

基于两步聚类 and 随机森林的乳腺腺管自动识别方法

王 帅 刘 娟 毕姚姚 陈 哲 郑群花 段慧芳

(武汉大学计算机学院 武汉 430072)

摘 要 腺管的自动识别在乳腺癌的组织病理学诊断中十分关键,因为腺管密度是乳腺癌分级中的一个重要因子。腺管由一个周围充满细胞质的中心管腔以及管腔周围均匀环绕的细胞核组成。若管腔、细胞质、细胞核在空间位置上接近,则意味着这可能是一个腺管,但是这种识别方法会因为乳腺组织切片中存在脂肪、气泡以及其他类似管腔的对象而出现假阳性错误。为了解决上述问题,提出基于二次聚类与随机森林的腺管自动识别方法。首先通过一次聚类和二次聚类构建出待分割图片;然后通过形态学操作对图片进行处理,并在此基础上进行分割,进而构建候选腺管,利用中心管腔与其周围细胞核的空间位置关系以及一些统计特征来描述腺管;最后通过随机森林分类算法进行分类。实验结果表明,所提算法可以达到 86% 以上的准确率,为乳腺癌的自动分级奠定了基础。

关键词 乳腺癌,病理图像,病理诊断,腺管分割,腺管识别,形态学操作,计算机辅助诊断

中图分类号 TP751 文献标识码 A DOI 10.11896/j.issn.1002-137X.2018.03.039

Automatic Recognition of Breast Gland Based on Two-step Clustering and Random Forest

WANG Shuai LIU Juan BI Yao-yao CHEN Zhe ZHENG Qun-hua DUAN Hui-fang

(Computer School, Wuhan University, Wuhan 430072, China)

Abstract Automatic recognition of the glands is critical in the histopathology diagnosis of breast cancer, as glandular density is an important factor in breast cancer grading. The gland is composed of a central lumen filled with cytoplasm and a ring of nuclei around the lumen. The spatial proximity of the lumen, cytoplasm, and nucleus may mean that it is a gland, but this method can lead to false-positive errors due to the presence of fat, bubbles and other lumen-like objects in the breast tissue section. In order to solve the above problems, this paper presented an automatic recognition method of breast gland based on two-step clustering and random forest. First, the images to be segmented are constructed by clustering and two-step clustering. A series of morphological operations are performed on the images to repair the objects. Then the segmentation is performed. After that, the method builds the candidate glands, and utilizes the spatial position relationship between central lumen and the nucleus around the lumen and some other features to describe glands. By using random forest classification algorithm, the experimental results show that more than 86% accuracy can be achieved. The result lays the foundation for breast cancer automatic grading.

Keywords Breast cancer, Histopathology image, Histopathology diagnosis, Gland segmentation, Gland recognition, Morphological operations, Computer-aided diagnosis

1 引言

近年来,我国乳腺癌的发病率和死亡率不断上升。乳腺癌是我国女性最常见的恶性肿瘤之一,约占女性癌症患者的 25%。乳腺癌是指乳房的腺体或导管上皮细胞发生癌变。乳腺癌的诊断方式包括影像学诊断、病理诊断、分子诊断。病理

诊断是指通过组织活检或手术切除的方法,取一小块病灶组织制成 HE 染色切片(见图 1),然后在显微镜下诊断是否存在癌细胞,若存在则判断癌细胞的分化程度、分级程度。组织病理学诊断是乳腺癌确诊和治疗的依据,是通过综合分析临床的各种信息及病理形态得出的最后诊断,此步骤为确诊乳腺癌的金标准。

到稿日期:2017-06-22 返修日期:2017-08-22

王 帅(1992—),男,硕士生,主要研究方向为计算机视觉、数字图像处理,E-mail:stamaimer@foxmail.com;刘 娟(1970—),女,博士后,博士生导师,CCF 高级会员,主要研究方向为数据挖掘、生物信息、自然语言处理、自然计算、机器学习,E-mail:liujuan@whu.edu.cn(通信作者);毕姚姚(1994—),女,硕士生,主要研究方向为机器学习、数字图像处理,E-mail:1311060843@qq.com;陈 哲(1993—),女,硕士生,主要研究方向为计算机视觉、数字图像处理,E-mail:785307750@qq.com;郑群花(1989—),女,硕士生,主要研究方向为机器学习、数字图像处理,E-mail:961290964@qq.com;段慧芳(1991—),女,硕士生,主要研究方向为机器学习、数字图像处理,E-mail:774570667@qq.com。

目前,乳腺癌的诊断由病理科的病理医生在显微镜下对病人的活组织样本切片进行视觉分析完成。人工诊断既耗时,又有很大的不确定性。为了克服人工诊断的弊端,基于计算机的辅助诊断系统被提出。辅助诊断系统可以在分割组织结构以及提取组织结构特征方面给予病理医生帮助,有助于缩短病理医生的诊断时间,并使得诊断结果相对客观。

浸润性导管癌(IDC)是一种最常见的乳腺癌。80%的乳腺癌都是浸润性导管癌。乳腺癌的组织学分级主要针对浸润性导管癌的浸润性癌部分。腺管的形成程度是组织学分级的一个重要因子。乳腺腺管通常是圆形或者椭圆形结构,管腔周围均匀地环绕着一层上皮细胞。在癌症的形成过程中,细胞之间的粘附性以及上皮细胞形成腺管结构的功能受到破坏,腺管结构随之消失。病理学家认为,腺管的形成反映了组织结构的分化程度。准确分割并识别腺管是构建基于计算机的辅助诊断系统的重要环节。

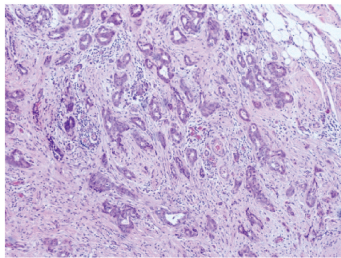


图1 HE组织病理图像(浸润性导管癌)

Fig. 1 HE histopathology image (invasive ductal carcinoma)

2 相关工作

针对乳腺癌的 HE 染色组织病理图像,国内外的研究者提出了大量的细胞核分割算法,如基于模糊聚类的细胞核自动分割算法^[1]、基于自适应阈值法的细胞核分割算法^[2]、基于活动轮廓模型和无监督学习的细胞核分割算法^[3]。因为腺管并非像细胞核那样独立存在,通常是圆形或者椭圆形结构,管腔周围均匀地环绕着一层上皮细胞,所以传统的细胞核分割算法不适用于腺管等复杂对象的识别问题。

针对腺管的结构特点,最近有不少学者提出了腺管的分割算法。例如,徐军等人^[4]提出了一种变种的活动轮廓模型,首先从病理图像中选择感兴趣的区域样本,然后使用带权重的均值漂移算法对像素点进行聚类,之后使用 N-Cut^[5]算法解决过分割的问题,用于产生活动轮廓模型算法的初始轮廓。实验结果显示,其分割效果优于基于区域的 Chan-Vese^[6]主动轮廓模型。Ajay 等人^[7]结合病理专业知识更深入地扩展上述工作,利用管腔和细胞核的空间位置关系区分腺管与其他管腔状对象。Maqlin 等人^[8]使用网格分析的方法获得水平集方法的初始轮廓,从脂肪、空腔中识别出腺管。Kien 等人^[9]在构建管腔之前,先将细胞核分成肿瘤细胞核和正常细胞核,使用肿瘤细胞核而不是所有细胞核来构建腺管并进行腺管分割。Korsuk 等人^[10]对 GlaS@ MICCAI' 2015 中使用的肠腺腺管识别方法 CUMedVision, CVML, ExB, Image

Analysis Lab Uni Freiburg, LIB, vision4GlaS 等进行概述总结,从 F1 分数、Object Dice、Object Hausdorff 3 个方面进行评估,其中 CUMedVision 的表现最好。Janowczyk 等人^[11]基于深度学习方法,使用不同的 Patch 选择技术进行细胞核分割、上皮细胞分割、腺管分割等。目前乳腺癌腺管分割的相关文献并不多,而与之非常相关的应用如前列腺癌组织病理图像中腺管分割的研究则相对广泛^[12-13],因此可对应用于前列腺癌腺管分割的方法进行适当修改后用于乳腺癌腺管分割。

现有方法大致可以分为两类:1)基于区域增长算法直接识别腺管;2)首先识别构成腺管的低层次对象(细胞核、管腔),然后结合病理专业知识,利用低层次对象的位置关系构建腺管,最后进行分类。上述文献中提出的方法需要根据图像特征手动设置参数,通用性不高;细胞核的分割结果不理想,同时用于寻找腺管轮廓的算法过于复杂,耗时较长。因此,根据乳腺 HE 染色组织病理图像的特点,提出基于二次聚类与随机森林的腺管自动识别方法。该方法主要包含如下 4 个部分(见图 2):首先,对图像进行第一次聚类,获得管腔结构;然后,基于第一次聚类的结果进行第二次聚类,获得细胞核信息;接着,基于两次聚类的结果确定候选腺管,并据此提取特征;最后,利用随机森林算法构造识别腺管的分类模型。与现有算法相比,该方法强调细胞核分割的准确性以及低层次对象识别的效率。

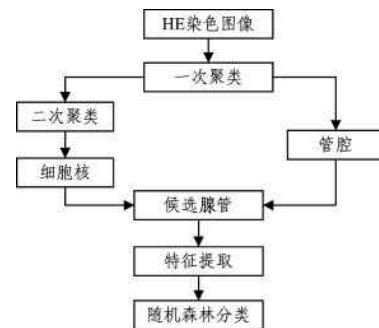


图2 系统概述流程图

Fig. 2 Flow chart of system overview

3 基于聚类和形态学操作的管腔识别

腺管由一个周围充满细胞质的中心管腔以及管腔周围均匀环绕的细胞核组成。若管腔、细胞质、细胞核在空间位置上接近,则意味着这可能是一个腺管。因此,腺管识别需要分别获取管腔和细胞核。乳腺组织 HE 染色病理图像由一些像素点组成,可以根据颜色把这些像素点分为 3 类:紫色的细胞核,粉色的细胞质,白色的管腔、脂肪、气泡。

使用 KMeans 聚类算法将像素点的像素值作为特征对像素点进行聚类,并提取出感兴趣的像素点来重构图像。通过计算每类像素点的灰度均值然后进行排序,即可将每类像素点对应到相应的对象上:灰度均值最低的类对应紫色的细胞核,灰度均值最高的类对应白色的管腔、脂肪、气泡。

KMeans 聚类结果如图 3 所示。

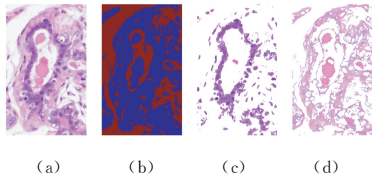


图 3 KMeans 聚类结果

Fig. 3 Result of KMeans cluster

聚类后,从聚类结果中得到重建的关于管腔、脂肪和气泡等的二值图像(见图 3(b))、细胞核图像(见图 3(c))、细胞质图像(见图 3(d))。本节的目的是从这张二值图像中识别出管腔状对象,结合图 3(a)和图 3(b)可知:蓝色为背景,红色为管腔、脂肪和气泡等对象。根据图 3(b)中对象的形态结构和位置,可以将图中对象分为 3 类:1)与图片边界相连的对象;2)琐碎细小的对象;3)不完整的圆形或椭圆形对象。其中第三类是管腔的可能性最大,其余两类均会影响识别结果,需要使用形态学操作来移除前两类对象或尽可能地减小其造成的影响,并对管腔内的二值图像进行形态学操作的各步结果如图 4 所示。

管腔识别的步骤如下:

1)对二值图像进行 clear_border 操作,该操作可移除二值图像中所有与图像边界相连的物体。

2)通过图 3(a)可以看到,腺管中的管腔出现了粉色的部分,在聚类过程中,这些粉色的部分被标记为细胞质,导致在重建关于管腔、脂肪和气泡等对象的二值图像时,管腔区域出现了孔洞。为了修复这些区域,对图 4(c)进行形态学闭操作。

3)形态学闭操作并未完全修复这些孔洞,需要在图 4(d)的基础上再进行填洞操作。

4)图中存在一些细小的对象,这些对象不是需要分割的管腔,同时在进行形态学闭操作时,这些对象改变了管腔的面积。鉴于上述两种问题,在图 4(e)的基础上进行形态学开操作。

5)在图 4(f)的基础上进行识别,使用 SuperPixel 的概念表示每个被识别出的管腔对象。针对每个被识别出的管腔状对象,计算圆度、周长、中心坐标、面积以及轮廓坐标等属性,用于后续的腺管识别。

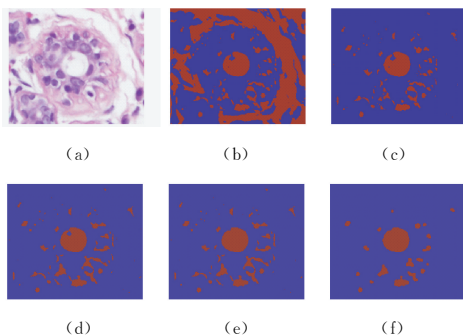


图 4 形态学操作的结果

Fig. 4 Result of morphological operations

4 基于再聚类的细胞核识别

在聚类结果中,把不属于细胞核类的像素点的像素值置成(255,255,255),重建关于细胞核的 HE 染色图像。HE 染色的基础是组织结构对不同染料的不同结合程度不同。苏木精可以将嗜碱性结构染成蓝紫色,而伊红可以将嗜酸性结构染成粉红色。嗜碱性结构通常包括含有核酸的部分,如核糖体、细胞核及细胞质中富含核糖核酸(RNA)的区域等。嗜酸性结构则通常包含细胞内及细胞间的蛋白质,如路易体(Lewy body)、酒精小体(Mallory body)、细胞质的大部分等。

针对一次、二次、三次聚类获得的细胞核的图像进行颜色反卷积的结果如图 5 所示。

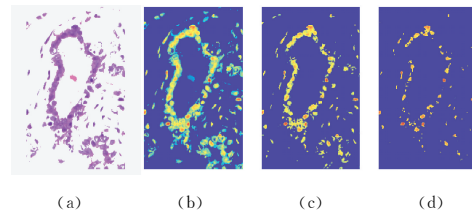


图 5 细胞核(颜色反卷积后的苏木精层)

Fig. 5 Nuclei(hematoxylin of color deconvolution)

从图 5(a)中可以看出,细胞核之间粘连严重,且细胞核的颜色深浅不一。

根据 HE 染色的特点,使用颜色反卷积算法提取染色图像的 H 通道。从图 5(b)中可以更明显地看出,亮黄色的细胞核周围包裹着一圈蓝绿色的区域。由此得出,第一次聚类的结果并不精准,需要进行二次聚类。

在重建的细胞核图像上再次进行聚类,把图像中的像素点聚成 3 类:背景、杂质、细胞核;然后同样进行重建和颜色反卷积。从图 5(c)中可以看出,细胞核之间的粘连程度有所下降。

更进一步地又进行一次聚类。对比图 5(c)和图 5(d)发现,后者细胞核之间的分离程度更高,但同时细胞核也更琐碎,不利于后续的形态学操作。因此选择图 5(c),即将进行过两次聚类的图像作为后续识别的基础。

图 5(c)中的细胞核之间仍然存在粘连,有的细胞核上还存在孔洞。同样地,需要在图 5(c)的基础上进行形态学操作,以消除细胞核之间的粘连并填充孔洞。

细胞核的识别步骤如下:

1)对图像进行 clear_border 操作。由于图中的细胞核已经存在粘连,因此不能在图像上执行形态学闭操作,因为形态学闭操作会在一定程度上连接邻近物体,导致细胞核的粘连程度进一步加深。

2)进行填洞操作,填充细胞核上的孔洞。

3)在上述结果的基础上进行形态学开操作,以解决细胞核之间的粘连问题。

4)基于上述结果进行识别操作,提取每个细胞核的中心坐标用于后续的腺管识别。

5 邻区构造与特征提取

识别出管腔和细胞核后,得到一个拥有所有管腔的集合,其中每个管腔使用圆度、周长、中心坐标、面积以及轮廓坐标来描述;对一个拥有所有细胞核的集合中的每个元素,使用中心坐标进行描述。由于腺管由管腔和环绕在管腔周围的细胞核构成,因此需要基于这两个集合构造腺管候选区域,然后在每个候选区域上提取特征用于腺管识别。

5.1 邻区构造

邻区被定义为一个管腔以及管腔周围环绕的一圈细胞核构成的区域。用 L 表示管腔集合,用 N 表示细胞核的集合。对于 L 中的每个管腔 l ,使用 l_c 表示管腔 l 的轮廓坐标集合。

遍历 l_c 中的每个轮廓坐标,在遍历过程中,以轮廓坐标 c 为圆心、 r 为半径画圆 C ,然后检查细胞核集合 N 中的每个元素,其中位于圆 C 内部的细胞核均被纳入到管腔 l 的邻区中。

构造完邻区之后,因此得到一个邻区集合,这个集合的元素个数与管腔集合 L 相同,其中每个元素包括在构造邻区的过程中获得的邻区细胞核坐标和管腔分割过程中获得的相关属性。

由于在遍历管腔的每个轮廓坐标的同时需要遍历细胞核的集合,因此整个过程耗时较长。在不影响邻区构造效果的前提下,对轮廓坐标进行等距采样:每隔 3 个轮廓坐标采集一次细胞核。这样既保持了精度,又减少了时间消耗。

5.2 特征提取

本文采用如下 10 个特征来识别腺管。

1) 邻区细胞核的个数与邻区管腔周长的比值

真正的腺管管腔周围环绕着一圈均匀分布的细胞核,而脂肪、气泡等类似管腔的物体周围则没有,因此使用细胞核的个数来区分真正的腺管和脂肪、气泡等物体。同时由于腺管大小不一,因此使用细胞个数除以管腔周长来进行归一化处理。

2) 邻区管腔的圆度

因为乳腺腺管通常是圆形或者椭圆形,且腺管的圆度可以由邻区管腔的圆度近似替代,所以使用邻区管腔的圆度作为特征之一来区分腺管与非腺管。

3) 邻区管腔中心到邻区细胞核的距离的统计特征

为了衡量邻区细胞核关于管腔中心的分布是否均匀,计算管腔中心到其邻区细胞核的欧氏距离:

$$D(o^l) = \{d(o^l, o_i^n) : \forall i \in \{1, 2, 3, \dots, N\}\} \quad (1)$$

使用其均值、标准差、最大值、最小值这 4 个统计量作为特征。其中, o^l 表示管腔中心坐标, o_i^n 表示第 i 个邻区细胞核的中心坐标。

4) 相邻细胞核之间角度的统计特征

腺管的一个重要特征是细胞核以均匀的间隔分布在管腔周围,可以利用相邻细胞核之间的夹角来量化这个特征。使用 $\vec{o^l o_i^n}$ 表示管腔中心 o^l 到细胞核 o_i^n 的向量,使用 $\vec{o_i^n o_j^n}$ 表示管腔中心 o^l 到细胞核 o_j^n 的向量。相邻细胞核之间夹角的集合可以表示如下:

$$A = \left\{ \arccos \left(\frac{\vec{o^l o_i^n} \cdot \vec{o^l o_j^n}}{\| \vec{o^l o_i^n} \| \cdot \| \vec{o^l o_j^n} \|} \right) : \forall i, j \in \{1, 2, 3, \dots, N\}, i \neq j \right\} \quad (2)$$

使用其均值、标准差这两个统计量作为特征。

5) 相邻细胞核之间距离的统计特征

另一种用于量化细胞核以均匀的间隔分布在管腔周围的指标是相邻细胞核之间的距离。相邻细胞核之间的距离集合可以表示如下:

$$D = \{ \| o_i^n - o_j^n \| : \forall i, j \in \{1, 2, 3, \dots, N\}, i \neq j \} \quad (3)$$

使用其均值、标准差这两个统计量作为特征。

6 基于随机森林的分类模型

随机森林分类算法属于集成学习。集成学习的一般结构为:先产生一组个体学习器,再用某种策略将它们结合起来。个体学习器一般由一个现有的学习算法从训练数据中产生,比如 C4.5 决策树算法。集成学习可获得比单一学习器显著优越的泛化性能。为了获得比单一学习器更好的性能,集成学习的个体学习器需要具有准确性和差异性。集成学习研究的核心是如何产生“好而不同”的个体学习器。根据个体学习器生成方式的不同,集成学习分为串行生成 (Boosting) 和并行生成两种 (Bagging 和随机森林)。

Bagging 直接基于自助采样法 (Bootstrap Sampling), 给定包含 m 个样本的数据集,先随机取出一个样本放入采样集中,再把该样本放回初始数据集,使得下次采样时该样本仍有可能被选中,这样经过 m 次随机采样操作后会得到含 m 个样本的采样集,初始训练集中约有 63.2% 的样本出现在采样集中。可采样出 T 个含 m 个训练样本的采样集,然后基于每个采样集训练出一个基学习器,再将这些基学习器进行结合 (投票法、平均法)。

随机森林分类算法是 Bagging 的一个扩展变体,它在以决策树为基学习器构建 Bagging 集成的基础上进一步在决策树的训练过程中引入了随机属性选择。具体来说,传统的决策树在选择划分属性时,是在当前结点的属性集合中 (假定有 d 个属性) 基于信息增益或者信息增益率选择一个最优属性;而在随机森林中,对于决策树的每个结点,先从该结点的属性集合中随机选择一个包含 k 个属性的子集,再从这个子集中选择一个最优属性用于划分。这里的参数 k 控制了随机性的引入程度;若令 $k=d$,则基决策树的构建与传统决策树相同;若令 $k=1$,则随机选择一个属性用于划分;一般情况下,推荐 $k = \log_2 d$ 。因此,随机森林分类算法同时拥有样本扰动和属性扰动两个优势。

在本文中 $d=10$, 因此 k 取 4。在利用投票法作出最终决策之前,需要确定子树的棵数,即使用几棵子决策树来集成随机森林分类模型。较多的子树可以使模型具有更好的性能,但耗时也更长,因此选择在选择处理器能承受的范围内最大的子树棵数,本文的随机森林分类算法包含 10 棵子树。

7 实验及结果讨论

Ajay 等人^[7]和 Maqlin 等人^[8]提出的方法与本文方法的思路相似。不同的是, Ajay 等人^[7]使用颜色反卷积算法以及

形态学开操作进行细胞核识别;使用 HNCut 及 CGAC 模型进行管腔识别;使用 O'Callaghan 近邻算法构建候选腺管。Maqlin 等人^[8]首先找出管腔附近的细胞核团,然后利用网格分析和水平集方法进一步分割细胞核,最后利用管腔周围细胞核分布的均匀程度筛选出真正的腺管。因此,本文重新实现了 Ajay 等人^[7]提出的细胞核识别、管腔识别和候选腺管构建方法,以及 Maqlin 等人^[8]提出的细胞核识别和腺管筛选方法。然后对上述两种方法采用与本文相同的随机森林模型预测真正的腺管。O'Callaghan 近邻算法通过距离和角度两种限制条件来构建候选腺管,但是由于管腔大小不一且管腔并不是一个规则的圆形,因此该算法中的距离限制不能固定,对比我们使用与当前管腔形态近似的椭圆的半长轴与半短轴的均值进行近似。为了比较的公平性,在相同的数据集上进行了对比实验,评估方法均采用 10 折交叉验证方法。数据来源于本实验室采集的 46 张乳腺组织病理图像,所有图像均为 HE 染色图像,放大倍数为 1040,图像大小为 4140 * 3096。

针对每张图像,采用本文方法获得 3456 个候选腺管,并用上文提出的 10 维特征向量进行描述。经过人工标注,其中正样本有 234 例,占 6.77%;负样本有 3222 例。采用 Ajay 等人^[7]提出的方法获得 2672 个候选腺管,每个候选腺管使用原文提出的 22 维特征向量进行描述。经过人工标注,其中正样本 206 例,占 7.71%;负样本 2466 例。采用 Maqlin 等人^[8]提出的方法获得 3275 个候选腺管,每个腺管使用表示管腔周围相邻细胞核之间距离的统计特征的 2 维特征向量进行描述。经

过人工标注后,正样本有 228 例,占 6.96%;负样本有 3047 例。

7.1 特征分析

本文提出的 10 个特征在不同类中的统计结果如图 6 所示。从图 6 中可以看出,本文选用的 10 个特征在正、负类的样本中具有较大的差异。其中角度标准差由于绝对数值较小,在图中的差异不明显。除此之外,其余特征的最小差异约 20%,最大差异约 50%。由此可见,本文提出的 10 个特征在正、负类样本之间具有较好的区分度。

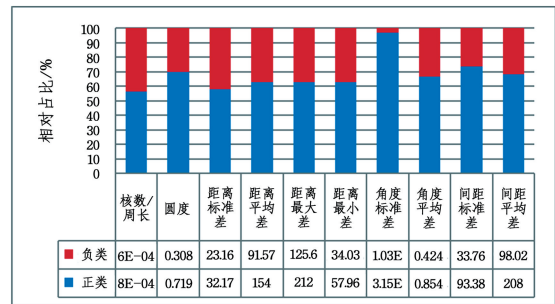


图 6 正类与负类之间的特征值比较

Fig. 6 Comparison of eigenvalues between positive class and negative class

7.2 分类精度

由于负样本远远多于正样本,因此在训练和评估随机森林模型时,分别抽取不同比例的负样本数据进行实验。不同比例的负样本得到的 10 折交叉验证实验的对比结果如表 1 所列。

表 1 3 种方法的对比分类结果

Tabel 1 Comparison results of three methods

(单位:%)

方法	1:1			1:2			1:3		
	pos	neg	tol	pos	neg	tol	pos	neg	tol
本文方法	94.35	92.61	92.48	92.17	95.22	94.20	86.52	96.29	93.87
Ajay 等人的方法 ^[7]	84.49	82.67	82.05	82.31	84.47	83.47	80.19	86.58	84.75
Maqlin 等人的方法 ^[8]	87.93	85.44	85.20	85.43	87.58	87.39	82.67	90.39	87.73

注:pos 表示正样本的分类精度,neg 表示负样本的分类精度,tol 表示所有样本的总体分类精度

从表 1 中的数据看出,本文提出的基于二次聚类和随机森林的乳腺腺管自动识别方法的准确率可以达到 86% 以上;但是随着负样本的比例逐渐增加,正负样本比例的不平衡度增大,虽然整体的准确率有所提高,但对正样本的分类准确率有所降低。因此,在实际应用中采用数量均衡的正负样本训练模型较为合适。

从候选腺管个数来看,相比 Ajay 等人^[7]提出的方法,本文提出的方法能筛选出更多的候选腺管,尽管正样本占比略低于后者,但本文方法尽可能不遗漏或少遗漏真正的腺管。分类准确率对比结果显示,本文提出的方法更优,说明本文提出的 10 个特征比 Ajay 等人^[7]选用的 22 个特征具有更好的区分能力。

在候选腺管和正样本相对占比方面,Maqlin 等人^[8]提出的方法与本文提出的方法基本持平。分类准确率对比结果显示,本文提出的方法更优,可能的原因是 Maqlin 等人^[8]提出的方法中用于筛选真正的管腔的条件过于简单。

结束语 本文根据乳腺 HE 染色组织病理图像的特点,提出了一种在乳腺组织病理图像中进行自动腺管识别的方法。采用聚类算法、分割算法、形态学操作和分类算法进行腺管的自动分割以及识别。首先,由于在 HE 染色组织病理图像中不同的组织结构具有不同的颜色,因此根据像素值通过一次聚类、二次聚类构建出待分割图像(细胞核和管腔);然后,对其进行形态学操作,并在此基础上进行分割;最后,构建候选腺管,利用中心管腔与管腔周围细胞核的空间位置关系以及一些其他特征来识别出真正的腺管。

本文提出的自动腺管识别的方法在乳腺组织病理图像中含较多明显腺管的情况下可以获得很好的准确率及稳定性;对已经中度分化或者已经完全分化的腺管不能很好地进行识别,原因在于上述两种腺管的形态已经发生变化,不再符合腺管管腔周围环绕着一圈均匀分布的细胞核这一描述。后续将从以下几个方面进行研究:1) 进一步提高分割准确率;2) 提出其他特征来描述候选腺管;3) 研究如何在乳腺组织病理图像

中识别中度分化和完全分化的腺管。目前正在尝试首先对 HE 图像进行三分类(由于第三类腺管形态完全消失,因此舍弃这些图片),然后通过细胞质环对第一、二类腺管形态进行腺管识别。

参 考 文 献

- [1] LATSON L, SEBEK B, POWELL K A. Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy[J]. *Analytical and Quantitative Cytology and Histology/the International Academy of Cytology and American Society of Cytology*, 2003, 25(6): 321-331.
- [2] PETUSHI S, GARCIA F U, HABER M M, et al. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer[J]. *BMC Medical Imaging*, 2006, 6(1): 1-11.
- [3] BAMFORD P, LOVELL B. Unsupervised cell nucleus segmentation with active contours[J]. *Signal Processing*, 1998, 71(2): 203-213.
- [4] XU J, JANOWCZYK A, CHANDRAN S, et al. A weighted mean shift, normalized cuts initialized color gradient based geodesic active contour model: Applications to histopathology image segmentation[C]// *Medical Imaging: Image Processing*. 2010.
- [5] SHI J, MALIK J. Normalized cuts and image segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22(8): 888-905.
- [6] CHAN T F, VESE L A. Active contours without edges[J]. *IEEE Transactions on image processing*, 2001, 10(2): 266-277.
- [7] BASAVANHALLY A, YU E, XU J, et al. Incorporating domain knowledge for tubule detection in breast histopathology using O'Callaghan neighborhoods[C]// *Proceedings of SPIE—The International Society for Optics and Photonics*. 2011.
- [8] MAQLIN P, THAMBURAJ R, MAMMEN J J, et al. Automatic detection of tubules in breast histopathological images[M]// *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*. India: Springer, 2013: 311-321.
- [9] NGUYEN K, BARNES M, SRINIVAS C, et al. Automatic glandular and tubule region segmentation in histological grading of breast cancer[C]// *SPIE Medical Imaging*. International Society for Optics and Photonics. 2015.
- [10] SIRINUKUNWATTANA K, PLUIM J P W, CHEN H, et al. Gland segmentation in colon histology images: The glas challenge contest[J]. *Medical Image Analysis*, 2016, 35: 489-502.
- [11] JANOWCZYK A, MADABHUSHI A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases[J]. *Journal of Pathology Informatics*, 2016, 7(1): 29.
- [12] PENG Y, JIANG Y, EISENGART L, et al. Segmentation of prostatic glands in histology images[C]// *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2011: 2091-2094.
- [13] NGUYEN K, SARKAR A, JAIN A K. Structure and context in prostatic gland segmentation and classification[M]// *Medical Image Computing and Computer-Assisted Intervention*. Springer Berlin Heidelberg, 2012: 115-123.
- (上接第 222 页)
- [10] WAN S, LAN Y, GUO J, et al. Informational friend recommendation in social media[C]// *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'13)*. ACM, 2013: 1045-1048.
- [11] CHEN C C, SHIH S Y, MENG L, et al. Who should you follow? Combining learning to rank with social influence for informative friend recommendation[J]. *Decision Support Systems*, 2016, 90: 33-45.
- [12] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. *Journal of the American Society for Information Science and Technology*, 2007, 58(7): 1019-1031.
- [13] JAMALI M, ESTER M. Trustwalker: a random walk model for combining trust-based and item-based recommendation[C]// *ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*. 2009: 397-406.
- [14] LESKOVEC J, HUTTENLOCHER D, KLEINBERG J. Predicting positive and negative links in online social networks[C]// *Proceedings of the 19th International Conference on World Wide Web(WWW'10)*. ACM, 2010: 641-650.
- [15] DENG S G, HUANG L T, XU G H. Social network-based service recommendation with trust enhancement[J]. *Expert Systems with Applications*, 2014, 41(18): 8075-8084.
- [16] SARUKKAI R R. Link prediction and path analysis using Markov chains[J]. *Computer Networks*, 2000, 33(1-6): 377-386.
- [17] MASSA P, AVESANI P. Trust-aware recommender systems[C]// *Proceedings of the 2007 ACM Conference on Recommender Systems*. 2007: 17-24.
- [18] ARMENTANO M G, GODOY D, AMANDI A. Topology-based recommendation of users in micro-blogging communities[J]. *Journal of Computer Science and Technology*, 2012, 27(3): 624-634.
- [19] WANG J, DE VRIES A P, REINDERS M J. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C]// *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2006: 501-508.
- [20] CHEN K, CHEN T, ZHENG G, et al. Collaborative personalized tweet recommendation[C]// *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'12)*. ACM, 2012: 661-670.