doi:10.3969/j.issn.1000-9760.2010.05.008

二维主成分分析在乳腺钼靶 X 线片钙化点感兴趣区域提取中的应用

张会如 马奎元 董 睿
(济宁医学院医学影像系,山东 济宁 272067)

摘 要 **目的** 针对乳腺钼靶 X 线影像,将基于二维主成分分析(Two-Dimensional Principal Component Analysis,2DPCA)的方法提取的图像特征用于乳腺感兴趣区域的自动提取,实现计算机辅助检测乳腺 X 线影像中微钙化点的前期预处理阶段。**方法** 对乳腺图像进行预处理,通过改进的 2DPCA 方法提取乳腺图像特征,利用边缘检测算法对乳腺图像进行边缘特征提取,最后利用神经网络分类器提取乳腺感兴趣区域。**结果** 实验结果表明该方法可以得到 95%的阳性检出率。**结论** 综合运用二维主成分分析方法、边缘特征提取方法和神经网络进行乳腺感兴趣区域提取,准确率更高。

关键词 二维主成分分析;神经网络;感兴趣区域
中图分类号:R737.9 **文献标志码**:A **文章编号**:1000-9760(2011)10-327-04

Extracting calcification ROIs in mammograms using 2DPCA

ZHANG Hui-ru, MA Kui-yuan, DONG Rui
(Department of Medical Imaging, Jining Medical University, Jining 272067, China)

Abstract: ***Objective*** In order to preprocess mammograms for diagnosing the early cases of breast cancer and realize the computer-aided detection of micro-calcifications in mammograms, this paper presented a method based on two-dimensional principal component analysis(2DPCA) to extract the region of interests(ROI) automatically. ***Methods*** First we preprocessed the mammograms, and then extracted mammography features by 2DPCA method and edge-detection algorithm. Finally, ROI was extracted by neural network classifier. ***Results*** The results showed that we obtained better positive detection ratio with this method. ***Conclusion*** Our method could obtain better extraction effect by integrating 2DPCA algorithm, edge-detection algorithm and neural network.

Key words: Two-Dimensional Principal Component Analysis; Neural network; Region of interests

乳腺感兴趣区域(Region of Interest,ROI)的提取是计算机辅助检测乳腺微钙化的前期预处理阶段,为了提高后续的微钙化点检测算法的处理速度、减小运算量,必须首先提取乳腺中含可疑微钙化点的区域。目前大多文献中通常使用人工方法确定 ROI,关于乳腺 ROI 的自动提取鲜有报道。但也有不少专家已经提出自动提取 ROI 的算法。Lee 等[1]提出基于块区域增长法与地毯覆盖法相结合的乳腺图像 ROI 自动提取算法,该方法获得了高达 90%的检出率。马振鹤[2]提出了基于传统图像处理和阈值化分类的乳腺感兴趣区域自动提取方法。王瑞平等[3]提出一种基于独立分量分析的 ROI 自动提取方法。彭镭[4]提出综合运用多种图像处理方法对乳腺图像中可能含有钙化点的感兴趣区域进行提取。上述自动提取 ROI 算法虽然得到较高的检出率,但还不能完全满足临床实际应用的要求。

主成分分析是一种常用的特征提取方法,它依据特征值的大小自动选择主成分,但是主成分分析算法对图像矩阵进行处理会带来很多不利的后果,Yang[5]提出了二维主成分分析(Two-dimension principal component analysis,2DPCA)方法提取图像特征,这种方法直接基于二维图像矩阵进行特征提取,与用主成分分析算法确定特征向量的时间相比,其所需的时间要小很多。本文将改进的 2DPCA 方法[6]用于乳腺图像的特征提取,并结合

利用图像的边缘特征训练神经网络分类器，提取乳腺感兴趣区域。

1 方法

1.1 获取训练样本

在有经验的放射科医生的指导下，对乳腺图像进行手工截取获得训练样本。具体方法为：如果钙化点区域的面积大于 128×128，则在此区域中切割出一幅 128×128 大小的图像；如果钙化点区域的面积小于 128×128，则直接选取包含钙化点的 128×128 大小区域作为一个 ROI 样本。针对乳腺背景区域直接截取 128×128 大小区域作为背景样本。

1.2 乳腺图像主成分特征提取

1.2.1 二维主成分分析算法 设 X 表示 n 维列向量，将 $m\times n$ 大小的图像矩阵 A 通过以下线性变换直接投影到 X 上：

$$Y=AX \tag{1}$$

得到一个 m 维列向量 Y，X 为投影轴，Y 称为图像 A 的投影特征向量。最佳投影轴 X 可以根据特征向量 Y 的散布情况来决定，采用的准则如下：

$$J(x)=tr(S_x) \tag{2}$$

其中 S_x 表示训练样本投影特征向量 Y 的协方差矩阵，$tr(S_x)$ 代表 S_x 的迹，当准则(2)式取得最大值时的物理意义是：找到一个将所有训练样本投影在上面的投影轴 X，使得投影后所得特征向量的总体散布矩阵(即样本类间散布矩阵)最大化。矩阵 S_x 可以记成下式：

$$\begin{aligned} tr(S_x)&=E(Y-EY)(Y-EY)^T\\ &=E[(A-EA)X][(A-EA)X]^T \end{aligned} \tag{3}$$

所以，

$$tr(S_x)=X^T[E(A-EA)^T(A-EA)]X \tag{4}$$

我们定义所有训练样本图像的总体散布矩阵(协方差矩阵)为：

$$G=E[(A-EA)^T(A-EA)] \tag{5}$$

从定义容易证明 G 是非负的，可以由原始的图像矩阵直接计算出来。假设共有 N 个训练样本，每个样本大小为 m×n，记作 A_j，所有样本的平均图像记作 A′，这样所有训练样本的协方差矩阵为：

$$G=\frac{1}{N}\sum_{j=1}^{N}(A_j-A')^T(A_j-A') \tag{6}$$

其中 $A'=\frac{1}{N}\sum_{j=1}^{N}A_j$ 为训练样本总体的均值矩阵。准则函数改写为：

$$J(X)=X^TGX \tag{7}$$

最大化该准则函数的单位向量 X 称为最优投影轴，即为协方差矩阵 G 的最大特征值所对应的单位特征向量。一般说来，单一的最优投影方向是不够的，需要寻找一组满足标准正交条件且极大化准则函数(7)的最优投影向量 $X_1,\cdots,X_d$。(引理[7] 最优投影向量组 $X_1,\cdots,X_d$ 可取为 G 的前 d 个最大特征值所对应的标准正交的特征向量。)

2DPCA 方法的特征提取，令 $P=[X_1,\cdots,X_d]$，称为最优投影矩阵，大小为 n×d；$B=[Y_1,\cdots,Y_d]$，称为样本图像 A 的特征矩阵或特征图像，$B=AP$。其中 $Y_j=AX_j(j=1,2,\cdots,d)$。

1.2.2 L-2DPCA 方法 由 2DPCA 算法可知，既然可以把图像的最佳投影轴进行右乘于图像矩阵，那么一定也可以进行左乘于图像矩阵，进行特征提取，这个算法叫做 L-2DPCA 算法，2DPCA 算法我们称为 R-2DPCA 算法。

协方差矩阵 可以用下式来表示：

$$\begin{aligned} S'_x&=E(Y'-EY')(Y'-EY')^T\\ &=E[(X'^TA-E(X'^TA)][(X'^TA-E(X'^TA)]^T\\ &=E[(X'^T(A-EA)][(X'^T(A-EA)]^T \end{aligned} \tag{8}$$

上式(8)中的 X′是新的投影轴，是 $m\times d$ 维的已经经过标准正交化后的矩阵，由上式得：

$$tr(S'_x)=X'^T[E(A-EA)(A-EA)^T]X' \tag{9}$$

那么图像的协方差矩阵为：

$$G'=E[(A-EA)(A-EA)^T] \tag{10}$$

和 2DPCA 算法一样，G' 的最大特征值所对应的特征向量就是最佳投影轴。单一的最优投影方向是不够的，需要寻找一组满足标准正交条件且极大化准则函数的最优投影向量组 $X'_1,\cdots,X'_e$，它可取为 G′的前 e 个最大特征值所对应的标准正交的特征向量。

L-2DPCA 的特征提取，令 $P'=[X'_1,\cdots,X'_e]$，称为最优投影矩阵，大小为 m×e；$B'=[Y'_1,\cdots,Y'_e]$称为样本图像 A 的特征矩阵或特征图像；图像 A 的特征矩阵为 $B'=(P')^TA$。

1.2.3 *LR-2DPCA* 方法 本文使用 *LR-2DPCA* 方法提取乳腺图像的主成分特征，它是上述两种方法的综合改进。方法如下：设乳腺图像 A 大小为 m×n，先使用 *2DPCA* 算法得到一个 n×d 维的特征投影矩阵 P，再采用 *L-2DPCA* 算法得到一个 m×e 维的特征投影矩阵 P′。用公式(11)可求得图

像 A 的特征矩阵：

$$Y=(P')^{T}AP \qquad (11)$$

特征矩阵 Y 作为图像 A 的主成分特征，用作后续神经网络分类器的输入，矩阵大小为 $e\times d$。这里取 $e=2,d=2$。这样每一幅乳腺图像的特征矩阵 Y 可以取 4 个数值，即输入神经网络的主成分特征为 Y_{2DPCA1}，Y_{2DPCA2}，Y_{2DPCA3}，Y_{2DPCA4}。

1.2.4 图像边缘特征提取　在乳腺样本图像中，微钙化点的边缘像素值普遍大于乳腺其它组织，并且微钙化点边缘都能被提取出来，因此，微钙化点边缘像素点的总个数和边缘像素最大值这两个特征可以作为对提取乳腺图像中微钙化点有贡献意义的特征输入神经网络分类器。本文利用文献[7]中边缘检测算法提取训练样本图像的边缘图像，并计算边缘像素点的总个数 N_i，根据样本的原始图像数据得到边缘图像中边缘像素的最大值 M_i；考虑到所得到的各个边缘特征取值范围各有不同，这不仅会使网络训练速度下降，而且容易造成网络训练失败，因此，需先对数据进行如下式的归一化处理：

$$N'_i=N_i/k_1 \qquad (12)$$

$$M'_i=M_i/k_2 \qquad (13)$$

其中，$k_1=\sum_{i=1}^{n}N_i$，$k_2=\sum_{i=1}^{n}M_i$，N'_i、M'_i 分别表示归一化后的边缘像素点的总个数和边缘像素最大值，i 表示输入样本序号，n 表示输入样本数量。

1.2.5 神经网络分类器　本文使用三层 BP 神经网络提取乳腺 ROI，如图 1 所示。网络输入层为 6 个节点，分别输入图像的主成分特征和边缘特征：$A_1=Y_{2DPCA1}$，$A_2=Y_{2DPCA2}$，$A_3=Y_{2DPCA3}$，$A_4=Y_{2DPCA4}$，$A_5=M'_b$，$A_6=N'_b$；输出层为 2 个节点：Y_1 和 Y_2；隐含层取 12 个节点。算法分为两个阶段：训练阶段和测试阶段。

1）训练阶段

对大小为 128×128 的训练样本图像，根据公式(11)提取主成份特征，根据边缘检测算法结合公式(12)和(13)提取图像边缘特征；将这些图像特征作为特征矢量输入神经网络输入层，经隐含层处理后传入输出层；若输出层未得到期望的输出结果，则进行误差的反向传播。网络根据反向传播的误差信号修改各层的连接权值，使网络预测与实际值之间的均方差最小。经过反复训练，使得网络输出是有分类能力的神经网络。

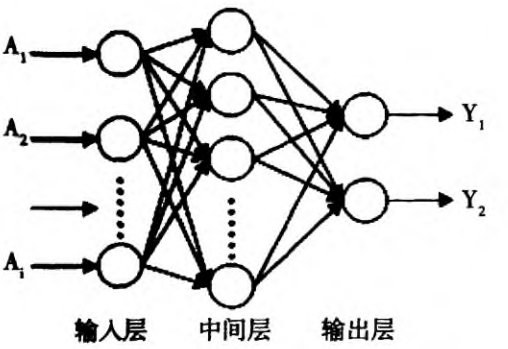

图 1　三层神经网络分类器

2）测试阶段将待检测的乳腺图像分成若干 128×128 大小的子区域；提取每个子区域图像的主成分特征和边缘特征，组成特征向量，输入训练后的神经网络进行分类，提取 ROI。

2　试验结果

所用数字乳腺 X 线图像来自山东省医学影像研究所，共 40 个乳腺癌病例，患者全部为女性，年龄 35～64，从这 40 个病例中选择了 101 幅乳腺钼靶 X 线影像进行试验。其中图像大小 1914×2294 像素，每像素 16 位存储，12 位位深，4096 灰度级，编程软件为微软 VC++ 6.0。

对 101 幅乳腺图像手工分割出乳腺区域，并从中选取 40 幅用于样本选取和神经网络训练，其它 61 幅用于神经网络测试提取 ROI。图 2 所示为 ROI 提取部分结果，图 2(a)是原始图像，(b)是提取的可疑病灶区域，(c)为(b)中方框所示感兴趣区域的放大图。

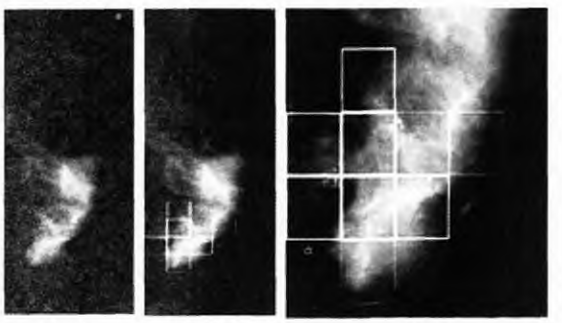

图 2　乳腺 ROI 提取结果

为了判断算法的优劣，使用目前国际上较为流行的 ROC 模型和 61 幅乳腺区域图像对本文 ROI 提取算法进行了测试。由医生或专家对 61 幅乳腺区域图像标记出感兴趣区域(含可疑病灶区域)，作

为 ROC 评价的金标准，使用本文算法对未标记的 61 幅乳腺区域图像进行 ROI 提取，将本文算法提取 ROI 的结果与金标准比较，作出 ROC 曲线，从图中可以看出，算法能在误检率为 14%时，得到 95%的阳性检出率。

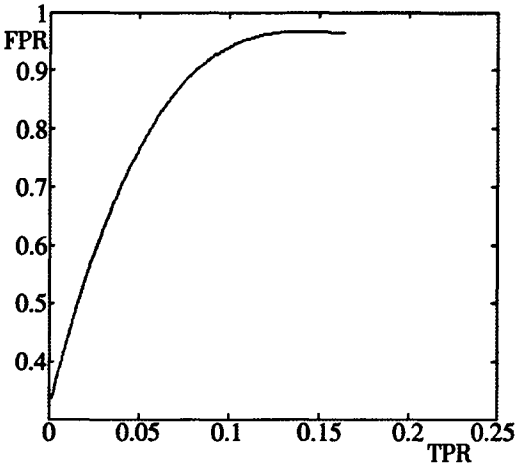

图 3 乳腺感兴趣区域提取 ROC 曲线

3 讨论与结论

本文提出了基于二维主成分分析特征提取方法用于乳腺感兴趣区域的提取。该方法在整个检测过程中，不需要人工参与，具有较高的智能性。由于二维主成分分析方法是直接基于图像矩阵的，因而大大减少了计算量，提高了神经网络训练的速度。本文综合了多种特征参数用于神经网络分类，这样取得了更好的分类效果。

从乳腺 X 线影像中提取含微钙化点区域，减小了后期乳癌计算机辅助诊断的运算量，提高了乳腺癌计算机辅助诊断系统的智能性，同时为医院数字化的发展奠定基础。

参考文献：

[1] Lee S K, Lo C S, Wang C M, et al. A computer-aided design mammography screening system for detection and classification of microcalcifications[J]. International Journal of Medical Informatics, 2000, 60: 29-57.

[2] Papadopoulos A, Fotiadis D I, Likas A. Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines[J]. Artificial Intelligence in Medicine, 2005, 34(2): 141-150.

[2] 马振鹤. 乳腺 X 线片中微钙化点感兴趣区域提取方法的研究[D]. 天津：天津大学，2003.

[3] 王瑞平. 独立分量分析在乳腺钼靶 X 片感兴趣区域提取中的应用[J]. 中国生物医学工程学报，2007, 26(4): 532-536.

[4] 彭镭，刘波峰，王家乐，等. 乳腺 X 射线数字影像中钙化点感兴趣区域提取方法[J]. 计算机系统应用，2010, 19(2): 83-85.

[5] Yang J, Zhang D, Frangi A F, et al. Two-dimensional PCA: a new approach to appearance-based face representation and recognition[J]. IEEE Trans, 2004, 26(1): 131-137.

[6] 张正. 直接基于二维图像的人脸识别技术研究[D]. 哈尔滨：哈尔滨工业大学，2006.

[7] Turk M, Pentland A. Eigenfaces for Recognition[J]. Cognitive Neuroscience, 1991, 3: 71-86.

[8] 张光玉，龚光珍，朱维乐. 基于克隆算法的彩色图像边缘检测新算法[J]. 电子学报，2006, 34(4): 702-707.

（收稿日期 2011-09-03）

·读者·作者·编者·

本刊对来稿中表、图的要求

来稿中的表、图均须置于正文中，切勿单独放于文后。每幅表、图应有言简意赅的题目。

统计表格一律采用“三线表”格式，不用纵线、斜线。要合理安排纵表的横标目，并将数据的含义表达清楚；若有合计或统计学处理行（如 F 值、P 值等），则在该行上面加一条分界横线；表内数据要求同一指标保留的小数位数相同。

图片应清晰，不宜过大。图的宽×高为 7cm×5cm，最大宽度半栏图不超过 7.5cm，通栏图不超过 17.0cm，高与宽的比例应掌握在 5∶7 左右。

本刊编辑部