基于主动支持向量机的乳腺癌微钙化簇检测

冯 筠[1,2] 姜 军[2] 叶豪盛[2] 王惠亚[3]
(西北大学信息技术学院 西安 710069)[1] (香港城市大学电脑科学系 香港)[2]
(西北大学数学系 西安 710069)[3]

摘 要 乳腺微钙化簇是早期乳腺癌的重要征象,计算机辅助的微钙化簇检测是医学影像领域的难题。为了提高检测系统的准确率,往往需要大量病灶标记,除了搜集样本本身的难度外,还需花费专家的大量时间。目前的研究工作很少涉及这个问题的解决方法。首次将基于主动学习的支持向量机技术应用到该领域,针对钙化簇感兴趣区域的特点,提出了选择训练集合的样本应该满足的基本条件。标准数据库上的实验证明,提出的方法能够大量地减轻样本标记的工作,并使乳腺癌微钙化簇检测系统的分类性能基本不变。

关键词 乳腺癌,计算机辅助检测,主动学习,支持向量机
中图法分类号 TP391 **文献标识码** A

Clustered Microcalcification Detection in Digital Mammograms Based on an Active Learning with Support Vector Machine

FENG Jun[1,2] JIANG Jun[2] Ip Ho-Shing Horace[2] WANG Hui-ya[3]
(College of Information Science and Technology, Northwest University, Xi'an 710069, China)[1]
(Department of Computer Science, City University of Hongkong, Hongkong, China)[2]
(Department of Mathematics, Northwest University, Xi'an 710069, China)[3]

Abstract Clustered microcalcification is an important signal for breast cancer in the early stages. However, computer aided detection of microcalcification is a challenge in the field of medical imaging. To improve the performance of the detection system, a large amount of lesion labeling is essential. Besides the difficulty on collecting samples itself, it also takes experts much time for manual labeling. Few state-of-the-art techniques take into account this problem. W first applied the techniques of active learning with SVM into this area to try to solve this problem. The basic conditions for the selected training set samples were proposed. The experiments on benchmark dataset show that our approach can reduce much works on labeling samples with holding the classification performance of the system of detecting interesting ROI regions.

Keywords Breast cancer, Computer aided detection, Active learning, Support vector machine

乳腺癌是妇女常见的恶性肿瘤,防治的关键在于早期诊断。微钙化点簇是早期乳腺癌的重要征象。乳腺钼靶X线片对钙化影像敏感、费用低廉,已成为诊断的重要工具。但即使是有经验的医生也很难迅速完全发现影像上早期乳腺癌的微小钙化点簇,以致延误病人的治疗时机。目前国内外已有市场化的乳腺计算机辅助检测系统[1,2],但费用昂贵,运算速度较慢,准确率有待提高。计算机辅助检测通常从原始图像开始,基本步骤为:(1)预处理,例如滤除背景噪声、增强图像对比度等。(2)感兴趣区域提取,利用信号处理或者图像分割的技术,找出疑似的病灶位置。一般情况下分割出的感兴趣区域(ROI)中含有大量的假阳性区域,所以需要对这些区域进行特征提取即步骤(3),亦即找出能够将病灶区域和正常组织区分开来的特征参数,例如形态学特征、灰度特征、几何特征以及纹理特征等,再根据特征参数进行模式分类即步骤(4),亦即减少第一步检测出的感兴趣区域中的假阳性区域。本文提出了基于主动支持向量机(SVM)乳腺癌微钙化簇检测方法,在分析了乳腺微钙化簇感兴趣区域特征空间的基础上,提出了适用于微钙化簇检测的基于半监督主动学习的SVM的模式分类策略。该方法的主要思想是将特征空间的先验知识融入到样本选择条件中,主动向专家反馈满足条件即需要标定的样本,从而达到标记少量典型训练样本就可得到最优分类器的目的。这对于标定工作费时、费力、费钱的乳腺癌微钙化簇检测领域具有积极的意义。实验证明,主动SVM在保持系统分类性能的同时,大量减少了标定样本的数目。由于有选择地使用样本,该方法还能达到有效筛选噪声点的效果。

1 文献回顾

由于乳腺图像中肿瘤的判定难有精确标准,在早期的自动化检测中,人工神经网络(Artificial Neural Network) ANN

到稿日期:2009-03-23 返修日期:2009-06-04 本文受陕西省教育厅科学研究计划基金(07JK381),中国博士后科学基金(20070421126)资助。
冯 筠(1972—),女,博士,副教授,CCF会员,主要研究领域为医学图像处理、三维重建、模式识别,E-mail:fengjun@nwu.edu.cn;姜 军(1971—),男,博士生,主要研究领域为机器学习和生物特征识别;叶豪盛(1959—),男,教授,主要研究领域为人工智能、模式识别;王惠亚(1980—),女,博士生,讲师,主要研究领域为统计决策理论、模式识别。

成为首选技术[3]；也有学者通过线性误差分类器来进行分类[4]；或者使用模板匹配技术来确定感兴趣区域内是否包含病灶区域[5]。然而，无论是人工神经网络、模板匹配还是线性分类方法，共同的理论基础是样本数目趋于无穷大时的渐进理论。对于只有有限样本集合的医学影像数据，理论上优秀的分类方法在实际应用中往往不尽人意。这些方法过于片面强调克服训练错误，因此得到的可能是局部最优解，而忽略了泛化性能的定量研究，产生的模型有时会产生过度拟合或拟合程度较差的现象。因此，寻找合适小样本的模式识别方法，成为乳腺疑似病灶区域模式分类的主要研究目标[6]。

支持向量机(Support Vector Machine) SVM 的分类目的是寻求泛化能力好的决策函数，即使由有限训练样本得到的决策规则对独立的测试集仍能够得到小的误差。对凸二次优化问题的求解也能够保证找到全局最优解。2002 年，Issam 等将经典 SVM 应用到乳腺 X 线照片的微钙化点检测[7]，他们用交叉检验法寻找最优参数，该算法得到了很好的分类效果。Bazzani 等将经典 SVM 中的错分因子的权值分为正类错分因子和负类错分因子，来处理微钙化点数据分类中的非平衡现象[8]。Papadopoulos 将 ANN 与 SVM 用于微钙化点检测并做对比，得出 SVM 较之具有更好的分类性能[9]。万柏坤等经过实验也得到同样的结论，并指出 SVM 具有更强的推广能力[26]。周伟达等专门对支持向量机的推广进行了分析，提出了可累积性学习的方法，对训练样本的选择很有指导意义[27]。

在传统的监督学习方法中，训练样本都是事先随机挑选出来的，然后进行标定，最后输入学习机进行训练，得到相应的分类器。对于 SVM 而言，由于其最后的分类器只依赖于少量的支持向量，这就意味着大部分样本的标定工作是无效的，因为它们对最终的分类器不产生任何影响。在乳腺癌检测系统中，标定工作需要组织高级放射科专家和医生进行反复辨识和手工圈定，费时、费力、费钱。为了解决这个问题，有学者提出了半监督学习方法[10]和主动学习方法[11]。主动学习方法的基本思路是，先随机挑选少量样本并标定，在训练标定好的样本集合上得到一个初始的分类器；然后根据训练好的分类器，以及标定和未标定样本的信息，从未标定的样本中按某种准则选取典型样本，提交给专家进行标定，标定之后再加入现有的训练样本集合；最后使用扩大的训练样本集合再训练；如此循环迭代，以达到用标定最少的样本得到最好的分类器的目的。其详细流程如图 1 所示。

可以看到，这种方法的关键在于如何设计样本选取准则，也就是如何定义评估函数 $EvalFun(x,L,h,H)$。早期的选取准则大都是基于统计的，譬如最小化方差(variance)[12]、偏置(bias)[13]和泛化误差(generalization error)[14]准则。虽然这些方法都有很强的理论基础，但由于必须计算样本的后验概率，因此在实际应用中受到了极大的限制。于是，一种基于版本空间(version space)理论的方法产生了，这种方法假设目标函数可以用版本空间的一个假设(hypothesis)来表达。这样，如果一个样本能最大地减少版本空间体积，它将被选择和标定，因为最小体积的版本空间意味着最小的泛化误差[15]。Query by Committee[16]和 SG 主动学习方法[17]就是基于上述理论的。但版本空间计算的复杂性限制了这类方法的应用。直到进入 21 世纪，这种理论与 SVM 方法结合，才使得它真正走向了实用。Tong 等[18]首次引进 SVM 到主动学习领域，提出了 SVM_{Active} 方法并应用于文本分类。但 SVM_{Active} 成批选择样本的效率不高，为此，一些学者引进了各种基于多样性(diversity)的策略，如基于角度(angle)[19]、内积(inner product)[20]和熵(entropy)[21]多样性的方法。详细的基于主动学习的 SVM 方法可以参考文献[22]。

初始化步：在初始标定的样本集合 L 上训练一个分类器 h

Step 1　学习机用函数 $EvalFun(x,L,h,H)$ 评估候选样本池(未标定样本集的子集或全部)中的每一个样本 x，挑选出具有最小值的样本 x^* 进行标定，并得到它的标记 y^*；

Step 2　学习机用扩大的训练集 $\{L+(x^*,y^*)\}$ 更新分类器 h；

Step 3　$L=\{L+(x^*,y^*)\}$，$Q=Q\backslash\{x^*\}$；

Step 4　重复 Step 1 到 Step 3，直到停止训练。

其中

$EvalFun(x,L,h,H)$：评估函数(此处假设最小最优)；

Q：待选择的未标定样本的集合；

L：当前标定好的样本集合；

H：假设空间集合是所有候选分类器的集合。

图 1　主动学习的框架图

2　基于主动学习的 SVM

如前所述，主动学习的关键是样本的选取准则。主动 SVM 选择样本的准则是基于最大化降低版本空间体积的原理。由于 SVM 的特殊性，使得主动 SVM 简单易行。本节在分析 SVM 的特殊性质的基础上，分别描述了单处理和批处理的主动 SVM 的方法，并针对本系统，对主动学习方法本身提出了一些新的改进。这里所谓单处理，就是每次只选择一个样本，批处理就是每次选择多个样本。

2.1　SVM 在版本空间的表示

从 20 世纪 90 年代发展到现在，SVM 有很多种变化，C-SVC[23]是最常用的 SVM 模型，它通过求解如下的最优化问题来得到分类器：

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C(\sum_{i=1}^{l}\xi_i)^k\\&s.t.\ y_i(w\cdot\Phi(x_i)+b)\geqslant 1-\xi_i\\&\xi_i\geqslant 0,i=1,\cdots,l\end{aligned}\tag{1}$$

其中，w 表示分类器 $c(x)=\mathrm{sgn}(w\cdot\Phi(x)+b)$ 的系数，$\frac{1}{2}\|w\|^2$ 表示结构风险，$(\sum_{i=1}^{l}\xi_i)^k$ 表示训练误差，l 是训练样本的数目，C 是这两项之间的平衡系数(trade-off parameter)。对于可分离的样本，上述模型可以简化为：

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}\|w\|^2\\&s.t.\ y_i(w\cdot\Phi(x_i)+b)\geqslant 1,i=1,\cdots,l\end{aligned}\tag{2}$$

$\Phi(x)$可以看作一种隐含的特征映射。在这个映射的高位特征空间中，分类器将变成线性的。$\frac{|w\cdot\Phi(x_i)+b|}{\|w\|}$表示点 $\Phi(x_i)$到分类面 $w\cdot\Phi(x)+b=0$ 的距离，因此当 $\|w\|$ 最小时，将得到最大的间隔(margin，它定义为$\frac{2}{\|w\|}$)，所以 SVM 又称作最大间隔分类器。当假设 $\|w\|=1$，模型(2)可以转换为[18]：

$$\max_{w\in W}\mathrm{margin}=\min_i\frac{\{|y_i(w\cdot\Phi(x_i)+b)|\}}{\|w\|}$$

$$= \min_i \{ y_i (w \cdot \Phi(x_i) + b) \}$$

$$s.t. \ \| w \| = 1 \tag{3}$$

$$y_i (w \cdot \Phi(x_i) + b) > 0, i = 1, \cdots, l$$

其中，W 表示参数空间，它和映射的高维特征空间 $F = \Phi(x)$ 有相同的维数。因为一个分类器对应参数空间的一个点 w，所以版本空间(所有符合限制条件点 w 的集合)可以表示为：

$$V = \{ w \in W \mid \| w \| = 1, y_i (w \cdot \Phi(x_i) + b) > 0, i = 1, \cdots, l \} \tag{4}$$

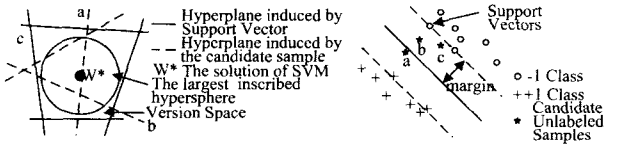
当采用高斯核函数 $k(x, y) = \exp(-\gamma \cdot \| x - y \|^2)$ 时，可以得到 $\| \Phi(x_i) \| = 1$，所以有：

$$\min_i \frac{\{ | (y_i \Phi(x_i)) \cdot w + y_i b | \}}{\| y_i \Phi(x_i) \|} = \min_i \{ | (y_i \Phi(x_i)) \cdot w + y_i b | \} = \min_i \{ y_i (w \cdot \Phi(x_i) + b) \} = \text{margin} \tag{5}$$

这意味着 SVM 的间隔也等于参数空间中的点 w 到超维面 $(y_i \Phi(x_i)) \cdot w + b = 0$ 的最小距离。所以，SVM 的模型可再次转换为：

$$\max_{w \in V} \text{margin} = \min_i \frac{\{ | (y_i \Phi(x_i)) \cdot w + y_i b | \}}{\| y_i \Phi(x_i) \|} \tag{6}$$

也就是说，SVM 的目的就是在版本空间内找一个最大的超维内切球，这个球的中心就是 SVM 的解，而球的直径就是分类器间隔的值(见图 2(a))。对于不可分的情况，可以加一个正则项校正因子，使得样本在新的映射空间变成可分的[24]。



(a) 在参数空间　　(b) 在映射的高维特征空间

(将选择样本"a")

图 2　基于距离的主动 SVM

2.2　基于距离的单处理主动 SVM

在映射的高维特征空间 F，SVM 分类器变成线性的。这就意味着在特征空间 F 和参数空间存在一种对偶性质：参数空间中的一个点对应特征空间的一个超平面，这个超平面将特征空间分成两部分；同样，特征空间中的一个点对应参数空间的一个超平面，这个超平面将参数空间分成两部分(如果这个超平面穿过版本空间，它也可将版本空间分成两部分)。Tong 和 Koller 首先提出如下的引理[18]。

引理 1　设存在一个有限维的输入空间 X 和参数空间 W，假设主动学习机 l^* 总是选择这样的样本，它在参数空间对应的超平面能够平分当前的版本空间，而 l 为其他的主动学习机。V_i^* 和 V_i 分别表示经过 i 次反馈迭代之后版本空间的体积，P 表示所有的 $p(y|x)$ 的集合，那么：

$$\forall i \in N^+ \sup_{P \in \mathcal{P}} E_P [Area(V_i^*)] \leqslant \sup_{P \in \mathcal{P}} E_P [Area(V_i)]$$

只要存在一个反馈迭代过程 j，l 选择的样本所对应的在参数空间的超平面不平分版本空间 V_{j-1}，那么不等式将严格存在。

他们同时认为在第 i 次选择的样本 x^* 应该满足如下的条件(将 $Area(V_{i+1})$ 简写为 V_{i+1})：

$$\min_{x \in Q} \sup_{p \in \mathcal{P}} [E_p (V_{i+1})]$$

其中，$E_p(V_{i+1})$ 表示在分布 $p(y|x)$ 条件下第 $(i+1)$ 次反馈迭代之后的期望体积，它可以通过下式进行计算：

$$E_p (V_{i+1}) = V_i^+ \cdot p(+1|x) + V_i^- \cdot p(-1|x) \tag{7}$$

其中，$p(+1|x)$ 和 $p(-1|x)$ 分别表示类标号 $y=1$ 和 $y=-1$ 的后验概率，V_i^+ 和 V_i^- 表示在第 i 次迭代过程中选择样本 x，并标记为 $+1$ 和 -1，然后再加入训练样本集后形成的新的版本空间的体积。所以 $\sup_{p \in \mathcal{P}} [E_p (V_{i+1})] = \max(V_i^+, V_i^-)$，因此所选择的样本 x^* 应该满足：

$$\min_{x \in Q} \max(V_i^+, V_i^-)$$

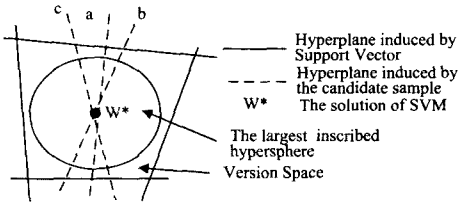
根据上述原则，所选择的样本应该尽可能去平分当前的版本空间。如果假设版本空间是对称的，那么结合对偶性和 SVM 的解是版本空间中最大的内切圆的中心两个性质，就可以得出应该选择距离当前决策面最近的样本。因为在所有的候选样本中，它最接近平分当前的版本空间(见图 2(b))，我们把这个方法称作基于距离的单处理主动 SVM。被选择的样本 x^* 应该满足 $\min_{x_i \in Q} | d(x_i) |$。

2.3　基于距离和多样性的批处理主动 SVM

当要同时选择多个样本时，文献[18]仅仅简单地选择那些最靠近当前决策面的样本。然而，加入这些样本不一定能保证最大限度地减少版本空间的体积。以图 2 为例，尽管每一个被选择的样本能够平分版本空间，但它们 3 个样本加在一起，也仅仅能够减少大约 1/2 的版本空间的体积，而不是 7/8。为了解决这个问题，文献[19]提出了一种新的基于角度多样性的方法，这也是目前最有效的方法之一。从图 2 可以观察到，3 个样本对应的超平面之间的角度小是造成它们效率低下的原因。设当前迭代中已被选择和标定的样本集合为 S，则新被选择的样本 x_q 应该满足如下条件：

$$\min_{x_i \in Q} \max_{x_j \in S} \frac{| k(x_j, x_i) |}{\sqrt{k(x_q, x_q) k(x_i, x_i)}} \tag{8}$$

其中，$k(.,.)$ 表示核函数，$\frac{| k(x_j, x_i) |}{\sqrt{k(x_j, x_j) k(x_i, x_i)}}$ 表示样本 x_j 和 x_i 对应的超平面之间角度的余弦值，因此叫它角度多样性准则。我们可以观测到，引入角度多样性后，同样多的样本能够减少更多的版本空间的体积，如图 3 所示，大概就能减少 7/8 的体积。



(样本"a","b","c"被选择反馈给用户)

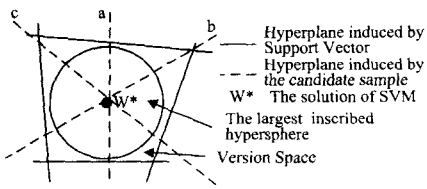
图 3　一个简单基于距离的批处理主动 SVM 的例子

在实际应用当中，应该同时考虑距离准则和角度多样性准则，所以新被选择样本 x_q 应满足的条件改为：

$$\min_{x_i \in Q \backslash S} ((1-\lambda) | d(x_i) | + \lambda \max_{x_j \in S} \frac{| k(x_j, x_i) |}{\sqrt{k(x_j, x_j) k(x_i, x_i)}}) \tag{9}$$

其中，$| d(x_i) |$ 表示样本 x_i 到当前分类决策面距离的绝对值；λ 是调节距离项和角度多样性项之间的权系数，它的取值范围为 $[0,1]$。由于真正微钙化簇区域的特征向量集合具有聚集性，所以我们采用高斯核函数，此时 $k(x_i, x_i) = 1$，因此上述准则可以简化为式(10)，如图 4 所示。

$$\min_{x_i \in Q \backslash S} ((1-\lambda) | d(x_i) | + \lambda \max_{x_j \in S} | k(x_j, x_i) |) \tag{10}$$



（样本“a”,“b”,“c”被选择反馈给用户）

图 4　一个基于角度多样性的批处理主动 SVM 的例子

然而，文献[19]的方法没有考虑到支持向量的多样性。在乳腺癌微钙化簇的检测中，发现经过感兴趣区域的提取步骤，大部分疑似 ROI 的特征向量在特征空间的位置比较靠近，给模式分类带来了很大的困难。实际上，我们并不希望被选择的样本距当前的支持向量太近，即新被选择样本对应的超平面和支持向量对应的超平面之间的距离不应该太小。因此，新被选择的样本 x_q 应该满足如下条件：

$$\min_{x_i \in Q \backslash S}((1-\lambda_1-\lambda_2)|d(x_i)|+\lambda_1 \max_{x_j \in S}|k(x_j,x_i)|+\lambda_2 \max_{x_j \in SV}|k(x_j,x_i)|) \tag{11}$$

其中，SV 表示当前的支持向量集合。根据核函数的性质，$k(x_j,x_i)$表示 x_i 和 x_j 映射到高维空间之后的两个点 $\Phi(x_i)$和 $\Phi(x_j)$之间的内积，而且对于高斯核函数，$\|\Phi(x)\|^2=1$。这意味着在高维映射空间中，两个点之间的内积越大，则它们的距离越小，所以用第 3 项表示支持向量的多样性。

这里还要提一个问题，它几乎被当前所有的主动学习方法所忽视，就是前面的算法都是假设训练样本是可分的。当然这可以通过调整正则项常数和核函数的参数来达到，然而在实际的应用中，为得到更好的泛化性能，总是允许少量错分样本的存在。错分样本不会分割版本空间，它们只会影响整个版本空间的空间位置。这样，在初始阶段，由于仅有少量样本，因此版本空间很容易漂移。而这只有靠引进更多的不受版本空间限制的点，即随机选择的样本来控制。因此，在本系统中，在主动学习的初始阶段，适当引入随机选择的样本，以保持版本空间的稳定性。

3　基于主动 SVM 的乳腺微钙化簇检测

根据上述主动 SVM 的工作原理，设计了一个乳腺癌微钙化簇检测系统（如图 5 所示）。在进行模式分类之前，首先进行了预处理、感兴趣区域提取和特征提取 3 个步骤。由于本文的重点是主动 SVM 模式分类，这里只对前 3 个步骤做简单描述，详细内容可参考文献[28]。

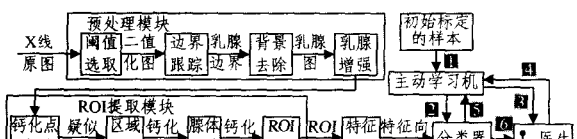


图 5　基于主动 SVM 的乳腺癌微钙化簇检测系统的工作流程

3.1　乳腺 X 线图像的预处理

乳腺钼靶 X 线摄影技术是目前为止早期发现和诊断乳腺癌的首选方法。在 X 线摄影过程中，不可避免地存在噪声点和杂质，并且乳腺中大多数软组织的 X 线影像特征上都具有近似的密度。预处理的目的就是提高图像质量，尽可能凸现出病灶区域，为后面的检测打下良好的基础。我们采用的预处理过程如图 5 左上部分所示。

3.2　感兴趣区域的提取

根据医学放射科对乳腺 X 线图像中的微钙化簇的定义，区域内含有 3 个或 3 个以上钙化点的称之为钙化簇，微钙化簇中的两个钙化点的距离小于 1mm。根据这两个先验知识，在预处理的二值化图像上计算各像素的贡献矩阵，并提取疑似钙化点。对每个检测出来的疑似钙化点进行计算，计算 2 * 2mm的邻域内疑似钙化点的数目，大于 3 个则认为该疑似钙化点成簇状分布。最后在保留邻域范围内疑似钙化点的基础上，去除大量的相对孤立的疑似钙化点和腺体。ROI 提取的过程如图 5 左中部分所示。

3.3　ROI 特征提取

选择能够有效地描述区域特性的影像特征，在 ROI 的模式分类中起着至关重要的作用。学者们曾经提出了许多有用的图像特征，这些特征主要有灰度特征、几何特征以及纹理特征。虽然这些特征都有着很好的数学理论基础，但还是无法完全涵盖医生所描述的所有关于病灶的信息。为了尽可能准确地用数学语言描述出肿块区域的特征，我们经过大量的实验选取了 63 个典型特征，如表 1 所列。

表 1　乳腺癌微钙化簇检测系统中采用的特征

特征类别	特征（维数）	
A. 灰度特征	1. 对比度；2. 灰度均值；3. 方差 4. 三阶矩；5. 四阶矩；6. 平均梯度；7. 区域边界的平均梯度；8. 不变矩（7 维）	
B. 几何特征	1. 圆形度；2. 紧缩度；3. 球状性；4. 傅立叶描述子	
C. 纹理特征	基于纹理能量图	1. 能量图均值；2. 能量图方差
	基于灰度共生矩阵	1. 能量；2. 熵；3. 对比度；4. 均匀性
	基于小波变换	1. 能量（9 维）；2. 熵（9 维）

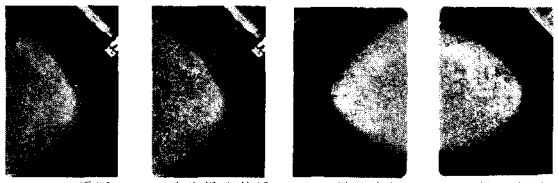
3.4　基于主动学习的疑似区域分类

在微钙化簇检测系统主动学习的过程中，系统的工作状态有两种：一种是在线学习状态，一种是脱机工作状态。在线学习状态时，系统只接受训练样本，含有以下几个过程：先将初始标定的样本输入主动学习机（如图 5 右半部分过程 1 所示）；主动学习机在当前标定好的训练样本集合上训练分类器（过程 2）；然后分类器检测所有未标定的样本，并将结果反馈给主动学习机（过程 5）；主动学习机根据一定的样本选择准则，从中挑选最优的样本，并反馈给医生（过程 3）；医生标定好样本之后，再返回样本给主动学习机（过程 4）；主动学习机接受医生反馈回的样本之后，将得到一个扩大的标定好的训练样本集合，重新回到过程 2。如此循环往复，直到得到满意的分类器。在脱机工作状态时，系统接受从病人而来的测试样本，分类器直接将分类的结果告诉医生（过程 6）。当然，在实际工作过程中，医生可以根据分类器的工作状况，通过过程 4 随时将样本反馈给主动学习机，要求重新训练分类器。

4　实验数据及结果分析

本文使用美国南佛罗里达州立大学 DDSM 数据库[25]，共测试了 239 幅图像。图 6(a)显示了一幅乳腺 X 光原图，图 6(b)显示了该图像乳腺边缘检测的结果。经过如 3.1 节—3.2 节所述的3 个步骤，一共提取出 1064 个 ROI，其中正样本 496 个、负样本 568 个。图 6(c)和图 6(d)显示了两幅 ROI 提取的结果实例。根据 3.3 节所述，共计抽取出 63 个特征。这些特征从某一特定角度描述了微钙化簇区域的特点，而且它们都

不受区域平移、旋转和尺度变化的影响。



(a)X-Ray 原图　(b)边缘提取结果　(c)ROI 提取实例 1 (d)ROI 提取实例 2

图 6　预处理和 ROI 区域提取实验结果

将样本随机分割，1/3 作测试样本（355 个）、2/3 作训练样本（709 个）。候选待选择样本的集合 Q 为整个训练样本集合。总共运行 10 次，每次训练的初始训练集由 20 个随机从 Q 选取的样本构成，其中至少要有一个正样本和一个负样本。采用批处理采样方式，每次选择 20 个样本。核函数采用高斯核函数 $k(x,y)=\exp(-\gamma\cdot\|x-y\|^2)$，$\gamma=0.03$，SVM 的平衡参数 C 取 100。通过计算 10 次运行的平均准确率、平均错误的正样本百分比（false positive）、平均错误的负样本百分比（false negative）来评估系统的性能。我们做了 3 批实验：(1)测试式(7)中样本选择准则中 λ 的影响（λ 分别等于 0.25，0.5，0.75，1.0，等价于将式(8)中 $\lambda_2=0$，而 λ_1 分别设为 0.25，0.5，0.75，1.0）（平均准确率结果见图 7）；(2)测试式(8)中的样本选择准则中 λ_2 的影响：（固定 $\lambda_1+\lambda_2$ 为 0.5，λ_2 分别为 0，0.1，0.2，0.3）（平均准确率结果见图 8）；(3)测试稳定版本空间策略的影响：在前 3 步中，每次引入 15 个随机选择的样本，而后面的步骤中，每次引入 3 个样本，其它参数与对比的方法一致；对比的方法采用式(7)中的选择准则，$\lambda=0.5$（平均准确率结果如图 9 所示）。

作为对比，用所有的 709 个训练样本，采用与主动学习过程一致的核函数和 C 去训练分类器，训练所得到的分类器的准确检测率为 80.56%。为观测方便，在表 2 详细列举了主动学习在第 2、5 和 10 次迭代后以及用全部训练样本训练所得到的分类器的结果。而如果用主动学习的方法，则在第 10 次迭代的时候，仅用 220 个数据，不到总训练样本的 1/3 准确检测率就接近了 80%，可见降低标定样本数目的效果是非常显著的。从表 2 中还可以看到，在主动学习的过程中，假阴比率（即错误的负样本百分比）得到迅速的降低，这在系统的实际应用中是有重要意义的。因为错误地判断假阴，会耽误病人的及时治疗。从图 9 和表 2 中可以观察到，在主动学习过程中引入稳定版本空间策略后，性能得到了显著的提高，尤其是初始阶段，这与 2.3 节的分析结果一致。而从图 7 和图 8 中以及表 2 中也可以看到，角度多样性和支持向量的多样性的效果都不是很明显。这可以从我们用整个训练集合训练得到的分类器得到解释，在这个分类器中，有很多支持向量的系数都为 100。也就是说，由于从 3.1 节—3.2 节中提取出的都是疑似 ROI 区域，在特征空间上比较聚集，而且训练集合中存在很多不可分的样本，因此这些都会降低角度多样性和支持向量多样性的效率。

图 10 是对 194 幅含有微钙化簇的图像进行测试得出的 FROC 曲线图。由 FROC 曲线可知，我们研制的微钙化簇 CAD 系统检出率最高可达 98%，平均每幅图像含有假阳性区域 2～3 个，达到或超过现有的乳腺 CAD 产品。

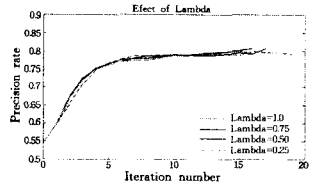


图 7　λ 对主动学习的影响

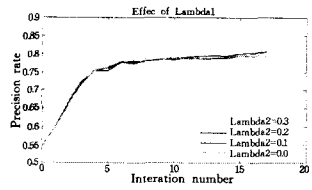


图 8　λ_2 对主动学习的影响（$\lambda_1+\lambda_2=0.5$）

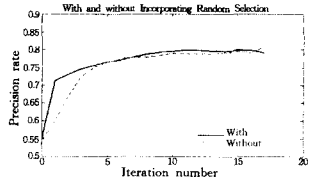


图 9　加入随机样本的影响

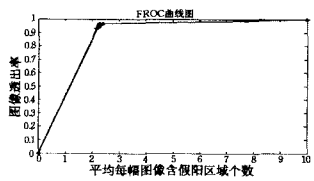


图 10　基于主动 SVM 乳腺微钙化簇检测结果 FROC 曲线图

表 2　主动支持向量机和一般支持向量机的对比

第一批实验	0.25	0	2	60	65.23	38.62	16.55
			5	120	76.64	27.02	16.32
			10	220	78.98	24.14	15.70
	0.5		2	60	66.67	37.51	15.61
			5	120	76.42	26.96	17.29
			10	220	79.01	23.88	16.20
	0.75		2	60	67.09	37.18	15.27
			5	120	76.11	27.91	15.66
			10	220	79.07	23.66	16.48
	1.0		2	60	66.81	37.41	15.39
			5	120	76.08	27.16	17.93
			10	220	78.76	24.04	16.61
第二批实验	0.5	0	2	60	66.67	37.51	15.61
			5	120	76.42	26.96	17.29
			10	220	79.01	23.88	16.20
	0.4	0.1	2	60	66.22	37.87	15.78
			5	120	75.38	28.29	17.28
			10	220	78.33	24.46	17.00
	0.3	0.2	2	60	66.14	37.96	15.53
			5	120	76.16	27.47	16.77
			10	220	78.59	24.06	17.08
	0.2	0.3	2	60	65.94	38.04	16.64
			5	120	76.02	27.97	15.82
			10	220	78.61	24.40	16.27
第三批实验	未引入稳定版本空间策略		2	60	66.67	37.51	15.61
	0.5	0	5	120	76.42	26.96	17.29
			10	220	79.01	23.88	16.20
	引入稳定版本空间策略		2	60	73.01	28.51	24.52
	0.5	0	5	120	76.61	24.66	21.48
			10	220	79.57	22.32	17.58
正常实验	$\gamma=0.03$，$C=100$			709	80.56	21.33	16.67

结束语　本文将主动学习的 SVM 算法首次引入到乳腺癌微钙化簇检测系统，在对现有的主动学习算法进行了详尽的分析之后，针对本微钙化簇疑似区域特征空间的特殊性，对当前的主动学习算法本身作了改进。通过实验发现，在保持系统性能基本不变的情况下，减少了大量需要标定的乳腺癌微钙化样本。本文提出的主动学习算法可以很容易推广到其他标定样本代价很高的领域。

（下转第 245 页）

结束语 本文在前人研究的基础上，提出了一类带交互时延和二次环境影响的 Swarm 模型，讨论了模型的稳定性问题，并得到了时延系统的稳定性及收敛的条件，此外具体考虑了在二次分布环境中时延 Swarm 的内聚性和其中心的运动情况。最后用仿真实验演示了结论的正确性并说明了在 Swarm 系统中由于时延的存在可能发生更复杂的动力学行为，其是否稳定或震荡，则主要取决于时延参数的取值范围。

参 考 文 献

[1] Reif J H, Wang Hongyan. Social potential fields : A distributed behavioral control for autonomous robots[J]. Robotics & Autonomous Systems, 1999, 27(3): 171

[2] Pachter M, Chandler P. Challenges of autonomous control[J]. Control Systems Magazine, IEEE, 1998, 18(4): 92-97

[3] Lawton J R T, Beard R W, Young B J. A decentralized approach to formation maneuvers[J]. IEEE Trans. Robot Automat., 2003, 19(6): 933-941

[4] Jin K, Liang P, Beni G. Stability of synchronized distributed control of discrete swarm structures[C] // Proc. IEEE Int. Conf. Robot. Automat. San Diego, 1994: 1033-1038

[5] Gazi V, Passino K M. Stability analysis of swarms[J]. IEEE Trans. Automat. Contr., 2003, 48(4): 692-697

[6] Gazi V, Passino K M. A class of attraction/repulsion functions for stable swarm aggregations[C] // Proceedings of the 41st IEEE Conference on Decision and Control. Los Vegas, Nevada USA, Dec. 2002: 2842-2847

[7] Gazi V, Passino K M. Stability analysis of social foraging swarm [J]. IEEE Trans. on Syst. Man, and Cybernetics-Part B: Cybernetics, 2004, 34(1): 539-557

[8] Liu Y, Passino K M, Polycarpou M M. Stability analysis of one-dimensional asynchronous swarms[J]. IEEE Trans. Automat. Contr., 2003, 48(2): 1848-1854

[9] Liu Y, Passino K M. Stable social foraging swarms in a noisy environment[J]. IEEE Trans. Automat. Contr., 2004, 49(1): 30-44

[10] Pedrami R, Gordon B W. Control and Cohesion of Energetic Swarms[C] // American Control Conference. Seattle, Washington, USA, June 2008: 129-134

(上接第 241 页)

参 考 文 献

[1] Astley S M, Gilbert F J. Computer-aided detection in mammography[J]. Clinical Radiology, 2004, 59(5): 390-399

[2] Cheng H D, Cai Xiao-peng, Chen Xiao-wei, et al. Computer-aided detection and classification of micro calcifications in mammograms: a survey[J]. Pattern Recognition, 2003, 36(12): 2967-2991

[3] Sahiner B, Chan H P, Petrick N, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images[J]. IEEE Trans. Med. Imag., 1996, 15(5): 598-610

[4] Wei D, Chan H P, Helvie M A, et al. Classification of mass and normal breast tissue on digital mammograms: multi resolution texture analysis[J]. Medical Physics, 1995, 22(5): 1501-1513

[5] Tourassi G D, Vargas-Voracek R, Catarious D M, et al. Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information[J]. Medical Physics, 2003, 30(8): 2123-2130

[6] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2: 121-167

[7] El-Naqa I, Yang Yong-yi, Wernick M N, et al. Support vector machine learning for detection of microcalcification in mammograms[J]. IEEE Transactions on Medical Imaging, 2002, 21(12): 1552-1563

[8] Bazzani A, Bevilacqua A, Bollini D, et al. Automatic detection of clustered micro calcifications using a combined method and an SVM classifier[A] // 5th International Workshop on Digital Mammography[C]. 2000: 161-167

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

[10] Chapelle O, Zien A, Scholkopf B, et al. Semi-supervised learning[M]. MIT Press, 2006

[11] Lewis D A, Gale W A. A Sequential Algorithm for Training Text Classifiers[A] // Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval[C]. 1994

[12] Cohn D A, Ghahramani Z, Jordan M I. Active Learning with Statistical Models[J]. Journal of Artificial Intelligence Research, 1996, 4: 129-145

[13] Cohn D A. Minimizing Statistical Bias with Queries [J]. Advances in Neural Information Processing Systems, 1997, 9

[14] Roy N, McCallum A. Toward Optimal Active Learning Through Sampling Estimation of Error Reduction[A] // Proceedings of 18th International Conference on Machine Learning[C]. 2001

[15] McAllester D A. Some PAC Bayesian Theorems[A] // Proceedings of the 11th Annual Conference on Computational Learning Theory[C]. Madison, Wisconsin, 1998

[16] Freund Y, Seung H S, Shamir E, et al. Selective Sampling Using the Query by Committee Algorithm[J]. Machine Learning, 1997, 28: 133-168

[17] Cohn D, Atlas L, Ladner R. Improving Generalization with Active Learning[J]. Machine Learning, 1994, 15: 201-221

[18] Tong S, Koller D. Support Vector Machine Active Learning with Application to Text Classification[J]. Journal of Machine Learning Research, 2001: 45-66

[19] Brinker K. Incorporating Diversity in Active Learning with Support Vector Machines[A] // International Conference on Machine Learning[C]. 2003

[20] Ferecatu M, Crucianu M, Boujemaa N. Reducing the redundancy in the selection of samples for SVM-based relevance feedback [R]. 2004

[21] Dagli C K, Rajaram S, Huang T S. Utilizing Information Theoretic Theoretic Diversity for SVM Active Learning[A] // International Conference on Pattern Recognition[C]. Hong Kong, 2006

[22] Jiang J, Ip H H S. Active Learning with SVM[M] // Rabuñal J R, Dorado J, Pazos A, eds. Encyclopedia of Artificial Intelligence: Information Science Reference, 2008

[23] Cortes C, Vapnik V. Support Vector Network [J]. Machine Learning, 1995, 20: 273-297

[24] Shave-Taylor J, Cristianini N. Further Results on the Margin Distribution[A] // Proceedings of the 12th Annual Conference on Computational Learning Theory[C]. 1999

[25] http://marathon.csee.usf.edu/Mammography/Database.html

[26] 万柏坤，王瑞平，朱欣，等. SVM 算法及其在乳腺 X 片微钙化点自动检测中的应用[J]. 电子学报，2004，34(4)：587-590

[27] 周伟达，张莉，焦李成. 支撑向量机推广能力分析[J]. 电子学报，2001，29(5)：590-594

[28] 王宇. 基于 SVM 的乳腺癌微钙化簇检测系统[D]. 西安：西安电子科技大学，2008