DOI:10.20033/j.1003-7241.(2025)12-0134-04

基于 Stacking 集成算法的医院财务数据智能分类研究

江 雨
(北京市第六医院,北京 100007)

摘 要:随着数据信息的膨胀式增长,医院财务管理面临巨大挑战。为了识别医院财务数据的异常情况,研究搭建初级分类器为随机森林、支持向量机、极端梯度提升树和 K 近邻模型,Logistic 回归为次级分类器的智能安全管理分类模型。结果表明,集成模型的分类器总性能最佳,为 94.02%。与极端梯度提升树模型对比,集成模型性能提升了 23%。基本收益作为明显特征,更为容易识别出医院财务数据异常。设计的模型具有较高的分类准确性和稳定性,这对于医院财务管理部门具有一定的理论意义和现实价值。
关键词:财务数据;异常识别;Stacking 算法;数据安全;智能分类;逻辑回归;随机森林;支持向量机
中图分类号:TP391.3;TN406.72 **文献标志码**:A **文章编号**:1003-7241(2025)12-0134-04

Intelligent classification of hospital financial data based on Stacking integration algorithm

JIANG Yu
(Beijing NO.6 Hospital, Beijing 100007, China)

Abstract: With the explosive growth of data information, hospital financial management is facing enormous challenges. In order to identify abnormal financial data in hospitals, this study constructs a primary classifier consisting of random forest, support vector machine, extreme gradient boosting tree, and K-nearest neighbor model, and Logistic regression as an intelligent security management classification model for secondary classifiers. The results show that the ensemble model's classifier had the best overall performance, at 94.02%. Compared with the extreme gradient boosting tree model, the integrated model improves performance by 23%. Basic income, as an obvious feature, makes it easier to identify anomalies in hospital financial data. The designed model has high classification accuracy and stability, which has certain theoretical significance and practical value for hospital financial management departments.
Keywords: hospital financial data; abnormal recognition; Stacking algorithm; data security; intelligent classification; logistic regression; random forest; support vector machine

0 引言

随着数字化程度的提高,财务数据信息在医院管理中的应用越来越广泛。Abeysinghe A 等[1]通过大数据技术,将各个部门的财务数据集成到一个平台上,实现全面的财务数据监控和分析,及时发现异常情况并采取相应措施。在此背景下,集成学习分类模型由于易于融合、强鲁棒性,以及适用于不同样本容量的优点,被广泛应用在金融、医疗和工业等领域。向峰等[2]为了预测复杂生产过程产品质量,通过改进随机森林(random forest, RF)算法和 Stacking 集成学习算法,解决误差累积,结果表明该方法准确性较高。而 Stacking 集成算法的优势在于其对不同基模型的兼容性,允许综合多种算法的优势以提升准确性。Kannan K 等[3]预测心血管疾病的风险因素,通过 Logistic 回归评估风险因子影响,并对比 K 近邻(K-nearest neighbor, KNN)算法、支持向量机(support vector machine, SVM)和 RF 等分类技术。Logistic 回归不仅计算简便,且具有良好的泛化能力,能够提供分类结果的概率解释,有助于防止过拟合。考虑到医院财务数据存在内部造假等重大挑战,研究样本选择各个医院的年度财务报表数据,搭建结合 Stacking 集成算法的医院财务数据智能安全管理分类模型。研究旨在识别异常财务数据,便于医院数据管理规范化,提高安全性能。

1 基于 Stacking 集成算法的医院财务数据分类模型

1.1 医院财务数据 Stacking 集成模型构建

医院财务数据安全管理分类直接关系到医院的资金运营和经营状况,对医院管理的各个方面都有着重要的影响[4]。作为一种机器学习方法,集成学习旨在通过组合多个基本模型的预测结果,从而提高模型的性能和泛化能力。因此,研究采用了 Stacking 集成算法,通过融合 RF、极端梯度提升(extreme gradient boosting, XGBoost)、KNN 和 SVM 四种初级分类器的预测结果,来执行终极分类任务。设计的 Stacking 算法模型架构,如图 1 所示。

* 基金项目:北京市科技计划资助项目(Z231100007423002)
收稿日期:2024-03-25

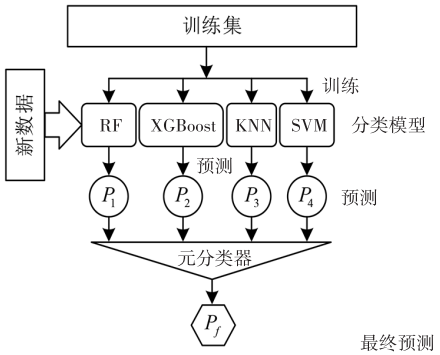


图 1 Stacking 算法模型架构

图 1 中,输入原始训练集,训练四种分类模型 RF、XGBoost、KNN 和 SVM。之后以每一个分类模型的训练集为完整的原始训练集,以分类模型在原始测试集上的预测为测试集。合并获得输出相对应的 4 个预测结果,以此为新训练集对元分类器进行训练。RF 是以决策树作为基础学习器的计算方法,能够处理不平衡的数据集和缺失值,评估特征的重要性,并具有高效性。其利用决策树法使得一组随机生长的决策树 $\{h(X,\theta_k),k=1,2,\cdots,\}$,生成大量决策树。之后对各决策树的评估,选择最佳结果完成预测。其中,二叉树构建的特征指标为 Gini 指数[5]。选择最小的 Gini 指数的特征生成二叉树,表达式为

$$\mathrm{Gini}(p)=\sum_{k-1}^{k}p_k(1-p_k) \tag{1}$$

式中,p 为样本数据。训练集随机从输入向量 $\boldsymbol{X}$ 中选取,边际函数为

$$\mathrm{mg}(\boldsymbol{X},Y)=\mathrm{av}_k I(h_k(X)=Y)-\max_{j\neq Y}\mathrm{av}_k I(h_k(X)=j) \tag{2}$$

式中,av_k 为 $\boldsymbol{X}$ 中正确分类 Y 的平均数,$I(h_k(X)=j)$ 为指示函数,j 为 $\boldsymbol{X}$ 中的随机分类。RF 泛化误差定义有

$$\mathrm{PE}^{*}=P_{\boldsymbol{X},Y}(\mathrm{mg}(\boldsymbol{X},Y)<0) \tag{3}$$

XGBoost 可以实现自动学习,处理稀疏数据和缺失值,这使得其在处理实际、复杂的数据时具有优势。为了避免过拟合,XGBoost 在传统的梯度提升框架基础上,加入正则化项,以此来控制树的复杂度和叶节点权重的惩罚项,并降低方差。该方法的数学描述和目标函数为

$$\begin{cases}\hat{y}_i=\sum_{k}^{K}f_k(x_i)\\ \mathrm{Obj}(\theta)=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k}^{K}\Omega(f_k)\end{cases} \tag{4}$$

式中,x_i 为第 i 个样本,n 为样本个数。K 为 CART 树的数量,f_k 为第 k 个 CART 树,y_i 和 $\hat{y}_i$ 分别为样本 i 的真实值和预测值。$\Omega(f_k)$ 为正则项,$L(y_i,\hat{y}_i)$ 为损失函数。KNN 在特征空间中,如果一个样本附近的 k 个最近,即特征空间中最邻近样本的大多数属于某一个类别,则该样本也属于这个类别[6]。XGBoost 在训练过程中支持并行计算,能够在多核 CPU 作用下加速。其在选择分割点时,引入贪心算法,利用比较分割前后的目标函数值,来确定最优的分割点。KNN 首先计算测试样本和训练样本中每个样本点的距离,标准化欧氏距离公式为

$$d=\sqrt{\sum_{k=1}^{n}\left(\frac{x_{1k}-x_{2k}}{s_k}\right)^2} \tag{5}$$

接着,对所有的距离值 d 进行排序,选前 k 个最小距离的样本,按照 k 个样本的标签进行投票,得到最后的分类类别。SVM 在分类问题中给定输入数据 $\boldsymbol{X}$ 和学习目标,若输入数据所在的特征空间存在作为决策边界的超平面,将学习目标按正类和负类分开,并使任意样本的点到平面距离大于等于 1[7-8]。超平面表达为

$$w^{\mathrm{T}}+b=0 \tag{6}$$

式中,参数 $\boldsymbol{w}$ 和 b 分别为超平面的法向量和截距,T 为超平面的转置。其中核函数选择 Sigmoid 函数,参数 $\eta>0$,$\theta<0$,tanh 为双曲正切函数[9]。核函数表达关系为

$$k(x,x_i)=\tanh(\eta<x,x_i>+\theta) \tag{7}$$

Stacking 集成算法可以充分通过 RF、XGBoost、KNN 和 SVM 四个模型的特点和优势,使得模型的泛化性和预测精度提升,并获得协同效应的效果[10]。

1.2 次级分类器

Stacking 模型的初级分类器确定之后,次级分类器选择 Logistic 回归。Logistic 回归利用给定的 n 组数据,即训练集来训练模型,在结束训练之后分类给定的一组或多组数据,即测试集,而 p 个指标构成每一组数据。Logistic 适合二分类问题,简单易于理解,速度快,可以直接看到各个特征的权重,并能够容易地更新模型吸收新的数据[11-12]。其一般步骤为寻找预测函数,完成损失函数构造,并将损失函数最小化,最终求得回归参数 θ。Logistic 函数形式为

$$\mathrm{Logistic}(z)=\frac{1}{1+e^{-z}} \tag{8}$$

对于线性边界的情况,边界形式为

$$z=\theta^{\mathrm{T}}x \tag{9}$$

预测函数 $h_\theta(x)$ 为

$$h_\theta(x)=g(\theta^{\mathrm{T}}x)=\frac{1}{1+e^{-\theta^{\mathrm{T}}x}} \tag{10}$$

Logistic 回归中,在每个特征上都乘以一个回归系数,然后将所有值累加,将总和代入 Sigmoid 函数中,得到一个范围在[0,1]的数值,小于 0.5 归为 0 类,等于 0.5 归为 1 类。将 $h_\theta(x)$ 看作一种概率,y_θ 为样本类别[13]。数据集包含如财务报表、患者账单信息、供应链成本、员工薪酬等主要方面。为了提高模型收敛速度,研究选择对原始数据进行 min-max 标准化处理[14]。min-max 标准化本质上将原始数据 x 参与线性变换,使得数组值映射至[0,1],转换后的数据 x^{*} 表达关系为

$$x^{*}=\frac{x-\min}{\max-\min} \tag{11}$$

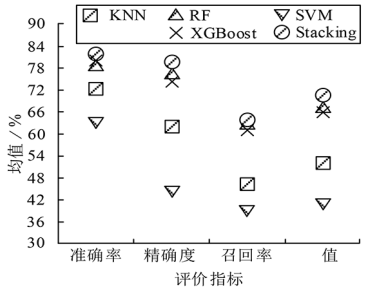
2 Stacking 集成分类模型结果分析

研究样本选择 2002 年~2022 年的医院年度财务报表数据，数据源自当地省级卫健委和医院公开发布。研究在 Jupyter Note book 平台进行，分析工具为 Python。实验环境配置，如表 1 所示。

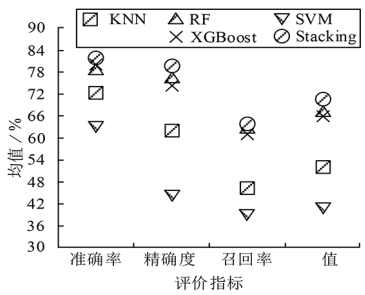
表 1 实验环境配置

类别	配置
操作系统	64-bit Windows 10
硬盘	931 G
安装内存	4.00 GB
处理器	Intel（R）Core（TM）i5-4590
Python 编程环境	PyCharmPython 3.7.4
R4.1.0	RStudio

实验指定学习器的数量为 150，为了避免过拟合，学习率为 0.3，gamma 值为 0。同时，考虑到 Stacking 集成模型的交叉验证，用 python 中 shuffleSplit( ) 函数分割 20%的测试集和 80%的训练集，以便模型训练和评估。为了验证设计的 Stacking 模型有效性，将其对比不同分类模型，模型准确率、精确度、召回率和 F 值的均值和方差结果对比，如图 2 所示。



（a）均值对比



（b）方差对比

图 2 不同模型性能对比

图 2（a）中，Stacking 模型的准确率、精确度、召回率和 F 值的均值均为最高，分别为 81.76%、79.76%、63.46%和 70.39%。这表示其可以准确地识别出异常财务数据，有助于及时发现和预防潜在的财务风险或不当行为，保障医院资金的安全运营。XGBoost 模型准确率为 80.00%，仅次于 Stacking 模型，对比之下 Stacking 模型提升了 2.2%，性能更好。图 2（b）中，SVM 模型评价指标的方差最大，表示其对训练数据过拟合，在未知数据上的泛化能力较差。Stacking 模型的召回率方差最低，仅为 7.30%，对比 KNN 模型的 9.04%，减少了 01.74%。为了评估 Stacking 模型性能，对比不同模型的受试者工作特征（receiver operating characteristic，ROC）曲线对比，如图 3 所示。

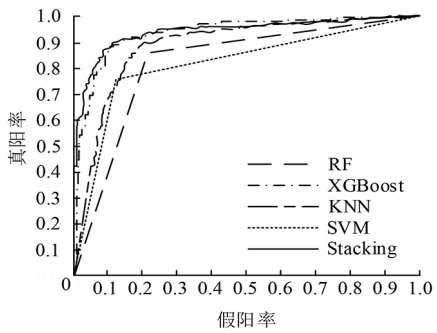


图 3 不同模型的 ROC 曲线对比

图 3 中，Stacking 模型曲线下面积（area under the curve，AUC）最大，为 0.94，因此该模型的分类器总性能最佳。与 XGBoost 模型对比，Stacking 模型性能提升了 23%。SVM 模型的 AUC 最小，为 0.81。RF、KNN 和 XGBoost 模型的分类程度分别为 82.01%、88.85%和 93.79%。由于 XGBoost 模型性能仅次于 Stacking 模型，对其进行特征重要性分析。特征排序结果，如图 4 所示。

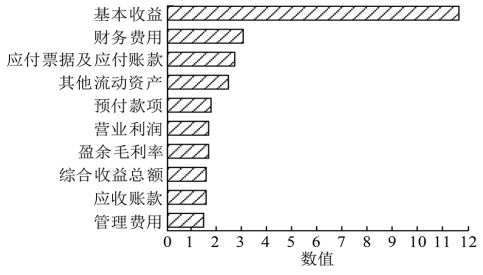
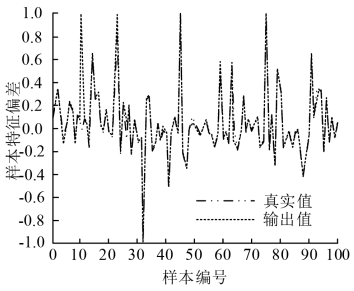


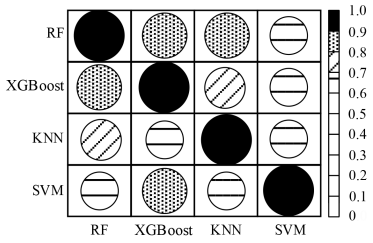
图 4 特征重要性分析

图 4 中，由于模型输出结果有多个特征，仅针对优先排序特征分析，数据集中基本收益、管理费用、财务费用、其他流动资产、综合收益等特征是判断结果的关键因素。基本收益重要性最大，为 11.67%。当基本收益处于异常，则容易作为明显特征识别财务数据异常。其次为财务费用，为 3.08%。由于设计的 Stacking 模型涉及多种模型，因此在评估两个模型之间的相关性时，采用 Pearson 相关性检验[15]。关于 RF、KNN、XGBoost 和 SVM 方法的参数，在优化之后，对比调参前后的模型效果。同时为了评估初级分类器之间的相关性，研究将 RF、XGBoost、KNN 和 SVM 的输出值，进行 Pearson 相关性分析，结果如图 5 所示。

图 5（a）中，调参前后的样本特征真实值和输出值偏差相差不大，均方根分别为 0.24 和 0.23，对比之下调参后的模型输出值偏差更小，效果更佳。图 5（b）中，XGBoost 与 RF、SVM 模型，KNN 与 RF 模型的 Pearson 相关系数均大于 0.80，表示具有极强相关性。其余模型对比之下的 Pearson 相关系数大于 0.60，表示他们之间具有强相关性。这四种模型能够捕捉到数据的关键特征，并且有助于提升模型对未见数据的泛化能力。研究选择这四种模型作为初级选择器，存在一定的合理性。



（a）调参前后对比



（b）Pearson 相关性检验

图 5　改进模型调参前后和 Pearson 相关性检验结果

3　结束语

随着医院财务数据量急剧增加，为了提高财务数据管理效率，研究结合 Stacking 集成算法，选择 Logistic 回归为次级分类器。结果表明，Stacking 模型在召回率方差上最低，为 7.30%，较 KNN 模型降低了 1.74%，展示出更高的稳定性。同时，XGBoost 与 RF 模型间 Pearson 相关系数为 0.862，存在极强相关性。Stacking 模型的准确率、精确度、召回率和 F 值的均值均为最高，分别为 81.76%、79.76%、63.46%和 70.39%，其准确率比 XGBoost 模型高 2.2%。因此，构建的 Stacking 集成模型不仅提高了数据处理的准确性和效率，还对提升医院整体财务管理水平具有重要意义。但研究仅考虑了医院数据，未来可以针对其他行业的数据进行相关研究。

参考文献

[1] ABEYSINGHE A, DE ZOYSA M T R, SAMUDITHA K M Y, et al. Security Operation Center for Healthcare Sector[J]. I nternational Research Journal of Innovations in Engineering and Technology, 2023, 7(11):299.

[2] 向峰, 杨磊, 张萌, 等. 基于模型融合的复杂生产过程产品质量预测[J]. 中国科学:技术科学, 2023, 53(7):1127-1137.

[3] KANNAN K, MENAGA A. Risk factor prediction by naive bayes classifier, logistic regression models, various classification and regression machine learning techniques[J]. Proceedings of the National Academy of Sciences, India Section B:Biological Sciences, 2022, 92(1):63-79.

[4] 赖尉文, 贺维. 一种基于自适应证据推理规则的集成学习方法[J]. 计算机应用研究, 2023, 40(8):2281-2285.

[5] 李苗. 基于数据分类的企业财务数据异常判定方法[J]. 自动化技术与应用, 2023, 42(10):91-94.

[6] 马宗方, 马祥双, 宋琳, 等. 异常信息的智能分类算法研究[J]. 计算机测量与控制, 2021, 29(10):164-169.

[7] 高媛媛. 基于多特征融合和机器学习的疾病基因检测大数据分类模型[J]. 微型电脑应用, 2023, 39(3):25-27, 39.

[8] 朱建霞. 基于聚类算法的海量医院财务数据精准分类方法[J]. 自动化技术与应用, 2023, 42(4):79-82.

[9] 孙丽娟, 徐伟, 胡艺宸. 基于加权 KNN 的医院财务信息自动分类系统[J]. 自动化技术与应用, 2022, 41(11):92-95.

[10] 曲福恒, 宋剑飞, 杨勇, 等. 基于 min-max 准则与区域划分的 I-k-means-+聚类算法[J]. 吉林大学学报:理学版, 2023, 61(5):1131-1138.

[11] 杨晶东, 李熠伟, 江彪, 等. 逐层 Transformer 在类别不均衡数据的应用[J]. 计算机应用研究, 2023, 40(10):3047-3052.

[12] 李昂, 韩萌, 穆栋梁, 等. 多类不平衡数据分类方法综述[J]. 计算机应用研究, 2022, 39(12):3534-3545.

[13] 黄晞. 基于 AHP-BPNN 的医疗监控数据信息化管理模式识别[J]. 自动化技术与应用, 2023, 42(8):165-169.

[14] 谷今一, 王梦莹, 奚蓓蓓, 等. 基于 HRP 系统的公立医院运营管理内部控制设计应用[J]. 中国医院管理, 2022, 42(11):76-78.

[15] 张永刚, 吕鹏飞, 张悦, 等. 基于 Stacking 集成学习的恶意 URL 检测系统设计与实现[J]. 现代电子技术, 2023, 46(10): 105-109.

作者简介:江　雨(1993—),女,本科,中级会计师,研究方向:财务管理,财务数据。

(上接第 97 页)

[7] MOKHTARI S A, SABZEHPARVAR M. Application of hilbert-huang transform with improved ensemble empirical mode decomposition in nonlinear flight dynamic mode characteristics estimation [J]. Journal of Computational and Nonlinear Dynamics, 2019, 14(1):011006

[8] LIU Q J, ZHANG H Y, GAO K T, et al. Time-frequency analysis and simulation of the watershed suspended sediment concentration based on the Hilbert-Huang transform (HHT) and artificial neural network (ANN) methods: A case study in the Loess Plateau of China [J]. Catena, 2019 (179):107-118.

[9] ALBIRUNI F, CHO Y, LEE J H, et al. Non-contact guided waves tomographic imaging of plate-like structures using a probabilistic algorithm[J]. Materials Transactions, 2012, 53(2):330-336.

[10] CHENG X, CAO X, WU Z, et al. A flexible conformal piezoresistive sensor based on electrospinning for deformation monitoring of carbon fiber-reinforced polymer [J]. Advanced Engineering Materials, 2023, 25(19):2300341.

[11] 李正, 杨庆生, 尚军军, 等. 面内随机堆叠石墨烯复合材料压阻传感机理与压阻性能[J]. 力学学报, 2020, 52(6):1700-1708.

[12] 郝罗亮, 毛小鑫, 李申宝, 等. 基于时域统计的压力传感器自动补偿方法[J]. 自动化技术与应用, 2023, 42(2):64-66, 93.

[13] YANG H, YUAN L, YAO X, et al. Piezor-esistive response of graphene rubber composites considering the tunneling effect [J]. Journal of the Mechanics and Physics of Solids, 2020 (139):103943.

[14] YANG J. Resolution-lossless ultrasound tomography for health monitoring of composite structures: from nanocomposite sensor network development tomachine learning-enabled imaging [D]. Hong Kong: The Hong Kong Polytechnic University, 2022.

作者简介:郑明铭(1999—),男,硕士研究生,研究方向:复合材料无损检测方面。