基于胸部CT图像的肺癌识别方法的研究

姜慧研,何 炜
(东北大学软件学院,辽宁沈阳 110004)

摘 要: 针对医学影像中小结节容易被漏诊的问题,提出了基于胸部CT图像的肺癌计算机辅助诊断新方法.首先从胸部CT图像分割出关心区域(ROI);然后提取ROI的特征;其次采用RS理论选择有效特征;最后基于这些有效特征建立面向不同需求的肺癌识别模型.即如果需要快速诊断,则利用SONN建立肺癌识别模型;如果需要进行准确诊断,则利用SPAM建立肺癌识别模型和非肺癌识别模型,并根据待识别样本与模型的相似度判断所属类别.但是当相似度较小时,则利用HMM进一步识别.通过实验验证了该方法的有效性.

关键词: 肺癌; 计算机辅助诊断; 自适应概率统计模型; 隐马尔可夫模型
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2009) 08-1664-05

Research of Lung Cancer Recognition Based on Chest CT Images

JIANG Hui-yan, HE Wei
(*School of Software, Northeastern University, Shenyang, Liaoning* 110004, *China*)

Abstract: To solve the missed diagnosis of small pulmonary nodules in medical images, a new approach on computer-aided diagnosis for lung cancer based on chest CT images has been proposed. The method firstly segments the Region of Interest(ROI), and extracts ROI's features. Then it selects effective attributes by theory of Rough Set(RS). Finally, it constructs a specific-demand oriented recognition model for lung cancer based on these effective features. Especially, we take the Self-Organizing Neural Network (SONN) to construct the recognition model of lung cancer for fast diagnosis. In order to perform the accurate diagnosis, we need to use the Self-Adaptive Probabilistic Model(SAPM) to build lung cancer and non-cancer recognition models respectively and we can identify the classification by the similarity of the recognition sample with the model. When the similarity is small, we re-identify the lung cancer by Hidden Markov Model(HMM). The experiment results proved that the approach mentioned in this paper can hold high efficiency.

Key words: lung cancer; computer-aided diagnosis(CAD); self-adaptive probabilistic model(SAPM); hidden markov model (HMM)

1 引言

肺癌作为人类癌症死亡的主要病因,成为医学界关注的焦点.肺癌的计算机辅助诊断(CAD)在准确排查病变、降低漏诊、降低劳动强度和提高阅片效率方面具有重要的研究意义和应用价值.

在CAD领域,很多学者开展研究并取得了一定的成果.例如,文[1]提出了利用阈值法检测孤立肺结节的方法;文[2]提出了利用改进的支持向量机进行癌症诊断的方法;文[3]提出了首先将粗糙集[4]运用于临床特征参数的约简,然后进行恶性脑瘤的MRI医学诊断方法;文[5]提出了根据粗糙集与概率型神经网络系统识别乳腺癌的方法;文[6]提出了利用大规模训练人工神经网络(MTANNs)降低肺结节的假阳性.

但是上述方法均是基于一幅图像的特征进行处理的,具有一定的片面性.事实上,医生进行影像诊断时,有时基于一幅图像就能快速诊断肺癌,有时需要观察多个相邻图像才能给出准确的诊断.根据医生诊断过程,本文分别提出了基于一幅CT图像快速诊断肺癌和基于序列CT图像准确诊断肺癌两种方案.即首先从胸部CT图像分割出关心区域(ROI);然后从ROI中提取特征;其次从中进行有效特征的选择;最后基于这些有效特征分别建立适合快速诊断和准确诊断的肺癌识别模型.

收稿日期:2008-04-29;修回日期:2009-03-20
基金项目:国家自然科学基金(No.60671050)

2 ROI 分割和特征提取

为了提取肺癌模型中的特征,本文基于文[7]提出的双快速行进法进行肺结节等关心区域(Region of Interest,ROI)的分割,分割结果如图 1 所示.

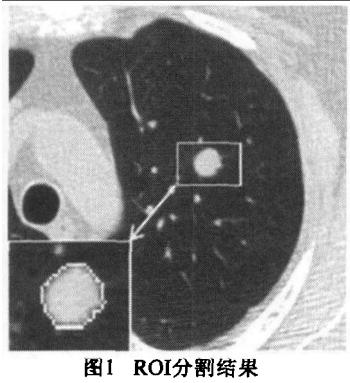


图1 ROI分割结果

为了较好地表现肺癌特征,本文从 ROI 提取了面积、欧拉数、周长、似圆度、矩形度、伸长度、质心横坐标、质心纵坐标、平均半径、半径方差(标准差)、方向角、偏心度等 16 个几何特征,以及平均灰度值、灰度方差、三阶矩、最大灰度值、最小灰度值、灰度熵、角二阶矩、惯量、逆差矩、熵、小梯度优势、二阶原点矩、不变矩等 34 个纹理特征,共计 50 个特征.

但是,50 个特征所构成的特征空间维数过高,为了从中提取对肺癌识别贡献较大的特征,本文基于粗糙集(Rough Sets,RS)理论[4]的知识约简进行有效特征的选择.粗糙集理论的基本思想是:在保持分类能力不变的情况下,通过知识约简,导出问题的决策或分类规则.

3 肺癌识别方法

医生读影时,往往先从一幅图像观察是否存在肺癌病灶,因此,本文首先将上一节提取的有效特征作为输入、基于自组织竞争型神经网络(SONN)[8]建立面向快速诊断需求的肺癌分类模型,进行特征明显的肺癌的诊断.

对于肺癌特征不明显的 CT 图像,临床医生往往通过观察和分析疑似病灶在相邻几个 CT 图像中的变化情况才能给出诊断.针对这种情况,本文提出了自适应概率统计模型(简称 SPAM 模型),利用序列图像进行肺癌的识别.

3.1 基于 SPAM 的肺癌识别方法

为了描述病灶在图像间的变化情况,本文定义了"断层特征"的概念.断层特征是指通过状态迁移将相邻断层 *ROI* 几何特征(质心和面积)的变化转化成状态值(不变、偏移、偏离等),即断层特征包括质心状态和面积状态.质心状态和面积状态特征的计算如表 1 所示.其中,P_i 表示第 i 个 CT 图像中 ROI 的质心坐标,A_i 表示第 i 个图像中 ROI 的面积,$V_j(j=1,\cdots,3)$表示状态间的阈值.

表 1 质心状态和面积状态

质心变化	面积变化	质心状态	面积状态
$\lvert P_i - P_{i+1}\rvert < V_1$	$\lvert A_i - A_{i+1}\rvert < V_3$	不变	不变
$V_1 < \lvert P_i - P_{i+1}\rvert < V_2$	$A_{i+1} - A_i < -V_3$	偏移	减小
$\lvert P_i - P_{i+1}\rvert > V_2$	$A_{i+1} - A_i > V_3$	偏离	增大

(1)肺癌训练模型

这里的"样本"是指某一组织的断层特征,即如果某一组织在连续 $N+1$ 个 CT 图像上出现,则对其中相邻的两个图像进行状态迁移,全部迁移后得到的断层特征形成该组织的一个"样本".由于每两个相邻断层之间只做一次状态迁移,所以迁移后有 N 个状态值.例如,统计 M 个肺癌样本,在状态序列 i 上,M_i 个样本共有 M_{i1} 个质心状态值为不变,M_{i2} 个质心状态值为偏移,M_{i3} 个质心状态值为偏离,$M_i = M_{i1} + M_{i2} + M_{i3}$,则肺癌模型在状态序列 i 上的质心状态概率的定义如下:

$$P_i(Status = '不变') = \frac{M_{i1}}{\sum_{j=1}^{3} M_{ij}} \tag{1}$$

$$P_i(Status = '偏移') = \frac{M_{i2}}{\sum_{j=1}^{3} M_{ij}} \tag{2}$$

$$P_i(Status = '偏离') = \frac{M_{i3}}{\sum_{j=1}^{3} M_{ij}} \tag{3}$$

其中,P_i 表示第 i 个状态序列上各类状态值出现的概率.

基于 SPAM 建立肺癌训练模型的算法如下:

① 初始化状态迁移数 N 及肺癌样本数 M.

② 提取第 i 个肺癌样本的质心和面积特征值($i \in [1,M]$).

③ 基于状态迁移的方法计算样本 i 在状态序列 j 上的质心状态值 VP_{ij} 和面积状态值 VA_{ij}.

④ 设质心状态矩阵为 ***MP***,面积状态矩阵为 ***MA***,***MP***$[j][VP_{ij}]$为肺癌样本 i 在状态序列 j 上出现 VP_{ij} 状态值的频率,***MA***$[j][VA_{ij}]$为肺癌样本 i 在状态序列 j 上出现 VA_{ij} 状态值的频率,将 ***MP***$[j][VP_{ij}]$和 ***MA***$[j][VA_{ij}]$分别加 1.

⑤ 令 $j = j+1$.如果 $j \leqslant N$,则转到③;否则,转到⑥.

⑥ 令 $i = i+1$.如果 $i \leqslant M$,则转到②;否则,结束.

非肺癌模型的建立算法与肺癌模型基本相同,限于篇幅,不再赘述.

(2)肺癌预测模型

定义待测样本和肺癌模型之间的相似度如下:

$$S = \sum_i \{P_j \mid j = Status(Y_i)\} \tag{4}$$

其中，S 为待测样本和肺癌模型的相似度，i 为状态序列编号，Y_i 为待测样本在状态迁移后的第 i 个状态序列的质心(或面积)状态值，$Status(Y_i)$ 表示肺癌模型在第 i 个状态序列状态值 Y_i 所属的状态，P_j 表示第 i 个状态序列的肺癌组织的质心(或面积)在状态值 j 等于 $Status(Y_i)$ 的概率.

待测样本和非肺癌模型之间的相似度定义与上式相同，但是，$Status(Y_i)$ 表示非肺癌模型在第 i 个状态序列状态值 Y_i 所属的状态；P_j 表示第 i 个状态序列的非肺癌组织的面积(或质心)状态值为 j 等于 $Status(Y_i)$ 的概率.

建立肺癌预测模型的算法如下：

① 初始化状态迁移数 N 和模型数 P，计算待测样本的质心和面积.

② 计算模型 i 的模型性质、训练样本数、状态迁移数等参数和质心与面积的状态矩阵，$i\in[1,P]$.

③ 基于状态迁移的方法，分别计算状态序列 j 时的样本质心状态值 VP_{ij} 和面积状态值 VA_{ij}.

④ 计算待测样本的质心和面积状态值在模型 i 对应序列 j 上出现的概率，$PP_{ij}=MP[j][VP_{ij}]/M$ 和 $PA_{ij}=MA[j][VA_{ij}]/M$，并根据式(4)累加到相似度 S_j 上.

⑤ 令 $j=j+1$，如果 $j\leqslant N$，转到④；否则，转到⑥.

⑥ 令 $i=i+1$，如果 $i\leqslant P$，执行②；否则，转到⑦.

⑦ 选取 S 集合中最大的 $S_j=S_{max}$，则模型 j 代表的类别即为待测样本所属的类别.

⑧ 如果样本与非肺癌模型的相似度值最大，则判断为非肺癌；否则，则判断为肺癌.

(3)自适应的肺癌训练模型

由于该方法尚需要确定质心阈值 V_1、V_2 和面积阈值 V_3，凭经验确定的阈值往往不是最优的，且随着样本变化而导致精度下降且识别结果不稳定.所以本文提出了一种自适应调整算法，使得模型在训练的过程中可以自动调整各参数，最终确定一组稳定的优选参数作为阈值.

建立自适应的肺癌训练模型算法如下：

① 根据经验选择初始质心、面积阈值.

② 读入当前样本，提取其特征，并做状态迁移，获得断层特征值.

③ 计算肺癌(或非肺癌)训练样本与肺癌(或非肺癌)模型的相似度.

④ 根据下式修正阈值.

$$\theta(t+1)=\theta(t)-\alpha\Delta E \tag{5}$$

其中，α 为学习因子，用于控制学习速率，$\alpha\in[0,10]$；V_1、V_2 和 V_3 二值化(大于中值的二值化为 1，小于中值的二值化为 0)，其中 V_1 和 V_2 的学习因子 α 相同，V_3 的学习因子和它们的 α 不同.t 代表时刻.ΔE 为训练误差，$\Delta E\ [-1,1]$；ΔE 越大，说明当前模式与监督信号的期望背离，故 ΔE 越小，阈值所做的调整越小.

当出现下面两种诊断错误的情况时，系统需要修正参数：一是当 $\Delta E<0$ 时，说明实际输出小于期望输出，期望输出 = 1 且实际输出 < 1，即实际为肺癌，检测结果为非肺癌；二是当 $\Delta E>0$ 时，说明实际输出大于期望输出，期望输出 = 0 且实际输出 > 0，即实际为非肺癌，检测结果为肺癌.

式(5)的展开式为：

$$\begin{aligned}&\theta(t+1)=\theta(t)-\alpha\left(\frac{1}{3}O'-O\right)\\&\Delta E=\frac{1}{3}O'-O,\Delta E\in[-1,1]\end{aligned} \tag{6}$$

其中，O 代表期望输出(监督信号)，O' 代表实际输出.当 $O=0$ 时表示非肺癌，$O=1$ 时表示肺癌，如式(7)所示.

$$\begin{aligned}&O'\sim S_{area}\text{ or }S_{pos},\\&S_{area},S_{pos}\in[0,3],O'\in[0,3]\end{aligned} \tag{7}$$

其中，S_{area} 和 S_{pos} 分别代表面积和质心的相似度.

⑤ 如果训练样本读取完毕，则结束；否则，转到②.

非肺癌训练算法与上述算法类似.基于该方法分别建立肺癌模型和非肺癌模型，通过计算样本与这两个模型的相似度来判断样本究竟和哪个模型更为“接近”.但通过实验发现，当样本与两个模型的相似度值均“很小”的情况下，该方法的识别误差较大.因此，本文提出了利用隐马尔可夫模型来解决这个问题.

3.2 基于 HMM 的肺癌识别方法

隐马尔可夫模型(HMM)[9]处理一维信号序列效果较好[10]，因此将状态迁移后的特征状态值进行序列编码，转换成适合 HMM 的信号序列，然后对其进行训练和预测.

与 SPAM 类似，HMM 的输入信号也是基于状态迁移后的断层特征的状态值，需要在质心和面积状态值的基础上进行合并和重新编码，如表 2 所示.

表 2 HMM 信号编码

合并后的状态值(质心|面积)	编码值
不变|不变	1
…	…
偏离|减小	9

这 9 组数据构成 HMM 的观测值，而状态值由“肺癌”、“可疑”、“非肺癌”组成.

如果将观测值记为 O，状态值记为 S，则 HMM 五元组中的状态转移矩阵 $\boldsymbol{A}$ 如表 3 所示.

概率分布矩阵 $\boldsymbol{B}$ 如表 4 所示.

表 3　状态转移矩阵

A	肺癌(阳性)	可疑	非肺癌(阴性)
肺癌	S_{11}	S_{12}	S_{13}
可疑	S_{21}	S_{22}	S_{23}
非肺癌	S_{31}	S_{32}	S_{33}

表 4　产生观测值的概率矩阵

B	①	②	…	⑨
肺癌	O_{11}	O_{12}	…	O_{19}
可疑	O_{21}	O_{22}	…	O_{29}
非肺癌	O_{31}	O_{32}	…	O_{39}

其中,O 和 S 下标 i 和 j 分别表示在状态 i 时,转移为状态 j 或观测值 j 的概率大小.

然后,采用 Baum-Welch 算法训练 HMM 模型,再用前向-后向算法进行预测.

4　实验结果与分析

(1)基于 SONN 的肺癌识别实验

本文选取 200 个样本进行实验,其中,肺癌个数为 50 个,非肺癌个数为 150 个.对样本进行 ROI 分割后提取原始特征 50 个,然后采用粗糙集进行属性约简,得到 7 组约简后的特征,选择的有效特征包括:面积、欧拉数、平均半径、灰度方差、惯量、熵和不变矩.通过特征选择实验发现:基于 50 个特征和约简后 7 个有效特征对肺癌的识别率分别是 81% 和 83%.

利用上述 7 个有效特征,本文又基于 BP、RBF、SONN 建立肺癌识别模型和进行识别实验,其识别率分别为 75%、77%、83%.

(2)基于 SPAM 的肺癌识别实验

选择 262 个图像作为训练样本,200 个图像作为测试样本.其中,训练样本集包含 94 个肺癌样本和 168 个非肺癌样本;测试样本集包含 56 个肺癌样本和 144 个非肺癌样本.

根据经验,N 取 4,即每个样本由同一组织在相邻连续 4 幅 CT 图像中的断层特征组成.建立的模型只包括肺癌模型和非肺癌模型,模型数 $P=2$.

对训练样本中的 94 个肺癌样本,基于 SPAM 建立肺癌模型,另外 168 个非肺癌样本建立非肺癌模型.再从另外 200 个待测样本中提取几何特征值,并做断层状态迁移.根据经验,质心状态初始阈值 V_1 为 3,V_2 为 10,面积状态初始阈值 V_3 为 5,样本总数 M 为 200,状态迁移数 L 为 3,计算状态值.然后利用式(4)计算每个测试样本与肺癌模型和非肺癌之间的相似度.

为了更精确地判断特征非常接近的肺癌、非肺癌组织,将样本性质分成五类,分别是:阳性 + +、阳性 +、可疑、阴性 +、阴性 + +.阳性 + + 的肺癌数表示实际为肺癌被 CAD 诊断为阳性 + + 的样本个数,阳性 + + 的非肺癌数表示实际为非肺癌被 CAD 诊断为阳性 + + 的样本个数;阳性 + 的肺癌数表示实际为肺癌被 CAD 诊断为阳性 + 的样本个数,阳性 + 的非肺癌数表示实际为非肺癌被 CAD 诊断为阳性 + 的样本个数;可疑的肺癌数表示实际为肺癌被 CAD 诊断为可疑的样本个数,可疑的非肺癌数表示实际为非肺癌被 CAD 诊断为可疑的样本个数;阴性 + 的肺癌数表示实际为肺癌被 CAD 诊断为阴性 + 的样本个数,阴性 + 的非肺癌数表示实际为非肺癌被 CAD 诊断为阴性 + 的样本个数;阴性 + + 的肺癌数表示实际为肺癌被 CAD 诊断为阴性 + + 的样本个数,阴性 + + 的非肺癌数表示实际为非肺癌被 CAD 诊断为阴性 + + 的样本个数.最终得出的肺癌诊断结果如表 5 所示,对应的 ROC(Receiver Operating Characteristic)曲线如图 2 所示.可以看出:利用 SPAM 方法对肺癌和非肺癌的总体识别率为 91%.

表 5　肺癌识别结果

	阳性 + +	阳性 +	可疑	阴性 +	阴性 + +
肺癌数	10	36	2	2	6
非肺癌数	4	4	36	66	34

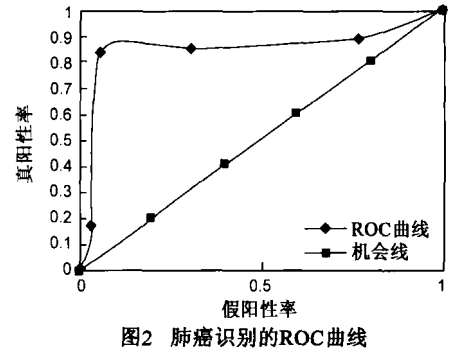


图2　肺癌识别的ROC曲线

(3)基于 HMM 的肺癌识别实验

HMM 的样本与 SPAM 样本相同,首先将样本转换成适合 HMM 的信号序列格式;然后根据 Baum-Welch 算法和已知是肺癌的 45 个样本建立肺癌模型;将 76 个非肺癌样本单独训练,建立非肺癌模型;其次,使用 97 个测试样本进行预测实验,其中的 51 个样本在 SPAM 试验中计算获得的相似度小于 1,其余 46 个样本的相似度则大于 1.经过前向-后向算法试验,基于 HMM 与 SPAM 建立模型的比较如表 6 所示.

表 6　基于 HMM 与 SPAM 建立模型的比较

	相似度	识别率(正确数/总数)
SPAM	相似度小于 1	88.24%(45/51)
	相似度大于 1	93.48%(43/46)
	整体	90.72%(88/97)
HMM	相似度小于 1	94.12%(48/51)
	相似度大于 1	89.13%(41/46)
	整体	91.75%(89/97)

可见,在相似度较小的情况下,HMM 所表现出来的识别效果更佳,因此,本文将 HMM 作为 SPAM 的补充.

5 结论

本文基于粗糙集理论进行特征选择,使特征空间维数从 50 压缩到 7;定义了断层特征,丰富了二维 CT 图像的信息;提出了基于一幅 CT 图像的快速识别肺癌方法和基于序列图像的准确识别方法.降低了运算成本.利用 HMM 对趋势变化的预测能力进一步识别相似图像.即利用断层特征建立肺癌模型和非肺癌模型,根据待测样本与这两个模型的相似度进行识别.对于相似度较小的样本,则采用改进后的 HMM 模型进行二次分类,提高了识别精度.

参考文献:

[1] 薛以锋,鲍旭东,马汉林,吴磊.基于 CT 图像的肺结节计算机辅助诊断系统[J].中国医学物理学杂志,2006,23(2):93 - 96.
Xue Yi-feng, Bao Xu-dong, Ma Han-lin, et al. Computer-aided diagnosis system for pulmonary nodules based on CT images [J]. Chinese Journal of Medical Physics, 2006, 23(2): 93 - 96. (in Chinese)

[2] 王晶,卫金茂.一种改进的支持向量机及其在癌症诊断中的应用[J].计算机应用,2006,26(2):508 - 511.
Wang Jing, Wei Jin-mao. Improved SVM algorithm application on Cancer Diagnoses [J]. Journal of Computer Applications, 2006, 26(2): 508 - 511. (in Chinese)

[3] X Wang, J Yang, R Jensen, X. Liu. Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma[J]. Computer Methods and Programs in Biomedicine, 2006, 83(2): 147 - 156.

[4] Z Pawlak. Rough set approach to knowledge-based decision support[J]. European Journal of Operational Research, 1997, 99(2): 48 - 57.

[5] K Revett, F Gorunescu, M Gorunescu, et al. A breast cancer diagnosis system: a combined approach using rough sets and probabilistic neural networks[A]. In Proceedings IEEE EUROCON[C]. Belgrade, Serbia, 2005, 2: 1124 - 1127.

[6] S Kakeda, J Moriya, H Sato1, et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system[J]. Am J Roentgenol, 2004, 182(2): 505 - 510.

[7] Z Y Cheng, H Y Jiang. Segmentation of pulmonary nodules based on improved dual fast marching method[J]. IEICE Technical Report, 2007, 106(509): 79 - 82.

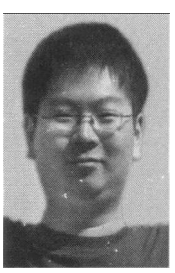
[8] 许少华,何新贵,李盼池.自组织过程神经网络及其应用研究[J].计算机研究与发展,2003,40(11):78 - 81.
S Xu, X He, P. Li. Research and applications of self-organization process neural networks[J]. Computer Research and Development, 2003, 40(11): 78 - 81. (in Chinese)

[9] L R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition[J]. Proceeding of the IEEE, 1989, 77(2): 257 - 286.

[10] J H Cai, Z Q Liu. Hidden markov models with spectral features for 2D shape recognition[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2001, 23(12): 1454 - 1458.

作者简介:

姜慧研 女,1963 年生于辽宁鞍山.副教授.2000 年和 2009 年在东北大学分别获得硕士和博士学位.2001 年 10 月 ~ 2002 年 9 月,在日本岐阜大学专门研修计算机辅助诊断技术.主要从事计算机辅助诊断、图像处理与分析、三维可视化、模式识别、优化控制、专家系统等方面的研究与教学工作.
E-mail: hyjiang@mail.neu.edu.cn



何 炜 男,1982 年生于浙江杭州.东北大学研究生.主要从事模式识别、图像处理等方面研究.