DOI:10.3969/j.issn.1005-202X.2024.07.018

医学人工智能

机器学习算法对心脏病预测效能的研究

蒋美艳[1,2],张辉[1,2]
1.广东医科大学第一临床医学院,广东 湛江 524023;2.广东医科大学广东省第二人民医院麻醉科,广东 广州 510317

【摘要】**目的:**探索基于机器学习的方法,包括判定树(DT)、随机森林(RF)、支持向量机(SVM)、K最近邻(KNN)和朴素贝叶斯(NB),构建心脏病预测模型,以实现心脏病的准确预测。**方法:**使用克利夫兰心脏病数据集作为数据源,通过皮尔逊相关系数选择显著特征,使用DT、RF、SVM、KNN和NB算法构建心脏病预测模型,通过准确度、精确度、召回率、F1分数和受试者工作特征曲线下面积(AUC)值等多项指标评估模型性能。**结果:**研究纳入303个样本,样本13个临床特征中有11个显著特征,RF预测模型获得最高的准确度(0.869)、召回率(0.906)、F1分数(0.879)和AUC值(0.93),NB预测模型获得最高的精确度(0.900)。**结论:**基于机器学习的方法能够有效进行心脏病预测,特别是RF预测模型具有显著优势,NB预测模型也表现出令人满意的效果。

【关键词】机器学习;心脏病预测;医疗大数据
【中图分类号】R318;TP391 【文献标志码】A 【文章编号】1005-202X(2024)07-0905-05

Efficacy of machine learning algorithms for heart disease prediction

JIANG Meiyan[1,2], ZHANG Hui[1,2]
1. The First Clinical Medical College, Guangdong Medical University, Zhanjiang 524023, China; 2. Department of Anesthesiology, Guangdong Second Provincial General Hospital, Guangdong Medical University, Guangzhou 510317, China

Abstract: Objective To explore the prediction of heart diseases using machine learning-based methods, including decision trees (DT), random forest (RF), support vector machine (SVM), K-nearest neighbors (KNN), and naive Bayes (NB). **Methods** The Cleveland heart disease dataset was utilized as the data source. Significant features were selected using Pearson correlation coefficients. Heart disease prediction models were constructed using DT, RF, SVM, KNN, and NB algorithms, separately, and the model performance was evaluated with multiple metrics, including accuracy, precision, recall rate, F1 score, and AUC value. **Results** The study included 303 samples, and among the 13 clinical features, 11 were found to be significant. RF prediction model achieved the highest accuracy (0.869), recall rate (0.906), F1 score (0.879), and AUC value (0.93), while NB prediction model obtained the highest precision (0.900). **Conclusion** Machine learning-based methods are promising in heart disease prediction, with the RF prediction model demonstrating significant advantages and NB prediction model exhibiting satisfactory performance.

Keywords: machine learning; heart disease prediction; medical big data

前 言

心脏病,包括冠心病、心肌梗塞、心力衰竭和心律失常等多种疾病,一直是我国居民的重大健康挑战。根据《中国心血管健康与疾病报告》报道,心脏病一直是我国居民的主要死亡原因之一,每年导致数百万人的死亡[1]。心脏病给社会医疗资源带来了巨大的压力[2]。

心脏病具有突发性和隐匿性。许多患者在发病前没有明显症状,因此往往错过了早期干预的机会[3]。此外,心脏病的发病风险受到多种因素的影响,包括年龄、性别、遗传、生活方式和健康状况等,使得心脏病的预测和诊断变得复杂且具有挑战性[4]。近年来,机器学习技术的快速发展为改进心脏病的早期诊断和风险评估提供了新的机会。机器学习能够处理大规模、多源、高维度的医疗数据,从中挖掘潜在的模式和关联[5]。通过建立精确的预测模型,机器学习方法有望提高心脏病的早期诊断准确性,并帮助医生们提供个性化的治疗方案。

【收稿日期】2024-01-20
【基金项目】广东省重点领域研发计划(2020B0101130020)
【作者简介】蒋美艳,硕士研究生,研究方向:麻醉、心脏病预测,E-mail: jmygmu@163.com
【通信作者】张辉,教授,研究方向:术后疼痛、机器学习,E-mail: zhanghui@gd2h.org.cn

本文使用具备广泛临床特征的克利夫兰心脏病数据集建立了多种机器学习预测模型，包括判定树（DT）、随机森林（RF）、支持向量机（SVM）、K最近邻（KNN）和朴素贝叶斯（NB），采用准确度、精确度、召回率和F1分数等多项指标评估模型预测性能。结果表明，机器学习方法在心脏病预测方面有着出色的表现。其中，RF模型在准确度、召回率和F1分数方面表现最佳，为患者提供可靠的心脏病风险评估，NB模型也表现出令人满意的性能。

1 数据与方法

1.1 数据集介绍

克利夫兰心脏病数据集由美国克利夫兰心脏病诊所的心脏病学家收集，并在UCI机器学习存储库中公开提供。该数据集包含303个样本，每个样本具备14个属性，分别为age、sex、cp、trestbps、chol、fbs、restecg、thalach、exang、oldpeak、slope、ca、thal以及target。样本各属性的具体描述如表1所示，数据集中各属性的密度分布如图1所示。

表1 样本各属性的描述

Table 1 Description of each variable of the samples

属性	含义	类型	取值范围
age	年龄	连续	29~77岁
sex	性别	离散	0表示女，1表示男
cp	胸痛经历	离散	1表示典型心绞痛，2表示非典型性心绞痛，3表示非心绞痛，4表示无症状
trestbps	静息血压	连续	94~200 Hg
chol	人体胆固醇	连续	126~564 mg/dL
fbs	空腹血糖	离散	0表示≤120mg/dL，1表示>120 mg/dL
restecg	静息心电图结果	离散	0表示正常，1表示有ST-T波异常，2表示明确的左心室肥厚
thalach	最大心率	连续	71~202 b/min
exang	运动诱发心绞痛	离散	0表示否，1表示是
oldpeak	运动相对于休息引起的ST段压低	连续	0~6.2 mV
slope	峰值运动ST段的斜率	离散	1表示上升，2表示平坦，3表示下降
ca	主要血管数量	连续	0~4条
thal	地中海贫血	离散	0表示缺失，1表示正常，2表示固定缺陷，3表示可逆缺陷
target	是否患有心脏病	离散	0表示否，1表示是

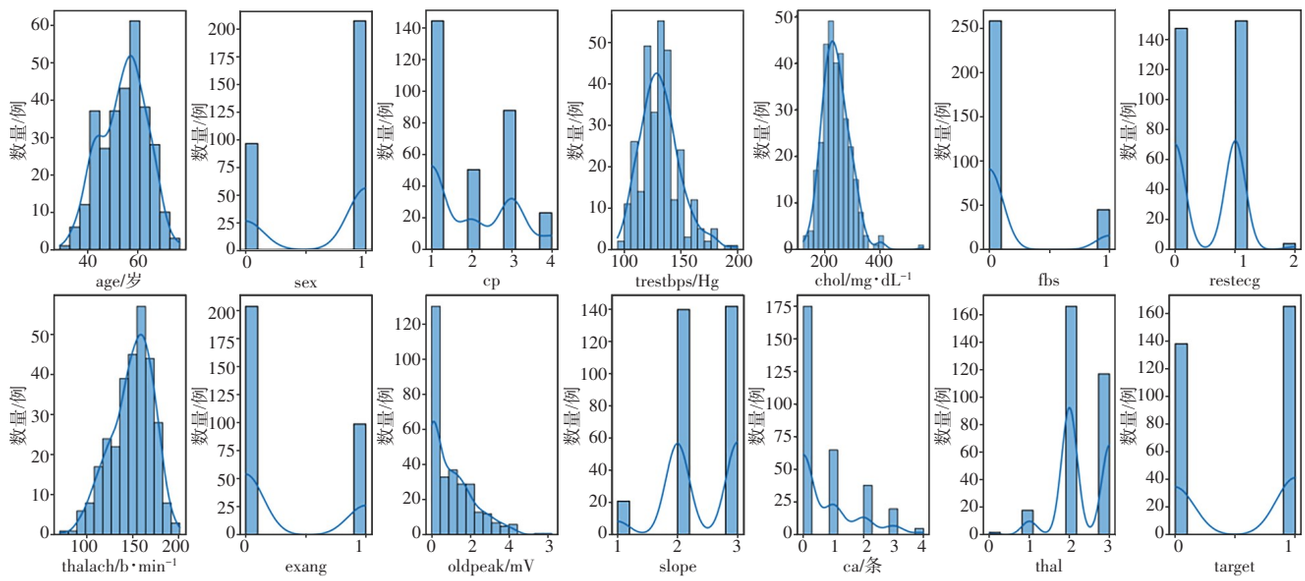


图1 数据集中各属性的密度分布

Figure 1 Density distribution of each variable in the dataset

1.2 方法

如图2所示，首先对数据集进行预处理，将样本target属性作为预测标签，其余属性作为样本特征，对样本特征进行Z-score标准化[6]；然后，计算各样本特征与预测标签间的皮尔逊相关系数，去除相关性低的特征；最后，按照8∶2的比例划分训练集和测试集，训练集样本分别通过DT、RF、SVM、KNN和NB算法构建心脏病预测模型，使用测试集样本对所构建的预测模型进行评估，计算混淆矩阵、准确度、精确度、召回率、F1分数和受试者工作特征（ROC）曲线下面积（AUC）。

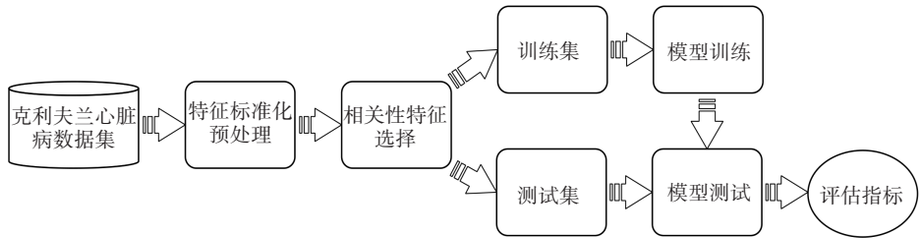


图2 方法流程图

Figure 2 Method flowchart

1.2.1 特征标准化 原始数据集中不同特征取值范围相差过大，可能导致预测模型更加偏向于取值更大的特征，需要对样本特征进行标准化预处理[7]。首先，对于每个临床特征X_i计算其均值μ和标准偏差σ：

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2} \tag{2}$$

其中，n表示样本数量，X_i表示第i个样本的特征值。然后，运用Z-score标准化公式计算得到标准化特征H_i：

$$H_i = \frac{X_i - \mu}{\sigma} \tag{3}$$

完成标准化后的每个特征都服从于均值为0，标准偏差为1的正态分布，这保证不同特征之间的数值范围相似，避免原始数据中数值较小但重要的特征被预测模型忽略。

1.2.2 相关性特征选择 样本13个特征中包含部分冗余特征，需要去除冗余特征避免其对预测结果的干扰。这里采用基于皮尔逊相关系数的特征选择方法[8]。计算各标准化特征与预测标签间的皮尔逊相关系数，去除与预测标签相关性小的特征，保留显著特征。皮尔逊相关系数可以衡量两个变量P和Q之间的线性相关性，其计算流程如下：

Step1：计算P和Q的均值$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} p_i$，$\bar{Q} = \frac{1}{n}\sum_{i=1}^{n} q_i$；

Step2：根据均值计算两个变量的协方差$\mathrm{Cov}(P,Q) = \frac{1}{n}\sum_{i=1}^{n}\left(\left(p_i - \bar{P}\right)\cdot\left(q_i - \bar{Q}\right)\right)$；

Step3：根据均值计算P和Q的标准差$S_P = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(p_i - \bar{P}\right)^2}$，$S_Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(q_i - \bar{Q}\right)^2}$；

Step4：计算皮尔逊相关系数$r = \frac{\mathrm{Cov}(P,Q)}{S_P \cdot S_Q}$。

1.2.3 预测模型构建 DT模型：构建一颗树，树的每个节点代表一个特征属性，每个分支代表一个属性值的判断，叶节点则代表最终的预测结果[9]。RF模型：构建多个DT模型，每个DT模型独立对样本进行预测，采用投票机制，选择获得最多投票的类别作为最终预测结果[10]。KNN模型：计算需要预测的样本与训练集中每个样本点的距离，选择距离最近的K个训练样本点，根据这K个最近邻居的类别使用多数投票原则来预测该样本的标签[11]。SVM模型：寻找一个超平面，能将不同类别的样本点分开，并且使最靠近这个超平面的样本点到该平面的距离最大化[12]。NB模型：基于贝叶斯定理，假设所有特征都是独立且对预测结果有相同的影响（朴素性假设），计算每个类别的后验概率，选择概率最高类别作为预测结果[13]。

2 结 果

2.1 显著特征

计算数据集中各属性间的皮尔逊相关系数，画出变量相关性热力图，如图3所示。皮尔逊相关系数绝对值$|r|$越大，图中颜色越深，表示变量间相关性越强。与target属性相关性最强的是exang属性，即是否患有心脏病与是否运动引发心绞痛密切相关。与target属性相关性最弱的是fbs属性，即是否患有心脏病与空腹血糖值相关性最弱。定义$|r| \leqslant 0.1$的两个变量为弱相关，那么target属性与chol属性、fbs属性弱相关。显著特征为去除弱相关特征后所余特征，显著特征为：age、sex、cp、trestbps、restecg、thalach、exang、oldpeak、slope、ca和thal。

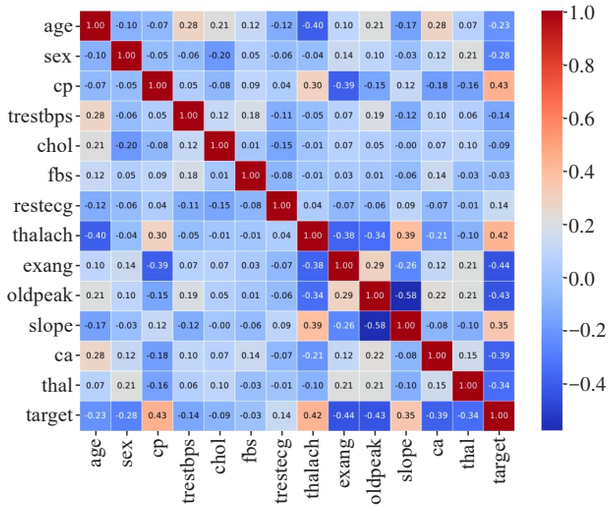


图3 变量相关性热力图

Figure 3 Variable correlation heat map

例(TP)、假正例(FP)、真负例(TN)和假负例(FN)的数量。TN表示预测模型测试过程中正确地将不患病样本预测为不患病的数量;FP表示预测模型测试过程中错误地将不患病样本预测为患病的数量;FN表示预测模型测试过程中错误地将患病样本预测为不患病的数量;TP表示预测模型测试过程中正确地将患病样本预测为患病的数量。画出各机器学习算法预测模型混淆矩阵图(图4),用颜色深浅表示数量多少,TN和TP方格颜色越深表示预测模型性能越好。可以大致看出RF模型和NB模型的性能要优于其他模型,但仍需结合具体的性能指标,给出更准确的评价。

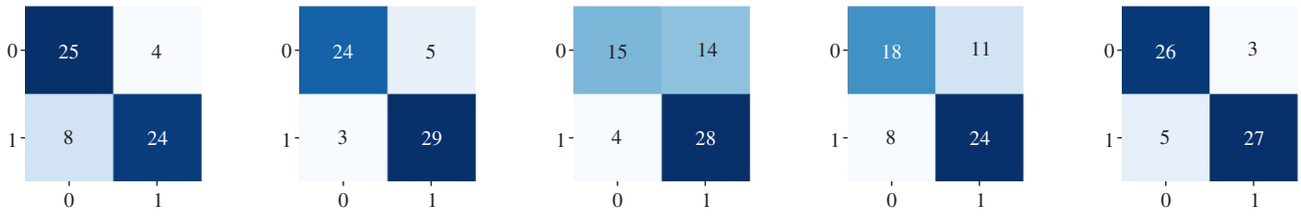
表2 混淆矩阵

Table 2 Confusion matrix

标签	预测不患病	预测患病
实际不患病	TN	FP
实际患病	FN	TP

2.2 混淆矩阵

如表2所示,混淆矩阵是一个表格,用于显示真正



a:DT混淆矩阵　b:RF混淆矩阵　c:SVM混淆矩阵　d:KNN混淆矩阵　e:NB混淆矩阵

图4 各机器学习算法预测模型混淆矩阵图

Figure 4 Confusion matrixes for heart disease prediction by machine learning algorithms

2.3 准确度、精确度、召回率和F1分数

根据各预测模型的混淆矩阵,计算各预测模型的准确度、精确度、召回率、F1分数,如表3所示。准确度是分类正确的样本数量与总样本数量的比率,准确度=(TP+TN)/(TP+TN+FP+FN)。精确度关注将样本分类为正类别时有多少是真正例,精确度=TP/(TP+FP)。召回率关注所有实际为正类别的样本中,有多少被正确地预测为正类别,召回率=TP/(TP+FN)。F1分数是精确度和召回率的调和平均值,它综合考虑了精确度和召回率,在精确度和召回率之间找到一个平衡点。

在准确度方面,RF和NB模型对心脏病预测的准确度最高,而SVM和KNN模型的准确度较差。通常来说准确度越高表示分类器性能越好,但数据集不平衡时,即某个类别的样本数量远远多于另一个类别时,模型可能倾向于预测为占多数的类别,从而导致其准确度偏高,但对于少数类别的预测性能可能很差。在精确度和召回率方面,RF模型的精确度低于NB模型,

表3 各算法模型的测试结果

Table 3 Test results of each algorithm model

模型	准确度	精确度	召回率	F1分数
DT	0.803	0.857	0.750	0.800
RF	0.869	0.853	0.906	0.879
SVM	0.705	0.667	0.875	0.757
KNN	0.689	0.686	0.750	0.716
NB	0.869	0.900	0.844	0.871

但RF模型的召回率在所有模型中表现最佳。在模型错误地将负类别样本预测为正类别(假正例)的代价很高时,通常需要更多地关注精确度。例如,在垃圾邮件检测中,假正例可能导致正常邮件被误标记为垃圾邮件,而提高精确度可以减少这种类型的错误。但在心脏病预测任务中,相对于精确率,应当更多地关注于模型的召回率。因为如果模型错误地将患心

脏病样本预测为不患心脏病样本，可能导致患者错失最佳治疗时机；而即使模型错误地将不患心脏病样本预测为患心脏病样本，那么也可以通过进一步的医疗检查进行排除。在召回率方面，RF模型无疑是心脏病预测的最佳模型。除此之外，RF模型的F1分数在5类模型中表现最佳。尽管SVM模型的召回率仅次于RF模型，但SVM模型的准确度、精确度和F1分数都偏低，难以达到实际应用的标准。NB模型的准确度和精确度都较为优秀，并且其召回率和F1分数也具有竞争力，可以作为心脏病预测任务中RF模型的备用模型。

2.4 ROC曲线及AUC值

为了更加有效地评估各预测模型的综合性能，画出各模型的ROC曲线对比图并计算各自的AUC值，如图5所示。ROC曲线是不同阈值下真正率和假正率之间的关系图，曲线越靠近左上角，模型性能越好；AUC值用于度量模型对正类别和负类别的分类性能，AUC越大表示性能越好。可以看到，RF模型的AUC值最大，达到0.93，其次依次为NB模型、SVM模型、DT模型和KNN模型。结合之前对各类性能指标的分析，可以得出结论：RF是预测心脏病的最佳模型，除此之外，NB也能表现出令人满意的效果。

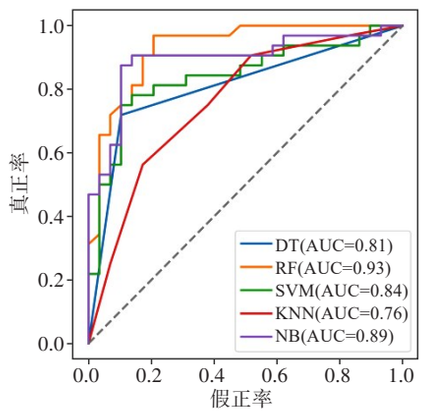


图5 ROC曲线及AUC值

Figure 5 ROC curve and AUC value

3 讨 论

心脏病是全球的重大健康威胁之一，早期的心脏病风险预测和诊断至关重要。本文采用了多种机器学习算法，包括DT、RF、SVM、KNN和NB，对心脏病预测进行了深入研究。通过对心脏病数据集的特征分析和模型训练，本文建立了多个预测模型，并通过绘制混淆矩阵图，计算准确度、精确度、召回率和F1分数等性能指标评估了模型性能。进一步绘制了ROC曲线以展示算法在不同阈值下的性能优势。

结果表明，DT模型尽管易于理解和解释，但其训练过程容易过拟合，虽然DT模型能达到0.857的精确度，但其召回率和AUC值却很低。KNN和SVM模型预测过程简单而直观，但却更适用于数据集分布不规则和不均衡的情况，在克利夫兰心脏病数据集上表现欠佳。RF模型作为DT模型的改进，能够有效克服DT算法的过拟合问题，准确实现心脏病预测，其AUC值达到了0.93。除此之外，尽管NB算法的朴素假设可能不符合实际情况[14]，但它简化了计算，也展现出了不错的性能，其精确度达到了0.900。这些结果为进一步研究和临床应用提供了有力的支持，有助于提高心脏病早期预测的效率和准确性。

但本研究仍然存在不足之处，本研究使用了克利夫兰心脏病数据集，其中包含303个样本。尽管数据集包含了多个临床特征，但样本数量相对较少。这可能对模型的泛化能力产生限制，因为在小规模数据上训练的模型可能难以在更广泛的人群中表现出色。将来的研究可以考虑收集更大规模的数据集，以进一步验证模型的性能。

总的来说，本研究证明了机器学习在心脏病预测中的潜在价值，有望为临床医生提供更准确的预测工具，对全球心脏病健康问题做出积极贡献。

【参考文献】

[1] 马丽媛, 王增武, 樊静, 等.《中国心血管健康与疾病报告2022》要点解读[J]. 中国全科医学, 2023, 26(32): 3975-3994.
Ma LY, Wang ZW, Fan J, et al. Interpretation of report on cardiovascular health and diseases in China 2022[J]. Chinese General Practice, 2023, 26(32): 3975-3994.
[2] 周晓艳, 王慧美, 林诗语, 等. 先天性心脏病患儿主要照顾者家庭管理水平及其影响因素的现况调查[J]. 全科护理, 2023, 21(9): 1259-1262.
Zhou XY, Wang HM, Lin SY, et al. Investigation on family management level and influencing factors of primary caregivers of children with congenital heart disease[J]. Chinese General Practice Nursing, 2023, 21(9): 1259-1262.
[3] Nouman A, Muneer S. A systematic literature review on heart disease prediction using blockchain and machine learning techniques[J]. Int J Comput Innovative Sci, 2022, 1(4): 1-6.
[4] 刘云龙, 周怡君, 罗晨. 基于GBM的特征选择在心脏病预测中的研究[J]. 现代电子技术, 2023, 46(19): 101-106.
Liu YL, Zhou YJ, Luo C. Research on feature selection based on GBM in heart disease prediction[J]. Modern Electronics Technique, 2023, 46(19): 101-106.
[5] 李彩, 范照. 基于机器学习的阿尔兹海默症分类预测[J]. 中国医学物理学杂志, 2020, 37(3): 379-384.
Li C, Fan Z. Classification and prediction of Alzheimer's disease based on machine learning[J]. Chinese Journal of Medical Physics, 2020, 37(3): 379-384.
[6] Yatsko VA. Patterns of using the Z-score for text classification purposes [J]. Autom Doc Math Linguist, 2022, 56(5): 245-250.
[7] Friedman L, Komogortsev OV. Assessment of the effectiveness of seven biometric feature normalization techniques[J]. IEEE Trans Inf Forensics Secur, 2019, 14(10): 2528-2536.
[8] Dufera AG, Liu TT, Xu J. Regression models of Pearson correlation coefficient[J]. Stat Theory Relat Fields, 2023, 7(2): 97-106.
[9] Costa VG, Pedreira CE. Recent advances in decision trees: an updated survey[J]. Artif Intell Rev, 2023, 56(5): 4765-4800.
[10] Dong YL, Zhang SF, Xu JC, et al. Random forest algorithm based on linear privacy budget allocation[J]. J Database Manage, 2022, 33(2): 1-19.
[11] Wang Z, Xu H, Zhou P, et al. An improved multilabel k-nearest neighbor algorithm based on value and weight[J]. Computation, 2023, 11(2): 32.
[12] Birzhandi P, Kim KT, Youn HY. Reduction of training data for support vector machine: a survey[J]. Soft Comput, 2022, 26(8): 3729-3742.
[13] Narayan S, Sathiyamoorthy E. Early prediction of heart diseases using Naive Bayes classification algorithm and Laplace smoothing technique [J]. Int J Grid High Perform Comput, 2023, 14(1): 1-14.
[14] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
Zhou ZH. Machine learning[M]. Beijing: Tsinghua University Press, 2016.

（编辑：薛泽玲）