

决策树在原发性肝癌鉴别诊断中的应用*

孙向东

董长征

陈晓妍 唐 玲 王 侃

(宁波市卫生信息中心 宁波
315010)(宁波大学医学院 宁波
315040)(宁波市卫生信息中心 宁波
315010)

〔摘要〕 以 95 例原发性肝癌和 190 例肝硬化病例的 CT 诊断、乙肝表面抗原、甲胎蛋白、癌胚抗原和岩藻糖苷酶等影像学 and 血清学标志物检测结果以及性别和年龄等为基本资料, 利用决策树算法和诊断指标建立判别规则, 辅助对原发性肝癌进行鉴别诊断。结果表明以决策树为代表的数据挖掘技术能够较好地辅助肝癌鉴别诊断。

〔关键词〕 决策树; 原发性肝癌; 判别规则; 鉴别诊断

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2015.09.013

Application of the Decision Tree in Differential Diagnosis of Primary Hepatocellular Carcinoma SUN Xiang-dong, Ningbo Health Information Center, Ningbo 315010, China; DONG Chang-zheng, Medical College, Ningbo University, Ningbo 315040, China; CHEN Xiao-yan, TANG Ling, WANG Kan, Ningbo Health Information Center, Ningbo 315010, China

〔Abstract〕 By use of the decision tree algorithm and diagnostic indexes, the paper sets up the discrimination rules to make differential diagnosis of Primary Hepatocellular Carcinoma (PHC) based on basic data of 95 patients with PHC and 190 patients with liver cirrhosis, including the CT diagnosis, testing results of imaging and serologic markers such as the HbsAg, AFP, CEA and AFU, sex and age, etc. As indicated by the results, the data mining technology represented by the decision tree can support the differential diagnosis of PHC.

〔Keywords〕 Decision tree; Primary Hepatocellular Carcinoma (PHC); Discriminant rules; Differential diagnosis

1 引言

原发性肝癌 (Primary Hepatocellular Carcinoma, PHC) 简称肝癌, 是我国 10 大恶性肿瘤之一, 死亡率排在恶性肿瘤第 3 位, 是危及生命的常见病^[1]。目前, 肝癌确切的病因及发病机制仍不明确, 一般来说, 男性发病率高于女性, 且发病率随着年龄增长而

上升, 乙肝表面抗原 (HbsAg)、饮酒和肝硬化等是肝癌的危险因素^[1]。肝癌的预后相对较差, 因此早期发现、早期诊断是早期治疗的基础和前提, 也是有效提高肝癌患者生存率、改善其预后的重要保证。

甲胎蛋白 (α -fetoprotein, AFP)、癌胚抗原 (Carcino-embryonic Antigen, CEA) 和 α -L-岩藻糖苷酶 (α -L-fucosidase, AFU) 是肝癌重要的血清分子标志物, 对于肝癌的早期诊断具有重要意义, 但是这 3 个肿瘤标志物的敏感性和特异性都不够高 (介于 40% ~ 85%), 因此主要用于肝癌的筛查^[2-3]。CT 和 B 超是肝癌的影像学检查方法, 是肝癌诊断的主要依据^[4-5]。肝癌与一般的良性肝病 (例如肝炎和脂肪肝) 比较容易鉴别, 主要鉴别对象是肝硬化, 但二者无论是肿瘤标记物水平还是影像学都

〔修回日期〕 2015-05-14

〔作者简介〕 孙向东, 副教授, 发表论文 18 篇。

〔基金项目〕 宁波市自然科学基金项目“早期肝癌辅助诊断系统数据挖掘方法及专家系统研究” (项目编号: 2012A610191)。

有相似之处^[4,6]。本文采用决策树（Decision Tree，DT）算法，综合考虑年龄、性别、乙肝表面抗原、CT 诊断、肿瘤标志物（甲胎蛋白、癌胚抗原和岩藻糖苷酶）等因素，为肝癌的鉴别诊断提供判别参考。

2 资料和方法

2.1 一般资料

病例资料来自于当地某三甲医院 2013 年 1 – 12 月住院资料，所有肝癌和肝硬化病例均通过病理诊断且资料齐全。肝癌组：共 95 例，均初次诊断为肝癌，排除复发和术后病例；其中，男性 81 例，女性 14 例，年龄介于 18 ~ 82 岁之间。肝硬化组：共 190 例，从肝硬化病例中随机选取，排除肝癌；其中，男性 134 例，女性 56 例，年龄介于 26 ~ 87 岁之间。采集病例年龄、性别（男性 1/女性 0）、乙肝表面抗原（阳性 1/阴性 0）、CT 诊断（阳性 1/阴性 0）、肿瘤标志物（甲胎蛋白、癌胚抗原和岩藻糖苷酶）等资料。

2.2 决策树算法

决策树采用二叉树来建立判别规则，树上的每个内部节点表示根据某个分类指标大小进行判断，判断的结果指向下游不同的决策树分支，每个树叶节点代表最终的某个判别分类，从起始节点到最终树叶节点即代表一条完整的判别规则，可供临床作为诊断参考标准。决策树的核心在于节点对于指标的选择，一般通过特点算法（如 CRT、C5 和 CHAID 算法）来选择节点分类指标，使得分类结果准确率达到最大化。本文采用分类和回归树（Classification and Regression Trees，CRT）算法，利用 Gini 系数分割构建二叉决策树，向后修建树结构。CRT 算法的好处是对连续变量或离散变量的指标都能很好地适应，自动测算分割阈值进行构树，与人类专家决策思路更加接近^[7]。节点最小个体数设置为 5，最大树深为 6 层，计算判别准确率和十倍交叉一致性（10 – fold crossvalidation）来评估模型判别效果。决策树算法通过 SPSS 17.0 统计软件实现。

3 结果

3.1 判别指标检测结果

如表 1 所示，肝癌组 85.3% 为男性，14.7% 为女性；肝硬化组 70.5% 为男性，29.5% 为女性。CT 检测结果，肝癌组 88.4% 为“确诊或首先考虑肝癌”，作为阳性结果，11.6% 为“考虑或不排除肝癌”，作为阴性结果；肝硬化组 98.9% 为“确诊肝硬化”，作为阴性结果，1.1% 为“考虑肝癌”，作为阳性结果。肝癌组 89.5% 乙肝表面抗原为阳性，肝硬化组则为 63.2%。由于经正态性检验，甲胎蛋白、岩藻糖苷酶和癌胚抗原这 3 个血清肿瘤标志物数据均不服从正态分布，因此一般描述采用中位数（下四分位数、上四分位数）形式表示。肝癌组甲胎蛋白、岩藻糖苷酶和癌胚抗原的中位数分别为 71.0、20.5 和 1.9，肝硬化组分别为 1.7、11.8 和 1.6，每个指标均存在一定交叉范围。肝癌组年龄中位数为 58，肝硬化组为 57。

表 1 判别指标检测结果

指标		肝癌组	肝硬化组
性别（人）	男	81（85.3%）	134（70.5%）
	女	14（14.7%）	56（29.5%）
CT（人）	阳性	84（88.4%）	2（1.1%）
	阴性	11（11.6%）	188（98.9%）
乙肝表面抗原（人）	阳性	85（89.5%）	120（63.2%）
	阴性	10（10.5%）	70（36.8%）
甲胎蛋白（μg/L）		71.0（4.2，1000）	1.7（1.1，2.3）
岩藻糖苷酶（U/L）		20.5（13.8，26.6）	11.8（10.4，13.5）
癌胚抗原（μg/L）		1.9（1.1，2.9）	1.6（0.9，2.4）

3.2 决策树判别结果

决策树判别结果，见图 1。285 例病例，共误判 5 人，判别准确率达到 98.2%，十倍交叉一致率达到 95.4%，说明决策树用于肝癌/肝硬化的判别分析取得了较好效果。具体判别规则，见表 2，判别肝癌的规则分别是规则 1、规则 2、规则 3 及规则 5，诊断准确率均为 100%。判别肝硬化的规则分别

是规则 4 和规则 6，诊断准确率分别为 94.4% 和 97.7%，分别有 1 例和 4 例肝癌被误诊为肝硬化。从误诊的结果来看，这 5 例肝癌均为不典型病例，一般肝癌病例 CT 诊断为阴性，甲胎蛋白和岩藻糖苷酶等标志物水平特别低。

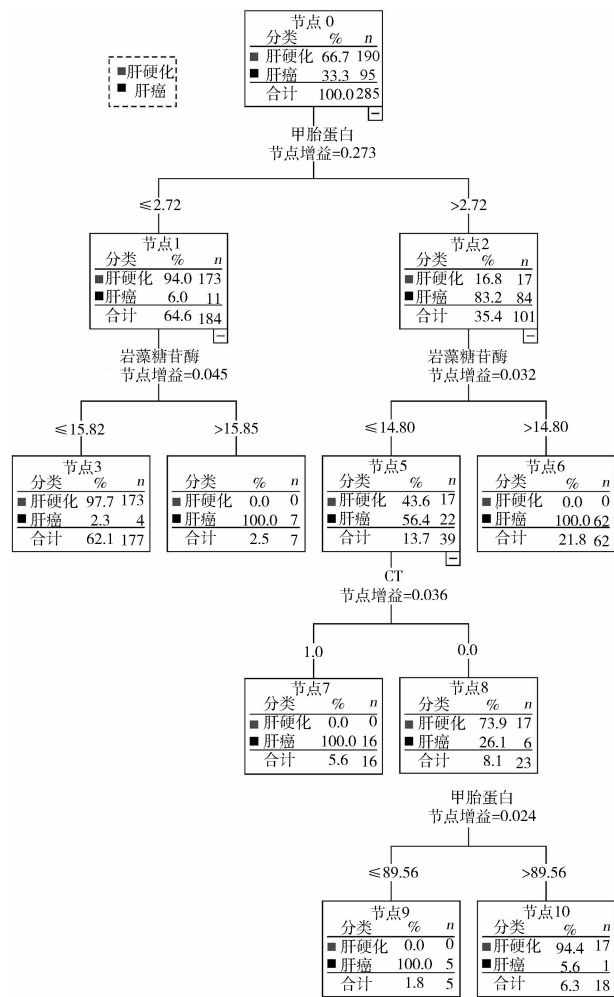


图1 决策树对肝癌/肝硬化判别结果

4 讨论与结果

4.1 数据挖掘技术的优势

在医学诊断（如肝癌诊断）中，常常需要考虑多个因素（例如影像学和血清学指标）进行综合判断，但是综合判断又会遇到一些困难。一方面，病例的指标不一定非常典型，例如本文中就有一些肝癌病例的肿瘤标志物水平非常低，甚至影像学上也没有明显的肿瘤特征；另一方面，综合判断需要丰富的临床经验，对于年轻医生来说，这是一个长期的积累过程。以决策树、神经网络、遗传算法和支持向量机为代表的数据挖掘技术，在机器学习和模式识别领域具有独特的优势。白雪峰等^[6]利用人工神经网络进行肝癌诊断^[6]，赵紫奉等^[7]基于决策树CRT算法对15种疾病进行诊断分析，魏小玲等^[8]利用决策树C5算法和生物标志物对肺癌进行辅助诊断，余鹏等^[9]基于遗传算法对骨肿瘤进行分类，王家祥等^[10]利用支持向量机和血清蛋白质指纹图谱来诊断甲状腺癌，这些都说明相较于传统的多元统计分析方法（如线性和logistic回归），数据挖掘技术由于采用非线性算法并通过交叉一致性来计算预测准确率，具有较强的灵活性、适应性和鲁棒性，不需要较为明确的统计分布来计算P值。决策树算法由于其判别规则与人类专家决策模式相近，而不像人工神经网络和遗传算法等不管是算法还是参数都相对抽象，在医学临床诊断中更容易得到应用。

表2 决策树判别规则及结果

顺序	判别规则	判别	实际结果	准确率（%）
1	甲胎蛋白 > 2.72 并且 岩藻糖苷酶 > 14.8	肝癌	肝癌 62 人；肝硬化 0 人	100
2	甲胎蛋白 > 2.72 并且 岩藻糖苷酶 < 14.8 并且 CT = 1	肝癌	肝癌 16 人；肝硬化 0 人	100
3	89.56 > 甲胎蛋白 > 2.72 并且 岩藻糖苷酶 < 14.8 并且 CT = 0	肝癌	肝癌 5 人；肝硬化 0 人	100
4	甲胎蛋白 > 89.56 并且 岩藻糖苷酶 < 14.8 并且 CT = 0	肝硬化	肝癌 1 人；肝硬化 17 人	94.4
5	甲胎蛋白 < 2.72 并且 岩藻糖苷酶 > 15.85	肝癌	肝癌 7 人；肝硬化 0 人	100
6	甲胎蛋白 < 2.72 并且 岩藻糖苷酶 < 15.85	肝硬化	肝癌 4 人；肝硬化 173 人	97.7

4.2 甲胎蛋白和岩藻糖苷酶对肝癌鉴别诊断的意义

在本文的判别诊断中, 判别准确率和交叉一致率分别为 98.2% 和 95.4%, 取得了较为理想的判别结果。需要注意的是, 如果单纯采用 CT 进行诊断, 诊断准确率能够达到 88.1%, 说明在肝癌诊断中, CT 仍是核心手段, 而其他指标的联合诊断将诊断准确率提高了约 10%。CT 结合造影能够较为清晰地观察到典型肝癌病例肝脏中肿瘤的形态、大小、包膜和病灶强化等, 但是肝硬化结合和小肝癌的影像学表现有时难以区分, 造影效果也跟造影技术以及病例肝脏血供等因素密切相关, 因此肿瘤标志物的辅助诊断对于肝癌的鉴别诊断具有重要的意义^[4-5]。

本文获得的对肝癌鉴别诊断的判别规则中, 除了 CT 诊断外, 甲胎蛋白和岩藻糖苷酶这两个肿瘤标志物起到了主要作用, 所有的判别规则都有这两个指标的参与。甲胎蛋白是胎儿发育早期肝脏和卵黄囊合成的一种血清糖蛋白, 中文全称为“甲型-胚胎蛋白”; 成人含量非常低, 肝癌和肝硬化会明显升高, 因此是肝癌首选诊断指标^[2-3,11]。岩藻糖苷酶是一种溶酶体酶, 在肝癌和一些良性疾病(如糖尿病和胰腺炎)都可能会升高^[2-3,11]。而且还可以进一步发现, 肝癌的主要诊断依据判别规则 1 “甲胎蛋白 > 2.72 并且 岩藻糖苷酶 > 14.8”, 肝硬化的主要诊断依据判别规则 6 “甲胎蛋白 < 2.72 并且 岩藻糖苷酶 < 15.85”, 这与甲胎蛋白和岩藻糖苷酶是主要的肿瘤标志物且具有相对较高的敏感性和特异性是一致的。其他一些指标如癌胚抗原, 由于与甲胎蛋白和岩藻糖苷酶存在一定相关性或者与肝癌的关系相对不密切, 均未被纳入到判别规则中。这样只需要考虑 CT 诊断、甲胎蛋白和岩藻糖苷酶这 3 个指标, 依据相对简单的判别规则就可以对肝癌和肝硬化做较为准确的鉴别。当然, 本文也存在一些限制, 如未对肝癌的类型(包括肝细胞癌、胆管细胞癌等)进行详细分类; 未对肝癌进行分期; CT 诊断只粗略考虑阳性/阴性, 未充分利用 CT 的详细数据等, 这影响了更加深入地指导临

床诊断和治疗方案, 在今后的研究中需要加以完善。

5 结语

本文构建的决策树在肝癌的鉴别诊断中获得了很高的判别准确率, 这说明以决策树为代表的数据挖掘技术能够较好地辅助肝癌鉴别诊断, 促进肝癌的早期诊断。

参考文献

- 1 叶家才, 崔书中, 巴明臣. 原发性肝癌的流行病学特征及其危险因素 [J]. 实用医学杂志, 2008, 24 (10): 1839 - 1841.
- 2 姚登福, 姚敏, 蔚丹丹, 等. 原发性肝癌早期诊断特异分子标志物的研究进展 [J]. 胃肠病学和肝病杂志, 2012, 21 (3): 202 - 205.
- 3 朱明, 徐吟亚. 肿瘤标志物联合检测对原发性肝癌诊断价值分析 [J]. 国际检验医学杂志, 2013, 33 (18): 2274 - 2275.
- 4 施昌村, 韩必亮, 黄崇权. CT 扫描联合血清甲胎蛋白检测在原发性肝癌诊断中的意义 [J]. 实用医药杂志, 2014, 22 (12): 744 - 746.
- 5 陈大同, 穆伟斌, 张淑丽, 等. PACS 系统数字化影像诊断结果联合肿瘤标记物对肝癌早期诊断的应用研究 [J]. 中国普通外科杂志, 2014, 23 (7): 986 - 988.
- 6 白雪峰, 王平瑜, 吴拥军. 基于人工神经网络技术的肿瘤标志联合检测在肝癌诊断中的应用 [J]. 实用医药杂志, 2012, 29 (4): 298 - 300.
- 7 赵紫奉, 李韶斌, 孔抗美. 基于决策树算法的疾病诊断分析 [J]. 中国卫生信息管理杂志, 2011, 8 (5): 67 - 69.
- 8 魏小玲, 谭善娟, 何其栋, 等. 决策树联合生物标志在肺癌辅助诊断中应用 [J]. 中国公共卫生, 2013, 29 (10): 1479 - 1479.
- 9 余鹏, 吴朝霞. 基于遗传算法的骨肿瘤分类方法研究 [J]. 西安交通大学学报, 2001, 35 (2): 166 - 170.
- 10 王家祥, 王利, 范应中, 等. 基于支持向量机的血清蛋白指纹图谱模型在甲状腺癌诊断中的应用研究 [J]. 中华医学杂志, 2006, 86 (14): 979 - 982.
- 11 赵运胜, 王猛, 崔辰莹, 等. 评价 8 种血清标志物在原发性肝癌诊断中的临床价值 [J]. 重庆医学, 2014, 43 (2): 214 - 216.