DOI:10.12113/j.issn.1672-5565.201711002

基于近邻成分分析算法的原发性肝癌精确放疗后HBV再激活分类预测

赵咏旺[1],刘毅慧[1*],黄 伟[2*]
(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:原发性肝癌(PLC)患者在精确放疗后乙型肝炎病毒(HBV)再激活是一种常见并发症,及时的预测防护能降低发病率、死亡率。研究表明:多余的特征变量会影响HBV再激活的预测精度。通过提出基于近邻成分分析(NCA)的特征选择方法找出HBV再激活的危险因素及特征组合。之后分别建立经Bayes优化前后的支持向量机模型(SVM)对这些关键特征子集及初始特征集进行分类预测。实验结果表:明HBV DNA水平、KPS评分、分割方式、外放边界、V25、肿瘤分期TNM、Child-Pugh等都是影响HBV再激活的危险因素。其中经NCA特征选择之后发现的V25是在乙型肝炎病毒再激活研究中首次提出的危险因素。10折交叉验证下特征组合HBV DNA水平、外放边界、V25的预测精度高达86.11%。支持向量机分类器可以很好的应用于乙型肝炎病毒再激活的研究,特征选择后的关键特征组合具有更优越的分类性能。

关键词:乙型肝炎病毒(HBV);近邻成分分析(NCA);特征选择;支持向量机
中图分类号:TP391 **文献标志码**:A **文章编号**:1672-5565(2018)03-163-07

Classification and prediction of HBV reactivation after radiotherapy of primary liver cancer based on neighborhood component analysis

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

Abstract: Hepatitis B virus (HBV) reactivation is a common complication in patients with primary liver cancer (PLC) after precise radiotherapy. Predictive protection can reduce morbidity and mortality. In this paper, we first found the risk factors and characteristics of HBV reactivation by a new feature selection method based on neighborhood composition analysis. Then the support vector machine classifier (SVM) was established before and after Bayes optimization. Finally, these key feature subsets and initial feature sets were classified and predicted. The results showed that HBV DNA level, KPS score, segmentation, extroversion border, V25, tumor staging TNM, and, Child-Pugh, etc. are the risk factors of HBV reactivation. V25 found in this paper after NCA feature selection was the risk factor firstly proposed in the study of HBV reactivation. The prediction accuracy of the characteristics of the combination of HBV DNA levels, extroverted border, and V25 under 10 fold cross validation were up to 86.11%. Support vector machine classifier can be applied to the study of HBV reactivation, and the key feature combination after feature selection has better classification performance.

Keywords: Hepatitis B virus (HBV); Neighborhood component analysis (NCA); Feature selection; Support vector machine

原发性肝癌(Primary carcinoma of liver)是我国 常见的恶性肿瘤之一。中国的原发性肝癌患者约占

收稿日期:2017-11-21;**修回日期**:2018-04-27.
基金项目:国家自然科学基金项目(No.81402538);国家自然科学基金项目(No.61375013);山东省自然科学基金项目(No.ZR2013FM020).
作者简介:赵咏旺,男,硕士研究生,研究方向:智能信息及图像处理技术;E-mail:624639406@qq.com.
*通信作者:刘毅慧,女,博士,教授,研究领域:生物计算,智能信息处理;E-mail:yxl@sdili.edu.cn.
黄伟,男,博士,副主任医师;研究方向:肿瘤精确放射治疗的临床与基础研究;E-mail:alvinbird@163.com.

世界的 55%,这类患者常伴有较高的死亡率。而原发性肝癌在接受精确放疗后易引起乙型肝炎病毒(HBV)再激活,所以找到 HBV 再激活的危险因素并通过进一步建立分类预测模型来进行研究对感染 HBV 的原发性肝癌患者具有重要的临床意义。国际上关于 HBV 再激活的研究不是很多,在国内,韩聚强等人指出 HBV 再激活与肿瘤直径大小及是否术前规范抗病毒治疗等因素有关[1]。汪孟森通过研究比较基线特征差异筛选出 HBV 再激活的可能危险因素,结果显示性别、年龄等指标无明显差异,肝功能 Child-pugh 分级可能是 HBV 再激活的危险因素[2]。黄伟在 69 例原发性肝癌患者接受精确放疗后致使乙型肝炎病毒再激活研究中发现基线血清 HBV DNA 水平和放疗剂量是 HBV 病毒再激活的独立危险因素[3]。吴冠朋在以前发现的危险因素的基础上随后又建立了 RBF 神经网络模型,识别率提高到 80%[4]。随后在论文[5-7]中通过遗传算法发现 HBVDNA 水平,肿瘤分期 TNM,Child-Pugh,外放边界,外放边界编码,V45 和全肝最大剂量是乙肝病毒再激活的危险因素。王会娜[8-11]研究表明利用随机森林方法选取的 HBV DNA 水平、TNM 肿瘤分期、V10、V20、外放边界这 5 个关键特征作为致使乙肝病毒再激活的危险因素组合,进行小波变换后,3 折交叉验证下预测精度最高达到82.96%。采用顺序后向选择方法发现 KPS 评分、HBV DNA 水平、外放边界、TNM、全肝最大剂量是乙肝病毒再激活的危险因素,采用 3 折交叉验证,预测精度达到 85.68%。而采用顺序前向选择方法发现性别、KPS 评分、HBV DNA 水平、HBeAg、外放边界两分类编码是乙肝病毒再激活的危险因素,5 折交叉验证下的贝叶斯分类预测精度达到 84.06%。

特征选择是从原始特征中选择出一些最有效特征以降低数据集维度的过程,是提高学习算法性能的一个重要手段,也是模式识别中关键的数据预处理步骤[12]。通过近邻成分分析法(NCA)来对原发性肝癌患者的原始数据集进行特征选择,然后通过分别建立经 Bayes 优化前后的支持向量机分类器(SVM)对特征数据集进行分类预测。

1 NCA 算法原理

NCA 算法就是一种简单有效的距离测度学习算法[13]。一个样本空间 $S^D=\{(x_i, y_i), i=1,2,3,4,\cdots, n\}$, x_i 是输入样本, y_i $(1,2,3,\cdots,c)$ 是分类标签。本研究中由山东省肿瘤医院提供的患者临床数据集中,$n=90$,(90 例患者样本),$c=2$(2 分类问题),$D=28$(28 个特征)。标签数据 y_1 表示 HBV 未激活,代表良性,标签数据 y_2 表示 HBV 再激活,代表恶性。

首先考虑一个随机分类器,随机从 S 选取一个点 $Ref(x)$,作为 x 的参考点,类似于 1-NN 分类器,参考点就是点 X 的最近邻。概率 $P(Ref(x_i)=x_j \mid s)$ 表示从 S 中选取的点 x_j 是离 x_i 最近的那个点,也就是 x_j 是 x_i 的参考点。这是由距离函数来判断的,如式(1)所示:

$$d_w(x_i, x_j)=\sum_{r=1}^{n} w_r^2 \mid x_{ir}-x_{jr} \mid \tag{1}$$

$$P(\mathrm{Ref}(x_i)=x_j \mid s)=\frac{k(d_w(x_i,x_j))}{\sum_{j=1}^{n} k(d_w(x_i,x_j))} \tag{2}$$

w_r 表示第 r 个特征向量的权重系数。k 是一个内核函数,$K(z)=\exp(\frac{-z}{\sigma})$, σ 是内核的宽度,而这个 σ 值就影响每个数据点被选为参考点的概率。当 σ 值 $\to 0$ 时,则只有离样本点最近的那个数据点被选作参考点,当 σ 值 $\to \infty$时,S 中所有的点都有相同概率被选作为参考点。

考虑经留一法交叉验证的随机分类器,点 x_j 是点 x_i 参考点的概率如式(3)所示:

$$\begin{cases} p_{ij}=P(Ref(x_i)=x_j \mid S)= \\ \dfrac{k(d_w(x_i,x_j))}{\sum_{j=1,j\neq i}^{n} k(d_w(x_i,x_j))}, \text{if} \quad i\neq j \\ 0, \text{if}_i=j \end{cases} \tag{3}$$

随机分类器正确分类观察样本 i 的概率 p_i 为式(4):

$$p_i=\sum_{j=1,j\neq i}^{n} p_{ij}\, y_{ij},\ y_{ij}=\begin{cases}1,(y_i=y_j)\\0,\text{otherwise}\end{cases} \tag{4}$$

使用随机分类器正确分类的平均概率 $F(w)$ 值取决于权重向量 w。近邻成分分析的目标就是最大化 $F(w)$,$F(w)=\sum_{i=1}^{n} p_i-\lambda\sum_{r=1}^{p} w_r^2=\sum_{i=1}^{n} F_i(w)$ 。λ 是正则化参数,作为一个重要的参数,在接下来的研究中,经优化的 λ 值将作为 NCA 预测分类的重要指标。

2 特征选择

2.1 数据选取

研究数据是采取的山东省肿瘤医院的 90 例经过精确放疗后原发性肝癌患者的临床资料作为研究样本,90 例样本中 20 例发生 HBV 再激活。每个样本包含性别、放疗前 TACE 次数、TNM、V20、HBV

DNA 水平、外放边界、分割方式等 28 个特征[14],组成 90×28 维大小的数据集,详细见表 1。

表 1　特征编号及分别对应的医学名称

Table 1　Feature numbers and the corresponding medical name

特征编号	医学参数
1	性别
2	年龄
3	KPS 评分
4	HbeAg
5	门脉癌栓有无
6	肿瘤分期 TNM
7	Child-Pugh
8	甲胎蛋白 AFP
9	HBV DNA 水平
10	放疗总剂量
11	等效生物剂量
12	放疗次数
13	放疗前 TACE
14	分割方式
15	GTV 体积(gross tumor volume)
16	PTV 体积(planning target volume)
17	外放边界
18	V5
19	V10
20	V15
21	V20
22	V25
23	V30
24	V35
25	V40
26	V45
27	全肝最大剂量
28	全肝平均剂量

2.2　参数调整

建立 NCA 模型训练预测计算 λ 值对应的损失函数值采取近邻成分分析(NCA)对原始特征空间进行特征选择,选出致 HBV 再激活的危险因素[14],组成新的关键特征子集。在这之前先采用 k 折交叉验证来调整 NCA 的正则化参数 λ ,具体工作如图 1 所示。

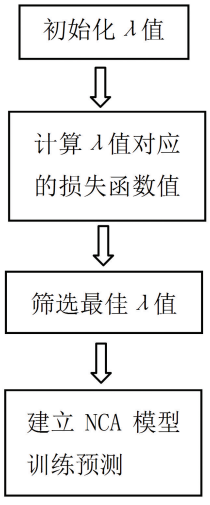

图 1　NCA λ 值调整

Fig.1　Process of adjusting the λ value

本文采取的 λ 值是从 0 开始到 $10/n$(5 折交叉验证,4/5 样本数据作训练,$n=72$)的等间距的 20 个点,步长就是 $10/n/(n-1)$, λ 值如下表 2 所示:

3　构建支持向量机(SVM)预测模型

3.1 模型训练

支持向量机(SVM)[15]是一种可用于二进制分类或回归的监督学习算法,属于一种机器学习算法,也称为内核机器。支持向量机(SVM)训练有两个阶段:

(1)将样本数据的特征向量转化为高维特征空间,这个过程就是内核技巧。

表 2　正则化参数 λ 值选择

Table 2　Regularization parameterλvalue to choose

编号	λ 值	编号	λ 值
1	0	11	0.083 5
2	0.0084	12	0.091 9
3	0.0167	13	0.100 3
4	0.025 1	14	0.108 6
5	0.033 4	15	0.117 0
6	0.041 8	16	0.125 3
7	0.050 1	17	0.133 7
8	0.058 5	18	0.142 0
9	0.066 8	19	0.150 4
10	0.075 2	20	0.158 7

(2)求解二次优化问题以适应最优超平面将变换后的特征分为两类。变换特征的数量由支持向量的数量确定。所谓支持向量是指那些在间隔区边缘的训练样本点。如图 2 所示。

在 SVM 理论中,需要考虑的就是能够让所有点中离它最近的点具有最大间距[16-17]。本文中乙型肝炎病毒激活人群的样本数量为 20,此类样本标记为 $Y=1$,未激活人群的样本数量为 70,此类样本标记为 $Y=2$。样本数量 $P=90$,超平面为 $wx+b=0$;样本点到超平面距离为:

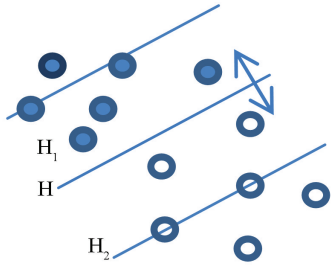

图 2　支持向量机模型

Fig.2　Support vector machine model

$$\frac{t_i \cdot t(x_i)}{\| w \|} = \frac{t_i \cdot (w^p \cdot \theta(x_i) + b)}{\| w \|} \tag{5}$$

首先构造并求解约束最优化问题：

$$\min (a): \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j t_i t_j(\theta(x_i) \cdot \theta(x_j)) - \sum_{i=1}^{p} a_i \tag{6}$$

求得最优解 a^*，然后计算 $w^* = \sum_{i=1}^{p} a_i^* t_i \theta(x_i)$，$b^* = t_i - \sum_{i=1}^{p} a_i^* t_i(\theta(x_i) \cdot \theta(x_j))$，最后求得分类决策函数：

$$f(x) = \text{sign}(w^* \theta(x) + b^*) \tag{7}$$

采用不同的核函数将导致不同的 SVM 算法。常用的核函数有：线性核函数，sigmoid 核函数，RBF 核函数，多项式核函数，二层感知器核函数等。本文通过分别建立优化前后的支持向量机（SVM）模型进行比较，前后采用的都是 RBF 核函数，对应函数为：

$$K(X_1, X_2) = \exp(-\frac{| X_1 - X_2 |^2}{2\sigma^2}) \tag{8}$$

3.2 Bayes 全局优化算法

贝叶斯全局优化算法目的是将有界目标函数 $F(X)$ 最大化，$F(X)$ 可以是确定的，也可以是不确定函数。根据已有采样点来构建一个高斯过程回归模型（Gaussian process）预估函数最大值的一个算法[18]。假设未知点也都服从多变量高斯分布，根据多变量高斯分布的一些性质，可以计算出这些点的均值 $\mu_i(x)$ 和标准差 $\sigma_i(x)$。根据加和公式（9）选择均值 $\mu_i(x)$ 和标准差 $\sigma_i(x)$ 的加和最小输入位置点作为下一个取样点。如果标准差值 $\sigma_i(x)$ 大，表示我们对该点了解甚少，多去采样类似点可以更好地确定目标函数形态。如果均值 $\mu_i(x)$ 大，表示该点可能是最大值位置，多去采样类似点可以帮助我们尽快锁定最大值[19]。而贝叶斯优化算法就是协调确定目标函数形态以及确定目标函数最大值这两个目标之间的矛盾。前期算法会采样标准差大的点来尽量确定目标函数形态。随着采样点增多，对函数大致熟悉之后，标准差值会下降。所以后期采样点会尽量选取均值大的点，这样就会有更大概率接近最大值[20]。贝叶斯全局优化算法寻找最优点过程图 3 所示。

加和公式为：

$$G(x_{i+1}) = \text{argmax}\, \mu_i(x) + \beta_t^{\frac{1}{2}} \sigma_i^2(x) \tag{9}$$

均值函数为：

$$\mu_i(x) = k^t(x)\, k^{-1} y \tag{10}$$

预测分布的标准差为：

$$\sigma_i(x) = \sqrt{k(x,x) - k^t(x)\, k^{-1} k(x)} \tag{11}$$

其中 $k(x)$ 是测试样本与训练样本输入值间的 $m \times 1$ 维协方差向量。$k(x,x)$ 是测试样本输入值和它自身的方差。K 为 $m \times m$ 维训练样本间的协方差矩阵。

最小化 $F(X)$ 的关键因素包括以下几个方面：

（1）$F(X)$ 是一个高斯过程模型

（2）通过先验值，可以决定下一个采样点所对应的 y 值

（3）已知点服从多变量高斯分布，假定 $A = \{(x_i, y_i)\}$，$i = 1,2,3\cdots,m$，$x_i \in R^v$，$y_i \in R$. m 表示训练样本个数，v 代表特征向量的维度，本文在 3 折、5 折、10 折交叉验证下 m 分别取 60，72，81，$v = 28$，贝叶斯优化是根据加和公式（9）来选择下一个采样点 x_{i+1}。

进行贝叶斯优化调整的 SVM 的参数主要有两个‘sigma’及‘box’[20]。在高斯 RBF 核函数中，‘sigma’的值就是内核的规模，Sigma 值越大，分离面就越平滑；Sigma 值越小，分离面就越细致。‘box’的值就是框式约束范围。这两个参数初始值设置的尽量广泛，因为具体的最优值不能确定。本文中‘sigma’及‘box’的初始范围都设置在 $10^{-4} \sim 10^4$。

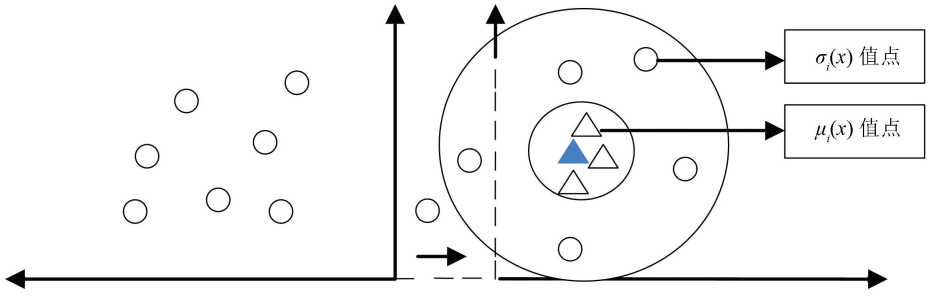

图 3 Bayes 全局优化算法

Fig.3 Bayesian global optimization algorithm

3.3　分类性能度量

本文主要采用三个分类性能指标,分别是准确性、特异性、灵敏性。准确性是指分类的正确预测值占样本实际值的比重。特异性是将实际无病的人正确判定为真阴性的比例。灵敏性是将实际有病的人正确判定为真阳性的比例。

4　实验结果及分析

利用近邻成分分析法(NCA)进行特征选择,得出所有最优特征子集规模为 1-5 的所有特征组合。分别将特征选择之后的最优特征子集代入优化前后的 SVM 分类器进行分类预测。为了验证特征选择对提高分类准确率的重要性,本实验加入对初始特征集的预测以求与最优特征子集的结果形成对比。实验分别采用 3 折、5 折、10 折交叉验证的方法对程序运行 50 次之后,选取每一个分类性能度量标准的平均值作为最终数据,具体实验结果见表 3~表 5。表 3 是优化前的 SVM 分类器针对最优特征子集的分类预测结果,加黑的数据代表分类性能比较好的特征组合。表 4 是优化后的 SVM 分类器针对表格 3 中分类表现比较好的几个特征子集分类预测的结果。表 5 列出了优化调整后的参数值。表格中出现的特征编号及所对应的医学参数详见表 6。

表 3　优化前的 SVM 分类预测结果

Table 3　Classification and prediction results of unoptimized SVM

最优特征子集 初始特征集	准确率(3 折/5 折/10 折)	特异性(3 折/5 折/10 折)	灵敏性(3 折/5 折/10 折)
9 17	**0.800 0/0.843 3/0.815 6**	**0.939 1/0.961 4/0.951 4**	**0.342 9/0.430 0/0.340 0**
9	0.755 3/0.747 8/0.753 3	0.969 6/0.937 1/0.962 9	0.051 4/0.085 0/0.020 0
3 9 17	**0.805 3/0.818 9/0.804 4**	**0.972 2/0.970 0/0.960 0**	**0.257 1/0.290 0/0.260 0**
6 9	0.800 7/0.815 6/0.806 7	0.967 0/0.972 9/0.965 7	0.254 3/0.265 0/0.250 0
6 7 9 17	0.786 0/0.803 3/0.784 4	0.955 7/0.938 6/0.931 4	0.228 6/0.330 0/0.270 0
3 9	**0.824 0/0.836 7/0.837 8**	**0.974 8/0.968 6/0.980 0**	**0.328 6/0.375 0/0.340 0**
3 9 14	**0.802 7/0.811 1/0.813 3**	**0.945 2/0.942 9/0.957 1**	**0.334 3/0.350 0/0.310 0**
9 14 17	0.792 0/0.808 9/0.833 3	0.950 4/0.945 7/0.954 3	0.271 4/0.330 0/0.410 0
6 9 17	0.779 3/0.795 6/0.817 8	0.950 4/0.944 3/0.960 0	0.217 1/0.275 0/0.320 0
6 9 14 17	0.784 0/0.793 3/0.802 2	0.960 0/0.958 6/0.965 7	0.205 7/0.215 0/0.230 0
3 6 9 17	0.792 7/0.802 2/0.802 2	0.967 0/0.957 1/0.960 0	0.220 0/0.260 0/0.250 0
6 7	0.788 7/0.810 0/0.808 9	0.957 4/0.952 9/0.962 9	0.234 3/0.310 0/0.270 0
7 9	0.766 7/0.782 2/0.786 7	0.944 3/0.947 1/0.934 3	0.182 9/0.205 0/0.270 0
3 6 7 9 17	0.770 7/0.791 1/0.788 9	0.973 0/0.982 9/0.982 9	0.105 7/0.120 0/0.110 0
6 7 9	0.756 7/0.795 6/0.795 6	0.933 9/0.955 7/0.960 0	0.174 3/0.235 0/0.220 0
3 6 9	0.773 3/0.784 4/0.773 3	0.950 4/0.947 1/0.951 4	0.191 4/0.215 0/0.150 0
9 14	0.748 7/0.730 0/0.726 7	0.960 0/0.922 9/0.928 6	0.054 3/0.055 0/0.020 0
9 17 22	**0.820 0/0.840 0/0.861 1**	**0.963 5/0.964 3/0.988 6**	**0.348 6/0.405 0/0.470 0**
3 6 9 14 17	0.795 3/0.806 7/0.822 2	0.967 0/0.967 1/0.974 3	0.231 4/0.245 0/0.290 0
3 7 9 17	0.786 0/0.793 3/0.804 4	0.969 6/0.968 6/0.982 9	0.182 9/0.180 0/0.180 0
6 7 9 14	0.745 3/0.756 7/0.731 1	0.933 0/0.934 3/0.911 4	0.128 6/0.135 0/0.100 0
初始特征集	0.766 7/0.777 8/0.777 8	1.000 0/1.000 0/1.000 0	0.000 0/0.000 0/0.000 0

表 4　Bayes 优化的 SVM 分类预测结果

Table 4　Classification and prediction results of Bayesian optimization SVM

最优特征子集	准确率(3 折/5 折/10 折)	特异性(3 折/5 折/10 折)	灵敏性(3 折/5 折/10 折)
9 17	0.812 7/0.831 1/0.855 6	0.944 3/0.957 1/0.974 1	0.380 0/0.390 0/0.450 0
9	0.766 7/0.774 4/0.777 8	0.963 5/0.995 7/1.000 0	0.054 3/0.005 0/0.000 0
3 9 17	0.798 0/0.821 1/0.824 4	0.985 2/0.985 7/0.988 6	0.182 9/0.245 0/0.250 0
3 9	0.829 3/0.838 9/0.844 4	0.973 0/0.968 6/0.961 0	0.357 1/0.385 0/0.440 0
3 9 14	0.815 3/0.826 7/0.842 2	0.944 9/0.962 9/0.961 9	0.351 4/0.390 0/0.430 0
9 17 22	0.820 7/0.843 3/0.835 6	0.923 5/0.922 9/0.914 3	0.480 0/0.480 0/0.510 0

表 5 经 Bayes 优化后的参数值

Table 5 Bayesian optimized parameter values

最优特征子集	*Sigma* 值	*Box* 值
9 17	17.998 4	$9.243\ 1\times10^{3}$
9	48.784 0	$3.878\ 5\times10^{3}$
3 9 17	$1.011\ 6\times10^{-4}$	0.619 9
3 9	$2.382\ 9\times10^{-4}$	$6.337\ 7\times10^{3}$
3 9 14	0.009 4	322.770 7
9 17 22	43.002 1	$2.825\ 6\times10^{4}$

表 6 中列出的这些医学参数就是经 NCA 特征选择出的对 HBV 再激活有着重要影响的危险因素。表 1 所列出的特征子集组合中,原始数据集下的分类预测精度、特异性、灵敏性都是最低的,预测精度在 77% 左右。特征编号 9 出现的频率是最多的,也就是说 HBV DNA 水平是影响适型放疗后 HBV 再激活的最关键因素。表 3 中在 5 折交叉验证下,特征子集组合 9、17 的分类预测精度为 84.33%,特征子集组合 3、9 的分类预测精度为 83.67%,特征子集组合 9、17、22 的分类预测精度为 84.00%。在 10 折交叉验证下,特征子集组合 3、9 的分类预测精度为83.78%,特征子集组合 9、17、22 的分类预测精度可达 86.11%。以上这几个特征子集组合的预测精度要明显高于其他特征子集组合的预测精度,由此可见影响 HBV 再激活的关键危险因素除了 HBV DNA 水平外还有 KPS 评分、分割方式、外放边界、V25 等。而通过临床灵敏度水平来看,特征子集组合为 9、17、22 的表现要优于其他特征组合,在 10 折交叉验证下平均精确度可达 47%。

表 6 危险特征编号及所对应的特征因子

Table 6 Danger signature number and corresponding feature factors

特征编号	特征因子
3	KPS 评分
6	肿瘤分期 TNM
7	Child-Pugh
9	HBV DNA 水平
14	分割方式
17	外放边界
22	V25

表 4 是关键特征子集在经 Bayes 优化之后的 SVM 模型中的分类表现,由数据可知,当 HBV DNA 水平作为独立危险因素来表现时,经过优化之后的 SVM 预测准确度在 3 折、5 折、10 折交叉验证结果下都要高于未优化的 SVM 预测准确度。其中 5 折交叉验证下,优化后比优化前提高了 2.66 个百分点。由 KPS 评分、HBV DNA 水平、外放边界组成的危险因素组合在 10 折交叉验证下,优化后比优化前提高了 2 个百分点。由 KPS 评分、HBV DNA 水平、分割方式组成的危险因素组合在 10 折交叉验证下,优化后比优化前提高了 2.89 个百分点。而由 HBV DNA 水平、外放边界组成的危险因素组合在 10 折交叉验证下,优化后比优化前提高了 4 个百分点。从临床灵敏度来看,在 10 折交叉验证下,所有关键特征子集组合在优化后的预测精度明显高于优化前,可见 Bayes 优化调整支持向量机参数对于提高真阳性病人的正确诊断率是十分必要的。综上所述, HBV DNA 水平可以当作影响 HBV 再激活的最危险因素,KPS 评分、外放边界、V25 是影响 HBV 再激活的关键因素,而分割方式、肿瘤分期 TNM、Child-Pugh 也是影响 HBV 再激活的重要因素。

5 结束语

前预防乙型肝炎病毒再激活,降低其发病概率,延长患者生命就需要找出原发性肝癌患者在接受精确放疗之后乙肝病毒再激活的关键特征,本文就是通过近邻成分分析算法发现 HBV DNA 水平、KPS 评分、分割方式、外放边界、V25、肿瘤分期 TNM、Child-Pugh 等都是影响 HBV 再激活的危险因素。再通过建立的优化前后的支持向量机模型分别对提取出的所有不同特征的组合进行分类预测,结果表明 HBV DNA 水平、外放边界、V25 组成的特征子集的分类表现要优于其他组合。10 折交叉验证下特征向量 HBV DNA 水平、外放边界、V25 组合的预测精度高达 86.11%。近邻成分分析法是一种有效的特征选择方法,可以对临床医学的研究提供一定的帮助。

参考文献(References)

[1]韩聚强, 任永强, 李国安. 原发性肝癌微创介入治疗术后 HBV 再激活及相关影响因素研究[J]. 中国医学前沿杂志:电子版, 2014, 6(3):27-30.
HAN Juqiang, REN Yongqiang, LI Guoan. Study on reactivation HBV and related influencing factors after minimally invasive interventional therapy for primary hepatic cancer[J]. Chinese Frontiers of Medicine: Electronic Etion,2014,6(3): 27-3.

[2]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].济南市:济南大学,2014.
WANG Mengsen. Study on reactivation of hepatitis B virus by three dimensional conformal radiotherapy for primary hepatic carcinoma[D]. Jinan: University of Jinan,2014.

[3]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.

[4]WU Guanpeng. Application of BP and RBF neural network in classification prognosis of hepatitis B virus reactivation[J]. Journal of Electrical and Electronic Engineering,2016,4(2):35.DOI: 10.11648/j.jeee.20160402.16.

[5]WU Guanpeng, LIU Yihui, WANG Shuai. The classification prognosis models of hepatitis b virus reactivation based on Bayes and support vector machine after feature extraction of genetic algorithm[C]. International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery,2016, 572-577.DOI:10.1109/FSKD.2016.7603236.

[6]吴冠朋, 刘毅慧, 王帅. 基于遗传算法特征选择的 HBV 再激活分类预测模型[J]. 生物信息学,2016,14(4):243-248.DOI:10.3969/j.issn.1672-5565.2016.04.08.
WU Guanpeng, LIU Yihui, WANG Shuai. Classification model of HBV reactivation based on genetic algorithm feature selection[J]. Chinese Journal of Bioinformatics, 2016,14(4):243-248.DOI:10.3969/j.issn.1672-5565.2016.04.08.

[7]吴冠朋, 王帅, 黄伟. 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.
WU Guanpeng, WANG Shuai, HUANG Wei. Classification model of hepatitis B virus reactivation based on BP neural network for radiotherapy of liver cancer[J]. Smart Computers and Applications, 2016,6(2):43-47.

[8]WANG Huina, LIU Yihui, HUANG Wei. The application of feature selection in hepatitis B virus reactivation[C]. IEEE International Conference On Big Data Analysis,2017.DOI:10.1109/ICBDA.2017.8078767.

[9]WANG Huina, HUANG Wei, LIU Yihui. Classification of hepatitis B virus reactivation after radiotherapy of primary liver cancer based on random forest[C]. International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD),2017 1th.

[10]王会娜, 黄伟, 刘毅慧. 原发性肝癌放疗后乙肝病毒再激活预测模型的特征降维分析[J].中国生物医学工程学报, 2017 (6):697-701.DOI:10.3969/j.issn.0258-8021.2017.06.009.
WANG Huina, HUANG Wei, LIU Yihui. The characteristic dimension reduction analysis of hepatitis B virus reactivation prediction model after radiotherapy for primary liver cancer [J]. Chinese Journal of Biomedical Engineering, 2017 (6):697-701.DOI:10.3969/j.is-sn.0258-8021.2017.06.009.

[11]王会娜, 黄伟, 刘毅慧. 基于连续小波和随机森林的原发性肝癌放疗后乙肝病毒再激活的分类预测[J]. 智能计算机与应用,2017,7(3):30-33.
WANG Huina, HUANG Wei, LIU Yihui. Classification prediction of hepatitis B virus reactivation after radiotherapy for primary liver cancer based on continuous wavelet and random forest[J]. Smart Computers and Applications, 2017,7(3):30-33.

[12]KIM B Y, DONG W C, WOO S R. Recurrence-associated pathways in hepatitis B virus-positive hepatocellular carcinoma[J]. BMC Genomics,2015,16(1):1-15.

[13]YANG Wei, WANG Kanquan, ZUO Wangmeng, et al. Neighborhood component feature selection for high-dimensional data[J]. Journal of Computers,2012,7(1):161-168. DOI:10.4304/jcp.7.1.161-168.

[14]WANG Shuai, WU Guanpeng, HUANG Wei, et al. The predictive model of hepatitis B virus reactivation induced by precise radiotherapy in primary liver cancer[J]. Journal of Electrical and Electronic En-Gineering,2016,4(2):31-34. DOI:10.11648/j.Jeee.20160402.15.

[15]ZHANG S, ZHANG S, JIN Z, et al. A novel SVM by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection [J]. Soft Computing, 2015, 19(5):1187-1199.

[16]宋晖, 薛云, 张良均. 基于 SVM 分类问题的核函数选择仿真研究[J]. 计算机与现代化,2011(8):133-136.DOI:10.3969/j.issn.1006-2475.2011.08.037.
SONG Hui, XUE Yun, ZHANG Liangjun. Research on selection of kernel function based on SVM classification problem[J]. Computer and Modernization,2011(8):133-136. DOI:10.3969/j.issn.1006-2475.2011.08.037.

[17]SARTAKHTI J S, ZANGOOEI M H, MOZAFARI K. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA) [J]. Computer Methods & Programs in Biomedicine, 2012, 108(2):570.

[18]ZHANG J, CHEN H, ZHOU H. Freeway travel time prediction research based on a deep learning approach[C]. International Conference on Advanced Materials and Information Technology Processing,2016,21-27. DOI:10.2991/amitp-16.2016.97.

[19]MARTINEZ-CANTIN R. BayesOpt: a Bayesian optimization library for nonlinear optmization, experimental design and bandits[J]. Journal of Machine Learning Research, 2014 , 15 :3735-3739.

[20]CARPIN M, ROSATI S, RIMOLDI B, et al. UAVs using Bayesian Optimization to Lo- cateWiFi Devices[C]. Bayesopt 2015, 2015.