三种决策树同源算法在肝部 B 超计算机辅助诊断中的应用比较

李宏彬[1],贺太平[2]
(1.咸阳职业技术学院医学院,陕西 咸阳 712000;
2.陕西中医药大学附属医院影像科,陕西 咸阳 712000)

摘要:目的 探索 CART、随机森林和极端随机树三种决策树同源算法结合不同的特征集进行肝部 B 超图像的分类效果,以期改善计算机辅助 B 超诊断肝脏疾病的准确率。**方法** 从陕西中医药大学附属医院 PACS 系统下载包含正常肝、脂肪肝、图像增粗、肝硬化、肝部占位和肝囊肿和非肝组织的 B 超图像,通过框选 B 超图像生成感兴趣区图块数据,分别对 CART、随机森林和极端随机树算法结合 7 种不同的特征集使用 10 折的交叉校验进行训练和测试,比较各算法的准确率。**结果** 在特定的 B 超切面,随机森林算法结合 C7 特征集"直方图 256 +Python 标准纹理 72+图块定位 3+补充纹理 48"的正确率最高,达 96.31%;另外,3 种算法结合 C7 都能获得很好的交叉校验正确率。**结论** 三种决策树同源算法结合 C7 特征集在 B 超计算机辅助诊断肝脏疾病中都有较高的价值,其中随机森林算法的表现最优。

关键词:B 超;CART;随机森林;极端随机树;肝脏疾病
中图分类号:TP391.41 **文献标识码:**A **DOI:**10.3969/j.issn.1006-1959.2021.19.003
文章编号:1006-1959(2021)19-0013-06

Comparison of Three Decision Tree Homology Algorithms in Computer Aided Diagnosis of Liver B Ultrasound

LI Hong-bin[1],HE Tai-ping[2]
(1.Medical college,Xianyang Vocational and Technical College,Xianyang 712000,Shaanxi,China;
2.Imaging department,Affiliated Hospital of Shaanxi University of Traditional Chinese Medicine,Xianyang 712000,Shaanxi,China)

Abstract:Objective To explore the classification effect of CART, random forest and extreme random tree decision tree homologous algorithm combined with different feature sets for liver B ultrasound images, in order to improve the accuracy of computer aided B ultrasound diagnosis of liver diseases.**Methods** The B-mode ultrasound images of normal liver, fatty liver, image thickening, cirrhosis, liver occupying, liver cyst and non-liver tissue were downloaded from PACS system of Affiliated Hospital of Shaanxi University of Traditional Chinese Medicine. The block data of interest area were generated by selecting B-mode ultrasound images. The CART, random forest and extreme random tree algorithms were trained and tested with seven different feature sets using tenfold cross-validation, and the accuracy of each algorithm was compared.**Results** In the specific sections of B ultrasound, random forest combined with C7 features set (Histogram 256 plus Python standard textures 72 plus block location 3 plus supplementary textures 48) had the highest accuracy, up to 96.31%, and these three algorithms combined with C7 all could achieve good cross validation accuracy. **Conclusion** These three decision tree homology algorithms combined with C7 features set could have good application prospects in computer-aided diagnosis of B ultrasound, and random forest is the best.

Key words:B ultrasound;CART;Random forest;Extreme random tree;Liver diseases

医学图像的计算机辅助诊断是一种机器学习过程，主要研究计算机如何模拟或实现人类的学习行为来获取新的知识或技能，并对现有的知识结构进行重组以提高其性能。计算机辅助诊断是通过对大量数据的分析,从中找出客观规律的技术,其主要包括 2 个步骤:数据准备和规则发现。数据准备是从相关的数据源中选择所需的数据；规则发现是以某种方式找出数据集中所包含的规则。B 超医师每天要处理大量的影像资料,工作强度大,可能会出现误诊和漏诊。但计算机不会因长期工作而产生疲劳,其分析结果是客观的、一致的。因此,计算机辅助诊断已成为 B 超诊断的迫切需求。目前,许多学者已对计算机辅助诊断在肝脏[1]、乳腺[2]、甲状腺疾病[3]等的 B 超诊断中的应用进行了系统深入的研究，但计算机辅助诊断的准确性仍有待提高。基于此,本文以陕西中医药大学附属医院 PACS 系统数据为来源，探索计算机辅助诊断对肝部 B 超图像的分类效果，以期提高计算机辅助 B 超诊断肝脏疾病的准确率。

1 资料与方法

1.1 数据来源 本研究数据来源于陕西中医药大学附属医院超声诊断科。机器学习前对下载的肝部 B 超声像图进行筛选,挑选同一机型(C5-1ABD)下超声探头位姿关联度高、肝区幅面相对比较大的 3 个右肝斜切面(右肋间经右肝隔顶部右肝斜切面、右肋间经第一肝门右肝斜切面、右肋缘下经第一肝门右肝斜切面）进行图像数据采样，共采集正常肝(91 个)、脂肪肝(82 个)、回声增粗(70 个)、肝癌(59 个)、肝囊肿(45 个)和其他肝外正常或病例(93 个)感兴趣区图块合计 440 个。肝部 B 超声像图采集切面见图 1，感兴趣区图块的图像数据的获取通过软件来完成,见图 2。

1.2 感兴趣区数据提取 运行软件后,通过对话框上载一副超声声像图。首先设置一个宽 1200 高 800 的

基金项目:陕西省教育厅专项科研计划资助项目(编号:18jk1212)
作者简介:李宏彬(1974.11-),男,山西大同人,博士,副教授,主要从事计算机辅助诊断方面的研究

注:A:右肋间经右肝隔顶部右肝斜切面;B:右肋间经右肝斜切面;C:右肋缘下经右肝斜切面

图 1 肝部 B 超声像图采集切面示意

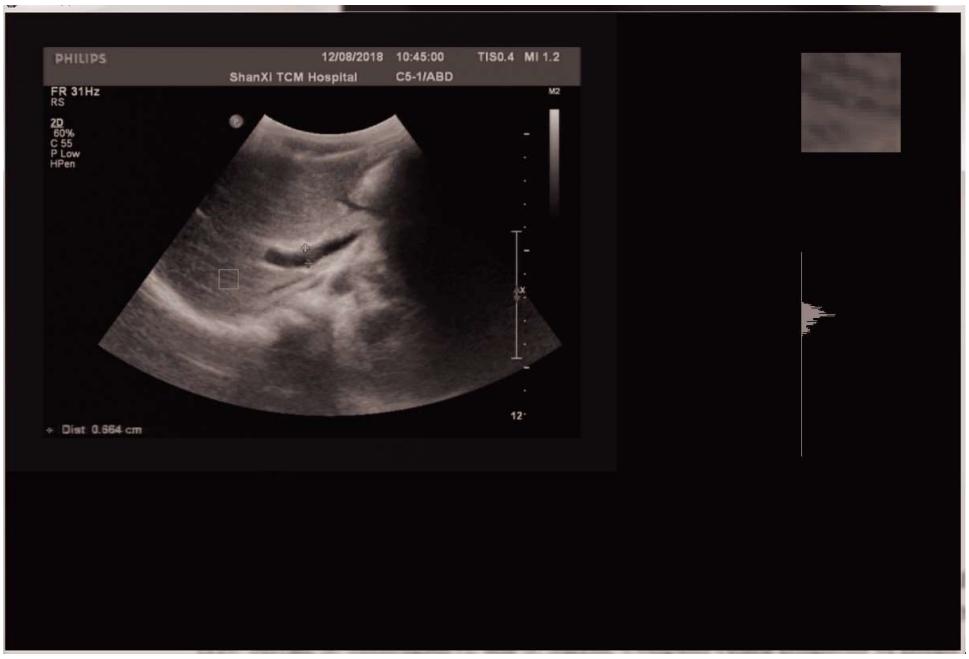

图 2 感兴趣区选取

窗口,窗口从左上角开始向下显示约 800×600 像素的 B 超图像;为使图像细节更细腻,可选择 B 超图像中 25×25 像素的 PNG 图像对感兴趣区图像进行框选,也可对感兴趣区图像进行放大,使其清晰显示;放大窗口下方为感兴趣图像灰阶直方图,从上到下分别代表 0~255 灰阶。当方框移动时,感兴趣区放大图像和直方图同步改变。由于菜单发命令的速度较慢,本软件因而采用快捷键方式发命令,其中键盘的上键、下键、左键、右键分别使方框向上、下、左、右运动 1 个像素的距离,为提高方框的运动速度,又设置键盘的 r、f、d、g 分别使方框向上、下、左、右运动 10 个像素的距离。其他的功能键分别为 q 键退出系统,n 键新建基于电子表格文件 XLS 的分类模式数据包,s 键将当前感兴趣图像的所有特征值数据和图块归类保存到自建的电子表格文件 XLS 数据包中。本软件可设置的类别分别包括正常肝、脂肪肝、回声增粗(肝炎)、肝硬化、肝囊肿、肝占位(肿瘤)和其他(肝外组织或较少见、病灶太小的肝病如肝脓肿,肝内钙化点和肝寄生虫)。当用户设置和确认后,合计 380 个特征值和图块类别写入到用户建立的数据包文件中,供选择其中的部分或全部进行训练。由于电子表格文件通用、直观不涉及数据库版权,在本软件中训练建树的感兴趣区多参数特征数据均以通用的电子表格 Excel XLS 格式保存在自建的数据包(小型数据库)中,涉及电子表格的建立,页(sheet)的建立,写入、覆盖和读。数据包包括 4 个页,第 1、第 2 页的每行用于保存感兴趣区的直方图 256 个灰阶值数据,第 3 页每行用于保存 Python 标准 72 纹理特征值、位置 3 特征值和补充 48 Haralick 纹理特征值,第 4 页用于保存当前块图的灰度共生矩阵,在每页首行各列保存特征的名称,首列各行代表不同图块序号,见图 3。

1.3 主要算法

1.3.1 纹理 纹理是一种反映图像中同质现象的视觉特征,体现了物体表面共有的内在属性,包含了物体表面结构组织排列的重要信息以及它们与周围环境的联系。纹理分析指通过一定的图像处理技术提取出纹理特征参数,从而获得纹理的定量或定性描述的处理过程。纹理分析广泛应用于遥感图像、医学影像、显微图像的解释和处理。如 B 超正常肝回声切

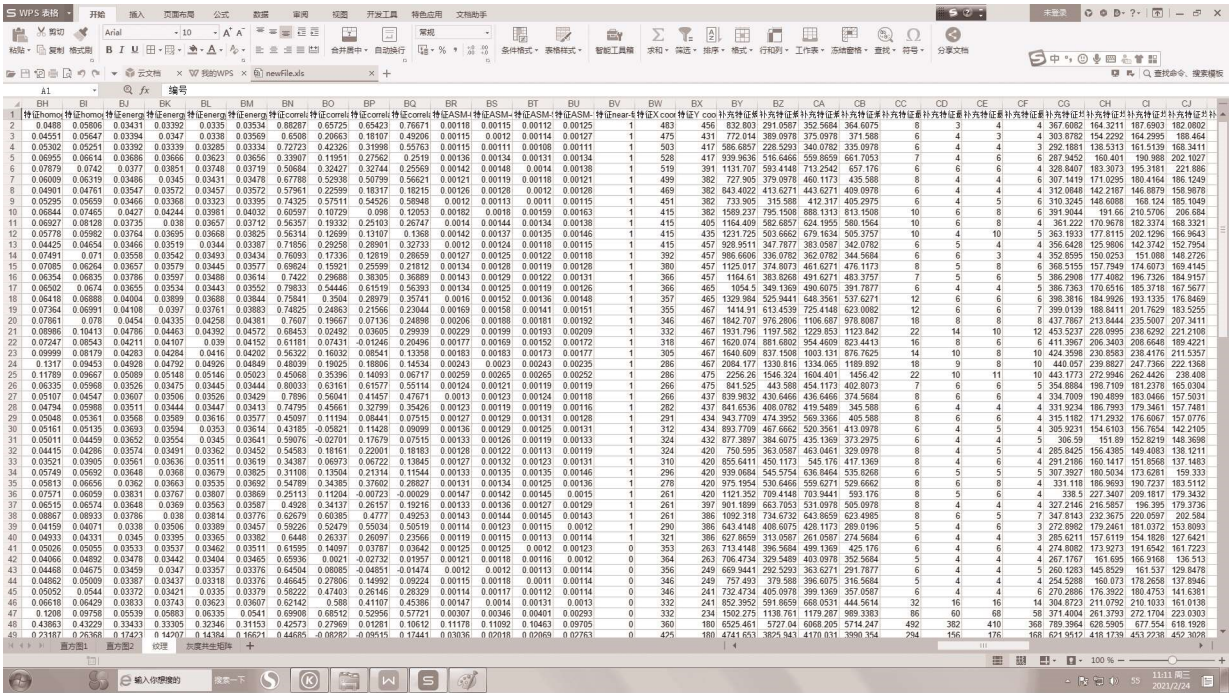

图 3 感兴趣区参数

面光点轻而细，分布均匀，而脂肪肝回声切面呈弥漫性增强，分布均匀致密。目前，图像分析中定义纹理的方法很多，常用的方法有灰度共生矩阵（GLCM）[4]、灰度梯度共生矩阵（GGCM）[5] 和 Laws 能量纹理[6]等。肝脏检查是 B 超检查中最常见的部位，对脂肪肝、肝炎、肝硬化、肝囊肿、肝占位性病变等肝脏疾病的诊断和评估有一定的参考价值。结合机器学习算法，纹理成为识别正常肝脏或异常肝脏疾病的重要手段。

1.3.2 决策树算法　决策树算法是一种通用而优秀的机器学习和模式识别算法，常用于分类和回归。其主要分为 3 种子算法：ID3[7]、C4.5[8]和 CART[9]。通过学习训练数据的特征值和类别建树，依据新数据的特征值并在决策树的引导下对其进行分类。决策树的每个节点表示一个属性，该属性可以是离散特征值（ID3）或连续特征值（C4.5 和 CART）。根据该属性的特征值划分，用户可以从根节点开始，在各个层次逐步访问父节点的子节点，直至到达树的末端叶节点，而每个叶节点均归类为一个特定的类别。决策树非常直观，易于理解和实现，用户在对训练数据学习过程中不需要了解太多的背景知识，只要经生成后的决策树解释后能够理解决策树的含义。决策树的树状结构使其分析速度非常快，能够在较短的时间内对大数据源得到有效的分析和识别。CART 是 Breiman 等于 1984 年提出的一种应用最广泛的分类回归决策树算法。与 ID3 决策树和 C4.5 决策树的多个子样本集节点不同，CART 采用相似的二叉树结构，每次对父样本集节点进行细分，生成 2 个兄弟样本集节点。CART 分类树的建立和改进分为两个步骤：①基于训练样本的数据递归地建立分类树；②利用测试样本的数据对分类树进行剪枝。

1.3.3 随机森林算法　随机森林（random forest）算法是 Breiman 在 2001 年[10]提出的一种基于决策树的机器学习算法。为了得到更稳定、更准确的预测结果，在随机森林中使用一定数量的决策树。随机森林通过从每个训练数据集中随机选取多个数据和属性集中的多个属性来构建决策树，并通过多次抽样替换来构建多个决策树，分类的最终结果由所有决策树中的大多数分类结果决定。随机森林算法的工艺流程如下：从样本集中随机选取 m 个样本，即随机样本；从所有属性中随机选取 k 个属性，构建一个决策树，即随机属性；重复上述两步 n 次，即构建 n 个决策树；以上 n 个决策树形成随机森林，将由 n 个决策树的大多数分类结果来决定新数据的类别。随机林的优点是：它具有非常高的预测准确率，对异常值和干扰具良好的容忍度，且很难出现过度拟合[11]。目前，一些学者使用随机森林算法进行了乳腺癌[11]、阿尔茨海默病[12]、肺结节[13]和宫颈病变[14]等的医学影像的计算机辅助诊断的研究。

1.3.4 极端随机树算法　极端随机树 [15]（Extremely randomized trees）算法与随机森林算法十分相似，都是由许多决策树构成。极端随机树与随机森林的主要区别：随机森林应用的是随机选取的部分样本，极端随机树使用的所有的样本，只是特征是随机选取的，因为分裂是随机的，所以在某种程度上比随机森林得到的结果更加好。随机森林在特征子集中选择

最优分叉特征，而极端随机树直接随机选择分叉特征。优缺点：基本与随即森林类似。由于极端随机树采用所有训练样本使得计算量相对随机森林增大，而采用随机特征，减少了信息增益比或基尼指数的计算过程，计算量又相对随即森林减少。

1.3.5 机器学习 机器学习属于人工智能范畴，其目的是基于现有的训练集来寻找函数，以便以极可能高的正确性预测新的测试集的类别。Python 是一种通用的面向对象编程语言，具有高效、灵活、开源、功能丰富、可扩展性强、表达力强和较高的可移植性等特点，广泛用于科学计算、数据分析与人工智能和机器学习领域。

1.4 特征集生成 GLCM 是对灰度图像某一区域中特定角度的灰度结构的频率进行统计。如对长宽为 m×n 像素的窗口进行纹理分析，概率矩阵 $P_{\Phi,d}(a,b)$ 用于描述窗口中出现的两个像素的频率，其灰度为 a 和 b，在 Φ 方向上和像素距离为 d。基于该矩阵，根据 GLCM 的各属性计算公式，Python 的机器学习库 scikit-learn 中设有灰度共生矩阵 GLCM 和相关纹理特征计算函数（http://tonysyu.github.io/scikit-image/api/skimage.feature.html），但它和上述的标准灰度共生矩阵 GLCM 又有区别，其灰度共生矩阵计算函数为 greycomatrix，用户输入灰度图像矩阵、间距、角度（0°，45°，90°，135°），就可以得到 256×256 的灰度共生矩阵，矩阵的两个维度代表 0~255 的灰度值。属性特征值计算函数为 greycoprops，用户输入 6 个属性的任意一个，就可得到属性的特征值，各属性的定义公式为：

‘Contrast’

$$\sum_{a,b}(a-b)^2P_{\Phi,d}(a,b)$$

‘Dissimilarity’

$$\sum_{a,b}|a-b|P_{\Phi,d}(a,b)$$

‘Homogeneity’

$$\sum_{a,b}\frac{P_{\Phi,d}(a,b)}{1+(a-b)^2}$$

‘Angular Second Moment’

$$\sum_{a,b}(P_{\Phi,d}(a,b))^2$$

‘Energy’

$$\sqrt{ASM}$$

‘Correlation’

$$\sum_{a,b}P_{\Phi,d}(a,b)\left[\frac{(a-\mu_x)(b-\mu_y)}{\sigma_x\sigma_y}\right]$$

当设置像素间距为 1、2 和 3，像素间角度为 0°，45°，90°，135°，使用 Python scikit-learn 库中 6 种标准 GLCM 纹理特征公式后，可得到 3×4×6 共 72 个纹理描述特征。另外，为了提高纹理特征的全面性，又补充 Haralick 的公式：

‘Entropy’

$$\sum_{a,b}P_{\Phi,d}(a,b)\log_2P_{\Phi,d}(a,b)$$

‘Maximum probability’

$$\max_{a,b}P_{\Phi,d}(a,b)$$

‘Inverse Difference Moment’

$$\sum_{a,b}\frac{P_{\Phi,d}(a,b)}{1+|a-b|^2}$$

构成补充纹理特征 3×4×3 共 48 个特征。除了纹理特征外，感兴趣区域图像的灰度直方图也反映图像间差异，为了研究直方图对分类准确性的影响，将其 0~255 的 256 个灰阶频率也列入特征集列入特征集。最后，由于 B 超图像近场图像细腻，远场图像较粗糙，为评估位置对分类准确性的影响，将感兴趣区横、纵坐标，近远场（近场即位于图像的上 1/2，远场位于图像的下 1/2）3 个特征也列入标准特征集。这样就产生了包括“C1:直方图”“C2:Python 标准 72”“C3:Python 标准 72+近远场 1”“C4:Python 标准 72+图块定位 3”“C5:Python 标准 72+图块定位 3+补充纹理 48”“C6: 直方图+Python 标准 72+图块定位 3”和“C7:直方图+Python 标准 72+图块定位 3+补充纹理 48” 7 个不同的特征集。

1.5 软件结构 软件设计主要使用 Python Numpy 库、Matplotlib 库、Pygame 库、Easygui 库、Sklearn 库、Xlrd+Xlwd+Xlutils 库，其中 Numpy 用于科学计算，Matplotlib 库用于绘图，Pygame 用于人机接口，Easygui 用于对话框设计，Sklearn 用于机器学习，Xlrd+Xlwd+Xlutils 用于 Excel 读写操作。软件由三部分自软件构成：其中子程序 pmain.py 用于感兴趣区块图框选和特征值计算和自建 XLS 收取库构建；子程序 pclassifybyCART、pclassifybyRF 和 pclassifybyET 分别用于根据用户使用 CART 决策树、随机森林和极端随机树 3 种算法和用户选择的特征集对未知声像图的感兴趣区进行模式预测即计算机辅助诊断；子程序 plearnbyCART、plearnbyRF 和 plearnbyET 分别用于根据用户 CART 决策树、随机森林和极端随机树 3 种算法和用户选择的特征集（C1~C7 共 7 种），使用户提供的特征数据包进行机器学习的训练、预测和准确率评估。CART 决策树、随机森林、极端随机树机器学习相关的 Python 相关函数如下：

CART 决策树：

建模：CART_decision_tree=tree.DecisionTreeClas-

sifier()

训练：CART_decision_tree.fit(Train_data,Train_label)

预测：predict_result=CART_decision_tree.predict(Test_data)

随机森林：

建模：Random_forest = RandomForestClassifier(n_estimators, random_state, n_jobs)

训练：Random_forest.fit(Train_data,Train_label)

预测:predict_result=Random_forestpredict(Test_data)

极端随机树：

建模：Extra_Trees= ExtraTreesClassifier (n_estimators, random_state)

训练：Extra_Trees.fit(Train_data,Train_label)

预测：predict_result=Extra_Trees.predict(Test_data)

交叉校验：

折数设定：kf = RepeatedKFold(n_splits)

分组校验循环：for train_number, test_number in kf.split (Data,Data_label)

为了评估算法结合特征集的分类准确率，本研究使用 10 折交叉校验方法。随机将用户数据包中的全部样本分成基本均等的 10 组，将其中 1 组设置为检验集，另外 9 组设置为训练集，并使用一定算法如 CART 和特征集如直方图进行训练，然后进行检验，该过程每轮进行 10 次，使 10 组中的每组都得到检验，每次都计算准确率，共进行 10 轮，共 100 次，计算最终的平均准确率。

2 结果

本次首先对全部数据进行训练，又利用全部数据分别使用 CART、随机森林和极端随机树 3 种算法，和 C1~C77 种不同的特征集进行全样本训练、测试和正确率评估，结果见表 1。随后，使用这 440 个样本数据又对上述 3 种算法和 7 种特征集进行 10 折的交叉校验正确率，结果见表 2。3 种算法结合不同的特征集都能无误的通过全样本校验，在交叉校验中，随机森林结合 C7 特征集的正确率最高，达 96.31%，其次是极端随机树结合 C7，达 95.11%，再次是 CART 决策树结合 C7 特征集，达 92.60%，即 3 种算法结合 C7（直方图+Python 标准 72+图块定位 3+补充纹理 48，共 380 个特征）都能获得很好的交叉校验正确率，提示 3 种算法都有很好的应用前景。将 CART 结合 C7 特征集 440 个样本训练和产生的决策树通过 gvedit.exe 软件转化为决策树图，见图 4，其中节点的第一排为特征在 380 个特征的排位。

3 讨论

本研究完成了支持 CART 决策树、随机森林和极端随机树 3 种决策树同源算法，和包含直方图、图块位置和纹理特征的 7 种特征集，采用面向对象和开放源码语言 Python，同时具有图像采集、机器学习和正确率评估多种功能的计算机辅助影像诊断软件的开发工作。

本研究结果表明，在使用若干特定肝部 B 超切面情况下，随机森林、极端决策树和决策树结合 C7 特征集都有很高的正确率，其中随机森林最高达 96.31%。该软件可用于如 B 超影像特别是大器官如肝、脾和肾脏的计算机辅助诊断、科学研究和 B 超影像医师的培训，软件目前还不能完全代替 B 超医师的工作，处于试验和完善阶段。但本研究以积累了大量的研究经验和数据，将医学诊断和人工智能相结合，将来完善后可减轻 B 超医师的工作强度。

本研究的不足之处：①目前软件还不能与 B 超机硬件相结合和获得接口，即用本软件替代 B 超工作站软件，只能对 B 超工作站产生的图像数据进行分析和研究；②机器学习只能针对相同的机型，不同机型的学习结果不能套用；③分析结果准确率受 B 超检查时探头切面位姿一致性的影响，主要因为切面差异会使 B 超图像产生非线性，降低了分析的准确性；④B 超机有多段局部的时间增益控制调节，可对声像图不同高度的亮度进行调节，这也会对图像灰阶和纹理产生不确定影响，增加了学习的难度。针对这些问题，将来可考虑用使用大数据技术来研究 B 超的计算机辅助诊断，以期解决上述问题。

表 1 各算法和特征集全样本校验正确率评估(%)

算法	C1	C2	C3	C4	C5	C6	C7
CART 决策树	100.00	100.00	100.00	100.00	100.00	100.00	100.00
随机森林	100.00	100.00	100.00	100.00	100.00	100.00	100.00
极端随机树	100.00	100.00	100.00	100.00	100.00	100.00	100.00

表 2 各算法和特征集 10 折交叉校验正确率评估(%)

算法	C1	C2	C3	C4	C5	C6	C7
CART 决策树	59.31	63.52	70.21	84.33	86.82	90.91	92.60
随机森林	81.50	79.73	84.30	90.41	86.24	95.23	96.31
极端随机树	84.24	88.12	84.02	90.10	88.71	95.32	95.11

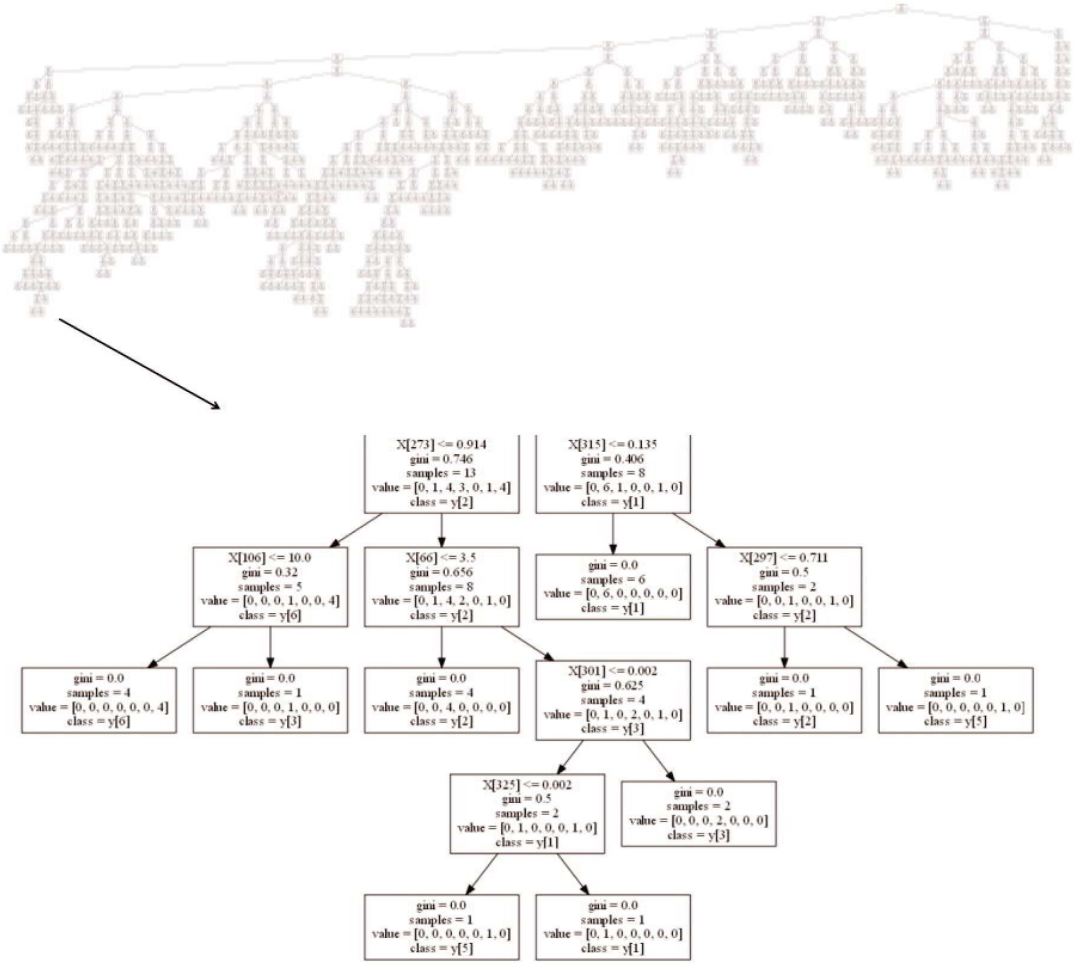

图4 CART算法结合C7特征集对440个肝部图块样本训练产生的决策树及局部放大图

综上所述,3种决策树同源算法结合C7特征集在B超计算机辅助诊断肝脏疾病中都有较高的价值,其中随机森林算法的表现最优。

参考文献:

[1]李海强.肝脏超声图像的计算机辅助诊断识别研究[D].青岛:青岛大学,2019.

[2]孙逸芳,董磊.计算机辅助系统联合超声对甲状腺结节的鉴别诊断价值[J].实用医药杂志,2019,36(1):30-32,36.

[3]刘隆忠,李擎,龙杏章,等.基于计算机辅助诊断技术的超声图像处理软件对甲状腺结节诊断效能的初步研究[J].中华医学超声杂志(电子版),2018,15(12):67-72.

[4]蒋恒,王磊,夏开建,等.基于灰度共生矩阵的ADC纹理分析鉴别直肠癌T3亚分期的临床价值[J].中国医学计算机成像杂志,2019,25(6):48-52.

[5]刘超然,王宁华,李威,等.年龄对肌肉超声图像纹理特征的影响[J].中国康复医学杂志,2020,35(1):33-39.

[6]李光,姜春雪,刘争战,等.Laws纹理能量结合灰度共生矩阵的遥感影像面状地物提取 [J]. 测绘与空间地理信息,2017,40(7):179-181.

[7]迟殿委.一种改进的决策树ID3算法的应用[J].现代计算机,2019(17):43-45.

[8]安葳鹏,尚家泽.决策树C4.5算法的改进与分析[J].计算机工程与应用,2019,55(12):169-173.

[9]Tang R,Zhang X.CART Decision Tree Combined with Boruta Feature Selection for Medical Data Classification [C]// 2020 5th IEEE International Conference on Big Data Analytics (ICBDA).2020.

[10]Ahmad MW,Mourshed M,Rezgui Y.Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption [J].Energy and Buildings,2017(147):77-89.

[11]全雪峰.基于随机森林的乳腺癌计算机辅助诊断[J].软件,2017,38(3):57-59.

[12]李长胜,王瑜,肖洪兵,等.基于随机森林算法的阿尔茨海默症医学影像分类[J].中国医学物理学杂志,2020,37(8):1005-1009.

[13]胡会会,龚敬,聂生东.基于集成随机森林模型的肺结节良恶性分类[J].计算机应用研究,2018,35(10):243-246,251.

[14]姚冰莹,李超,邹贵红.基于随机森林的宫颈病变识别应用研究[J].电脑与电信,2018(11):15-17.

[15]Soltaninejad M,Yang G,Lambrou T,et al.Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in FLAIR MRI[J].International Journal of Computer Assisted Radiology and Surgery,2017,12 (2):183-203.

收稿日期:2021-05-06;修回日期:2021-05-17
编辑/成森