DOI:10.16016/j.2097-0927.202310052

基于 SMOTE 算法和机器学习模型建立原发性肝癌术后的预后预测模型

潘 比*,余靖华*,黄译贤,伍亚舟,李 芳 400038 重庆,陆军军医大学(第三军医大学)军事预防医学系军队卫生统计学教研室

[摘要] **目的** 基于合成少数类过采样技术(synthetic minority over-sampling technique, SMOTE)算法和机器学习模型构建原发性肝癌术后的预后预测模型。**方法** 选取美国国立癌症研究所的监测、流行病学及最终结果(Surveillance, Epidemiology, and End Results, SEER)数据库中 4 297 例患者进行回顾性队列研究,通过独热编码和平均值插补法进行数据预处理,利用 SMOTE 算法解决数据类别不平衡问题,将临床变量纳入机器学习模型,基于决策树(decision tree, DT)、随机森林(random forest, RF)、梯度提升决策树(gradient boosting decision tree, GBDT)、极限梯度提升算法(eXtreme Gradient Boosting, XGBoost)方法构建预后预测模型(SMOTE+DT/RF/GBDT/XGBoost),通过比较多种模型的性能,筛选出最佳的预测模型。**结果** 组合模型 SMOTE+RF 展示出最优的预测性能,受试者工作特征曲线(receiver operating characteristic curve, ROC)下的面积(area under the curve, AUC)、准确率和精确率均高于其他模型,分别为 0.895、0.811、0.806。**结论** 基于 SMOTE+RF 算法的原发性肝癌的预后预测模型可有效预测原发性肝癌患者的生存结局。

[关键词] 原发性肝癌;少数类过采样技术算法;机器学习;预测模型
[中图法分类号] R319; R730.7; R735.7 [文献标志码] A

Construction of postoperative prognostic model for primary liver cancer based on SMOTE and machine learning

PAN Bi, YU Jinghua, HUANG Yixian, WU Yazhou, LI Fang Department of Health Statistics, Faculty of Military Preventive Medicine, Army Medical University(Third Military Medical University), Chongqing, 400038, China

[**Abstract**] **Objective** To construct a prognosis prediction model of primary liver cancer after surgical treatment based on synthetic minority over-sampling technique (SMOTE) algorithm and machine learning model. **Methods** A retrospective cohort study was conducted on 4 297 patients with primary liver cancer from the surveillance, epidemiology, and end results (SEER) database. One-Hot Encoding and Multiple Imputation were used to preprocess the collect data, and SMOTE algorithm was employed to solve the imbalance of data categories. The obtained clinical variables were included in the machine learning model. Based on decision tree (DT), random forest (RF), gradient boosting decision tree (GBDT) and eXtreme Gradient Boosting (XGBoost), a prognostic prediction model (SMOTE+DT/RF/GBDT/XGBoost) was build, and then the best prediction model was determined by comparing the performance of various models. Finally, a prognostic analysis system for primary liver cancer was developed based on the optimal model, which was then visualized. **Results** The combination model SMOTE+RF showed the best predictive performance, with higher area under the curve (0.895), accuracy (0.811) and precision (0.806) than those of other models in receiver operating characteristic curve (ROC) analysis. **Conclusion** The SMOTE+RF prognostic prediction model can effectively predict the survival outcome of patients with primary liver cancer.

[**Key words**] primary liver cancer; SMOTE algorithm; machine learning; prediction model

PAN Bi and YU Jinghua contributed equally to the article. Supported by the General Program of National Natural Science Foundation of China (82173621, 81872716). Corresponding author: LI Fang, E-mail: oucfli@163.com

* 同等贡献者

[基金项目] 国家自然科学基金面上项目(82173621, 81872716)

[通信作者] 李 芳,E-mail:oucfli@163.com

2097-0927/Copyright © 2024, 陆军军医大学(第三军医大学)。这篇开放获取文章遵循 CC BY 许可协议(https://creativecommons.org/licenses/by/4.0/)

原发性肝癌(primary liver cancer,PLC)是指发生在肝细胞或肝内胆管上皮细胞的恶性肿瘤[1],其恶性程度、肿瘤浸润性和转移性均处于较高水平。原发性肝癌中尤以肝细胞癌(hepatocellular carcinoma,HCC)最为常见。据2024年全球癌症统计数据,肝癌是第6大常见癌症且高居全球癌症死亡率第3位。原发性肝癌常见的临床表现有肝肿大、纳差以及消瘦等,采取的主要治疗措施为肝切除术和肝移植术。虽然肝切除术可阻止肿瘤侵袭,显著改善其预后,但术后患者5年内死亡率仍较高[2]。现在通常认为患者的肿瘤大小、美国癌症联合委员会(American Joint Committee on Cancer,AJCC)TNM分期等危险因素对患者术后的总生存时间(overall survival,OS)产生不利影响。因此,研究构建原发性肝癌术后的预后预测模型对患者术后总生存时间的预测以及手术获益具有重要的临床实践意义。

国内外临床研究表明,原发性肝癌患者的临床病理因素、社会人口学因素、病史和影像学检查结果与其术后生存具有相关性[3-7]。近年来,机器学习在构建预测模型方面取得了显著进展,这些模型广泛应用于疾病诊断、发病预测、疗效评估、生存分析以及复发预测等多个领域[8-11]。但面对原发性肝癌难以根治性切除、术后并发症较多、预后普遍较差、复发率和转移率较高的特点,其预后研究体系尚未完善。在资料选择方面,当前原发性肝癌预后研究选取样本量普遍较小、未避免混杂且组间均衡性欠佳;在模型构建方面,在模型构建方面,通常运用Cox比例风险回归模型或Logistic回归模型等经典的统计方法来分析生存数据,而通过机器学习方法构建的预后预测模型较少且性能不佳。因此,本研究利用原发性肝癌患者的临床病理资料,通过人工少数类过采样技术(synthetic minority over-sampling technique, SMOTE)平衡基线差异,并基于机器学习方法构建预后预测模型,旨在为原发性肝癌术后的预后研究提供理论依据,以辅助临床医生有效改善患者的预后生存状况。

1 资料与方法

1.1 研究对象

本研究为回顾性队列研究,通过SEER * Stat8.4.0软件从美国癌症研究所监测、流行病学和最终结果(Surveillance, Epidemiology, and End Results,SEER)数据库中提取诊断为原发性肝癌的患者资料。纳入标准:①年龄为18~85岁;②手术方式为肝切除术或肝移植术;③诊断年份为2010-2018年;排除标准:①非原发性肿瘤;②合并其他恶性肿瘤;③性别、种族等临床资料缺失;④生存时间、生存结局不完整。最终获得有效样本4 297例,随访的截止日期为2019年12月31日。

1.2 研究方法

1.2.1 临床资料收集 依据第7版AJCC临床指南,同时参考既往PLC相关文献,获取的临床病例信息主要有年龄、性别、诊断年份、种族、组织学分期、手术方式、生存时间、生存状态、术前甲胎蛋白(alpha-fetoprotein,AFP)状态、肿瘤分化程度、肿瘤大小、肿瘤数目、居住区域、家庭收入、淋巴结受累情况、AJCC分期等。

1.2.2 SMOTE算法 SMOTE算法是一种扩充样本量的方法。本研究通过对少数类样本点与其最近邻样本点间进行随机插值,实现对少数类样本的扩充,产生的新样本不会改变原样本集的空间边界,算法信度较好。具体做法为:①针对样本i,随机抽取其k个近邻点中的样本j,并计算i与j的差值d;②产生数值$\theta \in (0,1)$,新样本=样本i+$d\theta$。重复上述步骤,直至样本变量扩充完毕。

1.2.3 模型构建与评估 采用独热编码和平均值插补法对数据进行预处理,利用SMOTE算法对术后5年生存患者数据扩增至2倍,平衡与死亡患者数据的组间差异。将数据按照8:2的比例(4 258:1 064)随机分为训练集与测试集,训练集数据用于建立模型,测试集数据对所建立的模型进行预测性能评估,选择10折交叉验证方法调整模型参数。将临床变量纳入决策树(decision tree, DT)、随机森林(random forest, RF)、梯度提升决策树(gradient boosting decision tree, GBDT)、极限梯度提升算法(eXtreme Gradient Boosting, XGBoost)4种机器学习模型,基于原发性肝癌术后5年的生存结局构建预测模型,最后,通过计算受试者工作特征曲线下面积(area under curve,AUC)、准确率、精确率、召回率、F1值等评价指标,对各算法构建的原发性肝癌预后预测模型的性能进行比较和评价。

1.3 统计学分析

应用软件R4.3.0编程进行统计学分析。符合正态分布的计量资料以$\bar{x}\pm s$表示,采用独立样本t检验;不符合正态分布的计量资料以$M(P_{25},P_{75})$表示,采用Mann-Whitney U检验;计数资料用频数和百分比表示,组间比较采用χ^2检验。$P<0.05$表示差异具有统计学意义。

2 结果

2.1 患者的基线资料分析

本研究收集了9 526例PLC患者,剔除符合排除标准的患者,得到有效样本4 297例,采用平均值插补法来处理肿瘤大小的缺失数据。在有效样本4 297例患者中(包括1 025例术后5年生存的患者和3 272例

术后5年死亡的患者,比例约为1∶3),5年生存率显示为24%。对比2组患者临床一般资料,年龄、种族、肿瘤分化程度、肿瘤大小、术前AFP状态、居住区域、家庭收入、淋巴结受累情况、AJCC分期等差异均有统计学意义($P<0.05$),性别和肿瘤数量的差异无统计学意义。见表1。

表1 患者临床基本资料分析[$M(P_{25},P_{75})$,例(%)]

指标	生存组($n=1\,025$)	死亡组($n=3\,272$)	P
年龄/岁	61.0(56.0,66.0)	64.0(58.0,70.0)	<0.001
性别			0.503
男	266(26)	886(27.1)	
女	759(74)	2 386(72.9)	
种族			027
白种人	666(65.0)	2 042(62.4)	
黑种人	101(9.9)	431(13.2)	
其他	258(25.2)	799(24.4)	
肿瘤分化程度			<0.001
低分化	147(14.3)	728(22.2)	
中分化	571(55.7)	1 847(56.4)	
高分化	297(29.0)	637(19.5)	
未分化	10(1.0)	60(1.8)	
肿瘤大小/cm	3.2(2.2,5.3)	5.3(3.8,5.3)	<0.001
术前AFP			0.010
阴性	464(45.3)	1 331(40.7)	
阳性	561(54.7)	1 941(59.3)	
肿瘤数量/个	1(1.0,1.0)	1(1.0,1.0)	0.331
居住区域			0.012
大都市地区	987(96.3)	3 083(94.2)	
非大都市地区	38(3.7)	189(5.8)	
家庭收入/美元			0.022
≤60 000	324(31.6)	911(27.8)	
>60 000	701(68.4)	2 361(72.2)	
淋巴结受累情况			<0.001
阴性	600(58.5)	2 610(79.8)	
阳性	425(41.5)	662(20.2)	
T分期			<0.001
Ⅰ	615(60.0)	1 743(53.3)	
Ⅱ	332(32.4)	974(29.8)	
Ⅲ	70(6.8)	407(12.4)	
Ⅳ	8(0.8)	148(4.5)	
N分期			0.011
Ⅰ	1 020(99.5)	3 220(98.4)	
Ⅱ	5(0.5)	52(1.6)	
M分期			0.001
Ⅰ	1 020(99.5)	3 203(97.9)	
Ⅱ	5(0.5)	69(2.1)	

2.2 模型预测性能对比

在训练集中,RF模型的准确率最高,达到了0.817,但召回率仅为0.283,表明数据量不足,需要通过SMOTE扩充样本量提升机器学习模型的预测性能。经组合SMOTE算法,各模型在训练集上的AUC、准确率、精确率、召回率和F1值均得到提高,其中SMOTE+RF模型的结果最佳,AUC值为1.000,准确率为0.992,精确率为0.996,召回率为0.984,F1值为0.990(表2)。

表2 各模型在训练集上的预测性能比较

模型	AUC值	准确率	精确率	召回率	F1值
DT	0.719	0.780	0.587	0.255	0.355
RF	0.934	0.817	0.853	0.283	0.425
GBDT	0.819	0.791	0.649	0.271	0.382
XGBoost	0.842	0.797	0.654	0.311	0.421
SMOTE+DT	0.732	0.718	0.709	0.451	0.552
SMOTE+RF	1.000	0.992	0.996	0.984	0.990
SMOTE+GBDT	0.903	0.820	0.825	0.674	0.742
SMOTE+XGBoost	0.922	0.839	0.869	0.686	0.767

在测试集中,SMOTE+RF模型表现良好,展示出较高的预测性能,其AUC、准确率、精确率、召回率和F1值均高于其他模型,分别为0.895、0.811、0.806、0.671和0.732。从整体上看,组合SMOTE后,各机器学习模型的预测性能皆有显著提升(表3)。

表3 各模型在测试集上的预测性能比较

模型	AUC值	准确率	精确率	召回率	F1值
DT	0.750	0.776	0.559	0.302	0.392
RF	0.774	0.771	0.557	0.190	0.283
GBDT	0.789	0.773	0.550	0.268	0.360
XGBoost	0.791	0.774	0.552	0.283	0.374
SMOTE+DT	0.730	0.703	0.694	0.410	0.515
SMOTE+RF	0.895	0.811	0.806	0.671	0.732
SMOTE+GBDT	0.893	0.7979	0.793	0.644	0.711
SMOTE+XGBoost	0.894	0.8036	0.802	0.651	0.719

3 讨论

本研究利用回顾性队列研究,纳入患者年龄、性别、肿瘤大小、肿瘤数量、分化程度、淋巴结受累情况以及AJCC分期等指标,运用机器学习的方式,成功地构建了PLC患者预后预测模型。该模型的预测性能较以往研究有较大提升,能够有效预测原发性肝癌术后5年的生存结局。

PLC 临床特征表现隐匿,大多数病例在确诊时已处于癌症的中晚期阶段,导致其预后差,易复发,为患者带来较大的疾病负担,同时也给社会造成较大经济压力。[13]。既往研究发现,在成年人群中,HCC 的发病率与年龄相关,表现为随年龄增加而上升的趋势,且全球 PLC 发病率和预后存在显著的性别差异,男性患者相比女性预后较差[2,14-15]。AFP 具有免疫抑制剂的作用,可促进 PLC 发展过程中的恶性转化,并可能参与肝癌患者的多重耐药过程,是 PLC 诊断的重要生物标志物。因此,AFP 水平持续升高是 PLC 进展和预后不良的危险因素[16],许多 PLC 患者的血清 AFP 表达水平呈升高趋势[17]。同时,肿瘤大小、数量、分化程度、淋巴结受累情况以及 AJCC 分期都是影响 PLC 患者预后的重要因素[18-19]。本研究依据第 7 版 AJCC 临床指南并参考 PLC 相关文献纳入临床指标和社会学人口指标,以构建 PLC 患者预后预测模型。

临床研究中非平衡数据集的现象并不罕见,通过基于类与类之间样本量平衡的统计分析方法并不能达到满意的效果。近年来,在非平衡数据集上应用抽样方法改进数据集、改进分类算法逐渐广泛。不同于其他过采样方法简单地复制样本,SMOTE 算法对每个少数类样本在其近邻样本中随机选 1 个样本,然后在 2 个样本连线上随机选取 1 点作为新合成的样本,避免了过采样带来的数据规模扩大和模型训练复杂度增加的问题[20-21]。本研究利用 SMOTE 算法解决数据不平衡问题,增加少数类样本的数量,均衡组间数据,以提高组间可比性。结果显示,SMOTE 算法使模型的性能有了较大的提升。另外,SMOTE 算法增加少数类样本的数量,能够显著提升模型对少数类的识别能力,从而增强整体分类性能。

本研究利用 DT、RF、GBDT、XGBoost 4 种机器学习方法构建原发性肝癌患者的预后预测模型。结果表明,与以往采用机器学习方法建立的 PLC 术后预后预测模型相比,本研究建立的预后预测模型的预测性能和准确率较高,经 10 折交叉验证表明模型稳健性较好。各模型与 SMOTE 算法组合后,其在训练集和测试集上的拟合情况和预测性能呈现显著提升,能够有效预测原发性肝癌术后 5 年的生存结局,并进一步开发基于组合模型 SMOTE+RF 的原发性肝癌预后预测分析系统。为便于临床医生使用本模型,该模型可部署在网页上。临床医生可结合预测结果、患者具体情况和临床经验,制定个性化的治疗方案。同时,医生通过向患者展示基于该模型得出的预后预测结果,从而帮助患者更好地理解自身病情,提高治疗依从性。

本研究存在一些局限。首先,SEER 数据库缺乏放化疗方案、手术切缘、住院时间等更详细的临床资料,未能在模型中校正其对预后的影响,或许导致未知偏倚;其次,研究受中心限制,SEER 数据仅涵盖美国患者的信息,缺乏国内临床数据进行外部验证,结论外推时的代表性有待进一步验证;最后,本研究为大样本回顾性研究,可能存在一定的选择偏倚。未来可纳入更多相关临床资料同时采取多中心研究以做更深入的探讨。

参考文献:

[1] 中华人民共和国国家卫生和计划生育委员会. 原发性肝癌诊疗规范(2017 年版)[J]. 临床肝胆病杂志. 2017;33(8):1419-1431. DOI:10.3969/j.issn.1001-5256.2017.08.003.
National Health and Family Planning Commission of the People's Republic of China. Diagnosis, management, and treatment of hepatocellular carcinoma(V2017)[J]. Journal of Clinical Hepatology, 2017, 33(8): 1419-1431. DOI:10.3969/j.issn.1001-5256.2017.08.003.

[2] FORNER A, REIG M, BRUIX J. Hepatocellular carcinoma[J]. Lancet(London, England), 2018, 391(10127):1301-1314. DOI:10.1016/s0140-6736(18)30010-2.

[3] DING J, WEN Z. Survival improvement and prognosis for hepatocellular carcinoma: analysis of the SEER database[J]. BMC Cancer, 2021, 21(1): 1157. DOI:10.1186/s12885-021-08904-3.

[4] LIN E, ZOU B, ZENG G, *et al*. The impact of liver fibrosis on microvascular invasion and prognosis of hepatocellular carcinoma with a solitary nodule: a Surveillance, Epidemiology, and End Results (SEER) database analysis[J]. Annals of Translational Medicine, 2021, 9(16): 1310. DOI:10.21037/atm-21-3731.

[5] SHEN X H, LI H K, WANG F, *et al*. Clinical significance of lymph node metastasis in patients undergoing partial hepatectomy for hepatocellular carcinoma[J]. World J Surg, 2010, 34(5): 1028-1033. DOI: 10.1007/s00268-010-0400-0.

[6] ZHANG K, TAO C C, WU F, *et al*. A practical nomogram from the SEER database to predict the prognosis of hepatocellular carcinoma in patients with lymph node metastasis[J]. Ann Palliat Med, 2021, 10(4): 3847-3863. DOI: 10.21037/apm-20-1876.

[7] ZHENG Y S, ZHANG X, LU J S, *et al*. Association between socioeconomic status and survival in patients with hepatocellular carcinoma[J]. Cancer Med, 2021, 10(20): 7347-7359. DOI: 10.1002/cam4.4223.

[8] CHIU C C, LEE K T, LEE H H, *et al*. Comparison of models for predicting quality of life after surgical resection of hepatocellular carcinoma: a prospective study [J]. J Gastrointest Surg, 2018, 22(10): 1724-1731. DOI: 10.1007/s11605-018-3833-7.

[9] HO W H, LEE K T, CHEN H Y, *et al*. Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network[J]. PLoS One, 2012, 7(1): e29179. DOI: 10.1371/journal.pone.0029179.

[10] NOH B, PARK Y M, KWON Y, *et al*. Machine learning-based survival rate prediction of Korean hepatocellular carcinoma patients using multi-center data [J]. BMC Gastroenterol, 2022, 22(1): 85. DOI: 10.1186/s12876-022-02182-4.

[11] SHI H Y, LEE K T, LEE H H, *et al*. Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery[J]. PLoS One, 2012, 7(4): e35781. DOI: 10.1371/journal.pone.0035781.

[12] SATRIANO L, LEWINSKA M, RODRIGUES P M, *et al*. Metabolic rearrangements in primary liver cancers: cause and consequences [J]. Nat Rev Gastroenterol Hepatol, 2019, 16(12): 748-766. DOI: 10.1038/s41575-019-0217-8.

[13] ZHENG Y Y, LI Y, FENG J, *et al*. Cellular based immunotherapy for primary liver cancer[J]. J Exp Clin Cancer Res, 2021, 40(1): 250. DOI: 10.1186/s13046-021-02030-5.

[14] MCGLYNN K A, PETRICK J L, EL-SERAG H B. Epidemiology of hepatocellular carcinoma[J]. Hepatology, 2021, 73(Suppl 1): 4-13. DOI: 10.1002/hep.31288.

[15] RICH N E, MURPHY C C, YOPP A C, *et al*. Sex disparities in presentation and prognosis of 1110 patients with hepatocellular carcinoma[J]. Aliment Pharmacol Ther, 2020, 52(4): 701-709. DOI: 10.1111/apt.15917.

[16] WANG T, ZHANG K H. New blood biomarkers for the diagnosis of AFP-negative hepatocellular carcinoma [J]. Front Oncol, 2020, 10: 1316. DOI: 10.3389/fonc.2020.01316.

[17] LI W, LIU K, CHEN Y, *et al*. Role of alpha-fetoprotein in hepatocellular carcinoma drug resistance [J]. Curr Med Chem, 2021, 28(6): 1126-1142. DOI: 10.2174/0929867327999200729151247.

[18] JEONG S, LUO G J, GAO Q, *et al*. A combined Cox and logistic model provides accurate predictive performance in estimation of time-dependent probabilities for recurrence of intrahepatic cholangiocarcinoma after resection [J]. Hepatobiliary Surg Nutr, 2021, 10(4): 464-475. DOI: 10.21037/hbsn.2020.01.07.

[19] MAVROS M N, ECONOMOPOULOS K P, ALEXIOU V G, *et al*. Treatment and prognosis for patients with intrahepatic cholangiocarcinoma: systematic review and meta-analysis [J]. JAMA Surg, 2014, 149(6): 565-574. DOI: 10.1001/jamasurg.2013.5137.

[20] Chawla N V, Bowyer K W, Hall L O, *et al*. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357. DOI: 10.1613/jair.953

[21] NI Z, ZHU Y, QIAN Y, *et al*. Synthetic minority over-sampling technique-enhanced machine learning models for predicting recurrence of postoperative chronic subdural hematoma[J]. Front Neurol, 2024, 15: 1305543. DOI: 10.3389/fneur.2024.1305543.

(收稿:2023-10-13;修回:2024-07-02)

(编辑 邓强庭)