

乳腺癌复发因素预测模型构建*

黄昱诚

杨旭明

乔琼

(上海中医药大学附属曙光医院
上海 201203)(上海中医药大学针灸推拿学院
上海 201203)(上海中医药大学附属曙光医院
上海 201203)

〔摘要〕 构建乳腺癌复发因素预测模型, 分析最优化算法以及乳腺癌复发特征参数并对相关问题进行探讨, 包括随机对照试验应用情况、再复发重要因素以及预测模型算法确定、C5.0 算法比较研究等。

〔关键词〕 乳腺癌; 复发因素; 预测模型; SPSS Modeler; 神经网络

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.07.005

Building of the Prediction Model of Recurrence Factors of Breast Cancer HUANG Yucheng, Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China; YANG Xuming, School of Acumox and Tuina, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China; QIAO Qiong, Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

〔Abstract〕 The prediction model of recurrence factors of breast cancer is built. The paper analyzes the optimization algorithm and characteristic parameters of breast cancer recurrence, discusses the related issues, including the application situation of Randomized Controlled Trial (RCT), the determination of important factors of recurrence and prediction model algorithm, the comparative study of C5.0 algorithm, etc.

〔Keywords〕 breast cancer; recurrence factor; prediction model; SPSS Modeler; neural network

1 引言

乳腺癌发病与复发因素较复杂, 目前尚未确定。通常需要通过早期癌症筛查以实现早期发现、

诊断与治疗^[1-2]。建立预测模型是为了解决实际所遇到困难或是避免其再次发生, 即实现有效防控。当前运用最广泛的乳腺癌预测模型为 Gail 模型^[3]、Tyrer-Cuzick 模型^[4]与 BRCAPRO 模型^[5], 前两者是在危险因素基础上发展起来的, 后者建立在基因与疾病发病相关的基础上。目前较少有乳腺癌再复发因素风险预测模型构建。本文使用乳腺癌复发相关公开数据源 (<https://www.openml.org/d/13> 或 <http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer/>), 通过数据挖掘, 利用 SPSS Modeler 软件, 综合不同算法建立预测模型^[6], 发现潜在乳腺癌症复发特征参数, 以期对建立乳腺癌再复发预测最优化模型提供参考。

〔收稿日期〕 2020-09-18

〔作者简介〕 黄昱诚, 本科, 发表论文 1 篇; 通讯作者: 乔琼。

〔基金项目〕 国家体育总局健科技攻关项目“健身气功锻炼对中老年女性亚健康状况影响的数据挖掘研究——以马王堆导引术为例”(项目编号: G2018010)。

2 资料与方法

2.1 SPSS Modeler 软件

SPSS Modeler (18.0) 的前身是 Clementine 数据挖掘软件, IBM 公司将 Clementine 收购后整合到 SPSS 软件里推广到市场。其特色为可视化操作界面以及直观数据流, 可以让使用者清晰了解每一步骤及其意义。

2.2 数据清洗

利用“变量文件”方法将下载数据源载入 SPSS Modeler (以下简称 Modeler) 中。该数据源共有 285 条记录, 包含 10 个字段: (1) 类别。包含复发 (recurrence - events) 及未复发 (no - recurrence - events)。(2) 年龄。20 ~ 29 岁为一个区段, 依序到 70 ~ 79 岁。(3) 绝经期。包括未绝经 (premeno)、40 岁之后绝经 (ge40) 以及 40 岁之前绝经 (lt40)。(4) 肿瘤大小。0 ~ 4 为一区段, 依序到 50 ~ 54。(5) 受侵犯淋巴结数。0 ~ 2 个为一区段, 依序到 15 ~ 17 (另外有 1 条记录为 24 ~ 26)。(6) 有无结节。分为有 (yes) 或无 (no)。(7) 肿瘤临床分期。I 为最轻, III 为最严重。(8) 肿瘤位置。分为左 (left) 和右 (right)。(9) 肿瘤所在象限。分为左上 (left - up)、左下 (left - down)、右上 (right - up)、右下 (right - down) 以及中心型 (central)。(10) 是否接受放疗。分为是 (yes) 或否 (no)。全部字段为字符串类型, 只有“恶性肿瘤程度”为数值型, 将“类别”设置为目标变量, 其余字段设置为输入变量字段。“肿瘤所在位置”有 1 个数值为“?”, 占总记录的 0.35%; 字段“有无无结节”有 8 个数值为“?”, 占总记录的 2.81%。考量数据完整性, 将有缺失的数据删除。此外“受侵犯的淋巴结数”中“24 ~ 26”的数据只有 1 个, 占总记录的 0.36%, 且在该节字段中为离散数据, 通过“过滤器”予以删除。最终经数据清洗后得到 275 条记录。

2.3 样本平衡与建模前置

本案例中复发与未复发的样本数相差较多。在

此种情况下建模较易导致不能考虑少数样本的特征, 即无法客观评估两类因素, 可能产生非平衡类问题^[7], 因此使用随机欠采样方法重新采样。原始样本中未复发有 195 例而复发有 81 例, 对未复发案例进行 41% 的随机欠采样, 达成与复发病例相当数量的 1: 1。由于这是监督学习的模型建立, 需要将样本拆分为训练集与测试集以评估算法准确率, 可使用“平衡器”实现。本案例中 70% 的数据分为训练集, 30% 分成测试集, 对样本进行重复再抽样以最大限度增加样本量, 改善样本量较少问题^[8]。开始建模前的数据流显示, 见图 1。

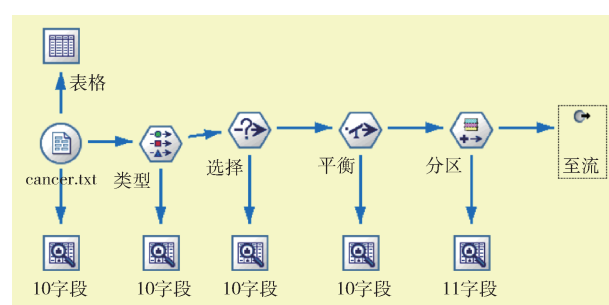


图 1 建模前的数据流

2.4 研究方法

2.4.1 损失矩阵 Modeler 具备丰富的分类算法。选取决策树、支持向量机、神经网络与最近邻分类 (K - nearest Neighbor Classification, KNN) 算法进行建模。决策树中包含 C5.0、CRT 算法、CHAID 算法以及 QUEST 算法; 支持向量机包括径向基函数 (Radical Basis Function, RBF), 多项式, Sigmoid 算法以及线形算法。在决策树系列算法中采用二分法或多分法进行类别分类, 并非每次分类都是 100% 的选择, 因此在某种程度上会有失真比例, 即医学领域的敏感性、特异性、误诊率以及漏诊率, 在本研究中体现为建模结果与实际复发、未复发的关系。实际情况中大多考虑敏感性以及特异性, 前者是检测出来为阳性而实际状况也为阳性, 后者是检测出来为阴性而实际状况也为阴性, 两者在建模中相加即为准确率。决策树在分类时并非完全归类, 不同算法分类和修剪方式也不同, 会出现因错判引起的损失问题, 即损失矩阵, 见表 1。 m

为弃真错误，实际为真而预测为假，即实际为乳腺癌复发但是模型预测未复发，也称为漏诊率； n 为取伪错误^[9]，即实际未复发但模型预测复发，也称为误诊率。将弃真损失设定为 1，取伪损失设定为 2，弃真错误由原来的 $(16/16 + 43 = 0.27)$ 变成 $(27/27 + 25 = 0.51)$ ，误诊率大幅增长；取伪错误由原来的 $(12/43 + 12 = 0.2)$ 变成 $(3/3 + 60 = 0.04)$ ，即漏诊率大幅下降。两者准确率无明显差异。

表 1 本案例的损失矩阵

实际状况	预测值	
	未复发	复发
未复发	0	n
复发	m	0

2.4.2 Boosting 技术 不同算法会产生不同拟合偏差，影响样本随机抽样。应用 Boosting 技术可

通过重复抽样现有样本而添加样本数，用以模拟增加的样本集进而提升准确率^[9]。

2.4.3 建模与预测变量重要性 使用 Modeler 内建自动建模器在训练集资料下选取总体精确性 > 90% 的算法，单独生成算法之后再行建模。通过自动分类器选取的算法为决策树中的 CRT 算法以及支持向量机 (Support Vector Machine, SVM) 中的 RBF 与多项式算法、神经网络中的多层感知器算法，见图 2。数据中央为超节点，是建模前数据流的浓缩以便简化数据流。由于建模前的数据经过两次随机采样，一是随机欠采样，二是分成训练集与测试集，所有算法得出的建模数据均经过 10 次采样平均。此外为了能够就本数据源进行乳腺癌复发变量预测，分别在两次随机采样的前中后针对所选取算法进行变量重要性预测。

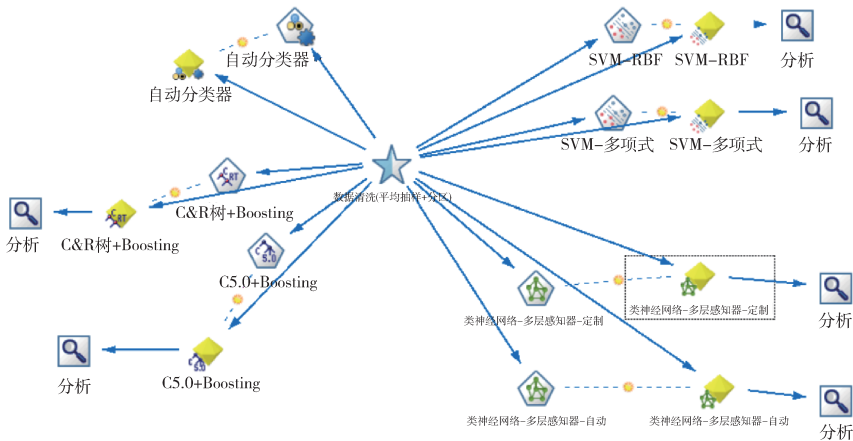


图 2 总体精确性 >90% 的算法 (C5.0 不在此限)

3 结果

3.1 最佳算法

通过自动分类器选出总准确率排名前 5 的算法，进行随机采样 10 次后平均的结果，见表 2。全部算法中选取最优解，即准确率最高情况，包括考虑损失矩阵以及 Boosting 增加模型准确率。使用受试者工作特征 (Receiver Operating Characteristic, ROC) 曲线下方面积 (Area Under the Receiver Operating Characteristic Curve, AUC) 进行计算，若所

得出值接近 0.9 则认为有较高分类性能^[10]。通过准确率以及 ROC 曲线下面积的比较，最适合的算法为使用神经网络的多层感知器 (Multi - Layer Perceptron, MLP) 且自动计算隐藏层的算法。

3.2 乳腺癌复发重要特征参数

本数据源各算法在两次随机采样的前中后所预测的重要性变量中，只有神经网络结合 MLP 的算法与决策树中的 CRT 算法所预测变量权重较大几率未改变 (>85%)。“肿瘤临床分期”可作为本数据源中乳腺癌复发的重要特征参数。其他算法随着随机采样

次数增多，所预测的变量重要性也随之改变。

表 2 各算法 10 次随机采样后的结果

算法	训练集				测试集			
	漏诊率 (%)	误诊率 (%)	准确率 (%)	AUC	漏诊率 (%)	误诊率 (%)	准确率 (%)	AUC
C5.0 ^{##}	52	3	74	0.630	50	1	73	0.643
CRT ^{##}	24	18	79	0.792	22	21	78	0.813
SVM – RBF	11	32	78	0.877	12	31	78	0.85
SVM – 多项式	16	35	75	0.881	17	40	71	0.858
类神经网络 – MLP – 定制 [#]	16	25	80	0.814	13	24	81	0.872
类神经网络 – MLP – 自动 [#]	7	24	84	0.869	7	27	83	0.894

注：##比较之前研究结果，列入 C5.0 算法；##使用损失矩阵，弃真损失 = 1，取伪损失 = 2，同时使用 Boosting 技术；#定制为指定隐藏层数为 1 个，自动为算法自动确定最佳隐藏层个数。

4 讨论

4.1 随机对照试验应用情况

数据挖掘技术的应用使在海量数据中寻找未知资讯并转化为有用知识变为可能^[11]。当前应用假设检验统计方法对人类观察或经验所得出变量进行处理而归纳出理论模型是预测模型的主要来源。其中随机对照试验（Randomized Controlled Trial，RCT）研究与临床决策相关度较高，其根据纳入排除标准在受试者中选取具有代表性的案例，在特定条件下采集疾病特征。RCT 存在样本选取局限性以及缺乏外部真实性的问题，而数据挖掘可以弥补此类不足。

4.2 再复发重要因素以及预测模型算法确定

本研究在乳腺癌治疗后的因素中寻找再复发的重要因素并得出预测模型算法。结果显示肿瘤临床分期占有较高权重，与刘胜、赵婧和刘佳等^[12]的研究结果相同。临床分期程度加重则 5 年复发转移同时增加，提示在临床中应给予较多关注。而在本数据源基础上对于多种算法进行评价，得出较佳算法为神经网络，其在低漏诊率、高准确率以及 AUC 方面获得较好表现。神经网络是模仿人类大脑处理信息方式的简化模型，其工作形式为模拟大量类似于神经元形式的互相联结单元，通过反复学习过程改变节点间联结权重，从而识别出数据中的模式。其中 MLP 允许构建较为复杂的关系，代价是更长的训

练与评分时间，反之 RBF 可以缩短训练与评分时间，但与 MLP 相比其预测能力较差。

4.3 C5.0 算法比较研究

本研究增加 C5.0 进行比较。C5.0 算法为罗斯·昆兰在 ID3（Iterative Dichotomiser 3）算法基础上所提出的 C4.5 改良版^[13]，主要改善 C4.5 的计算速度与运行内存。此外 C5.0 针对模型精准率增加 Boosting 方法，有研究指出使用 C5.0 误差率比 C4.5 的一半还低^[14]。在本研究中 C5.0 具有较低误诊率，但其漏诊率在全部算法中偏高且准确率、AUC 较其他算法较低。关于医疗模型评价，除了准确率以及 AUC 之外漏诊率以及误诊率是较重要考量因素，特别是漏诊及误诊的费用与时间所带来的放大效应，都是在数据挖掘中建立预测模型、选择算法时所要考虑的。

4.4 乳腺癌模型预测重点

当前乳腺癌模型预测重点在于影像学结合 Logistic 回归模型以建立预测模型^[15-16]，或者是在既有模型基础上增加新特征参数进行优化。例如石颖芳、张利群和徐胜昔利用人工绘制半定量方法，在 Gail 模型基础上结合乳腺密度检测对 40 例患者进行评估，其诊断率优于单纯使用 Gail 模型^[17]。Yala A、Lehman C 和 Schuster T 等对 89 112 个全视野数字化乳腺摄影样本结合深度学习方式建立模型，结果得出 AUC 为 0.7，优于单纯使用 Tyrer – Cuzick 模

型^[18]。上述两项研究所选用样本数量相差悬殊,反映出人工归纳与机器学习的差异。数据挖掘是在未明确假设即无预设立场的情况下,通过算法与模式匹配决定不同变量间的相关性,在大量、随机、有噪声数据中寻找特定规律,进而发现未知、潜在甚至与人类认知相反的知识与信息。数据挖掘能够利用大数据的海量性和整体性补充真实世界的特征信息以辅助临床决策,同时可以针对不同患者进行个体化预测^[19]进而建立个体化治疗方案^[20]。

5 结语

本研究在乳癌复发数据源基础上应用 Modeler 软件得出最佳算法,但复发与未复发每组样本数较少,数据源样本规模尚待进一步扩大。此外该数据源变量较少,不能完全作为反映实际情况的输入变量,所建立模型可能存在与现实脱节等问题。目前增加变量尚待检验手段与相应科技的进一步发展,以纳入更多、更广、更全面的参数特征,发挥数据挖掘优势,建立效果更好的预测模型,为乳腺癌复发因素预测提供更好的临床辅助决策支持。

参考文献

- 1 Bray Freddie, Ferlay Jacques, Soerjomataram Isabelle. Global Cancer Statistics 2018; GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. Cancer Journal for Clinicians, 2018, 68 (6): 394 - 424.
- 2 Li K, Anderson G, Viallon V, et al. Risk Prediction for Estrogen Receptor - specific breast cancers in two large prospective cohorts [J]. Breast Cancer Research, 2018, 20 (1): 147.
- 3 Gail M H, Brinton L A, Byar D P, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually [J]. Journal of the National Cancer Institute, 1990, 81 (24): 1879 - 1886.
- 4 Jonathan Tyrer, Stephen W Duffy, Jack Cuzick. A Breast Cancer Prediction Model Incorporating Familial and Personal Risk Factors [J]. Statistics in Medicine, 2004, 23 (7): 1111 - 1130.
- 5 Hall M J, Reid J E, Burbidge L A, et al. BRCA1 and BRCA2 Mutations in Women of Different Ethnicities Under-

- going Testing for Hereditary Breast - Ovarian Cancer [J]. Cancer, 2009, 115 (10): 2222 - 2233.
- 6 程国建, 张晗, 魏珺洁. 数据挖掘在乳腺癌复发预测中的应用研究 [J]. 智能计算机与应用, 2019, 9 (12): 104 - 107.
- 7 Guo H, Li Y, Shang J, et al. Learning from Class - imbalanced Data: review of methods and applications [J]. Expert Syst Appl, 2016 (73): 220 - 239.
- 8 Richard Simon. Resampling Strategies for Model Assessment and Selection [M]. Boston: Springer, 2017.
- 9 薛薇, 陈欢歌. SPSS MODELER 数据挖掘方法及应用 (第2版) (统计分析教材) [M]. 北京: 电子工业出版社, 2014.
- 10 Schleier P. ALA - based Fluorescence Diagnosis of Malignant Oral Lesions in the Presence of Bacterial Porphyrin Formation [J]. Proc SPIE, 2006 (6139): 46 - 53.
- 11 UN Global Pulse. Big Data for Development: challenges & opportunities [EB/OL]. [2020 - 05 - 01]. <http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment - UNGlobalPulseMay2012.pdf>.
- 12 刘胜, 赵婧, 刘佳, 等. 乳癌术后对乳腺癌术后5年复发转移率的影响 [J]. 中西医结合学报, 2008 (10): 1000 - 1004.
- 13 Quinlan J R. Induction of Decision Tress [J]. Machine Learning, 1986, 1 (1): 81 - 106.
- 14 赵玲萍. 多层次模糊综合评判法在中水工程中的应用 [J]. 甘肃科学学报, 2005, 17 (1): 111 - 113.
- 15 李蔓英, 李彬, 罗佳, 等. 基于灰阶超声的影像组学模型预测乳腺癌新辅助化疗效果 [J]. 中国医学影像技术, 2019, 35 (9): 1331 - 1335.
- 16 王鹏, 王阳, 吴华, 等. 基于声像图特点及人口学特征建立模型预测乳腺影像报告和数据系统分级 [J]. 中国医学影像技术, 2019, 35 (9): 1341 - 1345.
- 17 石颖芳, 张利群, 徐胜昔. 乳腺密度联合 Gail 模型对乳腺癌诊断的临床价值 [J]. 中国医学创新, 2014, 11 (33): 132 - 134.
- 18 Yala A, Lehman C, Schuster T, et al. A Deep Learning Mammography - based Model for Improved Breast Cancer Risk Prediction [J]. Radiology, 2019, 292 (1): 60 - 66.
- 19 秦文哲, 陈进, 董力. 大数据背景下医学数据挖掘的研究进展及应用 [J]. 中国胸心血管外科临床杂志, 2016, 23 (1): 55 - 60.
- 20 杨旭明, 杜小正. 中医药数据挖掘案例分析 [M]. 北京: 中医中医药出版社, 2016.