# 肺癌分类识别中的神经网络集成技术研究*)

杨育彬　李　宁　陈世福　陈兆乾

（南京大学计算机软件新技术国家重点实验室　南京210093）

## Neural Network Ensemble in Lung Cancer Cell Identification

YANG Yu-Bin　LI Ning　CHEN Shi-Fu　CHEN Zhao-Qian

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

**Abstract**　This paper describes a neural network ensemble method in lung cancer cell identification process. A neural network ensemble algorithm LCNE based on image feature extraction is proposed. Firstly, LCNE algorithm trains different neural network classifiers designed for shape and color features individually. Then, using neural network ensemble method, the final lung cancer cell identification results can be achieved. On the basis of LCNE algorithm, we implement a lung cancer cell identification system LCDS successfully.

**Keywords**　Neural network ensemble, Feature extraction, Image recognition, Lung cancer diagnosis

## 1　引言

肺癌已成为目前人类因癌症死亡的主要原因。有资料表明，我国肺癌发病率将在相当长时期内呈现显著上升趋势。同时，肺癌病例一经确诊，80%已属晚期，失去手术治疗的机会。运用计算机图像处理和模式识别技术对肺癌显微切片标本的彩色图像进行肺癌的分类诊断，将有助于及早发现癌变，实施准确的早期诊断和治疗。因此，肺癌图像的分类识别技术研究是图像处理与模式识别领域的一个具有重大现实意义的研究课题。

智能医学图像处理与识别及其在临床诊断中的应用，已经成为当前计算机界和医学界共同的研究热点，并且在国内外已有很多研究和应用。医学图像识别的对象包括CT(X光断测扫描图像)、MRI(核磁共振图像)、SPECT(单探头光子断层扫描图像)以及PET(正电子断层扫描图像)等多种医学成像技术[1~4]。这些系统使用的主要方法是：首先对原始图像进行图像分割，提取出特定的判别特征，最后由设计好的分类器进行分类识别。其中，分类器的选择极大影响系统的识别准确率。

与模式识别领域传统的分类器，如最小距离分类器、最近邻分类器等技术相比，人工神经网络技术表现出一定的优势。即使对某问题领域本身了解甚少，只要能给出较多数量的样本数据，就可以训练出一个有效的计算模型，从而大大放宽了传统模式识别方法所需的约束条件[5]。因此，神经网络技术被广泛使用于医学图像分析识别系统中[1~3]，取得了丰富的研究成果。但是，如何进一步提高系统的泛化能力和分析识别准确率，仍然缺乏一个较为有效的方法。1990年，Hansen和Salamon[6]开创性地提出了一种方法，即神经网络集成(neural network ensemble)，为上述问题的解决提供了一个简易可行的方案。使用这种方法，可以简单地通过训练多个神经网络并将其结果进行合成，显著地提高学习系统的泛化能力和分析识别的准确率。

本文提出一种结合彩色图像特征提取和多级神经网络集成技术的肺癌图像分类识别方法。该方法使用图像处理技术分割出图像中的细胞区域，提取其形态特征和颜色特征，并将这两类相互独立的特征分别作为单个神经网络的输入向量；然后在识别阶段，把这两级神经网络以分层互联的方式集成，进行肺癌分类，从而有效提高了肺癌分类识别的准确率，同时也加快了肺癌分类识别的速度。该方法目前已被成功地应用于我们研制的肺癌早期细胞病理诊断系统LCDS中，其效果很好。

## 2　肺癌图像识别系统框架

LCDS系统的框架结构如图1所示，主要由图像预处理、图像分割、特征提取和神经网络集成和分类模块组成，其各模块的功能如下：

(1)图像预处理：对采集到的原始RGB肺癌彩色图像，应用一个定制的投影算法将其从三维的RGB色彩空间投影到一维线性的256级灰度空间；再利用双阈值快速分割方法对灰度图像作阈值分割，从而得到效果较好的二值图像。

(2)图像分割：在图像预处理基础上，对二值化图像进行形态学滤波，改善切片图像内细胞区域的几何形状。由于形态滤波可以在一定程度上消除图像采集及转换过程中可能产生的毛刺及小孔状噪音，因此，分割细胞区域的准确性得到了保证。

形态滤波完成后，系统利用基于区域边界的八链码[7]表示法对肺癌图像进行分割，标记出图像中的细胞区域，得到其链码表示。

(3)特征提取：使用细胞区域的链码表示对二值图像进行边缘跟踪计算，可以得到细胞区域的一系列几何形状和纹理特征，包括细胞区域的周长、面积、似圆度和矩形度等；然后，利用彩色切片图像数据对细胞区域的颜色直方图进行统计分析，获取其颜色特征。

(4)神经网络集成和分类：分别以细胞区域的形态特征和颜色特征作为单个神经网络的输入向量，送入集成的神经网络进行肺癌细胞的分类识别。系统根据各级神经网络的输出

*)本文工作得到国家自然科学基金(60273033)、江苏省自然科学基金重点项目(BK2001202)的资助。

进行集成，得出最终的诊断结果。

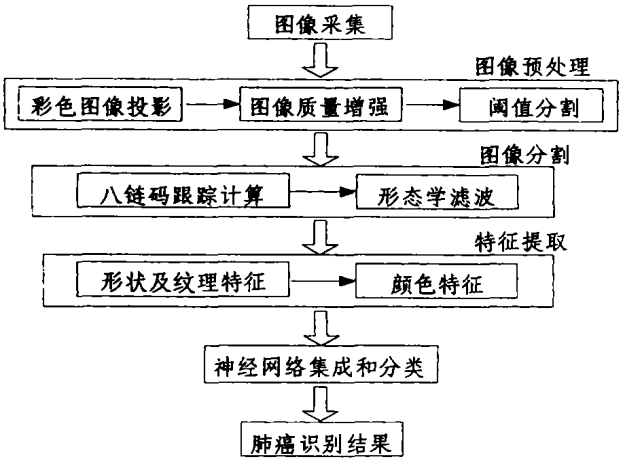


图1 LCDS 系统框架结构图

## 3 图像特征提取

在肺癌分类识别过程中，所依据的是肺癌细胞与正常细胞以及各类肺癌细胞之间在外形、纹理、颜色及其分布情况等多个方面的差异。通过对图像进行相应处理和计算后，提取出若干有代表性的图像特征分量来描述这些差异，并将其作为后续的肺癌分类器的输入特征。根据提取出的图像底层特征分量，训练好的分类器自动完成对图像中肺癌的分类识别，并得出最终结果。在详细介绍肺癌分类神经网络集成算法之前，先简单描述肺癌细胞的特征提取方法。

### 3.1 形状和纹理特征提取

LCDS 系统采用基于区域边界的八链码表示法[7]提取肺癌图像中细胞区域的形态和纹理特征，包括细胞面积、似圆度和像素密度等。对于图像中的像素点而言，均有8个方向的邻域，如图2所示。对每个方向赋以一种代码表示，图2中的8个方向分别对应于0，1，2，3，4，5，6，7，这些代码表示称为方向码。一条曲线的形状最终可以由下面的链码来表示：

$A_n=a_1a_2\cdots a_n \quad a_i\in\{0,1,2,\cdots,7\},i=1,2,\cdots,n$

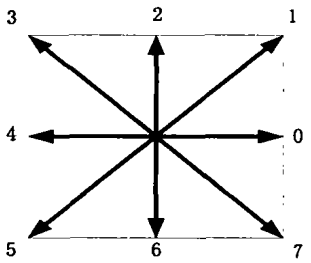


图2 八链码方向

这种链码表示既有利于有关形态特征的计算，也有利于节省存储空间，提高系统运行速度。用八链码对二值图像进行边缘跟踪可以得到计算以下的细胞区域几何形状和纹理特征，具体如下：

(1)链码所围细胞区域的周长 L

$$L=n_{even}+\sqrt{2}\cdot n_{odd} \tag{1}$$

周长 L 即链的长度。在公式(1)中，$n_{even}$表示链码中偶数码的个数，$n_{odd}$表示链码中奇数码的个数。

(2)细胞区域的宽度 W 和高度 H

定义方向码 $a_i$(i=0,…,7)在 X 轴与 Y 轴上的分量分别为 $a_{ix}$，$a_{iy}$，其取值如表1所示。

表1 $a_{ix}$与 $a_{iy}$的取值

| $a_i$ | $a_{ix}$ | $a_{iy}$ |
|---|---|---|
| 0 | 1 | 0 |
| 1 | 1 | 1 |
| 2 | 0 | 1 |
| 3 | −1 | 1 |
| 4 | −1 | 0 |
| 5 | −1 | −1 |
| 6 | 0 | −1 |
| 7 | 1 | −1 |

设 $x_0$与 $y_0$是起始点的坐标，则

$$\text{宽度 } w=\max_i\left(\sum_{k=1}^{i}a_{kx}+x_0\right)-\min_i\left(\sum_{k=1}^{i}a_{kx}+x_0\right) \tag{2}$$

$$\text{高度 } H=\max_i\left(\sum_{k=1}^{i}a_{ky}+y_0\right)-\min_i\left(\sum_{k=1}^{i}a_{ky}+y_0\right) \tag{3}$$

(3)细胞区域的面积

$$S=\max_{i=1}^{n}a_{ix}\left(y_{i-1}+\frac{1}{2}a_{iy}\right) \tag{4}$$

其中，n 为链码中方向码的个数，$y_i$ 使用迭代公式 $y_i=y_{i-1}+a_{iy}$计算。

(4)细胞区域的似圆度

似圆度用于描述细胞区域与标准圆形的偏离程度，是用于进行肺癌细胞分类识别的相当重要的一个量化指标。在相同面积的条件下，细胞区域边界光滑且为圆形，则周长最短，似圆度 C=1；否则，细胞区域形状越偏离圆形，则 C 值越小。其计算公式如下：

$$C=\frac{4\pi\cdot S}{L^2}=4\pi\cdot\frac{\text{面积}}{(\text{周长})^2} \tag{5}$$

(5)细胞区域的矩形度

$$R=\frac{S}{W*H} \tag{6}$$

矩形度用于描述细胞区域与矩形的偏离程度，用来作为衡量细胞区域异形性的指标。当细胞区域为矩形时，R 取最大值1。

(6)细胞区域的伸展度

$$E=\frac{\min(W,H)}{\max(W,H)} \tag{7}$$

伸展度特征用于度量细胞区域的狭长程度。细胞区域越呈细长形，则 E 越小；当细胞区域为圆形时伸展度具有最大值 E=1。

(7)细胞区域的均匀度

均匀度用于简单地衡量一个细胞区域的纹理特性，可以从一定程度上反映细胞区域中染色质颗粒的粗细及分布的均匀情况。其计算公式为：

$$u=\frac{N_f}{N_c}\cdot v \tag{8}$$

其中，$N_f$ 为一个细胞区域在对应的二值图像中的前景像素点个数，$N_c$ 为该细胞区域中的像素点总数，$v$ 为该细胞区域前景像素点集合的方差，按下式计算：

$$v=\frac{1}{N_f}\sum_{i=1}^{N_f}(x_i-\bar{x})^2 \tag{9}$$

由公式(8)和(9)可以推导出如下的均匀度计算公式：

$$u=\frac{1}{N_c}\sum_{i=1}^{N_f}(x_i-\bar{x})^2 \tag{10}$$

其中，$x_i\in[0,1]$，为该细胞区域前景像素点在对应灰度图像中的归一化灰度值，$\bar{x}$ 为其均值。

### 3.2 颜色特征提取

从病理专家的经验知识来看，细胞的染色特征在辨别癌细胞时起着非常重要的作用。单纯利用形态特征进行分类识别，可以找出绝大多数可疑的肺癌细胞，但同时也引起了大量的错误识别，例如将一些正常细胞和杂质错识为肺癌细胞，造成较高的假阳性率；同时，单靠形态特征还不能充分区分腺癌、鳞癌和小细胞癌这三种肺癌类型。因此，利用八链码跟踪计算分割出图像中的细胞区域后，结合其在原始肺癌彩色图像中的颜色空间分量值，可以提取用于肺癌分类识别的颜色特征。

为了保证颜色特征的有效性，首先需要确定适合于肺癌图像识别的颜色空间模型。由于不同颜色空间模型中的颜色分量所描述的色彩性质与肺癌诊断识别经验知识的吻合程度各不相同，有效的颜色特征分量可能分布于不同的颜色空间模型中。从实际效果来看，综合多种颜色特征识别要比单一颜色特征识别更符合人的视觉感受要求，因而识别效果更好。经过反复实验，LCDS 系统中提取的颜色特征如下：

(1){R,G,B}颜色空间　对肺癌图像中细胞区域，分别计算其包含的像素点集合在{R，G，B}颜色空间中的 R 分量、G 分量和 B 分量的平均值 $\bar{x}_R$，$\bar{x}_G$ 和 $\bar{x}_B$。由于平均值相当于颜色分量的一阶矩，因此它们可以看作是细胞区域的颜色矩特征，体现了细胞区域颜色值的统计特性。

(2){H,S,V}颜色空间　HSV 空间是一种符合人类视觉感知特征的颜色空间，特别适合于人类肉眼对颜色的识别，因此被广泛应用于计算机视觉领域[8]。它把彩色信号表示为三种属性：色调 H(Hue)、饱和度 S(Saturation)和亮度 V(Value)来表示。其中，色调 H 和饱和度 S 合起来定义了颜色的色度(Chromaticity)特性。因为 HSV 颜色空间根据色调 H 的值来区分不同的颜色，因此在 HSV 的三个分量中，H 分量是尤为重要的，它可以很好地模拟人类对颜色的识别和记忆过程。因此，应用 HSV 颜色模型更符合专家的肉眼诊断经验，可以获得更为接近人类观察的识别结果。

设 $V'=\max(R,G,B)$，定义 $R'$，$G'$，$B'$ 为：

$$R'=\frac{V'-R}{V'-\min(R,G,B)} \tag{11}$$

$$G'=\frac{V'-G}{V'-\min(R,G,B)} \tag{12}$$

$$B'=\frac{V'-B}{V'-\min(R,G,B)} \tag{13}$$

则从 RGB 空间向 HSV 空间的转换公式如下：

$$H=60*\begin{cases}5+B', & R=\max(R,G,B)\text{且}G=\min(R,G,B)\\ 1-G', & R=\max(R,G,B)\text{且}G\neq\min(R,G,B)\\ 1+R', & G=\max(R,G,B)\text{且}B=\min(R,G,B)\\ 3-B', & G=\max(R,G,B)\text{且}B\neq\min(R,G,B)\\ 3+G', & G=\max(R,G,B)\text{且}R=\min(R,G,B)\\ 5-R', & \text{其他}\end{cases} \tag{14}$$

$$S=1-\frac{3\min(R,G,B)}{R+G+B} \tag{15}$$

$$V=\frac{\max(R,G,B)}{255} \tag{16}$$

类似地，对肺癌图像中细胞区域，分别计算其包含的像素点集合在{H，S，V}颜色空间中的 H 分量、S 分量和 V 分量的平均值 $\bar{x}_H$，$\bar{x}_S$ 和 $\bar{x}_V$，作为分类识别的颜色特征。

(3)自定义颜色分量 C′　尽管{R,G,B}空间存在着各分量之间相关性强的缺点，但由于它是直接根据摄像镜头成像的特点定义的，因此很适合作为颜色识别的依据。考虑到偏蓝紫色是肺癌细胞的一般特性，而且三种不同的肺癌细胞其偏蓝紫色的程度有差别，因此，从{R，G，B}颜色空间中派生出一个新的颜色特征 C′，即 B 分量的比例值，其计算公式为：

$$C'=\frac{B}{R+G+B} \tag{17}$$

以上这些颜色特征和形态特征一起，共同组成了肺癌细胞区域的分类特征集合，作为神经网络分类器的输入向量。

## 4　肺癌分类神经网络集成

尽管使用单级神经网络分类器已经可以达到接近于实用的分类识别正确率，但如果要进一步提高则相当困难。而且，由于缺乏严密理论体系的指导，神经计算技术的应用效果完全取决于使用者的经验。虽然 Hornik 等[9]证明，只需一个具有单隐层的前馈网络就可以逼近任意复杂度的函数，但如何找到合适的网络配置却是一个 NP 问题。解决上述问题的一种行之有效的方法是把基于不同输入特征的神经网络进行集成以提高整个神经网络分类器的泛化能力。1990年，Hansen 和 Salamon[6]开创性地提出了神经网络集成(neural network ensemble)方法，使用这种方法，可以简单地通过训练多个神经网络并将其结果进行合成，显著地提高学习系统的泛化能力，其泛化能力优于单一的神经网络。自 Hansen 提出神经网络集成以来，很多研究人员对集成的理论基础进行了探讨，并将其应用到实际问题域中，取得了很好的效果[10,11]。大量研究结果表明，该方法不仅易于使用，还能以很小的计算代价显著提高泛化能力。因此，神经网络集成已成为目前神经网络界的研究热点。

在 LCDS 系统中，我们使用基于图像特征提取的神经网络集成技术进行肺癌细胞的分类识别，其实验和试用表明，效果很好。

### 4.1　神经网络集成结构

LCDS 系统中的神经网络识别部分由两个集成的神经网络结构所组成。根据专家识别肺癌细胞的流程，并保证各级神经网络输入特征之间的相互独立性，LCDS 系统的神经网络集成结构设计为分别以形状及纹理特征、颜色特征为输入特征的两级，其中，第一级为形状和纹理特征神经网络，第二级为颜色分量特征神经网络。

目前，在几十种神经网络模型中，使用最为广泛的是基于误差反向传播(Back-Propagation)算法的 BP 神经网络模型，其学习能力和容错能力强，适合于进行不确定性模式识别。LCDS 系统针对两种不同的输入特征，均建立了一个三层前馈型 BP 神经网络分类器，如图3所示。

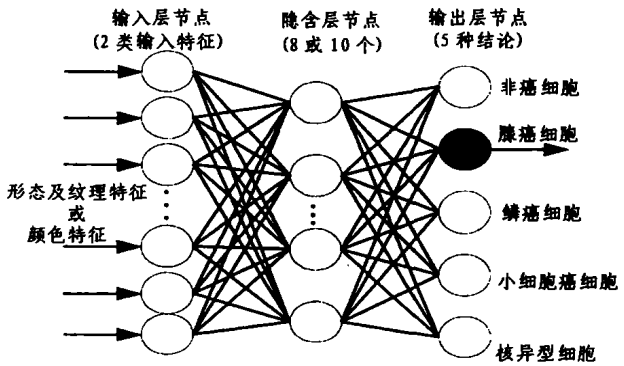


图3　LCDS 神经网络分类器

其中，每个 BP 神经网络以系统所提取的细胞区域的图像特征值为输入，输出为可能存在的细胞识别类型，包括非癌细胞、腺癌细胞、鳞癌细胞、小细胞癌细胞、核异型细胞5种，其中，腺癌、鳞癌和小细胞癌为肺癌的三种不同类型。隐含层节点根据实际经验设定，其中，第一级神经网络为8个，第二级神经网络为10个。

基于输入的图像特征分量，BP 神经网络模块在学习阶段对所有训练数据进行学习，而在分类识别阶段则对输入到本级的具体特征数据进行分类判别。每一个神经网络均基于一种特征分量，被第一级神经网络识别的细胞区域特征将被送给下一级神经网络，用颜色特征分量进行判别，由此，系统根据各级神经网络的输出进行集成，得出最终的肺癌识别结果。

### 4.2 LCNE 神经网络集成算法

目前，对神经网络集成实现方法的研究主要集中在两个方面，即怎样将多个神经网络的输出结论进行结合，以及如何生成集成中各网络的个体。

当以神经网络集成方法作为分类器时，集成的输出通常由各网络的输出投票产生。最后的输出要采用绝对多数投票法或相对多数投票法。理论分析和大量实验表明[6]，后者优于前者。而在生成集成中个体网络方面，最重要的技术是 Boosting[12]和 Bagging[13]。其中，在 Boosting 方法中，每一网络相应训练集的选取都基于整个集成网络中已产生的上级神经网络的表现。

LCDS 系统中神经网络集成算法 LCNE 的输出采用相对多数投票法，采用基于 Boosting 技术的训练集样本选取方法，其学习任务是利用 $N=2$ 个神经网络组成的集成对 $f:R^n \to R$ 进行近似。为了提高神经网络集成的整体性能，在第一级神经网络中，选用误识率相对较低的形状及纹理特征分量。

LCNE 神经网络集成算法步骤描述如下：

Step 1. 选定形态及纹理特征训练样本集合 $A_p^m=\{a_1^{(k)},\cdots,a_i^{(k)},\cdots,a_p^{(k)}\}_{k=1}^{m}$ 和颜色特征训练样本集合 $A_{q0}^{n_0}=\{a_1^{(k)},\cdots,a_i^{(k)},\cdots,a_p^{(k)}\}_{k=1}^{n_0}$，其中，m 为形状及纹理特征训练样本的总数，$n_0$ 为颜色特征训练样本总数，p 为每个样本中形态及纹理输入特征的种类数，q 为每个样本中颜色特征的种类数；

Step 2. 使用第一级神经网络对样本集合 $A_p^m$ 进行 BP 学习；

Step 3. 对第一级网络的每个训练样本 $A_i'(i=1,\cdots,m)$ 的输出结果进行校验，如果网络分类结果不正确，则将与该样本对应的颜色特征样本 $A_q'$ 加入二级网络的训练样本集合 $A_q^n$，即 $A_q^n=A_{q0}^{n_0}\cup\{A_q'\}$。其中，n 为生成的二级网络训练样本总数，q 为每个样本中颜色输入特征的种类数；

Step 4. 重复执行 Step 3，直至第一级神经网络训练完毕；

Step 5. 将得到的颜色特征训练样本集合 $A_q^n$ 使用第二级神经网络进行 BP 学习，直到训练完毕；

Step 6. 两级神经网络都训练完成后，可以集成各个网络的输出对肺癌图像进行自动分类识别。对提取出的细胞形态及纹理特征值，输入到第一级神经网络进行分类识别，如果识别为三种肺癌细胞类型之一，则执行 Step 7；否则，跳至 Step 9；

Step 7. 将该细胞对应的颜色特征输入到第二级神经网络继续进行分类识别。如果识别为三种肺癌细胞类型之一，则执行 Step 8；否则，跳至 Step 9；

Step 8. 神经网络输出的集成采用加权平均的方法。各网络分别被赋以权值 $w_\alpha(\alpha=1,2)$，并满足式(18)：

$$w_\alpha>0 \quad 且 \sum_\alpha w_\alpha=1 \tag{18}$$

设各网络的输入和输出分别为 $X_\alpha$ 和 $V^\alpha(X_\alpha)$，则按公式(19)计算神经网络集成输出：

$$\overline{V}(X)=\sum w_\alpha \cdot V^\alpha(X_\alpha) \tag{19}$$

根据神经网络的集成输出，系统得出最终的肺癌分类识别结果，并跳至 Step 10；

Step 9. 得出最终的肺癌分类识别结果为当前网络的分类识别结果；

Step 10. 输出最终的分类识别结果，分类识别完毕。

## 5 实验结果分析

对各级神经网络所使用的训练样本集包括1996～1999年间从中国人民解放军八一医院采集制作的550幅肺穿刺标本切片图像。这些训练样本图像的选取严格按照肺癌识别的5种结果在实际中的分布概率进行选取，其分布大致为：3种癌细胞比例为75%（其中，腺癌细胞占44%，鳞癌细胞占33%，小细胞癌细胞占23%），核异型细胞比例为10%，正常细胞比例为15%。其中，根据形状及纹理特征和颜色特征在对这些样本图像实际诊断中所起作用的大小，选取形状及纹理特征训练样本集合大小为321幅，颜色特征训练样本集合大小为229幅，即 $m=321$，$n_0=229$。

为了试验文中基于特征提取的神经网络集成方法的有效性，LCDS 系统在中国人民解放军八一医院进行了联合调试和试运行。在此期间，一共对该院1996～1999年的另外255幅肺穿刺标本彩色显微图像（采自119片显微镜切片）作为测试样本进行了分类识别。

以上述样本集合为输入，使用第一级和第二级神经网络进行训练学习的实验情况分析如表2所示。

表2 单级神经网络训练学习实验结果分析

| 网络级数 | 输入特征种类 | 训练样本数 | 训练轮数 | 测试样本识别正确率 | 送入下级网络的样本数 |
|---|---|---|---|---|---|
| 1 | 形状及纹理 | 321 | 2193 | 76% | 77 |
| 2 | 颜色 | 30 | 2879 | 68% | — |

从表2可以看出，第二级神经网络使用 Boosting 方法后其包含的训练样本数与第一级神经网络的训练样本数是基本一致的。但无论是单独使用何种特征，其识别正确率都相对偏低。

表3所示的是使用神经网络集成方法后 LCDS 系统对肺癌图像的分类识别结果分析。

表3 神经网络集成方法肺癌分类识别结果分析

| 切片年份 | 切片总数 | 图像总数 | 假阴性率(%) | 假阳性率(%) | 总错误率(%) |
|---|---|---|---|---|---|
| 1996 | 21 | 39 | 7.7 | 2.6 | 15.4 |
| 1997 | 26 | 65 | 6.2 | 3.1 | 13.8 |
| 1998 | 33 | 51 | 5.9 | 3.9 | 11.8 |
| 1999 | 39 | 100 | 5.0 | 2.0 | 9.0 |

其中，假阴性是指将肺癌细胞或核异型细胞错识为正常细胞的情况；假阳性则正好相反，指将正常细胞错识为肺癌细胞或核异型细胞的情况。由于在肺癌细胞中也存在不同肺癌

（下转第53页）

Data Bases, VLDB'96, Bombay, India, Sept. 1996
18 Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large databases. In: Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data, SIGMOD'98, Seattle, WA, June 1998. 73～84
19 Ester M, Kriegel H-P, Xu X. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In: Proc. 4th Int. Symp. Large Spatial Databases, SSD'95, Portland, ME, Aug. 1995. 67～82
20 Mehta M, Agrawal R, Rissanen J. SLIQ: A fast scalable classifier for data mining. In: Proc. 1996 Int. Conf. Extending Database Technology, EDBT'96, Avignon, France, Mar. 1996
21 Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. In: Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, SIGMOD'96, Montreal, Canada, June 1996. 103～114
22 Wang K, Zhou S, Han J. Pushing support constraints into association mining. In: Proc. 26th Int. Conf. Very Large Data Bases, VLDB'00, Cairo, Egypt, Sept. 2000
23 Pei J, Han J. Can We Push More Constraints into Frequent Pattern Mining? In: Proc. 2000 Int. Conf. Knowledge Discovery and Data Mining, KDD'00, Boston, MA, Aug. 2000
24 Lakshmanan L V S, Ng R, et al. Optimization of constrained frequent set queries with 2-variable constraints. In: Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data, SIGMOD'99, Philadelphia, PA, June 1999
25 Buttenfield B, Gahegan M, et al. Geospatial Data Mining and Knowledge Discovery. UCGIS White Paper on Emergent Research Themes, 2001
26 Ester M, Kriegel H-P, Sander J. Spatial data mining: A database approach. In: Proc. 5th Int. Symp. Large Spatial Databases, SSD'97, Berlin, Germany, July 1997. 47～66
27 Berchtold S, Bohm C, et al. Implementation of Multidimensional Index Structures for Knowledge Discovery in Relational Databases. In: Proc. 1999 Int. Conf. on Data Warehousing and Knowledge Discovery, DaWaK'99, 1999
28 Agrawal R, Shim K. Developing tightly-coupled data mining applications on a relational database system. In: Proc. 2th Int. Conf. Knowledge Discovery and Data Mining, KDD'96, Portland, OR, Aug. 1996
29 MacEachren A M, Wachowicz M, et al. Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualization with knowledge discovery in database methods. International Journal of Geographical Information Science, 1999, 13(5): 311～334
30 Miller H J, Han J. Discovering geographic knowledge in data-rich environments: [Report of NCGIA Varenius Workshop on Discovering geographic knowledge in data-rich environments]. Kirkland, Washington, March 1999
31 Abraham T, Roddick J F. Research issues in spatio-temporal knowledge discovery. In: Proc. SIGMOD'97 Workshop on Data Mining, Arizona, USA, 1997
32 May M. Spatial Knowledge Discovery: The SPIN! System: [GMD Technical Report]. 2000

(上接第42页)

种类之间的误识情况，例如将鳞癌细胞错识为腺癌细胞或小细胞癌细胞的情况，所以，在表3中，总错误率要高于假阴性率与假阳性率之和。

从表3可以看出，使用本文所提出的基于特征提取得神经网络集成方法后，对肺癌识别的准确率比单纯使用单级神经网络时有显著提高，其平均准确率在87.5%左右，假阳性率也得到了较为有效的抑制。而且，由于各级神经网络的输入针对的是某一类图像特征，其收敛速度也得到了提高，减少了训练代价。实验结果分析表明，对于肺癌图像分类识别，基于图像特征提取的神经网络集成方法是行之有效的，提高了识别准确率，达到了预期的效果。

**结论** LCDS系统在Windows 2000环境下，用C++编程实现。该系统可以辅助肺癌细胞病理专家提高对肺癌细胞的识别效率，特别是在缺乏细胞病理专家的医院，实用性很好。LCDS系统已在南京肿瘤医院、中国人民解放军八一医院等4家医院应用，其效果很好。目前，我们正从事利用神经网络集成方法抽取出用于肺癌分类识别的有效规则，既可以大大增加神经网络集成分类方法的可理解性，又有助于进一步提高识别准确率，提高系统的实际应用价值。本文的后续工作将致力于从神经网络集成中抽取肺癌分类识别规则的研究。

## 参考文献

1 Cox G S, Hoare F J, de Jager G. Experiments in lung cancer nodule detection using texture analysis and neural network classifiers. In: Third South African Workshop on Pattern Recognition, Nov. 1992. 136～142
2 Li X, Bhide S, Kabuka M R. Labeling of MRI Brain Images Using Boolean Neural Network. IEEE Transactions on Medical Imaging, 1996, 15: 628～638
3 Osareh A, Mirmehdi M, Thomas B, Markham R. Automatic Recognition of Exudative Maculopathy using Fuzzy C-Means Clustering and Neural Networks. In: Proc. Medical Image Understanding and Analysis Conf. BMVA Press, July 2001. 49～52
4 Shyu C, Brodley C E, Kak A, et al. Assert: a physician-in-the-loop content-based image retrieval system for hrct image databases. Computer Vision and Image Understanding, 1999, 74: 111～132
5 Zurada J M. Introduction to Artificial Neural Systems. West Publishing Company, 1992
6 Hansen L K, Salamon P. Neural Network Ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993～1001
7 Freeman. Computer Processing of Line-drawing Image. Computing Surveys, 1974, 6(1): 57～97
8 Miyahara M, Yasuhida Y. Mathematical Transform of (R, G, B) Color Data to Munsell(H, V, C) Color Data. Visual Communications and Image Processing'88, SPIE, 1988, 1001: 650～657
9 Hornik K M, Stinchcombe M, White H. Multilayer Feedforward Networks Are Universal Approximators. Neural Networks, 1989, 2(2): 359～366
10 Gutta S, Wechsler H. Face Recognition Using Hybrid Classifier Systems. In: IEEE Intl. Conf. on Neural Networks, NY: IEEE, 1996. 1017～1022
11 Shimshoni Y, Intrator N. Classification of Seismic Signals by Integrating Ensembles of Neural Networks. IEEE Transactions on Signal Processing, 1998, 46(5): 1194～1201
12 Schapire R. The Strength of Weak Learnability. Machine Learning, 1990, 5: 197～227
13 Breiman L. Bagging Predictors. Machine Learning, 1996, 24(2): 123～140