文章编号 1004-924X(2006)02-0327-06

# 利用多分辨率直方图特征分类数字 X 光乳腺图像

刘欣悦[1,2],黄廉卿[1]
(1.中国科学院 长春光学精密机械与物理研究所,吉林 长春 130033;2.中国科学院 研究生院,北京 100039)

**摘要**:提出了一种结合多分辨率直方图特征表示与核学习算法的数字 X 光乳腺图像的分类方法。该方法不依赖特征选择步骤,而是基于感兴趣区(ROI)的高维多分辨率直方图特征,通过从训练实例中学习,同时检测多种异常的 ROI。对该方法进行接收器工作特性(ROC)分析,敏感性约为 89%,ROC 曲线下面积(AUC)接近 0.91。与以前所提出的检测方法相比,该方法不需要针对特定类型病变选择特征表示,因此可以同时检测多种类型的病变,简化了检测过程,提高了检测效率,而且分类性能也达到或超过了以前方法的平均分类性能。结果表明,利用多分辨率直方图特征表示能够很好地区分乳腺图像中正常和异常区域,同时也显示了借助核学习算法消除或限制分类任务中特征选择步骤的可能性。

**关 键 词**:模式分类;计算机辅助诊断;多分辨率直方图;核学习算法
**中图分类号**:TP391.4 **文献标识码**:A

# Classification of digital mammograms using multi-resolution histogram features

LIU Xin-yue[1,2], HUANG Lian-qing[1]

(1. *Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun* 130033, *China*; 2. *Graduate School of the Chinese Academy of Sciences, Beijing* 100039, *China*)

**Abstract**: A classification approach of digital mammograms using multi- resolution histogram representation in conjunction with kernel-based learning methods was presented. The approach didn't rely on the feature selection step and learned to classify various kinds of Region Of Interest (ROI) as normal/abnormal using its high-dimensional multi-resolution histogram features. Receiver Operating Characteristic (ROC) analysis of classification performance of the proposed approach shows that the sensitivity is about 89% and the Area Under Curve (AUC) is nearly 0.91. Compared to previous approaches, the proposed approach does not need to select abnormality-specific features so that it can detect various kinds of abnormalities simultaneously, which simplifies the detection process and improves the detection efficiency. The results demonstrate that multi-resolution histogram features can clearly distinguish the normal or abnormal classes in mammograms and the feature selection step of certain classification tasks can be eliminated or limited by using kernel-based learning method.

**Key words**: pattern classification; computer-aided diagnosis; multi-resolution histogram; kernel- based learning method

**收稿日期**:2005-11-14;**修订日期**:2006-01-22.

## 1 引　言

近年来开展了很多致力于开发自动图像分析方法的研究工作，目的在于辅助医生识别乳腺图像中的病变[1-2]。用于图像分类的特征或者由医生指定[3-4]，或者利用小波、分形理论、模糊集、Markov 模型以及统计技术等，通过图像处理的方法提取[5]。常用的分类器包括神经网络、线性判别分析以及决策树等[5]。

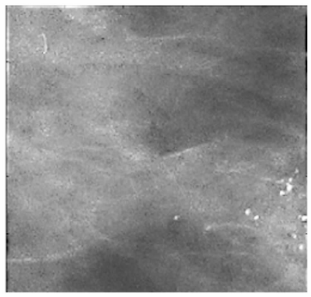
肿块和聚集微钙化是存在乳腺异常最常见的病变，如图 1 所示。这些病变在光学密度、形状、位置、尺寸和边缘特性等方面可能有很大的差异。由于病变在形状、尺寸以及细节方面具有的多样性，很难确定一组少量形态的、方向的或结构的特征，能够在任何尺度和形态下表征病变的特性。

对于计算机辅助诊断(CAD)系统，检测方法通常所使用的分类器性能受特征维数的影响很大，在同等条件下，随着特征维数的增长分类性能会严重恶化，因此各种检测方法的特征提取过程一般依赖特征选择的步骤，在特征生成步骤之后选取一组少量的最优特征表示病变的特性。但由于病变表现出的多样性，很难用少量特征有效建模各种类型的病变，因此各种方法都是针对单一类型病变进行检测，检测多种类型病变需要分别使用多种针对单一类型病变的检测方法，同时检测多种类型的病变对 CAD 系统是一项困难的任务。

本文提出了一种不依赖特征选择步骤的检测方法，能够同时检测多种类型病变。考虑到检测对象类型的复杂性，检测对象特性与周围环境的相似性，以及利用少量特征有效建模检测对象的难度，本文提出的方法不是从生成的特征中选择出少量最优特征表示检测对象，而是直接将生成的高维特征用于分类，通过从训练实例中学习检测异常的对象，所需要的只是用于训练的一组异常样本集合与正常样本集合。该方法将病变检测看成是一个两类模式分类任务，利用多分辨率直方图表示检测对象的特征，然后借助核学习算法处理提取大量信息。核学习算法是一类基于统计学习理论的技术，能够克服传统的学习技术在处理高维问题时所遇到的困难。正是由于核学习算法在高维特征空间很好的推广能力，消除或者限制分类任务的特征选择步骤才成为可能。核学习算法已被应用于检测特定类型的乳腺病变，而且性能明显优于之前公认的检测方法，如神经网络[6-8]。

(a)肿块

(a) mass



(b)聚集微钙化

(b) clustered micro-calcification

图 1　乳腺图像中的 ROI

Fig. 1　ROIs of mammograms

## 2 多分辨率直方图

灰度和彩色直方图广泛用于图像分类任务，与核学习算法结合应用于分类也获得了很好的结果[9]，但是简单直方图并不能提取图像空间变化的信息。对简单直方图进行扩展，通过计算图像在多个尺度下的直方图可以形成多分辨率直方图。多分辨率直方图具有简单直方图的许多优点，包括计算快速、空间效率高、对刚体运动的不变性、以及对噪声的健壮性。此外，多分辨率直方图还能够直接提取图像空间变化的信息。

图像空间变化的信息与 Fisher 信息测度直接相关，而 Fisher 信息测度是直方图密度变化率的加权平均值。多分辨率直方图可以变换为广义图像熵的向量表示，因此多分辨率直方图随图像

尺度的变化率可以变换为广义图像熵向量随图像尺度的变化率，后者是广义 Fisher 信息测度。广义 Fisher 信息测度是对图像锐度（空间变化）的非线性加权平均，由此可见多分辨率直方图能够提取图像空间变化的信息。详细介绍参见[10-11]。本文中，对 ROI 构造多分辨率直方图的过程如图 2 所示。

## 3　核分类算法

本文将正常和异常 ROI 分类看作一个两类模式分类问题。令 $x\in R^n$ 表示待分类的模式，标量 $y\in\{\pm 1\}$ 表示其类标识；此外，令 $\{x_i,y_i\},i=1,2,\cdots,N\}$ 表示训练样本集合。分类问题表述为如何确定能正确分类输入模式的决策函数 $f(x)$。下面简单介绍本文所使用的核分类算法。

### 3.1　支持向量机（SVM）

SVM 是一类源于统计学习理论[12]的构造性学习技术。基于结构风险最小化原则，SVM 的优化目标不是最小化训练集合上的均方误差，而是最小化推广误差限，因此 SVM 对训练集合外的样本具有很好的推广能力。


图 2　构造多分辨率直方图的过程

Fig. 2　Procedure of constructing the multi-resolution histogram

对于分类问题，SVM 首先将输入样本 $x$ 通过非线性映射 $\Phi(x)$ 映射到高维空间 $H$，然后在映射的特征空间进行线性分类。SVM 的分类函数可以表示为如下形式：

$$f_{\mathrm{SVM}}(x)=w^T\Phi(x)+b\ , \tag{1}$$

其中参数 $w$ 和 $b$ 通过在训练样本集合上最小化如下结构风险泛函确定：

$$J(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i$$
$$s.t.\quad y_i f_{\mathrm{SVM}}(x)\geqslant 1-\xi_i,\xi_i\geqslant 0;$$
$$(i=1,2,\cdots,N)\ , \tag{2}$$

其中 $C$ 是根据具体问题选择的正参数，$\xi_i$ 是松弛变量。通常使用 $w$ 的 $L_2$ 范式，式(2)是一个二次规划问题；如果使用 $w$ 的 $L_1$ 范式，式(2)就变成一个线性规划问题，相应的算法称为线性规划 SVM（LPSVM）[13]。

式(2)中的风险泛函是在经验风险（第二项训练误差）和模型复杂性（第一项）之间的折中考虑，参数 $C$ 控制折中程度。利用模型复杂性限制经验风险优化的目的在于避免过度拟合。过度拟合是指决策函数过于精确逼近于训练样本，而对训练集合之外的样本分类能力变差的现象。

满足 $y_i f_{\mathrm{SVM}}(x_i)\leqslant 1$ 的训练样本 $(x_i,y_i)$ 称为支持向量。引入核函数 $K(x,z)\equiv\Phi(x)^T\Phi(z)$ 可以将决策函数(1)改写为如下形式：

$$f_{\mathrm{SVM}}(x)=\sum_{i=1}^{N_s}\alpha_i K(x,\boldsymbol{s}_i)+b\ , \tag{3}$$

其中 $\boldsymbol{s}_i,i=1,2,\cdots,N_s$ 表示支持向量，通常支持向量只占训练样本集合的很小部分。

利用核函数 $K(\cdot,\cdot)$ 可以直接从式(3)中得到决策函数，而不需要考虑潜在的映射 $\Phi(\cdot)$。本文的学习算法使用 Gaussian 径向基函数（RBF）做为核函数：

$$K(x,z)=\exp(-\|x-z\|^2/2\sigma^2)\ , \tag{4}$$

其中 $\sigma>0$ 是控制核宽度的参数。

### 3.2　相关向量机（RVM）

RVM 是基于 Bayesian 估计的统计学习技术[14]，RVM 的重要特点在于得到的决策函数仅仅依赖于很少数量的训练样本（称为相关向量），因此可以得到更为稀疏的解。与 SVM 相比，RVM 在训练阶段不需要调整正则化参数 $C$。

对输入样本 $x$，RVM 利用 logistic 回归技术建模其类标识 $y\in\{\pm 1\}$：

$$p(y=1|x)=1/(1+\exp(-f_{\mathrm{RVM}}(x)))\ , \tag{5}$$

其中 $f_{\mathrm{RVM}}(x)$ 表示为

$$f_{\mathrm{RVM}}(x)=\sum_{i=1}^{n}\alpha_i K(x,x_i)\ , \tag{6}$$

其中参数 $\alpha_i$ 通过 Bayesian 估计确定。为确定 $\alpha_i$，首先在 $\alpha_i$ 上引入稀疏先验项。假定 $\alpha_i$ 之间统计独立，并且服从均值为 0 方差为 $\lambda_i^{-1}$ 的 Gaussian 分布。然后在方差 $\lambda_i^{-1}$ 上引入超先验项(基于 Gamma 分布)，目的在于强制 $\alpha_i$ 高度集中于 0 附近，从而使得 $f_{\mathrm{RVM}}(x)$只有很少的非 0 项。

通过在训练样本集合上最大化类标识的后验分布最终确定参数 $\alpha_i$，这等价于最大化如下目标函数：

$$J(\alpha)=\sum_{i=1}^{N}\log(y_i \mid x_i)+\sum_{i=1}^{N}\log(\alpha_i \mid \lambda_i^{*}) \quad (7)$$

其中第一项对应类标识似然项，第二项对应参数 $\alpha_i$ 先验项，$\lambda_i^{*}$ 表示超参数 $\lambda_i$ 的最大后验估计。在得到的解中，只有与非 0 系数 $\alpha_i$ 相关联的训练样本(相关向量)才会对决策函数 $f_{\mathrm{RVM}}(x)$产生影响。

## 4　实验结果

实验使用了 MIAS 图像库[15]，图像库中包括 322 幅经过临床验证的正常和几类异常图像。其中正常图像不包含病变，异常图像包含一处或者多处病变，病变类型包括聚集微钙化、多种类型的肿块等。从图像中手工提取出 1740 个 ROI(580 个异常样本/1160 个正常样本)，其中正常样本中只包含正常的组织，异常样本中除了包含完整的病变外也包含正常的组织。将样本随机分成两个部分分别用于训练和测试，每个部分所包括的异常样本数与正常样本数相等，如表 1 所示。

**表 1　实验中用于训练和测试的异常/正常样本数**

Tab. 1　Numbers of abnormal/normal samples used for training and testing in experiments

| | 异常样本数 | 正常样本数 | 总计 |
|---|---|---|---|
| 训练 | 290 | 580 | 870 |
| 测试 | 290 | 580 | 870 |

首先，将提取出的 ROI 经过均衡处理，构造 ROI 的多分辨率直方图。原始构造过程[10]生成几百维的特征直接用于匹配，由于应用方式的不同，本文方法对构造过程进行了修改，构造过程中使用了 4 个尺度等级，相邻尺度等级之间的欠采样因子为 2，最后形成 60 维的特征向量。

实验中分别利用 SVM、LPSVM、RVM 进行分类，三种分类器都使用 RBF 核函数。对每种分类器，在训练阶段利用 10 折交叉验证过程[16]进行模型选择，得到的最优参数在表 2 中列出。

**表 2　不同分类器的最优参数**

Tab. 2　Optimal parameters of different classifiers

| | SVM | LPSVM | RVM |
|---|---|---|---|
| 参数 | RBF 核 (sigma = 6) $C=50$ | RBF 核 (sigma = 6) $C=10$ | RBF 核 (sigma = 4) |

最后，通过接收器工作特性(ROC)分析评估分类性能。ROC 分析[17-18]给出分类器敏感性相对特异性的曲线，在很多分类任务中用于评估分类器性能。ROC 曲线是通过连续变化分类器决策函数的阈值，然后测量相应的敏感性和特异性得到的。作为总体分类性能的度量，ROC 曲线下面积(AUC)也被经常使用。不同核分类器的性能度量在表 3 中列出，图 3 是不同核分类器的 ROC 曲线，对应的 AUC 列于表 4。

**表 3　不同分类器的性能汇总**

Tab. 3　Performance summary of different classifiers

| | SVM | LPSVM | RVM |
|---|---|---|---|
| 准确度 (%) | 92.38 | 89.76 | 89.05 |
| 敏感性 (%) | 89.25 | 88.32 | 84.58 |
| 特异性 (%) | 90.78 | 91.26 | 93.69 |
| 正预测值 (%) | 90.95 | 91.30 | 93.30 |
| 负预测值 (%) | 89.05 | 88.26 | 85.40 |

注：准确度＝(TP＋TN)/(TP＋TN＋FP＋FN)，敏感性＝TP/(TP＋FN)，特异性＝TN/(TN＋FP)，正预测值＝TP/(TP＋FP)，负预测值＝TN/(TN＋FN)，TP＝真异常数，TN＝真正常数，FP＝假异常数，FN＝假正常数。

从 ROC 分析的结果可以看出，利用多分辨率直方图特征表示，三种核分类器对于本文的分类任务具有相似的性能。由于本文提出的方法用于同时检测多种类型的病变，与针对特定类型病变的检测方法[6-8, 19-20]很难进行客观的定量比较，但定性地利用综合的性能指标(敏感性、AUC 等)仍然可以说明本文方法的分类性能达到或者超过了以前方法的平均分类性能(针对特定类型病变方法的敏感性在 83%～95%之间，AUC 在0.83

～0.94 之间。这表明了多分辨率直方图在分类任务中的判别能力。

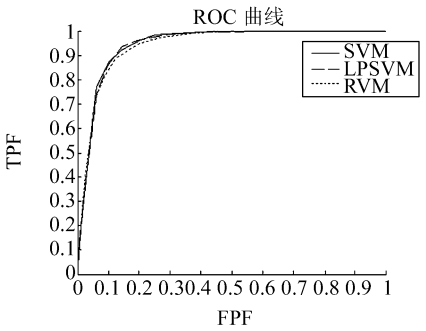



图 3　不同分类器的 ROC 曲线

Fig. 3　ROC curves of different classifiers

**表 4　不同分类器的 ROC 分析结果**

Tab. 4　ROC analysis results of different classifiers

| | SVM | LPSVM | RVM |
|---|---|---|---|
| 曲线下面积（AUC） | 0.9077 | 0.9059 | 0.8933 |

尽管实验中使用了高维特征以及相对较小的样本集合，但由于核学习算法在高维空间优异的推广能力，分类性能并未恶化。由于核学习算法的稀疏性，只有很少部分的训练样本对决策函数产生影响，SVM 和 LPSVM 的解中支持向量约占训练样本的 30%，而 RVM 的解中相关向量只占约 5%。这显示了借助核学习算法消除或限制分类任务中特征选择步骤的可能性。

## 5　结　论

本文提出了一种分类 X 光乳腺图像中 ROI 的方法，该方法通过构造 ROI 的多分辨率直方图特征表示，并基于此高维特征训练核分类器（SVM、LPSVM、RVM）检测图像中的异常。

与针对特定病变类型选择特征表示的检测方法相比，本文的方法不依赖特征选择步骤，能够同时检测多种类型的病变，而不需要分别使用不同的方法检测不同的病变。这大大简化了检测过程，提高了检测效率。ROC 分析的结果表明，该方法的分类性能达到或者超过了以前检测方法的平均分类性能。

与其他经过验证有效的特征相比，实验结果验证了多分辨率直方图特征同样具有很好的表示能力，而且在特征性质以及计算效率方面还具有一定优势。由于核学习算法在高维空间很好的推广能力，结合高维特征（相对以前的方法而言，以前的方法使用＜10 维的特征）与核学习算法仍可以获得满意的分类性能，这显示了消除或者限制分类任务中特征选择步骤的可能性。

**参考文献：**

[1] KARSSEMEIJER N, HENDRIKS J H. Computer- assisted reading of mammograms[J]. *Europe Radiology*, 1997, (7): 743-48.

[2] JIANG Y, NISHIKAWA R M, SCHMIDT RA, *et al*. Improving breast cancer diagnosis with computer-aided diagnosis[J]. *Academic Radiology*, 1999, 6, 22-33.

[3] Breast Imaging Reporting and Data System, Third Edition, American College of Radiology[Z]. 1998.

[4] BAKER J A, KORNGUTH P J, LO J Y, *et al*. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon[J]. *Radiology*, 1995, 196.

[5] SONKA M, FITZPATRICK J M. *Computer-aided diagnosis in mammography*[M]. Handbook of Medical Imaging. SPIE Press, 2000.

[6] El-NAQA I, YANG Y, WERNICK M N, *et al*. A support vector machine approach for detection of microcalcifications[J]. *IEEE Trans. Medical Imaging*, 2002, 21: 1552-63.

[7] CHANG R F, WU W J, MOON W K, *et al*. Support vector machines for diagnosis of breast tumors on US images [J]. *Academic Radiology*, 2003, 16: 189-197.

[8] WEI L Y, YANG Y Y, NISHIKAWA R M, *et al*. A Study on several machine-learning methods for classification of malignant and benign clustered microcalcifications[J]. *IEEE Trans. Medical Imaging*, 2005, 24: 371-80.

[9] CHAPELLE O, HAFFNER P, VAPNIK V. SVMs for histogram-based image classification[J]. *IEEE Trans. Neural Networks*, 1999, 10: 1055-1065.

[10] HADJIDEMETRIOU E, GROSSBERG M D, NAYAR S K. Multiresolution histogram and their use for recognition [J]. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2004, 26: 831-74.

[11] SPORRING J, WEICKERT. Information measures in scale-spaces[J]. *IEEE Trans. Information Theory*, 1999, 45: 1051-58.

[12] VAPNIK V. *Statistical learning theory*[M]. New York: Wiley, 1998.

[13] MANGASARIAN O L. *Generalized support vector machines*[M]. Advances in Large Margin Classifiers, MIT Press, 2000.

[14] TIPPING M E. Sparse Bayesian learning and the relevance vector machine[J]. *J. Machine Learning Research*, 2001, (1): 211-44.

[15] SUCKLING J, PARKER J, DANCE D R, *et al*. The mammographic image analysis society digital mammogram database[C]. *2nd International Workshop on Digital Mammography*, 375-78, *York*, *England*, *Elsevier*, 1994.

[16] MULLER K R, MIKA S, RATSCH G. An introduction to kernel-based learning algorithms[J]. *IEEE Trans. Neural Networks*, 2001, 12: 181-201.

[17] PEPE M S. Receiver operating characteristic methodology[J]. *J. American Statistical Association*, 2000, 95: 308-311.

[18] FAWCETT T. ROC graphs: notes and practical considerations for researchers[J]. *Technical Report*, *HP Laboratories*, *Palo Alto*, *CA*, *USA*, *April* 2004.

[19] CHU Y, LI L H, GOLDGOF D, *et al*. Classification of masses on mammograms using support vector machine[J]. *SPIE*, 2003, 5032: 940-948.

[20] SUN X J, QIANG W, SONG D S. Three-class classification in computer-aided diagnosis of breast cancer by support vector machine[J]. *SPIE*, 2004, 5370: 999-1007.

**作者简介**：刘欣悦(1973－)，男，辽宁大连人，博士研究生，主要研究方向为医学图像分析与机器学习。