张晓飞, 宋其江. 基于 RF-RFECV 和 Stacking 集成学习的脑卒中预测研究[J]. 智能计算机与应用,2024,14(5):252-256.
DOI:10.20169/j.issn.2095-2163.240536

基于 RF-RFECV 和 Stacking 集成学习的脑卒中预测研究

张晓飞, 宋其江
(东北林业大学 机电工程学院, 哈尔滨 150040)

摘 要: 脑卒中具有发病率高、死亡率高和致残率高的特点,提早发现和治疗显得至关重要。在脑卒中预测方法中,机器学习相对于其他方法具有更好的表现。针对传统的单一机器学习模型在预测的精度或稳定性上都存在局限性的问题,提出了一种基于 RF-RFECV 和 Stacking 集成学习的脑卒中预测方法。通过实验证明,该方法可以有效地降低特征维度,获得最优特征子集,与其他的单一模型以及其他集成算法模型相比,Stacking 模型的预测精度明显提升,可以更有效地预测脑卒中。

关键词: SMOTE 算法; RF-RFECV; Stacking 模型; 脑卒中; 机器学习
中图分类号: TP181 **文献标志码**: A **文章编号**: 2095-2163(2024)05-0252-05

Research on stroke prediction based on RF-RFECV and Stacking integrated learning

ZHANG Xiaofei, SONG Qijiang
(School of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: Stroke has the characteristics of high incidence rate, high mortality and high disability rate. Early detection and treatment are essential. Among stroke prediction methods, machine learning has better performance than other methods. Aiming at the limitation of traditional single machine learning model in prediction accuracy or stability, stroke prediction method based on RF-RFECV and Stacking ensemble learning is proposed. Experiments show that this method can effectively reduce feature dimensions and obtain the optimal feature subset. Compared with other single models and other integrated algorithm models, the prediction accuracy of Stacking model is significantly improved. The research can more effectively predict stroke.

Key words: SMOTE algorithm; RF-RFECV; Stacking model; stroke; machine learning

0 引 言

脑卒中是严重危害中国国民健康的重大慢性非传染性疾病,是国内成人致死、致残的首位病因,具有高发病率、高致残率、高死亡率、高复发率、高经济负担五大特点[1]。根据科学研究表明,脑卒中已经成为全球第二大死亡原因。在国内,脑卒中发病率为 39.3%[2],防治形势也日益严峻。由于现阶段缺乏对脑卒中的有效治疗手段,因此脑卒中的早期预防尤为重要。近年来,随着机器学习的不断成熟和发展,已被广泛应用于医学领域。利用机器学习对脑卒中进行及时的预测,具有重要的临床意义[3]。例如,常怀文等学者[4]构建支持向量机模型。李鹏等学者[5]构建逻辑回归模型。吴菊华等学者[6]构建神经网络模型对脑卒中进行风险预测,这些成果对脑卒中的防治具有不可忽视的积极意义。但是单一的机器学习模型存在预测精度低,泛化性差等问题。因此,本文研究提出基于 RF-RFECV 和 Stacking 集成学习的脑卒中预测方法,综合多个单一模型的优势,提升模型精度,为预防诊断提供有效辅助。

1 建模过程

1.1 实验数据

实验数据采用 kaggle 网站的脑卒中数据集。数据集包含 5 110 个样本,其中包括 10 个特征属性

基金项目: 中央高校基本科研业务费专项资金项目(2572020DR12)。
作者简介: 张晓飞(1998-),男,硕士研究生,主要研究方向:机器学习,模式识别。
通讯作者: 宋其江(1975-),男,博士,讲师,硕士生导师,主要研究方向:人工智能,大数据分析,深度学习。Email:2729505619@qq.com
收稿日期: 2023-04-20

和 1 个标签属性，此处提及的 10 个特征属性见表 1。

表 1　数据集特征属性

Table 1　Characteristic attributes of the dataset

属性名称	属性描述
Gender	性别
Age	年龄
Hypertension	高血压病史
Heart_disease	心脏病病史
Ever_married	婚姻史
Work_type	工作类型
Residence_type	居住类型
Avg_glucose_level	平均血糖水平
Bmi	身体质量指数
Smoking_status	吸烟状况

标签属性值为 0 和 1，其中 0 表示未患有脑卒中，1 表示患有脑卒中。经过分析统计，标签值为 0 的样本数为 4 861 个，标签值为 1 的样本数为 249 个。

1.2　数据预处理

本文对数据进行预处理主要包括 3 个方面：缺失值处理、数据编码以及数据不平衡处理。

首先，数据缺失会造成原始数据的信息价值的减少，导致预测结果不理想，而且不能采用常规统计分析方法。因此，需要对缺失的数据进行数据处理。缺失值处理常用方法有删除法、插补法。实验采用插值法以平均数对脑卒中缺失数据进行填充。

然后，由于数据集中存在类别特征，为了其能够在机器学习模型中得到更好的处理，需要将其转变为数值型的特征。数据编码的常用方法主要包含 OneHotEncoder、OrdinalEncoder。因为类别特征存在内在顺序，本文采用 OrdinalEncoder 法进行数据编码，进一步提高模型的泛化性能。

最后，从图 1 中看出卒中和非卒中患者分布差别明显，为了解决数据不均衡对建模分析的影响，采用 SMOTE 算法对不平衡数据进行过采样。算法通过对少数类别样本进行分析与模拟，将模拟得到的新样本纳入数据集，从而使原始数据中类别数量达到平衡[7]。处理后的患者分布如图 2 所示。

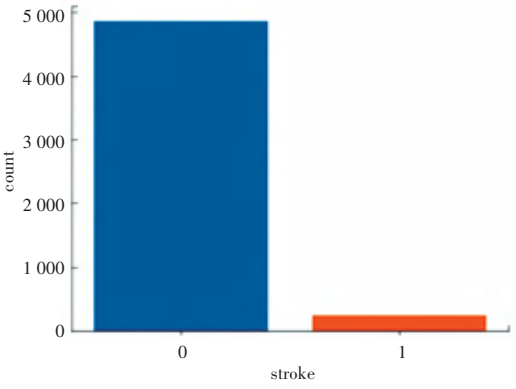

图 1　数据集标签分布

Fig. 1　Label distribution of the dataset

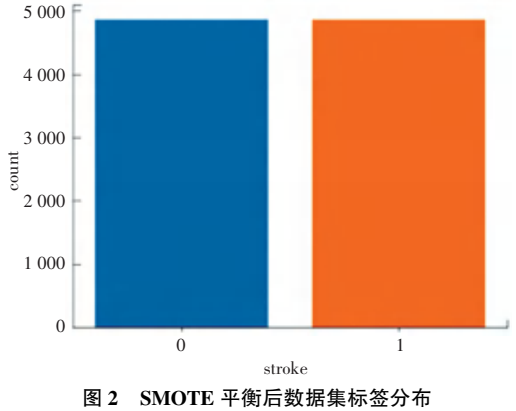

图 2　SMOTE 平衡后数据集标签分布

Fig. 2　SMOTE balanced dataset label distribution

1.3　特征工程

本文通过利用 Python 机器学习库 Sklearn 中的 RandomForestClassifier 和 RFECV 模块，实现特征重要度的分析和特征选择。递归特征消除是指在经过特征重要性排序后，剔除了重要性水平最低的特征并更新生成新的特征子集，然后再对随机森林模型继续进行训练和评估，可以保存原始数据信息和得到不同特征在预测结果上的重要性[8]。通过 15 折交叉验证法对每个特征子集相对应的随机森林模型重复运行 15 次，以 15 次预测结果的准确率的平均值为特征子集的评价指标[9]。最优特征子集的平均交叉验证准确率变化情况如图 3 所示。当特征子集数量小于 8 时，随着特征数量增多、平均交叉验证准确率也会逐渐提高，预测准确性也呈现总体提高态势。当特征子集数量为 8 时，平均交叉验证准确率达到最高，预测准确率也是最高。因此实验选择平均交叉验证准确率最高的 8 个特征组成的最优特征子集，计算各个特征对预测结果的影响程度，其排序如图 4 所示。

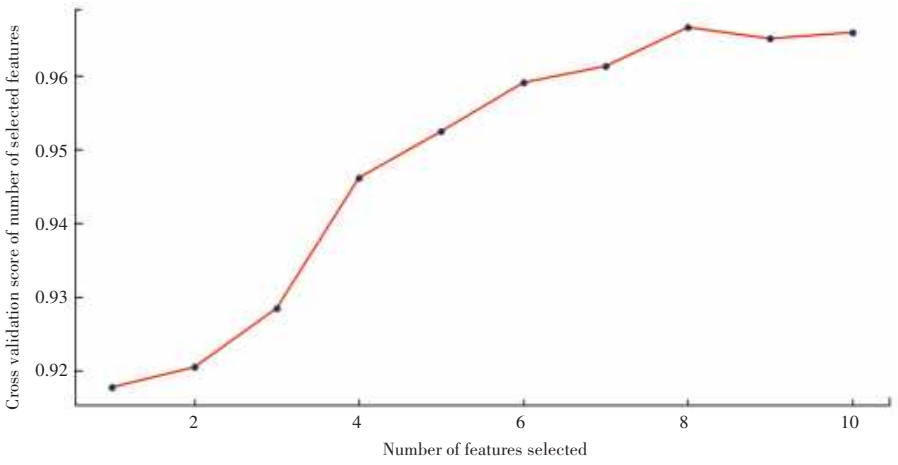

图 3 不同特征子集的平均交叉验证准确率

Fig. 3 Average cross-validation accuracy of different feature subsets

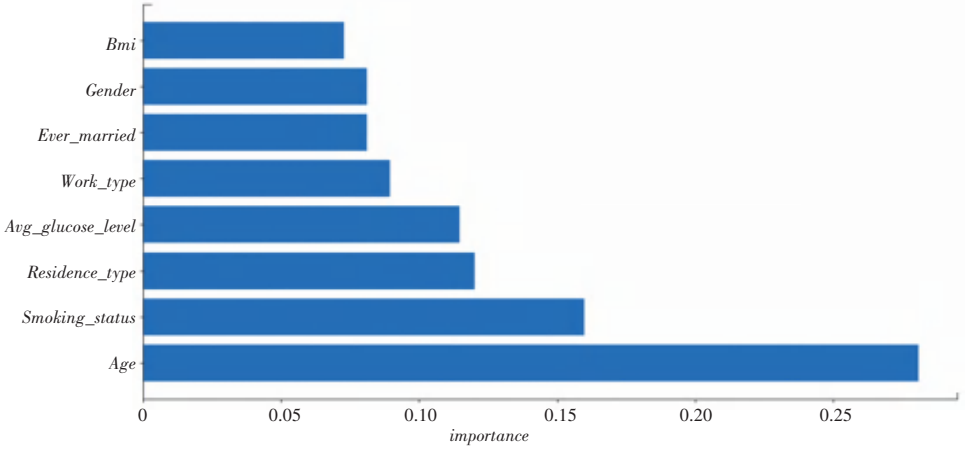

图 4 特征影响程度排名

Fig. 4 Ranking of influence degree of characteristics

1.4 Stacking 集成学习模型构建

由于单一模型存在预测精度差和泛化能力弱等问题,为了解决这类问题,建立具有 2 层结构的 Stacking 模型。Stacking 的本质是一种堆栈集成算法,可以通过整合多种不同类型的机器学习模型,使得模型的边界变得更稳定[10]。构建 Stacking 模型不仅仅要考虑到合适的基学习器和元学习器,还要考虑到不同方式的组合。

1.4.1 基分类器的选择

实验使用决策树(DT)、随机森林(RF)、梯度提升树(GBDT)、XGBoost、AdaBoost、支持向量机(SVM)、KNN、逻辑回归(LR)、朴素贝叶斯(BernoulliNB)等主流分类器进行研究分析。在进一步选择第一层的基分类器时,要综合考虑 2 个方面。一是 Stacking 模型的优越性主要体现在不同基学习器对原始数据的特征提取,不同基学习器获得的特征各不相同,为此要选择差异性较大的基学习器,才能最大程度地综合不同算法的优势;二是单个基学习器性能的优劣对融合模型的最终性能具有很大的影响,所以要选择学习能力强的基学习器。

为了得到性能较强的基分类器,需要对比各个模型单独预测的效果。单一模型的参数采用交叉验证结合网格搜索算法确定。各个模型参数以及预测性能见表 2。

由表 2 可得,经过网格搜索算法调参后,AdaBoost、GBDT、XGBoost 的表现是单一模型中相对出色的。为了增强 Stacking 提升效果,将 AdaBoost、GBDT、XGBoost 引入 Stacking 集成中,这几个算法虽然在训练方法上不同,但本质上都是基于 Boosting 算法的模型。因此,将基于感知机的 SVM 算法、基于距离的 KNN 算法以及基于 Bagging 算法的 RF 加入 Stacking 的基学习器中来确保第一层基学习器的多样性。综上所述,本文选择 RF、AdaBoost、KNN、SVM、GBDT、XGBoost 作为候选基学习器。

表 2　各模型的参数设置及预测结果

Table 2　Parameters setting and prediction results of each model

模型	超参数设置	*Precision*	*F*1
DT	*min_samples_split* = 1 *min_samples_leaf* = 2	0.951 664	0.951 427
RF	*criterion* = *gini* *n_estimators* = 1 000 *min_samples_leaf* = 1	0.974 358	0.973 819
GBDT	*min_samples_leaf* = 5 *max_depth* = 15 *learning_rate* = 0.1	0.974 745	0.974 453
XGBoost	*max_depth* = 8 *n_estimators* = 500	0.971 590	0.971 299
AdaBoost	*n_estimators* = 500 *learning_rate* = 0.1	0.976 038	0.975 714
SVM	*C* = 1 *Kernel* = *poly*	0.765 334	0.759 891
KNN	*n_neighbors* = 5	0.900 613	0.881 996
LR	*C* = 1	0.764 180	0.763 294
BernoulliNB	*alpha* = 100	0.766 093	0.757 366

1.4.2　元分类器的选择

由于第二层元分类器的输入数据是各个基分类器的预测结果，为了防止过拟合，第二层的元学习器一般选结构简单的 LR 模型，不但可以提高整体模型的稳定性，而且可以通过正则化避免过拟合。

1.4.3　Stacking 中不同学习器组合比较

实验设计对比不同组合下 Stacking 模型的预测效果，选择单一模型中预测性能最好的 AdaBoost 模型以及和 AdaBoost 关联度较小的且预测性能较好的 KNN、SVM 模型作为基学习器，并在此基础上结合了其他的学习器进行不同组合，得到实验结果见表 3。

表 3　不同模型组合的预测结果

Table 3　Prediction results of different model combinations

组合方式	基学习器	*Precision*	*F*1
1	GBDT	0.977 749	0.977 609
2	XGB	0.978 425	0.978 239
3	RF	0.979 359	0.979 185
4	GBDT、XGB	0.978 036	0.977 924
5	GBDT、RF	0.979 947	0.979 816
6	RF、XGB	0.978 146	0.977 923
7	GBDT、RF、XGB	0.981 172	0.981 078

由表 3 可知，在基学习器的不同组合下预测模型结果存在很大的不同。组合方式 1 的预测效果好于组合方式 2、3，说明 RF 作为基学习器增加输入元学习器的特征差异，从而增强了模型的泛化能力。组合方式 5 的预测效果好于组合方式 4、6，说明在保证基学习器差异性的前提下，选择基学习器的学习能力越强，融合模型的最终性能也越好。组合方式 7 的预测结果好于组合方式 3、5，说明基学习器的数量在一定程度上影响融合模型的预测性能。根据上述预测结果对比，选择组合方式 7 作为本实验 Stacking 模型基学习器组合如图 5 所示。

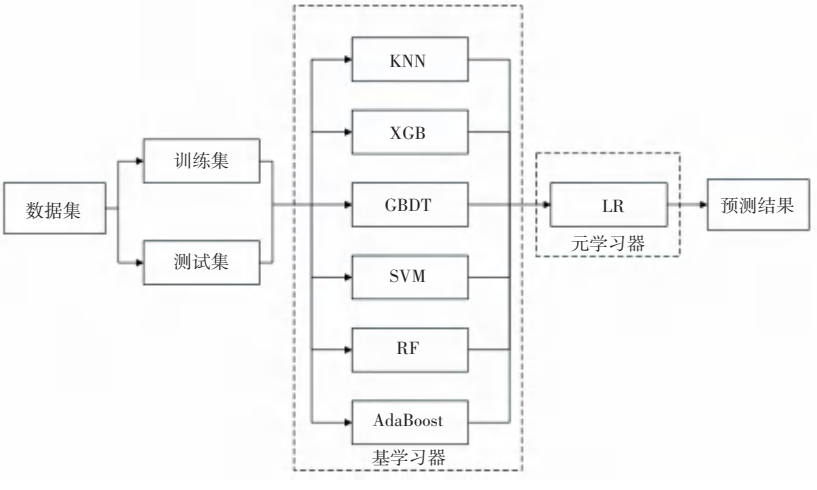

图 5　Stacking 集成学习模型流程

Fig. 5　Stacking integrated learning model process

2 实验结果与分析

根据前文的算法设计以及参数调整,在经过处理的数据集上进行实验验证,主要包括:特征选择方法 RF-RFECV 的有效性实验,不同预测模型与 Stacking 模型的对比实验。本文选取 *Precision* 和 *F*1 值作为评价指标。

2.1 特征选择前后模型性能对比分析

特征选择前后模型性能对比见表 4。根据表 4 所示,对比特征选取前后的模型性能,在多数情况下,模型性能均有较大提升说明通过 RF-RFECV 不仅减少数据的维度,同时获得的最优特征子集更有利于预测任务。

表 4 特征选择前后模型性能对比

Table 4 Model performance comparison before and after feature selection

模型	指标	特征选择前	特征选择后
DT	*Precision*	0.946 089	0.951 664
	*F*1	0.946 074	0.951 427
RF	*Precision*	0.971 242	0.974 358
	*F*1	0.970 665	0.973 819
SVM	*Precision*	0.765 966	0.765 334
	*F*1	0.762 089	0.759 891
KNN	*Precision*	0.903 034	0.900 613
	*F*1	0.884 905	0.881 996
XGBoost	*Precision*	0.969 037	0.971 590
	*F*1	0.968 777	0.971 299
AdaBoost	*Precision*	0.974 717	0.976 038
	*F*1	0.974 454	0.975 714
本文 Stacking 模型	*Precision*	0.978 706	0.981 172
	*F*1	0.978 555	0.981 078

由表 4 中也可以看出,本文 Stacking 集成模型相对于其他单一模型在性能上具有明显的提升。经过分析,Stacking 模型具有 2 层结构。第一层由不同基学习器组成可以综合不同模型的优势,提高模型的泛化能力。第二层的输入为上一层的输出结果,弥补第一层产生的误差,提高模型的预测精度。

2.2 Stacking 集成学习模型与其他模型性能分析

为了进一步说明本文模型的性能,将本文模型与 Voting 模型进行预测性能对比。不同模型的性能比较见表 5。为了保证对比实验的准确性,2 种模型使用相同数据并将 Voting 模型的基学习器与本文模型第一层基学习器保持一致。由表 5 可知,本文模型的性能好于 Voting 模型,因此,本文模型在脑卒中预测上具有良好的鲁棒性。

表 5 不同模型的性能比较

Table 5 Performance comparison of different models

模型	*Precision*	*F*1
Voting 模型	0.976 450	0.976 348
本文模型	0.981 172	0.981 078

3 结束语

由于脑卒中现阶段没有良好的治疗方法,所以提早预防至关重要。针对现有单一预测模型预测精度不高或者鲁棒性较差的问题,本文提出基于 RF-RFECV 和 Stacking 集成学习的脑卒中预测方法。实验结果表明,本文方法相比当前主流算法更加适用于脑卒中预测,具有更好的性能,证明本文方法的可行性。本文方法用于脑卒中预测与诊断,可以辅助医生提高脑卒中诊断效率。

参考文献

[1] 王陇德,彭斌,张鸿祺,等.《中国脑卒中防治报告 2020》概要[J]. 中国脑血管病杂志,2022,19(2):136-144.

[2] 刘冰,张艳,徐珏,等. 2014-2020 年杭州市居民脑卒中发病及死亡变化趋势分析[J]. 中国预防医学杂志,2023,24(7):726-731.

[3] 李洁洁,张雁儒,李昊,等. 机器学习在脑卒中预测中的研究进展[J]. 河南医学研究,2022,31(20):3832-3835.

[4] 常怀文,姚音. 基于机器学习构建江西地区缺血性脑卒中风险预测模型[J]. 西部医学,2022,34(8):1182-1186.

[5] 李鹏,闵慧,瞿昊宇,等. 基于逻辑回归模型的缺血性脑卒中发病率预测研究[J]. 医学信息学杂志,2020,41(6):28-32.

[6] 吴菊华,张烁,陶雷,等. 基于神经网络的脑卒中风险预测模型研究[J]. 数据分析与知识发现,2019,3(12):70-75.

[7] 于勤丽,于海征. 基于改进 SMOTE 自适应集成的信用风险评估模型[J]. 重庆理工大学学报(自然科学),2022,36(7):293-302.

[8] 张伟,王连彪,张广帅. 基于 RF-RFECV 和 PSO-SVM 的化工过程故障诊断方法[J]. 青岛科技大学学报(自然科学版),2022,43(5):101-108.

[9] 吴辰文,梁靖涵,王伟,等. 基于递归特征消除方法的随机森林算法[J]. 统计与决策,2017(21):60-63

[10] 艾成豪,高建华,黄子杰. 混合特征选择和集成学习驱动的代码异味检测[J]. 计算机工程,2022,48(7):168-176,198.