ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 5,Sep. 2015,pp. 993-1002
DOI:10.16337/j.1004-9037.2015.05.009

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

基于边缘邻域的乳腺肿块特征提取算法

叶鑫晶　李　洁　王　颖　高新波

(西安电子科技大学电子工程学院,西安,710071)

摘　要:乳腺癌是一种严重威胁人类生命健康的疾病。只有早发现和早治疗才不会错过治疗的最佳时机。乳腺肿块是乳腺癌最主要、最常见的病灶特征,研究乳腺图像中肿块的特征提取,有利于辅助医生诊断,提高医生阅片的效率和正确率。本文针对以往的特征提取方法没有考虑图像的空间信息,造成分类准确率不高的问题,提出一种基于边缘邻域的特征提取算法,使图像特征包含肿块边缘邻域空间信息,其基本思想结合了主动轮廓模型和词袋模型,利用参数控制并确定边缘邻域,对邻域内的特征进行组合或者加权。在保证分类器模型不变的情况下,通过与以往的特征提取算法相比,验证了本算法在分类准确率上优于其他特征提取算法。

关键词:乳腺肿块;特征提取;边缘邻域

中图分类号:TP391　　**文献标志码**:A

Mammographic Mass Feature Extraction Algorithm Based on Edge of Neighborhood

Ye Xinjing, Li Jie, Wang Ying, Gao Xinbo

(School of Electronic Engineering, Xidian University, Xi′an, 710071, China)

Abstract: Breast cancer is one of the most serious diseases greatly threatening human′s health. A patient will not miss the best time for treatment only with early detection and early diagnosis. Mass is the most important and common lesion of breast, so breast mass feature extraction is helpful to improve the efficiency and accuracy of diagnosis. The past algorithms do not consider spatial information of mass images, resulting in low classification accuracy. Aiming at this problem, a new breast mass feature extraction algorithm is proposed based on the edge of neighborhood. It combines the Chan-Vese active contour model with the bag of words. The adaptive parameters regulation methods are designed to control edges of mass images. The final representation can be obtained by combining or weighting those features in the neighborhood. Experimental results show that the proposed methods can achieve a better classification accuracy.

Key words: mammographic mass; feature extraction; edge of neighborhood

引　　言

图像的特征提取是指对图像信息进行测量并且量化的过程,它是数据挖掘领域中最关键的步骤

基金项目:国家自然科学基金青年基金(61201294)资助项目;中央高校基本科研业务费(K5051202048, JB140225)资助项目。
收稿日期:2014-03-25;**修订日期**:2014-05-12

之一,也一直是数据挖掘领域中的研究热点和难点。图像特征提取方法的好坏直接影响着后续图像处理及分析环节的执行结果,运用特征向量对图像进行表示,可以为后续的图像处理和分析奠定基础。

乳腺癌是一种严重威胁人类生命健康的疾病。在全球范围内,乳腺癌的发病数量从1980年的64.1万例,增长到2010年的160万例,每年的平均增长率超过3%,同时,乳腺癌的死亡病例也从1980年的25万例增长到2010年的42.5万例[1]。乳腺癌的诊断方法有很多,其中乳腺钼靶X线摄影由于简便、可靠、成本低、无创等优点,被公认为乳腺癌早期诊断的首选方法[2]。乳腺肿块是乳腺癌最主要也是最常见的病灶特征,有数据表明,85~90%的乳腺癌病例有致密的肿块阴影[3]。乳腺肿块大小不一,边缘特征多样,与乳腺软组织难以分辨,一直是乳腺癌诊断的难点,因此本文将探讨乳腺图像中肿块的特征提取算法。

已有很多研究者对乳腺肿块的特征提取方法进行了研究。文献[4]提取了乳腺肿块的分形纹理特征,分别研究了计盒法、差分计盒法、地毯法、傅里叶功率谱法及分形布朗运动法5种分形维数估计方法,作者利用估计的分形维数构成肿块图像的纹理特征,作为图像的特征表示。同时运用支持向量机(Support vector machine, SVM)分类器作为特征的性能评价标准,实验证明,采用分形布朗运动法估计分形维数产生的图像纹理特征具有最好的受试者工作特征(Receiver operating characteritic, ROC)曲线和良恶性分类准确率。文献[5]使用空间分集方法提取乳腺肿块的特征,并用SVM分类器对提取的特征进行有效性分析,得到了较好的分类效果。文献[6]研究了乳腺图像中肿块的62种纹理和光学特征,并利用线性分类器进行性能判定,证明其中的关联特征、重心特征、中低光密度对比度特征等6种特征能够更好的区分乳腺肿块正常与非正常区域。

尽管取得了一定效果,但这些学者大都通过组合多种图像的全局特征获取新的特征表示,全局特征种类繁多,获取复杂,对噪声、背景敏感,缺乏图像局部特征,无法完整、全面表征图像。为了使图像的特征提取变得简单高效,需要考虑图像的局部特性,目前有研究者将局部二值模式(Local binary pattern, LBP)、方向梯度直方图(Histogram of oriented gradients, HOG)[7]和词袋(Bag of words, BoW)等特征用于肿块图像的特征提取,其中以BoW特征效果最好。但BoW模型缺乏空间信息,容易丢失图像块之间的位置信息,而空间位置信息能够有效表征不同性质肿块区域的边缘分布特性。因此,本文提出一种基于边缘邻域的乳腺图像中肿块的特征提取算法,该方法结合图像的全局与局部特征,在特征表示中加入肿块的边缘空间信息,使乳腺图像的特征表示更为鲁棒。为了评价特征提取算法性能,本文利用支持向量机对肿块进行良恶性分类,实验结果验证了算法的有效性。

1 SIFT与词袋模型特征提取

1.1 SIFT特征

尺度不变特征变换(Scale-invariant feature transform, SIFT)[8]由Lowe于2004年提出,它基于图像局部梯度信息,能充分利用图像的尺度信息,而且对平移、旋转等具有很好的不变性,目前被运用在众多领域。

受HOG特征提取算法的启发,文献[9]提出了一种改进的SIFT特征算法Dense SIFT。这是一种基于网格划分的SIFT特征提取方法,主要思想是将图像分成均匀网格,对网格中的每一个单元提取SIFT特征,有多少网格就会生成多少个SIFT特征点。实验证明Dense SIFT能获得更好的分类结果,主要因为网格中计算的SIFT特征点能提供更多的信息,而经典的SIFT特征算法需要在尺度空间提取关键点,但实际上可能不存在或者少量存在极值点,使得只能提取到少量甚至无法提取到SIFT特征点,导致最后的图像表示较为稀疏。

Dense SIFT特征可以得到图像的局部特征,具有众多优点:(1) 由于Dense SIFT特征是基

于网格的,所以可以人为对网格大小进行设定,得到需要的特征维数;(2) 由于 Dense SIFT 特征是在局部网格内进行操作,所以对图像的光学和几何形变能保持不变性;(3) 对整幅图像进行操作,可以很好地保持图像的整体性和完整性。鉴于以上优点,本文选用 Dense SIFT 作为特征点检测方法。

1.2 词袋模型

BoW 模型最初用于文本处理领域,用来对文档进行识别或分类。由于简单有效的优点,BoW 模型被引入到计算机视觉领域,并且被广泛运用在图像处理领域,用于图像的特征提取和分类[10]。本文的 BoW 特征提取以 Dense SIFT 为特征点选取方法,具体可归纳为以下 4 个步骤。

(1) 特征点检测。在所有的特征提取方法中,感兴趣点的检测是一个区分度较好、稳定性较高的方法。常用的感兴趣特征点提取方法包括 SIFT 特征[8]、HOG 特征等。鉴于 Dense SIFT 特征对旋转、尺度缩放和亮度变化等都能保持局部不变性,并且能够得到更为稳定的特征点及其相应表示,被首先用来于提取肿块图像中的感兴趣特征点。

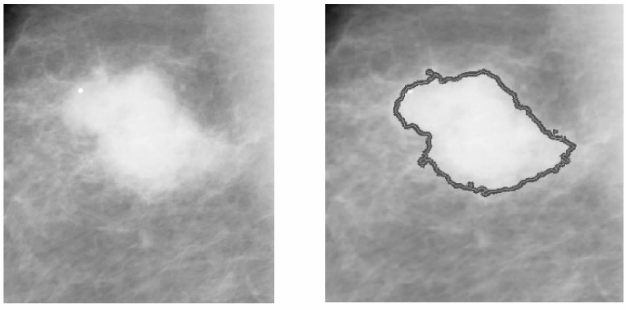
(2) 特征点描述。检测到图像的特征点之后,下一步就要利用特征点周围的局部信息对特征点进行描述,将特征点用高维空间的特征向量表示,得到特征描述子。特征描述子可以看作是文档中的单词。

(3) 生成字典。图像集合中得到的不同特征描述子可能成千上万,如果直接将不同的特征描述子当作字典,那么得到的特征表述将会十分稀疏。可以将特征相似的特征描述子聚为一类,将产生的聚类中心当作字典,即“视觉单词”,字典的个数就是聚类中心的个数。本文采用经典聚类方法 K-means 来提取字典。

(4) 特征表示。得到视觉单词后,就可以用视觉单词来表示每一幅图像,获取每幅图像的特征描述。针对每幅图像的每个特征描述子,寻找与其最接近的视觉单词,并将该特征描述子归属为此视觉单词,最后统计特征描述子在视觉单词上的直方图表示,即图像的 BoW 特征表示。

2 基于边缘邻域的乳腺肿块特征提取算法

在前期工作中获得了分割后的肿块图像[11],如图 1(a)所示。肿块图像的边缘具有很丰富的良恶性鉴别信息,恶性肿块边缘通常呈现毛刺发射状,良性肿块边缘通常圆润光滑,这是区分良恶性的关键。考虑到边缘邻域包含更多有用的信息,有必要通过权重分析来凸显肿块区域的特性。



(a) 原始图像 (a) Original image　　(b) 肿块边缘 (b) Mass edge

图 1 乳腺肿块的边缘提取

Fig. 1 Edge extraction of mammographic mass

2.1 肿块边缘提取

Chan-Vese 主动轮廓模型是基于区域的主动轮廓模型代表[12]，对于提取图像边缘轮廓有良好的效果。

假设轮廓线 C 将肿块图像划分为内部区域 inside(C)和外部区域 outside(C)，则能量泛函可由 $F(C)$ 表示

$$F(C)=\lambda_1\int_{\text{inside}(C)}|u(x,y)-c_1|^2\mathrm{d}x\mathrm{d}y+\lambda_2\int_{\text{outside}(C)}|u(x,y)-c_2|^2\mathrm{d}x\mathrm{d}y+\mu\text{Length}(C) \tag{1}$$

式中：$\lambda_1>0$，$\lambda_2>0$，$\mu\geqslant 0$；c_1，c_2 为曲线 C 内部区域和外部区域的平均灰度；$u(x,y)$ 为肿块图像灰度值。当 $C=C_1$ 时，$F(C)$ 取最小值，据此可定义边缘 C_1。其中 λ_1，λ_2 参数采用文献[12]的推荐取值 1，而 μ 与目标大小呈正比的关系[12]，本文选取 $\mu=1.5$。

在式(1)基础上，首先以肿块区域的核心密度区域初始化轮廓线 C_0，并据经验设置最大迭代次数，初始化参数 λ_1，λ_2 和 μ；再根据轮廓线 C_0 计算初始符号距离函数 φ_0，令 $\varphi^0=\varphi_0$，$n^0=0$。

根据当前的 φ^n，利用式(1)计算 $c_1(\varphi^n)$，$c_2(\varphi^n)$ 以及轮廓线曲率 $\kappa(\varphi^n)$

$$c_1=\frac{\int_\Omega u(x,y)H(\phi(x,y))\mathrm{d}x\mathrm{d}y}{\int_\Omega H(\phi(x,y))\mathrm{d}x\mathrm{d}y} \tag{2}$$

$$c_2=\frac{\int_\Omega u(x,y)(1-H(\phi(x,y)))\mathrm{d}x\mathrm{d}y}{\int_\Omega H(\phi(x,y))\mathrm{d}x\mathrm{d}y} \tag{3}$$

$$\kappa=\nabla\frac{\nabla\phi}{|\nabla\phi|} \tag{4}$$

继而产生迭代方程，求该点下一时刻的 φ 值，并更新轮廓线

$$\frac{\partial\phi}{\partial t}=\delta_\varepsilon(\phi)\left[\mu\nabla\cdot\frac{\nabla\phi}{|\nabla\phi|}-\mu-\lambda_1(u(x,y)-c_1)^2+\lambda_2(u(x,y)-c_2)^2\right] \tag{5}$$

$$\phi(0,x,y)=\phi_0(x,y) \tag{6}$$

最后通过判断迭代次数，或者计算当前轮廓线与旧轮廓线之间的迭代误差来输出最终的肿块边缘，如图 1(b)所示。

2.2 肿块边缘邻域提取

肿块通常边缘模糊、复杂，如恶性病变的放射状边缘，很难完全提取到，往往存在于肿块边缘附近，而良性病变的模糊边界，往往没有精确的边缘可以界定肿块。此外，从医生的角度来讲，肿块边缘附近的组织形变及分布状况，对于病变性质判别有着重要意义。因此，在尽可能准确地获取肿块边缘之后，需要扩充肿块边缘，获取边缘邻域区域。

首先要对肿块边缘图像进行处理，提取最大连通区域，剔除非重点区域，保证每幅图像只保留一个最大的连通区域，也即肿块区域。肿块区域确定之后，就可以在此区域边界的基础上，来自适应寻找不同性质肿块区域的边缘邻域。为了控制边缘邻域大小，定义外延伸参数 α 和内缩进参数 β，其中 α 表征图像外延伸像素比例，用于控制肿块边缘外延伸程度，β 表征图像内缩进像素比例，用于控制肿块边缘内缩进程度。α 和 β 的取值区间分别被设定为(0，0.2)和(0，0.09)。为了控制边缘邻域大小，并且自适应图像大小，需要将外延伸参数 α 分别乘以乳腺肿块图像宽度 w 和高度 h，得到在行、列上的外延伸像素点数目 p_1，p_2，同理，对内缩进参数 β 也作同样处理，乘以肿块图像的宽度和高度，得到在行、列

上的内缩进像素点数目 p_3,p_4。对于每一行,按照定义,p_1 为需要延伸的像素点数目,Chan-Vese 主动轮廓模型提取的肿块边缘位置范围则是$[a_1, a_2]$。假设边缘延伸后肿块像素位置范围是$[a'_1, a'_2]$,其中 a'_1 和 a'_2 分别定义为肿块延伸后边缘像素的最小与最大位置,则肿块内部区域 I_{out1} 的像素需要条件

$$\begin{cases} a'_1 = a_1 - p_1 \\ a'_2 = a_2 + p_1 \end{cases}$$
$$\begin{cases} a'_1 = 0 & a'_1 < 0 \\ a'_2 = w & a'_2 > w \end{cases} \tag{7}$$
$$a'_1 \leqslant I_{out1} \leqslant a'_2$$

每行内部区域 I_{out1} 组合得到图像行内部区域 I'_{out1}。对于每一列,同理,根据参数定义,肿块内部区域 I_{out2} 的像素需要满足条件

$$\begin{cases} a'_1 = a_1 - p_2 \\ a'_2 = a_2 + p_2 \end{cases}$$
$$\begin{cases} a'_1 = 0 & a'_1 < 0 \\ a'_2 = h & a'_2 > h \end{cases} \tag{8}$$
$$a'_1 \leqslant I_{out2} \leqslant a'_2$$

每列内部区域 I_{out2} 组合得到图像列内部区域 I'_{out2}。最后,取行列区域的并集可标记扩大后肿块的内部区域 I_{out}

$$I_{out} = I'_{out1} \cup I'_{out2} \tag{9}$$

这样得到的区域范围较大,可能包含大量背景区域,因此要进一步缩小肿块边缘邻域,已获得包含有用信息的更紧致的边缘邻域。与边缘缩小方法同理,获取延伸后肿块的内部区域 I_{in}。对每一行,肿块行内部区域 I_{in1} 满足条件

$$\begin{cases} a'_1 = a_1 + p_3 \\ a'_2 = a_2 - p_3 \end{cases}$$
$$\begin{cases} a'_1 \leqslant I_{in1} \leqslant a'_2 & a'_1 \leqslant a'_2;\ a'_1 < w;\ a'_2 > 0 \\ I_{in1} = 0 & \text{其他} \end{cases} \tag{10}$$

每行内部区域 I_{in1} 组合得到图像行内部区域 I'_{in1}。对于每一列,肿块内部区域 I_{in2} 满足条件

$$\begin{cases} a'_1 = a_1 + p_4 \\ a'_2 = a_2 - p_4 \end{cases}$$
$$\begin{cases} a'_1 \leqslant I_{in2} \leqslant a'_2 & a'_1 \leqslant a'_2;\ a'_1 < h;\ a'_2 > 0 \\ I_{in2} = 0 & \text{其他} \end{cases} \tag{11}$$

每列内部区域 I_{in2} 组合得到图像列内部区域 I'_{in2}。取行列的交集,标记缩小后的肿块内部区域 I_{in}

$$I_{in} = I'_{in1} \cap I'_{in2} \tag{12}$$

至此,就能够获取肿块边缘邻域 I

$$I = I_{out} - I_{in} \tag{13}$$

调整参数 α 和 β 大小,可以对区域大小进行调整。结果如图 2 所示,从图 2 中可以看出,参数能有效控制肿块边缘邻域的大小,调整合适的参数,算法可以较为准确地覆盖肿块边缘邻域。

2.3 基于边缘邻域的肿块特征提取

根据上述算法,能得到不同肿块区域的边缘邻域。为了进一步凸显肿块区域的特性,可以通过增加边缘邻域区域在整体图像中的权值,引入权重参数 λ,最终获得加权的邻域 BoW 特征。整体流程如

图 3 所示。基于边缘邻域的肿块特征提取方法可详细归纳如下：(1) 调整图像大小并初始化参数 α, β 和 λ；(2) 利用 2.1 节算法提取乳腺图像中肿块的边缘；(3) 利用 2.2 节算法提取肿块边缘邻域；(4) 提取整幅肿块图像的 BoW 特征，其中特征描述子选用 Dense SIFT，聚类算法使用经典的 K-means 聚类算法，得到图像集合的 BoW 特征 F_a；(5) 提取肿块边缘邻域的 BoW 特征 F_e；(6) 对边缘邻域特征加权，称为邻域加权算法，获得含有肿块边缘邻域信息的图像特征表示，得到最终的 BoW 特征表示 F

$$F = F_a + \lambda F_e \tag{14}$$

或者将加权的特征与原始特征组合，称为邻域组合算法，得到最终的 BoW 特征表示 F

$$F = F_a \cup \lambda F_e \tag{15}$$

(a) 边缘图像
(a) Edge image

(b) 边缘邻域提取结果 ($\alpha = 0.15$, $\beta = 0.045$)
(b) Feature extraction result ($\alpha = 0.15$, $\beta = 0.045$)

(c) 边缘邻域提取结果 ($\alpha = 0.075$, $\beta = 0.075$)
(c) Feature extraction result ($\alpha = 0.075$, $\beta = 0.075$)

(d) 边缘邻域提取结果 ($\alpha = 0.03$, $\beta = 0.03$)
(d) Feature extraction result ($\alpha = 0.03$, $\beta = 0.03$)

图 2 参数控制边缘邻域提取

Fig. 2 Edge of neighborhood extraction using parameter control

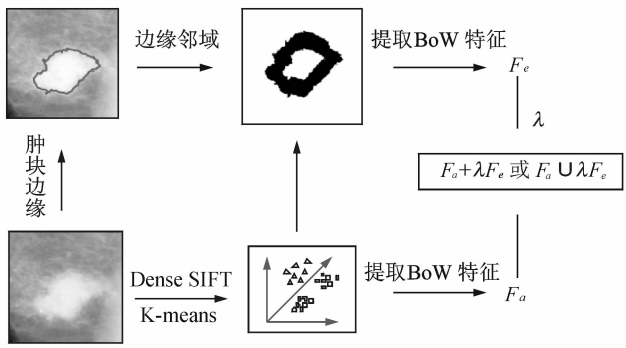


图 3 算法流程图

Fig. 3 Algorithm flow chart

3 实验结果与分析

为了验证算法的有效性，本文对大量乳腺线图像进行了测试实验，测试图像主要来自于 DDSM 数据库[14]中的乳腺钼靶 X 线图像。DDSM 数据库创建于 1999 年，包含了 2 620 个病例，每个病例有 4 个视角的图像，这个数据库创立的初衷就是为了为研究者提供一个评估和比较计算机辅助诊断系统性能的乳腺图像数据集。本文在课题组前期工作的基础上，获取了分割之后的肿块图像。实验的对象是其中的 600 幅图像，其中良、恶性肿块图像各 300 幅。

通过算法获取图像特征之后,本文选用 SVM 分类器对肿块图像进行分类,选定聚类中心数为 200,核函数为 RBF 核,利用基于网格的交叉迭代方法获取 SVM 的最优参数。实验中,随机选取 300 幅图像做训练集,其中良、恶性肿块图像各 150 幅,剩下 300 幅做测试数据集。训练集训练分类器,测试集测试分类结果。重复实验,迭代 100 次,得到迭代分类准确率的均值,作为衡量算法好坏的标准。

实验选择分类准确率及 ROC 曲线作为算法性能的评测指标。分类准确率表征分类器对测试集的分类结果与真值相符合的程度,在保证分类模型相同的情况下,分类准确率直接反映了图像特征的优劣,分类准确率越高,表明图像特征越好,而 ROC 曲线是通用的分类器评价工具,在保证分类模型一致的情况下,ROC 曲线间接反映了特征提取性能的好坏。ROC 曲线中,横坐标为假阳率(False positive rate, FPR),表征良性样本中,被错误判断为恶性的比率,纵坐标为真阳率(True positive rate, TPR),表征恶性样本中,被正确判定为恶性的比率,曲线下的面积越大,则表明特征提取的效果越好。

3.1 参数的影响

算法主要包含 3 个参数,分别为外延伸参数 α、内缩进参数 β 和权重参数 λ,参数的选择决定了效果的优劣。其中外延伸参数 α 和内缩进参数 β 的选择需要覆盖肿块边缘邻域,而这两个参数的选取受 Chan-Vese 主动轮廓算法影响较大,如果迭代次数较大,则可能发生过学习的情况,获取的边缘较大,此时 α 应相对取较小的值,β 取较大的值;如果主动轮廓算法欠学习,则边缘会相对较小,需要选取较大的 α 值和较小的 β 值。在实验中,Chan-Vese 主动轮廓算法的迭代次数设置为 1 000,此时能获取较为准确的边缘区域。相对来说,肿块内部区域较为光滑,信息较少,实验中 β 的取值相对于 α 较小。本文将 α 的取值区间设定为(0, 0.2);β 的取值区间设定为(0, 0.09); λ 的取值区间设定为(0, 5)。

对于本文提出的邻域加权算法,3 个参数对实验结果的影响可从图 4 中看出来。其中,图 4(a)为固定参数 $\beta=0.05$, $\lambda=1$ 的情况下,参数 α 对分类结果的影响;图 4(b)为固定参数 $\alpha=0.08$, $\lambda=1$ 的情况下,参数 β 对分类结果的影响;图 4(c)为固定参数 $\alpha=0.08$,$\beta=0.05$ 的情况下,参数 λ 对分类结果的影响。从图 4 可以看出,3 个参数分别在 $\alpha=0.08$, $\beta=0.05$, $\lambda=3$ 的时候,分类准确率达到峰值,其中 α 大于 β 表明肿块边缘外部信息比内部信息更为丰富,外延伸参数需要设定比内缩进参数更大的值,λ 取 3 表明对邻域特征进行加权的必要性。

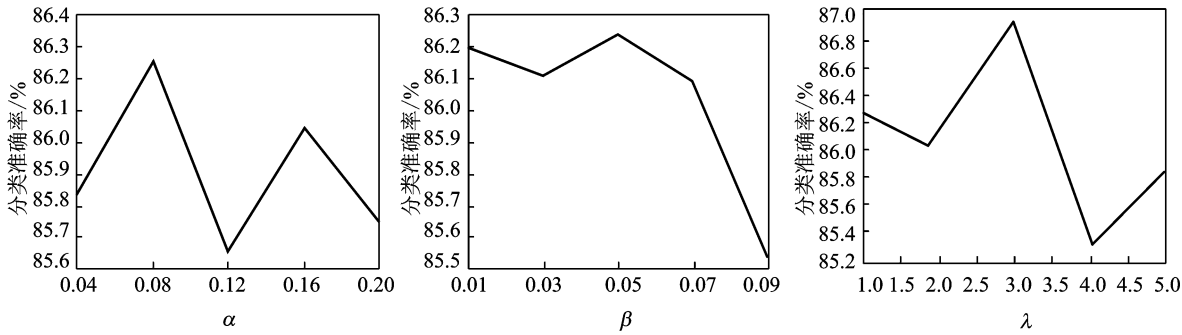
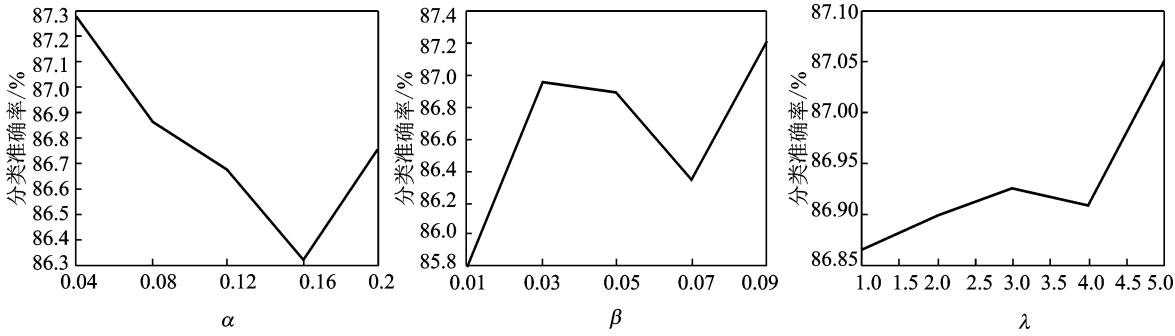


图 4 参数对邻域加权特征的影响

Fig. 4 Influence of parameters on the weighted neighborhood characteristics

对于本文提出的邻域组合算法,3 个参数对实验结果的影响可以分别从图 5 中看出来。与基于邻域加权的算法类似,图 5(a)为固定参数 $\beta=0.05$, $\lambda=1$ 的情况下,参数 α 对分类结果的影响;图 5(b)为固定参数 $\alpha=0.08$,$\lambda=1$ 的情况下,参数 β 对分类结果的影响;图 5(c)为固定参数 $\alpha=0.08$, $\beta=0.05$ 的

情况下，参数 λ 对分类结果的影响。从图中可以看出，3 个参数分别在 $\alpha=0.04$，$\beta=0.09$，$\lambda=5$ 时，分类准确率达到一个峰值。虽然如此，但实验中发现，当(α，β，λ)参数组合为(0.04，0.05，1)时，分类准确率达到最大值。



(a) α参数 (β=0.05, λ=1)
(a) Parameter α (β=0.05, λ=1)
(b) β参数 (α=0.08, λ=1)
(b) Parameter β (α=0.08, λ=1)
(c) λ参数 (α=0.08, β=0.05)
(c) Parameter λ (α=0.08, β=0.05)

图 5 参数对邻域组合特征的影响

Fig. 5 Influence of parameters on the neighborhood combination characteristics

3.2 实验结果对比

本文与多种特征提取算法进行了对比试验，对比的特征提取方法包括数学特征提取方法(Mathematical feature, MF)、潜在狄利克雷分布(Latent Dirichlet allocation, LDA)方法、空间词袋特征[19]方法(S-LDA)、HOG 特征提取方法和 BoW 方法。其中数学特征提取方法提取了图像的灰度均值、标准差、平滑度等 16 种数学特征，组成了 22 维的图像特征表示，该 16 种特征包括：(1)灰度特征。包括均值、标准差、方差、三阶矩、四阶矩、平滑度、一致性、峰值和不变矩(7 维)；(2)纹理特征。包括粗糙度、包括对比度、方向度、线性度、规则度、粗略度和熵。LDA 特征提取方法建立在 BoW 模型基础上，提取的是图像的主题特征，本文选择主题数为 80，此时 LDA 方法有较好效果；S-LDA 方法利用文献[19]算法，用弧形对图像进行分块，获取图像的空间词袋特征，由于该特征维数过大，容易造成“维数灾难”，产生各种不确定问题，所以用 LDA 方法进行降维，实验中设置主题数为 200，最终得到 200 维的图像主题特征；BoW 算法中聚类中心数目设置为 200。

对于邻域加权算法，选择参数 $\alpha=0.08$，$\beta=0.05$，$\lambda=3$ 进行实验，其分类准确率达到最大，为 86.67%。对于邻域组合算法，选择参数 $\alpha=0.04$，$\beta=0.05$，$\lambda=1$ 进行实验，其分类准确率达到最大，为 87.28%。分类实验结果如表 1 所示，ROC 曲线如图 6 所示。

由表 1 可见，MF 与 LDA 算法分类准确率都不高，LDA 特征提取算法建立在 BoW 算法的基础上，特征维数少于 BoW 算法，不可避免地损失了部分图像信息，分类准确率不高，而 S-LDA 由于加入空间信息，在分类准确率上优于 LDA 但低于原始 BoW 特征。HOG 算法是一种典型的基于局部图像的特征提取算法，但实验表明，BoW 算法具有更好的效果。而相比其他算法，本文提出的特征加权与特征组合算法具有最高的分类准确率，这是由于本文提出的算法包含了乳腺图像中肿块的边缘信息，凸显了肿块的边缘特性，能更为鲁棒地表征不同性质肿块区域的主要特性。图 6 中 ROC 曲线的对比则进一步验证了本文算法的有效性，本文提出的算法具有更好的 ROC 曲线。

从表 1 及图 6 中可以看到，邻域组合特征的性能稍好于邻域加权特征，这说明在乳腺肿块数据中，特征拼接效果好于特征叠加效果，特征拼接将肿块边缘特征独立出来与肿块图像的全局特征进行拼接，虽然增加了特征维数，但使最终的特征表示既包含边缘邻域的局部特征，又包含图像的全局特征，具有

更好的效果;特征叠加则是将边缘特征直接与图像全局特征进行叠加,虽然保持特征维数不变,但仍旧以图像的全局特征为主要表现形式,性能稍逊于特征拼接方式。尽管邻域组合与邻域加权算法各有优劣,但算法性能依旧明显优于其他特征提取算法。

表 1 分类准确率

Table 1 Classification accuracy

算法	MF	LDA	S-LDA	HOG	BoW	加权	组合
维数	22	80	200	81	200	200	400
准确率/%	73.77	75.43	80.10	84.08	85.10	86.67	87.28

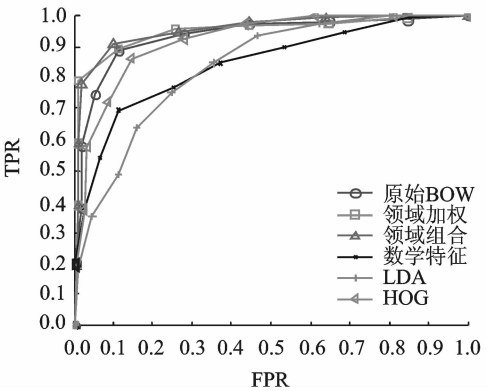


图 6 ROC 曲线对比

Fig. 6 ROC curve comparison

综上可以看出,由于本文提出的算法突出了包含丰富信息的乳腺肿块边缘,能够更好地描述肿块特征,在分类准确率上优于其他特征提取算法,对乳腺肿块图像的分类有更好的效果,有利于辅助放射科医生进行医学诊断。

4 结束语

本文在引入 BoW 词袋模型的基础上,提出了一种针对乳腺钼靶 X 线图像中肿块的特征提取算法,并对算法的参数选择进行了分析。与以往的针对肿块区域的特征提取算法相比,本文提出的方法获取了乳腺肿块的边缘邻域信息,使图像的特征表示具有边缘邻域的空间信息,能更为鲁棒地表示不同性质肿块区域的边缘分布性,提高了分类准确率。但是如何自动对参数进行选择,以及如何对组合特征进行有效降维,是值得进一步研究的课题。

参考文献:

[1] Institute for Health Metrics and Evaluation. The challenge ahead: Progress and setbacks in breast and cervical cancer[EB/OL]. http://www.healthmetricsandevaluation.org/publications/policy-report/challenge-ahead-progress-and-setbacks-breast-and-cervical-cancer. 2014-5-10.

[2] Miller A B. Mammography: Reviewing the evidence[J]. Epidemiology Aspect, Canadian Family Physician, 1993, 39: 85-90.

[3] 张召长. 基于 X 线乳腺肿块分割与检测的研究[D]. 南京:南京航空航天大学, 2009.
Zhang Shaochang. Research on mammographic mass segmentation and detection[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2009.

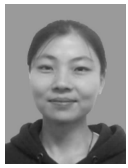
[4] Guo Q, Shao J, Ruiz V F. Characterization and classification of tumor lesions using computerized fractal-based texture analysis and support vector machines in digital mammograms[J]. International Journal of Computer Assisted Radiology and

Surgery, 2009, 4(1): 11-25.
[5] Braz J G, da Rocha S V, Gattass M, et al. A mass classification using spatial diversity approaches in mammography images for false positive reduction[J]. Expert Systems with Applications, 2013, 40(18): 7534-7543.
[6] Sameti M, Ward R K, Morgan-Parkes J, et al. Image feature extraction in the last screening mammograms prior to detection of breast cancer[J]. IEEE Journal of Selected Topics in Signal Processing, 2009, 3(1): 46-52.
[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005, 1: 886-893.
[8] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.
[9] Bosch A, Zisserman A, Munoz X. Scene classification via pLSA[C]//European Conference on Computer Vision. Graz, Austria: Springer Berlin Heidelberg, 2006: 517-530.
[10] Li F F, Perona P. A Bayesian hierarchical model for learning natural scene categories[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005, 2: 524-531.
[11] 王颖. 乳腺X线图像中肿块的计算机辅助检测与分析[D]. 西安:西安电子科技大学, 2010.
Wang Ying. Computer-aided detection and analysis of mammographic mass[D]. Xi'an: Xidian University, 2010.
[12] Chan T F, Vese L A. Active contours without edges[J]. IEEE Transactions on Image Processing, 2001, 10(2): 266-277.
[13] 职占江. 基于Chan-Vese模型的图像分割算法[D]. 郑州:河南大学, 2008.
Zhi Zhanjiang. Image segmentation algorithm based on Chan-Vese model[D]. Zhengzhou: Henan University, 2008.
[14] University of South Florida. DDSM: Digital database for screening mammography[EB/OL]. http://figment.csee.usf.edu/Mammography/Database.html, 2007-11-01/2014-3-20.
[15] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2006, 2: 2169-2178.
[16] Smart C R, Hendrick R E, Rutledge J H, et al. Benefit of mammography screening in women ages 40 to 49 years[J]. Current Evidence from Fandomized Controlled Trials, Cancer, 1995, 75(7): 1619-1626.
[17] Cheng E, Xie N, Ling H, et al. Mammographic image classification using histogram intersection[J]. IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2010: 197-200.
[18] 刘燕杰, 卢振泰, 冯前进, 等. 基于KL距离加权和局部邻域信息的CV模型[J]. 电子学报, 2011, 39(6): 1447-1451.
Liu Yanjie, Lu Zhentai, Feng Qianjin, et al. KL distance weighted CV model based on local neighborhood information [J]. Acta Electronica Sinica, 2011, 39(6): 1447-1451.
[19] Cao Y, Wang C, Li Z, et al. Spatial-bag-of-features[C]//IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, California, USA: IEEE, 2010: 3352-3359.

作者简介:

叶鑫晶(1987-),男,硕士研究生,研究方向:图像处理与分析,E-mail: yingwang@xidian.edu.cn。

李洁(1972-),女,教授,研究方向:机器学习、计算机视觉和影像处理与分析。



王颖(1981-),女,副教授,研究方向:模式识别、影像处理与分析。

高新波(1972-),男,教授,研究方向:计算机视觉、机器学习和智能信息处理。