(7):e011959.
- [9] 邓建国, 张素兰, 张继福, 等. 监督学习中的损失函数及应用研究[J]. 大数据, 2020, 6(1):60—80.
- [10] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016:1—91.
- [11] 史春奇, 卜晶玮, 施智平. 机器学习:算法背后的理论与优化[M]. 北京: 清华大学出版社, 2019:0—26.
- [12] 王磊, 王晓东. 机器学习算法导论[M]. 北京: 清华大学出版社, 2019:1—123.
- [13] 吉田拓真, 尾原飒. Numpy 数据处理详解[M]. 北京: 中国水利水电出版社, 2021:53—166.
- [14] 加文·海克(著), 张浩然(译). scikit-learn 机器学习[M]. 北京: 人民邮电出版社, 2019:77—128.
- [15] 吕云翔. 机器学习基础[M]. 北京: 清华大学出版社, 2018:31—38.
- [16] 鲁伟. 公式推导与代码实现[M]. 北京: 人民邮电出版社, 2022:30—96.
- [17] 闫瑞平, 王习亮, 姚粉霞, 等. 决策树模型与 Logistic 回归分析模型识别高血压危险因素的效果比较[J]. 中华疾病控制杂志, 2022, 26(2):218—222.