文章编号:1673-1549(2014)06-0038-05　　DOI:10.11863/j.suse.2014.06.10

基于粗糙集和支持向量机的肿瘤图像识别

张海燕[a], 蓝集明[b]

(四川理工学院 a. 理学院; b. 计算机学院, 四川　自贡　643000)

摘　要:设计了一种基于粗糙集与支持向量基的乳腺肿瘤图像识别方法。其基本思想是首先对图片进行降噪预处理,接着提取纹理和形状特征构成表征医学图像的特征矢量,然后将特征离散归一化处理,再用粗集方法进行特征属性约简,最后利用支持向量基进行识别。结果表明,该方法取得了比较理想的识别效果。

关键词:医学图像;肿瘤识别;粗糙集;支持向量机;特征提取;属性约简

中图分类号:TP31　　**文献标志码**:A

引　言

随着计算机可视化技术和医学影像学的不断发展,现在已经能够为疾病的检验提供多种医学影像,包括计算机断层扫描(CT)、正电子放射层析成像技术(PET)、单光子辐射断层摄像(SPECT)、磁共振成像(MR)、超声(Ultrasound)及其它医学影像设备所获得的图像。这些图像虽然能够提供大量有用的信息,但是由于图像生成机理的原因,图像中往往包含了许多噪声,使得图像的质量不高,再加上诊断医师的视觉疲劳等因素,可能会造成误诊。因此,如何充分利用上述图像中所包含的有用信息,降低误诊率,仍然是医学界的一个难题。

本文借助现代图像处理技术,并结合人工智能、模式分类等算法,提出了基于粗集(Rough Sets)理论与支持向量机(SVMs)的医学影像识别方法,可用来识别某一乳腺图像里的肿瘤是否为恶性肿瘤。该方法依次采用了中值滤波、模糊图像增强法和区域生长法将图像中感兴趣的区域提取出来,然后使用粗集理论对提取出的 13 个图像特征值做了简化,最后将简化后的特征值作为支撑向量机的输入,识别乳腺肿瘤的良性与恶性。其基本流程如图 1 所示[1]。

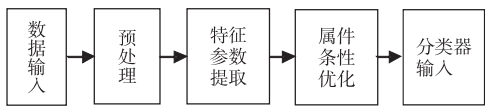

图 1　基于粗集理论和 SVMs 的乳腺图像识别流程

1　乳腺肿瘤图片的预处理

本文处理的图片来自 Mammographic Image Analysis Society(MIAS)数据集。MiniMIAS 数据库中的乳腺图像是由 X 线所得到的图像。由于 X 线的成像机理,其不可避免地使得图像中带有一定的噪声。图像去噪是医学图像识别中不可缺少的一步。先后采用了中值滤波、模糊图像增强法和区域生长法对乳腺肿瘤图片进行了预处理[2]。

图 2 是中值滤波之前和之后的乳腺部分图像对比。从图 2 可见,通过中值滤波后,图像中的肿瘤边界依旧保持良好。图 3 是将模糊增强算法应用到乳腺图像增强之前和之后的效果,可以看出,模糊图像算法增强了乳腺肿块与背景的对比度,其对边界的处理也较为理想,未出现边界模糊不清的现象。图 4 是使用区域生长

收稿日期:2014-06-06

基金项目:四川省教育厅科研项目(14ZB0214);自贡市科技局科研项目(2013ZC11);企业信息化与物联网测控技术四川省高校重点实验室项目(2014WYJ03);桥梁无损检测与工程计算四川省重点实验室科研项目(2014QZY01)

作者简介:张海燕(1977-),女,四川乐至人,讲师,硕士,主要从事应用数学和人工智能方面的研究,(E-mail)zhang_petrel@qq.com

法进行图像分割之前和之后的图像,可以看到对于病灶部位的提取是较为良好的。

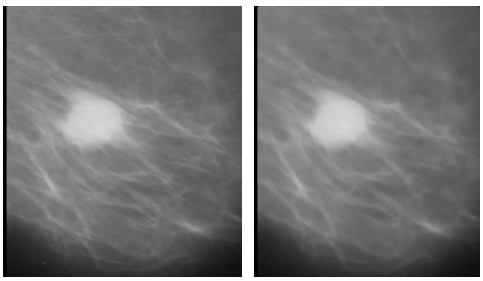

图 2　中值滤波之前和之后的乳腺部分图像

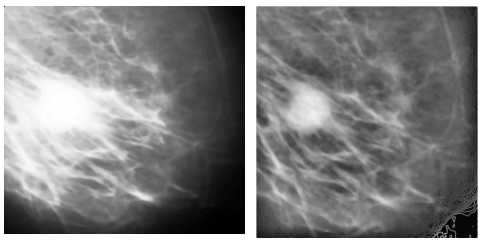

图 3　模糊增强之前和之后的乳腺部分图像

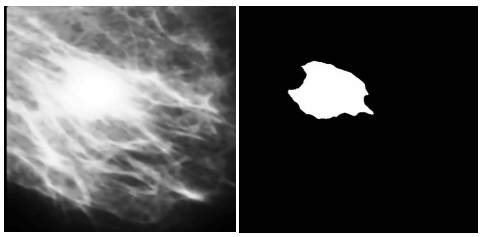

图 4　使用区域生长法之前和之后的乳腺部分图像

2　特征值的提取

在本文设计的乳腺肿瘤识别系统中,首先根据 MIAS 数据库中由专家指定的图像中乳腺肿瘤的中心位置为基准,人工切割包含乳腺肿块的一个大区域,经过图像去噪、图像增强后,利用计算机找出乳腺肿块的实际边界,然后在此基础上进行特征值的提取。

张红新等研究了乳腺肿瘤细胞核的形态,提出了用分形来提取乳腺肿瘤的特征[3]。本文主要从乳腺肿瘤的形状特征和纹理特征来提取乳腺肿瘤的特征[4]。

2.1　形状特征

乳腺肿瘤分为良性和恶性,对于良性肿瘤其通常是成椭圆状或圆形,而恶性肿瘤则通常表现为毛刺状或不规则状。乳腺图形的形状特征可以从乳腺肿瘤的边缘轮廓周长、肿瘤的面积、肿瘤的圆形度和椭圆紧致度等方面来概括乳腺肿瘤的特征。

其中,乳腺肿瘤的轮廓周长 p 和面积 s 可以直接统计分割好的图像上的像素点得到[5]。乳腺的圆形度 c 定义为:

$$c = 4\pi s/p^2$$

乳腺肿瘤的圆形度取值范围为 0 到 1,肿瘤越接近于圆形,圆形度越接近于 1。椭圆的紧致度 EC 定义为

$$EC = (m+n)\pi/p$$

其中, m,n 分别为对乳腺肿瘤拟合椭圆的长轴和短轴。椭圆紧致度表示的是拟合周长与原始边界周长的比,该值越小代表了乳腺肿瘤的毛刺程度越大,越有可能是恶性肿瘤。

2.2　纹理特征

纹理特征是一种区域描述方法,可以反映像素在空间上的分布属性。本文用灰度共生矩阵来提取图像的纹理特征。灰度共生矩阵是一种基于统计的分析纹理特征的方法,它根据在纹理中某一个灰度结构重复出现的情况而提取出来,不仅反映了图像的亮度分布特征,也反映了具有同样亮度或者接近亮度的像素点之间的位置分布特征[3]。

任何图像灰度表面都可以看成是三维空间中的一个曲面,在三维空间中相隔某一个距离的两个像素可能具有相同的灰度值。灰度共生矩阵就是要找出这样两个像素的联合分布的统计形式。灰度共生矩阵 G 的每一个元素 G_{ij} 表示一对灰度级 (i,j) 被给定的距离 d 分开的共生概率分布。它的计算就是从一个灰度值为 i 的像素点 (x,y) 出发,统计与其距离为 d, 方向为 θ, 灰度值为 j 的像素点 $(x+a,y+b)$ 同时出现的概率 $p(i,j,d,\theta)$。

$$p(i,j,d,\theta) = \Psi\{[(x,y),(x+a,y+b) \mid f(x,y) = i, f(x+a,y+b) = j]\}$$

其中,$f(\cdot)$ 表示该像素点的灰度值, $\Psi(\cdot)$ 表示该种情况出现的次数。一般地,可以有 0°、45°、90°、135° 四个方向来生成共生矩阵。即一个图像可以生成四个共生矩阵。

由生成的灰度共生矩阵,便可以得到图像的灰度特征。本文在四个方向各选取 9 个灰度特征,然后将其求平均值,得到本文的最终用于分类乳腺肿瘤的灰度特征。本文所选取的灰度特征有能量(T_1)、灰度平均(T_2)、梯度均值(T_3)、相关(T_4)、惯性(T_5)、相异性(T_6)、熵(T_7)、方差(T_8)和递差矩(T_9)。

$$T_1 = \sqrt{\sum_{i,j=1}^{L} p(i,j)^2}$$

$$T_2 = \sum_{i,j=1}^{L} i * p(i,j)$$

$$T_3 = \sum_{i,j=1}^{L} j * p(i,j)$$

$$T_4 = \sum_{i,j=1}^{L} \frac{p(i,j)(i-T_2)(j-T_3)}{\sigma_i \sigma_j}$$

$$T_5 = \sum_{m=0}^{L-1} m^2 \{ \sum_{i,j=1}^{L} p(i,j) \}$$

其中,$m = |i-j|$。

$$T_6 = \sum_{i,j=1}^{L} p(i,j) m$$

$$T_7 = -\sum_{i,j=1}^{L} p(i,j) \log(p(i,j))$$

$$T_8 = \sum_{i,j=1}^{L} p(i,j)(i-T_2)$$

$$T_9 = \sum_{i,j=1}^{L} \frac{p(i,j)}{1+m^2}$$

对图像预处理后,生成四个方向的灰度共生矩阵,依次提取各个灰度共生矩阵的上述 9 个特征。为避免由于图像旋转、移动等因素对特征值的影响,取四个共生矩阵各自生成特征值的平均值,得到最终的灰度特征值。

由此,本文共选取了 4 个形状特征和 9 个纹理特征共 13 个特征值作为识别乳腺肿瘤良性恶性的标准。由于特征值规模较大,为避免维数灾难,在进行识别之前有必要对特征值进行一定的简化。

3 基于粗集理论的肿瘤图片特征属性的优化

对于一个系统 $S = (U,C,D,V,f)$,这里 U 是对象的非空有限集合,$C \cup D = R$ 是属性集合,不相关的子集 C 和 D 分别称为条件属性集和决策属性集。考虑条件属性相对于决策属性的约简,条件属性集 C 中有些属性对分类规则的确定不起作用,是多余的,这些冗余属性要去掉,以达到属性的简化[6]。

根据粗集理论,定义系统中的条件属性集 C 和结果属性集 D 之间的关系依赖性。

$$k = \frac{card(POS_c(D))}{card(U)} \tag{1}$$

式(1)中,$card(POS_c(D))$ 表示基于条件属性集 C 的结果正域子集数目,$card(U) = |U|$ 表示对象论域数目。若 $k = 1$,则表明这个系统是协调的;若 $k_{R-r} = 1$,则表明去除条件属性 r 后系统仍是协调的。

根据粗集理论,定义系统中条件属性集 C 和结果属性集 D 等价关系之间的一致性。

$$Q_c = \frac{card(U|C)}{card(U)} \tag{2}$$

式中,$card(U|C)$ 表示基于条件属性集 C 的等价类数目,$card(U) = |U|$ 表示对象论域数目。若 $Q_c = 1$,则表明这个系统中条件属性集 C 和结果属性集 D 等价关系是一致的;若 $Q_{c-r} = 1$,则表明去除条件属性 r 后,该系统中条件属性集 C 和结果属性集 D 等价关系仍是一致的。

因此,利用粗集算法减少输入信息的表达特征数量,降低网络结构的复杂性,去掉冗余信息,将使网络的训练集合简化,减少了网络的训练时间,提高了网络的实时性[7]。

rosetta 软件提供了多种属性简约方法,包括遗传算法、Johnson 方法、Holte's 1R 等方法。本文采用 rosetta 遗传算法,它是一种自适应随机搜索方法,是基于达尔文生化进化论的自然选择学说和孟德尔的基因遗传学原理而建立的,具体步骤:

(1) 输入决策表 s。

(2) 求条件属性 C 相对于决策属性 D 的核 $core(C)$。令 $core(C) = \phi$,逐个去掉一个属性 $r \in C$,若 $pos_{C-\{r\}}(D) \neq pos_c(D)$,则 $core(C) = core(C) \cup \{r\}$。

(3) 随机产生 M 个长度为 n(原来决策表 s 中条件属性的个数)的二进制串组成初始种群,对于核 $core(C)$ 的属性,其对应位置取 1,其他位置随机取 0 或 1。

(4) 用公式 $f = e^{k-\frac{card(x)}{n}}$ 计算出每个个体的适应度值,其中 k 为决策属性对该个体所含条件属性的依赖度。采取轮盘赌方式进行选择操作,同时将最佳个体直接保留到下一代。

(5) 选择父代个体分别采用单点交叉、简单变异,自适应

$$p_c = \begin{cases} \dfrac{1}{1+e^{\beta(f_{avg}-f')}} & f' \geqslant f_{avg} \\ 1 & f' < f_{avg} \end{cases}$$

$$p_m = \begin{cases} \dfrac{1}{2} e^{\beta(f_{avg}-f)} & f \geqslant f_{avg} \\ 0.5 & f < f_{avg} \end{cases}$$

其中,f_{avg} 为群体的平均适应值,f' 为预交叉的两个个体中较大的适应度值,f 为欲变异的个体的适应度值,β 为(0,1)区间上的一个常数。

(6) 修剪相似个体,并动态补充新个体,得到下一代种群。

(7) 若连续若干代的最优个体的适应值不再提高,则终止算法,并输出最优个体,否则转到第 4 步。

(8) 输出属性的简约表。

所提取出的 13 个特征属性简化为了 9 个。它们分别是肿瘤的周长、圆形度、椭圆紧致度、灰度共生矩阵的能量、灰度平均、相关性、惯性、熵、方差。

4　基于 SVMs 的乳腺肿瘤识别

支持向量机是基于结构性风险最小化原则的理论,能够克服传统学习方法中经验最小化的局限,使得其在小样本的情况下也具有良好的泛化能力。从 MIAS 数据库中得到的乳腺图像,通过图像的预处理、特征值的提取以及粗集对于特征值的简化以后,便可以将其输入到支持向量机中进行训练[8-9]。这一过程的流程如图 5 所示。

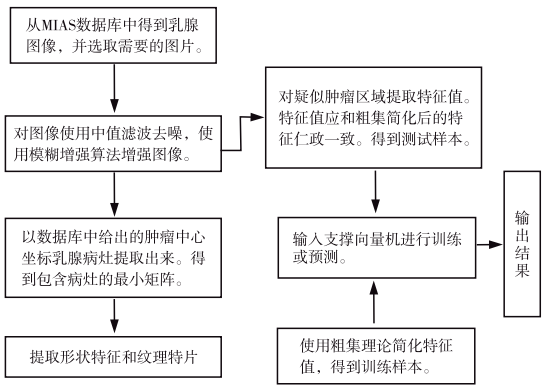

图 5　基于 SVM 的乳腺图像识别流程图

mini – MIAS 库提供的乳腺图片是 200 μm × 200 μm 的分辨率。库中有 322 个图片,分别来自 161 位病人。图片分三类:正常无肿块、正常肿块、异常。对于有肿块的乳腺,数据库中给出了由专家指定的乳腺肿块的位置和类别。数据库中的乳腺肿瘤又分为三类:圆形或椭圆形、不规则型、毛刺型。本文选取圆型乳腺肿瘤共 14 个,其中良性 10 个、恶性 4 个;不规则乳腺肿瘤共 15 个,其中良性 8 个、恶性 7 个;毛刺型乳腺肿瘤共 15 个,其中良性 7 个、恶性 8 个。选取的规则是要尽量考虑正负样本的比例,又要在数据库能够得到这个比例。

对所选取的 44 个训练样本按照流程图依次处理。对提取出的肿瘤区域提取其纹理特征和形状特征共 13 个。它们分别为肿瘤的周长、面积、圆形度、椭圆紧致度、灰度共生矩阵的能量、灰度平均、梯度平均、相关性、惯性、相异性、熵、方差和逆差距。经过离散化和粗集简化以后的特征值共 9 个,分别为肿瘤的周长、圆形度、椭圆紧致度、灰度共生矩阵的能量、灰度平均、相关性、惯性、熵和方差。

采用交叉训练的方式进行训练。将所有的训练样本随机均分为 4 个部分,依次使用其中三部分样本作为训练集,剩下的部分作为测试集进行训练。构造一个 4 分类的支持向量机模型,模型中的核函数采用高斯径向基核函数,模型中的参数经过试验选取为: $c = 100, \delta = 4$。

将交叉训练的四个模型得到的预测结果加总,得到最终的预测结果。本文所得到的预测精度见表 1。表 1 中数据显示对于恶性肿瘤的识别的识别率较高,但需要指出的是,为了包含各种形状的恶性肿瘤,已经将 MIAS 数据库中所有的样本都纳入了训练样本中。尽管如此,样本的数量还是偏少。下一步工作,可以考虑扩充数据库中各样本的数目,进一步验证预测精度。

表 1　模型预测识别率

名称	总个数	正确识别个数	识别率/%
良性肿瘤	25	21	84.00
圆形恶性肿瘤	4	4	100.00
不规则肿瘤	7	6	85.71
毛刺型肿瘤	8	8	100.00
恶性肿瘤	19	17	89.473

5　结束语

从医学透片中挖掘出肿瘤的特征,是一个从不准确、不完整、不确定的大量数据中发现知识的过程。在这方面粗糙集和支持向量机显示出了无穷的魅力。支持向量机可以实现有导师和无导师的学习,并且解决传统机器学习算法存在的一 些问题,如神经网络的过学习问题和局部最小问题,但不能确定哪些知识是冗余的,哪些属性是有用的;粗集理论方法可以描绘知识表达中不同属性的重要性,进行知识表达中不同属性的重要性,但粗糙集对适应变化的环境和系统本身的容错性都不如支持向量机。本文将两者结合起来,将粗糙集作为支持向量机的前置系统,对输入支持向量机的数据进行预处理,使用粗糙集对输入支持向量机的数据进行简约,减少了支持向量机的输入空间维数,去掉了训练集中的冗余信息,使训练集得到简化,训练时间更快,识别精度更高,模型验证的精度能够满足辅助医学诊断的目的。

参 考 文 献:

[1] Agrawal P, Vatsa M, Singh R. Saliency based mass detection from screening mammograms[J]. Signal Processing,2014,99:29-47.

[2] Maitra I K,Nag S,Bandyopadhyay S K. Technique for preprocessing of digital mammogram[J]. Computer Methods and Programs in Biomedicine.2012,107(2):175-188.

[3] 张红新,赵培荣,高冬玲,等.乳腺肿瘤细胞核形态的分形分析[J].郑州大学学报:医学版,2002(4):431-433.

[4] Wang X,Yang J,Teng X,et al.Feature selection based on

rough sets and particle swarm optimization[J]. Pattern Recognition Letters,2007,28(4):459-471.
[5] 华 翔,孙 蕾.基于 SVM 的医学图像分类器的设计[J].微电子学与计算机,2011(6):171-175.
[6] 蔡乐才,姚行艳.基于粗集理论的属性约简算法[J].四川理工学院学报:自然科学版,2009(1):34-37.
[7] 阎成栋.基于粗糙集的医学图像增强系统研究[D].太原:太原理工大学,2010.
[8] Eddaoudi F, Regragui F, Mahmoudi A, et al. Masses detection using SVM classifier based on textures analysis[J]. Applied Mathematical Sciences,2011,5(8):367-379.
[9] Mohamed H,Mabrouk M S,Sharawy A.Computer aided detection system for micro calcifications in digital mammograms[J]. Computer Methods and Programs in Biomedicine,2012,92:357-362.

A New Method for Tumor Images Recognition Based on Rough Sets and SVMs

ZHANG Haiyan[a], *LAN Jiming*[b]
(a. School of Science; b. School of Computer Science, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: In this research, a novel method is proposed which is used for breast tumor images recognition based on rough sets and support vector machines (SVMs). The basic idea is: firstly, the image is pre-processed for noise reduction. Secondly, the texture and shape features are extracted to constitute the feature vector that can represent the mammogram accurately. Next, the features are discrete normalized. Finally, attribute reduction by rough sets and classification recognition by SVMs are completed. The experimental results show that this method can achieve a satisfactory effect for mammographic recognition.

Key words: medical image; tumor recognition; rough sets; SVMs; feature extraction; attribute reduction

(上接第 37 页)

Simulation Research of a Vertical-Axis Wind Turbine Base on MATLAB

YANG Rui[1,2], *LI Jinlong*[1], *XIA Weiwei*[1], *LI Dandan*[1]
(1. Shool of Energy and Power Engineering, Lanzhou University of Technology, Lanzhou 730050, China; 2. Wind Power Technology Center, Lanzhou 730050, China)

Abstract: In order to describe the output characteristics of a vertical axis wind turbine, the wind speed model, drive system and speed control system of a vertical-axis wind turbine are established in MATLAB/SIMULINK. Particularly, taking a 5KW vertical axis wind turbine as example, the power coefficients under the different tip speed ratio are received by using Double-Multiple Streamtube (DMS) model. Then the data is imported into MATLAB, and the mathematical model of the 5 KW vertical axis wind turbine rotor is obtained by using the curve fitting toolbox, thus the SIMULINK model is built. Later the wind turbine model is formed by these above models and other models. Some operation parameters of wind turbine are given, and the data for analysis is received. Through the simulation study, MATLAB/SIMULINK can better simulate the overall performance of a wind turbine from wind turbines to the drive system, which provides a reference for the future research of the overall performance of vertical axis wind turbines.

Key words: wind turbine; vertical-axis; performance study; simulation; MATLAB