技术方法 文章编号:0253-3626(2009)05-0606-04

决策树在乳腺癌高位淋巴结转移判别诊断中的应用

易 静[1],苏新良[2],王润华[1]

(1.重庆医科大学公共卫生学院卫生统计与信息管理教研室,重庆 400016;2.重庆医科大学附属第一医院内分泌外科,重庆 400016)

【摘 要】目的:探讨决策树(Decision tree)模型从乳腺癌病例信息中自动提取诊断规则,用于乳腺癌腋窝高位淋巴结判别分类诊断研究的计算机辅助诊断系统。方法:利用重庆医科大学 1998~2002 年普外科 223 例手术证实的乳腺癌患者 11 项指标,采用 C5 算法、CR 算法、QUEST 算法及(Chi-square automatic interaction detector,CHAID)算法 4 种决策树算法,建立了乳腺癌腋窝高位淋巴结判别分类模型。结果:建立的 CHAID 判别分类模型的准确率为 83.79%,CHAID 模型所获知识的可靠程度以及对测试集合的准确率明显优于 Logistic 模型。结论:CHAID 模型可从病例自动提取诊断规则,具有较广泛的实用价值,可应用于乳腺癌腋窝高位淋巴结判别分类的诊断。

【关键词】决策树;乳腺癌;判别分类

【中国图书分类法分类号】R181.3+2 【文献标识码】A 【收稿日期】2008-11-21

Applications of decision tree model in classification of axillary III lymph nodes metastasis of breast cancer

YI Jing, et al

(*Department of Health Statistics and Information Management, Public Health School, Chongqing Medical University*)

【Abstract】*Objective*: To discuss the application of Decision Tree model and fit classification model of maxillary Ⅲ lymph nodes of breast cancer, which can abstract diagnostic knowledge used for computer-based diagnosis system. *Methods*: Four classification models of Decision Tree with the arithmetic of C5, CR, QUEST and CHAID were applied for classifying maxillary Ⅲ lymph nodes

作者介绍:易 静(1969-),女,博士,研究方向:人群健康评价。

通讯作者:王润华,男,教授,E-mail:wrhzyp@163.com。

尿 Cr 阳性似然比最大,肯定诊断的价值大;急性肾小球肾炎的尿 OCT/尿 Cr 阴性似然比最小,否定诊断的意义大,尿 OCT/尿 Cr 是诊断急性肾小球肾炎的早期指标(见表 2,表 3)。

机体尿流率的变化可影响尿 OCT 活力浓度,又因留取 24 h 尿测定比较麻烦,而且留尿时间过长,酶本身就会灭活;采用尿 Cr 比值来报告尿酶排出率,以避免随机尿测定带来的误差,并且减少尿液浓缩与稀释的影响。在急性肾小球肾炎患者的多次生化中发现,在病变早期许多生化指标升高不显著时,血清 OCT 活性与尿 OCT/尿 Cr 比值便有所升高,这表明在诊断急性肾小球肾炎,它们是灵敏指标。急性肾小球肾炎患者血清 OCT 活性与尿 OCT/尿 Cr 比值上升程度明显高于其它肾脏疾病。对急性肾小球肾炎治疗后的患者动态观察发现血清 OCT 活性与尿 OCT/尿 Cr 比值明显下降,可作为急性肾小球肾炎患者判断病情严重程度及治疗疗效的指标。因肝脏疾病也能导致血清 OCT 升高,需根据病人病史与体检及影像检查,排除肝脏疾病,而肝脏疾病患者的尿 OCT 不升高,所以尿 OCT 活性测定比血清 OCT活性测定对急性肾小球肾炎的诊断更有意义。肾细胞中 OCT 的具体位置未见报导,肾脏疾病间 OCT 的差异尚需进一步研究。

参 考 文 献

[1] 贾长恩.组织化学[M].第 3 版.北京:人民卫生出版社,2001.504-505,343-344.

[2] 徐金升.肾脏疾病[M].第 1 版.北京:中国医药科技出版社,2007.96-99.

[3] Frantisek K, Baudouin F, Leon E, et al. Isolation and characterization of ornithine transcarbamylase from normal Human liver[J]. The Journal of biological chemistry, 1978, 253(16): 3939-3944.

[4] 周爱儒.生物化学[M].第 6 版.北京:人民卫生出版社,2003.172-176.

(责任编辑:冉明会)

in 223 cases of breast cancers, confirmed by with 11 diagnostic indexes in the first affiliated hospital of chongqing medical university from 1993 to 2003. *Results*: CHAID classification model was better than other three models, according to the reliability of diagnostic knowledge and the accuracy of the test samples, the accuracy being 83.79%. *Conclusion*: The diagnostic knowledge of CHAID classification model automatically from the cases can be used to classify maxillary Ⅲ lymph nodes of breast cancer and other diseases. It is worth of widely practical use.

【Key words】 Decision Tree; Breast cancer; Classification

乳腺癌是严重危害妇女身心健康的最常见恶性肿瘤[1],在恶性肿瘤中的比例为 31%,全世界每年 120 万妇女发生乳腺癌,有 50 万人死于该病,其死亡率仅次于肺癌而位居第二。在西方国家,其发病率占妇女恶性肿瘤的首位。我国虽是乳腺癌低发区,但近年来发病率在逐年上升,从 1972 年的女性恶性肿瘤第四位上升到 1999 年女性恶性肿瘤首位,并且目前仍以 3%的速度增长;发病年龄趋于年轻化。

乳腺癌的主要致死原因是恶性肿瘤细胞的转移,而淋巴道转移[2-7]是主要转移途径之一。乳腺癌发生淋巴道转移,其淋巴液主要是流入患侧腋窝淋巴结约占 70%。腋窝淋巴结有 4 组,是乳癌发生转移的最早受累部位,转移的程度是指导乳腺癌的诊断、分期、外科治疗和放疗、化疗的重要参考依据;而且是最重要的独立的影响原发性乳腺癌预后即是否复发的因素。因此,术前做出正确判断,对乳腺癌的诊断具有重要价值[8]。

1 资料与方法

1.1 资料来源

资料来源于重庆医科大学 1998 ~ 2002 年普外科 223 例手术证实的乳腺癌患者。每一病例包含对应病人的 11 项指标,其中"高位淋巴结状况"真实反映了乳腺癌高位淋巴结是否转移的情况,其余 10 项指标用于诊断,它们是:"年龄"、"月经状况"(未停经、已停经)、"肿瘤大小"(T_{0-1}、T_2、T_3、T_4、T_x)、"肿瘤部位"(内上、外上、外下、内下、乳晕区、整个)、"临床检查腋窝淋巴结状况"(N_0、N_1、N_2、N_3)、"B 超检查腋窝淋巴结状"(肿大、不肿大)、"术中发现腋窝中低位淋巴结状况"(肿大、不肿大)、"手术方式"(典型根治术、改良根治术Ⅱ式、改良根治术加胸大肌开窗术)、"病理类型"(粘液癌、侵润性导管癌、侵润性小叶癌、管内癌早侵、乳头派杰氏病、伴有破骨局细胞的乳腺癌)、"Ⅰ、Ⅱ组淋巴结转移状况"(N_0、N_1~N_3、N_4~N_9、≥N_{10})。

对原始数据中每个分类变量指标数量化,对分类变量指标的可能属性赋值,例如"月经状况"属性用 0 表示"未停经、",1 表示"已停经";其中"肿瘤部位"、"临床检查腋窝淋巴结状况"、"手术方式"、"病理类型"等无序多分类变量采用哑变量赋值;决策树(Decision tree)[10,11]模型要求指标为分类变量指标,故对连续型指标"年龄"进行离散化,分为 6 个等级。

1.2 方法

采用 EXCEL 录入资料,应用 SPSS Clementin10.1 决策树工具箱对资料进行分析。

1.3 Logistic 模型

Logistic 回归分析在流行病学的病因研究中,是分析疾病与危险因素间联系的一种统计方法[9]。在这类研究中,所观察指标的值,常以二项反应变量取值,即生存与死亡,是否发病,是否接触危险因素等的反应变量 y 的取值是 0 或 1。非条件 Logistic 模型的定义:

$$P=\exp(\alpha+\sum_{i=1}^{m}\beta_i x_i)/[1+\exp(\alpha+\sum_{i=1}^{m}\beta_i x_i)] \quad i=1, 2, 3, \ldots, m \qquad (1)$$

$$1-P=1/[1+\exp(\alpha+\sum_{i=1}^{m}\beta_i x_i)] \qquad i=1, 2, 3, \ldots, m \qquad (2)$$

式中 P 表示疾病发生的概率,1-P 表示疾病不发生的概率,α,β_1,$\cdots\beta_m$ 是回归模型中的参数。

Logistic 模型参数估计方法有:最大似然估计法(Maximum likehood estimate)、对数似然函数、非线性迭代方法—Newton-Raphson 法。

1.4 决策树模型

决策树[10,11]是解决实际应用分类问题的一种数据挖掘方法,分类是把数据项映射到事先定义好的类的一个过程,由一组输入的属性值向量和相应的类组成,用归纳学习算法得出分类。决策树是一个类似于流程图的树结构,其中每个内部节点表示在一个属性上的测试,每个分枝代表一个测试输出,而每个树叶节点代表类或类分布,树的最顶层节点是根节点。

目前生成决策树的算法主要有 5 种:CART(Classification and regression trees)算法、CHAID(Chi-square automatic interaction detector)算法、ID3 算法,C4.5 算法、C5 算法。最著名的决策树算法为 Quinlan 的贪心算法(Iterative dichotomizer version 3, ID3),其他都是在 ID3 上的扩展。

CHAID 适用于分类和等级数据的分析,是一种以目标最优为依据,具有目标选择、变量筛选和聚类功能的分析方法。其核心思想是:根据给定的反应变量和经过筛选的解释变量对样本进行最优分割,按照卡方检验的显著性进行多元列联表的自动判断分组。

1.5 模型评价

算法获得的诊断模式的准确性是指采用这个算法所获模式对样本进行判别分类结果的正确性，用准确度（Accuracy）表示。以病例的“高位淋巴结状况”的实际分类结果为金标准，算法获得的诊断模式（知识）的判断分类结果与之比较，其计算式为：$accuracy=\frac{a+d}{a+b+c+d}$，式中 a、b、c、d 分别为算法获得的诊断模式的判断分类结果的真阳性数、阴性数、假阳性数和真阴性数。

表 1 诊断模式的准确性判别结果

金标准	新诊断模型	
	阳性	阴性
阳性	a	b
阴性	c	d

2 结果与分析

2.1 基础数据分析

223 例手术证实的乳腺癌患者，其中 65 例确诊为高位淋巴结转移，158 例确诊为高位淋巴结未转移，年龄在 24~84 岁，平均年龄为 47 岁。

由于收集到的病例数有限，采用 10-折交叉验证（10-fold cross-validation）方法：将原始样本集随机分成 10 个样本数基本相同且互不重叠的子集，依次取 1 个子集作为测试集合，其余 9 个子集合并为 1 个训练集合，这样就获得 10 组训练~测试数据，统一编号为 S1、S2、S3、S4、S5、S6、S7、S8、S9、S10。每种算法都采用这 10 组数据进行训练和测试。

2.2 Logistic 模型

采用逐步 Logistic 回归法进行分析，其结果见表 2、表 3（只列出 2 表），编号为 S1、S2、S3、S4、S5、S8、S9 的各组分析中，可知乳腺癌腋窝高位淋巴结转移与腋窝中低淋巴结状况、肿瘤大小有明显关系，S6、S7、S10 组的危险因素稍有差别，增加了临床检查腋窝淋巴结状况和手术方式两项危险因素；测

表 2 S1 乳腺癌高位淋巴转移多因素非条件 Logistic 分析及预测分类结果

因素	参数估计	标准误	*P* 值	训练样本分类准确率(%)	测试样本分类准确率(%)
常数	7.909	1.231	0.000	84.00	95.65
肿瘤大小	0.615	0.287	0.032		
中低淋巴节状况	1.899	0.325	0.000		

表 3 S7 乳腺癌高位淋巴转移多因素非条件 Logistic 分析及预测分类结果

因素	参数估计	标准误	*P* 值	训练样本分类准确率(%)	测试样本分类准确率(%)
常数	7.760	1.255	0.000	77.27	81.82
临床检查腋窝淋巴结状况	0.889	0.434	0.041		
中低淋巴节状况	2.077	0.390	0.000		

表 4 CHAID 决策树模型训练样本与测试样本预测结果

训练－测试样本编号	训练样本分类准确率(%)	测试样本分类准确率(%)
S1	83.00	91.30
S2	83.00	91.30
S3	86.00	82.61
S4	83.58	86.36
S5	83.58	86.36
S6	84.08	81.82
S7	84.58	77.27
S8	84.08	81.82
S9	83.58	86.36
S10	85.07	72.73
平均值	84.06	83.79

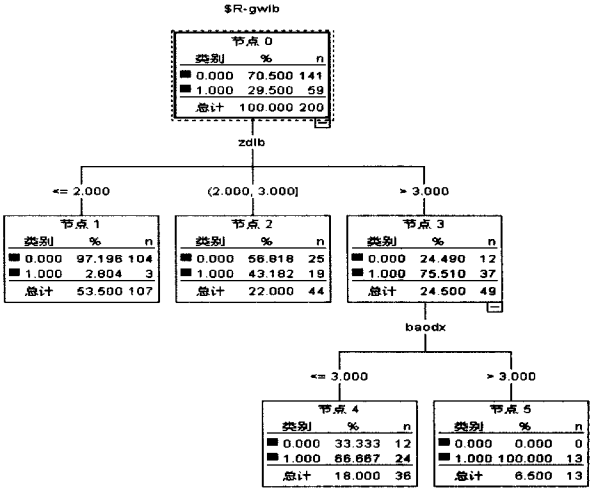


图 1 S1 决策树结构图

试结果除 S3 和 S10 两组的准确率为 78.26%和 68.18%,其他各组均达到 80%以上,平均的预测准确率为 83.34%。

2.3 CHAID 决策树模型

采用 C5 算法、CR 算法、QUEST 算法及 CHAID 算法,对 10 组训练~测试数据进行试验和筛选分析研究发现:CHAID 算法预测结果最好, 结果见图 1 (只列出 1 图)。

编号为 S1、S2、S7、S9、S10 的各组分析中可知"中低淋巴结状况"大于 3 且"肿瘤大小"大于 3,则"腋窝高位淋巴结状况"即可诊断为转移。

S2 分析可知"中低淋巴结状况"小于等于 2 且"肿瘤部位"大于 4,则"腋窝高位淋巴结状况"即可判别为无转移。"中低淋巴结状况" 小于等于 2 且"肿瘤部位"小于等于 3,则"腋窝高位淋巴结状况"有 99%的可能判别为无转移。

S3 分析可知"中低淋巴结状况"在 2~3 之间且"临床检查腋窝淋巴结状况"小于等于 0 且"手术方式"大于 1 且"肿瘤大小"大于 1,则"腋窝高位淋巴结状况"即可判别为无转移。

S7 分析可知"中低淋巴结状况"小于等于 2 且"临床检查腋窝淋巴结状况"大于 0 且"肿瘤部位"小于等于 4,则"腋窝高位淋巴结状况"有 96%的可能判别为无转移。

各组预测模型预测准确率结果见表 4。测试结果除 S7 和 S10 两组的准确率为 77.27%和 72.73%,其他各组均达到 80%以上,平均的预测准确率为 83.79%。

3 结 论

Logistic 模型、CHAID 模型的结果采用不同的表达方式,Logistic 模型采用 Logistic 回归系数表达诊断知识,不易被使用者解读;CHAID 模型采用了与自然语言相近的表达方式,提高了结果的可理解性。

两种模型从乳腺癌病例中提取乳腺癌腋窝高位淋巴结转移的诊断知识,对于测试样本,它们的平均准确率都达到 80%,达到了医生的一般要求。CHAID 模型所获模式(知识)的准确率较高,为 83.79%,Logistic 模型为 83.34%;就可靠程度来说,CHAID 模型所获模式(知识)的可靠程度较高,CHAID 模型所获知识的可靠程度以及对测试集合的准确率明显优于 Logistic 模型。

综合乳腺癌患者 10 项检查指标进行 CHAID 决策树分析,对乳腺癌的诊断正确率可达到83.79%;CHAID 方法提取的诊断规则描述简单,应用方便,可判断各诊断指标对乳腺癌诊断贡献的大小,从 CHAID 模型中可见,中低淋巴结状况对乳腺癌腋窝高位淋巴结转移诊断起决定性作用,肿瘤大小则可作为诊断的重要指标,因此,CHAID 方法是一种简便可行的计算机辅助诊断方法,可从病例自动提取诊断规则,具有较广泛的实用价值,可应用于其它疾病的诊断研究。

参 考 文 献

[1] 朴后男,崔兴烈.延边地区 1075 例乳腺癌统计分析[J].中华临床医学研究杂志,2005,11(21):3168-3169.

[2] 季晓昕,骆成玉,张 键,等.乳腺癌新辅助化疗后腋窝淋巴结的变化[J]. 中国微创外科杂志,2006,6(4):272-274.

[3] Veronesi U,Paganelli V,Viale G,et al. Sentinel-lymph-node biopsy as a staging procedure in breast cancer: update of a randomised controlled study[J]. Lancet Oncol,2006,7(12):983-990.

[4] 徐文通,李 荣.乳腺癌腋窝淋巴结转移的分析[J].中国现代医学杂志,2005,15(7):1029-1031.

[5] 苏新良,吴凯南.B 超检查预测乳腺癌腋淋巴结转移的价值[J].重庆医学,2004,33(2):244-245.

[6] Dabbs D J,Fung M,Landsittle D,et al.Sentinel lymph node micrometastasis as a predictor of axillary tumor burden[J]. Breast J,2004,10(2):101-105.

[7] Changsri C,Prakash C,Sandweiss L,et al.Prediction of additional axillary metastasis of breast cancer following sentinel lymph node surgery [J]. Breast J,2004,10(5):392-397.

[8] 毛利锋,瞿海斌.一种基于决策树的乳腺癌计算机辅助诊断新方法[J]. 江南大学学报:自然科学版,2004,3(3):227-229.

[9] 孙拯球.医学统计学[M].北京:人民卫生出版社,2002.257-270.

[10] Quinlan J R. Induction of decision tree[J]. Machine Learning,1986,(1):81-106.

[11] 王曙燕,耿国华,李丙春.决策树算法在医学图像数据挖掘中的应用[J]. 西北大学学报:自然科学版,2005,35(3):262-265.

[12] 叶晨洲,杨 杰,耿道颖.应用数据挖掘技术从大脑胶质瘤病例中获取诊断知识[J]. 生物医学工程学杂志,2002,19(3):426-430.

(责任编辑:关蕴良)