文章编号：2095-2163(2021)08-0035-07 中图分类号：TP181；R319 文献标志码：A

基于数据挖掘的肝癌早期复发预测与阈值研究

刘海钰，曲海成
（辽宁工程技术大学 软件学院，辽宁 葫芦岛 125105）

摘 要：肝癌术后的早期复发率极高，然而肝癌术后症状不明显，传统医学手段判断肝癌是否早期复发的准确率并不理想。针对这一问题，本文选取医院就诊病人信息作为数据对象，采用数据挖掘技术，探究机器学习方法对肝癌早期复发的疾病预测应用效果，以及肝癌早期复发的最优阈值。首先对病人信息数据进行数据清洗等数据预处理操作，运用多个特征工程方法增加预测的精准性。将逻辑回归、随机森林、SVM 支持向量机以及 GBDT 梯度提升树 4 种模型进行比较，最终选择 GBDT 梯度提升树建立预测模型。选择准确率、精确率、召回率和 AUC 4 个指标对所建立的模型进行评估，并得出了肝癌早期复发的最优阈值，为医学领域的相关研究与临床应用提供了一定的参考。

关键词：数据挖掘；疾病预测；机器学习

Prediction and threshold of early recurrence of liver cancer based on data mining

LIU Haiyu, QU Haicheng
(College of software engineering, Liaoning technological university, Huludao Liaoning 125105, China)

【Abstract】 The early recurrence rate of liver cancer after operation is very high. However, the symptoms of liver cancer after operation are not obvious, and the accuracy of traditional medical methods to determine whether liver cancer has early recurrence is not ideal. In order to deal with this problem, the information of hospital patients was selected as the data object, and data mining technology was adopted to explore the application effect of machine learning method on disease prediction of early recurrence of liver cancer and the optimal threshold value of early recurrence of liver cancer. Firstly, data preprocessing operations such as data cleaning were carried out on the patient information data, and multiple feature engineering methods were used to increase the accuracy of the prediction. Logical regression, random forest, SVM support vector machine and GBDT gradient lifting tree were compared. Finally, GBDT gradient lifting tree was selected to build the prediction model. Four indexes of accuracy, precision, recall and AUC were selected to evaluate the established model, and the optimal threshold for early recurrence of liver cancer was obtained, which provided a certain reference for related research and clinical application in the field of medicine.

【Key words】 data mining; disease predictions; machine learning

0 引 言

肝癌是威胁人类身体健康的主要恶性肿瘤之一，对其研究至今已有百年历史。目前，根治肝癌最有效的手段是手术切除，而对于广大患者而言，肝癌术后效果仍然欠佳，有报告指出肝癌在 5 年内的复发转移率可到 40%~70%[1]。肝癌手术过后有两个复发高峰，1 年内的复发定义为早期复发，之后的复发定义为晚期复发。

近年来数据挖掘技术发展迅速，医学数据挖掘的目的是从大量的医学数据中挖掘出潜在并且有效的知识、信息、模型、关联和变化等，从而帮助医生进行更加快速和准确的诊断[2]。通过利用大量病人的各项信息来进行分析并得出结论的方法在医学界已经得到了广泛认可。文献[3]中探讨了随机森林算法在心血管疾病预测中的应用效果，并对其性能进行了评价；文献[4]中以随机森林算法为基础，采用交叉检验和网格搜索寻找最佳参数，建立了心脏病预测模型；文献[5]探讨了随机森林算法在产后抑郁影响因素的筛选和风险预测中的应用效果；文献[6]利用机器学习方法构建心血管疾病的预测模型，对心血管疾病进行快速高效的预测；文献[7]基于机器学习算法，对医疗数据进行了处理和分析；文献[8]提出了基于 Choquet 积分的数据挖掘模型的预测算法和模型组合的特征筛选方法，利用体检数据对某一类疾病高血压做预测，制定了基于大数据的疾病风险预测模型；文献[9]采用基于机器学习的分类判断算法，建立慢性阻塞性肺疾病分期模型；文献[10]运用决策树机器学习算法，建立慢性肝硬化疾病预测模型，得到了预测准确率达到 98%肝硬

作者简介：刘海钰(1999-)，男，本科生，主要研究方向：机器学习；曲海成(1981-)，男，博士，副教授，主要研究方向：智能图像处理、机器学习等。
通讯作者：曲海成 Email：quhaicheng@ lntu.edu.cn

收稿日期：2021-06-09

化预测模型;文献[11]通过构建机器学习模型,预测了肾病在人群中的流行程度;文献[12]中将数据挖掘技术与机器学习算法相结合,对心脏病患者进行预测。

综上所述,可见数据挖掘技术的应用面非常广泛,在疾病预测方面的表现尤为突出。然而,关于肝癌早期复发预测问题上的研究却相对较少。为此,本文从数据的筛选、数据的预处理、特征工程、建立预测模型以及模型评价中得出预测模型,使其能够准确预测肝癌是否早期复发的结论,并经过数据分析得出了肝癌早期复发的最优阈值。

1 模型构建方法

1.1 网格搜索法

网格搜索(gridsearchCV)是一种指定参数值的穷举搜索方法,也是机器学习中一种常用的调参方法。指定需要调整的参数,使其在指定的参数范围内,通过遍历所有组合选定参数,选择能够让模型得到最优结果的那个参数组合作为最终结果。本项目应用网格搜索法,为 GBDT 梯度提升树选择了两个最佳参数:树的数量(n_estimators)和学习率(learning_rate)。通过该方法可以得到最优的参数。但该方法进行的是一种穷举操作,所以在时间耗费上会相对较长。本文实验项目在 n_estimators:[30, 50, 80, 100]以及 learning_rate:[0.1, 0.05, 0.01]范围内确定了 GBDT 的最优参数。

1.2 梯度提升决策树

梯度提升决策树(Gradient Boosting Decision Tree,GBDT),又称做多重累计回归树[13]。其内部子树为 CART 树,基于 Boosting 算法集成思想提出,是机器学习、数据分析中最常见的预测模型方法之一。该算法选择决策树作为弱学习器。回归树大致流程为:在每一次分支的时候寻找能够实现最优分支的节点,作为分裂节点。在分类决策树中使用的是基尼系数等,在回归树中使用的是均方误差,直到分裂完毕或者满足了一定的条件。CART 决策树结构示意如图 1 所示。

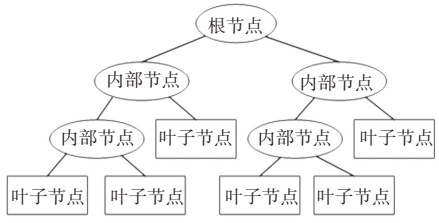

图 1 CART 决策树结构示意图

Fig. 1 The Structural diagram of CART decision Tree

Boosting 算法:使用已经给出的弱分类器线性组合,生成一个表现出强性能的强分类器的过程[14]。通过使用多个弱分类器,训练基分类器时采用串行方式,每个基分类器之间有依赖,其基本思路是将基分类器一个个叠加。每个基分类器在训练时,对前一个基分类器分错的样本给予更高的权重。测试时,根据各个分类器的结果加权得到最终结果。

GBDT 分类算法属于集成学习中的 Boosting 算法。其原理是:将所有弱分类器结果的总和作为预测值,下一个弱分类器去拟合误差函数对预测值的残差(残差就是预测值与真实值之间的误差)。其中弱分类器的表现形式就是各棵决策树[15],算法如下:

假设训练集样本:$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$,最大迭代次数为 T,输出的强学习器为 $f(x)$。则损失函数的表达为:

$$L(y,f(x))=\log(1+e^{-yf(x)}), \tag{1}$$

(1)初始弱学习器:

$$f(x)=\arg\min_c\sum_{i=1}^{m}L(y_i,c). \tag{2}$$

(2)对迭代次数 $t=1,2,\cdots,T$:

①对样本 $i=1,2,\cdots,m$ 计算负梯度误差:

$$r_{ti}=-\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}=\frac{y_i}{1+e^{(y_i,f(x_i))}}, \tag{3}$$

②利用 $(x_i,r_{ti})(i=1,2,\cdots,m)$,拟合一棵 CART 回归树,得到第 t 棵回归树。其对应的叶子节点区域为 $r_{tj}(j=1,2,\cdots,J)$。其中 J 为回归是 t 的叶子节点个数。

③对叶子区域 $j=1,2,\cdots,T$,计算最佳负梯度拟合值:

$$\begin{aligned}c_{tj}&=\arg\max_c M,\\ M&=\sum_{x_i=r_{ti}}\log(1+e^{-y_i(f_{t-1}(x_i)+c)})\\ &\approx\frac{\sum_{x_i\in r_{tj}}(r_{ti})}{\sum_{x_i\in r_{tj}}|r_{ti}|(1-|r_{ti}|)},\end{aligned} \tag{4}$$

更新强学习器:

$$f_t(x)=f_{t-1}(x)+\sum_{j=1}^{J}c_{tj}I(x\in R_{tj}), \tag{5}$$

得到强学习器表达式:

$$f(x)=f_T(x)=f_0(x)+\sum_{i-1}^{T}\sum_{j-1}^{J}c_{ij}I(x\in R_{ij}). \tag{6}$$

2　预测模型构建

2.1　数据预处理

在进行实验数据处理前，通过查看原始数据表，以此来确认数据表的格式、内容种类等信息，从而选择合适的数据处理方式。本文选择使用 Excel 查看数据，可以发现数据是标准的行列式表格数据。其中包含表示 *ID*、性别等信息的分类型变量，也有肿瘤数量、肿瘤大小等表示医学指标的数值型变量。由于这些数据变量的类型过于冗杂，因此需要将这些冗杂的变量类型统一为可以直接输入到预测模型中的连续型数值变量。部分预处理前的数据见表 1。

表 1　预处理前部分数据展示

Tab. 1　Partial data presentation before data preprocessing

ID	肿瘤大小 a/cm	肿瘤大小 c/cm	肿瘤数量/个
1	2.5	NaN	3.0
4	3.4	NaN	28.0
5	3.6	NaN	1.0
30	2.4	1.8	1.0

注：NaN 表示数据缺失。

2.1.1　数据清洗

数据清洗主要包括：将数据规范成合适的数据表现形式、去除空值、重复值、异常值、噪声数据剔除等。其中包含两方面的工作：一个是对数据进行异常值检测，查看数据是否在合理范围内变动，有无超出固定范围的数据，是否有矛盾数据以及可以替换掉的多余数据；二是选择合适的方式方法，来处理这些数据。

2.1.2　数据格式转换

通过查看原始数据表可以发现，初始数据类型大多为 float64 的浮点数格式，只有少量数据列为无格式类型 object。其中包括性别（以 0-1 值表达布尔数值），Child 分级（以字母 A-B 表达分级分类），严重并发症（以文本字符串表示内容）。对于这些非数值变量的数据列，需要将其转换为合理的浮点数值。部分数据类型展示见表 2。

表 2　部分数据的数据类型

Tab. 2　The data type of partial data

ID	性别	肿瘤大小	Child 分级	肿瘤数量
float64	Object	float64	Object	float64

2.1.3　数据异常值处理

识别异常值的方法主要有：基于统计学原理的散点图、四分位图、箱线图、正太分布图等方法；基于分布的异常点检测：根据已有数据建立模型，基于模型对数据进行检测，从而判断数据是否异常；基于聚类的方法找出那些零散的不能归为某一类别的数据，作为异常点等方法。

表 3　部分异常值数据展示

Tab. 3　Partial outlier data display

属性	血管癌栓	术前 AST	术前 AFP
Count：	1 733.000 0	1 381.000 0	1 231.000 0
Mean：	0.025 9	37.295 1	1 077.540 8
Std：	0.159 0	33.539 9	17 890.433 0
Min：	0.000 0	9.900 0	0.570 0
25%：	0.000 0	20.800 0	3.710 0
50%：	0.000 0	27.400 0	15.650 0
75%：	0.000 0	41.000 0	139.750 0
Max：	1.000 0	434.900 0	611 000.000 0

注：25%为第一四分点，75%为第三四分点。

由表 3 可以发现，数据中存在着大量的异常值。如：术前 AFP 指标，其均值为 1 000，而中位数仅为 16，且 75%分位点也仅有 139.75，但最大值竟然达到了611 000这种极度不合理的大数值，说明该数据列中的一些数据在录入时出现了错误，导致整个数据列的分布偏离了正常的分布。异常值由于数值问题，会在模型中产生极大的噪声，导致对包含异常值数据的样本预测难以继续，同时在对包含异常值的数据列进行归一化时也会出现分布不均的问题，因此需要采用合理的方式对异常值进行处理，降低其表达的信息量。

在异常值检测方面，本文采用了两种方式：一是正态分布的假设检验。即出现偏离均值超过方差 3 倍的值属于极小概率事件，记为异常值。但这种方式对于均值和方差均被显著提高的数据列检测效果不佳，易将一些数据漏算。第二种是计算数据分布的四分位点，认为比 $Q1$ 小 1.5 倍的 IQR 或者比 $Q3$ 大 1.5 倍的 IQR 的值为异常值（$Q1$ 为第一四分位数，$Q3$ 为第三四分位数，IQR 为四分位数极差，其值为 $Q3-Q1$），将两种方式相结合，即可最大程度的检测出异常值。

2.1.4　数据缺失值处理

一部分数据由于某些原因，导致数据缺失，需要对这部分数据给予适当的处理。数据集中部分数据的缺失，不但增大了数据集的不确定性，也影响了算法的执行。缺失值产生的原因主要来自机械和人为因素。从缺失值的分布来看，可以分为完全随机缺失、随机缺失和完全非随机缺失。从缺失值的所属

属性上讲,如果所有的缺失值都是同一属性,那么这种缺失称为单值缺失,如果缺失值属于不同的属性,则称为任意缺失。部分特征数据缺失值情况统计见表 4。

表 4 部分特征缺失个数

Tab. 4 The missing number of partial feature

ID	性别	肿瘤大小 a	肿瘤大小 c	术前 AFP
1 个	0 个	24 个	1 219 个	503 个

常用数据缺失值处理方法:

(1)置 0 处理法:将缺失值置为 0,这种方法实现起来比较简单,但是容易造成较大的误差。

(2)均值处理法:用某一列特征所具有数据的平均值填充这一列的空值。如果出现特征数据为非数据的形式,可以选择频次最高的数据作为数据的填充值。此种方法在数据挖掘中应用广泛,方法便捷。

(3)最近邻填充处理法:根据各种距离计算公式,计算两个样本之间的距离,确定空缺值所在的样本与其最接近的样本,对 k 个最接近的样本加权平均得到空缺值所需的数据。

(4)模型填充:把缺失值作为新的标签,基于已有的完整信息建立模型,对数据拟合,将训练好的模型预测缺失值进行填补。常用随机森林等拟合填充空值,线性回归预测空缺值。但是该方法的空缺值过多将会影响最终的预测结果。

(5)删除所有空缺值所在的属性列值:这种方式适合空缺值较多、属性多、被删除的特征属性具备较多空值的情况,否则将严重影响最终的预测结果。

在缺失值处理方法上,本文针对只有少量数据缺失的数据列采用了置 0 处理法和均值处理法。如:肿瘤大小 a、肿瘤数量等。对于像肿瘤大小 c 这样有大量数据均缺失的数据列,采用了直接删除属性列的方式。

2.1.5 数据归一化

数据清洗以及缺失值处理后,由于各项指标的区间不同,其表达的特征维度也都不同。为了使各列数据对于预测结果的贡献相同,模型训练的参数处于同一量级,需要将特征进行放缩到相同量级,进行数据归一化的结果就是将特征缩放到相同量级。

数据归一化常用方法有如下两种:

(1)最大最小标准化(Min-Max Normalization)

对数列 $x_1,x_2,\cdots,x_n$ 进行变换:

$$y=\frac{x_i-\min\{x_j\}}{\max\limits_{1\leqslant j\leqslant n}\{x_j\}-\min\limits_{1\leqslant j\leqslant n}\{x_j\}},\tag{7}$$

则新数列 $y_1,y_2,\cdots,y_n\in[0,1]$。其中,$\min\{x_j\}$ 为数列 x 中的最小值,$\max\{x_j\}$ 为数列 x 中的最大值。

(2)Z-score 标准化方法

对数列 $x_1,x_2,\cdots,x_n$ 进行变换:

$$y_i=\frac{x_i-\bar{x}}{s},\tag{8}$$

式中,$\bar{x}=\frac{1}{n}\sum\limits_{i=1}^{n}x_i$,$s=\sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$.

则得到的数列 $y_1,y_2,\cdots,y_n$ 均值为 0,方差为 1。

本文采用了最大最小标准化(Min-Max Normalization)方法对数据进行归一化处理。数据归一化后的部分结果见表 5。

表 5 部分数据 MinMax 数据归一化结果

Tab. 5 Partial data Minmax data normalization results

术前白细胞	术前 GT	腹水	是否早期复发
0.332 036	0.190 334	0.0	1.0
0.426 938	0.346 659	0.0	1.0
0.325 952	0.381 265	0.0	0.0
0.472 495	0.276 777	0.0	0.0

2.1.6 特征工程

特征工程的目的,是探索特征对于预测任务的重要性及影响,从而提取更加有效,对结果影响更加显著的特征,提高预测的稳定性、鲁棒性,避免引入过多噪声数据列。由于该数据集中包含了大量的特征,且在疾病预测的任务上许多医疗检测指标都没有实际的医学意义的影响,因此需要对数据进行特征工程,提取更加有效的特征作为预测模型的输入。

本文共采用了 6 个特征工程的方法,来对数据进行操作。

(1)Pearson 相关系数:Pearson 相关系数是一种能够帮助理解特征和响应变量之间关系的方法。该方法衡量的是变量之间的线性相关性。

(2)随机森林(Random Froest):随机森林就是通过集成学习的思想将多棵决策树组合在一起,其基学习器是决策树。

(3)逻辑回归(Logistic regression):逻辑回归主要思想是,根据现有数据对决策边界建立回归方程后,将回归方程映射到分类函数上,实现分类。

(4)平均准确度减少(Mean accuracy reduction):平均准确度减少是通过将某个特征随机打乱,使其表达的信息紊乱,计算在此情况下模型预测准确度减少的比率,来判断特征重要性的方法。平均准确

度减少也是随机森林特征工程方法中重要的度量方式，在实现时也使用随机森林作为基准模型。

(5)递归特征消除(Recursive feature elimination)：递归特征消除的主要思想是，反复构建模型并选出最好或最差的特征，消除此特征，并不断重复，直到所有特征都被遍历，这个过程中特征的消除次序表示了特征的排序，因此这也是一种寻找最优特征子集的贪心算法。

(6)互信息与最大信息树(Mutual information and maximum information tree)：通过寻找一种最优的离散化方式，将互信息取之转换为度量方式，对不同变量之间的距离进行度量，从而判断关系最密切的变量。

通过以上特征工程方法，可以得出肿瘤数量是最核心特征的结论。为了探索特征工程的有效性，本文采用了逻辑回归模型，得到不同特征下评价指标的结果见表 6。

表 6　不同特征下评价指标结果

Tab. 6　Evaluation index results in different characteristics

数据	Accuracy	Precision	Recall	AUC
肿瘤数量	0.690 1	0.751 1	0.556 9	0.732 3
其它特征	0.554 1	0.548 0	0.542 5	0.589 8
全部特征	0.675 9	0.699 7	0.601 3	0.740 5

从表 6 可以看出，相比于其它特征，肿瘤数量是最核心的特征，其它特征的重要性并不高，甚至会是噪音，造成误诊。

2.1.7　样本不均衡处理

数据不平衡是指在数据集中，不同类别的样本数量差距很大。如：在病人是否得癌症的数据集上，可能绝大部分的样本类别都是健康的，只有极少部分样本类别是患病的，这样会给预测带来极大地噪声。为了消除这些噪声，处理数据的过程，称之为数据不平衡处理。数据不平衡处理的常用方法有重采样、过采样和欠采样，使采样的样本标签均衡；有类别加权，调整不同的标签类别的权重来处理不平衡数据等。

本任务中，统计类别为 0，即无早期复发，与类别为 1，即早期复发的样本比例，负样本数量为 787，正样本数量为 765，二者比值约为 1∶1，说明数据相对均衡，无需进行数据不平衡处理。

2.2　模型构建

数据输入时，需要将预处理好的数据转换为模型可以接受的输入形式。首先要读取预处理好的数据，这里使用了文件存储读取的方式。由于数据中包含了许多随访时间未达到 24 个月的未复发样本，为保证数据分布的一致性，将其划定为非早期复发并将其过滤掉。通过特征工程得到的特征排序，可以选定重要的特征来进行预测，避免引入过多的噪声。最后，将处理好的数据转换为 NumPy 的数组形式，以便被预测模型接受并使用。

模型训练部分使用 sklearn 自带的机器学习模型，选择了逻辑回归、随机森林、SVM 支持向量机、GBDT 梯度提升树 4 种模型。

将这 4 种机器学习模型采用准确率、精确率、召回率和 AUC 4 个指标来评价，评价结果见表 7。

表 7　学习模型指标结果

Tab. 7　Learning model evaluation results

模型	*Accuracy*	*Precision*	*Recall*	*AUC*
逻辑回归	0.675 9	0.699 7	0.601 3	0.740 5
SVM	0.561 8	0.681 9	0.583 1	0.673 6
随机森林	0.699 1	0.709 6	0.658 8	0.743 6
GBDT	0.695 2	0.720 4	0.631 4	0.765 5

由于数据分布规则性较差，逻辑回归和 SVM 支持向量机表现相对较差。由表 7 可以看出，集成学习模型、随机森林和 GBDT 梯度提升树，在此任务上更加有效稳定，因此选择这两类模型来进行训练与测试。

为了探索肝癌早期复发的阈值，设计了对阈值进行选择的函数。通过设置不同的阈值，为数据进行重标签，使用此阈值条件下的数据得到的 AUC 分数结果作为该阈值的评价，来对阈值进行排序，得到最优的阈值。为了避免偶然性导致的结果不稳定，对每个阈值条件下的数据随机排序，进行了复数次计算，将多次得分的均值作为最终结果。

3　实验与结果分析

3.1　实验数据描述

本次实验分析所用到的原始数据是以 xls 文档的形式保存的，可以用 Python 自带的 Pandas 处理。数据每一行表示一个病人的随访记录，包含其基本信息以及检测的各项指标；每一列表示一类相关信息，包括年龄、性别等基本信息和肿瘤数量等病历信息与医学检测指标。数据基本描述见表 8。

从中可以看出，数据共有 1 734 行、47 列，患者基础数据有 3 列，基础检验指标有 39 列，肿瘤相关指标有 4 列。由此可见，数据数量较多，可以用来预测的数据列庞大，正负样本数量约为 1∶1，数据相对平衡。

表 8 数据的基本描述

Tab. 8 The basic description of the data

数据表属性	数量
数据表行数	1 734 行
数据表列数	47 列
患者基础数据列数	3 列
基础检验指标列数	39 列
肿瘤相关指标列数	4 列
正标签数	765 个
负标签数	787 个

3.2 模型参数与验证

机器学习模型需要进行调参来获取较好的预测结果,本文选择了网格搜索法来进行参数的搜索。在 GBDT 模型中,主要搜索决策树数量和学习率两个参数;在随机森林模型中,主要搜索决策树数量和划分标准两个参数。最优参数的评价指标选择了 *AUC* 数值,获取到最优的参数后,使用五折交叉验证的方式进行模型的训练与测试。

在模型评价方面,选择了准确率、精确率、召回率和 *AUC* 4 个指标进行评价。其中,准确率为预测正确数量的个数占总预测数量个数的比重;精确率为正确预测为正的数量占全部预测为正的数量比例;召回率为正确预测为正的数量占全部实际为正的数量的比例。*AUC* 定义为 ROC 曲线下的面积,这个面积的数值越接近 1(数值不会大于 1),则说明模型效果越好。其中对于疾病预测问题, *AUC* 和召回率的结果更加有价值,更适合此问题的评价指标。本文最终选择将 *AUC* 作为最主要评价指标。在 GBDT 和随机森林两种模型中,最终选择了 GBDT 梯度提升树,并在 0.05 的学习率、80 迭代次数的参数条件下进行验证。

3.3 肝癌早期复发预测结果

GBDT 梯度提升树在 0.05 的学习率,80 的迭代次数的参数条件下。准确率、精确率、召回率和 *AUC* 4 个指标的结果见表 9。

表 9 模型测试结果

Tab. 9 Model test results

Accuracy	*Precision*	*Recall*	*AUC*
0.695 2	0.720 4	0.631 3	0.765 5

通过模型训练与预测的结果可以看出,准确率的数值达到了 0.695,说明正确预测的概率可以达到近 70%;而精确率的数值达到了 0.720,说明有 72% 的概率可以正确预测结果。本实验采用的数据集正负样本比例为 1:1,不存在数据不平衡的情况,因此准确率则能较好的说明问题;召回率的数值达到了 0.631,说明有 63% 的概率可以正确估计正样本;而主要评价指标 *AUC* 的值达到了0.765 5,是一个较为不错的数值。因此,可以大致认为,在肝癌早期复发预测问题上,此数据可以得出一个可信的结果。

本文从数据而非医学的角度,对肝癌复发问题进行了预测研究。相对于传统医疗手段(包括但不仅限于 B 超检查、增强 CT 和核磁检查等)的预测方法,本实验仅依赖一份数据表便得出了极高准确度的预测结果,很大程度上节约了人力、物力、财力,对医学上的病人是否会复发肝癌提供了一种简单并且准确的判断方式,对医学上病人是否会复发的初步判断,提供了强有力的手段。同时,数据并不会产生任何的环境污染和能源消耗,相对于医学检查来说,极大地节约了资源和保护了环境,符合绿色发展的观念,具有非常深远的意义。

3.4 阈值选择结果

在各个阈值条件下,模型预测 *AUC* 的结果,整体呈现递增趋势,即阈值时间越靠后,预测的准确率越高,结果越可信。然而对于疾病预测这种特殊的任务,通常认为时间靠后导致病人就医的成本显著提升,因此简单地以预测结果的可信程度作为阈值选择的因素不符合现实。

为解决这一问题,本文选择得分相对提升最大的时间作为阈值,即最优阈值应当满足相比更小的阈值,得分提升尽可能大,而相比更大的阈值,得分降低尽可能小。经过反复实验后,得到的最优阈值为 12 个月的可能性最大,因此从数据的角度来看,肝癌的早期复发预测的阈值被设置为 12 个月。

现代医学上普遍将 12 个月作为肝癌早期复发的阈值,而本文在大量实验的基础上,得出肝癌复发预测的最优阈值为 12 个月的概率高达 84.615%,不管是从医学角度还是从数据角度,都说明了肝癌复发预测的最优阈值为 12 个月的结论。这对现代医学上肝癌早期复发的预测问题具有一定的参考和指导意义。如果在为病人问诊时医生能将这一因素的参考权重增大,就可能尽早发现病人的异常情况,从而辅助医疗决策,具有很强的实用意义和现实意义。

4 结束语

肝癌早期复发预测是肝癌术后护理的重要问题,本文应用数据挖掘技术,通过对数据进行缺失值补全,特征工程等处理,并搭建机器学习模型,对肝癌早期复发进行预测。实验结果表明,本文应用的

自动化数据挖掘技术，在疾病预测的准确性上实现了前沿的结果，可以为肝癌患者的预后诊断提供指导性意见。同时，本文探索的肝癌早期复发阈值是基于客观数据得到的，可以为医学视角下的肝癌复发病理研究提供帮助。

虽然本文在肝癌早期复发预测上准确率较高，但由于数据量与数据特征有限，在临床应用上依然不可避免地存在着偏置性。通过扩大数据量，引入更加专业相关度更高的医学指标，是后续研究重点关注的方向。

参考文献

[1] 中华人民共和国卫生和计划生育委员会医政医管局. 原发性肝癌诊疗规范(2017 年版)[J]. 中华消化外科杂志，2017，16(7)：705-720.

[2] 王逊. 数据挖掘在医学领域中的应用[D]. 成都：电子科技大学，2014.

[3] 石胜源，朱磊，叶琳，等. 基于随机森林算法的心血管疾病预测研究[J]. 智能计算机与应用，2021，11(4)：176-178，181.

[4] 赵金超，李仪，王冬，等. 基于优化的随机森林心脏病预测算法[J]. 青岛科技大学学报(自然科学版)，2021，42(2)：112-118.

[5] 肖美丽，晏春丽，付冰，等. 随机森林算法在产后抑郁风险预测中的应用[J]. 中南大学学报(医学版)，2020，45(10)：1215-1222.

[6] 王君贤. 基于随机森林与支持向量机的心血管疾病预测研究[D]. 天津：天津大学，2018.

[7] 王远旭. 基于机器学习算法的医疗数据处理与分析[D]. 厦门：厦门大学，2018.

[8] 崔晓旭. 基于数据挖掘的疾病预测组合模型研究[D]. 北京：北京交通大学，2019.

[9] 王哲，李琳，李丞，等. 基于机器学习方法的慢性阻塞性肺疾病分期预测[J]. 中国数字医学，2019，14(3)：38-40.

[10] 陈志屹. 基于机器学习的肝硬化疾病预测[D]. 长沙：湖南大学，2018.

[11] Krishnamurthy Surya，KS Kapeleshh，Dovgan Erik，et al. Machine Learning Prediction Models for Chronic Kidney Disease Using National Health Insurance Claim Data in Taiwan[J]. Healthcare，2021，9(5)：546.

[12] Mulyawan，Bahtiar Agus，Dwilestari Githera，et al. Data mining techniques with machine learning algorithm to predict patients of heart disease[J]. IOP Conference Series：Materials Science and Engineering，2021，1088(1)：012035.

[13] 周相广，李大伟. 应用梯度提升决策树算法预测套损[J]. 计算机应用，2018，38(S2)：144-147.

[14] Haewon Byeon. Predicting the Anxiety of Patients with Alzheimer's Dementia using Boosting Algorithm and Data-Level Approach[J]. International Journal of Advanced Computer Science and Applications (IJACSA)，2021，12(3)：2021.

[15] 陈锋，李张铮，庄毅莹. 基于 GBDT 算法的潜在 5G 用户预测研究与实现[J]. 邮电设计技术，2021(4)：45-49.

(上接第 34 页)

建形变特征提取模块提升对于目标特征提取的有效性，同时针对于形变卷积对特征提取网络模块进行优化，增强了特征信息的传递能力。经测试，优化后的 YOLO-sd 在针对于红外小目标的检测场景下检测精度有明显的提高。整体精度提升 1.05%，达到 83.09%。本文的网络对于夜间来往的行人、驾驶的车辆来说，有辅助参考价值，有助于提高安全性。

参考文献

[1] 崔美玉. 论红外热像仪的应用领域及技术特点[J]. 中国安防，2014(12)：90-93.

[2] CARLO C，SALVETTI O. Infrared：a key technology for security systems[J]. Advances in Optical Technologies，2012，2012：838752.

[3] 刘学，李范鸣，刘士建. 改进的 SSD 红外图像行人检测算法[J]. 电光与控制，2020，27(1)：42-46，59.

[4] VIOLA P，JONES M J，SNOW D. Detecting pedestrians using patterns of motion and appearance[J]. International Journal of Computer Vision. 2005，63(2)：153-161.

[5] DALAL N，TRIGGS B. Histograms of oriented gradients for human detection[C]//international Conference on computer vision & Pattern Recognition，2005：886--893.

[6] FELZENZWALB P F，GRISHICK R B，MCALLISTERD，et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2010，32(9)：1627-1641.

[7] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[8] 孙宁，陈梁，韩光，等. 深度分类网络研究及其在智能视频监控系统中的应用[J]. 电光与控制，2015，22(9)：77-82.

[9] 车凯，向郑涛，陈宇峰，等. 基于改进 Fast R-CNN 的红外图像行人检测研究[J]. 红外技术，2018，40(306)：70-76.

[10] JENSEN M B，NASROLLAHI K，MOESLUND TB. Evaluating state-of-the-art object detector on challenging traffic light data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.2017：9-11.

[11] GIRSHICK R. FastR-CNN[C]//Proceedings of the IEEE international conference on computer vision. 2015：1440-1448.

[12] REN S，He K，GIRSHICK R，et al. Faster r-cnn：Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015：192-199.

[13] LIU W，ANGUELOV D，ERHAN D，et al. Ssd：Single shot multibox detector[C]//European conference on computer vision. Springer，Cham，2016：122-137.

[14] REDMON J，DIVVALA S，GIRSHICK R，et al. You only look once：Unified，real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016：779-788.

[15] 李慕锴，张涛，崔文楠，等. 基于 YOLOv3 的红外行人小目标检测技术研究[J]. 红外技术，2020，42(2)：176-181.