浙江理工大学学报,第 51 卷,第 3 期,2024 年 5 月
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2024.03.012

# 基于 Relief-F 算法的心血管介入患者术后死亡风险预测

杨健斌[1],李　咏[1],夏淑东[2],齐鹏嘉[1],戴燕云[1],童基均[1]
(1.浙江理工大学信息科学与工程学院,杭州 310018;2.浙江大学医学院附属第四医院,浙江义乌 322000)

**摘　要**:针对心血管介入患者全周期病理数据普遍存在缺失、不连续、非结构化等问题,建立了心血管介入专病数据库,并采用基于 Relief-F 算法的预测方法,对心血管介入患者术后死亡风险进行预测。首先参照 HL7、CDISC 等国际心血管疾病统一标准对各数据源进行标准化处理,建立研究数据集,并对数据进行清洗和预处理;其次采用 Relief-F 算法对特征进行选择,最终保留 30 个特征变量;再次选择逻辑回归、支持向量机、随机森林等 3 种机器学习方法进行建模分析,并采用 10 折交叉验证方法对分类器进行训练;最后引入准确率等模型评价指标来评估各算法在数据集上的分类预测效果。实验结果表明:随机森林的分类效果在该研究数据集上的表现最佳,准确率达到 81.97%,精确率为 86.90%,召回率为 82.14%,$F_1$ 值为 0.8441。该研究提出的方法能够客观反映患者术后死亡风险,为心血管介入患者术后死亡风险预测提供了一种有效的解决方案。

**关键词**:心血管介入;术后死亡风险预测;Relief-F 算法;特征提取;机器学习;随机森林
**中图分类号**:TP391.4　　**文献标志码**:A　　**文章编号**:1673-3851 (2024) 05-0378-11
**引文格式**:杨健斌,李咏,夏淑东,等. 基于 Relief-F 算法的心血管介入患者术后死亡风险预测[J]. 浙江理工大学学报(自然科学),2024,51(3):378-388.
**Reference Format**: YANG Jianbin, LI Yong, XIA Shudong, et al. Prediction of postoperative death risk in patients with cardiovascular intervention based on the Relief-F algorithm[J]. Journal of Zhejiang Sci-Tech University,2024,51(3):378-388.

# Prediction of postoperative death risk in patients with cardiovascular intervention based on the Relief-F algorithm

*YANG Jianbin*[1], *LI Yong*[1], *XIA Shudong*[2], *QI Pengjia*[1], *DAI Yanyun*[1], *TONG Jijun*[1]
(1. School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2. The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu 322000, China)

**Abstract**: In view of the common problems such as missing, discontinuous and unstructured pathological data of patients with cardiovascular intervention throughout the whole cycle, a cardiovascular interventional disease database was established, and the prediction method based on the Relief-F algorithm was adopted to effectively predict the risk of postoperative death of patients with cardiovascular intervention. Firstly, all data sources were standardized according to HL7, CDISC and other international cardiovascular disease standards to obtain research data sets, and the data sets were cleaned and preprocessed. Secondly, the Relief-F algorithm was used to select the features, and 30 feature variables were retained in the end. Thirdly, logistic regression, support vector machine and random forest were selected for modeling and analysis, and the 10-fold cross-validation method was used to train the classifier. Finally, model evaluation indexes such as accuracy rate were introduced to evaluate the classification

收稿日期:2023－10－11　　网络出版日期:2024－03－13
基金项目:浙江省自然科学基金项目(LQ22F010006,LTGY23H170004)
作者简介:杨健斌(1999－　),男,浙江金华人,硕士研究生,主要从事信号处理和控制方面的研究。
通信作者:童基均,E-mail:jijuntong@zstu.edu.cn

prediction effect of each algorithm on the data set. The experimental results show that the classification effect of random forest has the best performance on the research data set, its accuracy rate is 81.97%, the accuracy rate is 86.90%, the recall rate is 82.14%, and the $F_1$ value is 0.8441. This study can objectively reflect the postoperative death risk of patients, and provides an effective solution for predicting the postoperative death risk of patients with cardiovascular intervention.

Key words: cardiovascular intervention; postoperative death risk prediction; Relief-F algorithm; feature extraction; machine learning; random forest

## 0 引　言

心血管介入治疗是指在医学影像设备引导下,经皮穿刺,将穿刺针、特制导管、导丝等精密器械引入体内血管,对心血管疾病进行微创诊断和治疗[1]。据国家心血管病中心发布的《中国心血管健康与疾病报告 2022》[2]显示,2022 年国内心血管病患者总人数已高达 3.3 亿,随着人口老龄化加速,心血管病患病率、死亡率仍在上升,疾病负担下降的拐点尚未出现。据《中国介入医学白皮书》2019 版[3]显示,心血管介入患者死亡率在众多介入治疗类型的患者中居于首位,较神经介入治疗、肝胆胰腺介入治疗、肿瘤介入治疗等其他类型平均高出 12.7%。因此,在众多心血管介入患者术后评价指标中,能够反映介入治疗质量、术后生存状况以及医疗干预效果的术后死亡风险,一直都是医生和患者关注的重要指标。然而,当前绝大数医院并未构建心血管介入专病数据库,患者全周期病理数据普遍存在缺失、不连续、非结构化等问题,导致医生对患者术后死亡风险的预测大多是基于医护人员的个人临床经验或相关统计学方法,预测准确率较低。随着医疗信息化建设的快速发展,上述方法已不适应当前的发展需求。因此,构建心血管介入专病数据库[4],通过信息技术帮助医生进行全周期记录,建立患者术后死亡风险预测模型,不仅有助于辅助医生发现患者数据的内在关联,对患者术后健康状况进行跟踪治疗,还能提醒医生对高死亡风险病人及时做出医疗干预,进而降低患者死亡率,具有重要的医疗研究意义和实际应用价值。

国内关于疾病专病数据库的建设起步较晚,直到 20 世纪 80 年代后期,中国逐步建立了医学注册系统[5-9],开始全面收集和管理疾病数据。在《"健康中国 2030"规划纲要》政策的指导下,国内的一些医院也开始着手建设针对特定病种的专病数据库。例如,2021 年,上海交通大学附属胸科医院袁骏毅等[10]以上海胸科医院冠心病专病数据为研究对象,基于临床数据中心的多源异构系统构建了冠心病专病数据库,为该院的冠心病临床科研提供了有力的数据支撑。2022 年,广东省第二人民医院的龙思哲等[11]借助双向语言模型从院内其他数据平台筛选出脑血管患者病理数据,构建了脑血管专病数据库,为建立脑血管科研学习平台提供数据支撑。同年,中国人民解放军总医院的赵前前等[12]以医院信息管理系统(Hospital information system, HIS)、实验室信息管理系统(Hospital information system, LIS)、电子病历(Electronic medical record, EMR)等临床业务信息系统中的数据和整合后的临床数据库为数据源,经抽取、转换、加载后形成疾病科研数据库,大大减轻了科研工作前期数据处理工作量,提高了科研效率。纵观国内专病数据库的构建与应用,由于肿瘤疾病、肝脏疾病等传统疾病类型有长期的数据支撑和科研基础,相关专病数据库的建设已经较为完善。但对于需要介入治疗的相关疾病,一方面由于介入治疗在我国起步较晚,医院在此方面的信息化建设尚未完善,导致很多患者病理数据难以统一;另一方面,介入治疗近年来受政策影响和需求驱动刚开始发展,医院对日益增长的数据尚未及时收集与整理,致使很多数据遗漏或残缺。因此,搭建一套心血管介入专病数据库,并结合信息技术对患者病情进行预测和危险因素分析,不仅可以帮助医生研究该类型疾病,还能辅助医生对高风险患者进行及时的医疗干预,降低死亡风险。

近年来,随着信息技术的快速发展,利用人工智能相关理论方法对患者病情进行预测和危险因素分析已成为当下医疗领域的研究热点。例如,Behera 等[13]采用支持向量机和改进的粒子群优化模型创建了一个混合模型,对患有心脏病和肝脏疾病的患者进行死亡风险因素分析。Theerthagiri 等[14]基于递归特征消除的梯度提升方法,通过评估患者的健康记录来避免心血管疾病的病发或降低心血管疾病的严重程度。Singh 等[15]基于支持向量机模型探讨了药物靶点预测降低背后的常见因素,进而预测肿

瘤的发生率。Islam 等[16]在监督学习环境中,选择了 12 种不同的机器学习分类器来对慢性肾脏病进行研究分析,得出在 XGBoost 分类器的准确率最高可达 0.983。Annamalai 等[17]借助基于最优拍卖机制的卷积神经网络对肺部疾病进行预测,发现所提出的方法可以从 X 射线图像中提取特征,并对肺部疾病进行准确预测。Sudha 等[18]使用混合 CNN-LSTM 模型对心脏病进行预测分析,并使用 $K$ 折交叉验证技术进行验证,最终混合模型的准确度达到 89%。Liang 等[19]提出了一种基于具有注意力机制的时序双向神经网络模型 tBNA-PR,选择患者的电子健康档案数据进行心力衰竭疾病预测和分层。Hao 等[20]利用多策略优化核极限学习机对心脏病和肝病的死亡率进行研究,在临床应用中取得良好的预测效果。路晓云等[21]基于机器学习预测算法对慢阻肺患者院后再入院风险进行预测,选择了 5 种预测模型进行对比分析,并基于 $K$ 均值聚类算法对患者再入院风险等级进行评估和分类。赵明诚等[22]以社区获得性肺炎患者为研究对象,基于长短期记忆网络对患者 30 d 的死亡率进行了预测,预测精确度达到 77.51%,能够帮助医生对社区获得性肺炎患者进行跟踪观察。

以上研究表明,建立专病数据库和预测模型,对患者病情进行预测是可行的。但是,目前关于心血管介入患者的相关预测研究仍是基于医生个人临床经验或者相关统计学方法,且研究所用的数据集仍停留在患者在院期间的记录数据,未覆盖到患者治疗全周期,具有一定的局限性。因此,本研究构建了心血管介入专病数据库,并针对难以区分重要特征和噪声特征等问题,采用改进的 Relief-F 算法选择预测特征变量;选择逻辑回归(Logistic regression,LR)、支持向量机(Support vector machine,SVM)、随机森林(Random forest,RF)等 3 种机器学习方法进行建模研究,并采用 10 折交叉验证方法对分类器进行训练;最终引入准确率等模型评价指标来评估各算法在本研究数据集上的分类预测效果,为心血管介入患者术后死亡风险预测提供一种有效的解决方案。

## 1 数据采集与处理

### 1.1 数据集

本研究的实验数据采集于浙江省某三甲医院,共获得 728 例心血管介入患者全周期病理数据,建立了心血管介入专病数据库。采集数据源包括医院的各业务系统(HIS、LIS、EMR 等)、各医疗表单(门诊病历、门诊医嘱、检验报告等)及医院的数据平台(人口学资料、就诊资料及随访资料)。具体信息包括患者基线信息(个人信息、病史信息、主诉和症状信息、生活方式信息等)、术前检查信息(临床评估、实验室检查、血管造影或介入性检查等)、术中手术信息(手术类型与时间、手术操作过程、手术过程中的观察和事件、手术结束情况和结论等)及术后随访信息(术后病情观察、生命体征监测、药物治疗等)。为了保证数据的一致性和互操作性,本研究在医生的建议下采用了中华医学会心血管病学分会牵头制定的《中国心血管病一级预防指南》[23]、HL7 卫生信息交换标准(Health level seven)和临床数据交换标准协会(Clinical data interchange standards consortium,CDISC)制定的全球临床研究的数据标准对各类数据源进行数据标准化处理。此外,鉴于各心血管介入患者之间存在较大差异,在医生的建议下还对心血管介入专病数据库中收录的患者数据制定了筛选标准,具体包括:a)患者年龄不低于 18 岁;b)患者在院建档入库,并完成了心血管介入手术治疗;c)患者术前、术中及院后随访信息连续无中断,不存在信息错误录入;d)各项病理数据结构化完整,且在院后随访生理体征数据记录完善;e)收集的患者个人信息符合隐私条例保护。

本研究与大多数研究类似,研究起点始于患者入院建立个人信息档案。然而,由于不同研究者根据其数据集实际情况或研究侧重点不同,所选择的研究终点也不尽相同,但均集中在术后到院后一年以内。虽然,在本研究数据集中存在个别患者入院后三年的随访信息,但是,由于时间跨度较长,部分记录信息出现缺漏或提前终止的现象,并不利于统计分析。而且,由于院后随访具有周期性,医院难以第一时间掌握到患者的死亡情况。因此,根据医生的建议,结合患者随访的实际情况,本研究选定术后六个月随访期间内发生死亡或未死亡为研究终点。基于以上标准和医生建议,本研究从心血管介入专病数据库中严格筛选了 638 例符合标准的患者全周期病例数据,共计 42746 条数据小项,其中:术后六个月死亡 41 例,未死亡 597 例。

### 1.2 数据清洗与处理

#### 1.2.1 缺失值和异常值处理

缺失值与异常值的处理是数据预处理过程中的关键步骤,具体处理方法需要根据实际情况进行确定。常见的缺失值处理方法包括:删除缺失值较多

的特征、采用众数填充文本类数据、采用均值或中位数填充连续型数据以及利用线性或指数插值法填充缺失值[24]。合理处理缺失值可以保持数据的连续性,并减少噪音特征对模型的干扰,从而提高计算效率。针对异常值,可以选择直接删除或将其视为缺失值处理,或者采用平均值修正等方法。虽然选择直接删除会减少数据量,但是可以有效避免异常值对模型的干扰。

对于少量缺失率较高的数据本研究选择直接删除,如基线信息中缺失率达 87.4%的左室后壁厚度。对于一些记录信息的缺失本研究采用众数来填充缺失值,如患者住院方式变量中,"门诊"方式占总样本的比例达到 96.23%,所以对于此类缺失值可直接将其填充为"门诊"。对于连续型变量,如主动脉舒张压、白细胞、血红蛋白等,其数据连续且完整性完好,缺失率极低,则直接使用该变量的平均值来填充。对于一些变量的极值或者离群值,当数据量较少时本研究选择直接删除,较多时则选择离异常值最近的正常范围来填充数据。为了能够快速、直观地了解数据集的完整性,本研究采用缺失值可视化工具库 Missingno 得到了特征缺失值矩阵图,结果如图 1 所示;图中左侧 1～638 为病人数,右侧为数据热力值,白色部分代表数据缺失值所在位置,且白色部分越多代表缺失情况越严重。

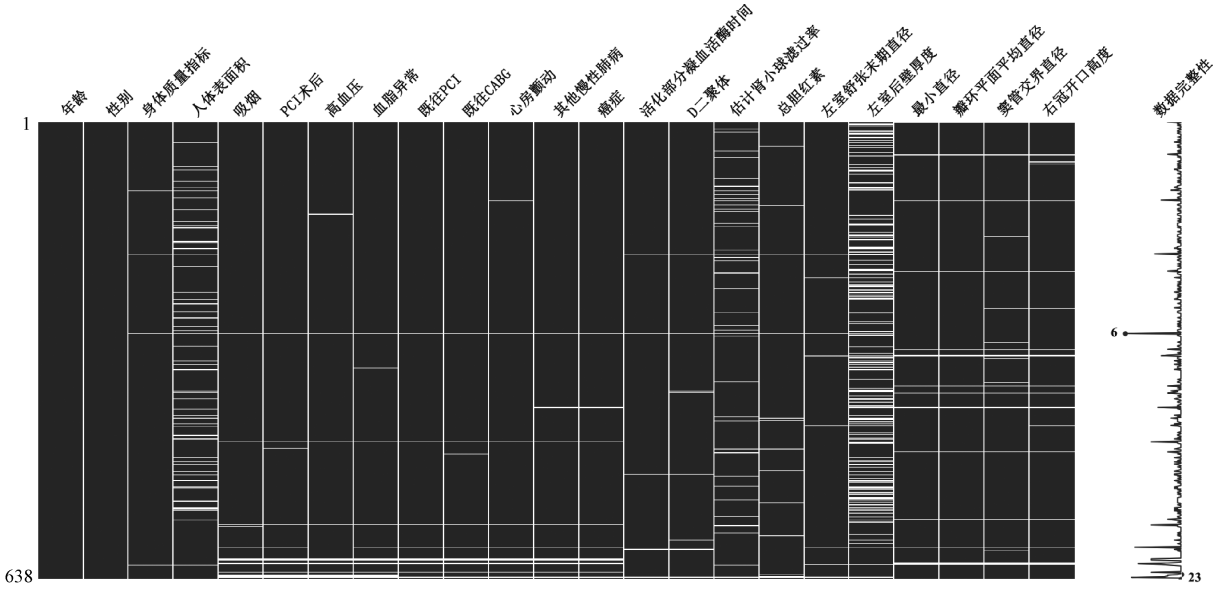


**图 1　特征缺失值可视化矩阵图**

### 1.2.2　特征向量化

由于大多数机器学习算法只能处理数值型数据,不能处理文字。所以,在训练和预测机器学习模型时,需要把这些特征进行编码,将字符型数据转换成数值型,这个过程可以让计算机更好地处理数据。合适的编码和量化方法,不仅可以提高模型的准确性和效率,还可以避免数据丢失或歪曲的情况。

本研究采用独热编码(One-Hot 编码)将离散型数据转换为数值型数据。例如,对于二类问题均采用 01 编码方式:男性患者编码为 1,女性患者编码为 0;患有糖尿病编码为 1,未患有编码为 0;术中出血编码为 1,未出血编码为 0;吸烟编码为 1,不吸烟编码为 0 等。此外,病例中的等级评分按英文等级划分。例如,患者的日常生活活动能力测定评分,可分 A～G 6 个等级来衡量。主要分类变量数值映射表如表 1 所示。

**表 1　本研究中离散型数据与数值型数据之间的映射关系**

| 序号 | 变量名 | 英文缩写 | 赋值 | 说明 |
|---|---|---|---|---|
| 1 | 性别 | Sex | 0/1 | 女/男 |
| 2 | 入院方式 | Admission type | 0/1/2 | 急诊/门诊/转院 |
| 3 | 麻醉方式 | Anesthetic type | 0/1 | 局麻/全麻 |
| 4 | 吸烟 | Smoker | 0/1 | 有/无 |
| 5 | 血脂异常 | Dyslipidemia | 0/1 | 无/有 |
| 6 | 再入院 | Re_admission | 0/1 | 无/有 |
| 7 | 心功能分级 | Katz_ADL | Ⅰ/Ⅱ/Ⅲ/Ⅳ | 心功能Ⅰ/Ⅱ/Ⅲ/Ⅳ级 |

1.2.3 合并症特征处理

通过分析心血管介入患者基线信息可以发现,大部分患者患有多种慢性疾病或基础疾病,如糖尿病、高血压等。如果对这些病症不做合并处理,直接作为待预测的特征使用,可能会造成数据维度过高和部分特征稀疏等问题,影响研究的科学性和准确性。查尔森合并症指数(Charlson comorbidity index,CCI)是由查尔森等在 1987 年提出的是一种评估患者合并症负担的指数。近年来,CCI 在临床实践中被广泛应用于预测患者的死亡风险、评估治疗效果、制定护理计划和手术决策等方面[25]。本研究按照 CCI 评价标准[25]对每名心血管介入患者是否患有对应合并症进行评分,并逐项相加,具体查尔森合并症指数评分表见表 2。

**表 2 查尔森合并症指数评分表**

| 序号 | 疾病名称 | 评分 | 序号 | 疾病名称 | 评分 |
|---|---|---|---|---|---|
| 1 | 老年痴呆 | 1 | 9 | 任何白血病 | 2 |
| 2 | 心肌梗死 | 1 | 10 | 偏瘫 | 2 |
| 3 | 脑血管疾病 | 1 | 11 | 中度至重度慢性肾脏病 | 2 |
| 4 | 慢性肺部疾病 | 1 | 12 | 淋巴癌或实体癌 | 2 |
| 5 | 充血性心衰 | 1 | 13 | 肝脏疾病 | 3 |
| 6 | 糖尿病 | 1 | 14 | 恶性肿瘤 | 6 |
| 7 | 溃疡病 | 1 | 15 | 艾滋病 | 6 |
| 8 | 结缔组织病 | 1 | | | |

## 2 实验与分析

### 2.1 基于 Relief-F 算法的特征选择

经过预处理后,数据集中仍包含冗余或无关变量,如果将这些特征变量直接输入分类器进行训练学习,则会对模型的训练结果造成较大影响。例如,术中信息中的“是否预扩张”和“预扩张次数”只需保留后者即可。对预处理后的数据进行降维处理,筛选出主要特征,去除冗余特征,减少数据噪声,降低模型学习难度,可以有效提高算法的准确度。常用的特征选择方法有过滤法、包裹法和嵌入法[26]。其中:过滤法是根据特征与目标变量之间的统计关系进行选择,筛选出与预测变量相关度较高的特征;其优点在于运算速度快,不需要进行模型训练,但无法考虑特征之间的关系。包裹法是通过穷举搜索或启发式搜索来选择最佳特征子集;其优点是考虑了特征之间的关系,但运算速度慢,可能出现过拟合等现象。嵌入法则是将特征选择作为学习模型训练的一部分,通过优化算法来选择最佳特征子集;其优点在于减少了特征选择和模型训练的时间,但可能会丢失有用特征。

与传统方法相比,相关特征(Relevant features,Relief)算法[27]是通过评估特征之间的关联程度来确定特征的重要性,不仅简单易实现、不依赖数据分布假设、适用于离散和连续特征,还对噪声和冗余特征具有一定的鲁棒性。但是,Relief 算法最初局限于解决二分类问题,无法有效去除高维数据的冗余特征。所以,本研究采用改进的 Relief-F 算法[28]来进行预测特征变量的选择。相对于传统的 Relief 算法,Relief-F 算法在计算特征权重和评估分数时引入了权重方差,能够更准确地估计特征的重要性,并对特征权重的稳定性进行评估,从而更好地区分重要特征和噪声特征。

Relief-F 算法每次从训练样本集中随机取出一个样本 $R$,然后从与 $R$ 同类的样本集中找出 $R$ 的 $k$ 个近邻样本(Near Hits),从每个 $R$ 的不同类的样本集中均找出 $k$ 个近邻样本(Near Misses),最后更新每个特征的权重;重复抽取 $m$ 次,其中单次权重迭代可用式(1)表示:

$$W(A)=W(A)-\sum_{j=1}^{k}diff(A,R,H_j)/(mk)+\sum_{C\notin class(R)}\left(\frac{p(C)}{1-p(class(R))}\sum_{j=1}^{k}diff(A,R,M_j(C))\right)/(mk) \tag{1}$$

其中:$A$ 为特征变量的个数;$H_j$ 为样本 $R$ 的 $k$ 个最近邻同类点;$diff(A,R,H_j)$为在特征 $A$ 上样本 $R$ 和 $H_j$ 的差;$M_j(C)$为异类样本点;$class(R)$为样本 $R$ 的类别;$p$ 为概率。$diff(A,R_1,R_2)$用式(2)算:

$$diff(A,R_1,R_2)=\begin{cases}\dfrac{|R_1[A]-R_2[A]|}{\max(A)-\min(A)},A\text{ 为连续值};\\ 0,A\text{ 为离散值且 }R_1[A]=R_2[A];\\ 1,A\text{ 为离散值且 }R_1[A]\neq R_2[A]\end{cases} \tag{2}$$

本研究在 Python3.9 环境下进行,通过 Relief-F 算法进行特征选择。以患者基线信息为例,特征权重曲线如图 2 所示,本研究选择了特征权重前 10 的特征变量作为患者基线信息。

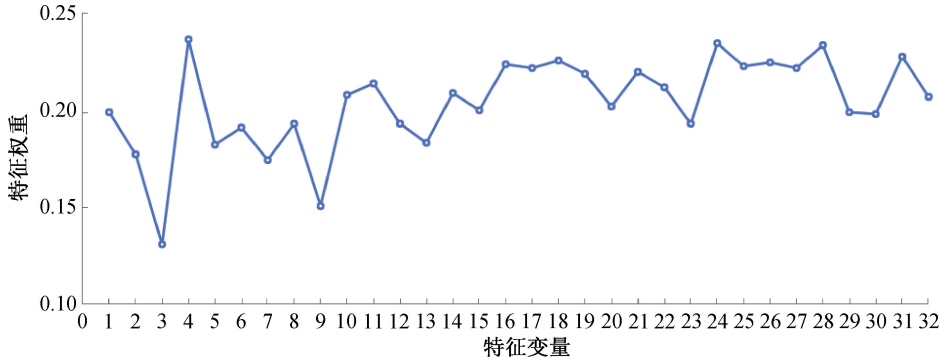


**图 2　患者基线信息特征权重曲线**

经过数据预处理和 Relief-F 算法筛选,并结合医生建议,本研究最终保留了 30 个特征变量,其中:患者基线信息保留了 10 个特征变量,术前检查信息保留了 7 个特征变量,术中手术信息保留了 5 个特征变量,术后随访信息保留了 8 个特征变量。患者特征变量说明见表 3。

## 2.2　模型构建与训练

本研究模型训练流程如图 3 所示。首先,将数据集按照训练集(70%)和测试集(30%)进行划分,其中训练集共 13398 条数据,测试集 5742 条数据。然后,先对训练集进行不同模型训练,为了评估不同模型在训练集上的表现,在训练过程中使用 10 折交叉验证方法和网格搜索方法来进行参数选择和参数优化,得到该模型下的最佳参数和训练效果。最后,利用测试集进行验证,得到不同模型的预测效果。基于处理后的数据集,本研究选择了 3 种在疾病预测研究中表现较佳的机器学习方法进行建模研究。

### 2.2.1　逻辑回归

LR 是一种用于分析因变量与一个或多个自变量之间的关系的统计方法。在疾病自动诊断领域,逻辑回归常被用于探讨引发某种疾病的危险因素,并基于这些因素预测疾病的发生概率。在本研究中,每个心血管介入患者 $y_i$ 术后六个月内发生死亡(记为 1)的概率可用式(3)表示:

$$P(y_i=1)=\frac{\exp(\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\varepsilon)}{1+\exp(\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\varepsilon)} \tag{3}$$

其中:$\beta_0\sim\beta_k$ 表示模型的估计参数,$x_1\sim x_k$ 表示模型的变量,$\varepsilon$ 为随机误差。

同时有:

$$\ln\left(\frac{P(y_i=1)}{1-P(y_i=1)}\right)=\beta_0+\beta_1\cdot x_{\text{Age}}+\beta_2\cdot x_{\text{Sex}}+\beta_3\cdot x_{\text{BMI}}+\cdots+\varepsilon \tag{4}$$

其中:$x_{\text{Age}}$、$x_{\text{Sex}}$ 及 $x_{\text{BMI}}$ 均为本研究中的特征变量。

本研究选择 Sklearn 库中的 Logistic regression 包来构建逻辑回归的分类器。在模型训练过程中,需要根据训练效果对模型的配置参数进行选择,采用网格搜索算法对 LR 的主要参数调优,LR 参数选择情况见表 4。

### 2.2.2　支持向量机

SVM 是一种通过最大间隔化思想来实现分类或回归问题的方法。对于二分类问题,SVM 的目标是找到一个超平面,使得两个类别的数据点都能够被正确地分开。如果数据集中有多个超平面可以完全分开两个类别,那么 SVM 会选择具有最大间隔的超平面作为最终分类器。本研究使用的是单核 RBF 的 SVM 模型。同时,为了避免过拟合现象,需要对 SVM 的参数进行优化选择,其中惩罚系数 $C$ 和控制高斯核宽度参数 gamma 的选择可以通过网格搜索来实现。本研究选择 Sklearn 库中的 SVC 包来构建 SVM 的分类器。采用网格搜索算法对 SVM 的主要参数选取。SVM 参数选择情况见表 5。

### 2.2.3　随机森林

RF 作为一种经典的集成学习方法,具有训练速度快、实现简单和泛化性能强等优点。RF 参数一般分为两类,一类是模型框架参数,如元分类器的个数等;另一类是决策树的参数,如树的深度等。本研究调用 Sklearn 库中的 Random Forest Classifier 包来构建随机森林的分类器。使用网格搜索算法对随机森林模型参数进行优化选择。随机森林 n_estimators 参数与模型准确率之间的关系如图 4 所示。当参数 n_estimators 在[20, 25]之间时,模型的准确率在[0.80, 0.82]范围内,当 n_estimators

超过 30 时,准确率趋于稳定。因此,选择 n_estimators 为 21,此时模型的准确率最高。RF 参数选择情况见表 6。

**表 3 本研究最终选取的 30 个特征变量说明**

| 信息类别 | 特征变量名称 | 死亡患者特征信息($N$=41) | 非死亡患者特征信息($N$=597) | 特征变量说明 |
|---|---|---|---|---|
| 患者基线信息 | 年龄 | 平均值:78.21 岁;标准差:5.17 岁 | 平均值:75.83 岁;标准差:6.95 岁 | 患者年龄分布 |
| | 性别 | 女:34.14%;男:65.86% | 女:41.23%;男:58.77% | 患者性别及其人数百分比 |
| | 身体质量指标 | 平均值:27.41 kg/m$^2$;标准差:1.55 kg/m$^2$ | 平均值:22.70 kg/m$^2$;标准差:3.49 kg/m$^2$ | 正常范围:18.5～23.9 kg/m$^2$ |
| | 吸烟 | 是:36.59%;否:63.41% | 是:16.14%;否:83.86% | 患者是否长期吸烟及其人数百分比 |
| | 查尔森合并症指数 | 平均值:4.75;标准差:2.32 | 平均值:2.71;标准差:1.54 | 评估患者合并症负担的指数 |
| | 高血压 | 是:58.53%;否:41.47% | 是:53.29%;否:46.71% | 患者是否患有高血压及其人数百分比 |
| | 心房颤动 | 是:29.26%;否:70.74% | 是:17.58%;否:82.42% | 患者血脂异常情况及其人数百分比 |
| | PCI 术后 | 是:7.31%;否:92.69% | 是:11.89%;否:88.11% | 患者前期是否接受经皮冠状动脉介入治疗及其人数百分比 |
| | 瓣环平面平均直径 | 平均值:23.97 mm;标准差:1.74 mm | 平均值:24.49 mm;标准差:2.51 mm | 正常范围小于 35 mm |
| | 二尖瓣反流 | 平均值:1.94;标准差:0.27 | 平均值:1.12;标准差:0.73 | 0,无;1,轻度;2,中度;3,重度 |
| 术前检查信息 | 白细胞计数 | 平均值:7.87/L;标准差:2.94/L | 平均值:7.41/L;标准差:3.67/L | 成人正常范围:3.5～9.5×10$^9$/L |
| | D-二聚体 | 平均值:0.2941 mg/L;标准差:0.2168 mg/L | 平均值:0.2647 mg/L;标准差:0.2466 mg/L | 成人正常范围:0～0.256 mg/L |
| | 血红蛋白 | 平均值:87.32 g/L;标准差:37.20 g/L | 平均值:108.85 g/L;标准差:71.77 g/L | 成人正常范围:110～160 g/L |
| | 血小板 | 平均值:214.72/L;标准差:101.32/L | 平均值:160.26/L;标准差:73.19/L | 成人正常范围:100～300 10$^9$/L |
| | PT 凝血酶原时间 | 平均值:16.31 s;标准差:6.27 s | 平均值:15.19 s;标准差:5.09 s | 成人正常范围:11～13 s |
| | INR 国际标准化比值 | 平均值:1.57;标准差:0.26 | 平均值:1.23;标准差:0.70 | 成人正常范围:0.8～1.5 |
| | ProBNP 脑钠肽水平 | 平均值:3875.2 pg/mL;标准差:4257.3 pg/mL | 平均值:2976.5 pg/mL;标准差:1939.7 pg/mL | 成人正常范围:50 岁及以下<450 pg/mL;50 岁以上<900 pg/mL |
| 术中手术信息 | 后扩展次数 | 平均值:0.672 次;标准差:0.049 次 | 平均值:0.581 次;标准差:0.025 次 | 患者术中进行后扩展次数 |
| | 瓣膜尺寸 | 平均值:27.34 mm;标准差:5.37 mm | 平均值:28.07 mm;标准差:7.41 mm | 患者术中植入心脏瓣膜的尺寸 |
| | 入路方式 | 0:37 人;1:4 人 | 0:573 人;1:24 人 | 患者手术入路方式。0,股动脉人数;1,心尖人数 |
| | 麻醉方式 | 1:38 人;2:3 人 | 1:580 人;2:17 人 | 患者手术麻醉方式。1,局麻人数;2,全麻人数 |
| | 鞘植入时长 | 平均值:147.71 min;标准差:43.61 min | 平均值:106.24 min;标准差:21.25 min | 患者手术过程中鞘植入时长 |

表 3(续)

| 信息类别 | 特征变量名称 | 死亡患者特征信息($N$=41) | 非死亡患者特征信息($N$=597) | 特征变量说明 |
|---|---|---|---|---|
| 术后随访信息 | 肌酐 | 平均值：61.25 μmol/L；标准差：6.35 μmol/L | 平均值：89.40 μmol/L；标准差：12.74 μmol/L | 成人正常范围：50～150 μmol/L |
| | 总胆红素 | 平均值：14.59 μmol/L；标准差：5.38 μmol/L | 平均值：9.92 μmol/L；标准差：3.74 μmol/L | 成人正常范围：1.71～21 μmol/L |
| | 白蛋白 | 平均值：42.31 g/L；标准差：2.59 g/L | 平均值：44.29 g/L；标准差：10.37 g/L | 成人正常范围：35～50 g/L |
| | 左房大小 | 平均值：4.23 cm；标准差：0.84 cm | 平均值：4.07 cm；标准差：0.61 cm | 成人正常范围：2.7～4.0 cm |
| | ALT 谷丙转氨酶 | 平均值：26.71 U/L；标准差：9.38 U/L | 平均值：21.23 U/L；标准差：18.84 U/L | 成人正常范围：0～50 U/L |
| | 左室舒张末期直径 | 平均值：41.25 mm；标准差：7.12 mm | 平均值：47.57 mm；标准差：4.29 mm | 成人正常范围：35～55 mm |
| | KATZ 评分 | A：2；B：8；C：13；D：15；E：2；F：1；G：0 | A：154；B：242；C：138；D：43；E：13；F：7；G：0 | 反映患者日常生活能力，分A～G 七个等级，等级越低表示生活能力越差，依赖他人程度越高 |
| | 华发林 | 0：70.74%；1：29.26% | 0：83.59%；1：16.41% | 患者是否长期服用华法林及其人数百分比。0，未服用人数百分比；1，服用人数百分比 |

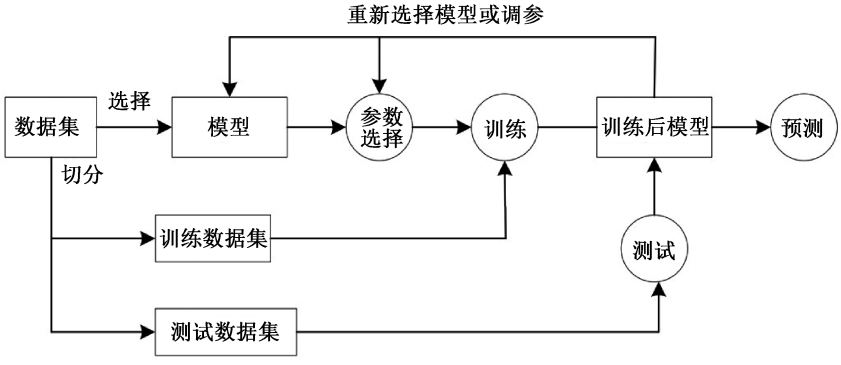


图 3　本研究模型训练流程图

表 4　本研究中 LR 的参数选择情况

| 参数名称 | 配置参数值 | 参数说明 |
|---|---|---|
| random_state | 22 | 随机种子 |
| penalty | L2 | 正则化参数 |
| solver | liblinear | 优化算法 |
| max_iter | 100 | 迭代次数 |
| $C$ | 0.1 | 正则化强度的倒数 |

表 5　本研究中 SVM 的参数选择情况

| 参数名称 | 配置参数值 | 参数说明 |
|---|---|---|
| $C$ | 0.7 | 惩罚系数 |
| kernel | RBF | 核函数 |
| gamma | 0.1 | 宽度参数 |
| class_weight | balanced | 类别权重 |

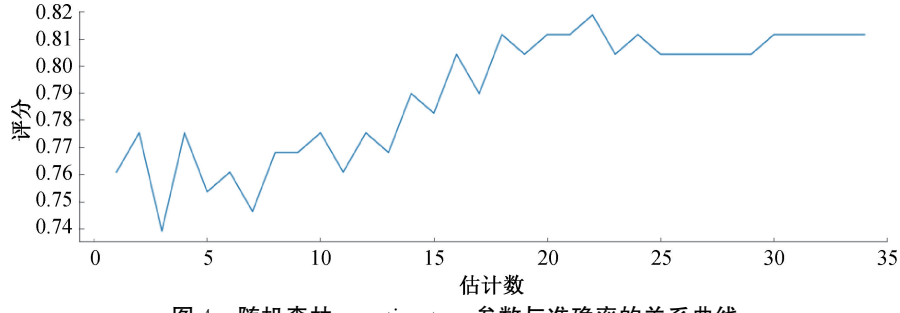


图 4　随机森林 n_estimators 参数与准确率的关系曲线

表 6 本研究中 RF 的参数选择情况

| 参数名称 | 配置参数值 | 参数说明 |
|---|---|---|
| bootstrap | True | 抽样方法 |
| max_depth | 50 | 决策树的最大深度 |
| min_samples_leaf | 5 | 叶节点所需最小样数 |
| max_features | 9 | 决策树最大特征数 |
| random_state | 37 | 随机数种子 |
| n_estimators | 21 | 决策树的数量 |
| class_weight | balanced | 类别权重 |

## 2.3 结果分析与讨论

算法模型的评估是验证模型构建效果的重要指标,用于确定模型的性能和可靠性。心血管介入患者术后死亡风险的预测研究实质上可映射成一种二分类问题进行研究,即将患者院后六个月随访期内发生死亡与否作为预测目标。对于二分类问题,可将预测样本划分为真正例(True positives)、假正例(False positives)、真反例(True negatives)、假反例(False negatives),分类结果可用混淆矩阵表示,混淆矩阵见表 7,其中:$T_P$ 表示真正例数;$F_P$ 表示假正例数;$T_N$ 表示真反例数;$F_N$ 表示假反例数。

表 7 二分类问题分类结果的混淆矩阵

| 真实情况 | 预测结果正例 | 预测结果反例 |
|---|---|---|
| 正例 | $T_P$(真正例) | $F_N$(假反例) |
| 反例 | $F_P$(假正例) | $T_N$(真反例) |

本研究从准确率(Accuracy)、精确率(Precision)、召回率(Recall rate)、$F_1$($F_1$ score)、AUC(Area under curve)等 5 个评价指标对建立的预测模型进行评估分析,具体公式如式(5)—(8)所示:

$$a_{CC}=\frac{T_P+T_N}{T_P+F_P+T_N+F_N} \tag{5}$$

$$P=\frac{T_P}{T_P+F_P} \tag{6}$$

$$R_R=\frac{T_P}{T_P+F_N} \tag{7}$$

$$F_1=\frac{2\cdot P\cdot R_R}{P+R_R} \tag{8}$$

其中:$a_{CC}$ 表示准确率;$P$ 表示精确率;$R_R$ 表示召回率;$a_{UC}$ 表示 AUC 值。

受试者工作特征曲线(Receiver operating characteristic curve,ROC)曲线[29]表示在不同的分类阈值下,真阳性率与假阳性率之间的关系,其中真阳率为 ROC 曲线的 $y$ 轴,假阳率为 $x$ 轴。

$a_{UC}$ 可以理解为 ROC 曲线下的面积,取值范围在 0 到 1 之间,用式(9)计算:

$$a_{UC}=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)(y_i+y_{i+1}) \tag{9}$$

具体来说,$a_{UC}=1$ 表示模型完美地对正负样本进行了区分,$a_{UC}=0.5$ 表示模型的预测性能与随机猜测相当,而 $a_{UC}<0.5$ 则表示模型的预测性能不佳。RF、SVM 和 LR 三种机器学习算法模型的各项评价指标对比见表 8。

表 8 RF、SVM 和 LR 的评价指标

| 模型名称 | 评价指标 | | | |
|---|---|---|---|---|
| | $a_{UC}$ | $P$ | $R_R$ | $F_1$ |
| LR | 0.7498 | 0.8355 | 0.7087 | 0.7668 |
| SVM | 0.7780 | 0.8411 | 0.7798 | 0.7612 |
| RF | 0.8197 | 0.8690 | 0.8214 | 0.8441 |

从表 8 中可得出,针对本研究的数据集而言,RF 在 3 种传统机器学习预测算法中表现最佳,其准确率达到 81.97%,精确率为 86.90%,召回率为 82.14%,$F_1$ 值为 0.8441,而 LR 的准确度最低,未能达到 75%。其主要原因在于,LR 是 3 种机器算法中唯一使用线性模型的,而 SVM 和 RF 均使用非线性模型,后者可以利用更复杂的数据,从而提高过采样数据的准确性。另外,由于 RF 采用了决策树的集成方式,每棵决策树都可以学习不同的特征和决策规则,并根据所有决策树的预测结果进行投票。所以,RF 能够更有效地捕捉特征之间的非线性关系,相比于 SVM 在本研究数据集上的表现效果更佳。

ROC 曲线下的面积 $a_{UC}$ 从大到小依次是 RF(0.8292)、SVM(0.7743)和 LR(0.7576),3 种机器学习算法的 ROC 曲线如图 5 所示。

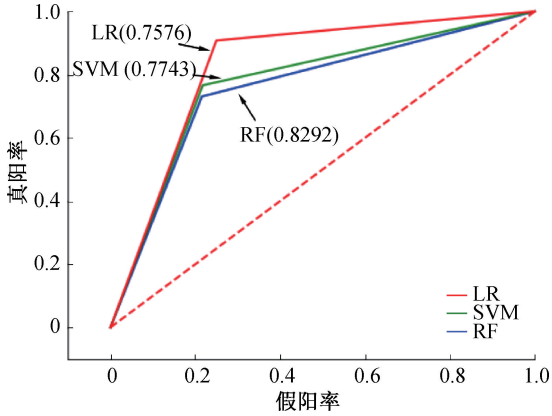


图 5 RF、SVM 和 LR 的 ROC 曲线

## 3 结 论

本研究建立了心血管介入患者专病数据库,并利用改进的 Relief-F 算法对心血管介入患者术后死

亡风险进行了预测研究。本研究在计算特征权重和评估分数时引入了权重方差,能够更准确地对特征重要性进行评估,通过数据预处理、Relief-F 算法筛选和医生标注,最终保留了 30 个特征变量,并对所有特征变量进行了分析解释,最后使用 LR、SVM 和 RF 三种机器学习算法训练得到预测结果。本研究采用的方法能够高效、准确地预测出具有高死亡风险的介入患者,辅助医生及时做出医疗干预,从而提高介入治疗质量并降低死亡率,具有较高的应用价值。

**参考文献:**


[1] 于波. 中国血管内影像学研究的进展与展望[J]. 中华心血管病杂志, 2019, 47(9): 722-725.

[2] 中国心血管健康与疾病报告编写组. 中国心血管健康与疾病报告 2022 概要[J]. 中国循环杂志, 2023, 38(6): 583-612.

[3] 中国医院协会介入医学中心分会.《中国介入医学白皮书》2019 版[J]. 中华介入放射学电子杂志, 2020, 8(1): 6-10.

[4] 李雪迎. 重视临床研究数据收集过程[J]. 中国介入心脏病学杂志, 2012, 20(5): 244.

[5] 吴燕秋, 黄伟, 刘慧鑫, 等. 医院创伤专病数据库建设与实践[J]. 医院管理论坛, 2021, 38(5): 79-82.

[6] 刘迷迷, 杜国霞, 周毅, 等. 专病数据库建设与应用研究[J]. 医学信息学杂志, 2021, 42(11): 81-86.

[7] 齐霜, 毛智, 胡新, 等. 基于专科信息系统建立的重症医学数据库: 大型三甲医院重症医学数据库的模式[J]. 中华危重病急救医学, 2020, 32(6): 743-749.

[8] Ruamtawee W, Tipayamongkholgul M, Aimyong N, et al. Prevalence and risk factors of cardiovascular disease among people living with HIV in the Asia-Pacific region: A systematic review[J]. BMC Public Health, 2023, 23(1): 477.

[9] 金涛, 王恺. 我国疾病数据库的建设情况概述[J]. 现代预防医学, 2018, 45(6): 1114-1117.

[10] 袁骏毅, 潘常青, 李榕, 等. 基于临床数据中心的冠心病专病数据库的构建与实现[J]. 中国卫生信息管理杂志, 2022, 19(5): 707-712.

[11] 龙思哲, 吴震天, 黎鹏安, 等. 基于数据治理的专病数据库建设实践[J]. 医学信息学杂志, 2022, 43(7): 20-25.

[12] 赵前前. 基于大数据科研平台的专病数据库建设及应用[J]. 中国数字医学, 2020, 15(12): 89-92.

[13] Behera M P, Sarangi A, Mishra D, et al. A hybrid machine learning algorithm for heart and liver disease prediction using modified particle swarm optimization with support vector machine[J]. Procedia Computer Science, 2023, 218(C): 818-827.

[14] Theerthagiri P, Ruby A U, Vidya J. Diagnosis and classification of the diabetes using machine learning algorithms[J]. SN Computer Science, 2022, 4(1): 72.

[15] Singh D P, Kaushik B. Machine learning concepts and its applications for prediction of diseases based on drug behaviour: An extensive review[J]. Chemometrics and Intelligent Laboratory Systems, 2022, 229: 104637.

[16] Islam M A, Majumder M Z H, Hussein M A. Chronic kidney disease prediction based on machine learning algorithms[J]. Journal of Pathology Informatics, 2023, 14: 100189.

[17] Annamalai B, Saravanan P, Varadharajan I. ABOA-CNN: auction-based optimization algorithm with convolutional neural network for pulmonary disease prediction[J]. Neural Computing and Applications, 2023, 35(10): 7463-7474.

[18] Sudha V K, Kumar D. Hybrid CNN and LSTM network for heart disease prediction[J]. SN Computer Science, 2023, 4(2): 172.

[19] Liang Y, Guo C H. Heart failure disease prediction and stratification with temporal electronic health records data using patient representation[J]. Biocybernetics and Biomedical Engineering, 2023, 43(1): 124-141.

[20] Hao Z Y, Ma J, Sun W J. The technology-oriented pathway for auxiliary diagnosis in the digital health age: A self-adaptive disease prediction model[J]. International Journal of Environmental Research and Public Health, 2022, 19(19): 12509.

[21] 路晓云. 基于机器学习的慢阻肺患者再入院预测和风险分类[D]. 广州: 广东工业大学, 2022: 41-54.

[22] 赵明诚. 基于长短期记忆网络的社区获得性肺炎死亡率预测模型研究[D]. 合肥: 安徽大学, 2021: 42-64.

[23] 刘静, 孙艺红, 彭道泉, 等. 中国心血管病一级预防指南[J]. 中华心血管病杂志, 2020, 48(12): 1000-1038.

[24] 刘晓玉, 李灯熬, 赵菊敏. 基于多核 SVM 的 AdaBoost 心力衰竭死亡率评估模型[J]. 太原理工大学学报, 2023, 54(5): 804-811.

[25] 缪慧, 吴震, 崔文佳. 查尔森合并症指数与中重度老年阻塞性睡眠呼吸暂停综合征患者全因死亡风险的相关性及性别差异分析[J]. 中国耳鼻咽喉头颈外科, 2023, 30(1): 45-50.

[26] Ben Jabeur S, Stef N, Carmona P. Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering[J]. Computational

Economics, 2023, 61(2): 715-741.

[27] Kushwaha N L, Rajput J, Suna T, et al. Metaheuristic approaches for prediction of water quality indices with relief algorithm-based feature selection[J]. Ecological Informatics, 2023, 75: 102122.

[28] Li L J, Xuan M L, Lin Q Z, et al. An evolutionary multitasking algorithm with multiple filtering for high-dimensional feature selection[J]. IEEE Transactions on Evolutionary Computation, 2023, 27(4): 802-816.

[29] Tatliparmak A C, Yilmaz S, Ak R. Importance of receiver operating characteristic curve and decision curve analysis methods in clinical studies[J]. The American Journal of Emergency Medicine, 2023, 70: 196-197.

(责任编辑:康 锋)