文章编号：2095-2163(2023)03-0231-05 中图分类号：TP181 文献标志码：A

基于机器学习的 SAE 患者 30 天死亡风险预测模型

刘 彬[1]，肖晓霞[1,2]，龚后武[3]，周 展[1]，郑立瑞[1]，谭建聪[1]
（1 湖南中医药大学 信息科学与工程学院，长沙 410208；2 湖南中医药大学 中医学国内一流建设学科，长沙 410208；
3 东华医为科技有限公司，北京 100089）

摘 要：脓毒症相关性脑病（SAE）是指在患脓毒症过程中发生的脑功能障碍，其与脓毒症患者短期死亡率的上升有关。本文从 MIMIC 数据库中抽取相关的脓毒症患者数据，其中 SAE 被定义为患脓毒症且 *GCS* 分数小于 15 分。使用 RFE 算法筛选出影响 SAE 患者 30 天死亡率的危险因素，对 SAE 病例数据采用逻辑回归、GBDT、XGBoost 三种算法建立 30 天死亡风险预测模型。实验结果表明，GBDT 算法的预测效果优于另外 2 种算法，其准确率为 78.6%，*AUC* 为 78.3%，该模型能够对 SAE 患者 30 天死亡情况进行较为准确的预测。

关键词：脓毒症；脓毒症相关性脑病；MIMIC 数据库；逻辑回归；随机森林

30-day mortality risk prediction model for SAE patients based on machine learning

LIU Bin[1], XIAO Xiaoxia[1,2], GONG Houwu[3], ZHOU Zhan[1], ZHENG Lirui[1], TAN Jiancong[1]
(1 School of Information Science and Engineering, Hunan University of Chinese Medicine, Changsha 410208, China;
2 The Domestic First-class Discipline Construction Project of Chinese Medicine, Hunan University of Chinese Medicine, Changsha 410208, China; 3 DHC Mediway Technology Co., Ltd., Beijing 100089, China)

【Abstract】 Sepsis related encephalopathy (SAE) refers to brain dysfunction occurring in the course of sepsis, which is related to the rise of short-term mortality in sepsis patients. In this paper, the data of sepsis patients are extracted from the MIMIC database, where SAE is defined as having sepsis and *GCS* score is less than 15. The RFE algorithm is used to screen out the important factors affecting the 30 day mortality of SAE patients, and the logistic regression, GBDT, XGBoost are used to establish the 30 day mortality risk prediction model for SAE patients. The experimental results show that the prediction effect of GBDT algorithm is better than other algorithms, with an accuracy of 78.6% and an *AUC* of 78.3%. This model can accurately predict the 30 day mortality of SAE patients.

【Key words】 sepsis; SAE; MIMIC database; logistic regression; Random Forest

0 引 言

脓毒症是由感染引起的全身炎症反应综合征，全球发病率较高，每年患脓毒症的人数约为 3 100 万，住院病死率约为 17%[1]。脓毒症相关性脑病（SAE）是指在患脓毒症过程中发生的脑功能障碍，是一种比较严重的脓毒症并发症，也是造成脓毒症患者死亡的独立危险因素[2]。并与人体行为、记忆、认知功能的长期损害密切相关，给患者的家庭和社会带来沉重的经济负担。仍需指出的是，SAE 患者的死亡率往往高于只患脓毒症的患者。格拉斯哥昏迷评分法（Glasgow Coma Scale，*GCS*）是一种用来评估病人昏迷程度的方法，满分为 15 分[3]，表示意识清楚；12~14 分表示轻度意识障碍；9~11 分表示中度意识障碍；8 分以下为昏迷。Eidelman 等学者[4]的研究表明脑病与医院死亡率的增加成正相关性，当格拉斯哥昏迷评分（*GCS*）为 15 分时，死亡率为 16%，而当 *GCS* 分数为 3 到 8 分时，死亡率为 63%。Sonneville 等学者[5]的研究也得出了类似的结论，研究显示当 *GCS* 分数为 15 时，患者 30 天生存率为 67%；当 *GCS* 分数为 3~8 分时，30 天生存率下降到 32%。即使发生轻度意识障碍（*GCS* 分数为

基金项目：2017 年科技部十三五重点研发计划（2017YFC1703300）；大规模跨模态医疗知识管理。

作者简介：刘 彬（1997-），男，硕士研究生，主要研究方向：数据挖掘、自然语言处理；肖晓霞（1977-），女，博士，副教授，主要研究方向：中医智能诊断、人工智能、生物医学工程等。

通讯作者：肖晓霞 Email：amily_x@ hnucm.edu.cn

收稿日期：2022-10-17

12~14)也是影响 30 天死亡的一个独立危险因素。综上表明,SAE 对于脓毒症患者短期死亡率的增加是有影响的,而这将进一步影响患者的健康,同时加重医疗资源的消耗。

基于上述问题,识别出短期死亡率较高的 SAE 患者,有利于及时进行医疗干预,对于改善这类患者的预后也具有重要的意义。因此本研究的主要目的是通过大型的临床数据库 MIMIC 去提取相应的 SAE 患者数据,然后通过 rfe 算法[6]对相应特征进行筛选,选出影响 SAE 患者 30 天死亡率的重要特征,最后基于这些特征构建机器学习模型,用于改善 SAE 患者的预后。

1 算法原理

1.1 RFE 特征筛选

特征递归消除(Recursive Feature Elimination, RFE)是一种用来衡量特征变量重要性的方法,通过重复构建模型,逐步迭代选出最重要的特征变量,能够寻找出最优的特征子集,剔除不重要的特征变量。具体运算步骤如下:

(1)设定需要进行选择的特征数。

(2)选择一个基模型来进行多轮训练,每次训练将 $J(k)=(w_k)^2$ 作为每个特征的排序准则,并且每次迭代去除排序最后需要移除的特征数量。

(3)基于新的特征集进行下一轮训练,直至特征个数为特征设定值。

本文选择的基模型为 XGBoost 模型,对总计 17 个特征进行筛选。

1.2 逻辑回归

逻辑回归[7]是一种广义的线性回归模型,属于机器学习中的监督算法,主要是用来解决二分类问题。该算法首先通过输入数据拟合出一条直线 $z=\boldsymbol{w}^{\mathrm{T}}x+\boldsymbol{b}$,显然这样的函数图像是一条斜线,难以达到最终想要的结果(0 或 1),于是要将 z 通过一个函数映射成 0 ~ 1 之间的数,这个函数就是 *sigmoid* 函数,式子如下:

$$g(z)=\frac{1}{1+e^{-z}} \tag{1}$$

然后,通过极大似然估计推导出损失函数:

$$J(\boldsymbol{w})=\min\left(-\frac{1}{n}\sum_{i=1}^{n}\left[y_i(\boldsymbol{w}^{\mathrm{T}}x+\boldsymbol{b})-\ln(e^{w^{\mathrm{T}}x+b}+1)\right]\right) \tag{2}$$

最后,通过梯度下降法求解出式(2)中的参数,从而解决了二分类问题。

1.3 GBDT

GBDT(Gradient Boosting Decision Tree)是一种基于决策树的集成算法。算法采用将基函数线性组合的方法[8],在训练过程中使得残差不断地减小,最终实现数据回归或者分类。GBDT 算法的训练过程具体如图 1 所示。

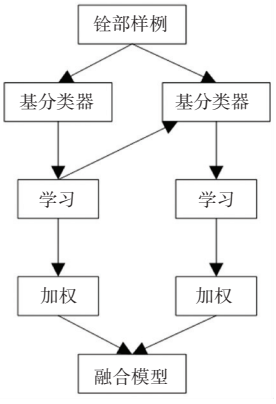

图 1 GBDT 算法训练过程

Fig. 1 GBDT algorithm training process

GBDT 通过多轮迭代,产生多个弱分类器,每个分类器在上一轮分类器的梯度(如果损失函数是平方损失函数,则梯度就是残差值)基础上进行训练。弱分类器一般会选择 CART TREE(分类回归树),这种树具有结构简单、高偏差、低方差的特点,因此十分适合用于 GBDT 算法的训练中。

1.4 XGBoost

XGBoost 算法[9]是在 GBDT 算法的基础上发展而来的,主要改进有:算法不仅可以使用 CART 分类回归树,还能使用线性基础模型;在目标函数中加入了正则化项,用来防止模型出现过拟合;借鉴了随机森林的原理,支持列抽样,不仅能降低过拟合,还能够减少模型的计算量;考虑到了训练数据为稀疏值的情况,能为缺失值指定分支的默认方向,从而提高算法效率。

2 数据与方法

2.1 数据来源

MIMIC[10](Medical Information Mart for ICU)是一个大型的、免费提供的数据库,其中包括来自美国马萨诸塞州波士顿贝斯以色列女执事医疗中心重症监护病房住院病人的高质量健康相关数据,数据包括生命体征、药物、化验数据、护理人员的观察和记录、输液、手术、诊断代码、成像报告、住院时间、生存数据。MIMIC 数据库到现在已经发布 4 个版本。MIMIC-II 中包含 2001 ~ 2008 年的数据,MIMIC-Ⅲ

包含 2001～2012 年的数据，MIMIC-IV 包含 2008～2019 年的数据。本文将基于 MIMIC-IV 数据库抽取相应的 SAE 患者数据。

2.2　数据抽取

SAE 被定义为脓毒症患者中 *GCS* 分数小于 15 的患者。研究使用的主要软件为 Navicat Premium (15.0.12 版本)，按照关键字[11]“s-epsis”、“severe sepsis”、“septic shoc-k”从数据库中搜索被诊断为“脓毒症”、“严重脓毒症”、“脓毒症休克”患者的原始数据。根据以往研究，确定好纳排标准后进一步筛选患者。患者筛选的详细过程如图 2 所示。

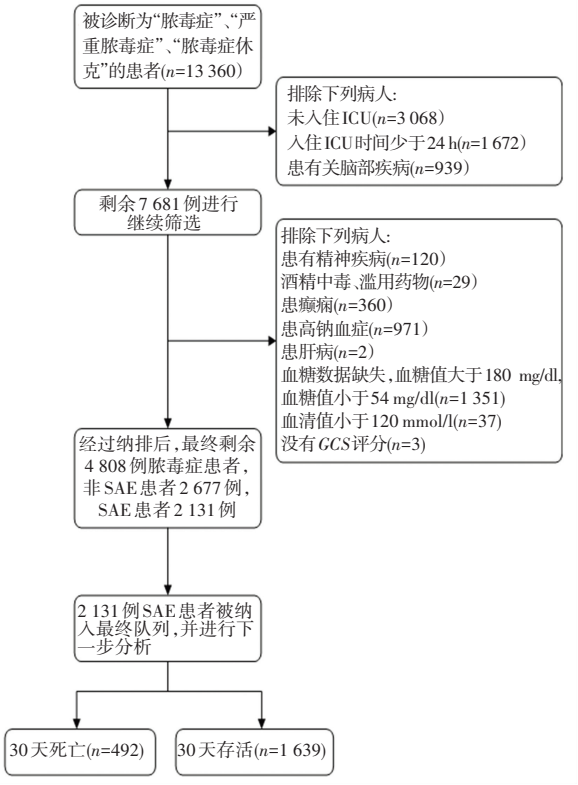

图 2　患者筛选图

Fig. 2　Patient screening

确定最终的 SAE 患者后，根据此前的研究文献，从 MIMIC 数据库中提取患者首次入院时对应的年龄(anchor_age)、性别(gender)、住院天数(day)、葡萄糖(glucose)、钠(sodium)、GCS 分数(gcs)、血小板(platelet)、肌酐(creatinine)、血红蛋白(hemoglobin)、钾(potassium)、血尿素氮(BUN)、白细胞(WBC)、乳酸盐(lactate)、血浆凝血酶原时间(PT)、心率(heart_rate)、血氧饱和度(spo2)、呼吸速率(respiratory_rate)、30 天是否死亡(morality)。数据总计 17 个特征属性，再加一个类别标签属性，其中类别标签表明患者是否在患病 30 天内死亡。

2.3　数据预处理

提取了数据后，对数据的缺失情况进行统计，结果见表 1。

表 1　数据缺失情况表

Tab. 1　Data missing table

特征	缺失数	缺失比例/%
血小板	6	0.12
肌酐	1	0.02
血红蛋白	5	0.10
血尿素氮	3	0.06
白细胞	6	0.12
血氧饱和度	13	0.27
乳酸盐	954	19.84
凝血酶原时间	365	7.59
心率	9	0.18
呼吸频率	12	0.25

从表 1 的结果中可以看出 10 个特征存在数据缺失的问题，缺失最多的特征是乳酸盐，缺失比例为 19.84%，缺失最少的是肌酐，仅缺失一例。根据文献[8]中对缺失数据的处理方法来看，缺失特征比例均小于 20%，予以保留，并统一采用平均值对其进行填补，在此基础上将对数据进行具体分析。

3　结果

3.1　纳入病例的基本信息

总计纳入 4 808 例脓毒症患者，其中 2 131 例为 SAE 患者。SAE 患者年龄为 19～91 岁之间，中位年龄数为 68 岁。男性为 1 127 例，女性为 1 004 例。30 天内死亡病例为 492 例，存活病例为 1 639 例，数据分布较为均衡。

3.2　筛选得到的特征变量

根据 RFE 特征筛选，每一轮筛选移去特征系数 $(w_k)^2$ 最小的特征，直到特征个数为设定值。结果显示，当特征数设定为 13 时，3 个模型中 GBDT 的 *AUC* 值最高，其在测试集上 *AUC* 为 0.783。此时选出的 13 个特征分别为：年龄、住院天数、钠、GCS 分数、血小板、肌酐、钾、血尿素氮、乳酸盐、血浆凝血酶原时间、血氧饱和度、心率、呼吸速率。

3.3　实验结果

将 SAE 数据集按照 7∶3 的比例随机划分为训练集和测试集进行训练。本文采用的评价指标为准确率、*P* 值、*R* 值、F_1 值、*AUC* 值。具体的实验结果见表 2、表 3。

表 2 未进行特征筛选结果

Tab. 2 No feature filtering results

算法	*Accuracy*	*Precision*	*Recall*	F_1 *Score*	*AUC*
LR	76.1	43.0	23.9	30.8	72.2
XGBoost	77.7	49.5	33.8	40.2	75.5
GBDT	79.1	54.7	33.1	41.2	77.4

表 3 特征筛选后结果

Tab. 3 Results after feature screening

算法	*Accuracy*	*Precision*	*Recall*	F_1 *Score*	*AUC*
LR	77.7	49.3	23.9	32.2	72.5
XGBoost	78.3	51.6	34.5	41.4	73.8
GBDT	78.6	52.9	31.7	39.6	78.3

从表 2 和表 3 中可以看出,数据集经过特征筛选后,3 个模型的某些指标得到了提高。逻辑回归模型的准确率提高了 1.6%、精度提高了 6.3%、F_1 值提高了 1.4%、*AUC* 值提高了 0.3%;XGboost 模型的准确率提高了 0.6%、精度提高了 2.1%、召回率提高了 0.7%、F_1 值提高了 1.2%;GBDT 模型的 *AUC* 值提高了 0.9%。

为了更直观地比较 3 个不同算法的性能,绘制的 *ROC* 曲线如图 3 所示。

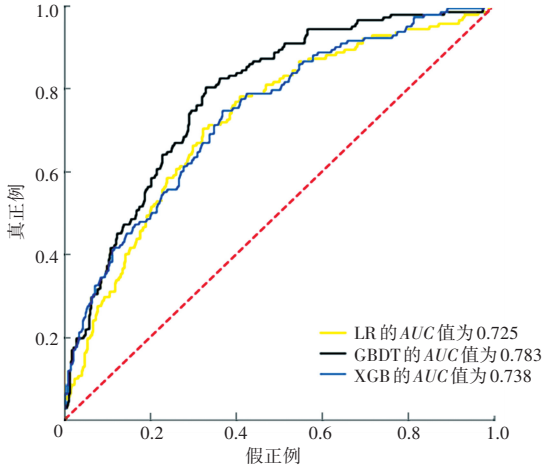

图 3 3 种分类算法的 *ROC* 曲线

Fig. 3 *ROC* curves of three classification algorithms

从图 3 中可以看出,在 3 个算法中 GBDT 算法的 *AUC* 值最大、为 0.783,说明 GBDT 算法性能最优,更适合用于 SAE 患者 30 天死亡预测。

4 分析与讨论

在这项基于 MIMIC-IV 数据库的研究中,从 MIMIC 数据库中抽取出对应的 SAE 患者数据,然后使用了 RFE 特征选择,筛选出了与 SAE 患者 30 天死亡率相关的危险因素,最后基于这些特征建立了 3 个机器学习模型去对 SAE 患者 30 天死亡进行预测。其中,GBDT 算法对于 SAE 患者 30 天死亡预测效果最佳,其精度为 52.9%,准确率为 78.6%、*AUC* 值为 78.3%,3 个指标均为不同算法中最高的。与其它研究方法进行对比,文献[3]提出的列线图模型在训练集上的 *AUC* 值为 0.763,在验证集上的 *AUC* 值为 0.753,均比本文提出的 GBDT 算法的 *AUC* 值略低。说明本文提出的模型性能更优、泛化能力也更强。目前,对于 SAE 的治疗是具有挑战性的,有许多关于脓毒症的指南列出了各种治疗脓毒症的建议,但却很少有治疗 SAE 的建议。有关 SAE 患者死亡预测的研究也较为匮乏,本研究很好地弥补了这方面的空白。从应用价值来看,本文提出的 GBDT 预测模型能够辅助临床医生去评估 SAE 患者的预后,从而制定出相应的治疗措施,降低患者死亡率。一旦研究出针对 SAE 的具体治疗方法,该模型的应用价值就会更高。未来可以开发一款能嵌入电子医疗系统的软件,该软件能够在不增加临床医生工作时间和负担的情况下,辅助临床医生及时治疗 SAE。

5 结束语

本文基于 MIMIC 数据库,提取相应的脓毒症患者数据,并通过 *GCS* 分数进一步筛选出 SAE 患者的数据。然后经过 RFE 特征筛选,筛选出 13 个重要的特征。使用逻辑回归、XGBoost、GBDT 三种算法基于筛选后的特征进行建模,实验结果表明,GBDT 算法更适合用于 SAE 患者 30 天死亡预测,其 *AUC* 值为 78.3%,高于其他 2 种算法,也比其他文献中的方法略好。对于 SAE 患者的预后具有一定的参考价值。

本次研究也存在局限性,即只对该数据库进行了内部验证,在今后的研究中还需要根据其它的数据进行外部验证,以进一步检验模型的鲁棒性和性能。

参考文献

[1] FLEISCHMANN C, SCHERAG A, ADHIKARI N K, et al. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations[J]. American Journal of Respiratory And Critical Care Medicine, 2016, 193(3):259-272.

[2] 周艺蕉,杨春燕,苏美仙. 脓毒症相关性脑病脑功能监测的研究进展[J]. 医学综述,2022(16):3246-3251.

(下转第 241 页)