文章编号：2095-2163(2020)01-0274-07 中图分类号：TP391.4 文献标志码：A

基于病理特征和改进随机森林的肺结节分类

孟晋洁，程远志
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：针对基于小样本训练的机器学习模型对肺结节良恶性分类精确度不高的问题，本文提出了一种基于病理特征和改进随机森林的肺结节良恶性分类方法。首先利用灰度级转换、两次区域生长和一次腐蚀膨胀，将肺结节周围的 CT 图像数据完整地分割出来，保留了特征细节，然后提取了语义特征、形态学特征、图形学特征、临床特征组合而成的病理特征，接着选取不同的特征分别训练 2 种随机森林分类器，最后将 2 种分类器进行集成加权得到肺结节良恶性的分类结果。通过 ROC 曲线、*AUC* 值和其他机器学习方法进行对比，表明本文的方法能有效提升肺结节良恶性的分类精确度。

关键词：肺结节良恶性分类；肺结节病理特征；改进随机森林；计算机辅助诊断

A novel pulmonary nodules classification method based on pathological features and improved Random Forests

MENG Jinjie, CHENG Yuanzhi
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 In order to solve the problem that the accuracy of machine learning models training on small dataset is not high in the classification of lung nodules, a new classification method based on pathological features and improved Random Forests for benign and malignant pulmonary nodules is proposed. First of all, the CT image data around the lung nodules are completely separated by gray level conversion, twice region growth and once corrosion expansion, and the feature details are retained. Then, the pathological features are extracted which consist of semantic features, morphological features, graphic features and clinical features. Next, two different types of Random Forests classifiers are trained. Finally, the two classifiers are integratedly weighted to get the classification results of benign and malignant pulmonary nodules. By comparing the ROC curves and *AUC* values with other machine learning methods, it is shown that the method can effectively improve the classification accuracy of benign and malignant pulmonary nodules.

【Key words】 classification of benign and malignant nodules; pathological features of pulmonary nodules; improved Random Forests; computer aided diagnosis

0 引 言

研究指出，肺癌已经成为中国发病率最高的癌症[1]。而且由于被检查出患有肺癌的患者大多数都已进入癌症晚期，致使肺癌的预后非常差，5 年存活率只有 16.1%[2]。但随着目前科技发展，已有文献表明：如果能尽早使用计算机断层扫描技术（Computed Tomography scans, CT）检查出肺部问题，就能够进行有效的治疗。

总地来说，肺结节是指直径在 3~30 mm 的肺内类球形结节，这是肺癌早期的明显特征，如果能及时发现并诊断其良恶性，根据后续跟踪情况辅以手术治疗，对恶性肺结节进行切除，5 年存活率能够提高到 60%[3]。有经验的医生可以对肺结节进行准确判断，但考虑到一张 CT 图像一般有 200 张以上的切片，长时间根据多个特征对肺结节进行判断容易产生视觉疲劳，很可能发生漏判、错判。因此，利用计算机辅助诊断技术（Computer Aided Diagnosis, CAD）提前给出肺结节的良恶性的分类结果供医生参考引证，帮助医生更全面地了解肺部病情，就能够提高效率和减少最终的误判风险。

目前，对肺结节良恶性分类存在着很多难点。首先，肺结节和周围血管、肌肉组织的灰度值十分接近，有的肺结节会粘连在血管、肺壁上，难以分割；其次，肺结节良恶性的区别特征不太明显，有些恶性结节和良性结节的影像非常类似，结节的形态、密度千变万化，有的血管和结节本身的毛刺特征十分类似，对分类是个很大的挑战；最后，如何利用小样本去构造效果较好的分类模型，也是需要攻克的研究难题。

国内外已然涌现对肺结节分类的许多研究。Dhara 等人[4]利用 57 个形状和纹理混合的特征，利用传统的支持向量机（Support Vector Machine, SVM），使用 871 个来自肺部影像数据库联盟（Lung

作者简介：孟晋洁（1995-），男，硕士研究生，主要研究方向：图像处理、机器学习；程远志（1976-），男，博士，教授，博士生导师，主要研究方向：图像处理、模式识别、计算机视觉等。

收稿日期：2018-04-23

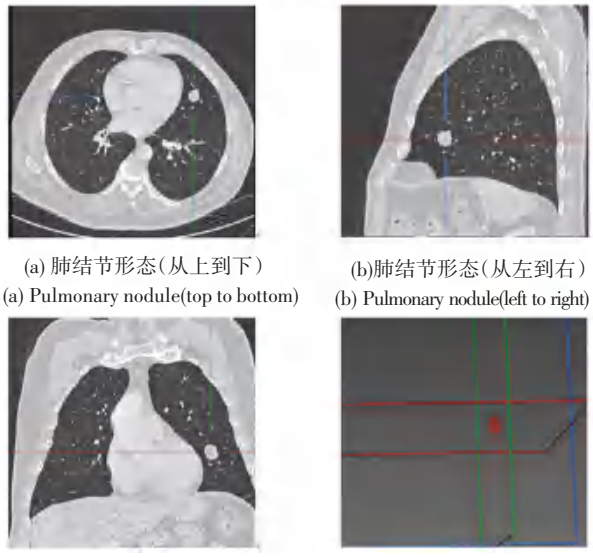
Image Database Consortium，LIDC）的肺结节，得到95.05%的分类准确度，但该方法特征过于稀少，并且容易产生过拟合。Kaya 等人[5]利用病理和其它共 155 个特征与基于潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）、支持向量机 SVM、k 近邻（k-Nearest Neighbour，kNN）、提升方法（Adaboost）、随机森林（Random Forest，RF）的集成分类器，对来自 LIDC 的 2 635 个肺结节进行分类，实验结果表明 RF 和 SVM 分类器的分类效果要明显好于 LDA、Adaboost、kNN，比较贴近本文提出的算法，但是该方法却需要大量的样本进行训练，并且也不适合样本数目较少、而特征较多的肺结节分类。Zhao 等人[6]放弃了传统算法，采用灵活卷积神经网络对 LIDC 的 743 个肺结节进行分类，取得了 82.2%的准确度，适用于小规模的数据库和小的肺结节，但却需要不断地去优化网络的参数才能达到最好的性能，训练比较费时。John 等人[7]对肺结节的特征进行了大量研究[7]，使用基础的阈值分割算法来体现特征的优越性，分割效果较好。Li 等人[8]利用深度卷积神经网络的新型算法对 LIDC 的 40 772 个肺结节、21 720个非肺结节进行训练，取得了较好的结果，但是同时也表现出深度模型比传统算法需要更多的数据、更长的时间去训练的弊端。Li 等人[9]提出了一种改进的 RF 算法分类肺结节，对 LIDC 的肺结节数据良恶性分类灵敏度取得了 92%的业界领先数值，对广州军区总医院的肺结节数据平均灵敏度也获得了 85%的较高结果值，说明 RF 对肺结节的分类是合理、且有效的，但该方法需要的数据集也十分庞大。

综上论述可知，本文针对小样本的胸部 CT 训练图像集，首先利用区域生长和形态学方法对肺结节 CT 图像进行了自动分割，然后根据肺结节形成和发展的特点，提取了语义信息、形态学信息、图像学信息组合成的肺结节良恶病理特征，最后利用改进的随机森林作为分类器进行肺结节良恶性分类，从而达到辅助诊断的目的。

1　胸部 CT 图像中肺结节特点研究

1.1　一般肺结节在 CT 图像中的形态

肺结节通过皮样细胞的堆积进行生长，一般由多核巨噬细胞、大量淋巴细胞组成，因此肺结节在 CT 图像中通常呈现为一个高密度的灰白色类球体，如图 1 所示，图（a）、图（b）、图（c）分别展示了肺结节在横向、侧向、纵向上的切片影像，在这些切片影像中肺结节为一个实心圆，图（d）展示了肺结节的 3D 立体形态，从模型中可以看出肺结节类似一颗肺中的小球，至此对肺结节有了感官上的认识。



(a) 肺结节形态（从上到下）
(a) Pulmonary nodule(top to bottom)

(b)肺结节形态（从左到右）
(b) Pulmonary nodule(left to right)

(c) 肺结节形态（从前到后）
(c) Pulmonary nodule(front to back)

(d) 肺结节形态（3D 建模）
(d) Pulmonary nodule(3D model)

图 1　一般肺结节在 CT 图像中的形态

Fig. 1　The shape of the pulmonary nodules in the CT image

1.2　肺结节良恶性特点

良性肺结节和恶性肺结节虽然难以做到完全区分，但也有一些比较清晰的基于病理学、统计学的辨识特征。根据尺寸大小划分，直径小于 5 mm 的肺结节通常为良性结节，直径大于等于 15 mm 的肺结节通常为恶性结节；根据分叶情况来区分，恶性结节通常沿各个方向生长的速度不一致，表面呈现凹凸结构分叶，越深的分叶恶性程度越高；根据钙化特征来区分，良性结节通常呈现中心、分层、爆米花式的钙化，钙化面积较大、较均匀，因而亮度很高，恶性结节通常在边缘呈针状、小面积、形状迥异的钙化；根据球形程度划分，一般来说良性结节球形程度较高，表面光滑，恶性结节球形程度较低，甚至发展成长条形，表面粗糙；根据毛刺多少来划分，恶性结节表面通常有针尖状的突起，像毛刷一样密集分布，这是恶性结节的典型特征；根据密度来划分，直径大于 8 mm的实性结节存在恶性的可能，直径大于5 mm的密度极小（类似磨玻璃看不清楚）的磨玻璃结节存在更大的恶性的可能；根据空洞程度来划分，恶性结节内部通常会产生坏死、液化，并将这些液化物质通过痰的形式排出体外，造成结节内空洞，空洞越严重说明恶化程度越高；根据边界清晰程度划分，一般肺结节周围出现模糊的现象，就是癌细胞开始侵入良性肺结节引起一系列反应并开始长出毛刺造成的，

这也是恶化的表现,边界越模糊、恶化程度越高;根据位置划分,位于肺上叶的肺结节恶性概率更高。经典的良性肺结节和恶性肺结节如图 2 所示。

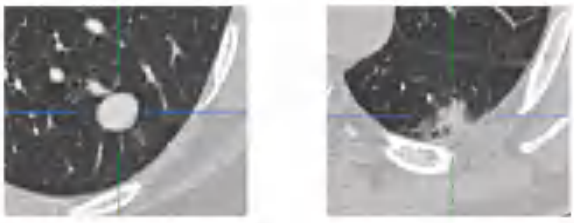


图 2 经典的良性肺结节(左)和恶性肺结节(右)切片图像

Fig. 2 Benign pulmonary nodules (left) and malignant pulmonary nodules (right)

梅奥临床人员提出了一个肺部结节恶性病变预测模型[10],研究 6 个独立的肺结节恶化因素,并可利用式(1)和式(2)计算出肺结节的恶性程度:

$$P(\text{肺结节为恶性}) = \frac{e\alpha}{1 + e\alpha}, \tag{1}$$

$$\alpha = -6.827\,2 + (0.039\,1 \times A) + (0.791\,7 \times S) + (1.338\,8 \times H) + (0.127\,4 \times D) + (1.040\,7 \times B) + (0.783\,8 \times U). \tag{2}$$

其中,e 为自然对数;A 为年龄(岁);S 为是否现在或者曾经吸烟(是 = 1,否 = 0);H 为是否有 5 年以上的胸腔恶性肿瘤史(是 = 1,否 = 0);D 表示肺结节的直径(mm);B 表示肺结节边缘是否有毛刺特征(是 = 1,否 = 0);U 表示肺结节是否位于肺上叶(是 = 1,否 = 0)。从前文设计模型中可以看出,这些病理特征对于判断肺结节的良恶性是非常有用的。

2 基于病理特征和改进随机森林的肺结节分类算法

2.1 CT 图像中肺结节的分割

首先,提取肺结节区域。假设给定一幅 CT 图像的所有横向切片,以 CT 图像左上角为原点,横向向右为 x 轴,纵向向下为 y 轴建立坐标系,同时给定肺结节中心点(cx, cy) 和肺结节所在切片 did。定义以肺结节中心点为中心,中心点所在切片为中心切片,以 $2 \times NW + 1$ 个单位为宽,$2 \times NW + 1$ 个单位为高,前后各 $PNUM$ 个切片、即 $2 \times PNUM + 1$ 为长所构成的长方体为肺结节区域。尤需一提的是,研究中需要重视解决边界问题。如果肺结节区域的宽、高超出CT图像边界,需要进行偏移处理,使得肺结节中心点和肺结节区域中心点重合。为表述方便,一幅CT横向切片中取得的肺结节区域切片即如图 3 所示,左图假设 y 方向两侧都触及了边界,造成原本应该取得 $2 \times NW + 1$ 的高,只取得了阴影面积的高,阴影面积的中心点(cx', cy') 未能与(cx, cy) 重合,为了使二者逐渐接近、直至重合,需要将 y_1 偏离一个 off,y_2 偏离一个 off,研究中推得的 off 的计算公式为:

$$offup = floor((2 \times NW + 1) - (y_2' - y_1' + 1)/2), \tag{3}$$

$$offdown = ceil((2 \times NW + 1) - (y_2' - y_1' + 1)/2). \tag{4}$$

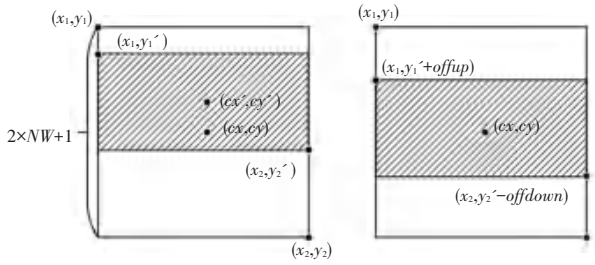


图 3 超出边界示意图(左)和进行偏移处理后(右)

Fig. 3 Beyond the boundary (left) and after the offset adjustment processing (right)

将提取的图像由高灰度级(一般为 -1 024 ~ 1 024)转为标准灰度级(0 ~ 255),并将超出边界的区域设置为背景(0),这样就将肺结节基本分割出来了,粗分割完成。

然后以中心点 (cx,cy) 为种子,以 $2 \times NW + 1$ 为宽,$2 \times NW + 1$ 为高,$theta$ 为灰度值相似度度量,肺结节区域中心切片进行区域生长,得到初步分割结果。然而结节旁边灰度值相似的血管、肺组织等物质也可能被区域生长选中,因此需要以 3×3 的模板进行一次腐蚀和膨胀,去除周围组织的连通性。腐蚀和膨胀只进行一次,这样可以很好地避免毛刺等特征在腐蚀和膨胀的过程中被丢弃。最后再次进行区域生长,得到最终的肺结节分割结果。分割肺结节完整流程图如图 4 所示。

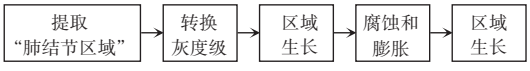


图 4 肺结节分割流程图

Fig. 4 Flow chart of pulmonary nodule segmentation

2.2 肺结节病理特征的选择和提取

基于前面的肺结节良恶性特点,首先可以提取出语义特征,详述如下:

(1)圆形程度,表明肺结节类圆的程度,从长条形(圆 = 0)变化到椭圆形(圆 = 2)、再变化到圆形(圆 = 4)。通常恶性肺结节呈现长条形,良性肺结节呈现圆形。

(2)贴壁程度,表明肺结节和肺部组织粘连的程度,从完全独立(贴=0)变化到完全粘连(贴=4)。

(3)毛刺突起程度,表明肺结节周围针状物质的粗细程度和数量,从无毛刺(毛=0)变化到有非常多细毛刺(毛=4)。通常恶性肺结节具有非常多细毛刺,良性肺结节则表面光滑。

(4)内部空洞程度,表明肺结节内部空洞的大小和数量,从无空洞(洞=0)变化到有非常多的大的空洞(洞=4)。通常恶性肺结节具有非常多的大的空洞,良性肺结节则实心的居多。

(5)固态程度,表明肺结节呈现固态的程度,从非固态(固=0)变化到固态(固=4)。通常恶性肺结节呈现液态,良性肺结节呈现固态。

(6)边缘清晰程度,表明肺结节边缘清晰的程度,从非常模糊(清=0)变化到非常清晰(清=4)。通常恶性肺结节边缘模糊,良性肺结节边缘清晰。

(7)分叶情况,表明肺结节分叶的数目和程度,从无分叶(叶=0)变化到有分叶(叶=4)。通常恶性肺结节有很深的分叶,而良性肺结节则均匀生长。

(8)切片均匀变化程度,表明肺结节的形状是否规则,从完全随机(均=0)变化到部分均匀(均=2),最后到非常均匀(均=4)。通常恶性结节呈现不规则立方体导致切片图像的尺寸、面积发生剧烈改变,良性结节从中心切片到两侧切片的尺寸和面积应该是从大到小均匀变化的。

其次,可以提取出一些形态学的数值特征,分述如下:

(1)坐标位置。表明肺结节在肺部的位置。位于肺上叶的肺结节恶性居多。

(2)切片位置。同样表明肺结节在肺部的位置。

(3)尺寸。可通过肺结节占整个肺结节区域宽高面积的比例来计算,通常恶性肺结节尺寸较大,良性肺结节尺寸较小。研究推得数学公式如下:

$$\text{尺寸} = \frac{pixelNum(\text{肺结节})}{(2 \times NW + 1)^2}. \tag{5}$$

(4)密度变化。表明肺结节的密度变化情况。通常良性肺结节密度变化较小,而恶性肺结节密度变化剧烈,通过灰度值的方差可以很好地反映这一情况,为了避免较少的灰度值产生较大的干扰,取$[0.1\times maxGrayLevel,\ maxGrayLevel]$之间的灰度值进行方差计算,相应数学公式可表示为:

$$\text{密度变化} = var(\text{肺结节灰度值向量}). \tag{6}$$

此外,也可以提取一些图像学特征。诸如,haar特征,表明了肺结节边界的变化情况,采用了5种模板生成的100个随机长宽、随机位置的haar特征,haar模板如图5所示。而haar模板的生成算法可参考如下设计代码。

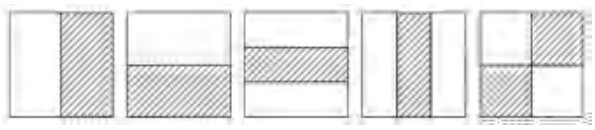


图5　5种haar模板

Fig. 5　5 kinds of haar templates

算法1　haar特征模板随机生成算法

$[haarID,\ leftupx,\ leftupy,\ w] = genHaarTemplate(px, py, haarIDs)$

$num \leftarrow 0$;

FOR $i \leftarrow 1$:$length(haarIDs)$

$ws \leftarrow$ 随机抽取$px \times py$矩形内不重复的、基于haarID模板的、合法的宽度

FOR $j \leftarrow 1$: $length(ws)$

$[leftupx, leftupy] \leftarrow$ 根据宽度、haarID模板,随机抽取不重复的合法的左上角坐标

$num \leftarrow num + 1$;

$rs(num) \leftarrow [haarIDs(i), leftupx, leftupy, ws(j)]$

ENDFOR

ENDFOR

RETURN rs

最后,如果有良好条件可以获取到一些临床信息,也可以加入作为一部分特征,可对此表述如下:

(1)患病年龄。低于35岁的患者肺部的肺结节大多为良性。

(2)吸烟。吸烟者肺结节恶化的可能性高于非吸烟者10~20倍。

(3)性别。一般男性肺结节为恶性的人数高于女性。

(4)生活环境。生活环境如果遭受污染,比如经常有PM2.5的天气,那么肺结节为恶性的可能性很大。

综合上述特征形成了肺结节的病理特征,见表1。

2.3　改进的随机森林肺结节良恶性分类器

肺结节灰度值变化范围较宽,为了避免haar特征对整个特征向量产生影响,并增加RF的稳定性,针对肺结节的特点,本文提出一种加权的随机森林

(Weighted Random Forest, WRF)算法。将特征分为 2 部分,一部分为全部的语义特征、全部的形态特征、全部的临床特征(可选)、haar 特征 α 个,另一部分完全为 haar 特征 β 个,两部分 haar 特征不重叠,分别训练 m 组 RF(记为 MRF)和 n 组 RF(记为 NRF)。每组 RF 中有 tn 棵相互独立的决策树,采用有放回的抽样形式处理训练数据,并且在 NRF 中对 β 个 haar 特征也采用随机抽取的方式,抽取 $fn(0.5tn < fn < tn)$ 个特征值来进行训练。在每棵决策树中,采用 Gini 指数[11]作为熵来进行阈值分割,每棵树的节点最小样本数目设置为 Δ。肺结节良恶性 WRF 的训练算法设计详见如下。

表 1 肺结节病理特征

Tab. 1 Pathological features of pulmonary nodules

特征类别	特征名称
语义特征	圆形程度
	贴壁程度
	毛刺突起程度
	内部空洞程度
	固态程度
	边缘清晰程度
	分叶情况
	切片均匀变化程度
图像特征	haar 特征
形态特征	坐标位置
	切片位置
	尺寸
	密度变化
临床特征(可选)	患病年龄
	吸烟
	性别
	生活环境

算法 2 肺结节良恶性加权随机森林分类器训练算法

$[WRF] = genHaarTemplate(trainFea, m, n, tn, fn, \Delta, \alpha, \beta)$

$[trainFeaM, trainFeaN] \leftarrow splitFea(trainFea, \alpha, \beta)$

FOR $i \leftarrow 1:M$

  FOR $j \leftarrow 1:TN$

    $dtm(i) \leftarrow trainCART(trainFeaM, \Delta)$

  ENDFOR

ENDFOR

FOR $i \leftarrow 1:N$

  FOR $j \leftarrow 1:TN$

    $[selNum, selTrainFeaN] \leftarrow randomSelectFea(trainFeaN)$;

    $dtn(i) \leftarrow trainCART(trainFeaM, \Delta)$

    $selNums(j) \leftarrow selNum$

  ENDFOR

ENDFOR

RETURN $[dtm, dtn, selNums]$

在测试阶段,将提取出来的特征向量分别输入所有的 MRF、NRF 的 RF 分类器中,假设 Y_{mi} 是每组 MRF 的分类结果(概率形式), Y_{ni} 是每组 NRF 的分类结果(概率形式),采用加权投票的方式获得最终的分类结果 C_i ,为此将用到如下数学公式:

$$C_i = sign(\frac{1}{2M}\sum_{mi=1}^{M} Y_{mi} + \frac{1}{2N}\sum_{ni=1}^{N} Y_{ni}). \quad (7)$$

其中, $C_i = 0$ 表示良性肺结节, $C_i = 1$ 表示恶性肺结节。肺结节良恶性 WRF 的测试算法的运行设计参见如下。

算法 3 肺结节良恶性加权随机森林分类器测试算法

$[C] = genHaarTemplate(testFea, WRF)$

FOR $i \leftarrow 1:M$

  $Ym(i) \leftarrow testMRF(testFea, WRF\{dtm\})$;

ENDFOR

FOR $i \leftarrow 1:N$

  $testNFea \leftarrow extFea(testFea, WRF\{selNums\})$

  $Yn(i) \leftarrow testNRF(testFea, WRF\{dtn\})$

ENDFOR

$C \leftarrow 0.5 * sum(Ym) / M + 0.5 * sum(Yn) / N$;

RETURN C

3 实验结果与分析

3.1 实验数据

本次实验采用了 Lungx 肺结节分类挑战赛(LUNGx SPIE - AAPM - NCI Lung Nodule Classification Challenge)的数据[12],训练数据来源于美国国家癌症研究所(National Cancer Institute, NCI) 2014 年 11 月 26 日的癌症影像档案(The Cancer Imaging Archive , TCIA)[13],一共 10 张胸部 CT 图像,5 名男性患者,5 名女性患者,中位年龄 65 岁。5 个良性肺结节,5 个恶性肺结节,共 10 个肺结节,良恶性判断结论是由多个放射科医生根据病例评估和后续跟踪检查得出的。测试阶段一共用到 60 张胸部 CT 图像,23 名男性患者,37 名女性患者,平均年龄 60.5 岁。37 个良性肺结节,36 个恶性肺结节,共 73 个肺结节[14]。CT 图像以 Dicom 的格式保存,每张切片大小均为 512×512 像素,附带肺结

节的位置和良恶性判断数据，统计结果见表 2。

表 2　训练和测试 CT 图像中的肺结节信息

Tab. 2　Pulmonary nodules informations for training and testing CT images

CT 编号	肺结节编号	肺结节中心点坐标	肺结节切片编号	诊断结果
LUNGx-CT001	1	135，303	142	Benign nodule

大赛的组织者精心设计了这些数据，良恶性比例都是被平衡过的，非常适合 WRF 算法。但同时也对分类带来了挑战，对此可做阐释分析如下。

(1)训练数据集非常少。

(2)包括肺结节的尺寸都是被平衡过的，恶性结节也会有很小的尺寸，良性结节也会有很大的尺寸，这是为了避免通过尺寸来构造简单的分类器进行良恶性判断，尽管这种尺寸判断在实际应用中分类精确度很高[15]。

(3)DICOM 头文件信息不允许使用，无法使用包括性别、年龄在内的分类特征，即便使用了也不会有太大效果，因为在性别和年龄上的数据也非常平衡。

3.2　实验结果

设定前后切片数目 $PNUM = 10$，半宽 $NW = 25$，CT 切片边界 $MAXW = 512$，$MAXH = 512$，$theta = 0.15$，对肺结节图像进行分割，分割的效果如图 6 所示，左侧为转换灰度级后的图像，还有很多粘连血管或者独立血管影像。中间图像为第一次区域生长后的图像，还有一部分粘连血管没有清除。右侧为腐蚀膨胀运算后再次区域生长的图像，可以看出粘连的血管已经被分割掉，并且完整地保留了肺结节边缘的形状信息和内部的灰度信息。

(a) 转换灰度级后的图像 (a) Converted grayscale image

(b) 第一次区域生长后的图像 (b) Image after the first region growth

(c) 腐蚀膨胀运算后再次区域生长的图像 (c) Image of region growth after corrosion expansion operation

图 6　训练图像的图像分割效果示意图

Fig. 6　The effect of training image segmentation

设定 $\alpha = 30$，$\beta = 70$。训练 $m = 1$ 组 MRF，$n = 5$ 组 NRF，每个 RF 中产生 $tn = 50$ 棵树，NRF 中随机选择 $fn = [30,60]$ 个 haar 特征，$\Delta = 1$，采用受试者工作特征曲线(receiver operating characteristic curve，ROC)来对比直接使用随机森林(所有参数一致)、使用 SVM(所有参数一致)和本文提出的 WRF 算法的分类效果，如图 7 所示。

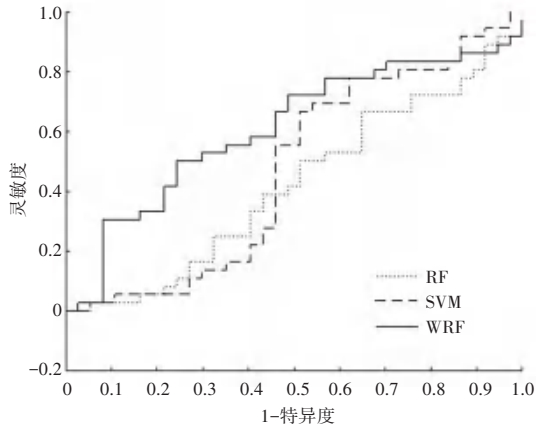


图 7　ROC 曲线对比图

Fig. 7　Comparison of ROC curves

由图 7 可以看出，使用相同数据、相同分割方法、相同特征的 WRF 相对于 RF、SVM 的效果都要好得多。使用相同数据的参赛者的曲线下面积(Area Under Curve，*AUC*)值范围为[0.50，0.68]，表 3 罗列了部分分类方法的 *AUC* 值，*AUC* 在[0.60，0.68]区间内的分类方法只有 3 组，而 WRF 的 *AUC* 值高达0.608 86，已经达到很高的水平，并且远远优于其中在分类器上同样使用随机森林算法进行分类的最高 *AUC* 值 0.56[15]，说明了本算法的优越性。

表 3　相同数据集分类方法效果对比

Tab. 3　Comparison of the effect of classification methods on the same data set

编号	分割方法	分类器	*AUC* 值
1	基于体素分割	SVM	0.50
2	区域生长	WEKA	0.50
3	图割	RF	0.56
4	区域生长+形态学	WRF(本文)	0.61
5	半自动阈值	SVR	0.68

4　结束语

本文提出了一种基于病理特征和改进随机森林的肺结节良恶性分类方法，首先利用肺结节及其周围组织的特性，采用灰度级转换、两步区域生长和一步形态学腐蚀膨胀将肺结节完整地分割出来，最大程度地保留了原始特征；然后逐步提取出语义信息、形态学信息、图像学信息、临床信息作为特征；最后为了充分利用这些小样本高维度特征，提出了一种集成加权的改进随机森林算法 WRF。在 Lungx 数据集上的实验结果表明，WRF 算法相比于其它算

法、尤其对于使用 RF 分类器的算法而言有更好的分类效果。在未来的工作中,需要提取更多在分割步骤中被忽略的特征,同时对比更多机器学习方法在肺结节良恶性分类上的效果,进一步探索 WRF 算法在肺结节辅助诊断中具有广阔重要前景的各类技术研发及实践应用。

参考文献

[1] CHEN Wanqing, ZHENG Rongshou, BAADE P D, et al. Cancer statistics in China, 2015[J]. CA Cancer Journal for Clinicians, 2016, 66(2):115-132.

[2] ZENG Hongmei, ZHENG Rongshou, GUO Yuming, et al. Cancer survival in China, 2003-2005: A population-based study[J]. International Journal of Cancer, 2015, 136(8):1921-1930.

[3] VAN RENS M T, DE LA RIVIÈRE A B, ELBERS H R, et al. Prognostic assessment of 2,361 patients who underwent pulmonary resection for non-small cell lung cancer, stage I, II, and IIIA[J]. Chest, 2000, 117(2):374-379.

[4] DHARA A K, MUKHOPADHYAY S, DUTTA A, et al. A combination of shape and texture features for classification of pulmonary nodules in lung CT images[J]. Journal of Digital Imaging, 2016, 29(4):466-475.

[5] KAYA A, CAN A B. A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics[J]. Journal of Biomedical Informatics, 2015, 56(C):69-79.

[6] ZHAO Xinzhuo, LIU Liyao, QI Shouliang, et al. Agile convolutional neural network for pulmonary nodule classification using CT images[J]. International Journal of Computer Assisted Radiology & Surgery, 2018,13(4):585-595.

[7] JOHN J, MINI M G. Multilevel thresholding based segmentation and feature extraction for pulmonary nodule detection[J]. Procedia Technology, 2016, 24:957-963.

[8] LI Wei, CAO Peng, ZHAO Dazhe, et al. Pulmonary nodule classification with deep convolutional neural networks on computed tomography images[J]. Computational and Mathematical Methods in Medicine,2016, 2016:6215085.

[9] LI Xiangxia, LI Bin, TIAN Lianfang, et al. Automatic benign and malignant classification of pulmonary nodules in thoracic computed tomography based on random forests algorithm[J]. Iet Image Processing, 2018,12(7):1253-1264.

[10] 周清华, 范亚光, 王颖,等. 中国肺部结节分类、诊断与治疗指南(2016 年版)[J]. 中国肺癌杂志, 2016, 19(12):793-798.

[11] BREIMAN L I, FRIEDMAN J H, OLSHEN R A, et al. Classification and regression trees (CART)[M]. Belmont, CA, USA:Wadsworth International Group , 1984.

[12] Armato III, Samuel G, Hadjiiski L, et al. SPIE-AAPM-NCI lung nodule classification challenge dataset[EB/OL]. [2015] Http://DOI.ORG/10.7937/K9/TCIA.2015.UZLSU3FL.

[13] CLARK K, VENDT B, SMITH K, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository[J]. Journal of Digital Imaging, 2013, 26(6):1045-1057.

[14] ARMATO S G, HADJIISKI L, TOURASSI G D, et al. LUNGx challenge for computerized lung nodule classification: Reflections and lessons learned[J]. Journal of Medical Imaging, 2015, 2(2):020103.

[15] ARMATO S G, DRUKKER K, LI Feng, et al. LUNGx challenge for computerized lung nodule classification[J]. Journal of Medical Imaging, 2016, 3(4):044506.

(上接第 273 页)

5 结束语

本文基于 Leap Motion 传感器采集了不同人的 4 组三维动态手势,采用 SVM 模型和经过优化的 PNN 神经网络模型进行识别。本文得到的研究结论可分述如下:

(1)对于三维动态手势,PNN 神经网络可以达到更高的准确率。

(2)多类手势分类的识别率与两类手势分类相比较低,但总体规律不变。

(3)动态手势识别前期采用 PCA 处理数据对 SVM 准确率有所提高,且可以大幅减少 PNN 的运行时间。

未来的工作可以研究新的分类算法,进一步提高其识别率,并探索在各个领域的应用等。

参考文献

[1] 吕蕾,张金玲,朱英杰,等. 一种基于数据手套的静态手势识别方法[J]. 计算机辅助设计与图形学学报,2015,27(12):2410-2418.

[2] 张岚. 基于数据手套的虚拟手人机交互的研究[D]. 武汉:武汉纺织大学,2016.

[3] 于汉超,杨晓东,张迎伟,等. 凌空手势识别综述[J]. 科技导报, 2017,35(16):64-73.

[4] 郭连朋,陈向宁,刘彬. Kinect 传感器的彩色和深度相机标定[J]. 中国图象图形学报,2014,19(11):1584-1590.

[5] 林书坦,尹长青. 基于 LeapMotion 的数字手势识别[J]. 电脑知识与技术,2015,11(35):108-109.

[6] 胡弘,晁建刚,杨进,等. Leap Motion 关键点模型手姿态估计方法[J]. 计算机辅助设计与图形学学报, 2015, 27(7):1211-1216.

[7] GAO Tingting, TIAN Yingjie, SHAO Xiaojian, et al. Accurate prediction of translation initiation sites by universum SVM[C]// The Second International Symposium on Optimization and Systems Biology(OSB2008). Lijiang, China: Asia-Pacific Operations Research Center,2008:279-286.

[8] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40(1):2-10.

[9] 陈晓琴. 基于概率神经网络的潜在客户数据挖掘应用研究[D]. 重庆:重庆交通大学,2011.