粗糙神经智能疑似乳癌图像分类方法研究

杨冬风[1],杨冬秀[2]
YANG Dong-feng[1],YANG Dong-xiu[2]

1.黑龙江八一农垦大学 信息技术学院,黑龙江 大庆 163319
2.大庆油田总医院 放射科,黑龙江 大庆 163411
1.Department of Information Technology,Heilongjiang Bayi Agriculture University,Daqing,Heilongjiang 163319,China
2.Department of Radiology,Daqing Oilfield Head Hospital,Daqing,Heilongjiang 163411,China
E-mail:yangsansun@sina.com

YANG Dong-feng,YANG Dong-xiu.Rough neural intelligent approach of mammogram image classification with suspected breast cancer.Computer Engineering and Applications,2010,46(12):188-191.

Abstract: A rough neural intelligent approach for rule generation and image classification is introduced.This method is a hybridization of intelligent computing techniques.Algorithms based on fuzzy image processing are first applied to enhance the contrast of the whole original image,to extract the region of interest and to enhance the edges surrounding that region.Then,this paper extracts features characterizing the underlying texture of the regions of interest by using the gray-level co-occurrence matrix.Then,the rough set approach to attribute reduction and rule generation is presented.Finally,rough neural network is designed for discrimination of different regions of interest to test whether they represent malignant cancer or benign cancer.To evaluate performance of the presented rough neural approach,this paper runs tests over different mammogram images.The experimental results show that the overall classification accuracy offered by rough neural approach is high compared with other intelligent techniques.
Key words: rough sets;neural networks;decision trees;fuzzy sets

摘 要:提出了一种用于乳腺X线图像分类的粗糙神经智能方法,该方法是一种混合智能计算技术。首先使用模糊图像处理算法来提高整个原始图像的对比度以提取感兴趣区域以及增强区域边缘;然后建立灰度共生矩阵,提取出表征感兴趣区域纹理的特征属性;接着使用粗糙集方法进行属性约简并产生规则;最后,设计出粗糙神经网络,用来将感兴趣区域区分为良性或是恶性。为了对所提出的粗糙集神经网络进行性能评价,对若干乳腺X线图像样本进行了测试,实验结果表明:用该方法进行乳癌识别的整体准确率要高于使用其他技术。
关键词:粗糙集;神经网络;决策树;模糊集
DOI:10.3778/j.issn.1002-8331.2010.12.056 **文章编号**:1002-8331(2010)12-0188-04 **文献标识码**:A **中图分类号**:TP391

1 引言

乳腺癌是女性最常见的恶性肿瘤之一。据统计,全球每年约有120余万妇女患乳腺癌,50万妇女死于乳腺癌。在美国,每年诊断出18.2万乳癌患者,其中有4.6万死亡。在我国,乳腺癌的发病率目前正以每年3%~4%的增长率急剧上升,每年约有18万妇女患乳腺癌,1.3万多妇女死于乳腺癌[1]。

目前,用于乳腺癌检测的钼钯射线摄影(mammography)(通常所说的X光照片)是临床上乳腺癌检测的最主要手段,这是因为乳腺X线图像的空间分辨率与灰度分辨率都很高,乳腺癌的两种主要病灶肿块与微钙化都比较明显。但是即使如此,乳腺癌的误诊率和漏诊率仍然无法得到较大的降低。据资料显示,只有70%~85%的乳腺癌病例能通过其乳腺X线图像被放射学家一次性检出;在剩下的病例中,又只有2/3能在二次检验中被检出。同时,乳腺X线图像的误诊率也并不低,约为4%~10%[2]。

这些主要是由于恶性病变的良性表象、征象与病灶过于细小、阅读图像的放射学家的视觉疲劳和疏忽等原因造成的。所以需要计算机辅助诊断技术来帮助他们对乳腺X线图像进行预检测。王曙燕等人研究模糊聚类分析在医学图像数据挖掘中的应用,利用决策树算法对乳腺癌图像数据进行分类[2];Antonie用神经网络和关联规则数据挖掘方法对乳腺X线图像进行分类,他们用神经网络方法能达到81.25%的分类精确度[3];A.E.Hassanien使用决策树对乳腺X线图像进行特征选择和分类[4]。

该文把模糊集、神经网络和粗糙集这三种技术集成起来,将粗糙集理论中的约简原理与神经网络方法相结合,减少了特征属性从而降低了数据训练时间,同时也避免了单独使用粗糙集分类的过度约简问题。

2 预处理阶段

在这个阶段,使用模糊图像处理技术增强整个图像的对比

作者简介:杨冬风(1977-),女,讲师,主要研究领域为图像处理,虚拟现实;杨冬秀(1975-),女,医师,主要研究领域为乳腺疾病诊断。
收稿日期:2009-10-30 **修回日期**:2010-01-05

度;提取出感兴趣的区域;增强感兴趣区域边缘。使用改良的标准模糊 C-mean 聚类算法初始化分割,它能加速操作并处理原始图像对噪声的敏感性。

2.1 模糊直方图均衡化算法(FHH)

2.1.1 基于不可分辨关系的子图划分

定义条件属性集 $C=\{c_1,c_2\}$,其中 c_1 是像素灰度值属性,c_2 是噪声属性。X 射线图像一般是由较亮区域和较暗区域组成,则它的直方图有两个峰,一个峰对应于亮区灰度值,一个峰对应于暗区灰度值,两峰之间选一个灰度值作阈值 P。灰度值属性 $c_1=\{0,1\}$,其中 0 代表 0~P 灰度值,1 代表(P+1)~255 灰度值,噪声属性 $c_2=\{0,1\}$,其中 0 代表 2×2,或者 4×4 像素组成子块 s 的平均灰度值与相邻子块平均灰度值之差的绝对值均小于某一阈值 Q,1 代表子块的差值绝对值均大于 Q。

(1)根据 c_1 划分子图

设 x 代表"较亮"的像素,等价关系 R_{c_1} 定义为:如果两个像素的灰度值都大于某个阈值 P,则两个像素是 R_{c_1} 相关的,即属于等价类,用公式表达:

$$R_{c_1}(x)|\{x:f(x)>P\}$$

$f(x)$ 表示像素 x 的灰度值,R_{c_1} 表示所有"较亮"的像素 x 组成的集合。R_{c_1} 的非集 $\overline{R_{c_1}}$ 则表示所有"较暗"的像素 x 组成的集合。

(2)根据 c_2 划分子图

定义等价关系为 R_{c_2}:子块 s_{ij} 与相邻子块的平均灰度值 $m(s)$ 之差的绝对值取整均大于某一阈值 Q,即:

$$R_{c_2}(s)=\bigcup_i\bigcup_j\{s_{ij}|:\mathrm{int}|m(s_{ij})-m(s_{i\pm1,j\pm1})|>Q,$$

$s_{i\pm1,j\pm1}$ 表示 s_{ij} 相邻的子块}

$R_{c_2}(s)$ 表示所有噪声像素组成的集合,子块 s_{ij} 与相邻子块 $s_{i\pm1,j\pm1}$ 构成宏块。令:

$$A_1=R_{c_1}(x)-R_{c_2}(s),A_2=\overline{R_{c_1}}(x)-R_{c_2}(s)$$

A_1 表示剔除噪声后所有"较亮"的像素 x 组成的集合。A_2 表示剔除噪声后所有"较暗"的像素组成的集合。

2.1.2 做增强变换

(1)补全子图 A_1,即在所有"较暗"的像素和噪声像素位置处,分别用阈值 P 灰度值和噪声子块处的宏块平均灰度值填充,构成 B_1。

(2)将子图 A_2 补全,即在所有"较亮"的像素和噪声像素位置处,分别用阈值 P 和宏块均值填充,构成 B_2。

(3)对 B_1 作直方图均衡变换,作 B_2 直方图指数变换。

(4)将变换后的图像作重叠,输出增强的图像。

2.2 改良的模糊 C-mean(M-FCM)聚类算法

大多数良胜肿瘤的边缘比较清晰和比较规则,而恶性肿瘤的边缘则模糊不清和不规则。然而,有些良性乳腺瘤如纤维性瘤和胞囊肿块的边界也是模糊不清的。因此肿瘤区域的分割是进一步精确分类良性肿瘤和恶性肿瘤的关键。使用标准的模糊 C-mean(S-FCM)聚类算法进行分割,方法如下[5]:

设 M 维数据空间中的有限样本数据集 $Y=\{y_1,y_2,\cdots,y_n\}$,n 为数据集中的元素个数,$y_k(k=1,2,\cdots,n)$ 为样本点。若 FCM 将样本数据聚类 c 个分类($2\leqslant c\leqslant n$),聚类中心 $P=\{p_1,p_2,\cdots,p_n\}$。聚类的分类矩阵为 $\boldsymbol{U}$,U_{ik} 是其元素,是样本 y_i 对聚类中心 p_k 的隶属度,加权指数为 $m(m\in[1,\infty))$。FCM 算法的目标函数描述如下[6]:

$$J=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m\|y_k-p_i\|^2$$

当较高的成员值被赋予强度接近于特定类的质心的像素时,并且较低的成员值被赋予强度与质心相差较大的像素时,目标函数被最小化了;上述算法计算费时并且易受噪声干扰。为了解决这些问题,对 FCM 算法进行改良[5]。

输入:增强的乳腺 X 线图像 I

设置质心 c 的初始值

输出:聚类区域

1:For i=1 to c do

2: 更新分割矩阵{成员计算}

3: $D_{ik}=\|y_k-\lambda_i*p_k\|^2$;

4: $\phi_i=\sum_{y_r\in N_k}\|y_k-\lambda_r*p_i\|^2$;

5: $p_{ik}^{*}=\dfrac{1}{\sum_{j=1}^{c}\left(\dfrac{D_{ik}+\dfrac{\alpha}{N_R}\phi_i}{D_{jk}+\dfrac{\alpha}{N_R}\phi\delta_j}\right)^{\frac{1}{P-1}}}$;

6:end for

N_R 表示 x_k 周围窗口中存在的邻集的基数;

N_R 的影响是由参数 α 控制的,α 取较高的值较好;

7:更新质心

8:估算 $p_i^{*}=\dfrac{\sum_{k=1}^{N}u_{ik}^{P}\left[(y_k-\lambda_k)+\dfrac{\alpha}{N_R}\sum_{y_r\in N_k}(y_r-\lambda_r)\right]}{(1+\alpha)\sum_{k=1}^{N}u_{ik}^{P}}$;

9:估算新的增益 $\lambda_j=\dfrac{\sum_{i=1}^{c}p_{ik}^{P}y_jv_i}{\sum_{i=1}^{c}p_{ik}^{P}}$

10:repeat 2

11:until $\|P_{new}-P_{old}\|<\varepsilon$;{$\varepsilon$ 是用户设置的极限值,V 是聚类中心的一个矢量}

3 特征提取阶段

乳腺 X 线图像包含了很多信息:各种组织、腺体、导管、乳腺边缘、肿块、钙化点等。表 1 给出标志乳癌的 X 线图像的形态特征。

表 1 乳腺癌钼钯射线图像特征

主要病灶	形状	主要病灶	形状
肿块	团块状	钙化	圆形
	星形尖刺状		卵圆形
	云片状		不规则多角形
	半球形		线形
	彗星形		分叉形
	弥漫结节		

经过预处理阶段,肿块和钙化点都被分割出来。但是,在分割过程中与肿块类似的假阳性区域,以及与微钙化类似的噪声点、划痕和导管这些膺像很容易在检测过程中被混淆。为了区分真正病灶与赝像,选取的特征如表 2 所示。

除了上述特征,还使用灰度共生矩阵(GLCM)来分析与二次特征相关的图像特征。为了降低计算复杂度,选择 Haralick[6] 提出 14 种统计属性中的几个属性:能(角二阶矩)、熵、对比和反差矩。

表 2 选取的特征

微钙化		肿块	
噪声	面积	假阳性区域	面积
	平均灰度		平均灰度
	对比度		对比度
	成簇性		灰度一致性
划痕	长宽比		分形维数
	类圆性		类圆性
导管	灰度一致性		长宽比
	边界凹点		边缘均方差
	比例空洞数目		边缘平均梯度
	开运算腐蚀量		方向熵

“能”,是对图像纹理一致性的度量。当灰度分布是连续的或呈周期分布时,“能“达到最大值。均质(均匀)的图像只包含几个主要的灰度跃迁,因此,图像的共生矩阵的规格化入口只有几个较大的入口,使得“能”的值也很大。相反,如果共生矩阵的规格化入口包含大量的小入口,那么“能”的属性值就较小。

“熵”度量图像的无序性。当规格化共生矩阵的入口元素都相等时,熵的值最大。当图像在组织上不一致,许多灰度共生矩阵值很小,熵会很大。因此,熵与灰度矩阵的“能“成反比。

“对比”是灰度共生矩阵的规格化入口的差方矩。它表示图像的局部差异大小。

“反差矩”标志图像的同质性。当多数共生集中在主对角线附近时,它达到最大值。反差矩与灰度共生矩阵的对比成反比。

4 粗糙集数据分析阶段

粗糙集数据分析包含下面步骤:

(1)离散图像属性并产生决策表;

(2)产生具有最小数目属性的约简;

(3)属性的重要程度:计算属性的权重;

(4)规则产生:产生一个规则表;

(5)计算规则精确性。

从决策表中计算出核心和约简是选择相关属性的一种方法[7]。从结果约简表示属性最小集的意义上说,这是一种全局方法。但要确保属性集中的属性要与原来整套属性在分类上具有相同的能力。选择相关属性的直接方法是给每个属性在相关性上的一个度量,并选择具有较高值的属性。基于约简系统,产生了规则列表,它将用作构建新对象的分类模型。

下面是规则产生算法:

输入:决策系统(U,C,D)

属性约简 $R\subseteq C$;$R=\{a_1,a_2,\cdots,a_m\}$;$m=|R|$

输出:决策规则集 RULES(R)

1:for $u\in U$

2: for $a_i\in R$

3: $v_i=a_i(u)$;

4: end for

5: $v_d=d(u)$;

6:$RULES(R)=RULES(R)\cup\{a_1=v_1\wedge a_2=v_2\wedge\cdots\wedge a_m=v_m\rightarrow d=v_d\}$;

7:end for

8:Return RULES(R);

5 粗糙神经分类器的设计

粗糙神经网络包括一个输入层、一个输出层和一个隐含层。输入层神经元由外界接受输入;输入神经元的输出作为隐含神经元的供给。隐含层的输入又作为输出层神经元的输入;最好输出到外界。隐含层神经元的数目是由下面的不等式决定的[8]:

$$N_{hn}\leqslant\frac{N_{ts}*T_e*N_f}{N_f+N_o}$$

N_{hn} 是隐含层神经元的数目,N_{ts} 是训练样本的数目,T_e 是公差,N_f 是属性的数目,N_o 是输出的数目。粗糙神经元的输出具有上限和下限,而传统的神经元的输出是一个单一的值。粗糙神经元是 Lingras 在 1996 年提出的,粗糙神经元是相对于上限(U_n)和下限(L_n)定义的,并且输入也是对于边界值的相对量。粗糙神经元有三种连接方法:

(1)输入–输出连接到 U_n;

(2)输入–输出连接到 L_n;

(3)连接到 U_n 和 L_n 之间。

设粗糙神经元 $R_n=(U_n,L_n)$,其中 U_n 和 L_n 分别是粗糙神经元上限和粗糙神经元下限。(Ir_{Ln},Or_{Ln})为输入/输出神经元下限,(Ir_{Un},Or_{Un})为输入/输出神经元上限。计算公式如下:

$$Ir_{Ln}=\sum_{j=1}^{n}w_{Lnj}Or_j \qquad Ir_{Un}=\sum_{j=1}^{n}w_{Unj}Or_j$$

$$Or_{Ln}=\min(f(Ir_{Ln}),f(Ir_{Un})) \qquad Or_{Un}=\max(f(Ir_{Ln}),f(Ir_{Un}))$$

粗糙神经元(Or_n)的输出计算如下:

$$Or_n=\frac{Or_{Un}-Or_{Ln}}{average(Or_{Un},Or_{Ln})}$$

分类算法如下:

输入:要分类的图像;特征集(属性);神经元输入;规则集

输出:最终的分类结果

1:for 属性集中的每个属性

2: 计算上、下粗糙神经元;

3:end for

4:建立粗糙集神经网络;

5:计算相对误差;

6:校准粗糙集神经网络;

7:repeat 4,5,6;

8:until 分类误差小于某个极小值

9:return 分类结果

6 结果与讨论

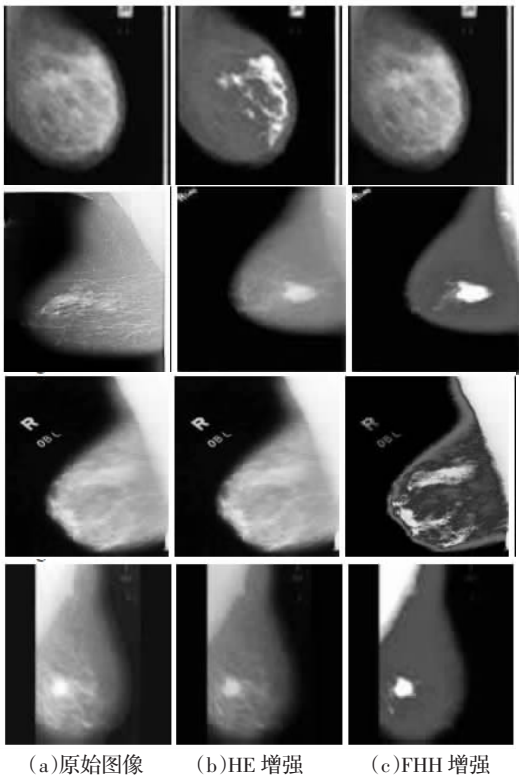
6.1 增强阶段结果评估

为了评估算法的视觉效果, 从某医院的 X 射线图像分析数据库中选择了一些图像并进行了处理(如图 1 所示)。其中:(a)是原始图像,(b)是直方图均衡化效果,(c)是模糊直方图增强效果。

表 3 给出了模糊参数(H)及熵(γ)的值在原始图像、HE 增强、FHH 增强的值。

表 3 模糊参数(H)及熵(γ)的值

样本	H			γ		
I	O	HE	FHH	O	HE	FHH
I1	0.289	0.278	0.210	0.217	0.198	0.009
I2	0.021	0.018	0.102	0.003	0.002	–0.005
I3	0.423	0.396	0.289	0.231	0.240	0.128
I4	0.053	0.036	0.018	0.410	0.390	0.210
I5	0.501	0.463	0.496	0.411	0.263	0.376
I6	0.313	0.118	0.002	0.232	0.198	0.189
I7	0.420	0.391	0.272	0.211	0.190	0.007
I8	0.344	0.301	0.276	0.301	0.285	0.211



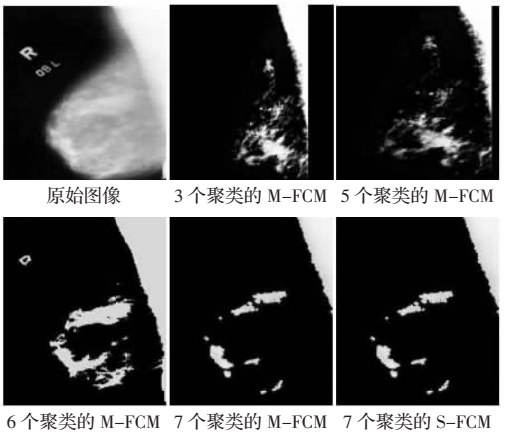
(a)原始图像 (b)HE 增强 (c)FHH 增强

图 1 图像增强效果对比

有效的图像增强技术要降低目标的熵值,并且要降低灰度范围。从实验结果中,也可看出图像增强之后,模糊参数和熵都降低了。

6.2 分割结果分析

图 2 表示 S-FCM 和 M-FCM 的带有不同初始参数的视觉聚类结果。价值函数的参数权值从 0.001 到 0.000 000 1,并且聚类数目从 3 到 8。



原始图像 3 个聚类的 M-FCM 5 个聚类的 M-FCM

6 个聚类的 M-FCM 7 个聚类的 M-FCM 7 个聚类的 S-FCM

图 2 不同聚类数目的 FCM 图像分割结果

从得到的结果看,M-FCM 和 S-FCM 算法都获得了较好的结果,具有较高的聚类数目和较小的价值函数的权重。算法的平均分割准确率大约 3.983 7%,这说明算法的鲁棒性很强。S-FCM 和改良的 M-FCM 算法都在 Matlab 中执行。

6.3 产生规则的数目和整体精度

表 4 显示了分别使用决策树、神经网络和粗糙神经网络这三种方法所产生规则的数目的对比。可以看到,所有算法产生的规则数目都很大,并且比对象的数目要多,因此分类速度很慢。

表 4 不同方法产生规则数目对比

使用方法	产生规则数目
决策树	1 393
神经网络	416
粗糙集神经网络	154

因此,有必要在规则产生时修改规则。图 3 显示了关于敏感度和专业性的分类精度,由图可以看出粗糙神经分类方法的精度要高于普通神经网络、粗糙集和决策树。而且,对于神经网络和决策树分类器,需要更多的鲁棒特性以提高性能。

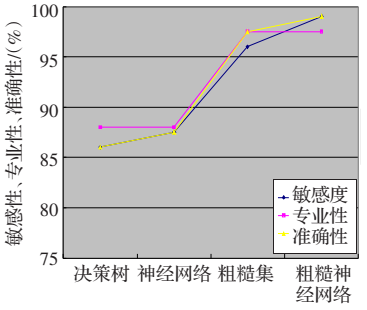


图 3 粗糙神经分类算法精度与决策树、神经网络和粗糙集的对比

7 结束语

讨论了将模糊理论、粗糙集和神经网络的优势结合起来形成融合的智能分类方法。使用 FHH 算法对图像增强的效果要优于传统的 HE 算法;使用改良的 M-FCM 算法对图像分割的效果要优于标准 S-FCM 方法;使用该文算法产生规则的数目为 154,要小于使用决策树和神经网络。对若干乳腺 X 射线样本进行测试的结果表明该方法的监测精度在 98%以上。

参考文献:

[1] 杨蕾,秦占芬,蒋湘宁.雌激素致乳腺癌机制的研究进展[J].肿瘤防治研究,2006,33(11):840-842.

[2] 王曙燕,周明全,耿国华.模糊聚类分析在乳腺癌图像分类中的应用[J].计算机应用与软件,2006,23(10):103-106.

[3] Antonie M L,Zaiane O R,Coman A.Application of data mining techniques for medical image classification[C]//Proc of Second Intl Workshop on Multimedia Data Mining(MDM/KDD'2001)in Conjunction with Seventh ACM SIGKDD,San Francisco,USA,2001:94-101.

[4] Hassanien A E.Classification and feature selection of breast cancer data based on decision tree algorithm[J].International Journal of Studies in Informatics and Control,2003,12(1):33-39.

[5] Ahmed M N,Yamany S M,Nevin M.A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data[J].IEEE Transactions on Medical Imaging,2003,21(3):193-199.

[6] Bezdek J C,Ehrlich R.FCM:The fuzzy c-means clustering algorithm[J].Computers and Geosciences,1984,10:191-203.

[7] Starzyk J A,Dale N.A mathematical foundation for improved reduct generation in information systems[J].Knowledge and Information Systems Journal,2000:131-147.

[8] Hu X,Lin T Y.A new rough sets model based on database systems[J].Fundamenta Informaticae,2004,59(23):135-152.