

# 基于迭代顺序滤波子空间约束的可拒识支持 向量机微钙化点检测

胡正平<sup>1,2</sup>, 吴燕<sup>1</sup>, 张晔<sup>2</sup>

(1. 燕山大学通信电子工程系, 河北秦皇岛 066004; 2. 哈尔滨工业大学信息工程系, 黑龙江哈尔滨 150001)

**摘要:** 为提高钙化点检测速度, 克服微钙化点检测中假阳性高的缺点, 本文构造了一种迭代顺序滤波子空间约束的可拒识-支持向量机分类器用于钙化点检测。训练时利用迭代顺序滤波检测作为钙化点的粗检测算子, 然后在其约束的子空间内收集非钙化点训练样本。对于输入模式, 首先利用基于最大软间隔超平面的支持向量分类器 (SVC) 进行分类判决; 然后对真实的钙化点样本特征空间求取最小的包含球形边界, 得到钙化点样本的球形支持向量域表示 (SVDD)。对于输入模式即可利用钙化点的支持向量域表示进行拒识或接受处理。仿真实验结果表明, 本文提出的算法在不影响微钙化点的检出率的情况下, 大大提高了检测速度, 部分解决了假阳性高的问题。

**关键词:** 支持向量分类器; 微钙化点检测; 支持向量域描述; 拒识性能; 迭代顺序滤波

**中图分类号:** TP301 **文献标识码:** A **文章编号:** 0372-2112 (2006) 02-0312-05

## Micro-calcifications Detection Based on Iterative Rank-order Filter Subspace and SVM with Rejection Feature

HU Zheng-ping<sup>1,2</sup>, WU Yan<sup>1</sup>, ZHANG Ye<sup>2</sup>

(1. Department of Communication and Electronic Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China;

2. Department of Information Engineering, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

**Abstract** To improve the speed and the problem of high false positive in micro-calcification detection, a novel micro-calcification detection method based on support vector classifier model with rejection feature and Iterative Rank-order Filter Subspace is proposed. In the training step, iterative rank-order filter is used as coarse detector and the non-micro-calcification training samples are selected from the constrained subspace. So it overcomes the effect of number difference between the negative and positive samples and at the same time it reduces the computing load in training stage. In the detection step, coarse detection is applied at first to find suspect micro-calcification region, then filters the first layer of support vector classifier (SVC) with maximum margin between two classes will be used for classification. The input pattern and the sphere support vectors of true micro-calcification points describing the distribution of the sample are obtained by searching all the sphere boundaries containing the samples of this class. So the input pattern of non-object classes can be rejected by the second support vector domain description (SVDD). Lastly the results of SVC and SVDD classifier are integrated to obtain the accurate results. Experimental results demonstrate that the new calcification detection method performs better in achieving lower false positive rate (FPR) and fast speed.

**Key words** support vector classifier; micro-calcification detection; support vector data description; rejection feature; iterative rank-order filters

### 1 引言

不断提高乳腺图像的微钙化点正确检测率是解决乳

腺癌诊断的关键, 为此成为众多学者追求的目标。微钙化点检测与识别就是从乳腺图像中抽取微钙化特征向量, 并在此基础上构建合适的分类器。常用的人工神经网络、贝

叶斯分类器等基于传统统计学的方法, 在样本数量有限的情况下表现不尽人意. 基于统计学习理论的支持向量分类器 (Support Vector Classifier, SVC) 因为其优越的性能成为近年研究的热点, 它在逼近与分类应用领域取得了极大的成功. SVC 算法是 Vapnik<sup>[1]</sup> 提出的一种有监督学习结构风险最小化的两类分类器. 一些学者<sup>[2,3]</sup> 将 SVC 引入微钙化点检测中并改进了检测效果. 在乳腺图像中钙化点主要表现为一系列较周围正常组织亮的小区域, 由于其半径小, 常常淹没于背景组织和噪声中, 这是导致假阳性高的原因. 经典的 SVC 分类器总是假定测试样本属于两类之一, 然而在钙化点分类判决问题中, 乳腺图像中往往存在许多种类的非目标样本, 这时经典 SVM 分类器缺乏拒识判决的能力, 所以不能很好解决假阳性高的问题. 为解决这个问题, 本文提出一种可拒识-双层支持向量机分类器用于钙化点检测, 它能够充分利用已知钙化点训练样本分布信息, 提高拒识非目标样本的能力. David M. J. Tax<sup>[4]</sup> 等针对单类数据提出了支持向量域描述 (Support Vector Domain Description, SVDD), 主要用来进行数据描述和剔除奇异点. SVDD 的基本思想就是计算包含一组数据的最小超球体边界来对数据的分布范围进行描述<sup>[4]</sup>. 该方法给我们的启发是: 如果能够知道一部分样本的分布范围, 就可以利用它构造可拒识支持向量分类器. SVDD 具有良好的数据选择能力, 但是缺乏对比鉴别分类能力 (因为它仅仅针对一类数据本身); 而 SVC 缺乏数据分布描述能力, 却具有最优的对比鉴别能力. 同时吸取二者的优点构造一种新的具有良好拒识能力的分类器是解决微钙化点检测假阳性问题的关键. 如果直接使用 SVC 解决微钙化点检测, 主要有两个问题: (1) 非钙化点样本数目不受限制时, 导致训练样本数量庞大; (2) 支持向量数目巨大导致检测时计算量巨大. 针对该问题, 本文提出将迭代顺序滤波与可拒识-双层支持向量机相结合的微钙化点检测方法. 该方法将迭代顺序滤波作为粗检测算子, 在粗检测算子限定的子空间内收集非钙化点样本, 训练第一层 SVM, 有效地解决了上面提到的两个问题, 并使得检测速度比单独的 SVM 方法快 5 倍左右.

针对可拒识分类问题, 不少学者也提出了有效的解决问题的思路. 文献 [5] 提出基于 (Support vector representation and discrimination machine) 支持向量表示鉴别机 (SVRDM) 的具有拒识能力的分类方法, 该方法的优点是一步实现可拒识分类器, 缺点是控制拒绝能力的门限参数选择比较复杂, 不具有自适应能力. 文献 [6] 提出另外一种类似于本文的两步实现策略, 首先将两类归于一类, 第一步决定是否拒绝, 第二步进行分类. 该方法缺点是第一步需要知道非目标样本并进行训练, 一方面计算量巨大, 另一方面在实际应用中往往不容易收集到全部的非目标样本. 而本文提出的方法不需要全部非目标样本, 只需要知道部分两类目标样本本身. 文献 [7] 提出了基于模板等

预处理结合 SVC 的分类方法, 预处理的过程就是数据粗选择过程, 该方法对于可模板化描述的问题是合理有效的, 缺陷是对于不能模板化的数据无能为力, 同时门限选择困难.

## 2 系统框图

本文构造的可拒识-双层支持向量机分类器如图 1 所示, 该系统主要由两个部分组成: 基于最大间隔的超平面鉴别分类器和基于 SVDD 的数据选择器, 各自的支持向量通过样本训练而得到.

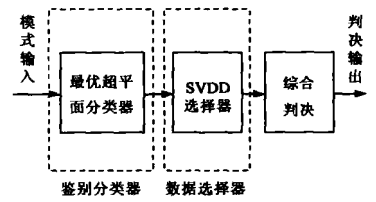


图 1 可拒识-双层支持向量模型分类器框图

本文构造的微钙化点训练与检测原理如图 2、3 所示. 首先从生成各种复杂度背景下的训练正样本训练图片由专家指导标定微钙化点正样本, 并形成训练 SVM 的正样本集合. 在迭代顺序滤波限定的子空间内收集非钙化点样本, 与前面的钙化点样本一起通过 (Sequential minimal optimization) SMO 算法训练第一层 SVC 分类器, 同时利用标定的钙化点样本训练第二层的 SVDD 数据选择器. 检测时, 首先对输入图像进行去噪等预处理操作, 然后分别利用迭代顺序滤波、CAFER 检测可疑钙化点的感兴趣区域, 最后利用 SVM 与 SVDD 进行综合判决.

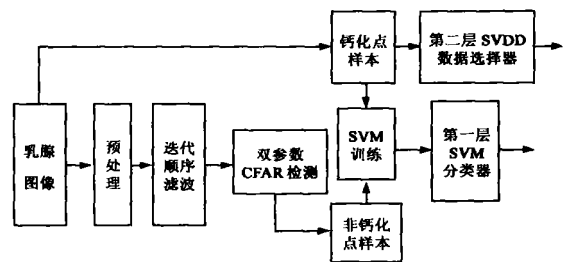


图 2 钙化点训练可拒识-双层支持向量分类器框图

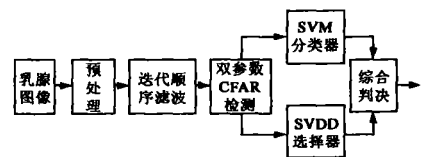


图 3 钙化点检测原理框图

## 3 可拒识-双层支持向量分类器

### 3.1 支持向量域数据选择器

SVDD<sup>[4]</sup> 通过寻找包含一组数据的最小半径的超球体对其进行描述, 可见它可用于对一组数据的分布进行描述. 对于奇异的数据点, 它应该位于超球体的外面. 为了减少奇异点数据的影响, 引入松弛变量  $\xi$  进行描述 (即允许存在错误), 求最小超球体体积可以转化为下面的二次规

划问题:

$$m \min [R^2 + C \sum_i \xi_i] \quad (1)$$

其约束条件为

$$(x_i - a)^T (x_i - a) \leq R^2 + \xi_i \quad (2)$$

其中  $\xi_i \geq 0$

上面的优化问题通过引入拉格朗日系数变为

$$m \min L = R^2 + C \sum_i \xi_i - \beta_i \sum_i \xi_i \sum_i \alpha_i (R^2 + \xi_i - (x_i - a)^T (x_i - a)) \quad (3)$$

其中  $\alpha_i, \beta_i \geq 0$

求式(3)的最小值可转化为如下的对偶问题

$$m \max F = \sum_i \alpha_i \langle x_i, x_i \rangle - \sum_i \sum_j \alpha_i \alpha_j \langle x_i, x_j \rangle \quad (4)$$

其中  $\sum_i \alpha_i = 1$   $0 \leq \alpha_i \leq C$ , 通常情况下, 即使去除奇异点数据, 数据依然不会呈现球状分布, 这时可以通过引入核函数将它映射到更高维空间进行讨论.

上面的 SVDD 数据选择器仅仅考虑一类样本, 如果利用到不同类别样本的信息, 其基本思想就是寻找包含本类数据的最小半径的超球体, 同时要求其它类样本位于超球体外. 其数据选择器的二次规划优化问题形式变为:

$$m \min [R^2 + C_1 \sum_i \xi_i + C_2 \sum_j \xi_j] \quad (5)$$

约束条件为:

$$\begin{aligned} (x_i - a)^T (x_i - a) &\leq R^2 + \xi_i, & (\text{本类样本}) \\ (x_j - a)^T (x_j - a) &\geq R^2 - \xi_j, & (\text{其他类样本}) \end{aligned}$$

其中  $\xi_i \geq 0, \xi_j \geq 0 \forall i, j$

通过引入拉格朗日系数, 上面的优化问题变为:

$$\begin{aligned} L = R^2 + C_1 \sum_i \xi_i + C_2 \sum_j \xi_j - \sum_i \gamma_i \xi_i - \sum_j \gamma_j \xi_j \\ - \sum_i \alpha_i [R^2 + \xi_i - (x_i - a)^T (x_i - a)] - \sum_j \alpha_j [(x_j - a)^T (x_j - a) - R^2 + \xi_j] \end{aligned} \quad (6)$$

令 L 对 R, a,  $\xi_i, \xi_j$  求导并等于 0 得到

$$\begin{aligned} \sum_i \alpha_i - \sum_j \alpha_j &= 1 \\ a = \sum_i \alpha_i x_i - \sum_j \alpha_j x_j \end{aligned}$$

其对偶形式为

$$\begin{aligned} L = \sum_i \alpha_i \langle x_i, x_i \rangle - \sum_i \alpha_i \langle x_i, x_i \rangle - \sum_{ij} \alpha_i \alpha_j \langle x_i, x_j \rangle \\ + 2 \sum_{ij} \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{lm} \alpha_l \alpha_m \langle x_l, x_m \rangle \end{aligned} \quad (7)$$

令  $\alpha'_i = \gamma_i \alpha_i$ , 则  $\sum_i \alpha'_i = 1, a = \sum_i \alpha'_i x_i$ , 式(7)可以简化为与式(4)相似的形式

$$m \max F = \sum_i \alpha'_i \langle x_i, x_i \rangle - \sum_i \sum_j \alpha'_i \alpha'_j \langle x_i, x_j \rangle \quad (8)$$

对于两类问题(也可扩展到多类问题中)的情况, 设有训练样本  $(x_i, y_i), i = 1, 2, \dots, n, y_i \in \{-1, +1\}$ , 分别对于  $y_i = 1$  和  $y_i = -1$  的样本建立超球体, 通过求解式(4)和式(8)分别计算其拉格朗日系数对应的支持向量.

如果一个测试钙化点样本 z 满足如下条件之一, 则这个

点被接受, 否则拒识判决.

$$K(z, z) - 2 \sum_i \alpha_i^{(1)} K(x_i, z) + \sum_{ij} \alpha_i^{(1)} \alpha_j^{(1)} K(x_i, x_j) \leq R_1^2$$

(钙化点数据选择器)

### 3.2 最优超平面鉴别分类器

设输入模式  $x_i$  为 p 维特征矢量,  $y_i \in \{+1, -1\}$  类标记.

如果输入数据线性可分, 决策函数为

$$D(x) = w^T g(x) + b \quad (9)$$

$g(x)$  为一非线性映射函数, 将 x (p 维) 映射到 l 维空间,  $w$  为 l 维矢量, b 为标量.

对于线性可分的数据, 决策函数满足

$$y_i (w^T g(x_i) + b) \geq 1 \quad i = 1, \dots, N \quad (10)$$

如果训练数据线性不可分, 引入松弛变量  $\xi_i (> 0)$ , 决策函数变为

$$\begin{aligned} y_i (w^T g(x_i) + b) &\geq 1 - \xi_i \quad i = 1, \dots, N \\ \xi_i &\geq 0 \quad i = 1, \dots, N \end{aligned} \quad (11)$$

一阶范数软间隔优化问题对应的拉格朗日函数为

$$\begin{aligned} m \max L = \frac{1}{2} \langle w, w \rangle + C \sum_i \xi_i - \sum_j \beta_j \xi_j \\ - \sum_i \alpha_i [y_i (\langle g(x_i), w \rangle + b) - 1 + \xi_i] \end{aligned} \quad (12)$$

式(12)中第一项控制分类器函数的复杂度(即分类函数的 VC 维), 第二项控制错误分类的经验误差, C 为常数,  $\alpha_i, \beta_i \geq 0$  对偶表示可以通过求对应于  $w, \xi, b$  偏导, 置零并带入式(12)得到.

$$m \max L = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} y_i y_j \alpha_i \alpha_j \langle g(x_i), g(x_j) \rangle \quad (13)$$

引入核函数, 上式变为

$$m \max L = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \quad (14)$$

其中  $\sum_i \alpha_i y_i = 0$  且  $0 \leq \alpha_i \leq C$

判决输出函数为

$$f(x) = \text{sign} \left( \sum_i y_i \alpha_i K(x, x_i) + b \right) \quad (15)$$

### 3.3 可拒识双层支持向量分类器判决规则

本文构造的双层可拒识支持向量分类器最终规则如下: 对于输入检测模式  $x$ , 首先利用 SVC 分类器进行分类, 如果判决为钙化点, 再利用钙化点 SVDD 数据选择器进行判决, 如果输入的钙化点样本通过两个判决器, 则判决为真正的钙化点, 否则判决为非钙化点.

### 4 钙化点粗检测

顺序滤波器可以表示为:

$$y = \frac{1}{N_p} \sum_{i=1}^{N_p} a_i \tilde{x}(i) \quad (16)$$

这里  $\tilde{x}(i), i = 1, \dots, N_p$  为窗口  $(M_1 \times M_2 = N)$  内图像灰度数据的排序结果,  $N_p \leq N$  为顺序滤波考虑的象素点, 本文关于  $a_i$  滤波器系数选择规则如下:

$$a_i = \begin{cases} 1 & i = \lfloor N/2 \rfloor, \lfloor N/2 \rfloor - 1 \\ 0 & \text{其他} \end{cases} \quad (17)$$

迭代顺序滤波的可疑目标判决函数为:

$$y_i = \begin{cases} 1, & x_o - \frac{1}{N_p} \sum_{i=1}^{N_p} a_i \tilde{x}(i) \geq th \\ 0, & x_o - \frac{1}{N_p} \sum_{i=1}^{N_p} a_i \tilde{x}(i) < th \end{cases} \quad (18)$$

这里,  $x_o$  为窗口中心点图像灰度,  $th$  为设定的门限. 迭代顺序滤波就是利用前面的处理结果分别处理后面的像素, 直到所有像素处理完毕.

对于前面检测到的可疑目标利用双参数 CFAR 检测进行二次判决, 其目的是粗检测能够适应背景的变化. 本文构造的双参数 CFAR 检测仅仅考虑两个滑动窗口: 目标窗口和背景窗口, 参见图 4 背景窗口主要用于统计背景组织、噪声的灰度分布, 其尺寸可以根据经验确定. 本文实验中背景窗口大小取 31

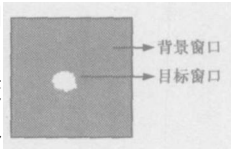


图 4 双参数 CFAR 检测窗口

设  $T$  为可疑目标窗口,  $B$  为背景窗口,  $m_T, m_B$  分别为可疑目标窗口和背景窗口的均值,  $\sigma_B^2$  为背景窗口的方差. 目标检测准则为:

$$\frac{m_T - m_B}{\sigma_B} \propto K_{cfar} \quad (19)$$

当  $\frac{m_T - m_B}{\sigma_B} > K_{cfar}$  时, 判定为目标;  $\frac{m_T - m_B}{\sigma_B} \leq K_{cfar}$  时, 判定为背景组织. 其中  $K_{cfar}$  为门限因子, 控制虚警概率. 通过上面的处理可以大大减少可疑的非钙化点样本数量.

### 5 仿真实验

为了验证本文提出的方法在乳腺钙化点自动检测中的效果, 本文选择 M I A S 数据库的钙化点进行仿真实验. 该数据库钙化区域由专家标定, 通过有经验的专家我们确定了 20 幅乳腺图像的 107 个钙化点, 利用粗检测算子收集到 201 个非钙化点, 作为训练样本. 同时我们选取 150 个钙化点可疑样本作为测试样本. 本文首先利用迭代顺序滤波器结合双参数 CFAR 检测对钙化点进行预处理作为粗检测算子, 然后对于这些可能的钙化区域分别抽取相对灰度、对比度、面积、方差、邻域可疑钙化点平均对比度、距离分布方差、面积方差 7 个特征参数组成特征矢量.

首先对 SVM 分类器进行训练得到其支持向量 (采用高斯核函数). 为了验证本文构造算法的效果, 分别将它与直接 SVM 检测方法、神经网络 (Artificial neural network ANN) 检测方法进行了对比实验, 对比结果如表 1 所示. 选择其中一幅乳腺图像, 其处理结果如图 5 所示.

### 6 结论

本文提出了基于迭代顺序滤波空间约束的可拒识-支持向量机分类器钙化点检测算法. 它将利用迭代顺序滤

波钙化点检测的粗选择环节, 主要有下面几方面的优点: (1) 它缩小了 SVM 面对的样本空间, 降低了 SVM 训练的难度. (2) 基于迭代顺序滤波约束的子空间的目标检测方法大大提高了检测速度, 同时该方法对于背景噪声的影响具有良好的适应性. 同时基于 SVDD 的球形支持向量具有很好的数据分布描述能力, 能够较好地提供每一类数据分布的紧凑描述; SVM 能够以最大间隔提供两类数据的分类鉴别能力, 其缺点是不能提供数据的分布范围描述. 本文将二者的优点相结合提出可拒识-双层支持向量分类器用于乳腺图像钙化点检测, 对比实验结果表明本文提出的方法不影响微钙化点的检出率的情况下, 由于充分利用了真实钙化点特征分布信息, 部分解决了假阳性高的问题.

表 1 几种钙化点检测方法性能对比

方法	直接 SVM	ANN	本文方法
检出率	96.2%	92.3%	96.0%
FPR	25.0%	29.0%	10.3%
检测时间 (s)	79.08	81.65	16.04

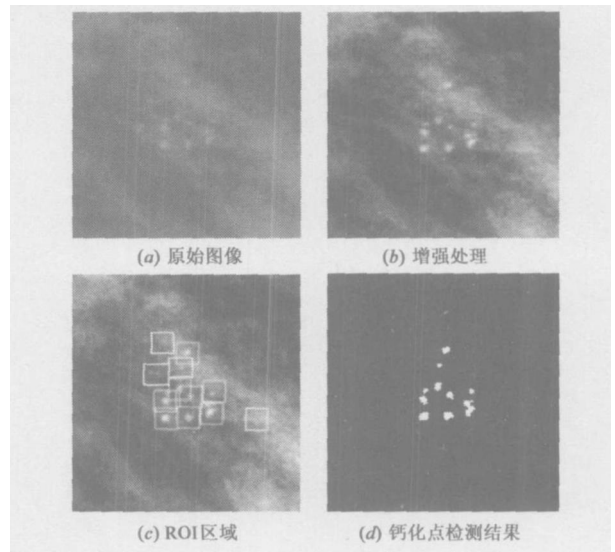


图 5 检测结果显示

### 参考文献:

[ 1 ] Vapnik. The Nature of Statistical Learning Theory [ M ]. New York, NY: Wiley, 1998: 5.

[ 2 ] Wan Baikun, Wang Rui-ping, et al. Principles of SVM and its application in micro-calcifications detection in mammogram [ J ]. Acta Electronica Sinica, 2004, 32(4): 587-590.

[ 3 ] El-Naqa Issam, Yang Yongyi, et al. A support vector machine approach for detection of microcalcifications [ J ]. IEEE Transactions on Medical Imaging, 2002, 21(12): 1552-1563.

[ 4 ] David M. J. Tax, Robert P. W. Duin. Support vector data description [ J ]. Machine Learning, 2004, 54(6): 45-66.

[ 5 ] Chao Yuan, David Casasent. A novel support vector classifier

fier with better rejection performance[ A ]. Proceeding of the IEEE computer society conference on computer vision and pattern recognition[ C]. Madison, WI United States Institute of Electrical and Electronics Engineers Computer Society, 2003. 1063- 1069

- [ 6 ] S A Rizvi T N Saasavi N M Nasabadi A clutter rejection technique for FLR imagery using region based principal component analysis[ J]. Pattern recognition, 2000, 33( 11): 1931- 1933
- [ 7 ] Yang G Z Huang T S Human face detection in a complex background[ J]. Pattern recognition, 1994, 27( 1): 58- 63

#### 作者简介:



胡正平 男, 1970年生于四川仪陇县, 哈尔滨工业大学博士研究生, 目前研究方向为统计学习理论与模式识别、医学图像处理。

E-mail: hnpochw@263.net

吴燕女, 燕山大学2003级信号信息处理专业硕士研究生, 研究方向为医学图像目标检测算法。

张晔男, 1960年生于辽宁北镇县, 教授, 哈尔滨工业大学博士生导师, 目前研究方向为遥感图像处理、模式识别。