TCGA数据库基因突变信息结合机器学习软件RapidMiner构建肝细胞癌患者复发模型

祁亮, 沈洁（南京大学医学院附属鼓楼医院 肿瘤中心 南京大学临床肿瘤研究所，南京 210008）

摘要：目的 通过TCGA数据库基因突变信息结合机器学习软件RapidMiner构建肝细胞癌患者复发模型。**方法** 首先通过TCGA数据库收集316例肝细胞癌患者的临床资料和全基因组测序的突变基因信息；然后利用R语言和SPSS19.0筛选出前127个高频突变基因和12个与无疾病生存期（disease-free survival period，DFS）显著相关的高频突变基因；通过RapidMiner8.0机器学习软件，利用316例患者的突变基因信息训练决策树和支持向量机（support vector machine，SVM）模型。**结果** 通过利用TCGA数据库筛选的基因构建的决策树模型准确率为77.42%，通过构建SVM模型佐证决策树模型的最大准确率为77.42%。**结论** 通过公共数据库构建的肝细胞癌患者的复发模型，可在临床上用来分析患者的基因检测报告，除了提供药物治疗靶点的信息外，还可初步判断患者的预后；此外，对于部分经济条件受限的患者可重点针对决策树中的基因进行检测，来预测预后及复发可能。

关键词：肝细胞癌；基因测序；决策树；支持向量机；机器学习

Construction of recurrence model of patients with hepatocellular carcinoma by gene mutation information in TCGA database combined with machine learning software RapidMiner

QI Liang, SHEN Jie *(Comprehensive Cancer Centre of Drum Tower Hospital, Medical School of Nanjing University & Clinical Cancer Institute of Nanjing University, Nanjing 210008, China)*

Abstract: Objective To investigate the construction of recurrence model of patients with hepatocellular carcinoma (HCC) by gene mutation information in TCGA database combined with machine learning software RapidMiner. **Methods** The clinical data and genome-sequenced mutant gene information of 316 patients with HCC were collected according to the TCGA database. The first 127 high frequency mutation genes and 12 high frequency mutation genes which had significant correlation with disease-free survival period (DFS) were screened by R language and SPSS 19.0. Mutated genetic information from 316 patients were applied to train decision trees and support vector machines (SVM) models by RapidMiner 8.0 machine learning software. **Results** The accuracy of the decision tree model constructed according to the TCGA database was 77.42%, and the maximum accuracy of the decision tree model by constructing the SVM model was 77.42%. **Conclusions** The recurrence model of patients with HCC constructed by public database can be used to analyze the gene detection report of patients in practice. In addition to providing information on drug treatment targets, it can also judge the prognosis of patients preliminarily. Some patients with limited economic conditions can focus on detecting genes in decision trees to predict the prognosis and recurrence.

Key words: Hepatocellular carcinoma; Gene sequencing; Decision tree; Support vector machine; Machine learning

肝细胞癌（hepatocellular carcinoma，HCC，以下简称肝癌）是全球最常见的恶性肿瘤，其发病率和病死率均位于所有肿瘤的前5位。可进行手术治疗的早期肝癌患者1年复发率高达50%以上，部分患者治疗后1～2个月内便出现转移，对于已发生转移的肝癌患者，目前有效的治疗手段为靶向治疗、化疗及局部放疗缓解症状，但再治疗的有效率低于10%。因此，在肝癌患者初诊时找到有效方法预测

DOI: 10.3969/j.issn.1674-7380.2018.03.003
基金项目：江苏省“十三五”科教强卫工程青年医学人才项目（QNRC2016043）；南京市医学科技发展重点项目（ZKX16032）；重大慢性非传染性疾病防控研究重点专项（2017YFC1308900）
通讯作者：沈洁 Email: shenjie2008nju@163.com

复发的风险，对治疗决策具有积极影响[1]。

诸多研究表明，肿瘤分期、大小、数目、癌栓、AFP及循环肿瘤细胞等可预测肝癌患者术后或综合治疗后的复发风险，但当这些因素出现阳性或水平升高时，肝癌患者可能已经发生了影像学上尚未能明确的微小转移，如何能在这些因素未出现波动时预测复发风险呢？基因组测序为这种预测提供了可能。借助美国TCGA数据库免费获得的经全基因组测序的376例肝癌患者的突变基因和临床资料等数据，通过SPSS 19.0统计软件生存曲线分析突变基因与无疾病生存期（disease-free survival period，DFS）的相关性，找出能预测DFS的高频突变基因，但这些突变基因在预测DFS中究竟可发挥多大作用，笔者希望能定量分析并构建复发模型。本研究中借助R语言（一种免费开源的大数据处理软件）和目前较为流行的人工智能学习软件RapidMiner来实现这种初诊肝癌患者复发模型的构建。

1 资料与方法

1.1 研究对象 从TCGA官网（https://cancergenome.nih.gov/）或cbioportal网站（www.cbioportal.org）下载376例肝癌患者的临床及基因突变信息，经数据处理后将其中非HCC病例及信息缺失的病例剔除，筛选出316例有完整DFS记录的HCC患者为研究对象。

1.2 高频突变基因的筛选 人类全基因组测序共2万多个基因，TCGA数据库中记录的这些基因中有9230个突变基因，将大量稀有突变基因纳入建模易导致过拟合（过拟合指用人工智能软件构建的模型在训练集上拟合度很好，但泛化能力较差，不具有实际应用价值），本研究选取了突变频率靠前的127个相对高频突变基因，利用SPSS 19.0软件进行生存分析，计算每个基因的Log-rank P值，从127个高频突变基因中筛选出12个与DFS有关的高频突变基因。

1.3 聚类和不聚类热图的绘制 通过R语言（版本：R3.4.2）利用pheatmap包绘制上述127个突变基因与患者DFS的聚类和不聚类热图（聚类是一种非监督学习算法，由于事先并不知要分析的属性间的相关性，通过聚类可发现这些属性间潜在的相关性）。由于数据体量的局限性，DFS作为连续型变量时，如未能明确发现这些高频突变基因突变与否和DFS的明确关系，可将DFS进行特征筛选，简化为二分类变量。

1.4 决策树的构建及支持向量机算法 使用RapidMiner8.0软件（人工智能领域较为常用的一种预测分析和数据挖掘软件）基于患者的基因突变信息和部分较为完整的临床数据如：年龄、种族和性别等，按分层抽样，以9∶1的比例随机分为训练组和测试组，先用训练组构建一个决策树A，然后用测试组测试这个模型，通过调整参数，计算该决策树的最高准确率。利用相同方法，构建12个与DFS有关的高频突变基因的决策树B，计算最高准确率。使用支持向量机（support vector machine，SVM）算法从另一个角度构建模型并计算准确率。本研究的流程图见图1。

1.4 统计学处理 采用SPSS 19.0统计软件进行Log-rank生存分析，以$P < 0.05$为差异有统计学意义。

2 结果

2.1 高频突变基因 本研究选取的127个相对高频突变基因分别为TTN、ALMS1、KMT2D、PKHD1L1、DMD、SVEP1、MT-ND6、ROBO2、TP53、HMCN1、SYNE1、FREM2、MT-CO1、HERC1、NEFH、CTNNB1、FRAS1、DSPP、MYO3A、KEAP1、LRP2、NFE2L2、MUC16、USH2A、BAP1、DNAH5、GCN1、FBN1、SPEG、ALB、MUC4、FAT4、DYNC2H1、HTT、PKHD1、FANCM、PCLO、FLG、CUBN、DOCK2、KIAA1109、UNC79、KMT2A、APOB、AHNAK2、DNAH9、RYR3、PTPRQ、DCHS1、MAP1B、RYR2、NBEA、SYNE2、MUC2、FBN2、POLQ、SACS、ND5、EYS、TCHH、HERC2、PREX2、DNAH2、DNAH17、CSMD3、CSMD1、ZNF469、DNAH10、FMN2、LAMA1、COL6A6、OBSCN、AXIN1、HSPG2、MUC17、CSMD2、PREX1、HECTD4、ABCA13、RB1、ZFHX4、LRP1、FASN、NEB、JAK1、ARID1A、DNAH7、UNC80、ANKRD12、FAT2、MYCBP2、NCAM1、CACNA1E、ADGRV1、COL11A1、ABCA12、COL6A3、UNC13C、MYO18B、LRP1B、CYTB、BIRC6、DCHS2、COL12A1、DSCAM、ITPR1、XIRP2、ARID2、KMT2B、PRUNE2、ATR、SDK1、ASCC3、SPTA1、DNAH6、WDR87、KMT2C、SETD2、PCDH15、DST、RYR1、FAT3、AHNAK、DNAH8、MDN1、KIF26B、TENM4。通过Log-rank生存分析筛选出12个与DFS有关的高频突变基因分别为TP53、APOB、ABCA13、FRAS1、CSMD1、RB1、DSPP、KMT2B、FREM2、DNAH8、ATR和ASCC3（$P < 0.05$），见表1。

2.2 聚类和不聚类热图 R语言绘制的316例患者中127个高频突变基因与DFS的聚类热图及不聚类热

图见图 2，结果表明上述 12 个基因突变中具有至少 1 个的患者，DFS 较短，易复发；具有高频突变基因数多的患者，DFS 相对较短，容易复发，高频基因突变越低，DFS 更倾向于半年以上复发。

2.3 决策树的构建和支持向量机算法 决策树 A 和决策树 B 的算法流程见图 3 和图 4，分类结果混淆矩阵见表 2 和表 3，决策树 A 的最高准确率为 77.42%，决策树 B 的最高准确率也为 77.42%。使用 SVM 算

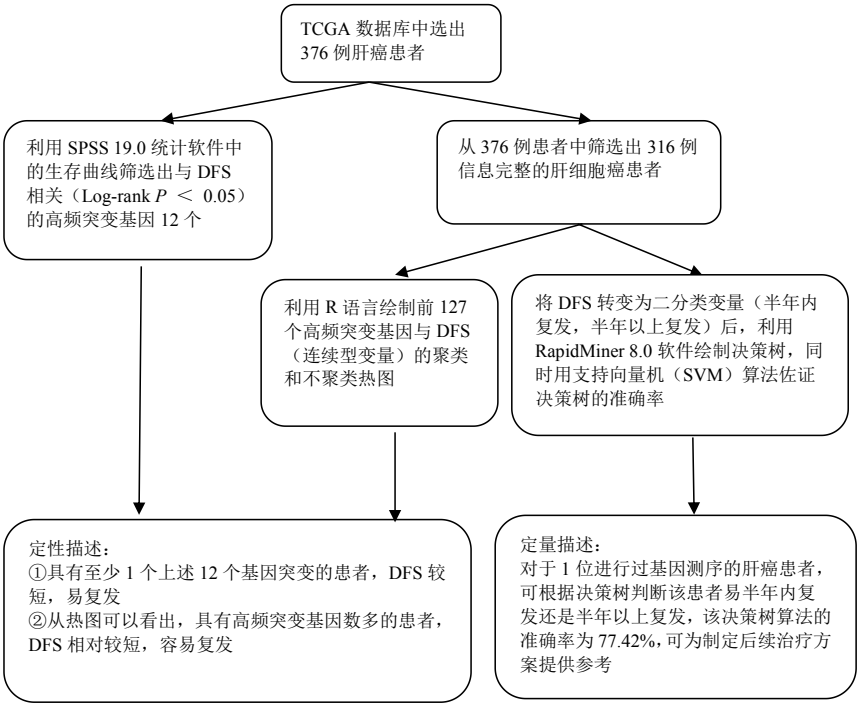


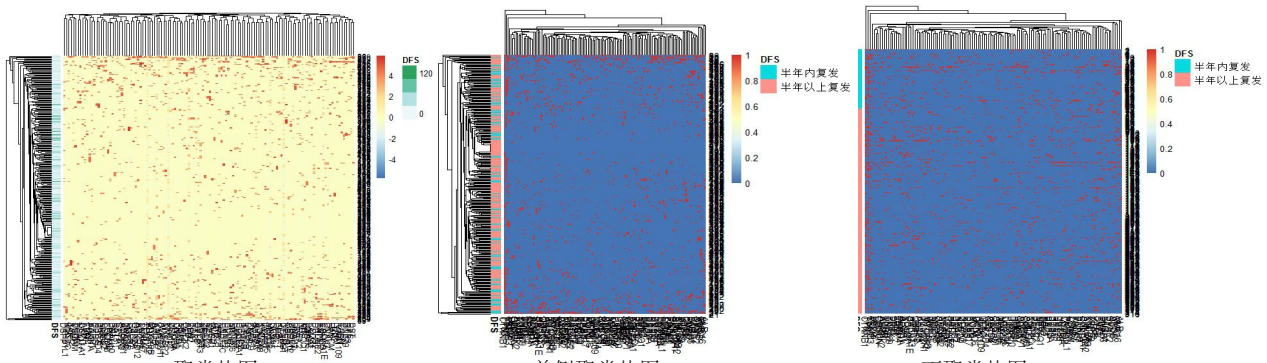
图 1 本研究流程图

表 1 376 例肝癌患者前 127 个高频突变基因中与 DFS 有关的 12 个高频突变基因

名称	染色体上位置	基因长度（nt）	有突变（例）	无突变（例）	有突变已死亡（例）	无突变已死亡（例）	Log-rank *P*值
TP53	17p13.1	19149	100	225	62	117	0.004
APOB	2p24.1	42645	35	290	23	156	0.035
ABCA13	7p12.3	476074	28	297	21	158	＜0.001
FRAS1	4q21.21	486947	20	305	14	165	0.034
CSMD1	8p23.2	2059683	21	304	17	162	0.012
RB1	13q14.2	178144	18	307	14	165	0.012
DSPP	4q22.1	8345	18	307	7	172	0.186
KMT2B	19q13.12	21063	13	312	10	169	0.008
FREM2	13q13.3	200096	17	308	13	166	0.003
DNAH8	6p21.2	315470	12	313	10	169	0.002
ATR	3q23	129592	12	313	11	168	0.005
ASCC3	6q16.3	373179	9	316	8	171	0.022

表 2 决策树 A 的分类结果混淆矩阵

真实情况	预测结果		特异性（%）
	6个月内复发（例）	6个月以上复发（例）	
6个月内复发	2	2	50.00
6个月以上复发	5	22	81.48
敏感性（%）	28.57	91.67	-



聚类热图　　单侧聚类热图　　不聚类热图

图 2 R 语言绘制的 316 例患者中 127 个高频突变基因与 DFS 的聚类热图、单侧聚类热图及不聚类热图

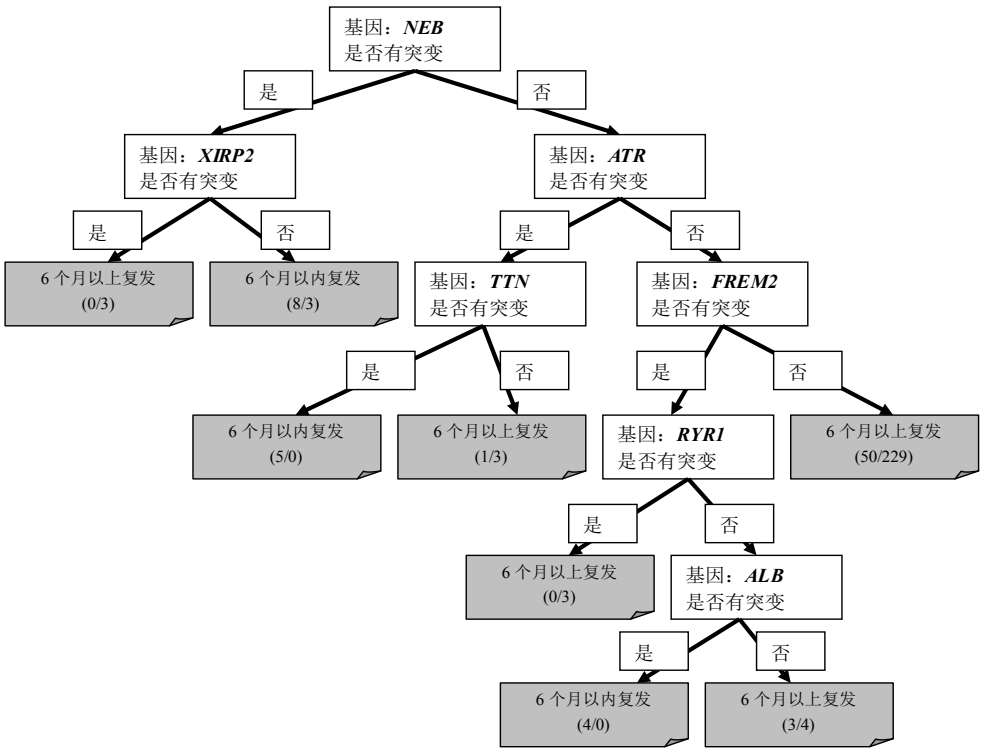


图 3 用决策树算法构建的 127 个高频突变基因预测复发的模型 A

表 3 决策树 B 的分类结果混淆矩阵

真实情况	预测结果		特异性（%）
	6个月内复发（例）	6个月以上复发（例）	
6个月内复发	1	1	50.00
6个月以上复发	6	23	79.31
敏感性（%）	14.29	95.83	-

注：“-”为无相关数据

表 4 支持向量机（SVM）的分类结果混淆矩阵

真实情况	预测结果		特异性（%）
	6个月内复发（例）	6个月以上复发（例）	
6个月内复发	0	0	00.00
6个月以上复发	7	24	77.42
敏感性（%）	0.00	100.00	-

注：“-”为无相关数据

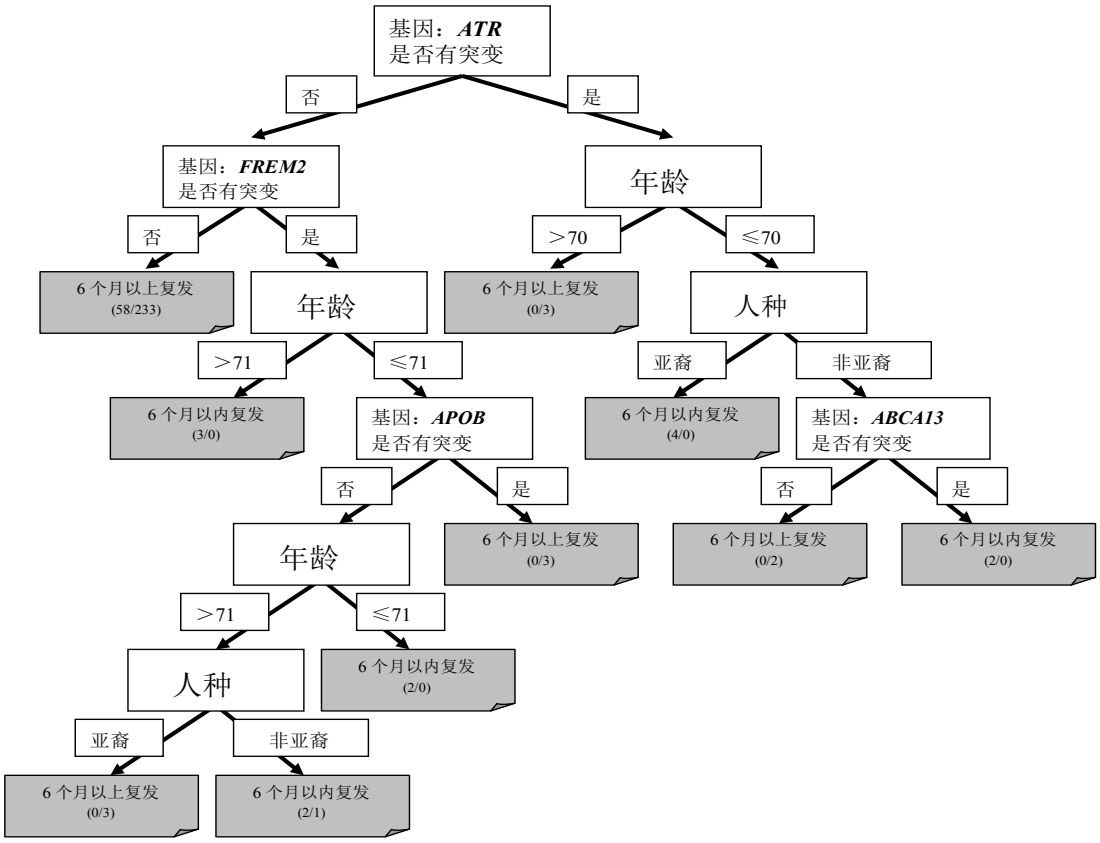


图 4　用决策树算法构建的 12 个与 DFS 有关的高频突变基因预测复发的模型 B

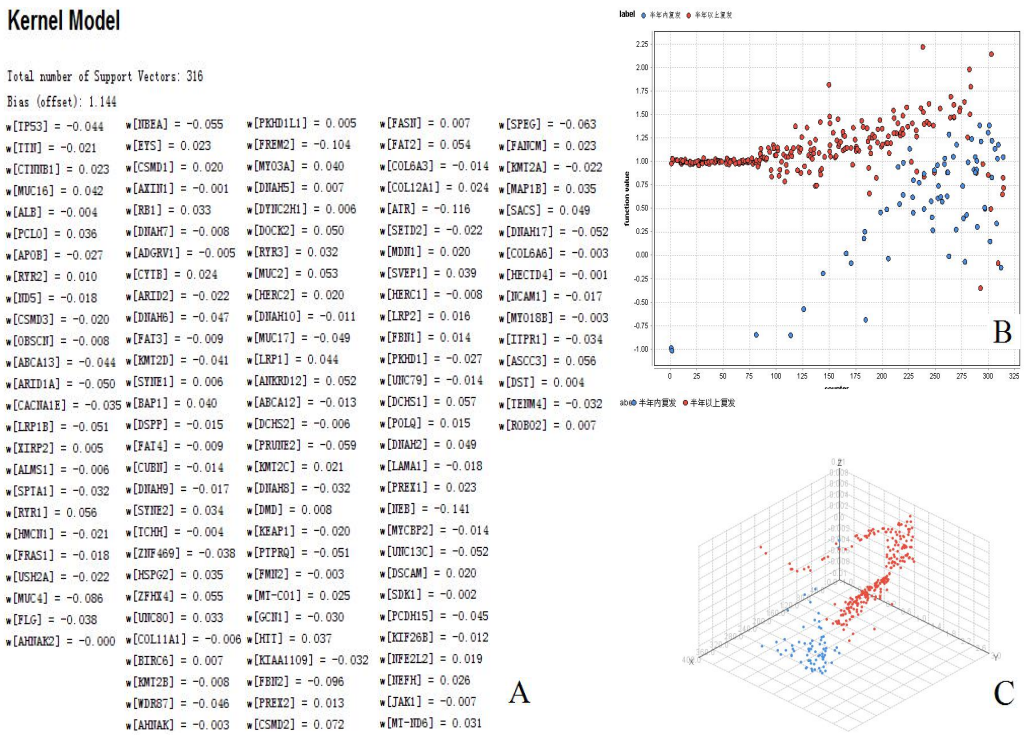


图 5　用支持向量机（SVM）构建的基因突变预测复发的模型

注：A 列出了每个基因在分类中的权重参数值，偏置参数为 1.144；B 显示在二维平面中 316 例患者的包含 127 个突变基因，即 127 个特征的函数分布；C 显示了在三维空间里，SVM 利用核函数将半年内复发（蓝色）和半年以上复发（红色）的患者区分开，准确率为 77.42%

法从另一个角度构建模型，见图5，分类结果混淆矩阵见表4，准确率也为77.42%。模型构建完成。

3 讨论

机器学习致力于研究如何通过计算手段并利用经验来改善系统自身的性能，在计算机系统中，“经验”通常以“数据”形式存在，因此，机器学习所研究的主要内容是在计算机上从数据中产生“模型”（Model）的算法，即“学习算法”（learning algorithm）。将经验数据提供给学习算法即可基于这些积累的大量数据产生模型，然后在面对新的肝癌患者时，模型会提供相应的判断（如预后、复发风险、疗效、影像等）。机器学习与十几年前出现的各种专家诊疗系统有本质区别，专家系统是将既往许多专家诊疗的病例汇集到一起，编成固定程序，当遇到新的肝癌患者时，专家系统会检索自己的数据库，找到最匹配的数据，然后给出诊疗建议；而机器学习是利用多种算法，让计算机自己学习既往诊疗的大量病例，把这些病例的所有特征或属性综合分析，使机器自己“习得”最佳的诊疗模型，然后去面对新的患者，在后续使用中，可根据新的病例继续学习、完善模型，这也是人工智能的体现[2-4]。

机器学习领域旨在开发经验丰富的计算机算法，其有望使计算机帮助人们分析大型复杂的数据集，如：肝癌影像学上的序列元素、肝癌的基因测序及表观遗传学、肝癌蛋白质组学和肝癌代谢组学[5,6]。肝癌诊疗领域常用的机器学习算法包括人工神经网络（artificial neural network，ANN）、决策树和支持向量机3种[2-9]。

决策树（decision tree）是一个树结构（可以是二叉树或非二叉树）。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。决策树的决策过程非常直观，易于理解。目前决策树已经成功运用于医学、制造产业、天文学、分子生物学以及商业等诸多领域。2015年Omran等[10]通过315例感染HCV的慢性肝病患者、116例肝硬化患者及135例HCC患者的临床资料，构建了预测肝癌患者预后的决策树模型，敏感性为83.5%，准确性为83.3%，并且通过机器学习的算法发现了可以独立预测肝癌发生风险的AFP临界值，提示机器学习与临床资料等大数据结合可发挥重要的预测功能，可辅助影像检查指导诊疗甚至独立预测肝癌的发生及预后，避免风险较大的有创检查或操作。Wang等[11]通过收集634例肝癌手术患者的资料，构建了肝癌肝切除术后患者发生肝衰竭的决策树。He等[12]利用决策树算法分析了肝癌患者确诊或治疗前后不同影像学评估的准确性和费用，选择最佳影像学检查手段，减轻了患者的经济负担。2013年Cao等[13]对50例肝癌患者术后的血清蛋白质谱训练决策树，之后用36例同质患者验证决策树的准确率，找到能够预测肝癌术后肝内复发的血清标记物。

SVM最早于1963年提出，是一种监督学习的分类方法，以统计学理论中的VC维理论以及结构化风险最小化为基础，引入最优分界面思想及核函数方法，对输入数据进行训练学习来对分类情况进行建模，对线性及非线性数据进行有效分类。SVM在解决高维、非线性以及小样本数据分类问题中具有较大优势。2012年Ho等[14]利用SVM算法和神经网络对482例接受肝癌切除术患者的临床数据资料训练机器学习模型，用于预测复发和生存，并通过对比不同模型的ROC曲线下面积评估了模型的优缺点。

张朋军等[15]对52例早期肝癌患者和34例健康对照人群的外周血基因数据进行分析，分别用逻辑回归和人工神经网络构建外周血多参数基因诊断模型，结果显示人工神经网络的灵敏度和特异度均高于逻辑回归（96% *vs* 94%；86% *vs* 80%），人工神经网络在肝癌的预测及早期检测中有更好的诊断价值。2017年Qin等[16]使用Illumina公司甲基化分析仪450K Beadchip对576例早期肝癌患者的基因CpG甲基化水平进行检测，所得数据采用机器学习中Lasso算法及SVM-RFE算法（Lasso算法用于精简数据，降低维度；SVM-RFE算法即支持向量机递归特征消除算法），建立了预测早期肝癌复发风险的甲基化标签。2018年Augello等[17]使用机器学习中的分类算法发现MICA基因的两个单核苷酸多态性位点rs2596542、rs2596538及“年龄”可用于肝硬化和肝癌的鉴别分类。Chandhary等[18]利用TCGA数据中的360例HCC患者的RNA测序、miRNA测序及甲基化数据，构建了一个含有3个隐藏层的多层人工神经网络模型，确定了2个不同生存期HCC患者的亚群分类。Liao等[19]对HCC患者的Dishevelled/EGL-10/Pleckstrin（DEP）结构域（DEPDC）蛋白质超家族进行研究，设计了一种分离DEPDCs和非DEPDCs的计算方法。首先，检查已知DEPDC的Pfam数，并使用每个Pfam的最长序列构建系统发育树；随后，提取DEPDCs和非DEPDCs的188维

（188D）和20D特征，用随机森林分类器进行了分类；最后，设计HCC及癌旁正常组织中人DEPDC表达水平的实验验证方法。结果表明，DEPDCs超家族可分为3类，188D和20D特征可用于有效区分两种蛋白质的类型。该研究成功构建了DEPDCs的二元分类器，并通过实验验证了其在人肝癌组织中的表达。Liang等[20]使用机器学习结合代谢组学从HCC患者的尿液中鉴定了15种HCC患者和匹配的健康对照者有差异的代谢物，涉及几种关键的代谢途径，其中5种代谢物对HCC有诊断价值，灵敏度为96.5%，特异度为83%。

既往研究利用患者的临床资料，包括性别、年龄、种族、HBV/HCV感染、Child-Pugh分级、TNM分期、BCLC分期、肿瘤大小、肿瘤数量、癌栓、ALT/AST、胆红素水平、血小板水平、肝纤维化程度和手术术式等信息训练了许多优秀的决策树模型，这些决策树模型的准确度为70%～95%[5-9]，但由于国内全基因组测序和全外显子组测序的患者数量尚在积累过程中，并未有大量数据可用来训练预测模型，故关于我国肝癌突变基因预测预后的研究非常少[21]。

本研究首次利用HCC患者的全基因组测序信息，从高频突变基因出发，预测患者的复发，两种决策树模型和支持向量机模型的总体准确率均为77.42%，但利用127个高频突变基因训练的决策树模型的查全率和查准率均高于12个高频突变基因训练的决策树和支持向量机模型，所以在临床实践中可使用决策树A分析患者的基因测序报告，在患者接受治疗前给予初步的预后预测及复发可能性评估，为制定个体化的综合治疗方案提供参考和依据；对于一部分受经济条件限制、不能行全基因组或全外显子组测序的患者，可根据决策树中的基因进行有针对性的检测。本研究也存在不足之处：①由于数据来自TCGA，该数据库HCC患者的资料无治疗信息，所以无法评估治疗方案对预后的影响；②目前除了TCGA数据库，国内各肿瘤中心尚未积累如此多的全基因组测序患者的信息，后续需积累更多中国HCC患者的全基因组测序信息，以完善预测复发的模型；③本研究中各种模型半年以上复发的敏感率和准确率均较高，但半年以内复发的敏感率和准确率非常低，考虑和样本量较少、半年内复发影响因素较多有关，后续将逐渐积累我国HCC患者的基因测序数据，完善半年内复发患者的预测模型。

参考文献

[1] 刘秀红, 赵一鸣, 赵晓飞, 等. 肝细胞癌诊断与治疗研究进展[J/CD]. 中国肝脏病杂志(电子版),2017,9(2):20-25.

[2] Vijay Kotu, Bala Deshpande. 预测分析与数据挖掘[M]. 北京:人民邮电出版社,2018:52-72.

[3] 周志华. 机器学习[M]. 北京:清华大学出版社,2016:73-95.

[4] 麻书琴. Relief特征选择与混合核SVM在疾病诊断中的研究[D]. 太原:太原理工大学,2017.

[5] Giger ML. Machine learning in medical imaging[J]. J Am Coll Radiol,2018,15(3 Pt B):512-520.

[6] Cao C, Liu F, Tan H, et al. Deep learning and its applications in biomedicine[J]. Genomics Proteomics Bioinformatics,2018,16(1):17-32.

[7] Lee JG, Jun S, Cho YW, et al. Deep learning in medical imaging: general overview[J]. Korean J Radiol,2017,18(4):570-584.

[8] Li S, Jiang H, Pang W. Joint multiple fully connected convolutional neural network with extreme learning machine for hepatocellular carcinoma nuclei grading[J]. Comput Biol Med,2017,84:156-167.

[9] Pang W, Jiang H, Li S. Sparse contribution feature selection and classifiers optimized by concave-convex variation for HCC image recognition[J]. Biomed Res Int,2017,2017:9718386.

[10] Omran DA, Awad AH, Mabrouk MA, et al. Application of data mining techniques to explore predictors of HCC in Egyptian patients with HCV-related chronic liver disease[J]. Asian Pac J Cancer Prev,2015,16(1):381-385.

[11] Wang XQ, Liu Z, Lv WP, et al. Safety validation of decision trees for hepatocellular carcinoma[J]. World J Gastroenterol,2015,21(31):9394-9402.

[12] He X, Wu J, Holtorf AP. Health economic assessment of Gd-EOB-DTPA MRI versus ECCM-MRI and multi-detector CT for diagnosis of hepatocellular carcinoma in China[J]. PLoS One,2018,13(1):e0191095.

[13] Cao XL, Li H, Yu XL, et al. Predicting early intrahepatic recurrence of hepatocellular carcinoma after microwave ablation using SELDI-TOF proteomic signature[J]. PLoS One,2013,8(12):e82448.

[14] Ho WH, Lee KT, Chen HY, et al. Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network[J]. PLoS One,2012,7(1):e29179.

[15] 张朋军, 田亚平. 外周血多参数基因诊断模型对于原发性肝细胞癌诊断价值的评价[J] 标记免疫分析与临床,2014,21(5):499-502.

[16] Qiu J, Peng B, Tang Y, et al. CpG methylation signature predicts recurrence in early-stage hepatocellular carcinoma: results from a multicenter study[J]. J Clin Oncol,2017,35(7):734-742.

[17] Augello G, Balasus D, Fusilli C, et al. Association between MICA gene variants and the risk of hepatitis C virus-induced hepatocellular cancer in a Sicilian population sample[J]. OMICS,2018,22(4):274-282.

[18] Chaudhary K, Poirion OB, Lu L, et al. Deep learning-based multi-omics integration robustly predicts survival in liver cancer[J]. Clin Cancer Res,2018,24(6):1248-1259.

[19] Liao Z, Wang X, Zeng Y, et al. Identification of DEP domain-containing proteins by a machine learning method and experimental analysis of their expression in human HCC tissues[J]. Sci Rep,2016,6:39655.

[20] Liang Q, Liu H, Wang C, et al. Phenotypic characterization analysis of human hepatocarcinoma by urine metabolomics approach[J]. Sci Rep,2016,6:19763.

[21] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics[J]. Nat Rev Genet,2015,16(6):321-332.

收稿日期：2018-04-28

祁亮, 沈洁. TCGA数据库基因突变信息结合机器学习软件RapidMiner构建肝细胞癌患者复发模型[J/CD]. 中国肝脏病杂志(电子版), 2018,10(3):13-19.