DOI:10.3969/j.issn.1672-5565.20161223001

CART 算法在原发性肝癌放疗后 HBV 再激活的应用

吴冠朋[1], 黄 伟[2], 刘毅慧[1*]
(1.齐鲁工业大学 信息学院, 济南 250353; 2. 山东省肿瘤医院放疗病区, 济南 250117)

摘 要:为了建立乙型肝炎病毒(Hepatitis B virus, HBV)再激活的预测模型,提出 CART(classification and regression tree)特征选择方法应用在原发性肝癌患者精确放疗后 HBV 再激活的危险因素分析中,进而建立基于 CART 和 Bayes 算法的 HBV 再激活预测模型。实验结果显示:CART 算法划分了多组具有优秀分类能力的特征节点集(危险因素),尤其当特征节点集为 HBV DNA 水平、外放边界、放疗总剂量、V20 和 KPS 评分时,在 CART 和 Bayes 预测模型中的分类正确性分别为 88.51%和 86.69%,得到 HBV 再激活正确性贡献度的排序为 KPS 评分>全肝平均剂量>V20>放疗总剂量>V10;当甲胎蛋白 AFP 出现时,增加了 HBV 再激活的预测正确性。

关键词:CART;特征选择;乙肝病毒再激活;危险因素;Bayes
中图分类号:TP391 文献标志码:A **文章编号**:1672-5565(2017)03-164-07

Application of HBV reactivation in primary liver carcinoma after radiotherapy based on CART algorithm

WU Guanpeng[1], HUANG Wei[2], LIU Yihui[1*]
(1.*School of Information*, *Qilu University of Technology*, *Jinan* 250353, *China*;
2. *Department of Radiation Oncology*, *Shandong Cancer Hospital*, *Jinan* 250117, *China*)

Abstract: To establish an excellent prediction model for Hepatitis B virus reactivation, the CART (classification and regression tree) feature selection method was applied to analyze the risk factors of Hepatitis B virus (HBV) reactivation in patients with primary liver cancer after precise radiotherapy, and then a prediction model of HBV reactivation was established based on CART and Bayes algorithm. The experimental results show that the CART algorithm split multiple sets of feature nodes (risk factors) with excellent classification ability. Especially when the feature set of nodes includes HBV DNA level, outer margin of radiotherapy, the total dose of radiotherapy, V20 and KPS score, the classification accuracy of CART and Bayes prediction models was 88.51% and 86.69% respectively. The decreasing order of accuracy contribution of HBV reactivation was: KPS score, mean dose of liver, V20, the total dose of radiotherapy and V10. The predictive accuracy of HBV reactivation was increased when the alpha-fetoprotein AFP appeared.

Keywords: CART; feature selection; HBV reactivation; Risk factors; Bayes

原发性肝癌[1-3]是全球第 5 大肿瘤疾病,中国原发性肝癌(primary liver carcinoma, PLC)患者众多,近年来精确放疗逐步成为治疗原发性肝癌的重要手段。2013 年,黄伟等[4]采用 logistic 回归分析 69 例经精确放疗的 PLC 患者发现基线血清 HBV DNA 水平是影响 HBV 再激活的独立危险因素,精确放疗后导致患者发生乙型肝炎病毒(Hepatitis B virus, HBV)再激活率达 25%,发生再激活的患者死亡率为 25%,HBV 再激活严重影响患者的生活质量以及生存周期。2014 年,Huang 等[5]又将临床剂量体积等因素纳入研究当中,发现 NLV(正常肝体积),V20,和 D-mean(平均剂量)与 HBV 再激活重

收稿日期:2016-12-23;**修回日期**:2017-02-18.
基金项目:国家自然科学基金(81402538,61375013);山东省自然科学基金(ZR2013FM020).
作者简介:吴冠朋, 男,硕士生,研究方向:智能信息及图像处理技术;E-mail:zbxywgp@163.com.
*通信作者:刘毅慧,女,博士, 教授,研究方向: 生物计算, 智能信息处理; E-mail: yxl@sdili.edu.cn.

要相关。2014 年,汪孟森[6]对山东省肿瘤医院治疗的 53 例原发性肝癌患者进行研究,推测肝功能 Child-Pugh 分级可能是发生 HBV 再激活的危险因素。2015 年,张晶晶等[7]研究发现 HBV 再激活患者和 HBV 未激活患者的 Child-Pugh 分级构成和 HBV DNA 水平差异具有统计学意义。吴冠朋等[8]对 90 例经精确放疗的原发性肝癌患者研究发现 HBV DNA 水平、外放边界和肿瘤分期 TNM 是致使 HBV 再激活的危险因素,并建立了基于 BP 和 RBF 神经网络的预测模型。随后 Wu 等[9]使用遗传算法应用在原发性肝癌患者精确放疗后的 HBV 再激活危险因素特征选择上,并建立了贝叶斯和支持向量机预测模型。临床上对原发性肝癌放疗后导致 HBV 再激活的危险因素有待进一步探究,且亟需建立更多的 HBV 再激活预测模型。

决策树算法包括 CART、ID3、C4.5 等, CART (classification and regression tree)算法是由 Breiman 等[10]提出的,是决策树中典型的二叉树,CART 算法有着较强的模式识别能力,并广泛应用在复杂的生物数据分析中。陈磊等[11]将 CART 算法用在肺癌微阵列数据上,并得到优秀分类能力的 CART 树模型。Kong 等[12]将 CART 算法用在乳腺癌分类上,提高了对乳腺癌的治疗质量。Gasparoviga-asite 等[13]将 CART 算法用于降低蛋白质维度特征,并得到分类任务中最有效的特征子集。本文把划分 CART 树的特征节点集作为 HBV 再激活的危险因素,然后用这些特征节点集建立基于 CART 和 Bayes 的 HBV 预测模型,最后得到基于 CART 和 Bayes 的 HBV 再激活预测结果。实验设计流程如图 1 所示。

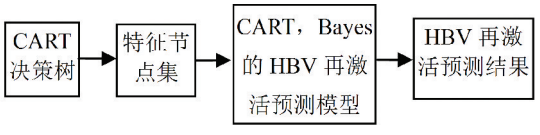

图 1　实验设计流程

Fig.1　Experimental design process

1　数据与相关原理

1.1　数据

实验数据来自山东省肿瘤医院收治的 90 例经精确放疗的原发性肝癌患者的临床资料,数据包含:年龄、HBV DNA 水平、外方边界、甲胎蛋白 AFP 和肿瘤分期 TNM 等 28 项特征属性,详见表 1。90 例患者中 20 例发生了 HBV 再激活,HBV 再激活率达 22.22%。对本组数据研究的意义在于如何从这些数据中找出 HBV 再激活的危险因素,并建立 HBV 再激活预测模型,从而指导个体病人在治疗过程中采用抗病毒治疗方法,避免发生 HBV 再激活,提高病人的生存质量及延长病人的生存周期。

表 1　特征节点及其所代表的特征属性

Table 1　Feature nodes and their attributes

特征节点	特征
x1	性别
x2	年龄
x3	KPS 评分
x4	HbeAg
x5	门脉癌栓有无
x6	肿瘤分期 TNM
x7	Child-Pugh
x8	甲胎蛋白 AFP
x9	HBV DNA 水平
x10	放疗总剂量
x11	等效生物计量
x12	放疗次数
x13	放疗前 TACE
x14	分割方式
x15	GTV 体积(gross tumor volume)
x16	PTV 体积(planning target volume)
x17	外放边界
x18	V5
x19	V10
x20	V15
x21	V20
x22	V25
x23	V30
x24	V35
x25	V40
x26	V45
x27	全肝最大剂量
x28	全肝平均剂量

1.2　CART 算法

1.2.1　构建 CART 决策树

CART[14]算法采用二分递归分割方法把数据中的特征作为二叉节点,CART 算法可建立带有特征选择的分类树和回归树,本文用于判别 HBV 是否激活,因此本文建立的是带有特征选择的分类树。CART 算法把非叶子结点作为判断类别的属性,叶子节点作为类别的标签,定义数据样本集 $\boldsymbol{I}$,属性向量集 $\boldsymbol{X}$ 和类别向量集 $\boldsymbol{Y}$ 为

$$\begin{cases}\boldsymbol{I}=(\boldsymbol{X},\boldsymbol{Y}),\\ \boldsymbol{X}=(x_1,x_2,\cdots,x_m),\\ \boldsymbol{Y}=(y_1,y_2,\cdots,y_c).\end{cases}$$

显然,本文数据中 $m = 28$, $c = 2$。根据所给数据集 $\boldsymbol{I}$ 建立并划分一棵二叉树,CART 算法使用 GINI 指数[15]划分一棵二叉树,对于一个 c 类样本集,用混合度 impurity(P) 来衡量节点的纯度(只包含同一类别的节点),GINI 指数定义为

$$\text{impurity}(P) = \text{GINI}(n) = 1 - \sum_{1}^{c} p^2(c).$$

式中,$P = (p_1, p_2, \cdots, p_c)$,其中 $p(c)$ 为第 c 类的概率,则节点 n 的混合度为

$$\text{impurity}(P) = \text{impurity}(p_1, p_2, \cdots, p_c).$$

当节点 n 是"纯"时 GINI 指数为 0,否则为正。对 CART 树而言,当节点 n 不满足属于同一类别或只有一个样本时,就需要对节点 n 进行划分,而划分时将混合度最大的进行划分,则得到最优分支。当节点 n 被划分成 n_1 和 n_2 后,则有

$$\begin{cases} p(n_1) = \dfrac{N_{n_1}}{N_n}, \\ p(n_2) = \dfrac{N_{n_2}}{N_n}. \end{cases}$$

式中:N_{n_1}、N_{n_2}、N_n 分别为 n_1、n_2、n 的样本数,划分后的混合度 $\Delta\text{impurity}(P)$ 为

$$\Delta\text{impurity}(P) = \text{impurity}(P) - p(n_1) \cdot \text{GINI}(n_1) - p(n_2) \cdot \text{GINI}(n_2).$$

其他节点亦重复以上划分过程,当 CART 树遇见以下情况时,停止划分为:

1)节点是纯的,即节点包含的样本属于同一类别。

2)属性集已划分完毕。

3)CART 树达到最大深度。

4)每个节点已达到允许划分的最小记录数。

1.2.2 CART 决策树的修剪

CART 决策树所选择的特征会影响预测结果,为了得到分类性能最好的 CART 树,对 CART 决策树的特征节点进行修剪,而修剪的方法包含前剪枝和后剪枝。前剪枝控制树的深度与叶子限制树的生长,后剪枝是在树完全生长后进行叶子与深度的再调整,较符合树的完全生长,本文以代价复杂性作为后修剪的策略,即

$$R_\alpha(T) = E(T) + \alpha |N_T|.$$

式中:$R_\alpha(T)$ 为树 T 的代价复杂性;$E(T)$ 为树 T 的误分类损失;α 为复杂性系数;$|N_T|$ 为叶子节点树,以代价复杂性最小选择出剪枝子树。

1.3 Bayes 分类模型

Bayes[16]分类器是基于先验概率求后验概率的一种统计分类器。假定总体样本第 i 类样本的先验概率 P_i,样品 x 属于 i 类样本的条件函数为 $f_i(x)$,则

$$f_i(x) = (2\pi)^{-\frac{e}{2}} |\boldsymbol{V}_i|^{-\frac{1}{2}} \cdot \exp\left(-\frac{d_i^2(x)}{2}\right).$$

式中:$\boldsymbol{V}_i$ 为联合协方差矩阵;$d_i^2(x)$ 为马氏距离,则基于 Bayes 理论判别 x 为 i 类样本的后验概率为

$$P(i|x) = \frac{P_i f_i(x)}{\sum P_i f_i(x)}, i = 1, 2, \cdots, n.$$

式中:P_i 为第 i 个总体的先验概率;n 为样本类别数量,显然本文中样本类别数量 n 为 2。

1.4 Hold-out 与 K 折交叉验证

为保证选取的特征以及预测结果不失泛化性,先采用不同 p 的 Hold-out 选择出划分 CART 的特征节点集,本文的 p 分别设为:0.7、0.8、0.9。例如,本文 90 个原发性肝癌数据,则有 $90\times p$ 个数据用于划分 CART 的特征节点集,每次为不同 p 的 Hold-out 运行 50 次,特征节点集用于建立 CART 和 Bayes 预测模型,随后再采用 K 折交叉验证取预测结果的平均 $\bar{A}_k$,即

$$\bar{A}_K = \frac{1}{K}\sum_{1}^{K} A_K,$$

式中,交叉验证的 K 设为 10。

1.5 预测模型性能评估

选用 3 个标准正确性(Accuracy)、灵敏性(Sensitivity)和特异性(Specificity)来评价所选特征的分类性能为:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN),$$

$$\text{Sensitivity} = TP/(TP+FN),$$

$$\text{Specificity} = TN/(TN+FP).$$

式中:Accuracy 为综合预测结果,作为预测结果的主要评判标准;Sensitivity 为再激活的预测结果;Specificity 为正常的预测结果。TP、TN、FP、FN 分别为真阳性(激活)、真阴性(正常)、假阳性和假阴性样本的数量。

2 结果与分析

文献[9]的正确性已经达到 82%以上,因此为了保证特征的意义,本文中预测结果选取正确性达到 80%的特征节点集,且将正确性达到 85%以上的特征节点集记为具有优秀的分类能力。

2.1 Hold-out 的 p 为 0.7 时 CART 选择的特征节点集及 CART 预测结果

运行了 50 次 Hold-out(p 为 0.7)划分 CART 树所选择的特征节点集在 10 折交叉验证下的 CART 预测结果见表 2。

在 CART 的特征选择中,CART 构建了易于理解的划分规则。例如表 2 第 1 组特征节点集:HBV DNA 水平(x9),外放边界(x17),KPS 评分(x3),V10(x19)和年龄(x2)的分类规则如图 2 所示,并最终得到激活(reactivation)和正常(normal)两种预测结果,其预测正确性达到 87.55%,灵敏性更是高达 98.49%,特异性达到 77.61%。

表 2　Hold-out 的 p 为 0.7 时 CART 所选的特征节点集及 CART 预测结果

Table 2　The set of feature nodes selected by CART and the prediction result of CART are obtained when the p value of Hold-out is 0.7

特征节点集	正确性/%	灵敏性/%	特异性/%
HBV DNA 水平,外放边界,KPS 评分,V10,年龄	87.55	98.49	77.61
HBV DNA 水平,外放边界,KPS 评分	85.16	97.80	60.91
HBV DNA 水平,外放边界,V20,甲胎蛋白 AFP	84.12	93.56	71.92
HBV DNA 水平,外放边界,V20	83.18	92.65	71.63
HBV DNA 水平,外放边界,V10	81.31	96.72	56.47

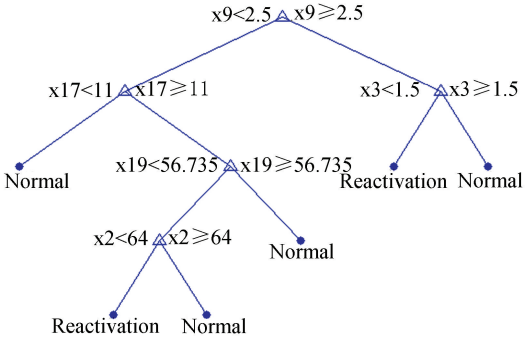

图 2　特征节点集:HBV DNA 水平(x9),外放边界(x17),KPS 评分(x3),V10(x19)和年龄(x2)的分类规则

Fig.2　The classification rule of feature nodes set of HBV DNA Level(x9), outer margin of radiotherapy(x17), KPS score(x3), V10(x19) and Age(x2)

表 2 第 2 组特征节点集:HBV DNA 水平(x9),外放边界(x17)和 KPS 评分(x3)的正确性为 85.16%,其分类规则如图 3 所示。

表 2 第 3 组特征节点集:HBV DNA 水平,外放边界,V20 和甲胎蛋白 AFP 的正确性为 84.12%,表 2 第 4 组特征节点集:HBV DNA 水平,外放边界和 V20 已在文献[5]和文献[8]中证明是影响 HBV 再激活的危险因素,其正确性为 83.18%,即"甲胎蛋白 AFP"的加入增加了 HBV 再激活正确性。

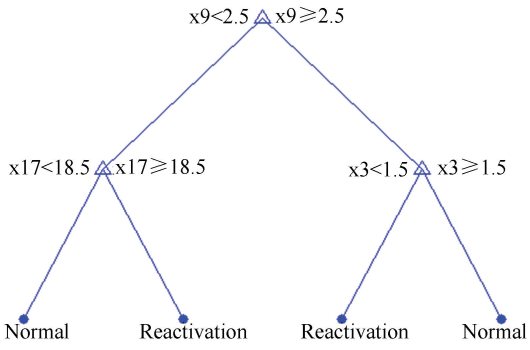

图 3　特征节点集:HBV DNA 水平(x9),外放边界(x17)和 KPS 评分(x3)的分类规则

Fig.3　The classification rule of feature nodes set of HBV DNA Level(x9),outer margin of radiotherapy(x17) and KPS score(x3)

表 2 第 5 组特征节点集:HBV DNA 水平,外放边界和 V10 的正确性为 81.31%。综合比较表 2 中的正确性,显然特征节点"KPS 评分"比特征节点"V20"和特征节点"V10"更能提高特征节点集的 HBV 再激活正确性,因此判定"KPS 评分"是影响 HBV 再激活的危险因素,且存在对 HBV 再激活的正确性贡献度:KPS 评分>V20>V10。

2.2　Hold-out 的 p 为 0.8 时 CART 选择的特征节点集及 CART 预测结果

运行 50 次 Hold-out(p 为 0.8)划分 CART 树所选择的特征节点集在 10 折交叉验证下的 CART 平均预测结果见表 3。

表 3　Hold-out 的 p 为 0.8 时 CART 所选的特征节点集及 CART 预测结果

Table 3　The set of feature nodes selected by CART and the prediction result of CART are obtained when the p value of Hold-out is 0.8

特征节点集	正确性/%	灵敏性/%	特异性/%
HBV DNA 水平,外放边界,放疗总剂量,V20,KPS 评分	88.51	97.74	74.54
HBV DNA 水平,外放边界,全肝最大剂量,甲胎蛋白 AFP	86.73	94.51	79.86
HBV DNA 水平,外放边界,全肝平均剂量	84.01	92.61	74.10

表 3 第 1 组特征节点集:HBV DNA 水平(x9),外放边界(x17),放疗总剂量(x10),V20(x21)和 KPS 评分(x3)的分类规则如图 4 所示,其正确性为 88.51%,灵敏性为 97.74%,特异性为 74.54%,该组特征节点集的正确性最好。

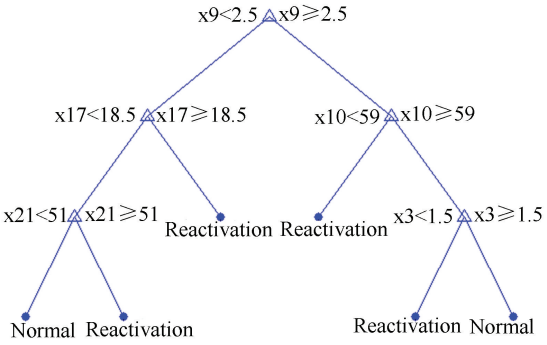

图 4 特征节点集:HBV DNA 水平(x9),外放边界(x17),放疗总剂量(x10),V20(x21)和 KPS 评分(x3)的分类规则

Fig.4 The classification rule of feature nodes set of HBV DNA Level(x9), outer margin of radiotherapy(x17), total dose of radiotherapy(x10), V20(x21) and KPS score(x3)

表 3 第 2 组特征节点集:HBV DNA 水平,外放边界,全肝最大剂量和甲胎蛋白 AFP 的分类正确性为 86.73%,表明该特征节点集也具有优秀的分类能力。

表 3 第 3 组特征节点集:HBV DNA 水平,外放边界和全肝平均剂量的正确性为 84.01%,高于表 2 第 4 组特征节点集:HBV DNA 水平,外放边界和 V20 的正确性,但低于表 2 第 2 组特征节点集:HBV DNA 水平,外放边界,KPS 评分的正确性,即存在对 HBV 再激活的正确性贡献度:KPS 评分>全肝平均剂量>V20。

2.3 Hold-out 的 p 为 0.9 时 CART 选择的特征节点集及分类预测结果

运行 50 次 Hold-out(p 为 0.9)划分 CART 树所选择的特征节点集,特征节点集在 10 折交叉验证下的 CART 平均预测结果见表 4。

表 4 Hold-out 的 p 为 0.9 时 CART 所选的特征节点集及 CART 预测结果

Table 4 The set of feature nodes selected by CART and the prediction result of CART are obtained when the p value of Hold-out is 0.9

特征节点集和初始特征集	正确性/%	灵敏性/%	特异性/%
HBV DNA 水平,肿瘤分期 TNM,外放边界,Child-Pugh	87.01	97.95	73.14
HBV DNA 水平,外放边界,肿瘤分期 TNM,KPS 评分	86.47	96.40	73.03
HBV DNA 水平,外放边界,放疗总剂量	81.52	96.79	64.68

表 4 第 1 组特征节点集:HBV DNA 水平(x9),肿瘤分期 TNM(x6),外放边界(x17),Child-Pugh(x7)的分类规则如图 5 所示,其正确性为 87.01%,灵敏性为 97.95%,特异性为 73.14%。

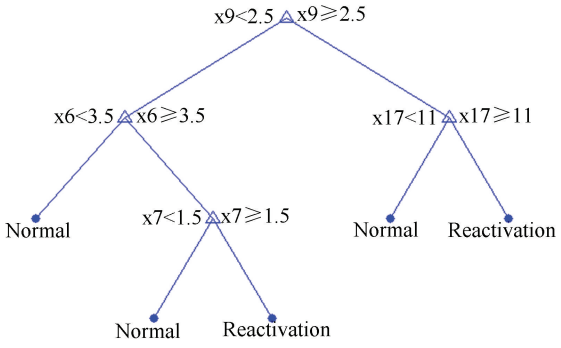

图 5 特征节点集:HBV DNA 水平(x9),肿瘤分期 TNM(x6),外放边界(x17),Child-Pugh(x7)的分类规则

Fig.5 The classification rule of feature nodes set of HBV DNA Level(x9), Tumor staging TNM(x6), outer margin of radiotherapy(x17), Child-Pugh(x7)

表 4 第 2 组特征节点集:HBV DNA 水平,外放边界,肿瘤分期 TNM 和 KPS 评分的正确性为86.47%,表明该特征节点集也具有优秀的分类能力。

表 4 第 3 组特征节点集:HBV DNA 水平,外放边界和放疗总剂量的正确性为 81.52%。低于表 2 第 4 组特征节点集:HBV DNA 水平,外放边界和 V20 的正确性,但高于表 2 第 5 组特征节点集:HBV DNA 水平,外放边界和 V10 的正确性,由此推出:对 HBV 再激活的正确性贡献度:V20>放疗总剂量>V10。

综上所述,得到一组对 HBV 再激活正确性贡献度的排序:KPS 评分>全肝平均剂量>V20>放疗总剂量>V10。

2.4 特征节点集的 Bayes 预测模型结果

Bayes 不考虑所选特征节点的先后顺序,因此可用于判定某些特征节点的加入或者替换对 HBV 再激活的影响。其初始特征集和特征节点集的结果见表 5。

在表 5 的 Bayes 预测模型中,特征节点集的分类性能相比初始特征集的都得到提高。但在相同特征节点集条件下,CART 的分类性能略优于 Bayes 的分类性能。

表 5 第 1 组特征节点集:HBV DNA 水平,外放边界,放疗总剂量,V20 和 KPS 评分在 Bayes 模型下的分类性能最好,正确性为 86.69%,灵敏性为96.36%,特异性为 74.86%,相比初始特征集其正确性,灵敏性和特异性分别提高:16.69%、21.36%和22.36%。并且该组特征节点集高于已知危险因素:HBV DNA 水平,外放边界和 V20 的正确性、灵敏性和特异性。证明了特征节点“放疗总剂量”和“KPS 评分”的加入提高了 HBV 再激活分类预测性能。

表 5　初始特征集和特征节点集的 Bayes 预测结果

Table 5　Bayes predictive results of original feature set and feature nodes set

初始特征集和特征节点集	正确性/%	灵敏性/%	特异性/%
初始特征集	70.00	75.00	52.50
HBV DNA 水平,外放边界,放疗总剂量,V20,KPS 评分	86.69	96.36	74.86
HBV DNA 水平,外放边界,KPS 评分,V10,年龄	86.51	97.33	77.57
HBV DNA 水平,肿瘤分期 TNM,外放边界,Child-Pugh	85.65	94.88	70.62
HBV DNA 水平,外放边界,肿瘤分期 TNM,KPS 评分	84.95	90.73	77.72
HBV DNA 水平,外放边界,KPS 评分	84.24	93.64	62.35
HBV DNA 水平,外放边界,全肝最大剂量,甲胎蛋白 AFP	83.82	91.16	60.47
HBV DNA 水平,外放边界,V20,甲胎蛋白 AFP	83.33	92.85	66.67
HBV DNA 水平,外放边界,全肝平均剂量	83.04	94.26	65.25
HBV DNA 水平,外放边界,V20	82.74	91.59	68.54
HBV DNA 水平,外放边界,放疗总剂量	80.95	90.77	62.72
HBV DNA 水平,外放边界,V10	80.03	94.51	55.81

表 5 第 2 组特征节点集:HBV DNA 水平,外放边界,KPS 评分,V10 和年龄的正确性为 86.51%。表 5 第 3 组特征节点集:HBV DNA 水平,肿瘤分期 TNM,外放边界和 Child-Pugh 的正确性为 85.65%。表 5 第 4 组特征节点集:HBV DNA 水平,外放边界,肿瘤分期 TNM 和 KPS 评分的正确性为 84.95%。表 5 中前 4 组特征节点集的正确性达到或接近 85%,即认为是具有优秀分类能力的特征节点集。

表 5 中第 4、5 组特征节点集中同时包含个特征节点:HBV DNA 水平,外放边界和肿瘤分期 TNM3 时,Child-Pugh 比 KPS 评分更能提升正确性,即存在对 HBV 再激活的正确性贡献度:Child-Pugh>KPS 评分。

表 5 第 7 组特征节点集含有“甲胎蛋白 AFP”,其正确性略微高于没有“甲胎蛋白 AFP”的第 9 组;第 8 组特征节点集也存在“甲胎蛋白 AFP”,其正确性也略高于没有“甲胎蛋白 AFP”的第 10 组,这证明了“甲胎蛋白 AFP”增加了对 HBV 再激活正确性,与之前 CART 中的“甲胎蛋白 AFP”增加了分类性能结论一致。

表 5 中第 6 组特征节点集: HBV DNA 水平,外放边界和 KPS 评分的正确性为 84.24%;第 9 组特征节点集: HBV DNA 水平,外放边界和全肝平均剂量的正确性为 83.04%;第 10 组特征节点集: HBV DNA 水平,外放边界和 V20 的正确性为 82.74%;第 11 组特征节点集: HBV DNA 水平,外放边界和放疗总剂量的正确性为 80.95%;第 12 组特征节点集: HBV DNA 水平,外放边界和 V10 的正确性为 80.03%。由此推出对 HBV 再激活的正确性贡献度:KPS 评分>全肝平均剂量>V20>放疗总剂量>V10,这与之前 CART 得出的正确性贡献度一致。

特征节点 KPS 评分越高则表明放疗后身体所能承受的副作用越强,致使 HBV 再激活的可能性越低,即预测结果表现为正常(Normal),反之为激活(Reactivation)。剂量参数 V20、V10 等代表了放疗与肝损伤的关系,V20、V10 分别指接受 20Gy 或 10Gy 以上放疗的体积占全肝体积比例,放射性损伤不仅与受到的肝放射性耐受剂量存在着紧密联系,而且与 HBV 再激活存在紧密联系。

对 HBV 再激活影响越大的危险因素被 CART 选作特征节点的可能性越大,实验中特征节点以及出现的次数见表 6。

表 6　特征节点与出现次数

Table 6　Feature nodes and the number of occurrences

特征节点	出现次数
HBV DNA 水平	11
外放边界	11
KPS 评分	4
V20	3
放疗总剂量	2
V10	2
肿瘤分期 TNM	2
甲胎蛋白 AFP	2
全肝平均剂量	1
Child-Pugh	1
全肝最大剂量	1
年龄	1

综上所述,不同特征节点集的分类性能不同,得到一个正确性较高的特征节点集:HBV DNA 水平,外放边界,放疗总剂量,V20 和 KPS 评分,并对实验中特征节点集的正确性比较后得到一组对 HBV 再激活正确性贡献度的排序:KPS 评分>全肝平均剂量>V20>放疗总剂量>V10。特征节点“甲胎蛋白 AFP”也增加了 HBV 再激活的正确性。已知的危险因素:HBV DNA 水平和外放边界在所有 CART 特征节点中都出现,证明了 CART 算法特征选择的有效性。

3　结　论

1)本文提出的 CART 算法应用在原发性肝癌患者精确放疗后致 HBV 再激活的特征节点集(危险因素)分析中,并建立了 CART 和 Bayes 预测模型。实验结果显示两种预测模型对原发性肝癌患者精确

放疗后 HBV 再激活有着较强的模式判别能力,且 CART 的分类性能优于 Bayes 的分类性能。CART 选择的特征节点集提高了 HBV 再激活分类性能,尤其特征节点集是: HBV DNA 水平、外放边界、放疗总剂量、V20 和 KPS 评分时的分类性能达到最优。经过实验结果的比较,得到了对 HBV 再激活正确性贡献度的排序:KPS 评分>全肝平均剂量>V20>放疗总剂量>V10。“甲胎蛋白 AFP”也会增加 HBV 再激活的正确性。已知的危险因素:HBV DNA 水平和外放边界在所有 CART 特征节点中都出现,证明了 CART 算法特征选择的有效性。

2)CART 的划分规则、特征节点的正确性贡献度、两种预测模型以及特征节点出现次数都可帮助医生对精确放疗的肝癌患者进行指导性治疗,并配合抗病毒和肝保护药物,防止 HBV 发生再激活,对提高患者的治疗效果,甚至防止 HBV 再激活导致的患者死亡具有重要意义。今后将继续研究其他特征选择方法和分类算法,致力于提高预测模型准确度。

参考文献 (References)

[1]EL-SERAG H B, RUDOLPH K L. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis[J]. Gastroenterology, 2007, 132(7): 2557-2576. DOI: 10.1053/j.gastro.2007.04.061.

[2]JUNG J H, YOON S M, KIM S Y, et al. Radiation-induced liver disease after stereotactic body radiotherapy for small hepatocellular carcinoma: clinical and dose-volumetric parameters[J]. Radiation Oncology, 2013, 81(1): 1-7. DOI: 10.1186/1748-717X-8-249.

[3]叶胜龙. 2013 年肝癌领域新进展[J]. 中华肝脏病杂志, 2014, 22(1):2-4. DOI: 10.3760/cma.j.issn.1007-3418.2014.01.002.
YE Shenglong. New advances in liver cancer research: A review of 2013[J]. Chinese Journal of Hepatology, 2014, 22(1): 2-4. DOI: 10.3760/cma.j.issn.1007-3418.2014.01.002.

[4]黄伟,卢彦达,张炜,等. 原发性肝癌精确放疗致乙型肝炎病毒再激活分析[J]. 中华放射肿瘤学杂志, 2013, 22(3):193-196. DOI: 10.3760/cma.j.issn.1004-4221.2013.03.006.
HUANG Wei, LU Yanda, ZHANG Wei, et al. Analysis of hepatitis B virus reactivation induced by precise radiotherapy in patients with primary liver cancer[J]. Chinese Journal of Radiation Oncology, 2013, 22(3): 193-196. DOI: 10.3760/cma.j.issn.1004-4221.2013.03.006.

[5]HUANG Wei, ZHANG Wei, FAN Min, et al. Risk factors for hepatitis B virus reactivation after conformal radiotherapy in patients with hepatocellular carcinoma[J]. Cancer Science, 2014, 105(6): 697-703. DOI:10.1111/CAS.12400.

[6]汪孟森. 原发性肝癌三维适形放疗致乙型肝炎病毒再激活相关研究[D]. 济南:济南大学, 2014.
WANG Mengsen. Reactivation of hepatitis B virus following three-dimensional conformal radiotherapy for primary hepatic carcinoma[D]. Jinan: University of Jinan, 2014.

[7]张晶晶, 曲颂, 余建荣,等. 原发性肝癌三维适形放疗致乙型肝炎病毒再激活相关研究[J]. 癌症进展, 2015, 13(2):183-187. DOI:10.11877/j.issn.1672-1535.2015.13.02.16.
ZHANG Jingjing, QU Song, YU Jianrong, et al. Related factors of reactivation of hepatitis B virus induced by three dimensional conformal radiotherapy in primary liver cancer [J]. Oncology Progress, 2015, 13(2): 183-187. DOI:10.11877/j.issn.1672-1535.2015.13.02.16.

[8]吴冠朋, 王帅, 黄伟,等. 基于 BP 神经网络的肝癌放疗致乙型肝炎病毒再激活分类预测模型[J]. 智能计算机与应用, 2016, 6(2): 43-47. DOI: 10.3969/j.issn.2095-2163.2016.02.014.
WU Guanpeng, WANG Shuai, HUANG Wei, et al. Classification prognosis model of hepatitis B virus reactivation after radiotherapy in patients with primary liver carcinoma based on BP neural network[J]. Intelligent Computer and Applications, 2016, 6(2): 43-47. DOI: 10.3969/j.issn.2095-2163.2016.02.014.

[9]WU Guanpeng, LIU Yihui, WANG Shuai, et al. The classification prognosis models of hepatitis b virus reactivation based on Bayes and support vector machine after feature extraction of genetic algorithm[C]//Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). Changsha: IEEE, 2016: 572-577. DOI: 10.1109/FSKD.2016.7603236.

[10]BREIMAN L I, FRIEDMAN J H, OLSHEN R A, et al. Classification and Regression Trees (CART)[J]. Biometrics, 1984, 40(3):17-23.DOI: 10.2307/2530946.

[11]陈磊, 刘毅慧. 基于 CART 算法的肺癌微阵列数据的分类[J]. 生物信息学, 2011, 9(3): 229-234. DOI:10.3969/j.issn1672-5565.2011.03.013.
CHEN Lei, LIU Yihui. Classification based on CART algorithm for microarray data of lung cancer[J]. China Journal of Bioinformatics, 2011, 9(3): 229-234. DOI:10.3969/j.issn1672-5565.2011.03.013.

[12]KONG A L, PEZZIN L E, NATTINGER A B. Identifying patterns of breast cancer care provided at high-volume hospitals: a classification and regression tree analysis[J]. Breast Cancer Research & Treatment, 2015,153(3):689-698. DOI: 10.1007/s10549-015-3561-6.

[13]GASPAROVICA-ASITE M, POLAKA I, ALEKSEYEVA L. The impact of feature selection on the information held in bioinformatics data [J]. Information Technology & Management Science, 2016, 18(1):115-121. DOI: https://doi.org/10.1515/itms-2015-0018.

[14]RICHETTE P, CLERSON P, BOUÉE S, et al. Identification of patients with gout: elaboration of a questionnaire for epidemiological studies [J]. Annals of the Rheumatic Diseases, 2014, 74(9):1684-1690. DOI: 10.1136/annrheumdis-2013-204976.

[15]DAVIS M, ABRAMS M T, WISSOW L S, et al. Identifying young adults at risk of Medicaid enrollment lapses after inpatient mental health treatment[J]. Psychiatric Services, 2014, 65(4):461-468.DOI: 10.1176/appi.ps.201300199.

[16]HOU Yi, EDARA P, SUN C. Modeling mandatory lane changing using bayes classifier and decision trees[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(2):647-655.DOI: 10.1109/TITS.2013.2285337.