

基于集成学习的肺癌存活性预测分析

李秀芹, 李琳, 张慢丽

(华北水利水电大学, 河南 郑州 450046)

✉ lixiuqin@ncwu.edu.cn; 1872134171@qq.com; 313364258@qq.com



摘要: 在我国, 恶性肿瘤死亡率最高的就是肺癌。为了提高肺癌存活性预测的准确性, 本研究以随机森林(Random Forest)、LightGBM(Light Gradient Boosting Machine)和CatBoost(Categorical Boosting)三种算法为基模型, 通过线性回归集成融合构建RF-LGC肺癌存活性预测模型, 运用分层十折交叉验证方法进行仿真实验。实验结果显示, RF-LGC组合模型的预测精度达到了98.0242%, 比单一的基模型提高了0.2%; 敏感性达到了89.3957%, 比单一的基模型提高了3%; 特异性达到了78.4848%, 比单一的基模型提高了1%。因此, 该集成融合模型是一种精确、方便的肺癌存活性预测模型。

关键词: 集成学习; 随机森林; 十折交叉验证; 癌症预后

中图分类号: TP311 **文献标识码:** A

Prediction and Analysis of Lung Cancer Survival based on Ensemble Learning

LI Xiuqin, LI Lin, ZHANG Manli

(North China University of Water Resources and Electric Power, Zhengzhou 450046, China)

✉ lixiuqin@ncwu.edu.cn; 1872134171@qq.com; 313364258@qq.com

Abstract: In China, lung cancer has the highest mortality rates among all of the malignant tumors. In order to improve the accuracy of lung cancer survival prediction, this paper proposes to use linear regression integration and fusion to build a lung cancer survival prediction model RF-LGC, which is based on three algorithms — Random Forest, LightGBM (Light Gradient Boosting Machine) and CatBoost (Categorical Boosting). Simulation experiments are carried out by using the hierarchical ten-fold cross-validation method. Experimental results show that the prediction accuracy of the RF-LGC combined model reaches 98.0242%, which is 0.2% higher than that of a single-based model; the sensitivity has reached 89.3957%, which is 3% higher than the single-based model; the specificity has reached 78.4848%, which is 1% higher than the single-based model. Therefore, the integrated fusion model is an accurate and convenient lung cancer survival prediction model.

Keywords: ensemble learning; random forest; ten-fold cross validation; cancer prognosis

1 引言(Introduction)

2021年最新的癌症死亡率调查显示, 男性和女性前十大癌症死亡率中肺癌均占第一位, 约为22%。另外, 在确诊五年后, 肺癌的生存率只有15%, 因此生存分析在医学研究中是最常见的课题之一。为了预测癌症的存活性, 一些相关变量被用来表明考虑因素, 如死亡或疾病的复发是否已经在特定的时间内发生^[1]。预测模型必须估计病人在诊断后是否能存活一

段特定的时间^[2]。癌症存活性预测是一项非常重要的工作, 疾病预后准确性越高, 医疗方面的决策就会越精准, 进而也会提高治疗效果和效率。所以, 提高癌症存活性预测的准确性非常重要。

医院大部分使用统计学方法对数据进行回顾性分析, 但对癌症预后的研究很少。现今机器学习技术的快速发展使得前瞻性的疾病预测成为可能, 一些大数据算法在医疗领域有

了很好的应用，如随机森林^[3-4]、人工神经网络^[5-6]和支持向量机^[7-8]等。王月等人利用最大最小爬山算法预测了五年后肺癌患者的生存情况^[9]；王宇燕等人运用遗传算法改进随机森林，构建GA-RF模型预测直肠癌的存活性^[1]；谭钰洁等人利用LASSO回归分析来建立IV期乳腺癌的生存预测模型^[10]；殷子博基于决策树算法构建癌症合成基因组的预后相关模型^[11]。

机器学习在医疗大数据研究方向上通常都是基于单一预测模型算法进行优化改进的，运行时间较长，泛化性差，很难达到理想的预测精度。本文运用集成融合思想，将好而不同的集成算法进行组合，融合单一基模型的优点，通过混合模型来弥补单一预测模型的驱动性不足。基于此，本研究以随机森林、LightGBM和CatBoost三种算法为基模型，提出构建集成学习融合RF-LGC模型来预测肺癌患者的存活性。

2 肺癌预测模型RF-LGC的构建(Construction of lung cancer prediction model RF-LGC)

2.1 数据来源及变量选择

本实验所使用的数据是来自美国SEER(The Surveillance, Epidemiology, and End Results)数据库中的肺癌数据，它收录了美国各个州几十年来的癌症病例相关信息，并且被认为是全世界相关癌症机构的质量标准^[12]。每个文件有149个属性，每个属性记录都与特定的癌症发病率有关。本实验在了解肺癌的相关资料，进行SEER数据库的初始预处理和数据清洗后，在数据集中选取了23个属性，如表1所示。

表1 肺癌数据集选取的属性

Tab.1 Attributes of lung cancer dataset selection

序号	变量名
1	Age recode with <1 year olds
2	Sex
3	Patient ID
4	Race/ethnicity
5	Primary site
6	Behavior code ICD-O-3
7	Histologic type ICD-O-3
8	Grade
9	Laterality
10	EOD 10-nodes (1988-2003)
11	RX Summ-Surg Prim Site (1998+)
12	Regional nodes examined (1988+)
13	Regional nodes positive (1988+)

(续表)

序号	变量名
14	Origin recode NHIA (Hispanic, Non-Hisp)
15	Reason no cancer-directed surgery
16	Diagnostic confirmation
17	Sequence number
18	EOD 10-extent (1988-2003)
19	AJCC stage 3rd edition (1988-2003)
20	Survival months
21	Survival months flag
22	EOD 10-size (1988-2003)
23	Vital status recode (study cutoff used)

对于肺癌存活性预测，以五年后生存情况作为评价指标，选择患者术后生存情况为结果变量，即若患者生存状况为“survive”则记录为1，生存状况为“dead”记录为0。

2.2 随机森林模型分析

随机森林属于Bagging算法，它通过Bootstrap(自助法)进行重采样，具有模型随机性强、不易过拟合、抗噪性强等优点。此外，它呈树形结构，模型可解释度高，能够执行回归和分类任务^[13]，同时它也是一种数据降维手段。该算法可定义如下：

(1)预设模型的超参数，设置决策树的个数、树的节点层数。

(2)对训练集随机采样生成决策树，然后进行训练，在决策树选择特征时，应选择基尼指数增益值最大的特征，作为该节点分裂条件，如式(1)所示：

$$Info(D) = - \sum_{i=1}^c p_i \log_2(p_i) \quad (1)$$

(3)输入测试样本到每个树中，再将每个树的结构进行整合。对于分类问题，按多棵树分类器投票决定最终分类结果；对于回归问题，由多棵树预测值的均值决定最终预测结果。

2.3 LightGBM模型分析

LightGBM属于Boosting算法，是一种提升算法模型，它是一个实现GBDT(Gradient Boosting Decision Tree)算法的框架。GBDT属于决策树的加法模型，如式(2)所示：

$$f_n(x) = \sum_{n=1}^N L(x; \theta) \quad (2)$$

其中， θ 表示决策树的参数， $L(x; \theta)$ 表示决策树， N 是决策

树的个数,依据算法地向前分布,第 n 步可以表示为式(3):

$$f_n(x) = f_{n-1}(x) + L(x; \theta) \quad (3)$$

设 h_i 是第 i 个样本的真实值, $f_n(x_i)$ 是其预测值, 则损失函数可以表示为式(4):

$$J(h_i, f_n(x_i)) = \frac{1}{2} (h_i - f_n(x_i))^2 \quad (4)$$

利用式(5)损失函数极小化得到参数 $\hat{\theta}$:

$$\hat{\theta} = \arg \min \sum_{i=1}^N J(h_i, f_{n-1}(x) + L(x; \theta)) \quad (5)$$

通过多次迭代, 回归树更新得到最终模型。LightGBM通过部分样本计算信息增益, 降低了每次拆分增益的成本; 并且内置特征降维技术, 降低内存使用, 处理困难样本能力强; 使用叶子节点直方图进一步加速计算; LightGBM支持高效并行, 降低并行学习的通信成本。

2.4 CatBoost模型分析

CatBoost是一种基于对称决策树为基学习器, 实现参数较少、支持类别型变量和高准确性的机器学习框架, 使用了组合类别特征, 丰富了特征维度; 采用排序提升的方法对抗训练集中的噪声点, 从而避免梯度估计的偏差, 它将样本的二进制特征存储在向量 C 中, 叶子节点中的值存储在 2^d 的向量中^[14]。对于样本 y , 建立二进制向量 C , 如式(6)所示:

$$C = \sum_{i=0}^{d-1} 2^i B(y, f(t, j)) \quad (6)$$

其中, $B(y, f)$ 是样本 y 上的特征 f 的值, $f(t, j)$ 是二进制特征的数目。构建向量是以数据并行方式, 使得速度高达三倍。

2.5 RF-LGC肺癌存活性预测模型的构建

本文引入了集成学习法, 集成学习不是一个特定的模型, 而是一种思想, 它结合了一组简单的基本模型的优点, 从而构建一个相对更强大的模型。Bagging算法和Boosting算法是集成算法的两大类, 两者存在异质性, 好而不同的算法融合集成起来会有更好的预测效果。Bagging算法的代表随机森林简单稳定, 具有高维运行数据的能力, 抗噪性强, 不易过拟合, 但并不具备处理困难样本的能力。Boosting算法的代表LightGBM是一种新型提升算法, 可以并行计算优化, 但它属于迭代算法, 使得模型偏差比较低, 对噪点较为敏感。CatBoost算法具有鲁棒性, 可以处理类别型、数值型特征, 但对于类别型特征的处理需要大量的内存和时间。本文综合三者优缺点, 使彼此相辅相成, 另外使用stacking的结合策略构建RF-LGC模型, 充分发挥每一个学习器的优点, 防止过拟合, 而且不用过多地调参数, 能够有效预测肺癌存活性。

本文融合构建的RF-LGC肺癌存活性预测模型如图1所

示, 过程采用分层十折交叉验证, 进行stacking的结合策略。

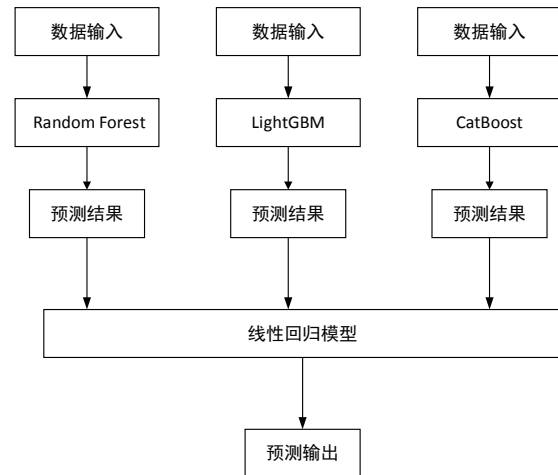


图1 肺癌存活性预测模型

Fig.1 Prediction model of lung cancer survival

(1)基于训练集训练模型。将90%的训练集分别输入随机森林、LightGBM和CatBoost三个基模型中, 在训练模型LightGBM模型构建过程中, 目标函数采用的是binary, 评价函数采用的是AUC, 目的是可以根据需要对评价函数作调整, 设定一个或者多个评价函数; CatBoost中eval_metrics参数设定为AUC; 随机森林模型采用Random Forest Classifier类构造函数, 并运用Predict_proba(X)预测函数, 同样也采用AUC来计算训练后的模型某一指定指标在每一轮迭代中的表现, 得到训练集学习结果预测矩阵。

(2)将剩余10%的测试集, 利用分层十折交叉验证, 先基于三个基模型训练出的模型进行测试, 得到测试预测结果矩阵。

(3)同样采用分层十折交叉验证, 将步骤(1)训练集预测结果作为新的训练集样本输入线性回归模型进行训练, 得到最终的训练集预测结果; 将步骤(2)测试集预测结果作为新的测试集样本输入线性回归模型进行测试, 得到最终预测结果。

基于本文的线性回归模型, 调用了LinearRegression包, 采用最小二乘法线性回归, 通过最小化误差的平方和寻找数据的最佳函数匹配, 我们可以通过残差之和最小化实现最佳拟合, 也就是说模型预测的值与训练集的数据最接近, 如式(7)所示:

$$ss = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (7)$$

最终, 在实现最佳拟合过程中, 利用AUC值来证明模型的可靠性, 实验结果如表2所示, 集成融合模型达到AUC最高, 具有可行性。

表2 算法AUC值

Tab.2 AUC value of algorithm

算法	AUC值
随机森林	0.961
LightGBM	0.970
CatBoost	0.967
RF-LGC	0.981

3 存活性预测实验分析(Experimental analysis of survival prediction)

3.1 分类性能评价标准

评估一个预测模型, 需要用合适的衡量手段, 对于本文所研究的医疗诊断预测领域, 选用的指标为敏感性^[15]、特异性^[16]和准确性^[17], 另外使用了分层十折交叉验证^[18]进行模型准确性测试。其中, 敏感性评价患者被查出患病的可能性, 如式(8)所示; 特异性是对阴性者被正确排除患病可能性的评价, 如式(9)所示; 而准确性是由特异性和敏感性一起决定的, 从整体角度来判断预测是否准确。

$$\text{敏感性} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{特异性} = \frac{TN}{TN + FP} \quad (9)$$

其中, TP 代表真正类^[19], FP 代表假正类, TN 代表真负类, FN 代表假负类。以上三个评价指标都是越大越好。分层十折交叉验证是估计分类方法精度的一种方法, 目的是测试所采用分类方法的泛化能力^[20], 解决数据不平衡性。分层十折交叉验证是把全部数据集分成基本一致的十个子集, 然后对模型进行十次训练、测试。每次使用九个不同的子集来训练模型, 剩下的一个作为测试集。之后将十次计算的正确率取均值作为本方法的估计精度值。分层十折交叉验证的正确率 A 公式如式(10)所示:

$$A = \frac{1}{10} \sum_{n=1}^{10} A_n \quad (10)$$

$n = 1, 2, 3, \dots, 10$

其中, A_n 是第 n 折计算得到的正确率。通过分层十折交叉验证计算得到的正确率, 可以知道模型在整体上的性能情况, 另外通过计算其平均值和标准差, 比较不同模型的稳定性。

3.2 实验结果与分析

本实验为了验证提出的集成模型具有更好的存活性预测性能, 将集成融合RF-LGC模型分别与单一的随机森林、LightGBM和CatBoost模型进行对比。表3是对肺癌数据分类的实验结果, 包括四种模型在肺癌数据集上进行分层十折交叉验证运算得到的各项指标、标准差和均值。

从表3中可以看到, 随机森林、LightGBM、CatBoost

和集成模型RF-LGC这四种模型敏感性的均值分别是0.863338、0.872936、0.898075、0.893957, 特异性的均值分别是0.774608、0.790723、0.780762、0.784848, 准确性的均值分别是0.978158、0.979854、0.980033、0.980242, 可以看出这三个指标在模型中的均值属集成模型RF-LGC最高, 次之是CatBoost; 四种模型的敏感性标准差分别是0.020570、0.020015、0.025285、0.023157, 特异性标准差分别是0.024972、0.017865、0.024759、0.017315, 准确性标准差分别是0.002559、0.002003、0.002390、0.001904, 可以看出这三个指标在模型中的标准差属集成模型RF-LGC最低, 次之是CatBoost。

表3 肺癌数据分类的实验结果

Tab.3 Experimental results of lung cancer data classification

折数	算法	敏感性	特异性	准确性
1	随机森林	0.881279	0.768924	0.978417
	LightGBM	0.863014	0.790795	0.979445
	CatBoost	0.899543	0.784861	0.980473
	RF-LGC	0.890411	0.792683	0.980730
2	随机森林	0.853881	0.745020	0.975334
	LightGBM	0.876712	0.780488	0.979188
	CatBoost	0.894977	0.774704	0.979445
	RF-LGC	0.894977	0.777778	0.979702
3	随机森林	0.885845	0.822034	0.982785
	LightGBM	0.908676	0.808943	0.982785
	CatBoost	0.917808	0.800797	0.982528
	RF-LGC	0.926941	0.802372	0.983042
4	随机森林	0.872146	0.789256	0.979702
	LightGBM	0.890411	0.809129	0.982014
	CatBoost	0.904110	0.801619	0.982014
	RF-LGC	0.899543	0.800813	0.981757
5	随机森林	0.840183	0.763485	0.976362
	LightGBM	0.844749	0.777311	0.977646
	CatBoost	0.885845	0.772908	0.978931
	RF-LGC	0.867580	0.775510	0.978417
6	随机森林	0.825688	0.765957	0.976099
	LightGBM	0.844037	0.786325	0.978412
	CatBoost	0.853211	0.759184	0.976613
	RF-LGC	0.853211	0.775000	0.977898

(续表)

折数	算法	敏感性	特异性	准确性
7	随机森林	0.885321	0.807531	0.981753
	LightGBM	0.885321	0.824786	0.983038
	CatBoost	0.931193	0.812000	0.984066
	RF-LGC	0.912844	0.812245	0.983295
8	随机森林	0.849315	0.756098	0.976099
	LightGBM	0.867580	0.769231	0.977898
	CatBoost	0.931507	0.733813	0.977127
	RF-LGC	0.922374	0.753731	0.978699
9	随机森林	0.876712	0.750000	0.976613
	LightGBM	0.881279	0.775100	0.978926
	CatBoost	0.894977	0.762646	0.978412
	RF-LGC	0.890411	0.773810	0.979183
10	随机森林	0.863014	0.777778	0.978412
	LightGBM	0.867580	0.785124	0.979183
	CatBoost	0.867580	0.805085	0.980725
	RF-LGC	0.881279	0.784533	0.979697
均值	随机森林	0.863338	0.774608	0.978158
	LightGBM	0.872936	0.790723	0.979854
	CatBoost	0.898075	0.780762	0.980033
	RF-LGC	0.893957	0.784848	0.980242
标准差	随机森林	0.020570	0.024972	0.002559
	LightGBM	0.020015	0.017865	0.002003
	CatBoost	0.025285	0.024759	0.002390
	RF-LGC	0.023157	0.017315	0.001904

为了更直观地展示集成模型RF-LGC的优越性，我们以图形的方式对实验结果进行展示。图2以折线图来展示这四种模型在分层十折交叉验证中的敏感性值，图3和图4分别是特异性和准确性折线图。从图2中可以看出，由于正类样本在验证中分布不均，敏感性的值变化较大，CatBoost和集成模型RF-LGC优劣难分，二者在整体上都要优于LightGBM，次之是随机森林。从图3中可以看出，集成RF-LGC模型的波动与LightGBM、CatBoost大致相似，LightGBM的特异性优于RF-LGC模型，RF-LGC模型的特异性优于CatBoost。随机森林波动过大，其稳定性相对于其三个模型较差。从图4中

可以看出，在准确性上可以排序为：集成模型RF-LGC优于CatBoost，CatBoost优于LightGBM，LightGBM优于随机森林。从以上分析结果可以看出，本研究所提出的集成模型得到的三个指标的均值都大于各个基模型，而标准差也都更小，证明了该模型的泛化能力更强。

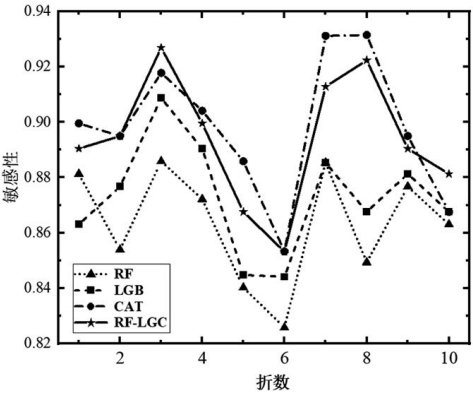


图2 十折交叉验证得到的敏感性对比图

Fig.2 Sensitivity comparison diagram obtained by ten-fold cross validation

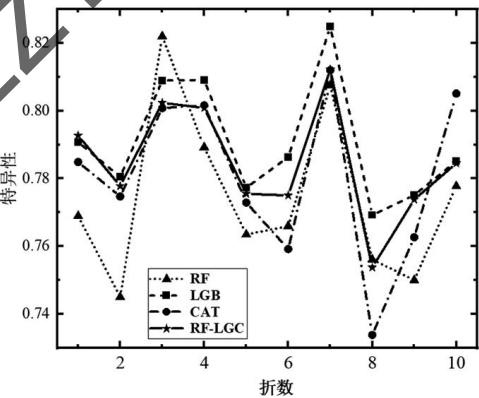


图3 十折交叉验证得到的特异性对比图

Fig.3 Specificity comparison diagram obtained by ten-fold cross validation

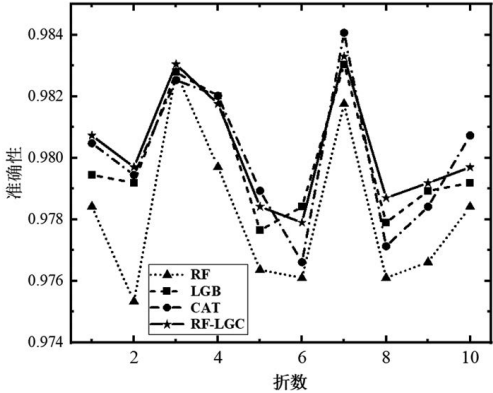


图4 十折交叉验证得到的准确性对比图

Fig.4 Accuracy comparison chart obtained by ten-fold cross validation

本研究所构建的集成融合模型不仅准确性最高，而且表

现出了更好的稳定性和泛化性,敏感性和特异性均值分别达到了0.893957和0.784848,这两者也共同为准确性提供了较高的均值,达到了0.980242,因此本研究提出的模型能够达到较好的预测结果,能够有效地协助医疗领域做出相应的决策。

4 结论(Conclusion)

本文提出了一种肺癌存活性预测的集成融合模型。该模型基于集成学习两大类不同的机器学习技术,充分发挥各个模型的优点,能够捕获数据中的复杂模式,高效而简洁,再利用线性回归来融合构建模型,增强了各个基模型的稳定性。本文使用了真实的肺癌数据,且结果表明所提出的集成模型能够达到理想的预测精度,稳定性和泛化能力都较强,因此能够推广到医疗领域为癌症病人预后预测提供决策,以弥补传统经验预测带来的不足,降低医疗成本,对癌症的治疗和预测都具有很大的现实意义。

参考文献(References)

- [1] 王宇燕,王杜娟,王延章,等.改进随机森林的集成分类方法预测结肠直肠癌存活性[J].管理科学,2017,30(1):95-106.
- [2] 郭占芳,张红武,杨如意,等.益气复脉对中晚期肺癌患者免疫功能和生存质量的影响[J].中国现代医学杂志,2017,27(6):88-92.
- [3] CHEN H, LIN Z, WU H G, et al. Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2015, 135A:185-191.
- [4] AZAR A T, ELSHAZLY H I, HASSANIEN A E, et al. A random forest classifier for lymph diseases[J]. Computer Methods and Programs in Biomedicine, 2014, 113(2):465-473.
- [5] DEVI M A, RAVI S, VAISHNAVI J, et al. Classification of cervical cancer using artificial neural networks[J]. Procedia Computer Science, 2016, 89:465-472.
- [6] LIN D, ATHANASIOS V V, TANG Y, et al. Neural networks for computer-aided diagnosis in medicine: A review[J]. Neurocomputing, 2016, 216(5):700-708.
- [7] 赵峦啸,刘金水,姚云霞,等.基于随机森林算法的陆相沉积烃源岩定量地震刻画:以东海盆地长江拗陷为例[J].地球物理学报,2021,64(2):700-715.
- [8] AZAR A T, EI-SAID S A. Performance analysis of support vector machines classifiers in breast cancer mammography recognition[J]. Neural Computing and Applications, 2014, 24(5):1163-1177.
- [9] 王月,赵茂先.基于最大最小爬山算法的肺癌预后模型[J].山东科技大学学报(自然科学版),2020,39(2):105-110.
- [10] 谭钰洁,何子凡,余运芳,等.首诊IV期乳腺癌生存预测模型建立并验证:一项基于机器算法的研究[J].岭南现代临床外科,2020,20(3):273-279.
- [11] 殷子博.基于决策树算法的癌症合成致死基因组合的预测及预后分析[D].南京:南京邮电大学,2020.
- [12] DELEN D, WALKER G, KADAM A. Predicting breast cancer survivability: A comparison of three data mining methods[J]. Artificial Intelligence in Medicine, 2005, 34(2):113-127.
- [13] 邱少明,杨雯升,杜秀丽,等.优化随机森林模型的网络故障预测[J].计算机应用与软件,2021,38(2):103-109,170.
- [14] 苏庆,林华智,黄剑锋,等.结合CNN和Catboost算法的恶意安卓应用检测模型[J].计算机工程与应用,2021,57(15):140-146.
- [15] 陈志君,朱振闯,孙仕军,等.Stacking集成模型模拟膜下滴灌玉米逐日蒸散量和作物系数[J].农业工程学报,2021,37(5):95-104.
- [16] 扈文秀,苏振兴,杨标.基于随机森林方法的投资者概念关注树概念指数收益预测及交易策略的研究[J].预测,2021,40(1):60-66.
- [17] 李昆明,厉文婕.基于利用BP神经网络进行Stacking模型融合算法的电力非节假日负荷预测研究[J].软件,2019,40(9):176-181.
- [18] 吴彤,李勇,葛莹,等.利用Stacking集成学习估算柑橘叶片氮含量[J].农业工程学报,2021,37(13):163-171.
- [19] MOLINA-MAYO C, HERNÁNDEZ-BORGES J, BORGES-MIQUEL T M, et al. Determination of pesticides in wine using micellar electrokinetic chromatography with UV detection and sample stacking[J]. Journal of Chromatography A, 2007, 1150(1/2):348-355.
- [20] 耿琪深,王丰华,金霄.基于Gammatone滤波器倒谱系数与鲸鱼算法优化随机森林的干式变压器机械故障声音诊断[J].电力自动化设备,2020,40(8):191-196,224.

作者简介:

李秀芹(1967-),女,博士,教授.研究领域:计算机网络,数据库与信息处理.

李琳(1995-),女,硕士生.研究领域:数据挖掘.

张慢丽(1994-),女,硕士生.研究领域:数据挖掘.